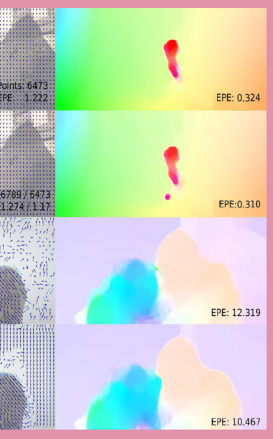
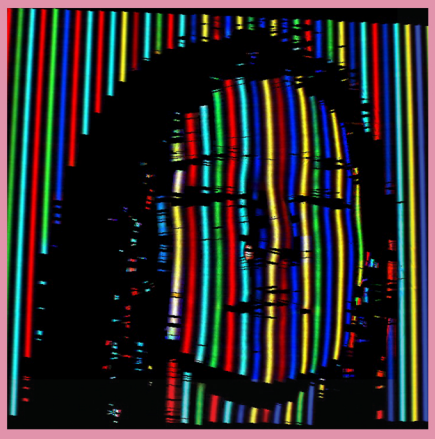
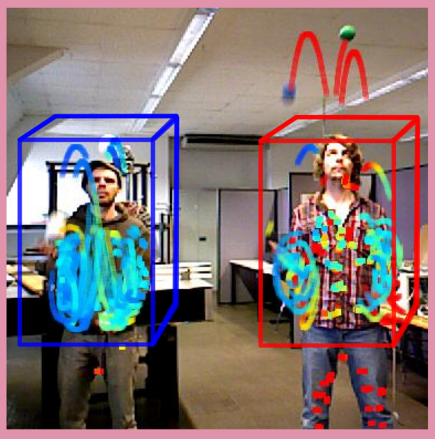


26th British Machine Vision Conference

Programme & Abstracts




Editors:
Xianghua Xie
Mark W. Jones
Gary Tam



BMVC 2015
7 - 10 SEPTEMBER, SWANSEA UK

26th British Machine Vision Conference

Swansea University
Prifysgol Abertawe

BMVA

BMVC 2015 would like to thank all our sponsors for their generous support.

Platinum

Movidius

Gold

QUALCOMM

Silver

Microsoft

ocado technology

NVIDIA

IET Journals
The Institution of Engineering and Technology

Google

Prize

WILEY

Springer

CRC Press
Taylor & Francis Group

THE COMPUTER JOURNAL 2015



bmvc2015.swan.ac.uk

BRITISH MACHINE VISION CONFERENCE 2015

Contents

Foreword	
People	
Tutorial Abstract	1
Keynote Abstracts	3
Paper Abstracts (Tuesday)	7
Paper Abstracts (Wednesday)	95
Paper Abstracts (Thursday)	184
Workshop Programmes (Thursday PM)	193
Author Index	
Venue Map & Local Information	

Download BMVC 2015 proceedings from the conference website:

bmvc2015.swansea.ac.uk

Foreword

It is our great pleasure to welcome all participants of the 26th British Machine Vision Conference (BMVC) to Swansea. Swansea University is set in a rolling parkland and enjoys a prime position overlooking Swansea Bay, the start of the famously dramatic Gower coastline.

BMVC is one of the top events in the field of Computer Vision and has always maintained a single track format. As its increasing popularity and quality, it has established as a prestigious event on the vision calendar. This year BMVC has attracted a record number of submissions. A total of 553 valid papers were received, which represents just over 25% increase from the last BMVC record achieved in 2013. The majority of the submissions are from overseas, that is 87% of the submissions, reflecting the international status of the conference.

For BMVC 2015, we have substantially expanded the international reviewer pool to 270, prior to the paper submission deadline. As a result, we were able to reduce the average paper load per reviewer to 6.3. Each paper had at least one area chair and in most cases was handled by two area chairs from our team of 54 area chairs. New to this year, we split the borderline rating to borderline accept and borderline reject to help the referees to decide which side they really wish to support. Accepted papers needed strong support from reviewers and area chairs. Papers that require extensive changes, new experiments, further comparisons to existing methods, improved clarity and other changes that would normally require a second review cycle were not accepted. We would like to thank all reviewers and area chairs for their diligent work and prompt responses.

In total, 186 papers are accepted, that is 33% of overall acceptance rate which is in line with recent BMVCs. Among those, 41 papers are selected as oral presentation, i.e. around 7% of the submitted papers that is among the lowest oral acceptance rates of BMVC. BMVC 2015 programme contains the largest number of technical papers since the BMVC inception in 1993. Using the first author's affiliation, 15% of the accepted papers are from a UK-based institute, 32% from Europe (excluding UK), 27% from Asia, 22% from North America, 3% from South America, 1% from Australia, and 1% from Africa. This year, BMVC publishes its proceedings entirely online, without the use of USB drives for environmental reasons. Three prizes will be selected and awarded at the conference. We also awarded 5 student bursaries to encourage participation, particularly where financial support is needed.

This year, we also introduced changes to the workshop programme. We have moved the student workshop from Friday to Thursday PM, which used to be a free afternoon, so that the conference is more compact. Meanwhile, we have expanded the workshop programme to allow four parallel sessions. These topics that are proposed by the organisers have enriched the conference and further encouraged participation. In total, the four workshops contain 6 keynote speeches and 33 accepted papers.

Up to the late registration deadline, BMVC 2015 has well over 300 registered delegates, which is another record in BMVC series. This year, there are strong interests from the industry to support the BMVC, with 11 sponsors providing financial and in-kind contributions. We are grateful to Movidius, Qualcomm, Microsoft, nVidia, IET, Ocado technology, Google, and other sponsors for their generous support.

We are very honoured to have three distinguished scholars as invited speakers. Dr. Andrew Fitzgibbon from Microsoft is going to deliver the conference tutorial on fitting models to data. Prof. Ron Kimmel from Israel Institute of Technology and Prof. Kristen Grauman from University of Texas at Austin are the two keynote speakers at the main conference. We are grateful to their inspiring contributions.

BMVC 2015 has been organised by the Computer Vision and Medical Image Analysis group and members of the Visual Computing group in the Department of Computer Science at the Swansea University. We are indebted to a number of people who have contributed generously to the organisation of the conference: Jingjing Deng, Sue Phillips, Mike Edwards, Julie Pellard, Robert Palmer, Jonathan Jones, David George, James Jones, Joss Whittle, Dean Thomas, and Neil Jenkinson. We thank BMVA executive committee members for their support and in particular Toby Breckon for his time in sorting out various finance matters so promptly. We also received generous support from the organisers of the previous BMVCs at Bristol University and Nottingham University, namely Dima Damen, Majid Mirmehdi, Michel Valstar, Andrew French, Susannah Lydon, and Tony Pridmore.

We sincerely hope you find BMVC 2015 intellectually stimulating, and we also wish you an enjoyable stay at Swansea.

Xianghua Xie, Mark Jones, Gary Tam

BMVC2015 Chairs

People

General Chair:

Xianghua Xie	Swansea University

Program Chair:

Mark Jones	Swansea University

Organisation Chair:

Gary Tam	Swansea University

Area Chairs:

Steven Beauchemin	Western University, Canada
Matthew Blaschko	Ecole Centrale Paris
Toby Breckon	Durham University
Mathieu Bredif	IGN
Matthew Brown	University of Bath
Tilo Burghardt	University of Bristol
Andrea Cavallaro	Queen Mary, University of London
Ondrej Chum	Czech Technical University, Prague
Adrian Clark	University of Essex
John Collomosse	University of Surrey
Tim Cootes	University of Manchester
Daniel Cremers	Technische Universitat Munchen
Dima Damen	University of Bristol
Andrew Fitzgibbon	Microsoft Research Cambridge
Shaogang Gong	Queen Mary, University of London
Lewis Griffin	University College London(UCL)
Peter Hall	University of Bath
Edwin Hancock	University of York
Richard Harvey	University of East Anglia
Adrian Hilton	University of Surrey
Herve Jegou	INRIA Rennes
Frederic Jurie	University of Caen
Fredrik Kahl	Lund University
Hongdong Li	Australian National University
Dave Marshall	Cardiff University
Stephen Maybank	Birkbeck, University of London
Stephen McKenna	University of Dundee
Majid Mirmehdi	University of Bristol
Srinivasa Narasimhan	Carnegie Mellon University
Nassir Navab	TU Munich
Matti Pietikäinen	University of Oulu
Stephen Pollard	Hewlett Packard Labs, UK
Guoping Qiu	University of Nottingham
Neil Robertson	Heriot Watt
Paul Rosin	Cardiff University
Sudeep Sarkar	University of South Florida
Ling Shao	Northumbria University
Paul Siebert	University of Glasgow
Jon Starck	The Foundry
Bjorn Stenger	Toshiba Research Europe
Manuel Trucco	University of Dundee
Kevin Wells	University of Surrey
Richard Wilson	University of York
Tao Xiang	Queen Mary, University of London
Xianghua Xie	Swansea University
Stefanos Zafeiriou	Imperial College London
Jianguo Zhang	University of Dundee
Zhao Zhang	Soochow University, China P.R.
Reyer Zwiggelaar	Aberystwyth University

Technical Committee:

Pablo Alcantarilla	Toshiba Research Europe
Jose Alvarez	National ICT Australia
Telmo Amaral	Newcastle University
Juan Andrade-Cetto	Institut de Robotica Informatica Industrial

Alexander Andreopoulos	IBM Research USA
Anelia Angelova	Google
Roland Angst	Stanford University
Gerardo Aragon-Camarasa	University of Glasgow
Relja Arandjelovic	INRIA
Ognjen Arandjelovic	Deakin University
Helder Araujo	University of Coimbra
Vasileios Argyriou	Kingston University
Antonis Argyros	University of Crete
Kalle Astroem	Lund University
Yannis Avrithis	National Technical University of Athens
Hossein Azizpour	Royal Institute of Technology, KTH
Andrew Bagdanov	Universitat Automoma de Barcelona
Xiang Bai	Huazhong University of Science and Technology
Haris Baig	Dartmouth College
Adrian Barbu	Florida State University
John Barron	University of Western Ontario Canada
Adrien Bartoli	Universite de Auvergne
Jorge Batista	Universidade de Coimbra
Rolf Baxter	Heriot Watt University
Jean-Charles Bazin	Disney Research Zurich
Loris Bazzani	Dartmouth College
Daniel Beale	University of Bath
Ardhendu Behera	Edge Hill University
Robert Bergevin	University Laval Canada
Fred Besse	The Foundry
Arnav Bhavsar	Indian Institute of Technology Mandi
Guillaume-Alexandre Bilodeau	Polytechnique Montreal Canada
Thomas Bishop	Heriot-Watt University
Arijit Biswas	University of Maryland
Miroslaw Bober	University of Surrey
Adrian Bors	University of York
Pierre Boulanger	University of Alberta Canada
Steven Branson	California Institute of Technology
Ernesto Brau	The University of Arizona
Michael Brown	National University of Singapore
Darius Burschka	TU Munchen
Martin Cadik	Brno University of Technology
Janko Calic	University of Surrey
Teofilo de Campos	University of Surrey
Barbara Caputo	IDIAP Research Institute
Joao Carreira	UC Berkeley
Jan Cech	Czech Technical University
Tat-Jen Cham	Nanyang Technological University
Guillaume Charpiat	INRIA
Rama Chellappa	University of Maryland
Chao Chen	Rutgers University
Xilin Chen	Institute of Computing Technology, Chinese Academy of Sciences
Xin Chen	University of Manchester
Wen-Huang Cheng	Academia Sinica
John Chiverton	Portsmouth University
Margarita Chli	University of Edinburgh
Minsu Cho	INRIA
Nam Ik Cho	Seoul National University
Sunghyun Cho	Samsung Electronics
Yung-Yu Chuang	National Taiwan University
Javier Civera	University of Zaragova
Nicholas Costen	Manchester Metropolitan University
Fabio Cuzzolin	Oxford Brookes University
Roy Davies	Royal Holloway, University of London
Andrew Davison	Imperial College London
Salil Deena	University of Sheffield
Jingjing Deng	Swansea University
Joachim Denzler	University of Jena
Konstantinos Derpanis	Ryerson University
Michael Donoser	TU Graz
Gianfranco Doretto	West Virginia University

Bruce Draper	Colorado State University
Kun Duan	University of Indiana
Lixin Duan	Institute for Infocomm Research
Anjan Dutta	Telecom ParisTech
Pinar Duygulub	Bilkent University
Khemraj Emrith	University of the West of England
Sergio Escalera	University of Barcelona
Ehab Essa	Swansea University
Francisco Estrada	University of Toronto
Giovanni Farinella	University of Catania
Paolo Favaro	Uinversity of Bern
Luis Ferraz	Universitat Pompeu Fabra
Corneliu Florea	University Politehnica of Bucharest
Francisco Florez	Kingston University
Per-Erik Forssen	Linkoping University
Charless Fowlkes	University of California, Irvine
Jan-Michael Frahm	University of North Carolina at Chapel Hill
Roberto Fraile	University of Hamburg
Andrew French	University of Nottingham
Mario Fritz	Max-Planck-Institut fur Informatik
Theo Gevers	University of Amsterdam
Simon Hadfield	University of Surrey
Junwei Han	University of Dundee
Allan Hanbury	Vienna University of Technology
Jari Hannuksela	University of Oulu
Sion Hannuna	University of Bristol
Michal Havlena	ETH Zurich
Shengfeng He	City University of Hong Kong
Xuming He	National ICT Australia
Janne Heikkil	University of Oulu
Edmond Ho	Hongkong Baptist University
Richang Hong	Hefei Technical University
Tim Hospedales	Queen Mary University
Gang Hua	Stevens Institute of Technology
Kaiqi Huang	Institute of Automation, Chinese Academy of Sciences
Mohamed Hussein	Alexandria University
Evren Imre	University of Surrey
Francesco Isgro	UNINA
Stuart James	University of Surrey
CV Jawahar	International Institute of Information Technology, Hyderabad
Hao Jiang	Boston College
Olaf Kahler	University of Oxford
George Kamberov	Stevens Institute of Technology
Meina Kan	Institute of Computing Technology, Chinese Academy of Sciences
Kenichi Kanatani	Okayama University
Konstantinos Karantzalos	NTUA
Vivek Kaul	Facebook
Edward Kim	Villanova University
Hansung Kim	University of Surrey
Dmitry Kit	University of Bath
Laurent Kneip	Australian National University
Piotr Koniusz	University of Surrey
Rigas Kouskouridas	Imperial College
Matej Kristan	University of Ljubljana
Roland Kwitt	University of Salzbury
Jan van Gemert	University of Amsterdam
Andrea Giachetti	Verona University
Philippe Giguere	Universite Laval
Arjan Gijsenij	AkzoNobel
Andrew Gilbert	University of Surrey
Monica Hernandez Gimenez	University of Zaragoza
Roland Goecke	University of Canberra
Albert Gordo	Xerox XRCE
Philippe-Henri Gosselin	ENSEA
Jean-Yves Guillemaut	University of Surrey
Hatice Gunes	Queen Mary, University of London
Frederic Labrosse	Aberystwyth University

Yukun Lai	Cardiff University
Jochen Lang	University of Ottawa
Cathy Laporte	Ecole de Technologie Superieure
Marius Leordeanu	Institute of Mathematics of the Romanian Academy
Vincent Lepetit	Ecole Polytechnique Federale de Lausanne
Frederic Lerasle	LAAS CNRS
Yongmin Li	University of Brunel
Joerg Liebelt	Airbus Group Innovations
James Little	The University of British Columbia
Jundong Liu	Ohio University
Yonghuai Liu	Aberystwyth University
Chen Change Loy	The Chinese University of Hong Kong
Enrique Sanchez Lozano	University of Nottingham
Imanol Luengo	University of Nottingham
John MacCormick	Dickinson College
Stefan Mairhofer	University of Nottingham
Yasushi Makihara	Osaka University
Ioannis Marras	Imperial College
Brais Martinez	Imperial College
Hongying Meng	University of Brunel
John McGonigle	Imperial College London
Gregor Miller	The University of British Columbia
Phiippos Mordohai	Stevens Institute of Technology
Greg Mori	Simon Fraser University
Jan Neumann	Comcast Labs
Mihalis Nicolaou	Imperial College London
Athanasios Nikolaidis	Technological Educational Institute of Central Macedonia
Mark Nixon	University of Southampton
Takayuki Okatani	Tohoku University
Soren Ingvor Olsen	University of Copenhagen
Eng-Jon Ong	University of Surrey
James Orwell	Kingston University
Alain Pagani	Intelligente Losungen fur die Wissensgesellschaft
Marco Paladini	Ocado Ltd
Ioannis Patras	Queen Mary, University of London
Luca del Pero	University of Edinburgh
Caroline Petitjean	Universite de Rouen
Stavros Petridis	Imperial College
Vladimir Petrovic	University of Manchester
Michael Pound	University of Nottingham
Victor Prisacariu	Oxford University
Nicolas Pugeault	University of Surrey
Srikumar Ramalingam	Mitsubishi Electric Research Lab
Narayanan Ramanathan	ObjectVideo
Paolo Remagnino	Kingston University
Carl Ren	University of Oxford
Emma Robinson	Imperial College
Emanuele Rodola	Technische Universitut Munchen
Arun Ross	Michigan State University
Peter Roth	Graz University of Technology
Anastasios Roussos	Queen Mary, University of London
Michel Roux	Telecom ParisTech
Sebastien Roy	Universite de Montreal
Yassine Ruichek	Universite de Technologie de Belfort-Montbeliard
Chris Russell	University College London
Albert Salah	Bogazici University
Jorge Sanchez	Universidad Nacional de Cordoba
Christian Scharfenberger	University of Waterloo Canada
Walter Scheirer	Harvard University
Frank Schmidt	University of Freiburg
Oliver Schreer	Fraunhofer
Falk Schubert	EADS Innovations Works
Samuel Schulter	TU Graz
Mikhail Sizintsev	York University, Canada
Hubert Shum	Northumbria University
Ran Song	Brighton University
Yizhe Song	Queen Mary, University of London

Hang Su	Carnegie Mellon University
Xianfang Sun	Cardiff University
Zhenan Sun	Institute of Automation, Chinese Academy of Sciences
Maria Jose Garcia Sopo	Holition
Sebastian Stein	University of Dundee
Harry Strange	Aberyswyth University
Yusuke Sugano	Max-Planck Institut fur Informatik
Deqing Sun	Harvard University
David Suter	University of Adelaide
Yuichi Taguchi	Mitsuishi Electric Research Lab
Gary Tam	Swansea University
Jun Tang	Anhui University
Barry Theobald	University of East Anglia
Radu Timofte	ETH Zurich
Bernard Tiddeman	Aberystwyth University
Akihiko Torii	Tokyo Institute of Technology
Yan Tong	University of South Carolina
Du Tran	Dartmouth College
Carole Twining	University of Manchester
Georgios Tzimiropoulos	University of Lincoln
Michel Valstar	University of Nottingham
Ashok Veeraraghavan	Rice University
Franck P. Vidal	Bangor University
Fabio Viola	The Foundry
Bo Wang	University of UTAH
Jingdong Wang	Microsoft Research Asia
Meng Wang	Hefei University of Technology
Qi Wang	Northwestern Polytechnical University
Roy Wang	University of Dundee
Wenwu Wang	University of Surrey
Xiaomeng Wang	University of Nottingham
Yang Wang	University of Manitoba Canada
Zhenhua Wang	Nanyang Technological University
Martin Weinmann	Karlsruher Institut fur Technologie
Andreas Wendel	Google Inc.
Aaron Wetzler	Technion
Alexander Wong	University of Waterloo Canada
Song Wu	Leiden University
Yi Wu	Nanjing University of Information Science & Technology
Yijun Michael Xiao	University of Glasgow
Xianghua Xie	Swansea University
Chenliang Xu	University of Michigan
Fei Yan	University of Surrey
Kuiyuan Yang	Microsoft Asia
Alper Yilmaz	Ohio State University
Qizhi Yu	University College London
Luca Zapella	Johns Hopkins University
John Zelek	University of Waterloo Canada
Hui Zhang	Hong Kong Baptist University
Jiaxiang Zhang	Corning
Huaizhong Zhang	University of Manchester
Weishi Zheng	Queen Mary, University of London
Huiyun Zhou	Queens University Belfast
Xiatian Zhu	Queen Mary University of London
Zeeshan Zia	Imperial College
Vasileios Zografos	Linkoping University

Additional Reviewers:

Hui Fang	Oxford University
Ioannis Kaloskampisl	Cardiff University
Jonathan Jones	Swansea University
Jing Wu	Cardiff University

Website Chair:

Jingjing Deng	Swansea University

Programme: Monday 7th September

Tutorial

13:00-14:30 Fitting models to data: Accuracy, Speed, Robustness 1
Andrew Fitzgibbon

14:30-15:00 Coffee

Tutorial

15:00-16:30 Fitting models to data: Accuracy, Speed, Robustness 1
Andrew Fitzgibbon

Welcome Reception (National Waterfront Museum)

19:00-21:00 The National Waterfront Museum at Swansea tells the story of industry and innovation in Wales, now and over the last 300 years. It is located at the Swansea Marina.



Programme: Tuesday 8th September

08:15 Registration (until 17:45)

08:45-09:00 Welcome

Keynote

09:00-10:00 A Spectral Perspective on Shapes 3
Ron Kimmel

Tracking

10:00-10:20 Online Domain Adaptation for Multi-Object Tracking 7
Adrien Gaidon, Eleonora Vig

10:20-10:40 Learning Optimal Parameters For Multi-target Tracking 8
Shaofei Wang, Charless Fowlkes

10:40-11:00 Coffee

Deep Learning

11:00-11:20 R-CNN minus R 9
Karel Lenc, Andrea Vedaldi

11:20-11:40 Convolutional Neural Networks for Direct Text Deblurring 10
Michal Hradiš, Jan Kotera, Pavel Zemčík, Filip Šroubek

11:40-12:00 Sketch-a-Net that Beats Humans 11
Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, Timothy Hospedales

12:00-12:20 Learning Deep Representations of Appearance and Motion for Anomalous Event Detection 12
Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, Nicu Sebe

12:20-12:40 Deep Perceptual Mapping for Thermal to Visible Face Recogntion 13
M. Saquib Sarfraz, Rainer Stiefelhagen

12:40-13:20 Lunch

Poster 1

13:20-14:20 An Efficient Algorithm for Learning Distances that Obey the Triangle Inequality 14
Arijit Biswas, David Jacobs

Diagnosing state-of-the-art object proposal methods 15
Hongyuan Zhu, Shijian Lu, Jianfei Cai, Guangqing Lee

WxBS: Wide Baseline Stereo Generalizations 16
Dmytro Mishkin, Jiri Matas, Michal Perdoch, Karel Lenc

Planar shape decomposition made simple 17
Nikos Papanelopoulos, Yannis Avrithis

Adaptation of Synthetic Data for Coarse-to-Fine Viewpoint Refinement 18
Pau Panareda Busto, Joerg Liebelt, Juergen Gall

Saliency Prediction with Active Semantic Segmentation 19
Ming Jiang, Xavier Boix, Juan Xu, Gemma Roig, Luc Van Gool, Qi Zhao

Using Near-Field Light Sources to Separate Illumination from BRDF 20
Jeroen Put, Nick Michiels, Philippe Bekaert

Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data 21
Shadi Albarqouni, Maximilian Baust, Sailesh Conjeti, Asharf Al-Amoudi, Nassir Navab

Spotlight the Negatives: A Generalized Discriminative Latent Model 22
Hossein Azizpour, Mostafa Arefiyan, Sobhan Naderi Parizi, Stefan Carlsson

Perceptual Dynamic Range for In-Camera Image Processing 23
Praveen Cyriac, David Kane, Marcelo Bertalmio

Robust Multiple Model Fitting with Preference Analysis and Low-rank Approximation 24
Luca Magri, Andrea Fusiello

Robust Global Motion Compensation in Presence of Predominant Foreground 25
Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa

Using Segmentation to Predict the Absence of Occluded Parts 26
Golnaz Ghiasi, Charless Fowlkes

Learning the Structure of Deep Architectures Using L1 Regularization 27
Praveen Kulkarni, Joaquin Zepeda, Frederic Jurie, Patrick Pérez, Louis Chevallier
Subspace Distribution Alignment for Unsupervised Domain Adaptation 28
Baochen Sun, Kate Saenko
Robust Spatial Matching as Ensemble of Weak Geometric Relations 29
Xiaomeng Wu, Kunio Kashino
Kernelized View Adaptive Subspace Learning for Person Re-identification 30
Qin Zhou, Shibao Zheng, Hang Su, Hua Yang, Yu Wang, Shuang Wu
Object localization in ImageNet by looking out of the window 31
Alexander Vezhnevets, Vittorio Ferrari
Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection 32
Huamin Ren, Weifeng Liu, Søren Ingvor Olsen, Sergio Escalera, Thomas B. Moeslund
Joint Calibration for Semantic Segmentation 33
Holger Caesar, Jasper Uijlings, Vittorio Ferrari
Camera Elevation Estimation from a Single Mountain Landscape Photograph 34
Martin Čadík, Jan Vašíček, Michal Hradiš, Filip Radenović, Ondřej Chum
Data-free Parameter Pruning for Deep Neural Networks 35
Suraj Srinivas, R. Venkatesh Babu
Real-Time Pedestrian Detection with Deep Network Cascades 36
Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, Dave Ferguson
Rule of thumb: Deep derotation for improved fingertip detection 37
Aaron Wetzler, Ron Slossberg, Ron Kimmel
Riesz-based Volume Local Binary Pattern and A Novel Group Expression Model for Group Happiness Intensity Analysis 38
Xiaohua Huang, Abhinav Dhall, Guoying Zhao, Roland Goecke, Matti Pietikäinen
Weakly Supervised Metric Learning towards Signer Adaptation for Sign Language Recognition 39
Fang Yin, Xiujuan Chai, Yu Zhou, Xilin Chen
Hashmod: A Hashing Method for Scalable 3D Object Detection 40
Wadim Kehl, Federico Tombari, Nassir Navab, Slobodan Ilic, Vincent Lepetit
Multi-Task Transfer Methods to Improve One-Shot Learning for Multimedia Event Detection 41
Wang Yan, Jordan Yap, Greg Mori
Appearance and Depth for Rapid Human Activity Recognition in Real Applications 42
Stavros Tachos, Konstantinos Avgerinakis, Alexia Briasouli, Ioannis Kompatsiaris
Robust Wearable Camera Localization as a Target Tracking Problem on SE(3) 43
Guillaume Bourmaud, Audrey Giremus
Joint Object-Material Category Segmentation from Audio-Visual Cues 44
Anurag Arnab, Michael Sapienza, Stuart Golodetz, Julien Valentin, Ondrej Miksik, Shahram Izadi, Philip H. S. Torr
Deep Face Recognition 45
Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman
Combinatorial Regularization of Descriptor Matching for Optical Flow Estimation 46
Benjamin Drayer, Thomas Brox
Discrete Light Source Estimation from Light Probes for Photorealistic Rendering 47
Farshad Einabadi, Oliver Grau
Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification 48
Elyor Kodirov, Tao Xiang, Shaogang Gong
Camera Pose Estimation from Lines using Plücker Coordinates 49
Bronislav Přibyl, Pavel Zemčík, Martin Čadik
Linear Global Translation Estimation with Feature Tracks 50
Zhaopeng Cui, Nianjuan Jiang, Chengzhou Tang, Ping Tan
Learning Discriminative Visual N-grams from Mid-level Image Features 51
Raj Kumar Gupta, Megha Pandey, Alex YS Chia
Minimizing the Number of Keypoint Matching Queries for Object Retrieval 52
Johannes Niedermayer, Peer Kröger
JH2R: Joint Homography Estimation for Highlight Removal 53
Sungmin Eum, Hyungtae Lee, David Doermann
Action Recognition based on Subdivision-Fusion Model 54
Zongbo Hao, Linlin Lu, Qianni Zhang, Jie Wu, Ebroul Izquierdo, Juanyu Yang, Jun Zhao
Mix and Match: Joint Model for Clothing and Attribute Recognition 55
Kota Yamaguchi, Takayuki Okatani, Kyoko Sudo, Kazuhiko Murasaki, Yukinobu Taniguchi
Describing Common Human Visual Actions in Images 56
Matteo Ruggero Ronchi, Pietro Perona
Searching for Objects using Structure in Indoor Scenes 57
Varun K. Nagaraja, Vlad I. Morariu, Larry S. Davis

Fast Inverse Compositional Image Alignment with Missing Data and Re-weighting 58
Vincent Lui, Dinesh Gamage, Tom Drummond
Automatic Age Estimation from Face Images via Deep Ranking 59
Huei-Fang Yang, Bo-Yao Lin, Kuang-Yu Chang, Chu-Song Chen
Feature Encoding of Spectral Signatures for 3D Non-Rigid Shape Retrieval 60
Frederico A. Limberger, Richard C. Wilson
A BoW-equivalent Recurrent Neural Network for Action Recognition 61
Alexander Richard, Juergen Gall
Depth Restoration via Joint Training of a Global Regression Model and CNNs 62
Gernot Riegler, René Ranftl, Matthias Rüther, Thomas Pock, Horst Bischof
Shape Detection with Nearest Neighbour Contour Fragments 63
Kasim Terzić, Hussein Adnan Mohammed, J.M.H. du Buf
Exploiting Image-trained CNN Architectures for Unconstrained Video Classification 64
Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, Ruslan Salakhutdinov
Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation 65
Ken Sakurada, Takayuki Okatani
Robust Multiview Registration of 3D Surfaces via ℓ_1-norm Minimization 66
Anil C. Raghuramu
Because better detections are still possible: Multi-aspect Object Detection with Boosted Hough Forest 67
Carolina Redondo-Cabrera, Roberto López-Sastre
Optimizing Partition Trees for Multi-Object Segmentation with Shape Prior 68
Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat
Face Painting: querying art with photos 69
Elliot J. Crowley, Omkar M. Parkhi, Andrew Zisserman
Cross-Domain Object Recognition Using Object Alignment 70
Pengcheng Liu, Chong Wang, Peipei Yang, Kaiqi Huang, Tieniu Tan
Dynamical Regularity for Action Analysis 71
Vinay Venkataraman, Ioannis Vlachos, Pavan Turaga
Efficient Spatio-Temporal Data Association Using Multidimensional Assignment in Multi-Camera Multi-Target Tracking 72
Moonsub Byeon, Songhwai Oh, Kikyung Kim, Haan-Ju Yoo, Jin Young Choi
A New Face Recognition Algorithm based on Dictionary Learning for a Single Training Sample per Person 73
Yang Liu, Ian Wassell
Robust Direct Visual Localisation using Normalised Information Distance 74
Geoffrey Pascoe, Will Maddern, Paul Newman
Joint Clustering and Classification for Multiple Instance Learning 75
Karan Sikka, Ritwik Giri, Marian Bartlett
Multi-Modality Feature Transform: An Interactive Image Segmentation Approach 76
Moustafa Meshry, Ahmed Taha, Marwan Torki
Multi-Shot Human Re-Identification Using Adaptive Fisher Discriminant Analysis 77
Yang Li, Ziyan Wu, Srikrishna Karanam, Richard J. Radke
Hierarchical Hybrid Shape Representation for Medical Shapes 78
Abhishek Kolagunda, Guoyu Lu, Chandra Kambhamettu
Data Separation of L1-minimization for Real-time Motion Detection 79
Yu Liu, Huaxin Xiao, Zheng Zhang, Wei Xu, Maojun Zhang, Jianguo Zhang
Color Constancy by Deep Learning 80
Zhongyu Lou, Theo Gevers, Ninghang Hu, Marcel P. Lucassen
Model-based 3D Hand Tracking with on-line Shape Adaptation 81
Alexandros Makris, Antonis Argyros
Globally Optimal DLS Method for PnP Problem with Cayley parameterization 82
Gaku Nakano
Joint Tracking and Event Analysis for Carried Object Detection 83
Aryana Tavanai, Muralikrishna Sridhar, Eris Chinellato, Anthony G. Cohn, David C. Hogg
Occlusion-Aware Object Localization, Segmentation and Pose Estimation 84
Samarth Brahmbhatt, Heni Ben Amor, Henrik Christensen
VESICLE: Volumetric Evaluation of Synaptic Inferfaces using Computer Vision at Large Scale 85
William Gray Roncal, Michael Pekala, Verena Kaynig-Fittkau, Dean M Kleissas, Joshua T Vogelstein, Hanspeter Pfister, Randal Burns, R Jacob Vogelstein, Mark A Chevillet, Gregory D Hager

Medical Applications

14:20-14:40 Epipolar Consistency in Fluoroscopy for Image-Based Tracking 86
André Aichert, Jian Wang, Roman Schaffert, Arnd Dörfler, Joachim Hornegger, Andreas Maier

14:40-15:00 Automatic Aortic Root Segmentation with Shape Constraints and Mesh Regularisation 87
Robert Ieuan Palmer, Xianghua Xie, Gary Tam
15:00-15:20 Anatomical triangulation: from sparse landmarks to dense annotation of the skeleton in CT images 88
Marie Bieth, Rene Donner, Georg Langs, Markus Schwaiger, Bjoern Menze
15:20-15:40 Primal-Dual convex optimization in large deformation diffeomorphic registration with robust regularizers 89
Monica Hernandez
15:40-16:00 Global Minimum for Curvature Penalized Minimal Path Method 90
Da Chen, Jean-Marie Mirebeau, Laurent D. Cohen

16:00-16:20 Coffee

3D Vision

16:20-16:40 Adaptive Contour Fitting for Pose-Invariant 3D Face Shape Reconstruction 91
Chengchao Qu, Eduardo Monari, Tobias Schuchert, Jürgen Beyerer
16:40-17:00 Freehand Laser Scanning Using Mobile Phone 92
Ron Slossberg, Aaron Wetzler, Ron Kimmel
17:00-17:20 Stochastic visibility in point-sampled scenes 93
Miles Hansard
17:20-17:40 MGM: A Significantly More Global Matching for Stereovision 94
Gabriele Facciolo, Carlo de Franchis, Enric Meinhardt

Programme: Wednesday 9th September

08:15 Registration (until 17:45)

Keynote

09:00-10:00 Action and Attention in First-person Vision 5
Kristen Grauman

Identification and Scene Annotation

10:00-10:20 Automated Identification of Individual Great White Sharks from Unrestricted Fin Imagery 95
Benjamin Hughes, Tilo Burghardt

10:20-10:40 Generating Multi-sentence Natural Language Descriptions of Indoor Scenes 96
Dahua Lin, Sanja Fidler, Chen Kong, Raquel Urtasun

10:40-11:00 Coffee

Learning and Recognition

11:00-11:20 Overlapping Domain Cover for Scalable and Accurate Regression Kernel Machines 97
Mohamed Elhoseiny, Ahmed Elgammal

11:20-11:40 Visual Comparison of Images Using Multiple Kernel Learning for Ranking 98
Amr Sharaf, Mohamed E. Hussein, Mohamed A. Ismail

11:40-12:00 Bridging the Domain Shift by Domain Adaptive Dictionary Learning 99
Hongyu Xu, Jingjing Zheng, Rama Chellappa

12:00-12:20 Joint Feature Selection with Low-rank Dictionary Learning 100
Homa Foroughi, Moein Shakeri, Nilanjan Ray, Hong Zhang

12:20-12:40 Scalable Visual Instance Mining with Instance Graph 101
Wei Li, Changhu Wang, Lei Zhang, Yong Rui, Bo Zhang

12:40-13:20 Lunch

Poster 2

13:20-14:20 Perceptually Motivated Benchmark for Video Matting 102
Mikhail Erofeev, Yury Gitman, Dmitriy Vatolin, Alexey Fedorov, Jue Wang

SDICP: Semi-Dense Tracking based on Iterative Closest Points 103
Laurent Kneip, Zhou Yi, Hongdong Li

Human Pose as Context for Object Detection 104
Abhilash Srikantha, Juergen Gall

Learning Depth Calibration of Time-of-Flight Cameras 105
David Ferstl, Christian Reinbacher, Gernot Riegler, Matthias Rüther, Horst Bischof

Low-Rank Spatio-Temporal Video Segmentation 106
Alasdair Newson, Mariano Tepper, Guillermo Sapiro

Fast Online Upper Body Pose Estimation from Video 107
Ming-Ching Chang, Honggang Qi, Xin Wang, Hong Cheng, Siwei Lyu

Simultaneous Inpainting and Super-resolution Using Self-learning 108
Milind G. Padalkar, Manjunath V. Joshi, Nilay Khatri

Manitest: Are classifiers really invariant? 109
Alhussein Fawzi, Pascal Frossard

MCSLAM : a Multiple Constrained SLAM 110
Datta Ramadasan, Marc Chevaldonné, Thierry Chateau

Incremental Dictionary Learning for Unsupervised Domain Adaptation 111
Boyu Lu, Rama Chellappa, Nasser M. Nasrabadi

StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation 112
Dan Levi, Noa Garnett, Ethan Fetaya

Segmenting natural images with the least effort as humans 113
Qiyang Zhao

Deep Fishing: Gradient Features from Deep Nets 114
Albert Gordo, Adrien Gaidon, Florent Perronnin

Multimodal Stereo Vision For Reconstruction In The Presence Of Reflection 115
Scott Sorensen, Philip Saponaro, Stephen Rhein, Chandra Kambhamettu

Real-time Human Detection based on Personness Estimation 116
Kyuwon Kim, Kwanghoon Sohn
Semantic Classification of Boundaries of an RGBD Image 117
Nishit Soni, Anoop M. Namboodiri, CV Jawahar, Srikumar Ramalingam
Elliptical ASIFT Agglomeration in Class Prototype for Logo Detection 118
Raluca Boia, Corneliu Florea, Laura Florea
Entire Reflective Object Surface Structure Understanding 119
Qinglin Lu, Olivier Laligant, Eric Fauvet, Anastasia Zakharova
TennisVid2Text: Fine-grained Descriptions for Domain Specific Videos 120
Mohak Sukhwani, CV Jawahar
Human activity recognition in the semantic simplex of elementary actions 121
Beaudry Cyrille, Péteri Renaud, Mascarilla Laurent
Character Identification in TV-series via Non-local Cost Aggregation 122
Ching-Hui Chen, Rama Chellappa
Prototypical Priors: From Improving Classification to Zero-Shot Learning 123
Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, Philip Torr
Fast Affine Template Matching over Galois Field 124
Chao Zhang, Takuya Akashi
Universal Hough dictionaries for object tracking 125
Fausto Milletari, Wadim Kehl, Federico Tombari, Slobodan Ilic, Seyed-Ahmad Ahmadi, Nassir Navab
3D Tracking of Human Hands in Interaction with Unknown Objects 126
Paschalis Panteleris, Nikolaos Kyriazis, Antonis A. Argyros
Handling Data Imbalance in Automatic Facial Action Intensity Estimation 127
Philipp Werner, Frerk Saxen, Ayoub Al-Hamadi
Hierarchical rank-based veiling light estimation for underwater dehazing 128
Simon Emberton, Lars Chittka, Andrea Cavallaro
The intrinsic error of exposure fusion for HDR imaging, and a way to reduce it 129
Raquel Gil Rodríguez, Javier Vazquez-Corral, Marcelo Bertalmío
Detecting Change for Multi-View, Long-Term Surface Inspection 130
Simon Stent, Riccardo Gherardi, Björn Stenger, Roberto Cipolla
Part Localization using Multi-Proposal Consensus for Fine-Grained Categorization 131
Kevin J. Shih, Arun Mallya, Saurabh Singh, Derek Hoiem
Gesture and Action Recognition by Evolved Dynamic Subgestures 132
Víctor Ponce-López, Hugo Jair Escalante, Sergio Escalera, Xavier Baró
Face Alignment Assisted by Head Pose Estimation 133
Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, Peter Robinson
Score Normalization in Multimodal Systems using Generalized Extreme Value Distribution 134
Renu Sharma, Sukhendu Das, Padmaja Joshi
Manifold-Regularized Selectable Factor Extraction for Semi-supervised Image Classification 135
Xin Shi, Chao Zhang, Fangyun Wei, Hongyang Zhang, Yiyuan She
Leaf Segmentation under Loosely Controlled Conditions 136
Simone Buoncompagni, Dario Maio, Vincent Lepetit
Shape from Focus with Adaptive Focus Measure and High Order Derivatives 137
Yuval Frommer, Rami Ben-Ari, Nahum Kiryati
A Segmentation Technique for Flexible Pipes in Deep Underwater Environments 138
Saulo Pessoa, Vinicius Cesar, Bernardo Reis, Judith Kelner, Ismael Santos
Structural Symmetries from Motion for Scene Reconstruction and Understanding 139
Natesh Srinivasan, Luca Carlone, Frank Dellaert
Stereo Tracking and 3D Reconstruction of Underwater Pipes 140
Vinicius Cesar, Bernardo Reis, Saulo Pessoa, Judith Kelner, Ismael Santos
Nonlinear Metric Learning for Alzheimer's Disease Diagnosis with Integration of Longitudinal Neuroimaging Features 141
Bibo Shi, Yani Chen, Kevin Hobbs, Charles D. Smith, Jundong Liu
Boosted Metric Learning for Efficient Identity-Based Face Retrieval 142
Romain Negrel, Alexis Lechervy, Frederic Jurie
Enforcing Point-wise Priors on Binary Segmentation 143
Feng Li, Fatih Porikli
Sparse Discrimination based Multiset Canonical Correlation Analysis for Multi-Feature Fusion and Recognition 144
Hongkun Ji, Xiaobo Shen, Quansen Sun, Zexuan Ji
Towards 4D Coupled Models of Conversational Facial Expression Interactions 145
Jason Vandeventer, Lukas Gräser, Magdalena Rychlowska, Paul L. Rosin, Dave Marshall
A Unified Bayesian Approach to Multi-Frame Super-Resolution and Single-Image Upsampling in Multi-Sensor Imaging 146
Thomas Köhler, Johannes Jordan, Andreas Maier, Joachim Hornegger

Boosting the Performance of Model-based 3D Tracking by Employing Low Level Motion Cues 147
Ammar Qammaz, Nikolaos Kyriazis, Antonis A. Argyros
Real-time RGB-D Tracking with Depth Scaling Kernelised Correlation Filters and Occlusion Handling 148
Massimo Camplani, Sion Hannuna, Majid Mirmehdi, Dima Damen, Adeline Paiement, Lili Tao, Tilo Burghardt
Experimental Evaluation of the Bag-of-Features Model for Unsupervised Learning of Images 149
Mariana Afonso, Luis F. Teixeira
Indoor Localisation with Regression Networks and Place Cell Models 150
Jose Rivera-Rubio, Ioannis Alexiou, Anil A. Bharath
Kinship Verification with Deep Convolutional Neural Networks 151
Kaihao Zhang, Yongzhen Huang, Chunfeng Song, Hong Wu, Liang Wang
Exploiting Low-rank Structure for Discriminative Sub-categorization 152
Zheng Xu, Xue Li, Kuiyuan Yang, Thomas Goldstein
Sparse 3D convolutional neural networks 153
Ben Graham
Multi-task Gaussian Process Regression-based Image Super Resolution 154
Xinwei Jiang, Jie Yang, Lei Ma, Yiping Yang
Spatiotemporal Stereo Matching with 3D Disparity Profiles 155
Yongho Shin, Kuk-Jin Yoon
Randomized Global Transformation Approach for Dense Correspondence 156
Kihong Park, Seungryong Kim, Seungchul Ryu, Kwanghoon Sohn
Achieving Turbidity Robustness on Underwater Images Local Feature Detection 157
Felipe Codevilla, Joel De O. Gaya, Nelson Duarte Filho, Silvia S. C. Costa Botelho
The Self-Equalizing De Bruijn Sequence for 3D Profilometry 158
Tomislav Petković, Tomislav Pribanić, Matea Đonlić
A fast and robust ellipse detector based on top-down least-square fitting 159
Yongtao Wang, Zheqi He, Xicheng Liu, Zhi Tang, Luyuan Li
Multiview Reconstruction of Complex Organic Shapes 160
Jasenko Zivanov, Thomas Vetter
Manifold Regularized Transfer Distance Metric Learning 161
Haibo Shi, Yong Luo, Chao Xu, Yonggang Wen
Top-down saliency with Locality-constrained Contextual Sparse Coding 162
Hisham Cholakkal, Deepu Rajan, Jubin Johnson
Very Efficient Training of Convolutional Neural Networks using Fast Fourier Transform and Overlap-and-Add 163
Tyler Highlander, Andres Rodriguez
Deep Q-learning for Active Recognition of GERMS: Baseline performance on a standardized dataset for active learning 164
Mohsen Malmir, Karan Sikka, Deborah Forster, Javier Movellan, Garison Cottrell
Camera Pose and Focal Length Estimation Using Regularized Distance Constraints 165
Ekaterina Kanaeva, Lev Gurevich, Alexander Vakhitov
Classifying Global Scene Context for On-line Multiple Tracker Selection 166
Salma Moujtahid, Stefan Duffner, Atilla Baskurt
Sketch based Image Retrieval using Learned KeyShapes (LKS) 167
Jose M. Saavedra, Juan Manuel Barrios
Online Visual Tracking via Coupled Object-Context Dictionary 168
Mingquan Ye, Hong Chang, Xilin Chen
Subspace Alignment Based Domain Adaptation for RCNN Detector 169
Anant Raj, Vinay P. Namboodiri, Tinne Tuytelaars
Time-slice Prediction of Dyadic Human Activities 170
Maryam Ziaeefard, Robert Bergevin, Louis-Philippe Morency
Adapting RANSAC SVM to Detect Outliers for Robust Classification 171
Subhabrata Debnath, Anjan Banerjee, Vinay Namboodiri
Beyond MSER: Maximally Stable Regions using Tree of Shapes 172
Petra Bosilj, Ewa Kijak, Sébastien Lefèvre
Exploring Locally Rigid Discriminative Patches for Learning Relative Attributes 173
Yashaswi Verma, CV Jawahar
Semantic description of medical image findings: structured learning approach 174
Pavel Kisilev, Eugene Walach, Sharbell Hashoul, Ella Barkan, Boaz Ophir, Sharon Alpert

Search and Detection

14:20-14:40 Latent Structure Preserving Hashing 175
Ziyun Cai, Li Liu, Mengyang Yu, Ling Shao

14:40-15:00 Local Feature Binary Coding for Approximate Nearest Neighbor Search 176
Li Liu, Mengyang Yu, Ling Shao

15:00-15:20 Revealing Smooth Structure of Visual Data by Permutation on Manifolds 177
Yi-Lei Chen, Chiou-Ting Hsu

15:20-15:40 Normalized Autobinomial Markov Channels For Pedestrian Detection 178
Cosmin Ţoca, Mihai Ciuc, Carmen Pătraşcu

15:40-16:00 Exploring Prior Knowledge for Pedestrian Detection 179
Yi Yang, Zhenhua Wang, Fuchao Wu

16:00-16:20 Coffee

Action and Event

16:20-16:40 APT: Action localization proposals from dense trajectories 180
Jan C. van Gemert, Mihir Jain, Ella Gati, Cees G. M. Snoek

16:40-17:00 Event Fisher Vectors: Robust Encoding Visual Diversity of Visual Streams 181
Markus Nagel, Thomas Mensink, Cees G. M. Snoek

17:00-17:20 Deep Structured Models For Group Activity Recognition 182
Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, Greg Mori

17:20-17:40 Fast Action Retrieval from Videos via Feature Disaggregation 183
Jie Qin, Li Liu, Mengyang Yu, Yunhong Wang, Ling Shao

Conference Banquet (Brangwyn Hall)

19:00-22:00 The Brangwyn Hall is a concert venue in Swansea. It is named after the artist Frank Brangwyn, whose famous panels, originally intended for the House of Lords, are displayed there.



Programme: Thursday 10th September

08:15 Registration (until 17:45)

Tracking and Pose Estimation

09:00-09:20 Pose Estimation of Kinematic Chain Instances via Object Coordinate Regression 184
Frank Michel, Alexander Krull, Eric Brachmann, Michael Ying Yang, Stefan Gumhold, Carsten Rother

09:20-09:40 Hybrid One-Shot 3D Hand Pose Estimation by Exploiting Uncertainties 185
Georg Poier, Konstantinos Roditakis, Samuel Schulter, Damien Michel, Horst Bischof, Antonis A. Argyros

09:40-10:00 Particle dynamics and multi-channel feature dictionaries for robust visual tracking 186
Srikrishna Karanam, Yang Li, Richard J. Radke

10:00-10:20 Collaborative Correlation Tracking 187
Guibo Zhu, Jinqiao Wang, Yi Wu, Hanqing Lu

10:20-10:40 Enable Scale and Aspect Ratio Adaptability in Visual Tracking with Detection Proposals 188
Dafei Huang, Lei Luo, Mei Wen, Zhaoyun Chen, Chunyuan Zhang

10:40-11:00 Coffee

Detection and Recognition

11:00-11:20 Multiple Frames Matching for Object Discovery in Video 189
Otilia Stretcu, Marius Leordeanu

11:20-11:40 Surface Based Object Detection in RGBD Images 190
Siddhartha Chandra, Grigorios G. Chrysos, Iasonas Kokkinos

11:40-12:00 Context Forest for Object Class Detection 191
Davide Modolo, Alexander Vezhnevets, Vittorio Ferrari

12:00-12:20 Spatiotemporal Deformable Prototypes for Motion Anomaly Detection 192
Robert Bensch, Thomas Brox, Olaf Ronneberger

Programme: Thursday PM 10th September

Workshop One

13:30-18:00 MVAB: Machine Vision of Animals and their Behaviour 193
Chairs: Telmo Amaral, Stephen Matthews, Thomas Plötz, Stephen McKenna, Robert Fisher

Workshop Two

13:30-18:00 Diff-CV: Differential Geometry in Computer Vision for Analysis of Shapes, Images and Trajectories 196
Chairs: Hassen Drira, Sebastian Kurtek, Pavan Turaga

Workshop Three

13:30-18:00 CVPPP: Workshop on Computer Vision Problems in Plant Phenotyping 198
Chairs: Sotirios Tsaftaris, Hanno Scharr, Tony Pridmore

Workshop Four

13:30-18:00 BMVW: Computer Vision Student Workshop 201
Chairs: Gary Tam

BMVC 2015 Tutorial: Fitting models to data: Accuracy, Speed, Robustness

Andrew Fitzgibbon Microsoft Research

1 Speaker Biography

Andrew Fitzgibbon is a principal researcher in the computer vision group at Microsoft Research Cambridge. He is best known for his work on 3D vision, having been a core contributor to the Emmy-award-winning 3D camera tracker "boujou" (www.boujou.com) and Kinect for Xbox 360, but his interests are broad, spanning computer vision, graphics, machine learning, and even a little neuroscience. He has published numerous highly-cited papers, and received many awards for his work, including nine "best paper" prizes, the Silver medal of the Royal Academy of Engineering, and the BCS Roger Needham award. He is a fellow of the Royal Academy of Engineering, the British Computer Society, and the International Association for Pattern Recognition. Before joining Microsoft in 2005, he was a Royal Society University Research Fellow at Oxford University, having previously studied at Edinburgh University, Heriot-Watt University, and University College, Cork.



2 Abstract

In vision and machine learning, almost everything we do may be considered to be a form of model fitting. Whether estimating the parameters of a convolutional neural network, target tracking, computing low-dimensional representations of datasets, computing structure and motion from image collections, estimating parameters for an inference model such as Markov random fields, or extracting shape spaces such as active appearance models, it almost always boils down to minimizing an objective containing some parameters of interest as well as some latent or nuisance parameters. This tutorial will describe several tools and techniques for solving such optimization problems, covering a broad syllabus: vector and matrix calculus; nonlinear optimization algorithms such as gradient descent and its many variants, Quasi-Newton and Gauss-Newton derivates such as LBFGS and Levenberg Marquardt; modelling options for curves and surfaces; how to deal with missing data; how to deal with outliers, including how to optimize functions with robust kernels. There will be maths, and lots of it, but I hope to make it clear enough that anyone with decent high-school linear algebra and calculus can benefit, while nevertheless showing mathematical experts a few useful tricks.

BMVC 2015 Keynote: A Spectral Perspective on Shapes

Ron Kimmel Technion, Israel Institute of Technology

1 Speaker Biography

Ron Kimmel is a Professor of Computer Science at the Technion where he holds the Montreal Chair in Sciences. He held a post-doctoral position at UC Berkeley and a visiting professorship at Stanford University. He has worked in various areas of image and shape analysis in computer vision, image processing, and computer graphics. Kimmel's interest in recent years has been non-rigid shape processing and analysis, medical imaging and computational biometry, numerical optimization of problems with a geometric flavor, and applications of metric geometry and differential geometry. Kimmel is an IEEE Fellow for his contributions to image processing and non-rigid shape analysis. He is an author of two books, an editor of one, and an author of numerous articles. He is the founder of the Geometric Image Processing Lab. and a founder and advisor of several successful image processing and analysis companies.



2 Abstract

The differential structure of surfaces captured by the Laplace Beltrami Operator (LBO) can be used to construct a space for analyzing visual and geometric information. The decomposition of the LBO at one end, and the heat operator at the other end provide us with efficient tools for dealing with images and shapes. Denoising, matching, segmenting, filtering, exaggerating are just few of the problems for which the LBO provides a convenient operating environment. We will review the optimality of a truncated basis provided by the LBO, and a selection of relevant metrics by which such optimal bases are constructed. A specific example is the scale invariant metric for surfaces, that we argue to be a natural choice for the study of articulated shapes and forms.

Prof. Kimmel will comment about Intel's RealSense geometry sensor, the intrinsic and extrinsic transformations applied to Alice in Wonderland, and the benefit of having an axiomatic model when valid.

BMVC 2015 Keynote: Action and Attention in First-person Vision

Kristen Grauman University of Texas at Austin

1 Speaker Biography

Kristen Grauman is an Associate Professor in the Department of Computer Science at the University of Texas at Austin. Her research in computer vision and machine learning focuses on visual search and object recognition. Before joining UT-Austin in 2007, she received her Ph.D. in the EECS department at MIT, in the Computer Science and Artificial Intelligence Laboratory. She is an Alfred P. Sloan Research Fellow and Microsoft Research New Faculty Fellow, a recipient of NSF CAREER and ONR Young Investigator awards, the Regents' Outstanding Teaching Award from the University of Texas System in 2012, the PAMI Young Researcher Award in 2013, the 2013 Computers and Thought Award from the International Joint Conference on Artificial Intelligence, and a Presidential Early Career Award for Scientists and Engineers (PECASE) in 2013. She and her collaborators were recognized with the CVPR Best Student Paper Award in 2008 for their work on hashing algorithms for large-scale image retrieval, and the Marr Best Paper Prize at ICCV in 2011 for their work on modeling relative visual attributes.



2 Abstract

A traditional third-person camera passively watches the world, typically from a stationary position. In contrast, a first-person (wearable) camera is inherently linked to the ongoing experiences of its wearer. It encounters the visual world in the context of the wearer's physical activity, behavior, and goals. This distinction has many intriguing implications for computer vision research, in topics ranging from fundamental visual recognition problems to high-level multimedia applications.

Prof. Grauman will present their recent work in this space, driven by the notion that the camera wearer is an active participant in the visual observations received. First, she will show how to exploit egomotion when learning image representations. Cognitive science tells us that proper development of visual perception requires internalizing the link between"how I move" and "what I see" - yet today's best recognition methods are deprived of this link, learning solely from bags of images downloaded from the Web. Prof. Grauman introduces a deep feature learning approach that embeds information not only from the video stream the observer sees, but also the motor actions he simultaneously makes. She will demonstrate the impact for recognition, including a scenario where features learned from ego-video on an autonomous car substantially improve large-scale scene recognition. Next, she will present their work exploring video summarization from the first person perspective. Leveraging cues about ego-attention and interactions to infer a storyline, the work automatically detects the highlights in long videos. Prof. Grumman will show how hours of wearable camera data can be distilled to a succinct visual storyboard that is understandable in just moments, and examine the possibility of person- and scene-independent cues for heightened attention. Overall, whether considering action or attention, the first-person setting offers exciting new opportunities for large-scale visual learning.

Online Domain Adaptation for Multi-Object Tracking

Adrien Gaidon
adrien.gaidon@xrce.xerox.com
Eleonora Vig
eleonora.vig@xrce.xerox.com

Computer Vision Group
Xerox Research Centre Europe
Meylan, France

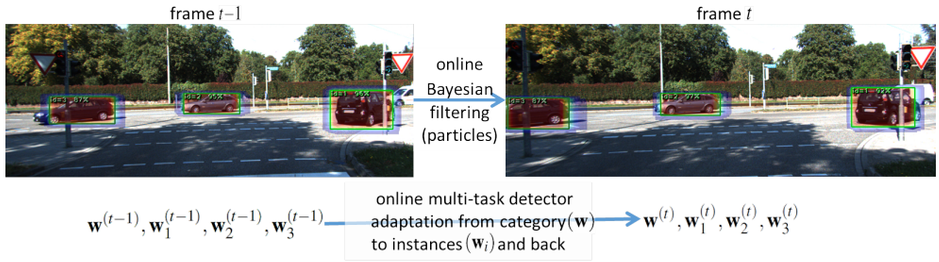


Figure 1: Online domain adaptation for MOT via Bayesian filtering coupled with multi-task adaptation of all detectors jointly.

Automatically detecting, labeling, and tracking objects in videos depends above all on accurate category-level object detectors. These might, however, not always be available in practice, as acquiring high-quality large scale labeled training datasets is either too costly or impractical for all possible real-world applications. A scalable solution consists in re-using object detectors pre-trained on generic datasets. This work is the first to investigate the problem of *on-line domain adaptation of object detectors for causal multi-object tracking (MOT)*. We propose to alleviate the dataset bias by adapting detectors from *category to instances, and back*: (i) we jointly learn all target models by adapting them from the pre-trained one, and (ii) we also adapt the pre-trained model on-line. Previous works investigated detector adaptation *or* on-line learning of appearance models, but not both *jointly*. Our approach can be interpreted as a generalization. We integrate our domain adaptation strategy in a novel motion model combining learned deterministic models with standard Bayesian filtering (*cf.* figure above) inspired by the popular Bootstrap filter. In particular, we leverage several techniques not widely used in MOT yet: (i) recent improvements in object detection based on object proposals, (ii) large-displacement optical flow estimation, (iii) the Fisher Vector representation, and (iv) ConvNet features for object detection. In addition, we use a Sequential Monte Carlo algorithm to approximate the filtering distribution of our Markovian motion model of the latent target locations.

Contrary to common practice in MOT, we here use object proposals, which we rank with a category-specific linear classifier parameterized by a vector $\mathbf{w}$. This classifier returns the probability that a candidate window $\mathbf{x}$, represented by a feature vector $\phi_t(\mathbf{x})$, contains an object of the category of interest at time t by $P(\mathbf{x}|\mathbf{w}) = \left(1 + e^{-\mathbf{w}^T \phi_t(\mathbf{x})}\right)^{-1}$. To represent proposals, we explore two common representations adapted to the computational constraints of tracking: Fisher Vectors with a single Gaussian and features from the memory-efficient pre-trained GoogLeNet ConvNet [6].

We propose a *convex multi-task learning objective* to *jointly adapt on-line* (i) all trackers from the pre-trained generic detector (*category-to-instance*), and (ii) the pre-trained category-level model from the trackers (*instances-to-category*). The first category-to-instance adaptation happens at the creation of a new track $\mathbf{w}_i^{(t_0)}$ by *warm-starting* its optimization from the category-level model $\mathbf{w}^{(t_0)}$, *i.e.* an already good solution. This leads to faster convergence and stronger regularization. The second category-to-instance adaptation consists in *updating all target models jointly* using multi-task learning. Given the stacked target models $\mathbf{W}^{(t)} = \{\mathbf{w}_1^{(t)}, \ldots, \mathbf{w}_{N_t}^{(t)}\}$ (N_t in total), and the training samples and labels $(\mathbf{X}^{(t)}, \mathbf{y}^{(t)})$ mined for all targets in frame t, updating all appearance models jointly amounts to minimizing the regularized empirical risk:

$$\mathbf{W}^{(t)} = \arg\min_{\mathbf{W}} L_t(\mathbf{X}^{(t)}, \mathbf{y}^{(t)}, \mathbf{W}) + \lambda \Omega_t(\mathbf{W}) \tag{1}$$

with the loss L_t and multi-task regularization term Ω_t defined as:

$$L_t(\mathbf{X}^{(t)}, \mathbf{y}^{(t)}, \mathbf{W}) = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{1}{n_i} \sum_{k=1}^{n_i} \ell_t(\mathbf{x}_{i,k}\ ,\ y_{i,k}\ ,\ \mathbf{w}_i) \tag{2}$$

method	MOTA↑	MOTP↑	MT↑	ML↓	Rec.↑	Prec.↑	FAR↓	IDS↓	FRG↓
DP_MCF† [5]	12.0%	68.5%	0.1%	80.2%	14.6%	**85.5%**	**7.7%**	**84**	**327**
G_TBD† [2]	17.5%	68.0%	0.9%	59.2%	30.0%	71.3%	37.6%	115	528
CFT [4]	17.6%	66.7%	1.8%	45.7%	33.5%	69.1%	47.2%	238	592
CIT [3]	22.8%	68.5%	**1.9%**	**43.4%**	33.9%	76.5%	32.6%	380	809
ODAMOT	**23.6%**	**68.7%**	1.8%	43.6%	**34.2%**	77.5%	31.1%	376	784

Table 1: MOT results on the PASCAL-to-KITTI domain adaptation dataset for the R-CNN-like detector. Methods with † are offline, the others are online.

$$\Omega_t(\mathbf{W}) = \frac{1}{2N_t} \sum_{i=1}^{N_t} \|\mathbf{w}_i - \bar{\mathbf{w}}^{(t-1)}\|_2^2, \tag{3}$$

where $\bar{\mathbf{w}}^{(t-1)}$ is the (running) mean of all previous instance models, and $\ell_t(\mathbf{x}, y, \mathbf{w})$ is the logistic loss.

The *instance-to-category* adaptation allows to continuously specialize the global appearance model to the specific video stream. Once the detectors $\mathbf{w}_i$ are updated in frame t, a new scene-adapted category detector is readily available as the running average of instance models:

$$\bar{\mathbf{w}}^{(t)} = \frac{1}{\bar{N}_{t-1} + N_t} \left(\bar{N}_{t-1} \bar{\mathbf{w}}^{(t-1)} + \sum_{i=1}^{N_t} \mathbf{w}_i^{(t)} \right), \text{ where } \bar{N}_{t-1} = \sum_{j=1}^{t-1} N_j. \tag{4}$$

Our multi-task formulation enforces parameter sharing between models to reduce model drift and robustly handle false alarms, while allowing for continuous domain adaptation to gradually decrease missed detections.

We evaluate our algorithm (ODAMOT) on the challenging KITTI car tracking benchmark [1]. On this dataset, ODAMOT achieves 57.06% MOTA and ranks third of all published methods despite its simple occlusion reasoning. We then quantitatively measure the benefit of our domain adaptation strategy on the new PASCAL-to-KITTI dataset we introduce to study the domain mismatch problem in MOT. The training set (the *source domain*) of this dataset consists of the training images of the standard Pascal VOC 2007 detection challenge, whereas the test set (the *target domain*) includes the 21 training videos of the KITTI tracking challenge. As expected, unrelated training data degrades MOT performance (*cf.* Table 1), however, our results show that domain adaptation partly mitigates this problem. Our multi-task adaptation from category-to-instances and back allows to improve overall MOT accuracy by increasing recall while maintaining high precision and limiting model drift.

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[2] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D Traffic Scene Understanding from Movable Platforms. *PAMI*, 2014.

[3] D. Hall and P. Perona. Online, Real-Time Tracking Using a Category-to-Individual Detector. In *ECCV*, 2014.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-Learning-Detection. *PAMI*, 2011.

[5] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

Learning Optimal Parameters For Multi-target Tracking

Shaofei Wang
shaofeiw@uci.edu
Charless C. Fowlkes
fowlkes@ics.uci.edu

Dept of Computer Science
University of California
Irvine, CA, USA

Multi-target tracking problems are traditionally tackled in two different ways. One way is to first group detections into candidate tracklets and then perform scoring and association of these tracklets [5, 6], this can be done in either an online/streaming fashion or an offline/batch fashion and it allows tracklets to be scored with richer trajectory and appearance models. Another approach is to attempt to include higher-order constraints directly in a combinatorial framework [1, 2]. In either case, there are a large number of parameters associated with these richer models which become increasingly difficult to set by hand and necessitate the application of machine learning techniques.

In this paper, we describe an end-to-end framework for learning parameters of min-cost flow multi-target tracking problem with quadratic trajectory interactions including suppression of overlapping tracks and contextual cues about co-occurrence of different objects. Our approach utilizes structured prediction with a tracking-specific loss function to learn the complete set of model parameters. Under our learning framework, we evaluate two different approaches to finding an optimal set of tracks under quadratic model objective based on an LP relaxation and a novel greedy extension to dynamic programming that handles pairwise interactions.

In a min-cost flow multi-target tracking problem, the set of optimal (most probable) tracks can be found by solving an integer linear program (ILP) over flow variables $\mathbf{f}$.

$$\min_{\mathbf{f}} \sum_i c_i^s f_i^s + \sum_{ij\in E} c_{ij} f_{ij} + \sum_i c_i f_i + \sum_i c_i^t f_i^t \tag{1}$$

$$\text{s.t.} \quad f_i^s + \sum_j f_{ji} = f_i = f_i^t + \sum_j f_{ij} \tag{2}$$

$$f_i^s, f_i^t, f_i, f_{ij} \in \{0,1\} \tag{3}$$

where E is the set of valid transitions between sites in successive frames. The costs c_i represent the negative log-likelihood ratio of an object appearing at a particular spatio-temporal location i based on image evidence, c_{ij} represents the cost of a transition between a location i in one frame and j in a subsequent frame and c^s and c^t are associated with the birth or death of a track. The flow conservation constraint (2) enforces that a detection at site i can only be active as part of a single contiguous track passing through that location.

It is also possible to capture interactions between multiple tracks by adding a pairwise cost term denoted q_{ij} for jointly activating a pair of flows f_i and f_j corresponding to detections at sites i and j. Adding this term to 1 yields an Integer Quadratic Program (IQP):

$$\min_{\mathbf{f}} \sum_i c_i^s f_i^s + \sum_{ij\in E} c_{ij} f_{ij} + \sum_i c_i f_i + \sum_{ij\in EC} q_{ij} f_i f_j + \sum_i c_i^t f_i^t \tag{4}$$

$$\text{s.t.} \quad (Eq.\ 2),\ (Eq.\ 3)$$

In our experiments, we investigate pairwise contextual interactions between pairs of sites in the same video frame which we denote by EC. The addition of quadratic terms makes the objective (4) NP-hard. We thus propose a novel greedy approximation based on repeated passes of dynamic programming and compare it with a standard LP relaxation-based approach.

We formulate parameter learning of tracking models as a structured prediction problem. Assume we have N training videos with detector outputs and corresponding ground-truth track associations specified by flow variables $\{(X_n, \mathbf{f}_n)\}$. We can parameterize the network flow costs c as a linear function of image and detection features where the parameters are specified by weight vector $\mathbf{w}$. We propose to discriminatively learn tracking model parameters $\mathbf{w}$ using a structured SVM with margin rescaling:

$$\mathbf{w}^* = \underset{\mathbf{w},\xi_n\geq 0}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \xi_n \tag{5}$$

$$\text{s.t. } \mathbf{w}^T\Psi(X_n,\widehat{\mathbf{f}}) - \mathbf{w}^T\Psi(X_n,\mathbf{f}_n) \geq L(\mathbf{f}_n,\widehat{\mathbf{f}}) - \xi_n \quad \forall n,\widehat{\mathbf{f}}$$

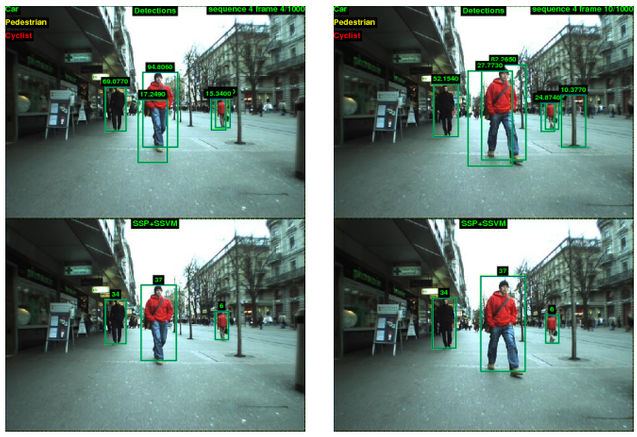


Figure 1: By learning a proper set of parameters, even a basic network-flow model without pairwise potentials can successfully prune away many false tracks by reasoning about detection confidence and transition smoothness.

where $\Psi(X_n,\mathbf{f}_n)$ are the features extracted from nth training video. $L(\mathbf{f}_n,\widehat{\mathbf{f}})$ is a loss function that penalizes any difference between the inferred label $\widehat{\mathbf{f}}$ and the ground truth label $\mathbf{f}_n$ and which satisfies $L(\mathbf{f}_n,\mathbf{f}_n) = 0$.

We use a standard cutting plane approach [4] to optimize the parameters $\mathbf{w}$ by repeatedly performing loss-augmented inference to find flows $\widehat{\mathbf{f}}$ that violate the constraint for each training example. Specifically, we propose to use a decomposable loss L for transition links that attempts to capture important aspects of multi-object tracking accuracy (MOTA) by taking into account the length and localization of transition links rather than using a constant (Hamming) loss on mislabled links.

Implementation details of approximate algorithms as well as the definition of tracking features and loss can be found in our full paper. Surprisingly, we found that with properly learned parameters, even the simple min-cost flow objective (1) yields better results than state-of-the-art methods on challenging MOT and KITTI benchmarks, while the quadratic terms improves the performance even further for tracking with ordinary, multi-category detector such as DPM [3].

[1] Asad A. Butt and Robert T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[2] Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, and Josef Sivic. On pairwise cost for multi-object network flow tracking. *CoRR*, abs/1408.3304, 2014.

[3] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[4] T. Joachims, T. Finley, and Chun-Nam Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[5] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang. Tracklet association with online target-specific metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[6] Bo Yang and Ram Nevatia. An online learned crf model for multi-target tracking. In *In CVPR*, 2012.

R-CNN minus R

Karel Lenc
http://www.robots.ox.ac.uk/~karel
Andrea Vedaldi
http://www.robots.ox.ac.uk/~vedaldi

Department of Engineering Science,
University of Oxford,
Oxford, UK.

Deep convolutional neural networks (CNNs) have had a major impact in most areas of image understanding, including object category detection. In object detection, methods such as R-CNN have obtained excellent results by integrating CNNs with region proposal generation algorithms such as selective search. In this paper, we investigate the role of proposal generation in CNN-based detectors in order to determine whether it is a necessary modelling component, carrying essential geometric information not contained in the CNN, or whether it is merely a way of accelerating detection. We do so by designing and evaluating a detector that uses a trivial region generation scheme, constant for each image. Combined with SPP, this results in an excellent and fast detector that does not require to process an image with algorithms other than the CNN itself. We also streamline and simplify the training of CNN-based detectors by integrating several learning steps in a single algorithm, as well as by proposing a number of improvements that accelerate detection.

Object detection is one of the core problems in image understanding. Until recently, the best performing detectors in standard benchmarks such as PASCAL VOC were based on a combination of handcrafted image representations such as SIFT, HOG, and the Fisher Vector and a form of structured output regression, from sliding window to deformable parts models. Recently, however, these pipelines have been outperformed significantly by the ones based on deep learning that acquire representations automatically from data using Convolutional Neural Networks (CNNs). Currently, the best CNN-based detectors are based on the R-CNN construction of [3]. Conceptually, R-CNN is remarkably simple: it samples image regions using a proposal mechanism such as Selective Search (SS; [6]) and classifies them as foreground and background using a CNN.

The first question that we address here is whether CNN contain sufficient geometric information to localise objects, or whether the latter must be supplemented by an external mechanism, such as region proposal generation. There are in fact two hypothesis. The first one is that the only role of proposal generation is to cut down computation by allowing to evaluate the CNN, which is expensive, on a small number of image regions. The second hypothesis is that, instead, proposal generation provides geometric information essential for accurate object localisation which is not represented in the CNN. This is not unlikely, given that CNNs are often trained to be highly invariant to even large geometric deformations and hence may not be sensitive to an object's location.

The second question is whether the R-CNN pipeline can be simplified. While conceptually straightforward, in fact, R-CNN comprises many practical steps that need to be carefully implemented and tuned to obtain a good performance. R-CNN builds on a CNN pre-trained on an image classification tasks such as ImageNet ILSVRC [1], such as the AlexNet network [5]. This CNN is ported to detection by: i) learning an SVM classifier for each object class on top of the last fully-connected layer of the network, ii) fine-tuning the CNN on the task of discriminating objects and background, and iii) learning a bounding box regressor for each object class. We simplify these steps, which require running a mix of different software on cached data, by training a single CNN addressing all required tasks.

The third question is whether R-CNN can be accelerated. A substantial speedup was already obtained in *spatial pyramid pooling* (SPP) by [4] by realising that convolutional features can be shared among different regions rather than being recomputed. However, this does not accelerate training, and in testing the region proposal generation mechanism becomes the new bottleneck. Our first improvement is that we are able to skip the SVM training step, which involves hard negative mining. Furthermore, at the test time the whole detector, including detection of multiple object classes and bounding box regression, *reduces to evaluating a single CNN* (implemented in MatConvNet [7]) which already brings a significant speedup shown in Table 1.

The Table 1 shows that for the SPP detector the main bottleneck is bounding box generation. We show, that picking a constant set of boxes (by clustering the ground truth boxes locations) and retraining the network and bounding box regressor, the drop in performance on PASCAL VOC 2007 data [2] is only about 6% mAP points and results in an overall detection speedup of more than 16×, from about 2.5s per image down to 160ms. The Figure 1 shows the performance of the resulting detectors versus the number of bounding box proposals.

Impl. [ms]	SelS	Prep.	Move	Conv	SPP	FC	BBR	Σ − SelS
SPP MS	$1.98 \cdot 10^3$	23.3	67.5	186.6	211.1	91.0	39.8	**619.2** ±118.0
OURS MS		23.7	17.7	179.4	38.9	87.9	9.8	**357.4** ±34.3
SPP SS		9.0	47.7	31.1	207.1	90.4	39.9	**425.1** ±117.0
OURS SS		9.0	3.0	30.3	19.4	88.0	9.8	**159.5** ±31.5

Table 1: Timing (in *ms*) of the original SPP-CNN and our streamlined full-GPU implementation, broken down into selective search (SS) and preprocessing: image loading and scaling (Prep), CPU/GPU data transfer (Move), convolution layers (Conv), spatial pyramid pooling (SPP), fully connected layers and SVM evaluation (FC), and bounding box regression (BBR).

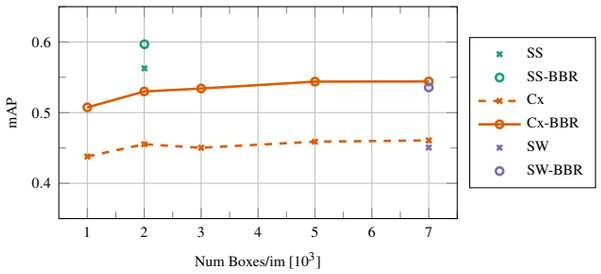


Figure 1: mAP on the PASCAL VOC 2007 test data as a function of the number of candidate boxes per image, proposal generation method, and using or not bounding box regression. In all cases, the CNN is fine-tuned for the particular bounding-box generation algorithm.

Our most significant finding is that current CNNs do contain sufficient geometric information for accurate object detection, although in the convolutional rather than fully connected layers. This finding opens the possibility of building state-of-the-art object detectors that rely exclusively on CNNs, removing region proposal generation schemes such as selective search, and resulting in integrated, simpler, and faster detectors. Our current implementation of a proposal-free detector is already much faster than SPP-CNN, and very close, but not quite as good, in term of mAP. However, we have only begun exploring the design possibilities and we believe that it is a matter of time before the gap closes entirely.

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR*, 2009.

[2] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool. The PASCAL visual obiect classes challenge 2007 (VOC2007) results. Technical report, Pascal Challenge, 2007.

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. ECCV*, 2014.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[6] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[7] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for MATLAB. *CoRR*, abs/1412.4564, 2014.

Convolutional Neural Networks for Direct Text Deblurring

Michal Hradiš[1]
ihradis@fit.vutbr.cz
Jan Kotera[2]
kotera@utia.cas.cz
Pavel Zemčík[1]
zemcik@fit.vutbr.cz
Filip Šroubek[2]
sroubekf@utia.cas.cz

[1] Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic

[2] Institute of Information Theory and Automation
Czech Academy of Sciences
Prague, Czech Republic

We investigate if convolutional neural networks (CNN) can learn to directly perform blind image deconvolution and restoration – that is if they can provide high-quality restored images directly from blurred inputs without any knowledge of the specific degradataion process. In our experiments on text documents, CNNs significantly outperformed existing blind deconvolution methods, including those optimized for text, in terms of image quality and OCR accuracy. In fact, the convolutional networks outperformed even state-of-the-art non-blind methods for anything but the lowest noise levels.

The architectures we use are inspired by the very successful networks that recently redefined state-of-the-art in many computer vision tasks starting with image classification on ImageNet by Krizhevsky et al. [2]. The networks are composed of multiple layers of convolutions and element-wise Rectified Liear Units (ReLU):

$$\begin{aligned} F_0(y) &= y \\ F_l(y) &= \max(0, W_l * F_{l-1}(y) + b_l), l = 1, \ldots, L-1 \\ F(y) &= W_L * F_{L-1}(y) + b_L \end{aligned} \tag{1}$$

The input and output are both 3-channel RGB images with values mapped to interval $[-0.5, 0.5]$. Each layer applies c_l convolutions with filters spanning all channels c_{l-1} of the previous layer. The last layer is linear (without ReLU).

As in previous works [5], we train the networks by minimizing mean squared error on a dataset $D = (x_i, y_i)$ of corresponding clean and corrupted image patches:

$$\arg\min_{W,b} \frac{1}{2|D|} \sum_{(x_i,y_i)\in D} ||F(y_i) - x_i||_2^2 + 0.0005||W||_2^2 \tag{2}$$

We evaluated the approach on a large set of documents from the CiteSeerX repository which we rendered at 120-150 DPI. A dataset was created by aplying random geometric transformations simulating deviations from optimal camera position to small rendered page regions and by convolving with realistic de-focus and camera-shake blur kernels (distribution of kernel sizes is show in Figure 1 right-bottom). We purposely limited the image degradations to shift-invariant blur and additive noise to allow for fair comparison with the baseline methods, which are not designed to handle other aspects of an imaging process.

In our experiments, deeper networks performed singnificantly better and restoration quality was fairly insensitive to other architectural choices. This can be clearly seen in Figure 1. Top-left shows best the benefit of deeper networks. Top-right shows that performance is insensitive to the size of the first layer filters. Bottom-left shows that networks with more channels per layer perform slightly better, but the benefit is negligable compared to increased network depth.

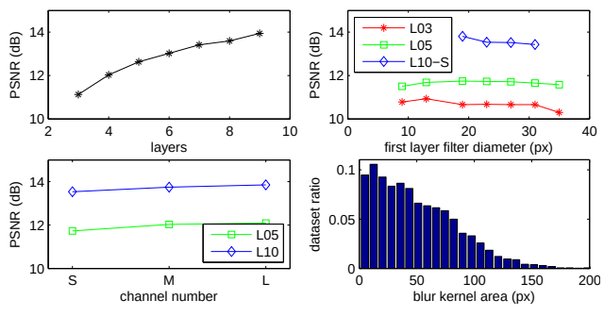


Figure 1: Different CNN architectures. The number after L (e.g. L10) indicates number of layers in the network. top-left – network depth; top-right – spatial support size; bottom-left – channel number; bottom-right – distribution of blur-kernel sizes in dataset

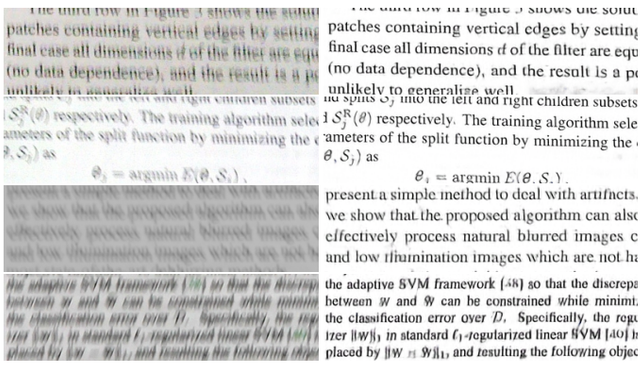


Figure 2: CNN deblurring results of challenging real images.

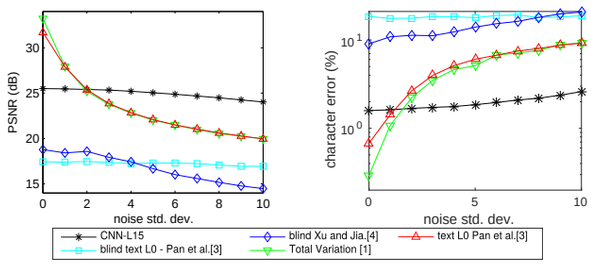


Figure 3: Blind deconvolution image quality (left) and OCR error on reconstructed images (right).

The largest and deepest network we trained has 15 layers with 2.3M parameters (PSNR 16.06 dB on the dataset from Figure 1). We compared this network with two blind [3, 4] and two non-blind [1, 3] methods on 200×200 image patches extracted from unseen documents (see Figure 3). The methods of Pan et al. [3] were designed specifically for text images. The CNN clearly outperforms the blind methods for all noise levels and the non-blind methods for all but the lowest noise levels. Surprisingly, the CNN maintains good quality even for noise levels higher than for what it was trained for.

[1] Stanley H. Chan, Ramsin Khoshabeh, Kristofor B. Gibson, Philip E. Gill, and Truong Q. Nguyen. An augmented Lagrangian method for total variation video restoration. *IEEE TIP*, 20(11):3097–3111, 2011.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 2012.

[3] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring Text Images via L0-Regularized Intensity and Gradient Prior. In *CVPR*, 2014.

[4] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, 2010.

[5] Li Xu, Jimmy SJ. Ren, Ce Liu, and Jiaya Jia. Deep Convolutional Neural Network for Image Deconvolution. In *NIPS*, 2014.

Sketch-a-Net that Beats Humans

Qian Yu*
q.yu@qmul.ac.uk
Yongxin Yang*
yongxin.yang@qmul.ac.uk
Yi-Zhe Song
yizhe.song@qmul.ac.uk
Tao Xiang
t.xiang@qmul.ac.uk
Timothy M. Hospedales
t.hospedales@qmul.ac.uk

School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, E1 4NS
United Kingdom

Sketches are very intuitive to humans and have long been used as an effective communicative tool. With the proliferation of touchscreens, sketching has become a much easier undertaking for many – we can sketch on phones, tablets and even watches. However, recognising free-hand sketches (e.g. asking a person to draw a car without any instance of car as reference) is an extremely challenging task. This is due to a number of reasons: (i) sketches are highly iconic and abstract, e.g., human figures can be depicted as stickmen; (ii) due to the free-hand nature, the same object can be drawn with hugely varied levels of detail/abstraction, e.g., a human figure sketch can be either a stickman or a portrait with fine details depending on the drawer; (iii) sketches lack visual cues, i.e., they consist of black and white lines instead of coloured pixels. A recent large-scale study on 20,000 free-hand sketches across 250 categories of daily objects puts human sketch recognition accuracy at 73.1% [2], suggesting that the task is challenging even for humans.

Prior work on sketch recognition generally follows the conventional image classification paradigm, that is, extracting hand-crafted features from sketch images followed by feeding them to a classifier. Most hand-crafted features traditionally used for photos (such as HOG, SIFT and shape context) have been employed, which are often coupled with Bag-of-Words (BoW) to yield a final feature representations that can then be classified. However, existing hand-crafted features designed for photos do not account for the unique abstract and sparse nature of sketches. Furthermore, they ignore a key unique characteristics of sketches, that is, a sketch is essentially an ordered list of strokes; they are thus sequential in nature (See Fig 1). In contrast with photos that consist of pixels sampled all at once, a sketch is the result of an online drawing process. It had long been recognised in psychology that such sequential ordering is a strong cue in human sketch recognition, a phenomenon that is also confirmed by recent studies in the computer vision literature [7]. However, none of the existing approaches attempted to embed sequential ordering of strokes in the recognition pipeline even though that information is readily available.

Index	Layer	Type	Filter Size	Filter Num	Stride	Pad	Output Size
0		Input	-	-	-	-	225×225
1	L1	Conv	15×15	64	3	0	71×71
2		ReLU	-	-	-	-	71×71
3		Maxpool	3×3	-	2	0	35×35
4	L2	Conv	5×5	128	1	0	31×31
5		ReLU	-	-	-	-	31×31
6		Maxpool	3×3	-	2	0	15×15
7	L3	Conv	3×3	256	1	1	15×15
8		ReLU	-	-	-	-	15×15
9	L4	Conv	3×3	256	1	1	15×15
10		ReLU	-	-	-	-	15×15
11	L5	Conv	3×3	256	1	1	15×15
12		ReLU	-	-	-	-	15×15
13		Maxpool	3×3	-	2	0	7×7
14	L6	Conv(=FC)	7×7	512	1	0	1×1
15		ReLU	-	-	-	-	1×1
16		Dropout (0.50)	-	-	-	-	1×1
17	L7	Conv(=FC)	1×1	512	1	0	1×1
18		ReLU	-	-	-	-	1×1
19		Dropout (0.50)	-	-	-	-	1×1
20	L8	Conv(=FC)	1×1	250	1	0	1×1

Table 1: The architecture of Sketch-a-Net.

HOG-SVM [2]	Ensemble [5]	MKL-SVM [6]	FV-SP [7]	
56%	61.5%	65.8%	68.9	
AlexNet-SVM [3]	AlexNet-Sketch [3]	LeNet [4]	Sketch-a-Net	Human [2]
67.1%	68.6%	55.2%	**74.9%**	73.1%

Table 2: Comparison with state of the art results on sketch recognition



Figure 1: Illustration of stroke ordering in sketching with the Alarm Clock category. Each sketch is split into three parts according to stroke ordering.

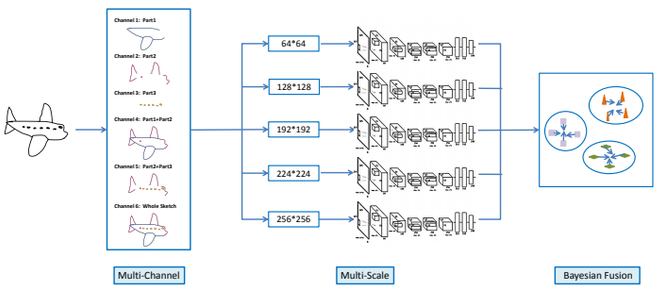


Figure 2: Illustration of our overall framework.

In this paper, we propose a novel deep neural network (DNN), Sketch-a-Net (See Fig 2), for free-hand sketch recognition, which is specifically designed to accommodate the unique characteristics of sketches including multiple levels of abstraction and being sequential in nature. Our contributions are summarised as follows: (i) for the first time, a representation learning model based on DNN is presented for sketch recognition in place of the conventional hand-crafted feature based sketch representations (Details are listed in Table 1); (ii) we demonstrate how sequential ordering information in sketches can be embedded into the DNN architecture and in turn improve sketch recognition performance; (iii) we propose a multi-scale network ensemble that fuses networks learned at different scales together via joint Bayesian fusion [1] to address the variability of levels of abstraction in sketches. Extensive experiments on the largest hand-free sketch benchmark dataset, the TU-Berlin sketch dataset [2], show that our model significantly outperforms existing approaches and can even beat humans by 1.8% at sketch recognition (See Table 2).

[1] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, 2012.

[2] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? In *SIGGRAPH*, 2012.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] Y. LeCun, L. Bottou, G. B. Orr, and K. Müller. Efficient backprop. *Neural networks: Tricks of the trade*, pages 9–48, 2012.

[5] Y. Li, Y. Song, and S. Gong. Sketch recognition by ensemble matching of structured features. In *BMVC*, 2013.

[6] Y. Li, T. M. Hospedales, Y. Song, and S. Gong. Free-hand sketch recognition by multi-kernel feature learning. *CVIU*, 2015.

[7] R. G. Schneider and T. Tuytelaars. Sketch classification and classification-driven analysis using fisher vectors. In *SIGGRAPH Asia*, 2014.

* These authors contributed equally to this work

Learning Deep Representations of Appearance and Motion for Anomalous Event Detection

Dan Xu[1], Elisa Ricci[2]
danxuhk@gmail.com,eliricci@fbk.eu
Yan Yan[1,3], Jingkuan Song[1], Nicu Sebe[1]
yan@disi.unitn.it,jingkuan.song@unitn.it,sebe@disi.unitn.it

[1] DISI, University of Trento, Italy
[2] Fondazione Bruno Kessler, Trento, Italy
[3] Advanced Digital Sciences Center, UIUC Singapore, Singapore

Introduction. A fundamental challenge in intelligent video surveillance is to automatically detect abnormal events in long video streams. This problem has attracted considerable attentions in recent years. In this paper we propose a novel Appearance and Motion DeepNet (AMDN) framework for discovering anomalous activities in complex video surveillance scenes. Opposite to previous works [1, 2], instead of using hand-crafted features to model activity patterns, we propose to learn discriminative feature representations of both appearance and motion patterns in a fully unsupervised manner. A novel approach based on stacked denoising autoencoders (SDAE) [3] is introduced to achieve this goal.

Contributions. i) As far as we know, we are the first to introduce an unsupervised deep learning framework to automatically construct discriminative representations for video anomaly detection. ii) We propose a new approach to learn appearance and motion features as well as their correlations. Deep learning methods for combining multiple modalities have been investigated in previous works. However, to our knowledge, this is the first work where multimodal deep learning is applied to anomalous event detection. iii) A double fusion scheme is proposed to combine appearance and motion features for discovering unusual activities. iv) Our method is validated on challenging anomaly detection datasets and we obtain very competitive performance compared with the state-of-the-art.

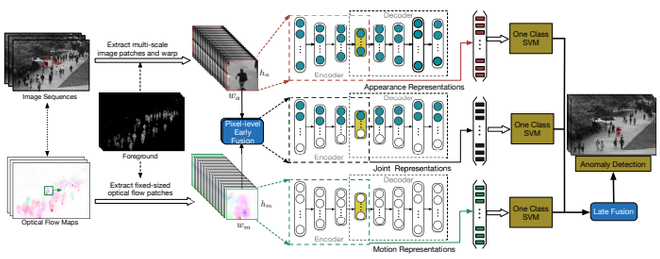


Figure 1: Overview of the proposed AMDN method for anomalous event detection. The proposed AMDN structure consists of three SDAE pipelines corresponding to different types of low-level inputs.

AMDN for Abnormal Event Detection. An overview of the proposed AMDN is shown in Fig. 1. Low-level visual information including still image patches and dynamic motion fields represented with optical flow is used as input of two separate networks, to first learn appearance and motion features, respectively. To further investigate the correlations between appearance and motion, early fusion is performed by combining image pixels with their corresponding optical flow to learn a joint representation. Finally, for abnormal event prediction, a late fusion strategy is introduced to combine the anomaly scores predicted by multiple one-class SVM classifiers, each corresponding to one of the three learned feature representations.

To learn the feature representations, we use three SDAE pipelines corresponding to different types of low-level inputs. The three SDAE networks learn appearance and motion features as well as a joint representation of them. We show the basic structures of the proposed SDAE networks in Fig. 2 (a) and (b). Each SDAE consists of two parts: encoder and decoder. For the encoder part, we use an over-complete set of filters in the first layer to capture a representative information from the data. Then, the number of neurons is reduced by half in the next layer until reaching the "bottleneck" hidden layer. The decoder part has a symmetric structure with respect to the encoder part.

We train the AMDN with two steps: pretraining and fine-tuning. The layer-wise pretraining learns one single denoising auto-encoder at a time with sparsity constraints. The input is corrupted to learn the mapping function, which is then used to produce the representation for the next layer with uncorrupted inputs. By using a greedy layer-wise pretraining, the denoising autoencoders can be stacked to build a multi-layer feedforward deep neural network, *i.e.* a stacked denoising autoencoder. Then fine-tuning is used to adjust parameters over the whole network.

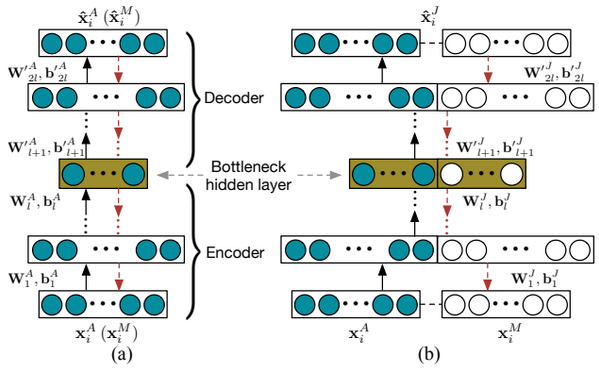
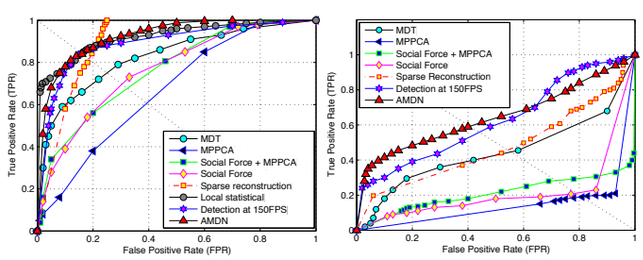


Figure 2: The structure of (a) the appearance and motion, and (b) the joint representation learning pipelines.

We formulate the video anomaly detection problem as a patch-based binary categorization problem, *i.e.* given a test frame we obtain $M_I \times N_I$ patches via sliding window with a stride d and classify each patch as corresponding to a normal or abnormal region. Specifically, given each test patch t we compute three anomaly scores $A^k(\mathbf{s}_t^k)$, $k \in \{A,M,J\}$, using one-class SVM models and the features representations $\mathbf{s}_t^k$ computed with the SDAEs. The three scores are then linearly combined to obtain the final anomaly score $\mathcal{A}(\mathbf{s}_t^k) = \sum_{k\in\{A,M,J\}} \alpha^k A^k(\mathbf{s}_t^k)$ ($k \in \{A,M,J\}$ corresponds to appearance, motion and joint representation, respectively). The weight vector α^k is automatically learned via an unsupervised late fusion scheme. Then for each patch t, we identify if it corresponds to an abnormal activity by computing the associated anomaly score $\mathcal{A}(\mathbf{s}_t^k)$ and comparing it with a threshold η, *i.e.* $\mathcal{A}(\mathbf{s}_t^k) \underset{abnormal}{\overset{normal}{\lessgtr}} \eta$.



(a) Frame-level ROC curve of PED1 Dataset (b) Pixel-level ROC curve of PED1 Dataset

Figure 3: UCSD dataset (Ped1 sequence): comparison of frame-level and pixel-level anomaly detection results with state of the art methods.

Results. Two publicly available datasets, the UCSD (Ped1 and Ped2) dataset and the Train dataset are used to evaluate the performance of the proposed approach. Fig. 3 (a) and (b) show the frame-level and pixel-level detection results on Ped1. The ROC curve is produced by varying the threshold parameter η. It is evident that our method outperforms most previous methods and that its performance are very competitive with the best two baselines. The proposed method is also evaluated on the Train dataset showing promising results (see the main paper).

[1] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013.

[2] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010.

[3] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.

Deep Perceptual Mapping for Thermal to Visible Face Recognition

M. Saquib Sarfraz
https://cvhci.anthropomatik.kit.edu/~ssarfraz
Rainer Stiefelhagen
http://cvhci.anthropomatik.kit.edu

Institute of Anthropomatics & Robotics
Karlsruhe institute of Technology
Karlsruhe, Germany.

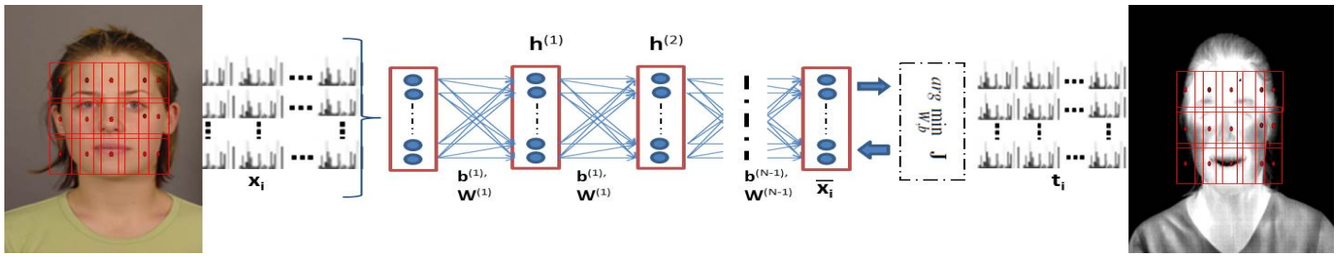


Figure 1: Deep Perceptual Mapping (DPM): densely computed features from the visible domain are mapped through the learned DPM network to the corresponding thermal domain.

Cross modal face matching between the thermal and visible spectrum is a much desired capability for night-time surveillance and security applications. Due to a very large modality gap, thermal-to-visible face recognition is one of the most challenging face matching problem. In this paper, we present an approach to bridge this modality gap by a significant margin. Our approach captures the highly non-linear relationship between the two modalities by using a deep neural network. Our model attempts to learn a non-linear mapping from visible to thermal spectrum while preserving the identity information. We show substantive performance improvement on a difficult thermal-visible face dataset (UND-X1). The presented approach improves the state-of-the-art by more than 10% in terms of Rank-1 identification and bridge the drop in performance due to the modality gap by more than 40%.

The goal of training the deep network is to learn the projections that can be used to bring the two modalities together. Typically, this would mean regressing the representation from one modality towards the other.

We construct a deep network comprising $N+1$ layers with $m^{(k)}$ units in the k-th layer, where $k = 1, 2, \cdots, N$. For an input of $x \in \mathbb{R}^d$, each layer will output a non-linear projection by using the learned projection matrix $\mathbf{W}$ and the non-linear activation function $g(\cdot)$. The output of the k-th hidden layer is $h^{(k)} = g(\mathbf{W}^{(k)} h^{(k-1)} + \mathbf{b}^{(k)})$, where $\mathbf{W}^{(k)} \in \mathbb{R}^{m^{(k)} \times m(k-1)}$ is the projection matrix to be learned in that layer, $\mathbf{b}^{(k)} \in \mathbb{R}^{m^{(k)}}$ is a bias vector and $g : \mathbb{R}^{m^{(k)}} \mapsto \mathbb{R}^{m^{(k)}}$ is the non-linear activation function. Similarly, the output of the most top level hidden layer can be computed as:

$$\mathbf{H}(x) = h^{(N)} = g(\mathbf{W}^{(N)} h^{(N-1)} + \mathbf{b}^{(N)}) \quad (1)$$

where the mapping $\mathbf{H} : \mathbb{R}^d \mapsto \mathbb{R}^{m^{(N)}}$ is a parametric non-linear perceptual mapping function learned by the parameters $\mathbf{W}$ and $\mathbf{b}$ over all the network layers. To determine the parameters $\mathbf{W}$ and $\mathbf{b}$ for such a mapping, our objective function must seek to minimize the perceptual difference between the visible and thermal training examples in the least mean square sense. We, therefore, formulate the DPM learning as the following optimization problem.

$$arg \min_{W,b} \quad \mathbf{J} = \frac{1}{M} \sum_{i=1}^{M} (\bar{x}_i - t_i)^2 + \frac{\lambda}{N} \sum_{k=1}^{N} (\|\mathbf{W}^{(k)}\|_F^2 + \|\mathbf{b}^{(k)}\|_2^2) \quad (2)$$

The first term in the objective function corresponds to the simple squared loss between the network output $\bar{x}$ given the visible domain input and the corresponding training example t from the thermal domain. The second term in the objective is the regularization term with λ as the regularization parameter. $\|\mathbf{W}\|_F$ is the Frobenius norm of the projection matrix $\mathbf{W}$. Given a training set $\mathbf{X} = \{x_1, x_2, \cdots, x_M\}$ and $\mathbf{T} = \{t_1, t_2, \cdots, t_M\}$ from visible and thermal domains respectively, the objective of training is to minimize the function in equation 2 with respect to the parameters $\mathbf{W}$ and $\mathbf{b}$.

The network is trained on densely computed feature representations (SIFT vectors) from overlapping small regions in the images. This proves effective, as the model is able to capture the differing local region's perceptual differences well. The training set comprises of these vectors coming from the corresponding patches from the images of the same identity. Using the corresponding images of the same identity ensures that the model will learn the only present differences due to the modality. Figure 1 capsulizes this process.

Effect of Modality gap: Performance with 1 Gallery image/subject			
Thermal-Thermal	Thermal-Visible	Thermal-Visible (via DPM)	Modality-gap bridged
89.47	30.36	55.36	$\sim 42\%$

Table 1: Performance drop due to Modality gap: Rank-1 identification using 1 image/subject as gallery in Thermal-Thermal and Thermal-Visible matching using baseline features.

After obtaining the mapping from visible to thermal domain, we can now pose the matching problem as that of comparing the thermal images with that of mapped visible data. The presented set-up is ideal for the surveillance scenario as the gallery images can be processed and stored offline while at test time no transformation and overhead is necessary. We use a simple matrix vector multiplication to compute the similarity enabling us to match the probes in real-time.

We report evaluations using typical identification and verification settings. As baseline we use the same concatenated SIFT features but without the DPM mapping. This enables us to directly compare and see the effectiveness of the proposed model. Our results show (see paper) that we improve the state-of-the-art best published results by more than 10% in all cases.

Effect of modality gap: We also present the experiment to measure the effect of modality gap. Keeping everything fixed *i.e.* using the same baseline features and settings, we compute the rank-1 identification score within the same modality. As shown in Table 1, we obtain 89.7% rank-1 score in the Thermal-Thermal identification scenario and 30.3% in the corresponding Thermal-visible scenario (using the same baseline features). This amounts to the performance drop, purely due to modality change, of about 59%. This reflects the challenging nature of the problem and the existing research gap to tackle this. As shown, with DPM on the same features, the performance is improved by 25%. This amounts to bridging the existing modality gap of 59% by more than 40%.

Computational Time: Training the DPM on 12 cores 3.2-GHz CPU takes between $1-1.5$ hours on MATLAB. Preprocessing, features extraction and mapping using DPM only takes $45ms$ for one image. This is even less in the testing case since no mapping is required for thermal images. It is, therefore, very fast and capable of running in real-time at ~ 28 fps.

Conclusively, the presented DPM approach is very effective, easy to train, real-time capable and provides a practical solution for a large surveillance and military application industry.

An Efficient Algorithm for Learning Distances that Obey the Triangle Inequality

Arijit Biswas
http://www.xrci.xerox.com/profile-main/67
Xerox Research Centre India
Bangalore, India

David Jacobs
http://www.cs.umd.edu/~djacobs/
Computer Science Department
University of Maryland
College Park, USA

Semi-supervised clustering of images has been an interesting problem for machine learning and computer vision researchers for decades. Pairwise constrained clustering is a popular paradigm for semi supervision that uses knowledge about whether two images belong to the same category (must-link constraint) or not (can't-link constraint). Performance of constrained clustering algorithms can be improved if the supervision on some image pairs is used to modify the pairwise distances of other image pairs, for which no supervision is available. There are several excellent metric learning approaches when the image distances can be represented as Euclidean distances in vector spaces [6]. However, in many cases distances are computed robustly [3], on manifolds [4] or as the output of algorithms [1] or classifiers. These do not lead to a natural embedding in a metric space, and may not even obey the triangle inequality. We are particularly interested in semi-supervised clustering of fine-grained categories. If we look at the top 10 distance measures in two important domains (LFW faces and leaf shapes) related to fine-grained classification, we find that 80% of the methods use non-vector space distances. These distances can be used for clustering images, but are not suitable for existing metric learning algorithms.

In order to propagate constraints from supervised to unsupervised pairs, some structure must be assumed on the set of possible distances. Otherwise, the distance between supervised pairs could be altered without affecting the distance between unsupervised pairs. Perhaps the weakest assumption that we can make about a distance is that it obeys the triangle inequality. Enforcing the triangle inequality allows us to propagate constraints; if a constraint alters one distance, other distances must also change to maintain triangle inequalities. For many interesting distances, the triangle inequality is not guaranteed to hold. However, we empirically find that the triangle inequality almost always holds for distances computed for fine-grained classification even when not explicitly enforced. This strongly motivates us to enforce the triangle inequalities when we alter distances to incorporate the pairwise constraints. We then find empirically that by enforcing the triangle inequality we can improve performance on several real world datasets.

Our main contribution is to formulate distance learning with pairwise constraints as a metric nearness problem[1] [2] and then provide an efficient algorithm to solve metric nearness for clustering. First, we formulate a quadratic optimization problem, where the pairwise distances between images are modified such that pairwise constraints and triangle inequality constraints are satisfied as much as possible. Since enforcing $O(N^3)$ triangle inequalities is computationally expensive, we propose a graph based approach, where only $O(n(M+C))$ triangle inequalities are sampled for use in the QP (N is the total number of images, n is the number of nearest neighbors in the n-nearest neighbor graph, M and C are the number of must-link and can't-link constraints respectively). We empirically show that this sampling approach works well in practice. We use the distances obtained by our approach along with a constrained clustering algorithm [5] to achieve state-of-the-art clustering results.

We theoretically analyze a simplified case in which only one pairwise constraint is present, to gain insights into our fast approach. Our sampling approach is based on the intuition that clustering is predominately affected by small distances, and is not sensitive to the exact value of larger distances. We prove that our sampling approach produces the same set of small distances that would be obtained by enforcing all constraints. We perform experiments on leaf and face image/video datasets and show that distances obtained by our method achieve state-of-the-art clustering results.

Formulation: We begin with a set of N unlabeled images $\mathcal{U}$ ($\boldsymbol{x} \in \mathcal{U}$) from K classes. We are also provided with initial distances between all the image pairs, i.e., $d_I(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is given $\forall i,j$ (note that d_I denotes initial distance). However we do not have any vector representation of the images. We have access to a set of must-link and can't-link constraints. Let $\mathcal{M}$ denote the set of must-link constraints such that any pair $(\boldsymbol{x}_i,\boldsymbol{x}_j) \in \mathcal{M}$ implies that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class. Similarly $\mathcal{C}$ denotes the set of can't-link constraints such that any pair $(\boldsymbol{x}_i,\boldsymbol{x}_j) \in \mathcal{C}$ implies that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to different classes.

We learn a new set of distances for all the image pairs given the pairwise constraints $\mathcal{M}$ and $\mathcal{C}$. A quadratic optimization problem is formulated to find these distances. Let us assume that the set of final distances are given by $d_F(\boldsymbol{x}_i,\boldsymbol{x}_j)$, which we can obtain by solving the following quadratic optimization problem:

$$\begin{aligned} \underset{d_F}{\text{minimize}} \quad & \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\notin\mathcal{M}\cup\mathcal{C}} (d_F(\boldsymbol{x}_i,\boldsymbol{x}_j) - d_I(\boldsymbol{x}_i,\boldsymbol{x}_j))^2 \\ \text{subject to} \quad & (i)\ \ d_F(\boldsymbol{x}_i,\boldsymbol{x}_j) \le U, \quad (\boldsymbol{x}_i,\boldsymbol{x}_j) \in \mathcal{M} \\ & (ii)\ \ d_F(\boldsymbol{x}_i,\boldsymbol{x}_j) \ge L, \quad (\boldsymbol{x}_i,\boldsymbol{x}_j) \in \mathcal{C} \\ & (iii)\ \ d_F(\boldsymbol{x}_i,\boldsymbol{x}_j) + d_F(\boldsymbol{x}_j,\boldsymbol{x}_k) \ge d_F(\boldsymbol{x}_i,\boldsymbol{x}_k), \forall i,j,k \\ & (iv)\ \ d_F(\boldsymbol{x}_i,\boldsymbol{x}_j) \ge 0, \forall i,j \end{aligned} \tag{1}$$

In the formulation in Eq. 1, distances between all image pairs are modified such that the constraints (i) to (iv) are satisfied. The objective function minimizes the total sum of changes in pairwise distances. Constraint (i) causes distances corresponding to must-link constraints to be reduced so they are upper-bounded by U, a user defined constant. Similarly, constraints in (ii) move can't-link image pairs as far as possible. The triangle inequality constraints are added in (iii). These triangle inequality constraints propagate information about must-link and can't-link constraints to other image pairs. The final set of constraints in (iv) ensure that all the quadratic optimization variables, i.e., the pairwise distances, remain non-negative.

In this QP formulation, there are large number of QP variables ($O(N^2)$) and triangle inequality constraints ($O(N^3)$). Also with all the pairwise and triangle inequality constraints there may not exist a feasible solution. To avoid these issues we reduce the size of the QP significantly by determining which triangle inequalities are crucial for clustering using a novel graph-based formulation (please see the paper for further details).

[1] Given a dissimilarity matrix, find the "nearest" matrix of distances that satisfy the triangle inequalities.

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape context: A new descriptor for shape matching and object recognition. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *NIPS*, pages 831–837. MIT Press, 2000.

[2] J Brickell, I. S. Dhillon, Suvrit Sra, and J Tropp. The metric nearness problem. *SIAM J. Matrix Analysis and Applications*, April 23 2008. URL http://eprints.pascal-network.org/archive/00004443/;http://link.aip.org/link/?SML/30/375.

[3] David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell*, 2000. URL http://doi.ieeecomputersociety.org/10.1109/34.862197.

[4] F. R. Schmidt, M. Clausen, and D. Cremers. Shape matching by variational computation of geodesics on a manifold. In *DAGM*, pages 142–151, 2006. URL http://dx.doi.org/10.1007/11861898_15.

[5] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained K-means clustering with background knowledge. In *ICML*, 2001.

[6] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart J. Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002. URL http://books.nips.cc/papers/files/nips15/AA03.pdf.

Diagnosing State-Of-The-Art Object Proposal Methods

Hongyuan Zhu[1]
zhuh@i2r.a-star.edu.sg
Shijian Lu[1]
slu@i2r.a-star.edu.sg
Jianfei Cai[2]
ASJFCAI@ntu.edu.sg
Guangqing Lee[3]
134174B@mymail.nyp.edu.sg

[1] Institute for Infocomm Research, A*Star, Singapore
[2] Nanyang Technological University, Singapore
[3] Nanyang Polytechnic, Singapore

Recent top performing methods in PASCAL VOC [6] and ImageNet [13] make use of object proposal to replace exhaustive window search. Object proposal's effectiveness is rooted in the assumption that there are general cues to differentiate objects from the background. Since the very first work by Alexe et al. [1], many object proposal methods have been proposed [2, 3, 4, 5, 7, 8, 10, 11, 12, 14, 15] and tested on various large scale datasets [6, 9, 13], and their overall detection rates versus different thresholds or window number have also been reported. Yet such partial performance summaries give us little idea of a method's strengths and weaknesses for further improvement, and users are still facing difficulties in choosing methods for their applications. Therefore, more detailed analysis of existing state-of-the-arts is critical for future research and applications.

Our contributions can be summarized in three aspects. First, we investigate the influence of object-level characteristics over state-of-the-art object proposal methods for the first time. Although there are some similar works in categorical object detection, few research has been conducted on object proposal side to the best of our knowledge. Second, we introduce the concept of localization latency to evaluate a method's localization efficiency and accuracy. Third, we create a fully annotated PASCAL VOC dataset with various object-level characteristics to facilitate our analysis. The annotations take us nearly one month's time which will be released to facilitate further related research.

Our experiments are based on PASCAL VOC2007 test set, which has been widely used in evaluating object proposal methods. A proposed window B is treated as detected if its Intersection-over-Union (IoU) with a ground truth bounding box $\bar{B}$: $IoU(\bar{B},B) = \frac{area(B \cap \bar{B})}{area(B \cup \bar{B})}$ is above a certain threshold T. We first study the localization accuracy of the existing methods. The region based methods have higher localization accuracy than window based methods. MCG and SelectiveSearch are the top performing region based methods, though window based EdgeBox shows comparable performance. The localization accuracy for region based methods are similar. One potential explanation is that all region based methods follow similar pipeline by grouping superpixels with either learned or handcrafted edge measures.

A good object proposal method should not only produce candidates with high accuracy, but also use as less windows as possible. To summarize a method's performance in terms of the accuracy and window number, we propose the **localization latency** metric:

$$LocalizationLatency = \frac{1}{|T|}\sum_{T}\log_2(\frac{1}{M}\sum_{K}\sum_{m}OBJ(m,T,K)) \quad (1)$$

where K ($K \in \{100, 1000, 10000\}$) are the number of windows $W_1^I, ..., W_K^I$ in returned descending order. Let $OBJ(m,T,K)$ be the total number of sampled windows to localize all instances of the target category in image m under varying threshold $T \in \{0.5, 0.7, 0.85\}$ and K (i.e. $OBJ(m,T,K) = \sum_i min\{k \in K : IOU(W_k^m, B_i^m) > T\}$). If using k windows still can't localize the object, k is set to $K+1$. M is the number of images contain the instances. From our experiments, the window based methods have lower localization latency than region based methods. Our interpretation for such result is that window based approach inherently assumes that objects are spatially compact.

We also examine the impact of additional properties: natureness, iconic view, object size, aspect ratio, color contrast, shape regularity and textureness. Our study reveals that the cues of existing methods are generalized to both 'nature' and 'man-made' objects. While, exsitng methods are sensitive to the changes in viewpoint, object size, aspect ratio, color contrast and shape regularity. While the impact of of textureness is pronounced.

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *CVPR*, 2010.

[2] Pablo Andrés Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[3] João Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[4] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. BING: binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014.

[5] Ian Endres and Derek Hoiem. Category independent object proposals. In *ECCV*, 2010.

[6] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[7] Ahmad Humayun, Fuxin Li, and James M. Rehg. RIGOR: reusing inference in graph cuts for generating object regions. In *CVPR*, 2014.

[8] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *ECCV*, 2014.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[10] Santiago Manen, Matthieu Guillaumin, and Luc J. Van Gool. Prime object proposals with randomized prim's algorithm. In *ICCV*, 2013.

[11] Esa Rahtu, Juho Kannala, and Matthew B. Blaschko. Learning a category independent object detection cascade. In *ICCV*, 2011.

[12] Pekka Rantalankila, Juho Kannala, and Esa Rahtu. Generating object segmentation proposals using global and local search. In *CVPR*, 2014.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL `http://arxiv.org/abs/1409.0575`.

[14] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

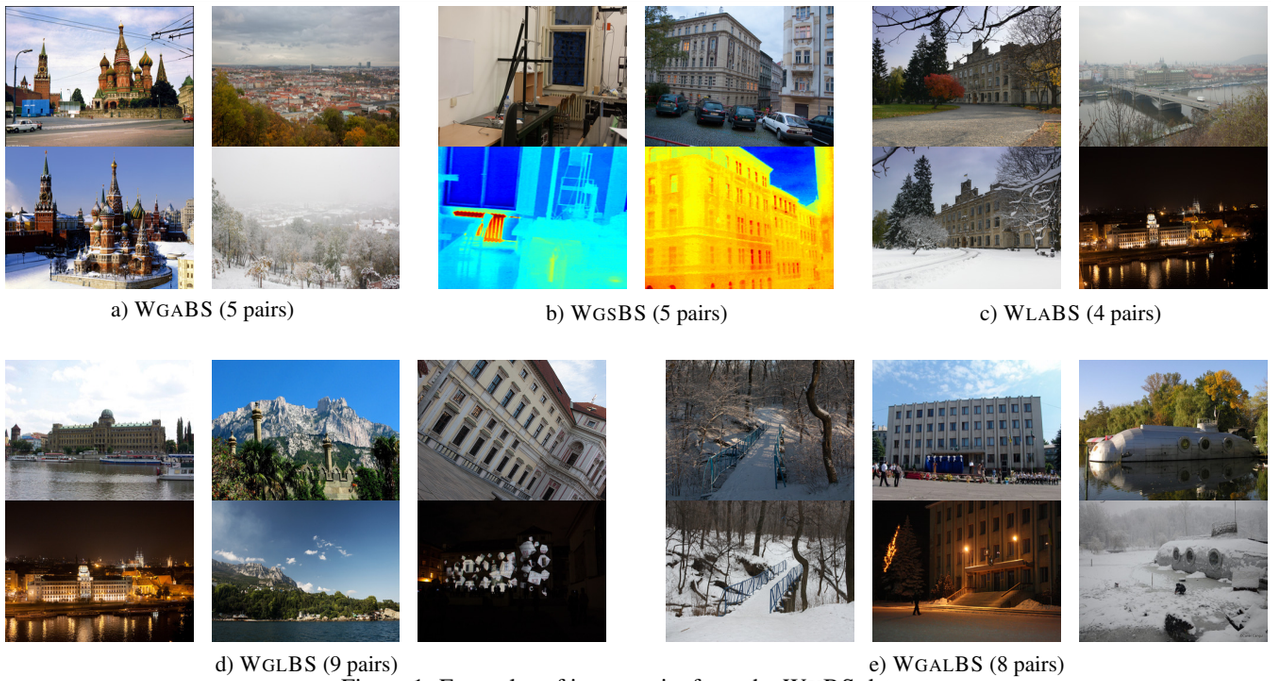
[15] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.

Tuesday Poster 1

WxBS: Wide Baseline Stereo Generalizations

Dmytro Mishkin[1]
ducha.aiki@gmail.com
Jiri Matas[1]
matas@cmp.felk.cvut.cz
Michal Perdoch[2]
perdoch@vision.ee.ethz.ch
Karel Lenc[3]
karel@robots.ox.ac.uk

[1] Center for Machine Perception
Czech Technical University in Prague Czech Republic
[2] Computer Vision Laboratory, ETH Zurich, Switzerland
[3] Visual Geometry Group
Department of Engineering Science
University of Oxford, Oxford, UK



a) WGABS (5 pairs) b) WGSBS (5 pairs) c) WLABS (4 pairs)

d) WGLBS (9 pairs) e) WGALBS (8 pairs)

Figure 1: Examples of image pairs from the WxBS dataset

We consider the generalization of Wide (geometric) Baseline Stereo to WxBS, a two-view image matching problem where two or more of the image formation and acquisition properties significantly change, i.e. they have a *wide baseline*.

The following single wide baseline stereo problems and their combinations are considered: illumination (WLBS) – difference in position, direction, number, intensity and wavelength of light sources; geometry (WGBS) – difference in camera and object pose, scale and resolution - the "classical" WBS; sensor (WSBS) – change in sensor type: visible, IR, MR; noise, image preprocessing algorithms inside the camera, etc; appearance (WABS) – difference in the object appearance because of time or seasonal changes, occlusions, turbulent air, etc.

We present a new public dataset (see Figure 1) with ground truth which combines the above-mentioned challenges and contains both W2BS image pairs including viewpoint and appearance, viewpoint and illumination, viewpoint and sensor, illumination and appearance change and W3BS – problems where viewpoint, appearance and lighting differ significantly.

We propose a novel algorithm for two-view matching in challenging conditions – WxBS-MODS (Algorithm 1). It significantly outperforms the state-of-the-art matchers: ASIFT [2], Dual Bootstrap (DBstrap) [3] and MODS [1] on various WxBS problems without a significant loss of speed (Table 1).

Algorithm 1 MODS-WXBS – a matcher for wide multiple baseline stereo

Input: I_1, I_2 – two images; θ_m – minimum required number of matches; $S_{\max}$ – maximum number of iterations.
Output: Fundamental or homography matrix F or H; a list of corresponding local features.

while ($N_{\text{matches}} < \theta_m$) *and* (Iter $< S_{\max}$) **do**
  for I_1 and I_2 separately **do**
    1 Generate synthetic views acc. to the scale-tilt-rotation-detector setup for Iter.
    2 Detect local features using adaptive threshold.
    3 Extract rotation invariant descriptors with:
    3a rSIFT and **3b hrSIFT**
    4 Reproject local features to I_1.
  end for
  5 Generate tent. corresp. based on the first geom.inconsistent rule for rSIFT and hrSIFT separately using kD-tree
  6 Filter duplicates
  7 Geometric verification of *all* TC with modified DEGENSAC estimating F or H.
  8 Check geom. consistency of the LAFs with est. F or H.
end while

[1] Dmytro Mishkin, Michal Perdoch, and Jiri Matas. Mods: Fast and robust method for two-view matching. *CoRR*, abs/1503.02619, 2015.
[2] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.
[3] Gehua Yang, Charles V Stewart, Michal Sofka, and Chia-Ling Tsai. Registration of challenging image pairs: Initialization, estimation, and decision. *PAMI 2007*, 29(11):1973–1989, 2007.

Table 1: Comparison of MODS-WxBS, ASIFT and Dual Bootstrap on public datasets. The number of matched image pairs (left) and the average running time (right).

Alg. / Dataset	EF		EVD		MMS		WGABS		WGALBS		WGLBS		WGSBS		WLABS		Past		OxAff		SymB		GDB	
	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]	#	time [s]
	33		15		100		5		8		9		5		4		172		40		46		22	
State-of-art matchers																								
ASIFT	23	27	5	12	18	3.2	0	52	0	32	0	35	0	12	1	30	62	32	40	102	27	14	15	41
MODS	33	4.8	15	11	27	11	2	41	2	31	1	46	0	17	1	26	94	27	40	3.4	42	18	18	11
DBstrap	31	26	0	18	79	9.3	0	11	0	13	0	13	0	4.7	0	15	16	28	36	24	38	21	16	17
Proposed matcher																								

Planar shape decomposition made simple

Nikos Papanelopoulos
papanelo@image.ntua.gr
National Technical University of Athens
Athens, Greece

Yannis Avrithis
http://image.ntua.gr/iva/iavr/
University of Athens
Athens, Greece

THE psychophysical, ecological, and computational aspects of planar shape decomposition into parts have been studied for more than five decades [9]. Although a complete theory of object recognition remains an impossibility, it is believed that our ability to recognize objects by their silhouette alone is related to simple rules by which the visual system decomposes shapes into parts [4]. In computer vision, object detection and recognition has deviated from such studies, but understanding visual perception towards learning better representations is always relevant.

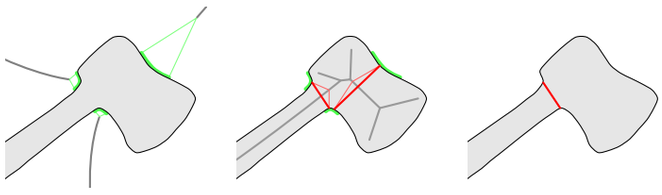
Recent work on the subject has introduced ever more complex computational models relying on combinatorial optimization [6, 7]. The main focus of such models is *convexity*, although the support from psychophysical studies is limited or absent [8]. The most recognized rules underpinning shape decomposition are the *minima rule* [4] and the *short-cut rule* [11], along with the definition of *part-cuts* [10]. However, attempts to reflect these rules into simple computational models still resort to optimization and new ad-hoc rules [5]. Although the medial axis has been one of the first representations used even before the formulation of these rules [1], it is not frequently used today. On the other hand, quantitative evaluation has been practically impossible until recently [3].

In this work, we revisit the problem assuming the medial axis representation and introduce a new computational model referred to as *medial axis decomposition* (MAD). Contrary to common belief [5], we argue that this representation is both efficient and robust, at least as far as decomposition is concerned, and as long as a part hierarchy [9] is not sought. We show that it is possible to incorporate all rules suggested by psychophysical studies into a computational model that is so simple that one nearly "reads off" part-cuts from the medial axis. In doing so, we suggest a stronger definition of part-cuts concerning local symmetry such that the list of candidate cuts is linear in the number of minima. We also shed more light into the relation of minima to convexity by relaxing the latter to *local convexity*. Contrary to global optimization models, this guarantees robustness [9].

The main ideas of our work are illustrated in Fig. 1. As in most related work, a shape is decomposed into parts by defining a number of part-cuts which are line segments contained in the shape. According to the minima rule [4], the part-cut endpoints are points of *negative minima of curvature* of the shape boundary curve. But it is known [2] that such points are exactly projection points (boundary points of minimal distance) of end vertices of the exterior medial axis (the medial axis of the complement of the shape). Moreover, as shown in Fig. 1a, one may get from a medial axis vertex not just one boundary point but an entire arc. We call this arc a *concave corner* or simply *corner*. It is readily available and involves no differentiation, contrary to all previous work. We show there are advantages over the common single-point approach.

There is no constraint as to which pairs of minima (corner points) are candidate as part-cut endpoints, hence all prior work examines all possible pairs. On the contrary, as shown in Fig. 1b, we only consider pairs of points that are projection points of the same point of the interior medial axis (of the shape itself). Similarly to *semi-ligatures* [1] and single-minimum cuts [5], a cut may also have only one corner point as endpoint [11]. In either case, endpoint pairs are readily available by a single traversal of the medial axis. Comparing to the conventional definition, which requires part-cuts to cross an axis of local symmetry [10], this is a stronger definition in agreement with the definition of *necks* [9]. Contrary to common belief, we show that it can actually be in accordance to psychophysical evidence [3]. For each corner, we only select one cut per medial axis branch; this is a simple and intuitive rule that has not been observed before.

Now, given a candidate list of cuts, the *short-cut* rule [11] suggests that priority be given to the shortest over all cuts incident to each corner point; but it does not specify how many should be kept. On the other hand, convexity-based approaches attempt to find a minimal number of cuts such that each shape part is convex [7]. Clearly, a concave smooth boundary curve segment would require an infinite partition, so convexity is only sought approximately. But negative minima of curvature are points where the shape is locally maximally concave. They are therefore the first points where one should establish convexity by cutting. Hence we introduce a *local convexity* rule whereby the minimal number of cuts is selected such that the interior angle of each part is less than π (up to tolerance) at each corner. Selection is linear in the number of candidate cuts and again, all information is merely read-off from the (exterior) medial axis. The final cuts are shown in Fig. 1c.



(a) exterior (b) interior (c) cuts

Figure 1: Main elements of our method. (a) Exterior medial axis and concave corners (in green) as boundary arcs that are each the projection of one medial axis end vertex (minima rule). (b) Interior medial axis and candidate cuts (in red) whose endpoints are contained in corners and are projection points of the same medial axis point; only one such cut is selected per corner and medial axis branch. (c) Final cuts according to short-cut and convexity rules: the shortest cuts are selected for each corner such that each shape part is locally convex at the corner, roughly forming an interior angle less than π (up to tolerance).

[1] Jonas August, Kaleem Siddiqi, and Steven W Zucker. Ligature instabilities in the perceptual organization of shape. *CVIU*, 76(3):231–243, 1999.

[2] H.I. Choi, S.W. Choi, and H.P Moon. Mathematical theory of medial axis transform. *Pacific Journal of Mathematics*, 181(1):57–88, 1997.

[3] Joeri De Winter and Johan Wagemans. Segmentation of object outlines into parts: A large-scale integrative study. *Cognition*, 99(3): 275–325, 2006.

[4] Donald D Hoffman and Whitman A Richards. Parts of recognition. *Cognition*, 18(1):65–96, 1984.

[5] Lei Luo, Chunhua Shen, Xinwang Liu, and Chunyuan Zhang. A computational model of the short-cut rule for 2d shape decomposition. *IEEE Transactions on Image Processing*, 24(1):273–283, 2015.

[6] Chang Ma, Zhongqian Dong, Tingting Jiang, Yizhou Wang, and Wen Gao. A method of perceptual-based shape decomposition. In *ICCV*, 2013.

[7] Zhou Ren, Junsong Yuan, Li Chunyuan, and Liu Wenyu. Minimum near-convex decomposition for robust shape representation. In *ICCV*, 2011.

[8] Paul L Rosin. Shape partitioning by convexity. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 30(2):202–210, 2000.

[9] Kaleem Siddiqi and Benjamin B Kimia. Parts of visual form: Computational aspects. *PAMI*, 17(3):239–251, 1995.

[10] Manish Singh and Donald D Hoffman. Part-based representations of visual shape and implications for visual cognition. *Advances in Psychology*, 130:401–459, 2001.

[11] Manish Singh, Gregory D Seyranian, and Donald D Hoffman. Parsing silhouettes: the short-cut rule. *Perception and Psychophysics*, 1999.

Adaptation of Synthetic Data for Coarse-to-Fine Viewpoint Refinement

Pau Panareda Busto[1,2]
pau.panareda-busto@airbus.com
Joerg Liebelt[1]
joerg.liebelt@airbus.com
Juergen Gall[2]
gall@iai.uni-bonn.de

[1] Airbus Group Innovations
Munich, Germany

[2] Computer Vision Group
Institute of Computer Science III
University of Bonn, Germany

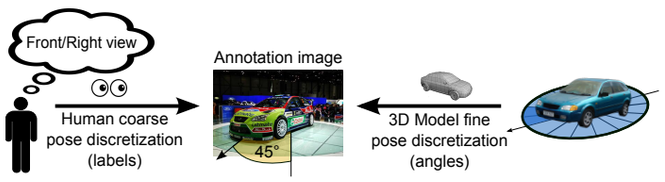


Figure 1: Humans are perfect for annotating coarse viewpoints of objects in real images, but fail to estimate pose accurately at a fine level. 3D graphic models can be used to synthesize data at very accurate fine angles, but it is time-consuming to model all appearance variations present in real images. We therefore propose to leverage the abilities of humans of estimating coarse viewpoints and the pose accuracy of synthetic data.

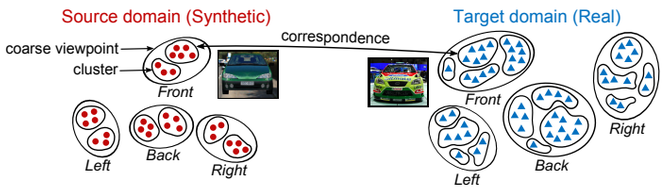


Figure 2: Synthetic data (source) is domain-adapted towards the real data (target) with a transformation estimated by minimizing the distance of correspondences between both domains. Each cluster in the target domain is assigned to a source cluster that belongs to the same coarse viewpoint. In this example, for an 8-view refinement: $V_i = 2$ and $K_i = 4$.

		8 views	16 views	32 views
	gt	*80.06*	*73.57*	*60.59*
w/o DA	syn	65.98	60.92	46.55
	real	76.04	65.46	49.90
	joint	72.52	63.81	50.04
with DA	syn	74.62	67.01	51.06
	real	**78.37**	69.04	**55.22**
	joint	75.73	**71.93**	53.00

Table 1: Pose estimation accuracy on test data using synthetic data, viewpoint refined real data or both training sets. In EPFL Dataset [4].

The quality of learning-based pose estimation still heavily relies on manual training data annotations. However, the manual labeling of large datasets is costly [6] and frequently limited to a few coarse viewpoint annotations of varying accuracy [5]. In this work, we propose to refine such coarse pose annotations with a domain adaptation approach, where the source domain consists of fine-grained pose annotations generated from synthetic computer graphics models, and the target domain of coarse manual pose annotations of a real dataset (see Figure 1), *i.e.*, we ask humans to annotate only four coarse views, namely front, back, left and right view.

Our domain adaptation technique starts by clustering the source and target domains and establishes correspondences between the clusters for each coarse viewpoint (see Figure 2). To cluster the synthetic data, we use the known fine-grained poses where each pose can be associated with one of the four coarse viewpoints $i = \{\text{front}, \text{back}, \text{left}, \text{right}\}$, *i.e.*, $V = \sum_i V_i$ where V is the total number of fine-grained poses. For the target domain, we only have the coarse viewpoints and thus we cluster the N_i training samples of one viewpoint further by K-Means where the number of clusters for each coarse viewpoint is given by K_i, *i.e.*, $K = \sum_i K_i$ and $V_i \leq K_i \leq N_i$. We represent each image by a HOG feature vector and append the aspect ratio of the bounding box surrounding the object. To this end, we represent each cluster by its centroid. The sets of centroids are denoted by $\hat{S}^i = \{\hat{s}^i_1, ..., \hat{s}^i_{V_i}\}$ and $\hat{T}^i = \{\hat{t}^i_1, ..., \hat{t}^i_{K_i}\}$. The correspondences are then established by solving a bipartite matching problem:

$$\begin{aligned} &\underset{e_{vk}}{\operatorname{argmin}} \sum_{v=1}^{V_i} \sum_{k=1}^{K_i} e_{vk} \left\| \hat{s}^i_v - \hat{t}^i_k \right\|_2^2 \\ &\text{subject to } \sum_v e_{vk} = 1 \ \ \forall k \ , \ \ \sum_k e_{vk} = a_v \ \ \forall v \ \text{ and } \ e_{vk} \in \{0,1\} \ \ \forall v,k \ . \end{aligned} \tag{1}$$

It assigns to each cluster in the target domain a unique cluster in the source domain. Since there can be more clusters in the target domain than in the source domain, each source is associated to $a_v = \frac{K_i}{V_i}$ target clusters. We use the Hungarian algorithm [3] to solve the problem.

The correspondences are then used to learn a mapping from the source domain, $\mathcal{S} \in \mathbb{R}^D$, to the target domain, $\mathcal{T} \in \mathbb{R}^D$, where D denotes the dimensionality of the features. We consider a linear transformation, which is represented by a matrix $W \in \mathbb{R}^{D \times D}$, *i.e.*, $t = Ws$. Let $S = \{s_1, ..., s_M\}$ and $T = \{t_1, ..., t_N\}$ denote the training samples of the source and target domain, respectively. M and N are the total amount of samples of each domain and we can assume that $M \geq N$ since we can always generate more synthetic data than annotated real images. Given the correspondences $C = \{c_1, ..., c_K\}$ with (s_{c_k}, t_k) and $K \leq N$, W can be learned by minimizing the objective

$$f(W) = \frac{1}{2} \sum_{k=1}^{K} ||W s_{c_k} - t_k||_2^2 \tag{2}$$

Lastly, the viewpoint refinement of the real training images is seen as a classification problem, where we train on the transformed synthetic samples a linear SVM for each of the fine viewpoints $v = \{1, ..., V\}$. Then, we apply the linear SVMs corresponding to the coarse viewpoint i of the real image and assign the fine pose with the highest scoring function:

$$f(x, i) = \underset{v=\{1,...,V_i\}}{\operatorname{argmax}} \ w_v^T x + b_v, \tag{3}$$

where w_v and b_v are the weights and bias of the linear SVM for the fine viewpoint v. The performance of our viewpoint refinement is evaluated in the paper on well-known car datasets with annotated poses, outperforming popular domain adaptation techniques (*e.g.* Geodesic flow kernel [1], Maximum margin domain transform [2]).

For further pose estimation on real test images, we also use linear SVMs in the same one-vs-all classification procedure. For each fine viewpoint, we evaluate trained SVMs using the real training images with refined pose labels, the synthetic training images, which have been transformed by domain adaptation, and both combined. As observed in Table 1, the accuracy increases in all setups when using domain adaptation.

[1] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[2] J. Hoffman, E. Rodner, J. Donahue, B. Kulis, and K. Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *IJCV*, 2014.

[3] H. W. Kuhn. The hungarian method for the assignment problem. *NRLQ*, 1955.

[4] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[5] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[6] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond Pascal: A benchmark for 3D object detection in the wild. In *WCACV*, 2014.

Saliency Prediction with Active Semantic Segmentation

Ming Jiang[1]
mjiang@u.nus.edu
Xavier Boix[1,3]
elexbb@nus.edu.sg
Juan Xu[1]
jxu@nus.edu.sg
Gemma Roig[2,3]
gemmar@mit.edu
Luc Van Gool[3]
vangool@vision.ee.ethz.ch
Qi Zhao[1]
eleqiz@nus.edu.sg

[1] Department of Electrical and Computer Engineering
National University of Singapore
Singapore

[2] CBMM, LCSL
Massachusetts Institute of Technology
Istituto Italiano di Tecnologia
Cambridge, MA

[3] Computer Vision Laboratory
ETH Zurich
Switzerland

Semantic-level features have been shown to provide a strong cue for predicting eye fixations. They are usually implemented by evaluating classifiers everywhere in the image. As a result, extracting the semantic-level features may become a computational bottleneck that may limit the applicability of saliency prediction in real-time applications.

To tackle this problem, we show a saliency prediction algorithm that uses active semantic segmentation [3] to exploit semantic-level features in a computationally efficient way (see Figure 1). The active semantic segmentation framework assumes that evaluating classifiers everywhere in the image is computationally more expensive than inferring the semantic labeling from a probabilistic model. Given a budget of time, active semantic segmentation evaluates classifiers in a subset of regions, and propagates the semantic information to the regions that have not been observed. In this paper, we introduce new semantic-level features for saliency prediction based on the probabilistic output from active semantic segmentation. To carry the analysis of our model, we collected eye tracking data on two popular segmentation datasets, the PASCAL VOC07 [1] and the MSRC-21 [4]. Experiments show that the semantic-level features extracted from active semantic segmentation improve the saliency prediction from low- and regional-level features, and it allows controlling the computational overhead of adding semantics to the saliency predictor.

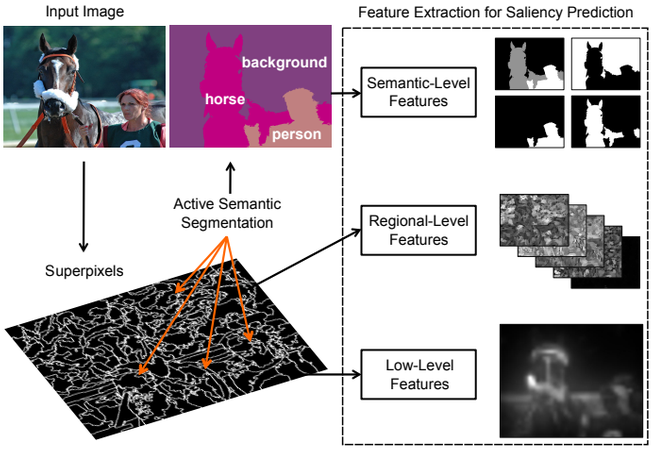


Figure 1: We propose a saliency prediction framework incorporating low-level, regional-level and new semantic-level features based on an active semantic segmentation of the input image. This allows to control the computational overhead of adding semantics to the saliency prediction.

Method

We use a Support Vector Regression (SVR) to learn a feature integration for saliency prediction. Our proposed model incorporates various saliency features at three levels:

Low-Level Features. We incorporate the state-of-the-art GBVS [2] method to calculate a low-level saliency feature, with combined color, intensity and orientation channels.

Regional-Level Features. Suggested by the Gestalt principles, locally coherent regions or proto-objects, are more likely to attract attention than others depending on the size and shape. We use five regional-level features (*i.e. size*, *solidity*, *convexity*, *complexity*, *eccentricity*) proposed by [5], which are independent of the semantics.

Semantic-Level Features. Based on the output of the active semantic segmentation (*i.e.* a set of semantic labelings of the full image that define the probability distribution of the semantics given few observations on the image), we compute a set of semantic-level features, including *label probability*, *semantic uncertainty*, *semantic rarity*, and *object center*.

Experiments and Dataset

We conducted eye-tracking experiments on the PASCAL VOC07 [1] and the MSRC-21 [4] datasets. Evaluation results demonstrated the effectiveness of the semantic features for saliency prediction under computational time constraints (see Figure 2). The eye-tracking data is publicly available at `http://www.ece.nus.edu.sg/stfpage/eleqiz/bmvc15.html`.

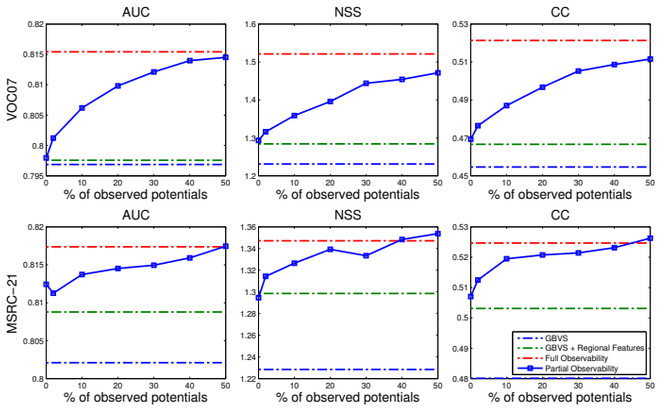


Figure 2: Performance evaluation of saliency prediction in the VOC07 and MSRC-21 datasets, with various percentages of observations. The semantic-level features are able to capture useful semantic information for saliency prediction, given a budget of time.

[1] M. Everingham, L. Van Gool, C. KI Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. Journal of Computer Vision*, 88(2):303–338, 2010.

[2] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2007.

[3] G. Roig, R. Boix, X.and de Nijs, S. Ramos, K. Kühnlenz, and L. Van Gool. Active map inference in crfs for efficient semantic segmentation. In *Proc. IEEE Int. Conf. on Computer Vision*, 2013.

[4] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. Journal of Computer Vision*, 2009.

[5] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *J. of Vision*, 14(1):1–20, January 2014.

Dr. Qi Zhao is the corresponding author.

Using Near-Field Light Sources to Separate Illumination from BRDF.

Jeroen Put
jeroen.put@uhasselt.be
Nick Michiels
nick.michiels@uhasselt.be
Philippe Bekaert
philippe.bekaert@uhasselt.be

Hasselt University - tUL - iMinds
Expertise Centre for Digital Media
Wetenschapspark 2
3590 Diepenbeek, Belgium

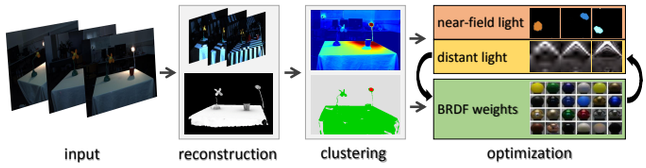


Figure 1: Schematic overview of our method.

Simultaneous estimation of lighting and BRDF (Bidirectional Reflectance Distribution Function) from multi-view images is an interesting problem in computer vision. It allows for exciting applications, such as flexible relighting in post-production, without recapturing the scene. The ability to alter scenes after they have been filmed has the potential to greatly reduce the number of costly recapturing iterations. Unfortunately, the estimation problem is made difficult because lighting and BRDF have closely entangled effects in the input images. This paper presents an algorithm to support both the estimation of distant and near-field illumination. Previous techniques are limited to distant lighting. We contribute by proposing an additional factorization of the lighting, while keeping the rendering efficient and additional data compactly stored in the wavelet domain. We reduce complexity by clustering the scene geometry into a few groups of important emitters and calculate the emitting powers by alternately solving for illumination and reflectance. We demonstrate our work on a synthetic and real datasets and show that a clean separation of distant and near-field illumination leads to a more accurate estimation and separation of lighting and BRDF.

The input of our algorithm is a sparse set of (usually 5-16) different views of a scene with unknown reflectance, captured under unknown illumination. This makes our method very broadly applicable. The outputs of our algorithm are the estimated wavelet coefficients of the distant lighting environment map, the estimated emitting powers for every emitting cluster in the scene and a set of per-vertex material weights that determine a mixture of BRDF models [3].

We can summarize our contributions in a few key points:

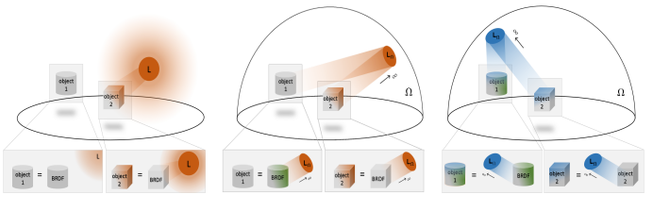
1. A factorization of illumination in distant and near-field lighting. This allows us to model local lighting effects, many of which are ignored by current methods (Figure 2).

2. A clean separation of distant and near-field lighting allows more accurate estimation of illumination and materials in a scene and permits the manipulation of local light sources. It also allows using knowledge of the light sources in a scene. For example, light source positions can be identified by a HDR clustering procedure and help improve the estimation of other factors.

3. We formulate the problem so that the wavelet coefficients of the distant lighting environment map and the emitting powers of the near-field lighting can be solved together. Any standard quadratic programming framework can be used to solve this problem robustly and efficiently.

4. We propose the use of clustering to group the most significant emitting triangles into dominant light sources. This grouping constrains the solver to more plausible solutions. We make the assumption that the distribution of light is roughly uniform along the emitting surface.

We propose factorizing illumination in distant and near-field lighting components. The distant illumination is expressed in wavelet coefficients [2], which are estimated directly. Then, we further factorize the



(a) (b) (c)

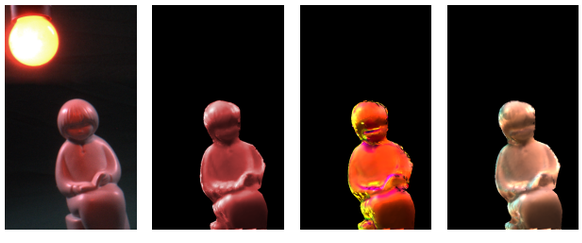
Figure 2: (a) A scene with 2 objects and a red local light source L. (b) Current techniques model local lighting effects with distant light sources at infinity. This causes erroneous correction colors (shown here as green) in the BRDF. (c) Relighting reveals the wrongly estimated materials.

near-field lighting term at every vertex. Rendering with the factorized representation is equivalent to adding an extra double product integral of the BRDF and the irradiance maps to the rendering equation. Because Haar wavelets are an orthogonal basis, this double product integral term reduces to a simple dot product calculation [2]:

$$B(v,\omega_0) = \underbrace{\sum_k \sum_l \sum_m C_{klm}\rho_k V_l \tilde{L}_m}_{\text{distant lighting}} + \underbrace{\sum_c \sum_k \rho_k \cdot (I_c P_c)}_{\text{near-field lighting}} \quad (1)$$

For every cluster c, visible over the hemisphere of vertex v, its irradiance map I_c is multiplied by the estimated power P_c. Finally, it is multiplied by the BRDF ρ at v. At the end, all this information is assembled in bilinear equations and finally into matrix form. The unknowns are then solved for with a quadratic programming solver. Figure 1 gives a summary of our algorithmic pipeline.

Figure 3 shows the improvement that our new factorization offers. By taking near-field lighting effects into account, material and lighting properties of objects can be more faithfully recovered.



(a) (b) (c) (d)

Figure 3: A white statuette under a red light source dataset. (a) One of the input images. (b) Reconstruction after optimization. (c) The technique of Haber et al. [1] reveals compensation colors in materials when rendered under white lighting. (d) Material estimated by our technique rendered under white lighting.

[1] T. Haber, C. Fuchs, P. Bekaert, H.-P. Seidel, M. Goesele, and H.P.A. Lensch. Relighting objects from image collections. In *CVPR 2009*, pages 627–634, June 2009.

[2] Ren Ng, Ravi Ramamoorthi, and Pat Hanrahan. Triple product wavelet integrals for all-frequency relighting. *ACM Trans. Graph.*, 23(3):477–487, August 2004. ISSN 0730-0301.

[3] R. Peter Weistroffer, Kristen R. Walcott, Greg Humphreys, and Jason Lawrence. Efficient basis decomposition for scattered reflectance data. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, EGSR'07, pages 207–218, 2007. ISBN 978-3-905673-52-4.

Multi-scale Graph-based Guided Filter for De-noising Cryo-Electron Tomographic Data

Shadi Albarqouni[12]
http://campar.in.tum.de/Main/ShadiAlbarqouni
Maximilian Baust[1]
http://campar.in.tum.de/Main/MaximilianBaust
Sailesh Conjeti[1]
http://campar.in.tum.de/Main/SaileshConjeti
Ashraf Al-Amoudi[2]
https://www.dzne.de/en/sites/bonn/research-groups/al-amoudi.html
Nassir Navab[13]
http://campar.in.tum.de/Main/NassirNavab

[1] Chair for Computer Aided Medical Procedure (CAMP), Technische Universität München (TUM) Munich, Germany
[2] Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany
[3] Chair for Computer Aided Medical Procedure (CAMP), Johns Hopkins University Baltimore, MD, USA.

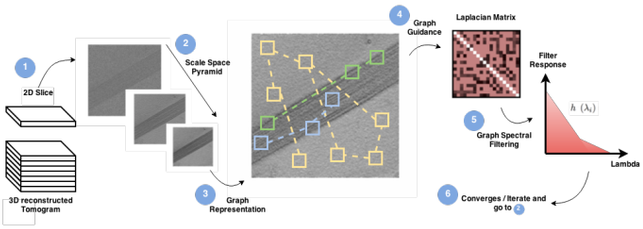


Figure 1: **MG^{2}F Framework:** A noisy image slice from the 3D reconstructed tomogram is fed to the algorithm, where the graph is built on a selected scale space image (i.e. coarse grid) acting as a guidance for the regularized graph spectral filter.

1 Introduction

Cryo-electron tomography (CET) is a powerful imaging technique in biological sciences which bridges the gap between the molecular and the cellular structural biology [5], giving a better understanding of protein interactions and thus better drug delivery strategies. In principle, similar to Computed Tomography (CT) in Medical Imaging, the acquired projections at limited angles are reconstructed back to create the 3D object, however, these projections are extremely noisy and have a low contrast. Therefore, many conventional filters failed in smoothing the background while preserving edges and interesting objects, which makes developing a denoising algorithm is very desirable for better interpretation.

We show in this paper how our methodology meets the hypothesis: a) By using a multi-scale pyramid for guidance we are able to detect meaningful scales and use them for guidance without oversmoothing fine scale structures. b) Using a patch-based approach, we can take advantage of redundant structures in the whole image rather than using a pre-defined spatial window for averaging similar pixels or patches. This way, we can preserve the local and global consistencies. c) By deriving explicit solution formulas for computing the intermediate filtering results we obtain an efficient algorithm.

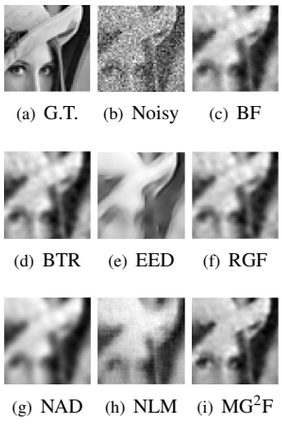
2 Methodology

Given a noisy image I_η, we collect N overlapping patches, which can be seen as data points $v = \{v_1, v_2, ..., v_N\} \in \mathbb{R}^{n\times N}$ lying on a manifold $\mathcal{M}$ embedded in $\mathbb{R}^n$ space such that $v = EI_\eta$, where E is an operator collecting patches and vectorize it, cf. Figure1. The relation between the data points can be represented by a k-NN connected, undirected, and weighted graph $G = \{v, \varepsilon, \omega\}$, where ε is the set of edges, and ω is the set of edge weights.

These weights are assigned using a heat kernel, however, the distance between these patches is computed on a certain structure scale σ_s where the noise manifest itself and can be used as a guidance for the graph spectral filter $h(\lambda_i)$, which is computed based on the eigenvalue decomposition of the normalized Laplacian matrix $\tilde{L_{\sigma_s}} := U\Lambda U^T$. This way, we can formulate the denoising problem as follows:

$$\hat{I_f} = \arg\min_{I_f} \left\{ \frac{1}{2}\|I_f - I_\eta\|_2^2 + \alpha S_{\sigma_s}(I_f) \right\}, \qquad (1)$$

where $\alpha > 0$ is the regularization parameter and $S_{\sigma_s}(I_f) = \frac{1}{2} Tr\left(vL_{\sigma_s}v^T\right)$ is the graph guidance regularization term.



Algorithm Parameters	PSNR (dB)
Bilateral (BF) [7] (σ_i=0.5, σ_r=1.5, W=10)	17.49
Beltrami (BTR) [2] (δ=0.1, *iter*=10)	17.37
EED [6] (ρ=4, *iter*=30)	11.27
NAD [4] (*iter*=10, κ=0.3)	16.50
NLM [1] (P=7, W=21, σ_s=4σ_n)	12.11
RGF [3] (σ_i=0.5, σ_r=1.5, *iter*=10)	17.49
MG^{2}F (α=0.8, *iter*=4, σ_h=0.1)	**17.78**

Figure 2: **Photographic Image:** Results of different algorithms on Lena image(128X128, SNR=7) along with a tabulated comparison to the proposed MG^{2}F filter.

The closed form solution can be written as

$$\hat{I_f} = E^T\left(\sum_{i=1}^{N} \frac{1}{(1+\alpha\lambda_i)} u_i \hat{v}_i\right) = E^T\left(\frac{1}{I+\alpha\tilde{L}_{\sigma_s}}\right) EI_\eta, \qquad (2)$$

where E^T denotes the reshaping process of the previously vectorised patches, and the spectral response of the filter $h(\lambda_i) = 1/(1+\alpha\lambda_i)$ controls the frequency decay and thus the degree of smoothness. A connection to classical filters and the sensitivity analysis are discussed in details.

3 Results

To give a good illustrative example, we run the algorithm on Lena image, which corrupted by an (i.i.d) Gaussian noise resulting in SNR of 7. Different algorithms are applied on this image, results are shown in Figure 2 for the cropped images. It is clear that our method gives an outperforming PSNR indicating for better contrast. A simulated and real CET data experiments in 2D and 3D are presented in the paper. Using the gold-standard metrics, we show that our denoising algorithm significantly outperforms the state-of-the-art methods such as NAD, NLM and RGF in terms of noise removal and structure preservation.

4 References

[1] J. Darbon, A. Cunha, T.F. Chan, S. Osher, and G.J. Jensen. Fast nonlocal filtering applied to electron cryomicroscopy. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 1331–1334, May 2008. doi: 10.1109/ISBI.2008.4541250.

[2] Jose-Jesus Fernandez. Tomobflow: feature-preserving noise filtering for electron tomography. *BMC bioinformatics*, 10(1):178, 2009.

[3] David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars, Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. Rolling guidance filter. In *Computer Vision –ECCV 2014*, volume 8691 of *Lecture Notes in Computer Science*, pages 815–830. Springer, 2014. ISBN 978-3-319-10577-2.

[4] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629–639, Jul 1990. ISSN 0162-8828. doi: 10.1109/34.56205.

[5] Carol V. Robinson, Andrej Sali, and Wolfgang Baumeister. The molecular sociology of the cell. *Nature*, 450(7172): 973–982, 12 2007. URL http://dx.doi.org/10.1038/nature06523.

[6] Bart ter Haar Romeny, Luc Florack, Jan Koenderink, Max Viergever, and Joachim Weickert. *A review of nonlinear diffusion filtering*, volume 1252, pages 1–28. Springer Berlin Heidelberg, 1997. ISBN 978-3-540-63167-5. doi: 10.1007/3-540-63167-4{_}37. URL http://dx.doi.org/10.1007/3-540-63167-4_37.

[7] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846, Jan 1998. doi: 10.1109/ICCV.1998.710815.

Tuesday Poster 1

Spotlight the Negatives: A Generalized Discriminative Latent Model

Hossein Azizpour[1]
www.csc.kth.se/~azizpour/
Mostafa Arefiyan[2]
mostafa@brown.edu
Sobhan Naderi Parizi[2]
http://vision.lems.brown.edu/students/Parizi
Stefan Carlsson[1]
stefanc@csc.kth.se

[1] CVAP
Royal Institute of Technology (KTH)
Stockholm, Sweden
[2] LEMS
Brown University
Providence, USA

Discriminative latent variable models (LVM) are frequently and successfully applied to various visual recognition tasks [1, 2, 3, 4, 5, 6, 7]. In these systems the latent (hidden) variables provide a formalism for modeling structured variation of visual features. These variations can be the result of different possible object locations, deformations, viewpoint, subcategories, etc. Conventionally, these have all been defined only based on the foreground (positive) class. We call such latent variables "positive". In this work we introduce generalized discriminative LVM (GLVM) which use both "positive" and "negative" latent variables. Negative latent variables are defined on the background (negative) class and provide *counter evidence* for presence of the foreground class. They can, for instance, learn mutual exclusion constraints, model scene subcategories where the positive object class is unlikely to be found, or capture specific parts which potentially indicate the presence of an object of a similar but not the same class.

The objective of the proposed framework is to learn a model which maximizes the evidence (characterized by positive latent variables) for the foreground class and at the same time minimizes the counter evidence (characterized by negative latent variables) lying in background class. Thereby GLVM empowers the latent variable modeling by highlighting the role of negative data in its own right. An interesting analogy is with game theory; when playing a game a simple strategy is to maximize your score. This resembles LVM's objective which is to maximize evidence for the presence of the foreground class. A better strategy, however, is to maximize your scores at the same time as minimizing the opponents' scores. This is analogous to GLVM's objective which takes into account counter evidence emerging from negative latent variables while looking for evidence from the foreground class. The score of the opponents is counter evidence in our hypothetical example.

The concept of *negative parts* was noted in [3]. However, [3] focuses on automatic discovery and optimization of a part based model with negative parts. In this paper we extend the notion of negative parts to negative latent variables and propose a framework for defining models that use both positive and negative latent variables.

Many different modeling techniques benefit from latent variables. Felzenszwalb *et al*. [1] introduced latent SVM to train deformable part models, a state of the art object detector [2]. Yu and Joachims [7] transferred the idea to structural SVMs. Yang *et al*. [6] proposed a kenelized version of latent SVM and latent structural SVM and applied it to various visual recognition tasks. Razavi *et al*. [4] introduced latent variables to the Hough transform and achieved significant improvements over a Hough transform object detection baseline. And-Or trees use latent variable modeling for object recognition and occlusion reasoning [5].

In this paper we introduce a novel family of models which generalizes discriminative latent variable models. We review LVMs and formulate our generalization of LVMs. Then, we discuss training of GLVMs and propose an algorithm for learning the parameters of a GLVM. Then we discuss how various computer vision models can benefit from GLVM. Finally, we experiment on generalized DPMs.

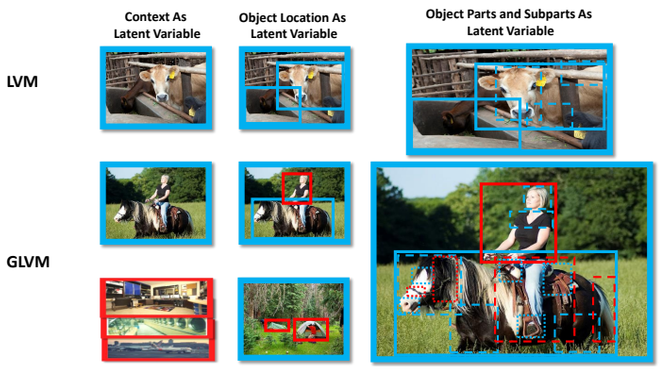


Figure 1: **Intuitive example of LVM and GLVM.** Assume the task of *cow* image classfication. This figure visualizes examples of possible latent variables for LVM and GLVM models. A latent variable can be assigned at image level such as the background scene and then gradually go down a hierarchy of objects, parts and subparts. Positive latent variables (modeling evidences for cow) of an LVM may correspond to different scenes where cow is likely to be seen *e.g*. stable, meadow (blue boxes in the first column). Another positive latent variable can be assigned to the location of the cow in the image (second column, first row). In addition, location of different parts of cow *e.g*. head, back, legs can be modeled as other latent variables dependent to the latent variable of cow location (top right). All these examples of latent variables in LVMs look for evidences for the existence of cow. In GLVM we encourage the modeling of counter evidences through negative latent variables. For example at high level a negative latent variable may correspond to scenes where cow is unlikely to be found *e.g*. office, laundromat, or airport (bottom left). Meadow is an evidence for cow, but cows are less likely to be seen in a meadow beside camping tents (a dependent negative latent variable to positive latent variable meadow). Horses are visually similar to cow but people usually do not ride cows thus a human on top of cow can be negative evidence (a negative latent variable dependent to the positive latent variable of cow location). Furthermore, human, a negative latent variable for cow, can have its own positive latent variable as parts *e.g*. face and shoulders (bottom right).

[1] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.
[2] Ross B. Girshick, Forrest N. Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015.
[3] Sobhan Naderi Parizi, Andrea Vedaldi, Andrew Zisserman, and Pedro F. Felzenszwalb. Automatic discovery and optimization of parts for image classification. In *ICLR*, 2015.
[4] Nima Razavi, Juergen Gall, Pushmeet Kohli, and Luc J. Van Gool. Latent Hough transform for object detection. In *ECCV*, pages 312–325, 2012.
[5] Xi Song, Tianfu Wu, Yunde Jia, and Song-Chun Zhu. Discriminatively trained and-or tree models for object detection. In *CVPR*, pages 3278–3285, 2013.
[6] Weilong Yang, Yang Wang, Arash Vahdat, and Greg Mori. Kernel latent SVM for visual recognition. In *NIPS*, pages 818–826, 2012.
[7] Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176, 2009.

Perceptual Dynamic Range for In-Camera Image Processing

Praveen Cyriac
praveen.cyriac@upf.edu
David Kane
david.kane@upf.edu
Marcelo Bertalmío
marcelo.bertalmio@upf.edu

Department of Information and Communication Technologies,
Universitat Pompeu Fabra,
Barcelona, Spain

Image sensors in digital cameras capture values which are proportional to light intensity and within a range of between 3-4 orders of magnitude. These values are typically passed through a non-linearity prior to quantisation to a range of around 2 orders of magnitude. The purpose of the non-linearity is to ensure that the quantisation rate is approximately proportional to human sensitivity (see [2] and references therein). The problem of reducing the dynamic range of an image while preserving the perceived detail, is known as tone mapping. Digital cameras generally do this by means of gamma correction, which is a very simple, image independent approach, designed to produce good results on average.

We propose an adaptive in-camera non-linearity that ensures that the detail and contrast visible in the processed image match closely with the perception of the original scene. The method has been developed to emulate basic properties of the human visual system including contrast normalisation and the efficient coding of natural images via adaptive processes which ensures that the output values are more evenly distributed over the available range. The transformations applied to the image are based on natural image statistics, psychophysical data and neurophysiological models. We have followed the same approach as Ferradans et al. in [4] and our method consists of two main stages, a global one followed by a local contrast enhancement one as shown in Figure 1:

1. Using natural image statistics, a function γ is defined by estimating its parameters (γ_L, γ_H and M) from the cumulative histogram of the intensity image in log-log coordinates.

$$\gamma(I) = \gamma_H + (\gamma_L - \gamma_H)(1 - \frac{I^n}{I^n + M_{lin}^n}), \tag{1}$$

   This function γ is used to perform a transform of the intensity values (Eq. (2)) in a manner that complies with users' preference data as obtained through psychophysical tests [6].

$$I_1(x) = (I(x))^{\gamma(I(x))}. \tag{2}$$

2. The output of the previous stage is passed through an additional contrast normalisation procedure that replicates efficient coding behaviour that occurs both in the retina and cortical areas of the human visual system.

$$O(x) = \mu(x) + (I_1(x) - \mu(x)) * k/\sigma, \tag{3}$$

   where x is a pixel, $I_1(x)$ is the value at pixel x computed at the previous stage of our method, $\mu(x)$ is the local mean of I_1, k is a constant, σ is the standard deviation of I_1, and $O(x)$ is the final output value of our method for pixel x.

Both equations (Eqs. (2) and (3)) are applied separately to each of the red, green and blue colour channels, which have previously been normalised into the range $[0,1]$ and where values above the 99 percentile have been clipped.

For video sequences, applying our method to each frame separately may result in flickering artefacts due to the possibility of a sudden change in the tone curve from one frame to the next. Therefore for video we adopt a two pass approach. In the first pass, we estimate γ_L, γ_H and M for each frame separately, then we apply a temporal low pass filter to these values. In the second pass, we apply our method (Eqs. (2) and (3)) with the new parameters.

In this paper we illustrate the potential of our method by comparing the regular JPEG output of the camera versus the result of applying our method to the corresponding RAW image. The results are illustrated

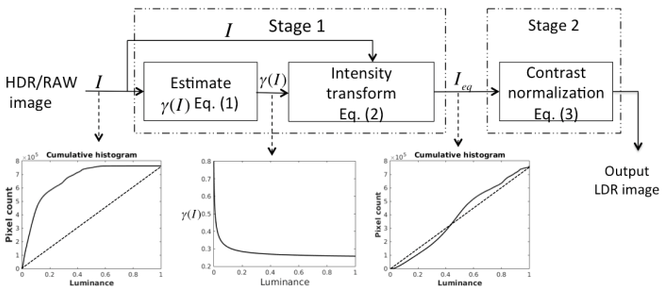


Figure 1: Block diagram of the proposed method.

with images from consumer, smartphone and cinema cameras. The results have natural appearance, enhanced contrast and absence of halos, spurious colours or visual artefacts of any kind. Our method can also be used offline as a tone mapping operator, applied to HDR images and we illustrated our result on two HDR data set [3][5]

We also perform a quantitative evaluation of our method. We use the Fairchild dataset [3] of HDR images and compare our algorithm with six state-of-the-art tone mapping operators, using the metrics DRIM [1] and TMQI [8]. According to DRIM, the amount of distortions produced by our approach matches that of Mantiuk et al. [7] and is significantly less than the error introduced by all other tested methods. In terms of TMQI, our method ranks as the best in terms of overall quality (Q), and matches to Mantiuk et al. [7] for structural fidelity (S). It is worth noting that our results have a very high naturalness (N) index.

The proposed approach is of low computational complexity and therefore it is a very good candidate for real-time applications.

[1] Tunç Ozan Aydin, RafałMantiuk, Karol Myszkowski, and Hans-Peter Seidel. Dynamic range independent image quality assessment. SIGGRAPH '08, pages 69:1–69:10, 2008.

[2] Marcelo Bertalmío. *Image Processing for Cinema*. CRC Press, 2014.

[3] Mark D Fairchild. The hdr photographic survey. In *Color and Imaging Conference*, volume 2007, pages 233–238, 2007.

[4] Sira Ferradans, Marcelo Bertalmio, Edoardo Provenzi, and Vicent Caselles. An analysis of visual adaptation and contrast perception for tone mapping. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):2002–2012, 2011.

[5] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays, 2014.

[6] David Kane and Marcelo Bertalmío. Is there a preference for linearity when viewing natural images? In *SPIE/IS&T 2015 Image Quality and System Performance XX*, 2015.

[7] RafałMantiuk, Scott Daly, and Louis Kerofsky. Display adaptive tone mapping. *ACM Trans. Graph.*, 27(3):68:1–68:10, August 2008.

[8] H. Yeganeh and Zhou Wang. Objective quality assessment of tone-mapped images. *Image Processing, IEEE Transactions on*, 22(2): 657–667, Feb 2013.

Robust Multiple Model Fitting with Preference Analysis and Low-rank Approximation

Luca Magri
luca.magri@unimi.it

Andrea Fusiello
andrea.fusiello@uniud.it

Department of Mathematics
Università degli Studi di Milano
Via Saldini, 50 - 20133 Milano, IT

DIEGM
Università degli Studi di Udine
Via delle Scienze, 208 - 33100 Udine, IT

Geometric multi-model fitting aims at extracting parametric models from unstructured data in order to organize and aggregate visual content in suitable higher-level geometric structures. This ubiquitous task can be encountered in many Computer Vision applications, for example in 3D reconstruction, in the processing of 3D point clouds, in face clustering, in body-pose estimation or video motion segmentation, just to name a few.

In practice, it is necessary to overcome the "chicken-&-egg dilemma" inherent to this problem: in order to estimate models one needs to first segment the data, but in order to segment the data it is necessary to know the models associated with each data point. The presence of multiple structures hinders robust estimation, which has to cope with both *gross* outliers and *pseudo*-outliers. Two somehow orthogonal strategies have been proposed in the literature in order to adress this challenging problem: consensus analysis and preference analysis. Consensus based methods, building on the RANSAC paradigm, instantiate a pool of tentative models and extract the strucutures that have maximal consensus. Preference oriented algorithms [2, 3] instead tackle this problem by the data point of view. Residuals between point and putative models are used in order to build a conceptual space in which points are portrayed by their preferences with respect to the instantiated strucures, The multi model fitting problem is then solved by clustering points in this preference space.

The method we present reduces the multi-model fitting task to many easier single robust model estimation problems, by combining preference analysis and robust low rank approximation. Three main step can be single out in our appraoch. At first data points are shifted in a conceptual space, where they are framed as a preference matrix Φ as shown in Fig. 1. Our conceptual representation makes use of an M-estimator in order to

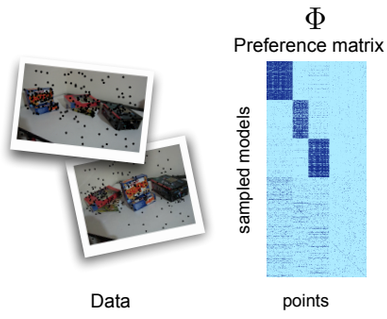


Figure 1: Preference representation

model points preferences, in this way a first protection against outlier is achieved. The preference space is then equipped with a kernel, based on the Tanimoto distance, in this way an affinity matrix K, which measures the agreement between the preferences of points, is derived.

The second step is devoted to robustly segment points explointing the information encapsulated in K. This stage can be thought as a sort of "*robust* spectral clustering". It is well known that spectral clustering produces accurate segmentations in two steps: at first data are projected on the space of the first eigenvectors of the Laplacian matrix and then k-means is applied. The shortcoming of spectral clustering however is that it is not robust to outliers. We propose to follow the same scheme enforcing robustness exploiting the low rank nature of the problem. As pictorially illustrated in Fig. 2, we decompose the affinity matrix as

$$K = UU^\top + S. \tag{1}$$

The matrix S models the sparse preferences expressed by outliers, and is obtained by appling Robust PCA, which replaces the eigen-decomposition

$$K = UU^\top + S$$

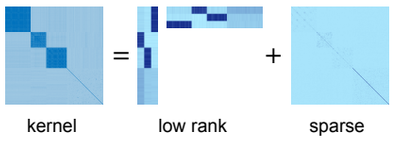


Figure 2: Robust low rank analysis

step of spectral clustering. The low rank part of K, representing symilarity between inliers, is hence decomposed as $UU^\top$ taking advantage of Symmetric NMF [1], which plays the role of k-means. The obtained matrix U represents a soft segmentation of the data in which outliers are underweighted.

Finally, models are extracted inspecting the product of the preference matrix with a thresholded U, mimicking the MSAC strategy. The use of robust statistics for adaptatively estimate the inlier threshold constituites a third guard against the presence of outliers.

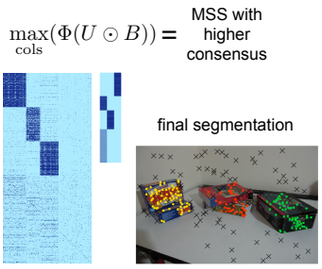


Figure 3: Models and segments extraction

We deal with two applications of geometric multi model fitting on real data: motion segmentation and plane segmentation. In the *motion segmentation* experiments, given two images of the same scene composed by several objects moving independently, the aim is to fit fundamental matrices to subsets of point matches. In *plane segmentation* scenario, given two uncalibrated views of a scene, the aim is to recover the multi-planar structures by fitting homographies to point correspondences. The experiments are carried on the AdelaideRMF [4] dataset, composed of 38 image pairs (19 for motion segmentation and 19 for plane segmentation) with matching points corrupted by gross outliers and have provided evidence that our method compares favourably with state of the art competing algorithms.

[1] Da Kuang, Sangwoon Yun, and Haesun Park. Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering. *Journal of Global Optimization*, pages 1–30, 2014.

[2] Luca Magri and Andrea Fusiello. T-linkage: A continuous relaxation of j-linkage for multi-model fitting. In *Computer Vision and Pattern Recognition*, pages 3954–3961, June 2014.

[3] R. Toldo and A. Fusiello. Robust multiple structures estimation with J-Linkage. In *European Conference on Computer Vision*, volume 5302, pages 537–547, 2008.

[4] H. S. Wong, T.-J. Chin, J. Yu, and D. Suter. Dynamic and hierarchical multi-structure geometric model fitting. In *International Conference on Computer Vision*, 2011.

Robust Global Motion Compensation in Presence of Predominant Foreground

Seyed Morteza Safdarnejad
https://www.msu.edu/~safdarne/
Xiaoming Liu
http://www.cse.msu.edu/~liuxm/
Lalita Udpa
http://www.egr.msu.edu/ndel/profile/lalita-udpa

Michigan State University
East Lansing
Michigan, USA

The objective of global motion compensation (GMC) is to remove *intentional* (due to camera pan/tilt/zoom) and *unwanted* (e.g., due to hand shaking) camera motion. GMC is utilized in applications such as video stitching, or as pre-processing for motion-based video analysis. Normally, GMC estimates the homography transformation between two consecutive frames by matching keypoints on the frames, and mapping the frames to a global coordinate. To remedy outliers in keypoint matches, robust techniques are proposed for homography estimation, e.g., RANSAC [1], by assuming the number of outliers to the correct homography is much less than inliers. However, in the presence of *predominant foreground*, i.e., moving objects and people, a larger proportion of the putative matches are mismatches. Predominant foreground may result from a higher percentage of coverage by foreground pixels, or occlusion, textureless and non-informative background, blurred background, or a combination of these reasons. In presence of predominant foreground, the common variations of RANSAC have little chance of selecting a minimal set of background keypoints by random sub-sampling in a limited number of iterations.

In this paper, we propose a robust GMC (RGMC) method for suppressing foreground keypoint matches and mismatches, enabling a reliable homography estimation in presence of predominant foreground and textureless background (Fig. 1). We perform *foreground suppression* by clustering motion vectors computed from keypoint matches and identifying potential clusters corresponding to the background. We use SURF algorithm for keypoint detection and description and to detect sufficient background keypoints even for textureless backgrounds, the Fast-Hessian keypoint detection threshold is decreased drastically. Since motion vectors on the background result from camera motion and are more *consistent* than foreground motion vectors, clustering will likely lead to some candidate regions from the background (see Fig. 1 (a)). Each cluster is analyzed separately by random subsampling of matches in that cluster and evaluating the resultant homography against the cost function, discussed later. However, background motion vectors may be assigned to multiple clusters. To merge background clusters, based on the estimated homography and cost function value (CFV) of each cluster, a subset of the best clusters are selected to be merged in a greedy algorithm (Fig. 1(b)).

To evaluate the estimated homography from a quadruplet of keypoints matches, we derive a cost function that unifies the keypoint matching score, edge matching score, and the information from compensating previous frames. Denote the matching frames as $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$, their candidate homography as θ_t, and the set of keypoint matches under study as $\mathbf{D}$. In Bayesian framework, θ_t can be estimated by maximizing

$$p(\theta_t|\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1},\theta_{t-1}) = \frac{p(\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t,\theta_{t-1})p(\theta_t|\theta_{t-1})}{p(\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_{t-1})}, \quad (1)$$

where θ_{t-1} is the obtained prior homography of frames $\mathbf{I}_{t-1}$ and $\mathbf{I}_{t-2}$. The $p(\theta_t|\theta_{t-1})$ is the conditional probability of θ_t given the prior homography θ_{t-1}. The denominator of Eqn. 1 is constant w.r.t. θ_t. By expanding the likelihood term, the homography can be verified using

$$p(\theta_t|\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1},\theta_{t-1}) \propto p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t,\theta_{t-1})p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t,\theta_{t-1})p(\theta_t|\theta_{t-1}). \quad (2)$$

The term $p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t,\theta_{t-1}) = p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t)$ and represents how well the keypoint matches $\mathbf{D}$ extracted from $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$ are matched by θ_t. Knowing $\mathbf{I}_t$ is independent from θ_{t-1}, $p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t,\theta_{t-1}) = p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t)$, reflects how well the frame $\mathbf{I}_t$ transformed under θ_t matches $\mathbf{I}_{t-1}$. Thus, the homography is estimated by minimizing,

$$f(\theta_t) = -\ln(p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t)) - \ln(p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t)) - \ln(p(\theta_t|\theta_{t-1})). \quad (3)$$

We use the keypoint matching error formulation of Yan et al. [2] for $p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t)$. For appearance consistency under θ_t transformation,

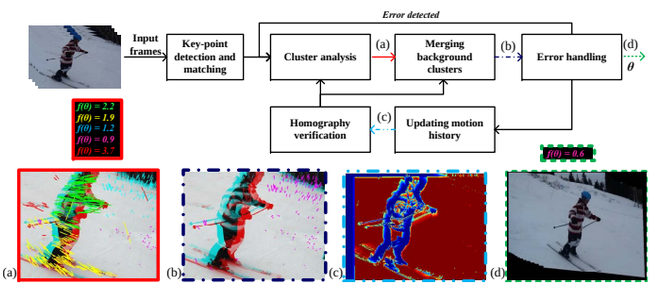


Figure 1: RGMC algorithm flowchart: (a) color indicates various motion vector clusters, (b) the merged cluster of background, (c) the motion history, and (d) the motion compensated video.

$p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t)$, we propose edge-based matching for its superior accuracy and efficiency.

To insure that the edge matching score reflects how well the background, not foreground, of the two frames match, we iteratively update a motion history $\mathbf{M}_t$. We define the edge matching score (EMS) as

$$E(\mathbf{I}_1,\mathbf{I}_2,\mathbf{R}) = \frac{2\|\Phi(\mathbf{I}_1)\bigodot\Phi(\mathbf{I}_2)\bigodot\mathbf{R}\|_1}{\|\Phi(\mathbf{I}_1)\bigodot\mathbf{R}\|_1+\|\Phi(\mathbf{I}_2)\bigodot\mathbf{R}\|_1+c}, \quad (4)$$

where Φ is edge detection operator, $\bigodot$ is element-wise multiplication, $\mathbf{R}$ denotes the mask specifying the region of interest for EMS calculation, $\|\cdot\|_1$ computes the L_1 matrix norm, and $c(=0.001)$ is a constant to avoid division by zero. We use thresholded motion history $\mathbf{M}$ as region of interest in our homography verification model.

To utilize the prior information for a stable homography estimation, we decompose the homography model into translation, scale, and rotation models and model the difference between these values over consecutive frames.

We empirically learn the various probability models used in our formulations. For this learning, we manually stitch 250 pairs of consecutive frames to find the best homography estimate. By plugging the probability models to Eqn. 3 and ignoring the constants, the final cost function is,

$$f(\theta_t) = \frac{\sum_{i=1}^{N_{in}}\frac{d_i}{b}+\sum_{i=1}^{N_{out}}\ln(S^2)}{N_{in}+N_{out}} + \frac{(E(\mathbf{I}_t,\mathbf{I}_{t-1},\mathbf{M})-\mu_E)^2}{2\sigma_E^2} + \lfloor\frac{(\Delta s-\mu_{\Delta s})^2}{2\sigma_{\Delta s}^2}\rfloor_T + \lfloor\frac{(\Delta\alpha-\mu_{\Delta\alpha})^2}{2\sigma_{\Delta\alpha}^2}\rfloor_T + \lfloor\frac{|\Delta t_x-\mu_{\Delta t_x}|}{b_{t_x}}\rfloor_T + \lfloor\frac{|\Delta t_y-\mu_{\Delta t_y}|}{b_{t_y}}\rfloor_T, \quad (5)$$

where N_{in} and N_{out} represent number of inliers and outliers, respectively, and $\lfloor x\rfloor_T = \min(x,T)$ restricts the impact of prior information. The homography θ_t is estimated by

$$\theta_t^* = \arg\min(f(\theta_t)). \quad (6)$$

Extensive experiments and comparison with manually matched ground truth and baseline methods demonstrate the superiority of RGMC.

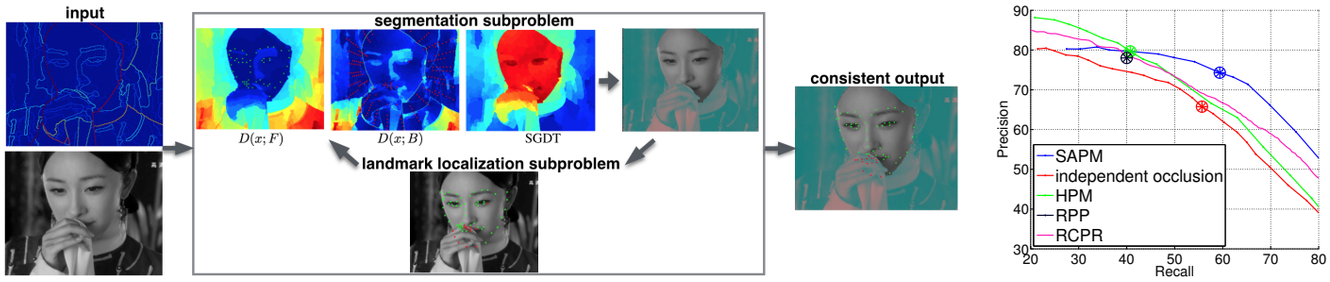
[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *ACM Trans. Communications*, 24(6):381–395, 1981.

[2] Qing Yan, Yi Xu, Xiaokang Yang, and Truong Nguyen. Heask: Robust homography estimation based on appearance similarity and keypoint correspondences. *Pattern Recognition*, 47(1):368–387, 2014.

Using Segmentation to Predict the Absence of Occluded Parts

Golnaz Ghiasi
gghiasi@ics.uci.edu
Charless C. Fowlkes
fowlkes@ics.uci.edu

Dept. of Computer Science
University of California
Irvine, CA



(a) (b)

Figure 1: (a) Overall landmark localization and face mask prediction pipeline. Given an input image and superpixel boundary weights, our method alternately optimizes a landmark localization subproblem and segmentation subproblem. (b) Landmark occlusion prediction accuracy for our segmentation-aware part model (SAPM) and baseline methods on the COFW test data. Enforcing segmentation consistency (SAPM) yields improved performance over an equivalent model without segmentation (independent).

Occlusion, arising from interactions between objects in cluttered scenes or between different parts of a single articulated object, creates drastic changes in local visual appearance. This poses a significant barrier to correctly detecting and performing fine-grained analysis of objects in images. We focus on the problem of detecting and localizing faces where the parts in consideration are represented by facial keypoints. Recent work has shown that pose-regression techniques [1] or deformable part models [5] that include explicit states representing part occlusion can yield better detection and localization accuracy.

A key difficulty with part-based occlusion reasoning is that such methods rely heavily on local image appearance in the vicinity of a part to predict whether it is occluded. This is easily confounded by lighting or other appearance variations and ignores long-range dependencies in patterns of occlusion (e.g., the occluding object comprises an extended region of the image demarcated by enclosing contours). We argue that bottom-up segmentation provides a valuable mechanism by which subsets of occluded or visible keypoints can be grouped in a way that is not easily captured by standard pose-regression or deformable part models. We formulate a joint objective that simultaneously attempts to localize parts and determine their occlusion state in a manner that is consistent with image segments suggested by edges in the image.

The contribution of this paper is in combining explicit part occlusion in a detection model with object-specific segmentation using a simple alternating minimization. Unlike previous work that focused primarily on segmenting objects from background, our model solves the problem of identifying occluders with high accuracy. We do not perform inference over a large pool of segmentation proposals (unlike [3, 4]), instead generating a consistent segmentation with only two iterations of segmentation and detection, even when the occluder has similar texture and color to the object.

Figure 1(a) gives an overview of our model which carries out an initial detection followed by alternating segmentation and refinement of detector keypoint locations. These two models are coupled by a unary potential function that enforces agreement between the location and occlusion state of keypoints in the detector and face/non-face assignment of superpixels in the segmentation model.

We use a deformable part model framework to capture the relative layout of facial keypoints. We use the same tree topology used in the hierarchical part model (HPM) described in [5]. Unlike the HPM work we do not restrict the possible occlusion patterns for the parts. Each keypoint can be visible or occluded independent of the other keypoints, allowing us to represent a much larger space of possible occlusion patterns. Instead we make use of bottom-up segmentation to guide the detector towards consistent patterns of occlusion.

To generate a high-quality segmentation, we train a boundary detector to specifically detect those edges relevant for face segmentation. We train a structured random forest [2] on images from the COFW training dataset in which we manually labeled face foreground masks. The ground-truth boundaries for each training patch were computed based on the segmentation masks. Hence, edges arising from parts of the face such as eyes and lips are labeled as non-boundaries and the boundary detector learns to not return strong boundaries for them. The detected boundaries and the current estimation of the keypoint locations and occlusion state are used to estimate a segmentation for the image. With just two iterations between these two models, we achieve state of the art results for landmark localization, occlusion prediction and face segmentation on the challenging COFW dataset.

Precision/recall of occlusion prediction on the COFW dataset for our model and some baselines are shown in Fig. 1(b). Our model yields substantial gains in the precision of keypoint occlusion prediction. Figure 2 shows example outputs of our model run on several COFW images. Our model produces both a foreground mask and estimates of the keypoint locations and occlusion states.



Figure 2: Examples of landmark localization, occlusion, and segmentation mask estimation for images from the COFW test data produced by our model.

[1] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[2] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, pages 1841–1848, 2013.

[3] Jian Dong, Qiang Chen, Shuicheng Yan, and Alan Yuille. Towards unified object detection and semantic segmentation. In *ECCV*, pages 299–314. 2014.

[4] Sanja Fidler, Roozbeh Mottaghi, Alan Yuille, and Raquel Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, pages 3294–3301, 2013.

[5] Golnaz Ghiasi and Charless C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, pages 1899–1096, 2014.

Learning the Structure of Deep Architectures Using ℓ_1 Regularization

Praveen Kulkarni[1]
Praveen.Kulkarni@technicolor.com
Joaquin Zepeda[1]
Joaquin.Zepeda@technicolor.com
Frederic Jurie[2]
frederic.jurie@unicaen.fr
Patrick Pérez[1]
Patrick.Perez@technicolor.com
Louis Chevallier[1]
Louis.Chevallier@technicolor.com

[1] 975 avenue des Champs Blancs,
CS 17616, 35576 Cesson Sévigné, France
http://www.technicolor.com

[2] University of Caen Basse-Normandie,
CNRS UMR 6072, ENSICAEN, France http://www.unicaen.fr

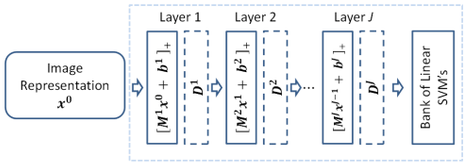


Figure 1: **Proposed deep processing pipeline**. Given an image representation, e.g., the output of the convolutional part of a pre-trained state-of-art DNN, J fully connected layers, each involving a diagonal matrix that controls its effective dimensions, are jointly learned with final linear SVM classifiers.

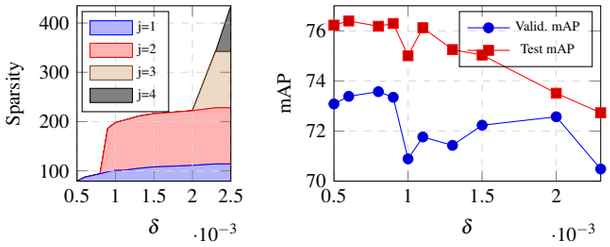


Figure 2: Effect of the penalty weight δ on *(left)* the number of zero diagonal entries of $\mathbf{D}^j$, $j = 1,\ldots,4$, and *(right)* on the classification performance as measured by mAP. The zero diagonal entries are presented as stacked plots so that the vertical displacement of any shaded regions corresponds to the number of zero diagonal entries of $\mathbf{D}^j$ for the corresponding layer.

Our proposed approach is illustrated in Fig. 1. The architecture we consider consists of a sequence of fully-connected layers, with a diagonal matrix between them. We present a method that automatically selects the size of the weight matrices inside fully-connected layers indexed by $j = 1,\ldots,J$. Our approach relies on a regularization penalty term consisting of the ℓ_1 norm of the diagonal entries of diagonal matrices inserted between the fully-connected layers. Using such a penalty term forces the diagonal matrices to be sparse, accordingly selecting the effective number of rows and columns in the weights matrices of adjacent layers. We present a simple algorithm to solve the proposed formulation and demonstrate it experimentally on a standard image classification benchmark.

We can express the architecture in Fig. 1 as a concatenation of units of the following form:

$$f^j(\mathbf{x}) = \mathbf{D}^j \left[\mathbf{M}^j\mathbf{x} + \mathbf{b}^j\right]_+, \tag{1}$$

where $[\mathbf{z}]_+ = [\max(0, z_i)]_i$ is the commonly used Rectified Linear Unit (ReLU) non-linearity. Each layer is defined by diagonal matrix $\mathbf{D}^j$ and matrix $\mathbf{M}^j$. A deep architecture can be derived from (1) using the standard stacking approach. Letting $\circ$ denote the composition operator such that $f \circ g(\mathbf{x}) = f(g(\mathbf{x}))$, this can be denoted as

$$f^J \circ \ldots \circ f^1(\mathbf{x}), \tag{2}$$

where, in this case, the vector $\mathbf{x}$ denotes the representation of the image at the input of the architecture. The image representation can consist of a direct re-ordering of the RGB values in the image [1, 2], or it can be a feature derived from the image [3], which is the approach we follow in the present work.

Besides the variables $(\mathbf{M}^j, \mathbf{D}^j, \mathbf{b}^j)_{j=1}^J$ in (1), one needs to learn the vectors $\{\mathbf{w}^k\}_{k=1}^K$ that define the SVM classifiers for the K classes. We will learn these variables from an annotated training set comprised of N training images $\mathbf{x}_i, i = 1,\ldots,N$, each with K labels $y_i^k \in \{-1, 1\}, k = 1,\ldots,K$ indicating whether image i belongs to class k or not. Given such a training set, our approach consist of minimizing the following objective over all the variables $\{(\mathbf{M}^j, \mathbf{b}^j, \mathbf{D}^j)\}_{j=1}^J$ and all the classifiers $\{\mathbf{w}^k\}_{k=1}^K$:

$$\frac{1}{K}\sum_{k=1}^{K}\left(\|\mathbf{w}^k\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} l\left(y_i^k \left(f_J \circ \ldots \circ f_1(\mathbf{x}_i)\right)^\top \mathbf{w}^k\right)\right) + \delta\sum_{j=1}^{J}\|\mathbf{D}^j\|_* + \mu\sum_{j=1}^{J}\|\mathbf{M}^j\|_F^2. \tag{3}$$

In the above expression, we have used *(i)* $l(x)$ to denote the hinge loss, given by $\max(0, 1-x)$; *(ii)* $\|\mathbf{D}\|_*$ to denote the trace norm, given by $\sum_i |D_{ii}|$ for the case of diagonal $\mathbf{D}$; and *(iii)* $\|\mathbf{M}\|_F^2$ to denote the squared Frobenius norm $\sum_{ij} M_{ij}^2$.

To illustrate the motivation behind this learning objective, we note first that the terms inside the summation over k in (3) are recognizable as an SVM objective for class k, where the scalar C is the SVM regularization parameter. The feature vectors used within this SVM objective are given by $f_J \circ \ldots \circ f_1(\mathbf{x}_i)$, which depends on $\{(\mathbf{M}^j, \mathbf{b}^j, \mathbf{D}^j)\}_{j=1}^J$. Hence we are learning the classifiers jointly with the feature extractor used to represent the input images.

The two regularization terms comprised of summations over j in (3) have two important purpose. First is to keep the SVM terms from decreasing indefinitely. A second important purpose is to automatically select the shapes of the weights matrices $\mathbf{M}^j$ and $\mathbf{b}^j$ and ℓ_1 norms applied to diagonal matrices such as $\mathbf{D}^j$ are sparsity inducing norms.

In order to minimize (3), we will employ a block-coordinate SGD approach. We evaluate our method on Pascal VOC 2007 dataset and compare it against various state-of-the-art algorithms. Our learning algorithm is governed by three important terms: the penalty weights μ and δ and and the number of training epochs T. The number of training epochs is determined using the validation set. Further, we evaluate the performance of our method as a function of the number of layers J in the architecture.

In Fig. 2, we plot both the sparsity for all layers and the corresponding test and validation mAPs when varying the penalty weight δ. Note that increasing δ drastically increases the number of zero diagonal entries in the architecture while only slightly affecting the classification performance.

[1] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference*, 2014. URL `http://arxiv.org/abs/1405.3531`.

[2] Alex Krizhevsky, I. Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems*, pages 1–9, 2012.

[3] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. *Computer Vision and Pattern Recognition*, 2014.

Subspace Distribution Alignment for Unsupervised Domain Adaptation

Baochen Sun
bsun@cs.uml.edu
Kate Saenko
saenko@cs.uml.edu

Computer Science Department
University of Massachusetts Lowell
Lowell, Massachusetts, US

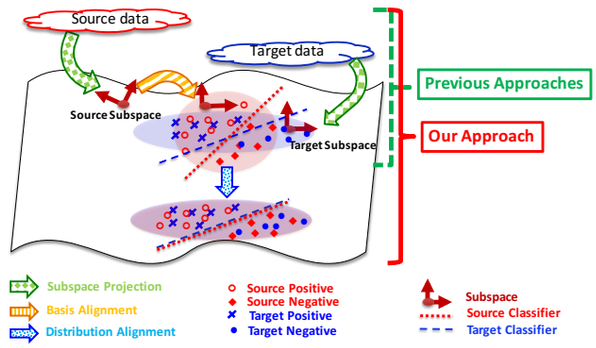


Figure 1: Existing approaches for domain adaptation via subspace mapping fail to properly align the source and target distributions. Our proposed approach improves domain alignment by taking the difference of the source and target distributions into account.

Abstract. Traditional machine learning algorithms often fail to generalize to new input distributions, causing reduced accuracy. Domain adaptation attempts to compensate for the performance degradation by transferring and adapting source knowledge to target domain. Existing unsupervised methods project domains into a lower-dimensional space and attempt to align the subspace bases, effectively learning a mapping from source to target points or vice versa. However, they fail to take into account the difference of the two distributions in the subspaces, resulting in misalignment even after adaptation. We present a unified view of existing subspace mapping based methods and develop a generalized approach that also aligns the distributions as well as the subspace bases.

Background. Domain adaptation, or covariate shift, is a fundamental problem in machine learning, and has attracted a lot of attention in the machine learning and computer vision community. Domain adaptation methods for visual data attempt to learn classifiers on a labeled source domain and transfer it to a target domain. There are two settings for visual domain adaptation: (1) unsupervised domain adaptation where there are no labeled examples available in the target domain; and (2) semi-supervised domain adaptation where there are a few labeled examples in the target domain. Most existing algorithms operate in the semi-superised setting. However, in real world applications, unlabeled target data is often much more abundant and labeled examples are very limited, so the question of how to utilize the unlabeled target data is more important for practical visual domain adaptation. Thus, in this paper, we focus on the unsupervised scenario.

Most of the existing unsupervised approaches have pursued adaptation by separately projecting the source and target data into a lower-dimensional manifold, and finding a transformation that brings the subspaces closer together. This process is illustrated in Figure 1. Geodesic methods [2, 3] find a path along the subspace manifold, and either project source and target onto points along that path [3], or find a closed-form linear map that projects source points to target [2]. Alternatively, the subspaces can be aligned by computing the linear map that minimizes the Frobenius norm of the difference between them, a method known as Subspace Alignment [1].

Approach. The intuition behind our approach is that although the existing approaches might align the subspaces (the bases of the subspaces), it might not fully align the data distributions in the subspaces as illustrated in Figure 1. We use the first- and second-order statistics, namely the mean and the variance, to describe a distribution in this paper. Since the mean after data preprocessing (i.e. normalization) is zero and is not affected by subspace projection, there is no need to align it. However, neither normalization nor subspace basis alignment aligns the covariances.

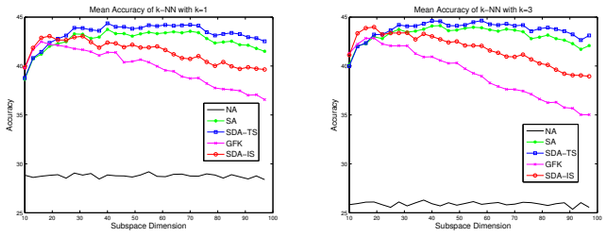


Figure 2: Mean accuracy across all **12** experiment settings (domain shifts) of the k-NN Classifier on the **Office-Caltech10** dataset. Both our methods SDA-IS and SDA-TS outperform GFK and SA consistently. **Left:** k-NN Classifier with **k=1**; **Right:** k-NN Classifier with **k=3**.

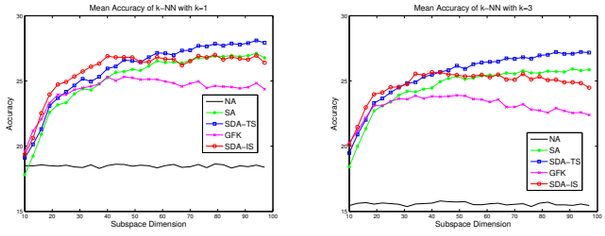


Figure 3: Mean accuracy across all **6** experiment settings (domain shifts) of the k-NN Classifier on the **Office31** dataset. Both our methods SDA-IS and SDA-TS outperform GFK and SA consistently. **Left:** k-NN Classifier with **k=1**; **Right:** k-NN Classifier with **k=3**.

We propose Subspace Distribution Alignment (SDA), which properly aligns the source and target data in the subspaces, namely, it aligns the distributions as well as the subspace bases. SDA constructs the mapping between source and target points to be

$$M_s = S_s T A S_t^T \tag{1}$$

where S_s and S_t are the d dimensional source and target subspaces respectively, T is the subspace basis alignment matrix, and A is the subspace distribution alignment matrix. Suppose W_s and W_t are the square roots of the covariance matrices of the source and target data in the subspaces respectively. Then A could simply be $W_s^{-1} W_t$ since W_s^{-1} transforms the source data into zero mean and unit covariance while W_t adds the covariance of target data. Thus, A transforms the source subspace distribution to the target one. We provide two variants of SDA: SDA-TS aligns the source and target data between two subspaces (source subspace and target subspace); SDA-IS aligns them between an infinite number of subspaces (all the subspaces on the geodesic flow). We show that Subspace Alignment (SA) [1] and Geodesic Flow Kernel (GFK) [2] are special cases of our generalized method.

Experiments. We follow the protocol of [1, 2, 3, 4] and compare our methods to their special cases as well as the no adaptation baseline (NA). Extensive experimental results on standard benchmarks (Figure 2 and 3) demonstrate the advantage of our approach over published methods.

[1] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[2] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[3] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

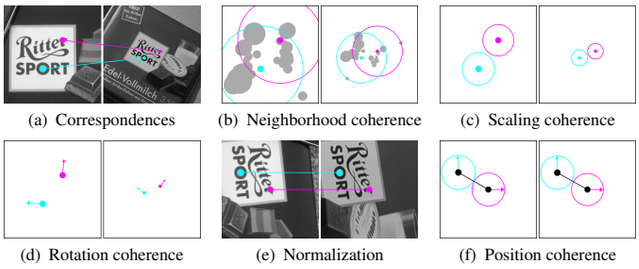
[4] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*. 2010.

Robust Spatial Matching as Ensemble of Weak Geometric Relations

Xiaomeng Wu
wu.xiaomeng@lab.ntt.co.jp
Kunio Kashino
kashino.kunio@lab.ntt.co.jp

NTT Corporation
3-1, Morinosato Wakamiya Atsugi-shi
Kanagawa, Japan 243-0198

Spatial matching methods including RANSAC, Hough transform and spatial context methods have been shown to be successful in addressing the mismatches resulted from the matching of local features in object retrieval. In these methods, true correspondences are identified by imposing a constraint on one or two classes of geometric coherences, e.g. in terms of spatial contexts or between-image transformations. These methods are potentially less discriminative due to the limited number of coherence classes, while forcibly enhancing the strength of constraints leads to the incorrect rejection of true correspondences. Spatial matching still faces a difficult trade-off between flexibility and discriminative power.



(a) Correspondences (b) Neighborhood coherence (c) Scaling coherence

(d) Rotation coherence (e) Normalization (f) Position coherence

Figure 1: A correspondence pair and four geometric coherences.

To address this issue, we regard spatial matching as an ensemble of geometric relations on the set of feature correspondences. A *geometric relation* is a set of pairs of correspondences, in which every correspondence is associated with every other correspondence if and only if the pair satisfy a geometric constraint. We design a novel, unified collection of weak geometric relations that fall into four fundamental classes of geometric coherences (Figures 1(b)-1(d) and 1(f)) in terms of both spatial contexts and between-image transformations. By a *weak geometric relation*, we mean a sufficiently flexible constraint which, nevertheless, may offer only a limited discriminative power. The spatial similarity reduces to the cardinality of the conjunction of all geometric relations. The flexibility of weak geometric relations makes our method robust as regards rejections of true correspondences, and the conjunctive ensemble provides a high discriminative power in terms of mismatches.

An image is represented by a set P of features. For each feature $p \in P$ we are given its visual word $u(p)$, position $\mathbf{t}(p) = [x(p)\ y(p)]^{\mathrm{T}}$, scale $\sigma(p)$ and orientation $\mathbf{R}(p)$. p can be mapped, from a unit circle heading a reference orientation, by a 3×3 transformation matrix $\mathbf{F}(p)$:

$$\mathbf{F}(p) = \begin{bmatrix} \mathbf{M}(p) & \mathbf{t}(p) \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{1}$$

where $\mathbf{M}(p) = \sigma(p)\mathbf{R}(p)$ is a linear transformation and homogeneous coordinates are to be used for the mapping. If $\sigma(p)$ is given by a real scalar, $\mathbf{F}(p)$ specifies a similarity transformation. $\mathbf{R}(p)$ is an orthogonal 2×2 matrix with $\det \mathbf{R}(p) = 1$, represented by an angle $\theta(p)$. Given two images P and Q, a correspondence $c \triangleq (p,q)$ is a pair of features $p \in P$ and $q \in Q$ with $u(p) = u(q)$. We assume $|C| \geq 2$ with $C = \{c\}$ and:

$$c = \big(u(c), \mathbf{t}(p), \sigma(p), \theta(p), \mathbf{t}(q), \sigma(q), \theta(q)\big). \tag{2}$$

Suppose that P and Q are related as regards a common near-rigid object and an unknown transformation $\mathbf{F}$. It can be inferred that all parts of the object obey the same transformation. Given a correspondence set C constructed from P and Q, there is a subset $C_{\mathbf{F}} \subseteq C$ of correspondences that lie inside the object and show similarity in terms of their local transformations. Spatial matching is to identify such a subset, whose cardinality provides evidence for the belief that P and Q include the same object.

We focus on the Cartesian product $C^2 = C \times C$, i.e. the set of all pairs (c_a, c_b) where $c_a, c_b \in C$. A function $h : C^2 \to \{0,1\}$ is defined, which maps any arbitrary (c_a, c_b) to one if a geometric constraint is satisfied, and zero otherwise. A geometric relation G is thus a subset of C^2 such that $\forall (c_a, c_b) \in G,\ h(c_a, c_b) = 1$. If h is sufficiently well-defined and if the geometries in Eq. 2 are accurately given, we have $G \approx C_{\mathbf{F}}^2$. The spatial similarity can be formulated by the cardinality of G instead of that of $C_{\mathbf{F}}$.

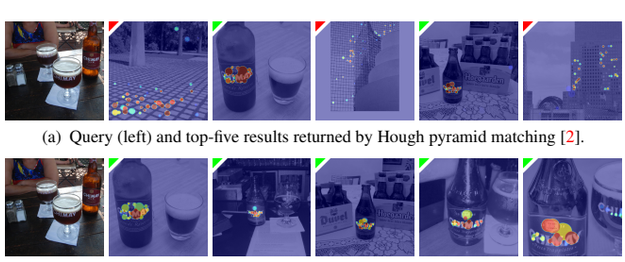
Instead of a single h, we build a set $H = \{h\}$ of weak geometric constraints, resulting in a set $\mathcal{G} = \{G\}$ of geometric relations. Each $h \in H$ should be flexible as regards feature detection errors, but is allowed to offer a limited discriminative power. A conjunctive ensemble of such relations (Eq. 3) creates a single strong constraint that is expected to be discriminating in terms of mismatches. The spatial similarity thus becomes $|\hat{G}|$.

$$\hat{G} = \bigcap_{G \in \mathcal{G}} G = \left\{ (c_a, c_b) \in C^2 \,\middle|\, \left(\prod_{h \in H} h(c_a, c_b) \right) = 1 \right\} \tag{3}$$

More detail on our carefully designed weak geometric constraints is presented in the paper. Table 1 presents the reported MAPs of spatial matching methods on the OB, Paris and OB+F100K datasets. Our method outperforms all methods on all datasets. Our search time per query and per 1K images was 19.0 msec for OB. The shortest corresponding time so far reported by Shen et al. [4] was 17.6 msec. This reveals the high competitiveness of our scalability. Figure 2 compares HPM [2], the best baseline we know in this field, and our method.

Table 1: Comparison with reported, state-of-the-art MAPs.

Methods	OB	Paris	OB+F100K
Our Method	**.827**	**.766**	**.769**
HPM [2]	.789	.725	.730
Perdoch et al. [3]	.789	n/a	.726
Shen et al. [4]	.752	.741	.729
Arandjelovic and Zisserman [1]	.720	n/a	.642
Zhang et al. [5]	.713	n/a	.604



(a) Query (left) and top-five results returned by Hough pyramid matching [2].

(b) Query (left) and top-five results returned by our method.

Figure 2: Comparison of HPM [2] and our method. The green and red colors of the upper-left corners of the images indicate positive and negative results. Identified correspondences are highlighted in colors.

[1] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[2] Yannis S. Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, 2014.

[3] Michal Perdoch, Ondrej Chum, and Jiri Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, pages 9–16, 2009.

[4] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Spatially-constrained similarity measure for large-scale object retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1229–1241, 2014.

[5] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.

Kernelized View Adaptive Subspace Learning for Person Re-identification

Qin Zhou[12]
zhou.qin.190@sjtu.edu.cn
Shibao Zheng[12]
sbzh@sjtu.edu.cn
Hang Su[34]
suhangss@gmail.com
Hua Yang[12]
hyang@sjtu.edu.cn
Yu Wang[12]
txtxs@sjtu.edu.cn
Shuang Wu[12]
shuangwu@sjtu.edu.cn

[1] Department of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2] Shanghai Key Lab of Digital Media Processing and Transmission
[3] Department of Computer Science and Technology, Tsinghua Univelrsity
[4] Robotics Institute, Carnegie Mellon University

Person re-identification refers to the task of recognizing the same person under different non-overlapping camera views and across different time and places. We propose a novel algorithm coined as Kernelized View Adaptive Subspace Learning (KVASL), which tries to learn different projection matrices for each camera view to compensate for the specific distortion brought by different views. The kernel trick is adopted to catch more information such that nonlinear transformation is possible. We present the motivating example in Figure 1, which not only demonstrates the benefit of adopting different projection matrices, but also illustrates the necessity of the kernel trick.

Denote $|S|$,$|D|$ as the number of matched and mismatched image pairs in the training set. Our kernelized loss function is formulated as follows:

$$\begin{aligned}\ell_K(L_A^K, L_B^K) &= \frac{1}{|S|} tr(K_A{}^T K_A Q_A^T Q_A - 2K_B{}^T K_A Q_A^T Q_B + K_B{}^T K_B Q_B^T Q_B) \\ &+ \mu_A ||L_A^K||_F^2 + \mu_B ||L_B^K||_F^2 - \frac{\lambda}{|D|} tr((|S|-1)K_A{}^T K_A Q_A^T Q_A) \\ &+ tr(2K_A{}^T X K_B Q_B^T Q_A - (|S|-1)K_B{}^T K_B Q_B^T Q_B)\end{aligned} \tag{1}$$

where $K_A = \phi(A_{train})^T \phi(A_{train}) \in R^{N\times N}$, $K_B = \phi(B_{train})^T \phi(B_{train}) \in R^{N\times N}$ are symmetry matrices, $\phi(A_{train}) = \left[\phi(x_{1A}^{train}), \cdots, \phi(x_{NA}^{train})\right]$, $\phi(B_{train}) = \left[\phi(y_{1B}^{train}), \cdots, \phi(y_{NB}^{train})\right] \in R^{m\times N}$ are the matrices formed by feature vectors in the kernel space of the corresponding camera views, $tr(\cdot)$ indicates the trace of the matrix and X corresponds to the matrix whose diagonal elements are all zeros and the other elements are all ones.

We adopt an alternately iterative gradient descent method to optimize our loss function. The gradients of Q_A,Q_B in Eq.(1) can be derived as follows:

$$\begin{aligned}\frac{\partial \ell}{\partial Q_A} &= \frac{2}{|S|}(Q_A K_A K_A - Q_B K_B K_A) \\ &- \frac{2\lambda}{|D|}[(|S|-1)Q_A K_A K_A - Q_B K_B X K_A] + 2\mu_A Q_A K_A \\ \frac{\partial \ell}{\partial Q_B} &= \frac{2}{|S|}(Q_B K_B K_B - Q_A K_A K_B) \\ &- \frac{2\lambda}{|D|}[(|S|-1)Q_B K_B K_B - Q_A K_A X K_B] + 2\mu_B Q_B K_B\end{aligned} \tag{2}$$

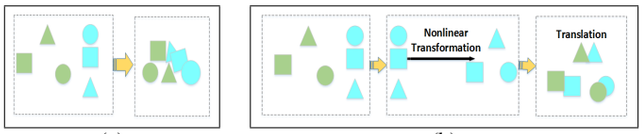
where X is the same matrix as in Eq.(1). In this case, the matrices to be learned can be updated in the following way:

$$Q_A(t+1) = Q_A(t) - \eta_{Q_A}\frac{\partial \ell}{\partial Q_A}, \quad Q_B(t+1) = Q_B(t) - \eta_{Q_B}\frac{\partial \ell}{\partial Q_B} \tag{3}$$

where $\eta_{L_A}, \eta_{L_B}, \eta_{Q_A}, \eta_{Q_B}$ are the learning rate corresponding to each transform matrix. Once Q_A, Q_B are learned, the distance between two test images can be derived as follows:

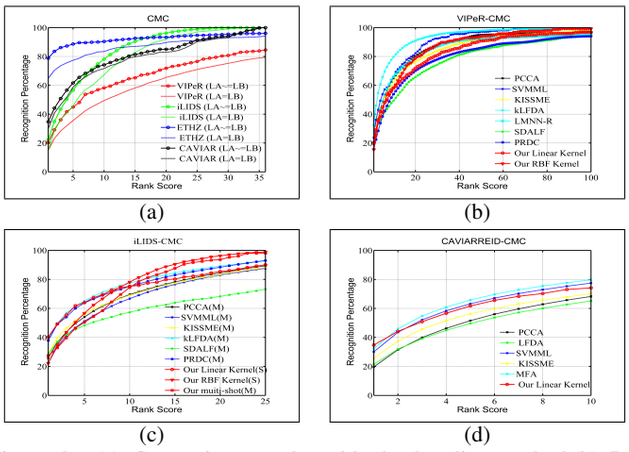
$$\begin{aligned}d_K^2(x_{iA}^{test}, y_{jB}^{test}) &= e_i^T K_A^{test} Q_A^T Q_A (K_A^{test})^T e_i \\ &- 2e_j^T K_B^{test} Q_B^T Q_A (K_A^{test})^T e_i + e_j^T K_B^{test} Q_B^T Q_B (K_B^{test})^T e_j\end{aligned} \tag{4}$$

We implement our algorithm on four publicly available datasets, the comparison results with the baseline method demonstrate the superiority of view-adaptive projection matrices over using the same projection matrices. We also compare our algorithm with some state-of-the-art algorithms (SVMML [2], KISSME [1], kLFDA [3] et al.), which demonstrates the effectiveness of our algorithm. Detailed comparison results are shown in Figure 2 and Table 1.



(a) (b)

Figure 1: A conceptual illustration of how different projection functions can be more flexible than the same ones. (a)The same transformation; (b) Different transformations.



(a) (b) (c) (d)

Figure 2: (a) Comparison results with the baseline method.(b) Results on VIPeR dataset.(c) Results on iLIDS dataset.(d) Results on CAVIAR4REID dataset. Red lines correspond to results of our algorithm.

Method	r=1	r=5	r=10	r=15
KISSME [1]	65.51	83.67	87.34	89.59
Our Linear Kernel	**78.77**	**88.63**	**90.41**	**92.60**
Our RBF Kernel	70.55	85.75	89.45	91.23

Table 1: Comparison of our algorithm with [1] on the ETHZ dataset.

[1] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, pages 2288–2295, 2012.

[2] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith. Learning locally-adaptive decision functions for person verification. In *Proc. CVPR*, pages 3610–3617, 2013.

[3] Fei Xiong, Mengran Gou, Octavia I. Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Proc. ECCV*, pages 1–16, 2014.

Object localization in ImageNet by looking out of the window

Alexander Vezhnevets
vezhnick@gmail.com
Vittorio Ferrari
vferrari@gmail.com

University of Edinburgh
Edinburgh, Scotland, UK

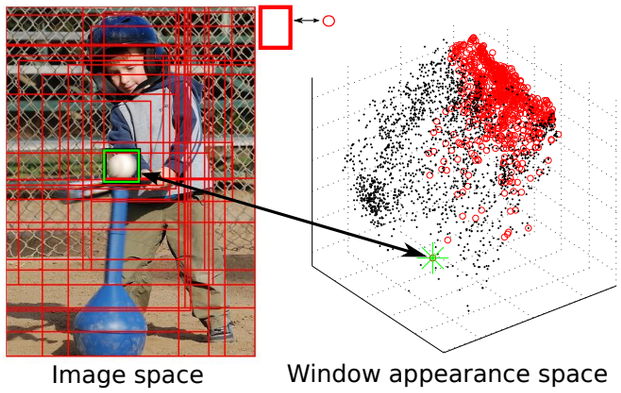


Figure 1: **Connecting the appearance and window position spaces.** *A window tight on the baseball (green star in the appearance space plot) and some larger windows containing it (red circles in the appearance space). Black points in appearance space represent all other candidate windows. The appearance space plots are actual datapoints, representing windows in 3-dimensional Associative Embedding of SURF bag-of-words.*

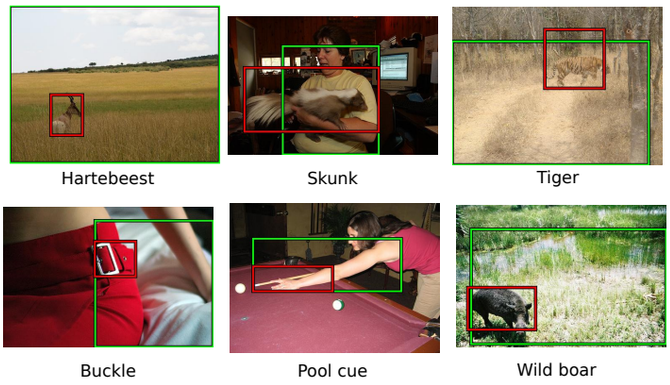


Figure 2: **Results of our method (red) vs a prior method [10] (green).** *Notice, how our method is able to detect small, off-center objects despite occlusion (pool cue) or the object blending with its surroundings (tiger).*

The ImageNet database [4] contains over 14 million images annotated by the class label of the main object they contain. However, only a fraction of them have bounding-box annotations (10%). Automatically annotating object locations in ImageNet is a challenging problem, which has recently drawn attention [6, 10]. These annotations could be used as training data for problems such as object class detection [3], tracking [7] and pose estimation [1]. Traditionally, object localization is cast as an image window scoring problem, where a scoring function is trained on images with bounding-boxes and applied to ones without. The image is first decomposed into candidate windows, typically by object proposal generation [8]. Each window is then scored by a classifier trained to discriminate instances of the class from other windows [3, 5, 8, 9] or a regressor trained to predict their overlap with the object [2, 10]. Highly scored windows are finally deemed to contain the object. In this paradigm, the classifier looks at one window at a time, making a decision based only on that window's appearance.

We believe there is more information in the collection of windows in an image. By taking into account the appearance of all windows *at the same time* and connecting it to their spatial relations in the image plane, we could go beyond what can be done by looking at one window at a time. Consider the baseball in fig. 1. For a traditional method to succeed, the appearance classifier needs to score the window on the baseball higher than the windows containing it. The container windows cannot help except by scoring lower and be discarded. By considering one window at a time with a classifier that only tries to predict whether it covers the object tightly, one cannot do much more than that. The first key element of our work is to predict richer spatial relations between each candidate window and the object to be detected, including part and container relations. The second key element is to employ these predictions to reason about relations between different windows. In this example, the container windows are predicted to contain a smaller target object somewhere inside them, and thereby actively help by *reinforcing* the score of the baseball window. By considering the configuration of all the windows in appearance space together we can reinforce its score.

In a nutshell, we propose to localize objects in ImageNet by scoring each candidate window in the context of all other windows in the image, taking into account their similarity in appearance space as well as their spatial relations in the image plane. To represent spatial relations of windows we propose a descriptor indicative of the part/container relationship of the two windows and of how well aligned they are. We learn a windows appearance similarity kernel using the recent Associative Embedding technique [10]. We describe each window with a set of hyper-features connecting the appearance similarity and spatial relations of that window to all other windows in the same image. These hyper-features are indicative of the object's presence when the appearance of a window alone is not enough (e.g. fig 1). These hyper-features are then linearly combined into an overall scoring function. We devise a fast and exact procedure to optimize our scoring function over all candidate windows in a test image, and we learn its parameters using structured output regression.

We evaluate our method on a subset of ImageNet containing 219 classes with more than 92000 images [6, 10]. The experiments show that our method outperforms a recent approach for this task [10], an MKL-SVM baseline [9] based on the same features, and the popular UVA object detector [8]. Figure 2 presents some qualitive results of our method compared to results of [10].

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[2] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.

[3] N. Dalal and B. Triggs. Histogram of Oriented Gradients for human detection. In *CVPR*, 2005.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[6] M. Guillaumin, D. Küttel, and V Ferrari. ImageNet auto-annotation with segmentation propagation. *IJCV*, 2014.

[7] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.

[8] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 104(2): 154–171, 2013.

[9] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *ICCV*, pages 606–613, 2009.

[10] A. Vezhnevets and V. Ferrari. Associative embeddings for large-scale knowledge transfer with self-assessment. In *CVPR*, 2014.

Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection

Huamin Ren[1]
hr@create.aau.dk
Weifeng Liu[2]
weifeng.liu@nbi.ku.dk
Søren Ingvor Olsen[3]
ingvor@di.ku.dk
Sergio Escalera[4]
sergio@maia.ub.es
Thomas B. Moeslund[1]
tbm@create.aau.dk

[1] Department of Architecture, Design and Media Technology
Aalborg University
Aalborg, Denmark

[2] Niels Bohr Institute
University of Copenhagen
Copenhagen, Denmark

[3] Department of Computer Science
University of Copenhagen
Copenhagen, Denmark

[4] Dept. Applied Mathematics, University of Barcelona, Computer Vision Center
Barcelona, Spain

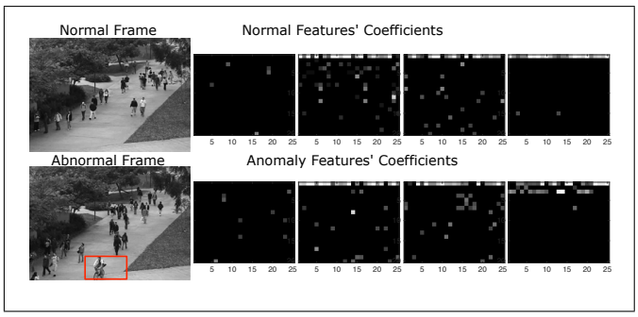


Figure 1: Non zero distribution in coefficients of features from normal and abnormal frames. From left to right: Lu13 [3] and our three frames: D_{basis}, $D_{feature}$ and $D_{missedAtoms}$. Accumulated non zeros in each dictionary are normalized in [0,1]; therefore, the whiter, the more non zeros.

Abnormal event detection is an important issue in video surveillance applications. The goal is to detect abnormal or suspicious behaviors while given training samples that contain only normal behaviors. Sparse representation has showed its effectiveness in abnormal event detection [2, 3, 4, 5], where a dictionary is commonly learned during training and anomalies are detected based on reconstruction error from the learned dictionary. Note that only a small proportion of the data is used for training - relatively to the huge amount of surveillance data, there is a high risk of incomplete normal patterns in the training data. Consequently, dictionary learning is crucial to the overall abnormality detection performance.

We propose a Behavior-Specific Dictionary (BSD) algorithm, which takes into consideration the relation of atoms in one dictionary without prior knowledge. In our algorithm, atoms in the same dictionary are related to one type of behavior. Normal features could be sparsely represented by one dictionary with a similar patten, while anomalies could not. To further reduce false alarms, features with compact non zero coefficients in each dictionary (called missed atoms) are added as new atoms in order to better distinguish between anomalies and infrequent behaviors. The non zero distribution of coefficients for normal and abnormal frames are demonstrated in Fig. 1, compared with Lu13 [3]. It can be seen that our frames could have a more concentrated non zero distribution for normal features, meanwhile obtaining spread out non zeros for abnormal features. Our proposed algorithm is shown in Fig.2.

Experiments are carried out on two datasets: UCSD Anomaly dataset and Anomaly Stairs dataset. The former dataset is used to get a comparable results with other state-of-art methods; while the latter one is used to test the effectiveness of our method when the training data is incomplete. On UCSD Anomaly dataset, our BSD obtains a AUC of 56.17% on pixel-level evaluation, which achieves the best result (promoting AUC by 10% compared to the best score (46.1%) published so far [2]), yet has a satisfactory result on frame-level evaluation (AUC is 70.69%). On Anomaly Stairs dataset, seen from Tab. 1, our BSD achieves the best detection with the least false alarms.

	# Correctly detected abnormal videos (TP)			# False alarms (FA)		
Pattern and training size	Lu13 [3]	Bao14 [1]	BSD (ours)	Lu13 [3]	Bao14 [1]	BSD (ours)
One group (1%)	11	12	**15**	**13**	16	14
One group (10%)	**17**	12	**17**	21	16	**15**
Two groups (1%)	14	13	**17**	**15**	19	17
Two groups (10%)	14	**19**	15	17	18	**12**
Three groups (1%)	14	**18**	**18**	**15**	19	16
Three groups (10%)	14	**19**	**19**	**17**	18	**17**
Average TP / FA	14.0	15.5	**16.8**	16.3	17.7	**15.2**
Average TPR / FPR	70.0%	77.5%	**84.2%**	77.8%	84.1%	**72.2%**

Table 1: Results on Anomaly Stairs dataset by varying the size of the training data.

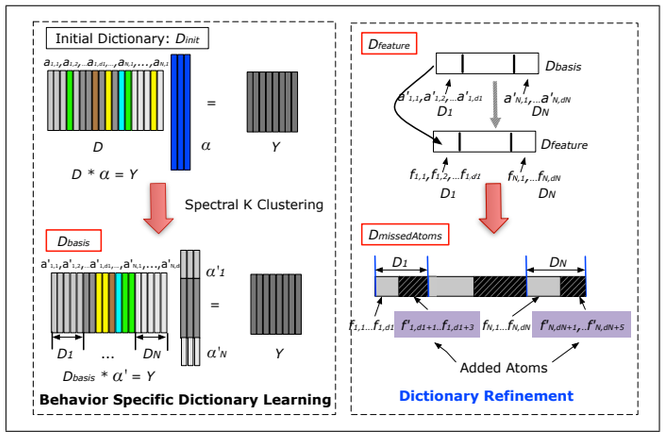


Figure 2: Behavior-specific dictionary learning (left) and refinement procedure (right). During BSD learning, multiple dictionaries are obtained, each of which corresponds to one particular normal behavior. Refinement is to find missed atoms to improve current dictionaries.

Our proposed algorithm learns behavior-specific dictionaries through unsupervised learning. The experiments proved the effectiveness of the proposed method, especially when the data available during the training is limited.

[1] Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *CVPR*, pages 3858–3865, June 2014.

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pages 3449–3456, 2011.

[3] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013.

[4] H. Ren and T. B. Moeslund. Abnormal event detection using local sparse representation. In *AVSS*, pages 125–130, 2014.

[5] B. Zhao, F. Li, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, pages 3313–3320, 2011.

Joint Calibration for Semantic Segmentation

Holger Caesar
holger.caesar@ed.ac.uk
Jasper Uijlings
jrr.uijlings@ed.ac.uk
Vittorio Ferrari
vittorio.ferrari@ed.ac.uk

School of Informatics
University of Edinburgh
Edinburgh, UK

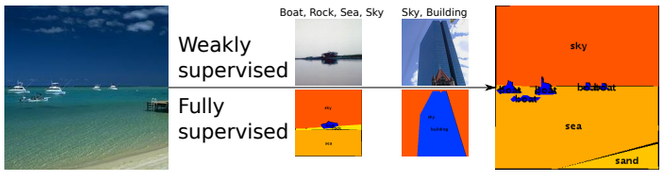


Figure 1: Semantic segmentation is the task of assigning class labels to all pixels in the image. Depending on the setting, we have either pixel-level or image-level ground-truth labels available at training time.



Figure 2: Fully supervised semantic segmentation on SIFT Flow. We show an example image, the target output, our uncalibrated output and our jointly calibrated output.

Semantic segmentation is the task of assigning a class label to each pixel in an image (Fig. 1). In the fully supervised setting, we have ground-truth labels for all pixels in the training images. In the weakly supervised setting, class-labels are only given at the image-level. We tackle both settings in a single framework which builds on region-based classification.

Our framework addresses three problems common to region-based semantic segmentation. First of all, objects naturally occur at different scales within an image [3]. Performing recognition at a single scale inevitably leads to regions covering only parts of an object which may have ambiguous appearance, and to regions straddling over multiple objects, whose classification is harder due to their mixed appearance. Therefore many recent methods operate on pools of regions computed at multiple scales, which have a much better chance of containing some regions covering complete objects [1, 2]. However, this leads to overlapping regions which may lead to conflicting class predictions at the pixel-level. These conflicts need to be properly resolved.

Secondly, classes are often unbalanced [2, 4]: "cars" and "grass" are frequently found in images while "tricycles" and "gravel" are much rarer. Due to the nature of most classifiers, without careful consideration these rare classes are largely ignored: even if the class occurs in an image the system will rarely predict it. Since class-frequencies typically follow a power-law distribution, this problem becomes increasingly important with the modern trend towards larger datasets with more and more classes.

Finally, classes compete: a pixel can only be assigned to a single class (e.g. it can not belong to both "sky" and "airplane"). To properly resolve such competition, a semantic segmentation framework should take into account predictions for multiple classes jointly.

In this paper we address these three problems with a joint calibration method over a set of SVMs.

Model. We represent an image by a set of overlapping regions [3] described by CNN features [1]. Our semantic segmentation model infers the label o_p of each pixel p in an image:

$$o_p = \arg\max_{c,\, r \ni p} \sigma(w_c \cdot x_r,\, a_c, b_c) \qquad (1)$$

As appearance models, we have a separate linear SVM w_c per class c. These SVMs score the features x_r of each region r. The scores are calibrated by a sigmoid function σ, with different parameters a_c, b_c for each class c. The arg max returns the class c with the highest score over all regions that contain pixel p. This involves maximizing over classes for a region, and over the regions that contain p.

Joint Calibration. The final goal of semantic segmentation is to output a pixel-level labeling, which is evaluated in terms of pixel-level accuracy. Now, using uncalibrated SVMs is problematic. SVMs are trained to predict class labels at the region-level, not the pixel-level. However, different regions have different area, and, most importantly, not all regions contribute all of their area to the final pixel-level labeling: Predictions of small regions may be completely suppressed by bigger regions. In other cases, bigger regions may be *partially* overwritten by smaller regions. Furthermore, the SVMs are trained in a one-vs-all manner and are unaware of the competition between classes. To address these problems we propose to jointly calibrate SVMs for the final evaluation measure. We do this by applying sigmoid functions σ to all SVM outputs. We calibrate the parameters a_c, b_c for all classes jointly by minimizing a loss function that depends on the pixel labeling output of our method and the ground-truth labeling. Since we optimize for the accuracy of this final output labeling, and we do so jointly over classes, our calibration procedure takes into account both problems of conflicting class predictions between overlapping regions and competition between classes. Moreover, we also address the problem of class imbalance, as we compensate for it in our loss functions. We minimize our loss functions using coordinate descent. We iteratively apply line search to optimize the loss over a single parameter at a time, keeping all others fixed.

Experiments. We evaluate our method using class-average pixel accuracy (Cl. Acc.) on the challenging SIFT Flow dataset. Table 1 shows that we outperform the state-of-the-art for the default setting in semantic segmentation by 2.8% in the fully supervised and 2.4% in the weakly supervised setting. Fig. 2 shows an example of our method that is able to detect even small and rare objects in an image.

Method	Cl. Acc.	Method	Cl. Acc.
Tighe ECCV 2010	29.1%	Vezhnev. ICCV 2011	14.0%
Pinheiro ICML 2014	30.0%	Vezhnev. CVPR 2012	21.0%
Shuai CVPR 2015	39.7%	Zhang TM 2014	27.7%
Tighe CVPR 2013	41.1%	Xu CVPR 2014	27.9%
Kekeç BMVC 2014	45.8%	Zhang CVPR 2015	32.3%
Sharma NIPS 2014	48.0%	Xu CVPR 2015	35.0%
Yang CVPR 2014	48.7%	Xu CVPR 2015	41.4%
George CVPR 2015	50.1%	(transductive)	
Farabet PAMI 2013	50.8%		
Long CVPR 2015	51.7%		
Sharma CVPR 2015	52.8%		
Ours SVM	28.7%	Ours SVM	21.2%
Ours SVM+PS	27.7%	Ours SVM+PS	16.8%
Ours SVM+JC	**55.6%**	Ours SVM+JC	**37.4%**

Table 1: Class-average pixel accuracy in the fully supervised (left) and the weakly supervised setting (right) setting. We show results for our model on the test set of SIFT Flow using uncalibrated SVM scores (SVM), traditional Platt scaling (PS) and joint calibration (JC).

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] A. Sharma, O. Tuzel, and D. W. Jacobs. Deep hierarchical parsing for semantic segmentation. In *CVPR*, 2015.

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[4] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015.

Camera Elevation Estimation from a Single Mountain Landscape Photograph

Martin Čadík[1]
cadik@fit.vutbr.cz
Jan Vašíček[1]
xvasic21@stud.fit.vutbr.cz
Michal Hradiš[1]
ihradis@fit.vutbr.cz
Filip Radenović[2]
filip.radenovic@cmp.felk.cvut.cz
Ondřej Chum[2]
chum@cmp.felk.cvut.cz

[1] CPhoto@FIT, http://cphoto.fit.vutbr.cz/elevation/,
Faculty of Information Technology,
Brno University of Technology,
Brno, Czech Republic

[2] Center for Machine Perception,
Department of Cybernetics,
Faculty of Electrical Engineering,
Czech Technical University in Prague,
Prague, Czech Republic

In outdoor environments one of the most important and informative attributes is the elevation: the height of a geographic location above the sea level. However, almost all currently available photos and videos lack elevation information. Moreover, a majority of them do not even contain the GPS coordinates. This work addresses the problem of camera elevation estimation from visual information contained in a landscape photograph.

We introduce a new **Alps100K dataset** of annotated (GPS coordinates, elevation, EXIF if available) outdoor images from mountain environments. We create a list of all hills and mountain peaks located in the seven Alpine countries from the OpenStreetMap database. The list of hill names is used to query the Flickr photo hosting service. The final collection contains *98136 outdoor images* that span almost all possible elevations observed in the Alps [0, 4782m] and covers vast geographic area of the Alps. The images span all the seasons of the year and exhibit high variation in landscape appearance, see Fig. 1.



Figure 1: A sample from the new benchmark dataset Alps100K. Image credits - flickr users: Allie_Caulfield, Erik, Guillaume Baviere, antoine.pardigon, Karim von Orelli.

To measure the ability of humans to estimate camera elevation from an image, an **experiment with 100 subjects** is conducted. The participants are asked to estimate the *camera* elevation for each of the 50 test images using a web-based interface. The overall root-mean-squared error of human elevation predictions is RMSE$(H) = 879.95$m. The predictions for each test image along with the ground truths are plotted in Fig. 2.

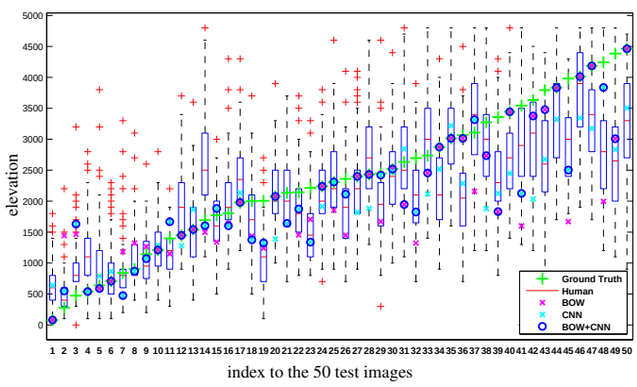


Figure 2: Comparison of the elevation estimation by humans and the proposed methods. Blue boxes show the span of the human predictions: the red mark is the median, the edges of the box are the 25th and 75th percentiles respectively, the whiskers extend to extreme human guesses that are not considered outliers, and outliers are plotted individually as '+'.

First proposed automatic approach to elevation estimation from image content is based on **convolutional neural networks (CNN)**. We initialize CNNs from a network previously trained on the Places205 dataset [2]. The basic network architecture is the same as the one used in the Caffe reference network and the Places-CNN network [2]. The main building blocks of the network are convolutions followed by Rectified Linear Units. First, second, and fifth convolutional layers are followed by max-pooling, each reducing resolution by a factor of two. The activations of the first and second convolutional layers are locally normalized. The output of the convolutional part of the network is fed into a fully connected layer with 4096 units. The network is trained by mini-batch Stochastic Gradient Descent with momentum.

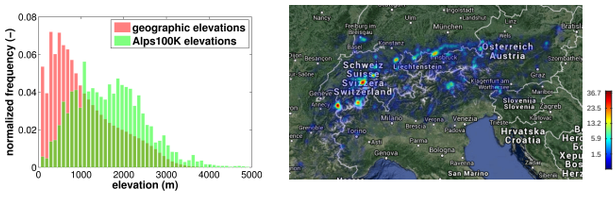


Figure 3: Left: normalized elevation histogram of the Alps mountain range (red) and the distribution of elevations in the Alps100K dataset (green). Right: geographic coverage of the new dataset.

An alternative bag-of-words approach is based on the k-NN classifier, two efficient methods of obtaining k-NN are considered: **sparse high-dimensional bag-of-words (BOW)** based image retrieval, and image retrieval with **compact image representations (mVocab)** [1]. The **BOW** approach using inverted file to efficiently retrieve images has been shown to perform well for specific object and place recognition, especially when combined with a spatial verification step, while the **mVocab** approach using a joint dimensionality reduction from multiple vocabularies shows certain level of generalization power. Therefore, we also propose a **hybrid method** that first tries to estimate the elevation by recognizing the location (using BOW), and if that fails, *i.e.*, no spatially verified image is retrieved, then by a secondary estimator: either mVocab or CNN.

A *test set* of 13148 images is randomly selected from Alps100K, the rest of the dataset (*i.e.*, 84988 images) is used for *training*. The selected measure of performance is an overall RMSE of elevation predictions with regards to the known ground truth elevations, see Tab. 1. The results for a subset of 50 images selected from the test dataset (used in user experiment), which compares the performance of automatic elevation estimation to the performance of humans, are shown in Fig. 2 and Tab. 1. Generally, all of the proposed methods achieve better scores than humans.

Table 1: Results (overall root-mean-square error in meters)

Method	test dataset (13148 images)	user experiment (50 images)
Baseline	801.49; *786.42*	1383.64; *1154.43*
Human	-	**879.95**
CNN	537.11	709.10
BOW	601.63	757.76
mVocab	610.36	811.00
BOW+mVocab	564.14	646.89
BOW+CNN	**500.44**	**531.05**

[1] F. Radenović, H. Jégou, and O. Chum. Multiple measurements and joint dimensionality reduction for large scale image search with short vectors. In *Proc. ICMR*. ACM, 2015.

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.

Data-free Parameter Pruning for Deep Neural Networks

Suraj Srinivas
surajsrinivas@ssl.serc.iisc.in
R. Venkatesh Babu
venky@serc.iisc.in

Video Analytics Lab,
Supercomputer Education and Research Centre,
Indian Institute of Science

Deep Neural nets (NNs) with millions of parameters are at the heart of many state-of-the-art computer vision systems today. However, recent works have shown that much smaller models can achieve similar levels of performance. In this work, we address the problem of pruning parameters in a trained NN model. Instead of removing individual weights one at a time as done in previous works, we remove one neuron at a time. We show how similar neurons are redundant, and propose a systematic way to remove them. Unlike previous works, our pruning method does not require access to any training/validation data.

Wiring similar neurons The main principle that we use in this paper is the fact that similar neurons are redundant, as shown in Figure 1. That is, if we find such a similar weight pair anywhere in a neural network, one of them can effectively be removed. Of course, while doing this we also need to account for the weights in the next layer, as shown in Figure 1. This observation also resonates with the well-known Hebbian principle, which roughly states that neurons that fire together ($W_1 = W_2$), wire together ($a_1 = a_1 + a_2$).

Wiring dis-similar neurons The above principle cannot be used as is in real NNs, for one simple reason - weight-sets are seldom equal in value. What do we do when $\|W_1 - W_2\| = \|\varepsilon_{1,2}\| \geq 0$? Let z_n be the output neuron when there are n hidden neurons. Let us consider two similar weight sets W_i and W_j in z_n and that we have chosen to remove W_j to give us z_{n-1}. Using some approximate analysis, we derive a simple rule to find which weight-sets to remove. The final equation is

$$min(E(z_n - z_{n-1})^2) \leq min_{i,j}(a_j^2 \, \|\varepsilon_{i,j}\|_2^2) \; E\|X\|_2^2 \qquad (1)$$

We aim to minimize the expected value of the squared difference between the output neurons. Using the expected error instead of the empirical error is what makes it a **data-free** parameter pruning method. We define the saliency of two weight-sets in (i, j) as $s_{i,j} = \langle a_j^2 \rangle \, \|\varepsilon_{i,j}\|_2^2$, which is exactly the term inside the $min(\cdot)$ in Equation 1. Intuitively, saliency between two weight-sets is low when they have very similar values. Equation 1 tells us that we need to start removing lowest-saliency neuron to minimize the expected squared difference.

We elucidate our procedure for neuron removal here:

1. Compute the saliency $s_{i,j}$ for all possible values of (i, j). It can be stored as a square matrix M, with dimension equal to the number of neurons in the layer being considered.

2. Pick the minimum entry in the matrix. Let it's indicies be (i', j'). Delete the j'^{th} neuron, and update $a_{i'} \leftarrow a_{i'} + a_{j'}$.

3. Update M by removing the j'^{th} column and row, and updating the i'^{th} column (to account for the updated $a_{i'}$.)

Connections to other methods Our method relates to the popular weight-pruning method called Optimal Brain Damage (OBD) [3]. In fact, our method is equivalent to OBD if change in output activation produces proportional change in test error. Unfortunately, this is almost never the case for neural networks. Our method also weakly relates to Knowledge Distillation (KD) [1]. The idea in KD was to minimize the empirical difference in output neurons between a large "teacher" network and a smaller "student" network. In our case the small student network is the pruned version of the larger teacher network.

Determining number of neurons to prune One way to use our technique would be to keep removing neurons until the test accuracy starts going below certain levels. However, this is quite laborious to do for large networks with multiple layers. We ask whether it is possible to somehow determine the number of removals automatically. Specifically, we ask whether the saliency values of pruned neurons can be used as a proxy for test error. Figure 2 shows that it is indeed true. Unfortunately, we found that this does not hold for large networks.

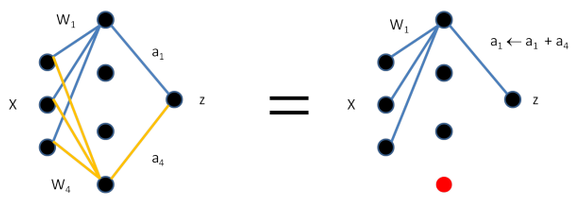


Figure 1: A toy example showing the effect of equal weight-sets ($W_1 = W_4$). Weights of the same colour in the input layer constitute a weight-set.

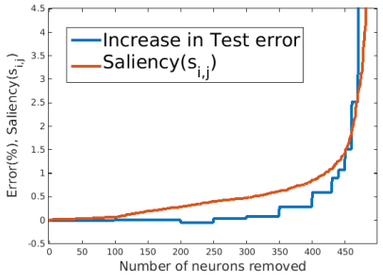


Figure 2: Scaled appropriately, the saliency curve closely follows that of increase in test error

Experiments

Toy dataset We show results on a toy dataset, comparing our method to Optimal Brain Damage and Optimal Brain Surgery. We see that our method performs better than the other two approaches, and is about 5000× faster than OBD, and 180000× faster than Optimal Brain Surgery.

MNIST-trained network We show results on an LeNet-like network trained on MNIST dataset. We were able to remove about 85% of the weights in the network, reducing the accuracy by only 1%.

AlexNet We try our method on AlexNet [2], which is a large-scale network trained on Imagenet database. We were able to remove about 35% of the total weights in the network, reducing the accuracy by 2.2%. In other words, we were able to remove about 21 million weights in a network of 60 million weights, with little effect on accuracy.

References

[1] Geoffrey E Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*, 2014.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[3] Yann LeCun, John S Denker, Sara A Solla, Richard E Howard, and Lawrence D Jackel. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pages 598–605, 1989.

Real-Time Pedestrian Detection With Deep Network Cascades

Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke
anelia,akrizhevsky,vanhoucke@google.com
Google Research
Mountain View, CA, USA
Abhijit Ogale, Dave Ferguson
ogale,daveferguson@google.com
Google X
Mountain View, CA, USA

Pedestrian detection has been an important problem for decades, given its relevance to a number of applications in robotics, including driver assistance systems, road scene understanding and surveillance systems. The two main practical requirements for fielding such systems are very high accuracy and real-time speed: we need pedestrian detectors that are accurate enough to be relied on and are fast enough to run on systems with limited compute power. This paper addresses both of these requirements by combining very accurate deep-learning-based classifiers within very efficient cascade classifier frameworks.

Deep neural networks (DNN) have been shown to excel at classification tasks [5], and their ability to operate on raw pixel input without the need to design special features is very appealing. However, deep nets are notoriously slow at inference time. In this paper, we propose an approach that cascades deep nets and fast features, that is both very fast and accurate. We apply it to the challenging task of pedestrian detection. Our algorithm runs in real-time at 15 frames per second (FPS). The resulting approach achieves a 26.2% average miss rate on the Caltech Pedestrian detection benchmark, which is the first work we are aware of that achieves high accuracy while running in real-time.

To achieve this, we combine a fast cascade [2] with a cascade of classifiers, which we propose to be DNNs. Our approach is unique, as it is the only one to produce a pedestrian detector at real-time speeds (15 FPS) that is also very accurate. Figure 1 visualizes existing methods as plotted on the accuracy - computational time axis, measured on the challenging Caltech pedestrian detection benchmark [4]. As can be seen in this figure, our approach is the only one to reside in the high accuracy, high speed region of space, which makes it particularly appealing for practical applications.

Fast Deep Network Cascade. Our main architecture is a cascade structure in which we take advantage of the fast features for elimination, VeryFast [2] as an initial stage and combine it with small and large deep networks [1, 5] for high accuracy. The VeryFast algorithm is a cascade itself, but of boosting classifiers. It reduces recall with each stage, producing a high average miss rate in the end. Since the goal is eliminate many non-pedestrian patches and at the same time keep the recall high, we used only 10% of the stages in that cascade. Namely, we use a cascade of only 200 stages, instead of the 2000 in the original work.

The first stage of our deep cascade processes all image patches that have high confidence values and pass through the VeryFast classifier. We here utilize the idea of a tiny convolutional network proposed by our prior work [1]. The tiny deep network has three layers only and features a 5x5 convolution, a 1x1 convolution and a very shallow fully-connected layer of 512 units. It reduces the massive computational time that is needed to evaluate a full DNN at all candidate locations filtered by the previous stage. The speedup produced by the tiny network, is a crucial component in achieving real-time performance in our fast cascade method.

The baseline deep neural network is based on the original deep network of Krizhevsky et al [5]. As mentioned, this network in general is extremely slow to be applied alone. To achieve real-time speeds, we first apply it to only the remaining filtered patches from the previous two stages. Another key difference is that we reduced the depths of some of the convolutional layers and the sizes of the receptive fields, which is specifically done to gain speed advantage.

Runtime. Our deep cascade works at 67ms on a standard NVIDIA K20 Tesla GPU per 640x480 image, which is a runtime of 15 FPS. The time breakdown is as follows. The soft-cascade takes about 7 milliseconds (ms). About 1400 patches are passed through per image from the fast cascade. The tiny DNN runs at 0.67 ms per batch of 128, so it can process the patches in 7.3 ms. The final stage of the cascade (which is the baseline classifier) takes about 53ms. This is an overall runtime of 67ms.

Experimental evaluation. We evaluate the performance of the Fast Deep Network Cascade using the training and test protocols established in the Caltech pedestrian benchmark [4]. We tested several scenarios by training on the Caltech data only, denoted as DeepCascade, on an independent dataset (completely disjoint from Caltech), DeepCascadeID, and on Caltech data plus some extra data from the publicly available Daimler and Eth datasets, DeepCascadeED.

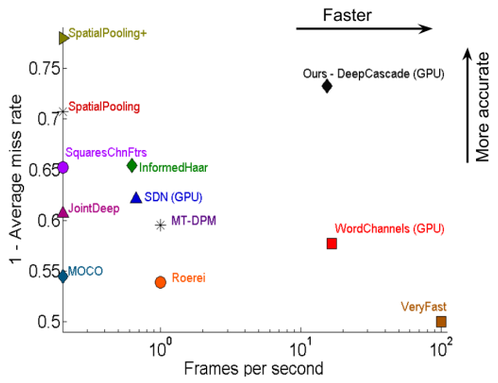


Figure 1: Performance of pedestrian detection methods on the accuracy vs speed axis. Our DeepCascade method achieves both smaller miss-rates and real-time speeds. Methods for which the runtime is more than 5 seconds per image, or is unknown, are plotted on the left hand side. The SpatialPooling+/Katamari methods use additional motion information.

Our method performs with average miss rates of 31.11%, 30.17%, and 26.21%, for DeepCascade, DeepCascadeID and DeepCascadeED respectively, and outperform most approaches. The best approaches, known on this dataset, perform at 36% [1], 29% [6] and at 22% for SpatialPooling+ and Katamari [3, 6] which use additional motion features. Our results also point to the strengths of DNNs, namely, achieving higher accuracy by simply incorporating more and higher quality data.

Apart from achieving very good accuracy, our method is much faster and runs at 15 FPS, which is real-time performance. Other real-time algorithms, we are aware of, VeryFast at 100 FPS and WordChannels at 16 FPS (on GPU) have high average miss rate of 50% and 42%, respectively. Previous methods, e.g. SDN, JointDeep have similarly HOG-based cascade and a deep network have runtime of at least 1-1.5 seconds on GPU. We further note that our algorithm is implemented using the publicly available 'cuda-convnet2' [5] and the VeryFast 'Doppia' code [2].

Conclusion. The main contribution of this work is a pedestrian detection system that is both accurate and runs in real-time. As such, it can be practically deployed within a real-life pedestrian detection system. No other prior work has demonstrated such capabilities. We expect our work to impact future methods by providing a simple to implement, accurate and effective real-time solution. Thus, future methods can continue to further push the boundaries in accuracy in pedestrian detection, while simultaneously keeping the methods fast and practically relevant.

[1] A. Angelova, A. Krizhevsky, and V. Vanhoucke. Pedestrian detection with a large-field-of-view deep network. *ICRA*, 2015.

[2] R. Benenson, M. Matthias, R. Tomofte, and L. Van Gool. Pedestrian detection at 100 frames per second. *CVPR*, 2012.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? *2nd Workshop on Road scene understanding and Autonomous driving, ECCV*, 2014.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *CVPR*, 2009.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[6] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection. *ECCV*, 2014.

Rule Of Thumb: Deep derotation for improved fingertip detection

Aaron Wetzler[1]
www.cs.technion.ac.il/~twerd
Ron Slossberg[1]
www.cs.technion.ac.il/~ronslos
Ron Kimmel[2]
www.cs.technion.ac.il/~ron

[1] Faculty of Electrical Engineering,
Technion Israel Institute of Technology
[2] Faculty of Computer Science,
Technion Israel Institute of Technology

In this paper we propose DeROT, a method for in-plane derotation of depth images using a deep convolutional neural network. The method is aimed at normalizing out the effects of rotation on highly articulated motion of deforming geometric surfaces such as hands. To support our approach we also describe a new pipeline for building a very large training database using high accuracy magnetic annotation and labeling of objects imaged by a depth camera. he proposed method reduces the complexity of learning in the space of articulated poses which is demonstrated by using two different state-of-the-art learning based hand pose estimation methods applied to fingertip detection. Significant classification improvements are shown over the baseline implementation. Our framework involves no tracking, kinematic constraints or explicit prior model of the articulated object.

DeROT: removing in-plane rotation Changing the global rotation of an object directly increases the variation in appearance of the object parts. For markerless situations, removing variability through partial canonization can significantly reduce the space of possible images used for pose learning instead of trying to explicitly learn the rotational variability through data augmentation. We therefore remove the variability as a preprocessing step during *both* a training phase *and* at run-time. To this end we propose to learn the rotation using a deep convolutional neural network (CNN) in a regression context based on a network similar to that of [4]. We show how this can be used to predict full three degrees of freedom (3 DOF) orientation information by training on a large database of hand images captured by a depth sensor. This is then combined with a useful insight which we call "Rule of thumb": there is almost always an in-plane rotation which can be applied to an image of the hand which forces the base of the thumb to be on the right side of the image. Synthetic and real examples of the results of applying DeROT to images of a hand can be seen in Figure 1.

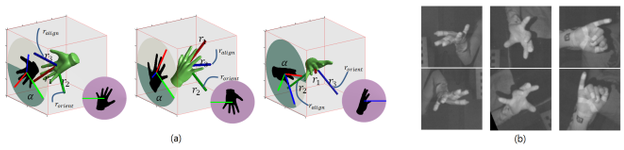


Figure 1: Synthetic and real examples of DeROT. a) The depth projection of the virtual hand before applying DeROT can be seen on the left wall of the cube representing the camera plane. The axis marked r_{orient} is projected onto the camera plane and used in DeROT to define the angle α. The purple circle contains the resulting image of the hand after applying derotation by angle α. b) The top row of images are un-derotated. The bottom row have been derotated by α obtained by DeROT. Note that the thumb is consistently on the right of the image.

Fingertip detection. In this work we specifically focus on per frame fingertip detection in depth images without either tracking or kinematic modeling. We propose useful modifications to the popular machine learning based methods of Keskin *et al*. [3] and Tompson *et al*. [4]. Our preprocessing step then involves cropping input images of hands and rotating them about their center of mass using the predicted angle of derotation produced by DeROT.

The HandNet database No currently available hand data-sets (*e.g*. [5],[2],[4]) include accurate full 3 DOF ground truth hand orientations on a large database of real depth images. A significant contribution of this paper is therefore the creation of a new, large-scale database of fully annotated depth images of hands that we call HandNet[1]. To build and annotate this database we use a RealSense camera combined with 2*mm* TrakStar [1] DC magnetic trackers. Sensor slippage is prevented by affixing the sensors inside tight elastic seams. The close fitting pockets also prevent the hand depth profile from being affected by the attached sensors. To

[1]To advance research in the field this database and relevant code is available at `www.cs.technion.ac.il/~twerd/HandNet/`.

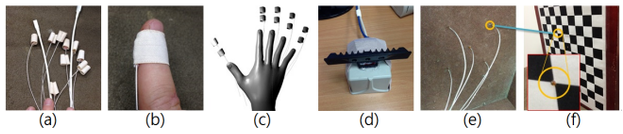


Figure 2: The data capture setup. a) 2mm magnetic sensors. The larger rectangular sensors are not used. b) A fingertip sensor inside the inner seam. c) Virtual model used for planning a multi-sensor setup. We only use 5 sensors. d) The RealSense camera rigidly fixed to the TrakStar transmitter. e) The back of the wooden calibration board where the glass sensor housings are firmly pushed through. f) The front of the calibration board where the glass sensor housings are visible on the corners as seen in the inset.

calibrate between the camera and sensor frames we position the magnetic sensors on the corners of a checkerboard pattern to create physical correspondence between the detected corner locations and the actual sensors. The setup can be seen in Figure 2. Sensors are modeled as 3D oriented ellipsoids and ray-cast into the camera frame. Discrete fingertip labels as well as heat-maps and orientation information are then trivially associated with each input image. The database is created from 10 participants in total who perform random hand motions with extensive pose variation. In total, after quality filtering, HandNet contains 212928 unique poses which is to the best of our knowledge the largest annotated, front facing database of real hand images currently available.

Experiments We perform our experiments using our HandNet database and the publicly available database NYUHands [4]. All experiments are performed separately on the two data sets. Using our deep derotation method we show up to 20.5% improvement in mean average precision (mAP) over our baseline results for both fingertip detection methods. We also compare our results to a non-learning based method similar to PCA and show that it produces inferior results. This shows that using derotation, specifically DeROT, significantly improves the localization ability of machine-learning based per-frame fingertip detectors by reducing the variance of the observed image space. Furthermore we find that this procedure works despite the extremely high range of potential poses. We see this approach as an alternative to data augmentation and as a useful preprocessing step in pipelines dedicated to articulated object pose extraction such as hands. For our experiments a single random decision tree mostly outperforms a convolutional neural network. Although they are trained with different data and objectives it hints that there is no silver bullet to determining which machine learning approach is more appropriate.

Acknowledgments This research was supported by European Community's FP7- ERC program, grant agreement no. 267414.

[1] Ascension TrakStar. `http://www.ascension-tech.com/`, 2015.

[2] Q. Chen, S. Xiao, W. Yichen, T. Xiaoou, and S. Jian. Realtime and robust hand tracking from depth. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1106–1113. IEEE, 2014.

[3] C. Keskin, F. Kiraç, Y. Emre Kara, and L. Akarun. Real time hand pose estimation using depth sensors. In *Consumer Depth Cameras for Computer Vision*, pages 119–137. Springer, 2013.

[4] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (TOC)*, 33, 2014.

[5] W. Zhao, J. Chai, and Y. Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Symposium on Computer Animation*, pages 33–42, 2012.

Riesz-based Volume Local Binary Pattern and A Novel Group Expression Model for Group Happiness Intensity Analysis

Xiaohua Huang[1]
huang.xiaohua@ee.oulu.fi
Abhinav Dhall[2,3]
abhinav.dhall@canberra.edu.au
Guoying Zhao[1]
gyzhao@ee.oulu.fi
Roland Goecke[2,3]
roland.goecke@ieee.org
Matti Pietikäinen[1]
mkp@ee.oulu.fi

[1] Center for Machine Vision Research
Department of Computer Science and Engineering
University of Oulu, Finland

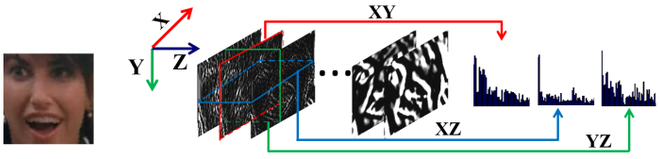
[2] Human-Centred Technology Research Centre
University of Canberra, Australia

[3] IHCC Group
Research School of Computer Science
Australian National University, Australia

Automatic emotion analysis and understanding has received much attention over the years in affective computing. This paper aims to infer the emotional intensity of a group of people. For group emotional intensity analysis [1], feature extraction and group expression model (GEM) are two critical issues.

Feature extraction: For 2D image, the 1st-order Riesz transform enables the rotationally invariant analysis of i1D structure, while the 2nd-order Riesz transform can characterize an i2D image structure [4]. Therefore, we employ the 1st-order and 2nd-order Riesz transforms with log-Gabor filters at three scales and four orientations to describe the structure of a facial expression image.

Its Riesz face images are resembled to form the Riesz volume. We furthermore employ the LBP operator on XY, XZ and YZ of volume-based Riesz face respectively. The histograms H_{XY}, H_{XZ} and H_{YZ} are concatenated into one feature vector H. The procedure is shown in Fig. 1. Five histograms $H_x, H_y, H_{xx}, H_{xy}, H_{yy}$ are concatenated into one feature vector F for representing the face. Based on a low-dimensional space $\mathbf{U}$, the low-dimensional features can be formulated as $\widetilde{F} = \mathbf{U}'F$.

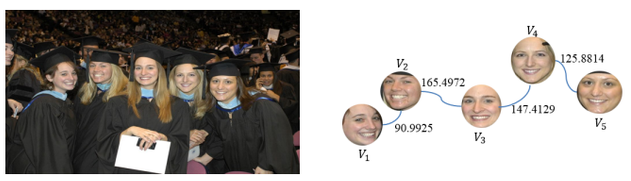


(a) (b)

Figure 1: Riesz-based volume local binary patterns for R_x: (a) A face image; (b) The 1st-order Riesz face R_x and its feature extraction.

Group expression model: GEM explores the relationship between faces and intensity label in a group image. We formulate a novel GEM based on Continuous Conditional Random Fields (GEM_{CCRF}) to combine global and local attributes.

For the global attribute, the minimal spanning tree algorithm [3] is employed to obtain a fully connected graph $\mathcal{G} = (V, E)$, which can provide the location and minimally connected neighbours of a face. An example is shown in Fig. 2. Based on the graph, the global attribute can be expressed by relative size S_i and relative distance δ_i. For the local attribute, it contains the local features $\widetilde{F}$ for faces.



(a) (b)

Figure 2: Illustration of obtaining minimally connected neighbours for a face: (a) Original group image, (b) Min-span tree depicting connection between faces.

Global and local attributes are concatenated into $\mathcal{F} = [\widetilde{F}, S, \delta]$, providing sufficient and useful information for CCRF. We suppose that a group image contains n faces, which is represented by $\mathcal{F}_i, i = 1, \ldots, n$. We transform $\mathcal{F}_i$ into t_i by using Support Vector Regression, which represents the relevance factor of one subject for CCRF. Therefore, CCRF model for a group image is a conditional probability distribution with the probability density function as $P(z \mid \mathbf{T}) = \frac{\exp(\sum_{i=1}^{n}\sum_{k=1}^{m}\mu_k f_k(z_i,\mathbf{T}_{i,k}) + \sum_{i,j}\nu g(z_i,z_j,\mathbf{T}))}{\int \exp(\sum_{i=1}^{n}\sum_{k=1}^{m}\mu_k f_k(z_i,\mathbf{T}_{i,k}) + \sum_{i,j}\nu g(z_i,z_j,\mathbf{T}))d\mathbf{z}}$, where $\mathbf{T}$ is the set of input feature vector, $\mathbf{Z}$ is the intensity label of faces in a group image, $f(.)$ and $g(.)$ are vertex and edge features, respectively. Following the work of [2], we pick μ and ν that maximise the conditional log-likelihood $(\mu^*, \nu^*) = \arg\max_{\mu,\nu} \sum_{q=1}^{M} \log P(\mathbf{z}^{(q)} | \mathbf{T}^{(q)})$, where M expresses the number of group images.

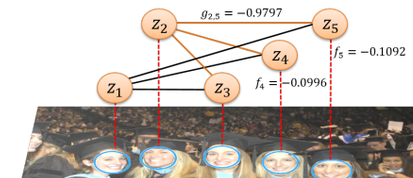


Figure 3: GEM_{CCRF}: The blue circle on a facial image represents the extracted content, z_i is an happiness intensity label, an edge (a solid line) between z_i and z_j means the dependency between intensities of two faces, an edge (a dash line) represents the dependency of an intensity label on its content.



(a) (b)

Figure 4: Examples of inferring two group images on HAPPEI database. For each image the texts at the bottom of an image indicate the estimated happiness intensity results and the ground truth, respectively

Given a group image, containing n faces, the intensity can be estimated by using $\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n} \arg\max_{z_i} P(z_i | \mathbf{T}_{test})$. An example of the CCRF model in group happiness intensity analysis is depicted in Fig. 3.

Implementation of this framework by group-level happiness intensity estimation, as shown in Fig. 4, is described in the paper. Our conclusion is that GEM_{CCRF} helps to predict the group mood better. Moreover, the feature indeed affects the performance of recognition for GEM models.

[1] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Transaction on Affective Computing*, 6(1): 13–26, 2015.

[2] V. Imbrsaitė, T. Baltrušaitis, and P. Robinson. Emotion tracking in music using continuous conditional random fields and relative feature representation. In *IEEE International Conference on Multimedia and Expo Workshops*, pages 1–6, 2013.

[3] R. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957.

[4] L. Zhang and H. Li. Encoding local image patterns using Riesz transform: With applications to palmpring and finger-knuckle-print recognition. *Image and Vision Computing*, 30(12):1043–1051, 2012.

Weakly Supervised Metric Learning towards Signer Adaptation for Sign Language Recognition

Fang Yin[1,2]
fang.yin@vipl.ict.ac.cn
Xiujuan Chai[1,2]
xiujuan.chai@vipl.ict.ac.cn
Yu Zhou[3]
zhouyu@iie.ac.cn
Xilin Chen[1,2]
xlchen@ict.ac.cn

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS
Beijing, 100190, China

[2] Cooperative Medianet Innovation Center, China

[3] Institute of Information Engineering, CAS
Beijing, 100093, China

In this paper, we introduce metric learning into Sign Language Recognition(SLR) for the first time and propose a signer adaption framework to address signer-independent SLR. For adapting the general model to the new signer, both clustering and manifold constraints are considered in the adaptive distance metric optimization.

The contribution of our work mainly lies in three-folds. Firstly, a Weakly Supervised Metric Learning(WSML) framework is proposed, which combines the clustering and manifold constraints simultaneously. Secondly, the general framework is applied to signer adaptation and achieves good performance. Thirdly, a fragment based feature is designed for sign language representation and the effectiveness is verified in large vocabulary datasets.

Our proposed WSML framework can be decomposed into two key steps. The first one is to learn a generic metric from the given labeled data. Then the second step is to realize the distance metric adaptation by considering the clustering and manifold constraints with the unlabeled data.

To learn a generic distance metric, the labeled data are used under clustering assumption with classical large margin hinge loss. Specifically, the distances between data points within the same cluster(with same label) should be minimized and the distances between data points from different clusters(with different labels) should be maximized. Here we define the index set with same labels as $S_g = \{(i,j)|y_i = y_j, x_i, x_j \in X_l\}$ and the index triplet $B_g = \{(i,j,k)|y_i = y_j, y_i \neq y_k, x_i, x_j, x_k \in X_l\}$. The objective function is

$$\begin{aligned} \min J_g = & \min_{M_g} \sum_{(i,j)\in S_g} D_g(x_i,x_j) + \sum_{(i,j,k)\in B_g} [1 + D_g(x_i,x_k) - D_g(x_i,x_j)]_+ \\ & s.t. M_g \geq 0, \end{aligned} \tag{1}$$

where the term $[z]_+ = max(z,0)$ denotes the standard hinge loss and M_g is the metric matrix for distance metric D_g.

After learning the generic distance metric, clustering constrain on both the labeled and unlabeled data can be considered. First of all, with the generic distance metric D_g determined by matrix M_g and the labeled data set X_l, a classifier is trained. Thus the labels of the unlabeled data can be predicted. We hope that the generic distance metric can be adaptive to the data from a different distribution through the predicted labels, although they are uncertain. For each unlabeled data x_i, the classifier gives a predicted label y_i and belief b_i, where $0 \leq b_i \leq 1$. Here the label y_i is not precise so the metric learning method with these uncertain labels is weakly supervised. The objective function of clustering constraint is

$$J_c = \sum_{(i,j)\in S_c} b_i D(x_i,x_j) + \sum_{(i,j,k)\in B_c} b_i [1 + D(x_i,x_k) - D(x_i,x_j)]_+ \tag{2}$$

where S_c is a index set defined as $S_c = \{(i,j)|y_i = y_j, x_i \in X_u, x_j \in X_l\}$. While B_c is a triplet set $B_c = \{(i,j,k)|y_i = y_j, y_i \neq y_k, x_i \in X_u, x_j, x_k \in X_l\}$.

To further investigate the topological structure of the unlabeled data, manifold assumption is considered. Specifically, if two data points x_i and x_j are close in the intrinsic geometry of Euclidean distance, they should also be close to each other in the Mahalanobis distance. A straightforward expression for manifold assumption is

$$J_m = \sum_{i=l+1}^{u} \sum_{j=l+1}^{u} D(x_i,x_j) W_{ij}, \tag{3}$$

where W is a weight matrix. $W_{ij} = 1$ if x_i is among the $k-$nearest neighbors of x_j or x_j is among the $k-$nearest neighbors of x_i. Otherwise, $W_{ij} = 0$.

The final objective function of metric adaptation can be formulated by incorporating expressions (2) and (3) into (1).

$$\min J = \min(J_g + \alpha J_c + \beta J_m), \tag{4}$$

where α and β are the weights corresponding to clustering constraint and manifold constraint respectively.

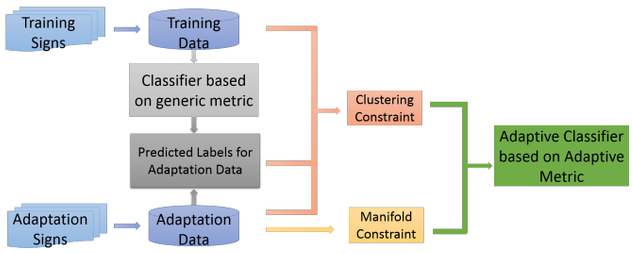


Figure 1: The pipeline of the signer adaptation.

Another main contribution of our work is applying WSML to signer-independent SLR. Figure 1 illustrates the pipeline of the signer adaptation using WSML. Concretely speaking, the features are extracted from the labeled signs and the unlabeled adaptation signs. With the features of labeled training data, a generic distance metric can be learnt, with which a classifier is trained. Then the labels of unlabeled data are predicted with believes. In the adaptation stage, both the labeled and unlabeled training data are used for adaptive metric learning. Specifically, the labeled data and the unlabeled data with their predicted labels are used together for clustering constraint, while the unlabeled data are used for manifold constraint. The classifier based on the adaptive metric is suitable for the new signer and is used for recognition.

WSML is compared with HMM[1], DTW, ARMA[2] and generic metric learning (ML) on our own dataset in the experiments. The accuracies are shown in Figure2(a). Figure2(b) illustrates the accuracy trend with the increasing of adaptation data size.

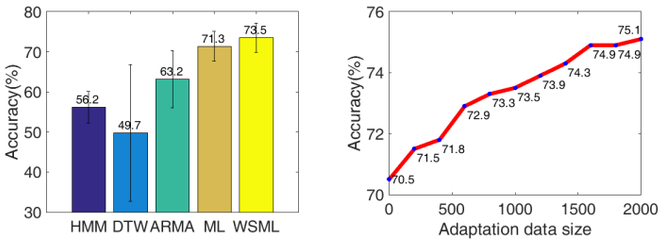


Figure 2: The accuracies of our proposed method and the baseline methods are given in (a). (b) shows the accuracies of our proposed method with different adaptation data sizes.

[1] Chunli Wang, Wen Gao, and Shiguang Shan. An approach based on phonemes to large vocabulary chinese sign language recognition. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 411–416. IEEE, 2002.

[2] Chunyan Xu, Tianjiang Wang, Junbin Gao, Shougang Cao, Wenbing Tao, and Fang Liu. An ordered-patch-based image classification approach on the image grassmannian manifold. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(4):728–737, 2014.

Hashmod: A Hashing Method for Scalable 3D Object Detection

Wadim Kehl[1]
kehl@in.tum.de
Federico Tombari[12]
federico.tombari@unibo.it
Nassir Navab[1]
navab.cs.tum.edu
Slobodan Ilic[3]
slobodan.ilic@siemens.com
Vincent Lepetit[4]
lepetit@icg.tugraz.at

[1] Computer-Aided Medical Procedures, TU Munich, Germany
[2] Computer Vision Lab (DISI), University of Bologna, Italy
[3] Siemens AG Research & Technology Center Munich, Germany
[4] Institute for Computer Graphics and Vision, TU Graz, Austria

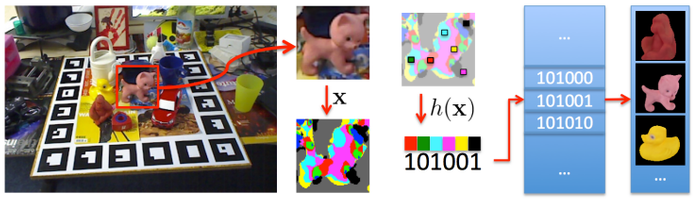


Figure 1: A sliding window evaluates learned hash functions h computed on extracted LineMOD features x to efficiently index into subsets of candidate views for further matching.

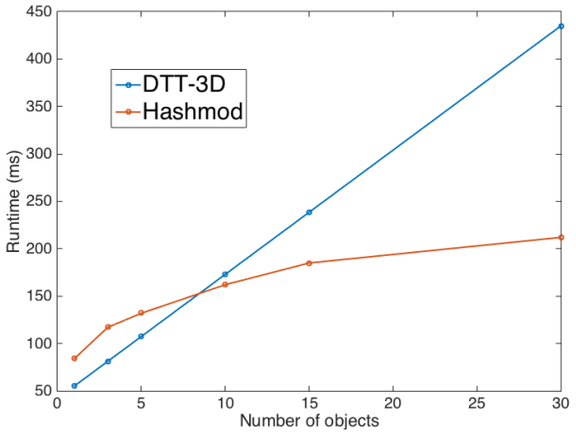


Figure 2: Runtime comparison with the state-of-the-art template matching DTT-3D [2]. Our method overtakes due its sublinear time complexity.

We present a scalable method for detecting objects and estimating their 3D poses in RGB-D data. To this end, we rely on an efficient representation of object views and employ hashing techniques to match these views against the input frame in a scalable way. While a similar approach already exists for 2D detection, we show how to extend it to estimate the 3D pose of the detected objects. In particular, we explore different hashing strategies and identify the one which is more suitable to our problem. We show empirically that the complexity of our method is sublinear with the number of objects and we enable detection and pose estimation of many 3D objects with high accuracy while outperforming the state-of-the-art in terms of runtime, as depicted in Figure 2.

Descriptor computation Given a database of M objects, we synthetically create N views for each object from poses regularly sampled on a hemisphere of a given radius. From this, we compute a set $\mathcal{D}$ of d-dimensional binary descriptors

$$\mathcal{D} = \{\mathbf{x}_{1,1}, ..., \mathbf{x}_{M,N}\}, \quad (1)$$

where $\mathbf{x}_{i,j} \in \mathbb{B}^d$ is the descriptor for the i-th object seen under the j-th pose. We use LineMOD [1] in practice to compute these descriptors and concatenate the binary representation to obtain the binary strings $\mathbf{x}_{i,j}$.

Hash learning We learn several hashing functions whose purpose is to immediately index into a subset, often called a "bucket", of $\mathcal{D}$ when applied to a descriptor $\mathbf{x} \in \mathbb{B}^d$ during testing (see Figure 1). These buckets are filled with descriptors from $\mathcal{D}$ with the same hash value so that we can restrict our search for the nearest neighbor of $\mathbf{x}$ to the bucket retrieved via the hashing function instead of going through the complete set $\mathcal{D}$. It is very likely, but not guaranteed, that the nearest neighbor is in at least one of the buckets returned by the hashing functions. In practice, a careful selection of the hashing functions is important for good performance. Since the descriptors $\mathbf{x}$ are already binary strings, we design our hashing functions $h(\mathbf{x})$ to return a short binary string made of b bits directly extracted from $\mathbf{x}$. This is a very efficient way of hashing and we will refer to these short strings as hash keys. We evaluated multiple strategies and settled for the one below that is inspired by greedy tree growing for Randomized Forests:

Starting with a set of descriptors at the root, we determine the bit that splits this set into two subsets with sizes as equal as possible, and use it as the first bit of the key. For the second bit, we decide for the one that splits those two subsets further into four equally-sized subsets and so forth. We stop if b bits have been selected or one subset becomes empty. This procedure alone yields a balanced tree with leafs of similar numbers of elements. Each hash key can be regarded as a path down the tree and each leaf represents a bucket. Note that such a balanced repartition ensures retrieval and matching at a constant speed. We now further adapt the strategy to our problem: to improve detection rates we favor similar views of the same object to go into different branches. The idea behind this strategy is to reduce misdetections due to noise or clutter in the descriptors, leading to different buckets during testing. Formally, the j-th bit B of the key is selected by solving:

$$\arg\min_B \frac{1}{|N_i|} \sum_i \left| |\mathcal{S}_L^B(N_i)| - |\mathcal{S}_R^B(N_i)| \right| + \frac{1}{|N_i|^2} \left(P(\mathcal{S}_L^B(N_i)) + P(\mathcal{S}_R^B(N_i)) \right), \quad (2)$$

where $N_i \subset \mathcal{D}$ is the set of descriptors contained by the i-th node at level j, and $\mathcal{S}_{\{L,R\}}^B(N_i)$ are the two subsets of N_i that go into the left and right child induced by splitting with B. The first term in above equation balances equal splits whereas the second penalizes similar views of the same object falling into the same side of the split.

Results We ran our method on the LineMOD ACCV12 dataset [1] consisting of 15 objects. We are able to consistently detect at around 95% – 96% accuracy on average which is slightly worse than LineMOD and DTT-3D. However, we are always faster than LineMOD and overtake DTT-3D at around 8 objects where our constant-time hashing overhead becomes negligible and the methods' time complexities dominate. This is important to stress since real scalability comes from a sublinear growth.

[1] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradsky, K. Konolige, and N. Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In *ACCV*, 2012.

[2] R. Rios-Cabrera and T. Tuytelaars. Discriminatively trained templates for 3D object detection: A real time scalable approach. In *ICCV*, 2013.

Multi-Task Transfer Methods to Improve One-Shot Learning for Multimedia Event Detection

Wang Yan
wyan@sfu.ca
Jordan Yap
jjyap@sfu.ca
Greg Mori
mori@cs.sfu.ca

School of Computing Science
Simon Fraser University
Burnaby, BC, CANADA

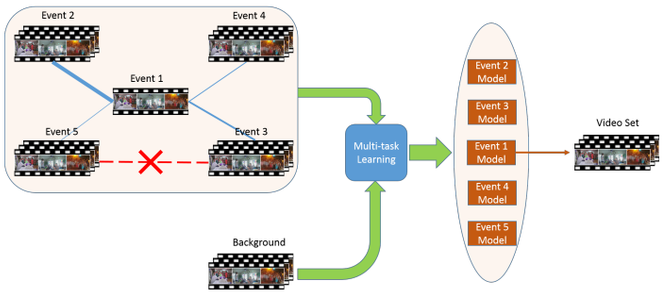


Figure 1: Overview of the proposed method – one-shot learning leveraging other categories. Given one example for an event of interest (Event 1), we implicitly infer the relevance between it and other events, and emphasise more on the most relevant ones in the multi-task learning. The learned classifier for the event of interest is applied to retrieve instances from a video test set.

This paper proposes a new multi-task learning method with implicit inter-task relevance estimation, and applies it to complex Internet video event detection, which is a challenging and important problem in practice, yet seldom has been addressed. In this paper, "detection" means to detect videos corresponding to the event of interest from a (large) video dataset, not to localize the event spatially or temporally in a video. In the problem definition, one positive and plenty of negative samples of one event are given as training data, and the goal is to return the videos of the same event from a large video dataset. In addition, we assume samples of other events are available.

Fig. 1 shows an overview of the proposed methods. The widths of the lines between the one-exemplar event and others represent the inter-event relevance, which is unknown a priori in our problem settings. However, the proposed method can implicity infer the relevance and utilize the most relevant event(s) more in multi-task learning, where the shared information from the relevant events helps to build a better model from the one exemplar. The proposed method does not assume the relevance between other events, as indicated by the red line. Although the learning algorithm outputs models of all input events, only that of the one-exemplar event is applied to detect videos of the event of interest from the video set.

Our method builds on the approach of graph-guided multi-task learning [1], which is described first. The training set $\{(\mathbf{x}_{ti}, y_{ti}) \in \mathbb{R}^D \times \{-1,+1\}, t = 1,2,\ldots,T, i = 1,2,\ldots,N_t\}$ is grouped into T related tasks, which are further organized as a graph $G =< V,E >$. The tasks correspond to the elements in the vertex set V, and the pairwise relevance between Task t and k are represented by the weight r_{tk} on edges $e_{tk} \in E$. The more relevant the two tasks are, the larger the edge weight is. The graph guided multi-task learning algorithm learns the corresponding T models jointly, by solving the optimization problem

$$\min_{\mathbf{W},\mathbf{b}} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \text{Loss}(\mathbf{w}_t^T \mathbf{x}_{ti} + b_t, y_{ti}) + \lambda_1 \|\mathbf{W}\|_F^2 + \lambda_2 \Omega(\mathbf{W}), \quad (1)$$

where $\mathbf{w}_t \in \mathbb{R}^D$ and $b_t \in \mathbb{R}$ are the model weight vector and bias term of Task t, respectively, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T)$ is the matrix whose columns are model weight vectors, $\mathbf{b} = (b_1, b_2, \ldots, b_t)$, $\|\mathbf{W}\|_F^2 = \text{Trace}(\mathbf{W}^T\mathbf{W})$ is the squared Frobenius norm,

$$\Omega(\mathbf{W}) = \sum_{e_{tk} \in E, t<k} r_{tk} \|\mathbf{w}_t - \mathbf{w}_k\|_2^2 \quad (2)$$

is the graph-guided penalty term. For the significantly relevant tasks, their model weight vectors are forced to be similar due to the large edge weights, and the information could be transferred between relevant tasks. $\text{Loss}(\cdot,\cdot)$ is the loss function. In our work we use logistic loss

$$\text{Loss}(s,y) = \log(1 + \exp(-ys)), \quad (3)$$

which is smooth and leads to an easier optimization problem compared to the hinge loss.

In the one-shot learning setting for event detection, it is hard to learn a good model for the specific event of interest from the only one positive sample. However, due to the potential relevance between events, one can expect a better model by applying multi-task learning to the event with one positive sample and some other events. In the multi-task setting, each of these other events corresponds to one task.

Different external events may share different common "part" with the event of interest. Consider that "birthday party" as the event of interest, and it is relevant to "parade" since both have lots of people inside, and it may be also relevant to "preparing food" due to the food itself. However, there is little relevance between "parade" and "preparing food". Therefore, cluster-based multi-task learning may not be used to learning the event of interest, because not all external events relevant to the event of interest are relevant enough to each other to fit into one cluster. In contrast, graph-guided multi-task does not assume the clustered structure of the tasks, and it a better choice for this task.

Without losing the generality, we assume Event 1 is the event of interest and others are external event in the following. Given the graph in Fig. 1, the formulation in (1) is not directly applicable because the pairwise relevance is unknown. However, the min operation can be added to select the most relevant tasks automatically, and they are all equally weighted, i.e. using

$$\Omega(\mathbf{W}) = \sum_{t \in \mathcal{T}_K} \|\mathbf{w}_1 - \mathbf{w}_t\|_2^2 \quad (4)$$

instead of (2), where $\mathcal{T}_K$ is the set of indices corresponding to the minimum K elements in $\{\|\mathbf{w}_1 - \mathbf{w}_t\|_2^2\}_{t=2}^T$. The most relevant tasks and task models are jointly optimised in the training process. The minimum operation can be further replaced by softmin function to make the objective smooth, i.e.

$$\Omega(\mathbf{W}) = -\log \sum_{t=2}^{T} \exp(-\|\mathbf{w}_1 - \mathbf{w}_t\|_2^2). \quad (5)$$

This penalty focuses more on the smallest inter-model distance, which is slightly different from the former one with equally weighted K smallest distances. The experiments show that we can get good results with the smooth penalty. In addition to squared l_2 distance, one can also use the penalty term represented by correlations. The term still focuses more on most relevant tasks, but softmax is used in the representation, i.e.

$$\Omega(\mathbf{W}) = -\log \sum_{t=2}^{T} \exp(\mathbf{w}_1^T \mathbf{w}_t). \quad (6)$$

We use the first option in the following experiments. All penalties in this subsection make the objective non-convex, but one can still get good results empirically. The objective is optimized by Quasi-Newton Soft-Threshold (QNST) method [2].

[1] Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[2] Mark Schmidt. *Graphical model structure learning with l_1-regularization*. PhD thesis, UNIVERSITY OF BRITISH COLUMBIA, 2010.

Appearance and Depth for Rapid Human Activity Recognition in Real Applications

Stavros Tachos
http://www.iti.gr/iti/people/Stavros_Tachos.html
Konstantinos Avgerinakis
http://www.iti.gr/iti/people/KAfgerinakis.html
Alexia Briasouli
http://www.iti.gr/iti/people/Alexia_Briasouli.html
Ioannis Kompatsiaris
http://www.iti.gr/iti/people/Ioannis_Kompatsiaris.html

Centre for the Research & Technology Hellas (CERTH)
Information Technologies Institute (ITI)
Thessaloniki, Greece

Introduction: As the demand for better assisted living and smart home environments continuously rises, the development of highly accurate and efficient human activity recognition algorithms becomes a necessity. We propose a novel technique for activity localisation and recognition from colour-depth sequences recorded with the Kinect sensor, specifically tailored for the recognition of Activities of Daily Living (ADLs). Comparative analysis with SoA [1, 2] on three challenging ADL datasets indicates that our algorithm is very appropriate for real life scenarios as it achieves SoA accuracy while performing 10-20 times faster. Our RGB-D video processing framework consists of four stages (Fig.1): i) the depth image is refined in order to fill the missing values produced by the sensor. ii) Activity is detected on a grid of Spatio-Temporal Activity Cells (STACs) over the video sequence. iii) Activity Representation takes place by extracting features from the 3D volume comprised of all STACs that contain activity, iv) the features are encoded with Fisher Vectors of fixed size and activity recognition is implemented with a SVM classifier.

Depth Refinement: Depth data provided by low cost colour-depth sensors contains missing values (depth holes) often caused by light reflection or high frequency light sources that add noise to the IR signal. Our refinement strategy operates pixelwise by maintaining a small set of the most recent depth values that are used to compute a median depth value for each pixel. Afterwards, holes are eliminated by filling missing values with the computed medians. Since this technique is being applied at every frame, the median values change through time and adapt to future variations on the depth image.

Activity Localisation: A grid is applied throughout the video sequence, comprised of *STACs* of size 24x24xW_{STAC}, with W_{STAC}=3 frames. Each grid cell over time contains a spatiotemporal $STAC_i$, which is characterised by a HOG and a Histogram of Depth (HoD), extracted around its center. Localisation is performed by our adaptive background model that exploits two statistical criteria to determine whether a $STAC_i$ is "active" (i.e foreground) or "inactive" (i.e. background): i) the minimum Chi-Square distance and ii) the homogeneity criterion. More specifically, two History Volumes of temporal size $N = 2 * fps$ are maintained for every $STAC_i$: the Foreground History Volume ($FgHV_i$) containing the N most recent "active" *STACs* and the Background History Volume ($BgHV_i$) containing the N most recent "inactive" STACs. The minimum Chi-square distance criterion indicates whether a $STAC_i$ can be considered as background based on its corresponding $BgHV_i$, while the homogeneity criterion serves to detect foreground *STACs* with no significant changes over time and need to become part of the background.

Activity Recognition: The STACs characterised as "active" through a video sequence form the 3D activity volume. This volume is continuously sampled, and sampled points are tracked over time using the KLT tracker, as long as they remain inside the volume. HOG (Histogram of Oriented Gradients) and HOSNP (Histogram of Surface Normal Projections) descriptors that represent the activity performed are extracted around every tracked point, concatenated with the points' 3D-Trajectory. The HOSNP, a 9-bin histogram of the orientation of the projections of surface normals offers significant surface information at a very low computational cost, mainly because it is extracted from finite depth differences of pixels within a small neighbourhood. Fisher encoding is applied on the HOG+HOSNP+3Dtrajectory descriptors computed in the former stage in order to get a compact representation of the activity performed. Final activity recognition is carried out by the use of a multiclass SVM.

Results: Experiments took place on three datasets (DemCare1, DemCare2, DemCare3) of elderly people performing ADLs, available for benchmark purposes upon request (http://www.demcare.eu/results/datasets). The

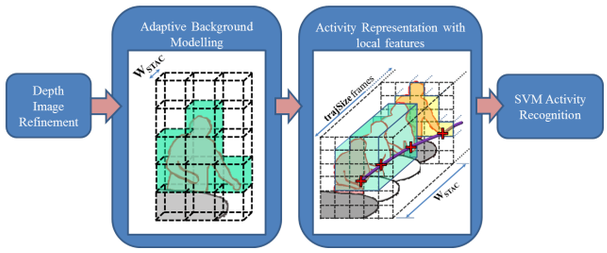


Figure 1: Overview of our Activity localisation and Recognition solution: (from left to right) a) depth frame refinement corrects noisy depth values. b) Adaptive background modelling uses HOG and HoD to separate *active* from *inactive* STACs. c) HOG, HOSNP and 3D trajectories are accumulated over time to represent human activities. d) Fisher encoding over the whole video trains a multiclass SVM model.

Datasets	HOG	HOSNP	HOG+ HOSNP	HOG+ HOSNP+ 3D Traj	[1]	[2]
DemCare1	70.2	81.4	85.3	89.6	85.1	**90.2**
DemCare2	66.9	69.9	75.0	79.9	**83.2**	79.9
DemCare3	80.7	79.1	88.6	**94.5**	93.3	91.7

Table 1: Average Accuracy of the proposed method for different combinations of descriptors and average accuracy of SoA methods[1,2]

640x480 videos of these datasets contain a variety of activities (e.g. Drink Beverage, Eat Snack, Talk to Visitor, Start Phonecall, End Phonecall, Prepare Hot Tea, Read Article, etc) and many anthropometric differences between subjects. Moreover, each dataset is recorded in a different environment at a unique sampling rate.

Results demonstrate that our method achieves accuracy that is highly competitive to SoA algorithms [1, 2] that make use of Optical Flow, while maintaining a very low computational cost. More specifically, the proposed method resulted in a -0.6% accuracy compared to SoA for D1 while performing 14.8 **times faster**. Similarly, a -3.3% accuracy deficit from SoA was reported for DemCare2 with a 11.8 **faster** computation, and lastly, our algorithm outperformed SoA on DemCare3 ($+1.2\%$ accuracy) while performing 21.4 **times faster**.

Further analysis exposes the value of the descriptors chosen for activity representation. Activity Recognition was carried out with different combinations of descriptors and as shown in table 1. The combination of HOG and HOSNP significantly increases the mean accuracy, demonstrating that these descriptors incorporate different aspects of the video data. Lastly, the concatenation of 3D trajectories boosts the accuracy even further, as more motion information is introduced to the final descriptor.

Conclusions: Our work proposes a new approach for feature-based activity localisation and recognition from RGB-D image sequences. Our proposed algorithm achieves SoA accuracy while maintaining a low computational cost. The development of a full spatio-temporal activity localisation system is scheduled as a future work.

[1] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Recognition of activities of daily living for smart home environments. In *9th International Conference on Intelligent Environments (IE2013)*, 2013.

[2] H Wang and C. Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558, 2013.

Robust Wearable Camera Localization as a Target Tracking Problem on SE(3)

Guillaume Bourmaud, Audrey Giremus

Univ. Bordeaux, CNRS, IMS, UMR 5218, F-33400 Talence, France

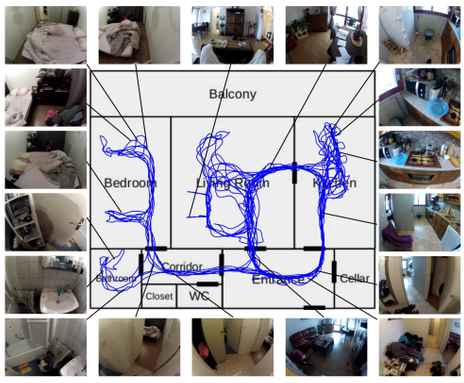
1 Introduction

In this paper, we are interested in Visual Indoor Localization (VIL) for challenging video sequences coming from a single monocular camera where the person wearing the camera performs daily living activities (see Fig.1(a)). The difficulty of this problem resides in the fact that: i) hand-held objects are frequently interposed between the camera and the environment; ii) strong motion blur and differences in illumination occur; iii) the environment changes between the images of the database and the video frames to localize, and the viewpoints can be significantly different.

We wish to develop a method that:
- relies only on the images coming from the wearable camera, i.e no other sensor such as Inertial Measurement Units should be used
- estimates the camera position with a sub-meter level accuracy as well as its orientation
- is consistent with the topology of the environment, i.e the camera trajectory should not cross walls
- is able to detect when the data is not sufficient to disambiguate the situation, i.e when the posterior distribution of the camera trajectory is multi-modal and/or too dispersed.



(a) Video frame examples of a person performing daily living activities. Hand-held objects are frequently interposed between the camera and the environment.



(b) Illustration of a database of images of an apartment annotated with map coordinates. This database of 6000 images was generated automatically from a training video sequence. The blue line corresponds to the 2D positions of all the images of the database. Note that each image is actually annotated with a 6-dof pose (3D position and 3D orientation) but only the 3D position projected onto the plan of the apartment is presented in this figure.

Figure 1: (Best seen in colors) Left: Examples of video frames to localize. Right: Illustration of the database

2 Contribution

In this context, we propose a novel VIL framework which is able to satisfy the previous technical specifications. The contribution of this paper is threefold:

1. We formulate the localization problem as a target tracking problem on the Lie group of camera motions SE(3) , where the measurements are map coordinates obtained by applying a Content Based Image Retrieval (CBIR) algorithm to the video frames.

2. In order to solve this problem, we derive a novel Rao-Blackwellized Particle Smoother on Lie Groups (LG-RBPS), which builds upon the recently proposed Extended Kalman Filter on Lie groups [1] and the Rauch-Tung-Striebel Smoother on Lie groups that we also derive in this paper.

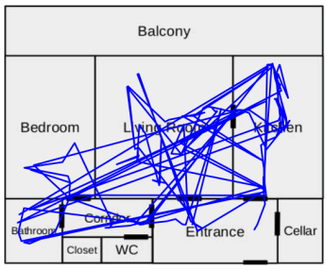
3. To take into account the topology of the environment, we propose to introduce Virtual Measurements (VM) that guide the particles and prevent them from crossing walls.

3 Results

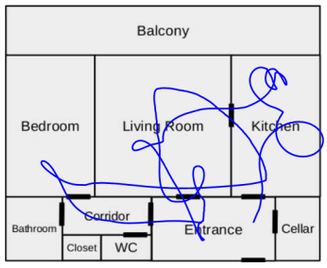
The proposed VIL framework is evaluated experimentally on several challenging video sequences where the person wearing the camera performs daily living activities.

	GO80	GO81	GO82	GO83	GO84	GO85
CBIR only (similar to [2])	1.7	1.3	2.4	2.2	1.3	2.2
CBIR + LG-RBPS No VM	**0.5**	0.7	1.7	0.9	**<0.5**	1.3
CBIR + LG-RBPS With VM	**0.5**	**<0.5**	**0.8**	**0.7**	**<0.5**	**0.9**

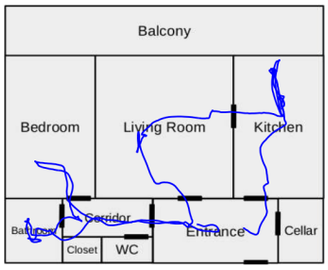
Table 1: Results on 6 challenging video sequences (GO80,..., GO85). Examples of video frames from these videos are presented Fig.1(a). The figures represent the RMSE in meter of the estimated trajectories w.r.t the ground truth which has an accuracy of 0.5m.



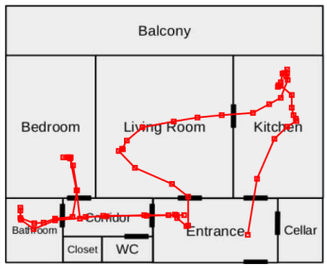
(a) Result of the CBIR algorithm (nearest neighbor) similar to [2]



(b) CBIR+LG-RBPS without virtual measurements



(c) CBIR+LG-RBPS with virtual measurements



(d) Ground truth

Figure 2: Illustration of estimated trajectories on the video sequence GO82. Only the projection of the 3D position onto the apartment plan is presented.

[1] G. Bourmaud, R. Mégret, A. Giremus, and Y. Berthoumieu. Discrete extended Kalman filter on Lie groups. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, 2013.

[2] C. Ó. Conaire, M. Blighe, and N. E. O'connor. Sensecam image localisation using hierarchical SURF trees. In *Advances in Multimedia Modeling*, pages 15–26. Springer, 2009.

Acknowledgments The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007- 2013) under grant agreement 288199 - Dem@Care.

Joint Object-Material Category Segmentation from Audio-Visual Cues

Anurag Arnab, Michael Sapienza, Stuart Golodetz[1]
{anurag.arnab,michael.sapienza,stuart.golodetz}@eng.ox.ac.uk
Julien Valentin, Ondrej Miksik, Philip H. S. Torr[1]
{julien.valentin,ondrej.miksik,philip.torr}@eng.ox.ac.uk
Shahram Izadi[2]
shahrami@microsoft.com

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] Microsoft Research
Redmond, US

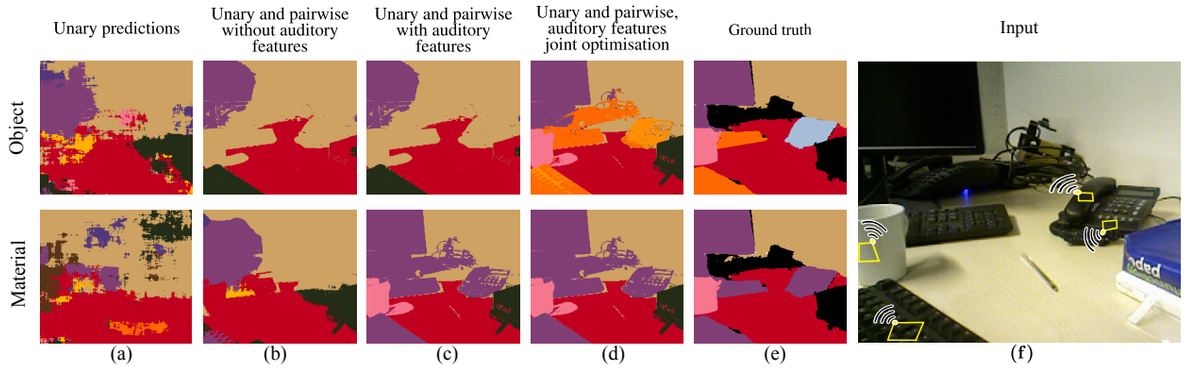


Figure 1: **(a)** The noisy predictions made by the per-pixel unary classifiers. **(b)** The output of the CRF using only visual features. **(c)** The use of auditory information improves material labeling. **(d)** Finally, joint optimisation between object and meterial categories improves object labelling as well. **(e)** The ground truth. **(f)** The input image, showing the locations where sound information is present.

It is not always possible to recognise objects and infer material properties for a scene from visual cues alone, since objects can look visually similar whilst being made of very different materials. In this paper, we therefore present an approach that augments the available dense visual cues with sparse auditory cues in order to estimate dense object and material labels. Since estimates of object class and material properties are mutually-informative, we optimise our multi-output labelling jointly using a random-field framework. We evaluate our system on a new dataset with paired visual and auditory data that we make publicly available. We demonstrate that this joint estimation of object and material labels significantly outperforms the estimation of either category in isolation.

By using sound, we are able to infer information about an object's material properties that would be difficult or impossible to obtain by visual means. This is evident from Figure 1 where we can see that the table, mug and wall all have similar local colour and texture, even though the table is made from wood, the wall from gypsum and the mug from ceramic. This leads to various object and material class labels being incorrect (Figure 1b). However, when we tap various objects in the scene and incorporate the resulting auditory information into our segmentation process, our predicted material label significantly improve (Figure 1c). We can then use these predictions to improve our object class predictions as well (Figure 1d).

Existing segmentation datasets do not provide audio-visual annotations as ground truth. Furthermore, it is not possible to simply augment them audio data, since we would need the original objects in the dataset to extract sound. As a result, we create our own dataset which we make publicly available[1]. In contrast to previous segmentation datasets, we annotate ours in 3D. We captured 9 different long sequences using a consumer-grade depth camera (ASUS Xtion Pro) and then reconstructed the 3D scene using the system of [4, 5]. This reconstructed scene was then annotated in 3D using an interactive scene segmentation framework [2]. This method allowed us to significantly decrease the annotation time since a typical sequence of 2000 frames could be fully annotated in about 45 minutes, which is far less than the 20-25 minutes per frame required to label each frame of the CamVid dataset by hand [1].

We captured our sound data using a portable condensor microphone (Samson GoMic). Due to the localised nature of sound, we can only associate sound data with the points at which the object was struck. This was done by annotating the approximate location at which the object was struck in the 3D reconstruction.

[1] http://www.robots.ox.ac.uk/~tvg/projects/AudioVisual/

Since auditory information obtained by tapping objects is only available at sparse locations in an image, we need a method of propagating this information to the whole image. To this end, since estimates of object and material properties can be mutually informative, we use a two-layer CRF to model the joint estimation of object and material labels, and allow the two types of estimate to influence each other by connecting the two layers of the CRF with joint potentials. We minimise the energy,

$$E(\mathbf{x}|\mathbf{D}) = E^O(\mathbf{o}|\mathbf{I}) + E^M(\mathbf{m}|\mathbf{I},\mathbf{A}) + E^J(\mathbf{o},\mathbf{m}|\mathbf{I},\mathbf{A}), \quad (1)$$

where $\mathbf{x}$ is an assignment to the random variable $\mathbf{X}$ that takes a label $[o,m]$ from the product label space of object and material labels, $\mathcal{O}\times\mathcal{M}$. The energy is conditioned on the visual and auditory data $\mathbf{D}=\{\mathbf{I},\mathbf{A}\}$. $E^O(\mathbf{o}|\mathbf{I})$ is the energy for the object labelling, conditioned on image data $\mathbf{I}$, $E^M(\mathbf{m}|\mathbf{I},\mathbf{A})$ is the energy for the material labelling, conditioned on image data $\mathbf{I}$ and audio data $\mathbf{A}$, and $E^J(\mathbf{o},\mathbf{m}|\mathbf{I},\mathbf{A})$ is the joint energy function linking the object and material domains.

The final joint energy function takes correlations between objects and materials into account, and encourages consistency between the two label categories. The joint potentials were learnt from the conditional distributions of the two labels in the training set. The first two energy functions consist of unary and pairwise potentials. The per-pixel unary potentials are obtained from a joint boosting classifier whilst the pairwise potentials takes the form of a mixture of Gaussian kernels to facilitate efficient mean-field inference [3, 6].

Using auditory data, the mean intersection-over-union (IoU) for material classification improves by 3.5% over the baseline which used only visual information. By employing joint optimisation between object and material classes, a further 4.1% improvement was obtained for object classification.

[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009.

[2] Stuart Golodetz, Michael Sapienza, Julien P C Valentin, Vibhav Vineet, Ming-Ming Cheng, Victor A Prisacariu, Olaf Kähler, Carl Yuheng Ren, Anurag Arnab, Stephen L Hicks, David W Murray, Shahram Izadi, and Philip H S Torr. SemanticPaint: Interactive Segmentation and Learning of 3D Worlds. Demo in SIGGRAPH ET, 2015.

[3] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.

[4] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 2013.

[5] V. A. Prisacariu, O. Kähler, M. M. Cheng, C. Y. Ren, J. Valentin, P. H. S. Torr, I. D. Reid, and D. W. Murray. A Framework for the Volumetric Integration of Depth Images. *ArXiv e-prints*, 2014.

[6] Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *IJCV*, 2014.

Deep Face Recognition

Omkar M. Parkhi
omkar@robots.ox.ac.uk
Andrea Vedaldi
vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

The goal of this paper is face recognition – from either a single photograph or from a set of faces tracked in a video. Recent progress in this area has been due to two factors: (i) end to end learning for the task using convolutional neural networks (CNNs), and (ii) the availability of very large scale training datasets. We make two contributions: first, we show how a very large scale dataset (2.6M images spanning more than 2.6K identities) can be constructed by semi-automatic annotations with humans in the loop, investigating the trade-off between annotation purity and cost; second, we introduce a very deep convolutional neural network and a corresponding training procedure that achieve face recognition accuracy comparable to the current state of the art on public benchmarks such as "Labelled Faces In the Wild" and "YouTube Faces Dataset", while at the same time using a fraction of the data used by competitors.



Figure 1: Example images from our dataset.

The availability of large scale datasets such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [4], MIT places [10] and Microsoft COCOA [2] have been instrumental in the giant CNN based leapforward we have witnessed in the community. Such datasets however, were missing from the face recognition domain. Most recent advancement in the field have been from the internet giants like Facebook and Google [5, 8, 9]. For example, the most recent face recognition method by Google [5] was trained using 200 million images and eight million unique identities. The size of this dataset is almost *three orders of magnitude* larger than any publicly available face dataset (see Table 1). The first part of this paper proposes a procedure to create a reasonably large face dataset whilst requiring only a limited amount of person-power for annotation. One of the key ideas was to use weaker classifiers to rank the data presented to the annotators. This procedure has been developed for faces, but is evidently suitable for other object classes as well as fine grained tasks. We employ this procedure to build a dataset with over two million faces, and will make this freely available to the research community.

Dataset	People	Images
LFW	5,749	13,233
WDRef [1]	2,995	99,773
CelebFaces [7]	10,177	202,599
Ours	**2,622**	**2.6M**
FaceBook [8]	4,030	4.4M
Google [5]	8M	200M

Table 1: **Dataset comparisons:** Our dataset has the largest collection of face images outside industrial datasets by Goole, Facebook, or Baidu, which are not publicly available.

The second part of this paper investigates various CNN architectures for face identification and verification, including exploring face alignment and learning the task specific embeddings, using the novel dataset for training. Many recent works on face recognition have proposed numerous variants of CNN architectures for faces, and we assess some of these

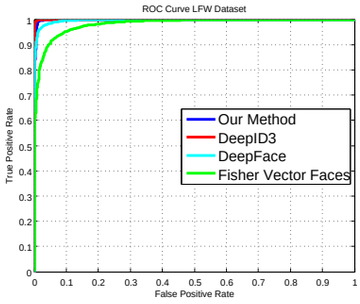


Figure 2: ROC curve for the LFW dataset unrestricted protocol setting.

modeling choices in order to filter what is important from irrelevant details . The outcome is a much simpler and yet effective network architecture without any embellishments but with appropriate training, achieving near state-of-the-art results on all popular image and video face recognition benchmarks using a single network (see Table 2 and 3). Again, this is a conclusion that may be applicable to many other tasks.

No.	Method	Images	Networks	Acc.
1	Fisher Vector Faces [6]	-	-	93.10
2	DeepFace [8]	4M	3	97.35
3	Fusion [9]	500M	5	98.37
4	DeepID-2,3		200	99.47
5	FaceNet [5]	200M	1	98.87
6	FaceNet [5] + Alignment	200M	1	99.63
7	Ours	2.6M	1	98.95

Table 2: **LFW unrestricted setting.** Left: we achieve comparable results to the state of the art whilst requiring less data (than DeepFace and FaceNet) and using a simpler network architecture (than DeepID-2,3). Note, DeepID3 results are for the test set with label errors corrected – which has not been done by any other method.

No.	Method	Images	Networks	100%- EER	Acc.
1	Video Fisher Vector Faces [3]	-	-	87.7	83.8
2	DeepFace [8]	4M	1	91.4	91.4
3	DeepID-2,2+,3		200	-	93.2
4	FaceNet [5] + Alignment	200M	1	-	95.1
5	Ours	2.6M	1	92.8	91.6
6	Ours + Embedding learning	2.6M	1	97.4	97.3

Table 3: **Results on the Youtube Faces Dataset, unrestricted setting**. The value of k indicates the number of faces used to represent each video.

[1] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. ECCV*, pages 566–579, 2012. 1

[2] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1

[3] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *Proc. CVPR*, 2014. 1

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, S. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and F.F. Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[5] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, 2015. 1

[6] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE PAMI*, 2014. 1

[7] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014. 1

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deep-Face: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014. 1

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *Proc. CVPR*, 2015. 1

[10] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. *NIPS*, 2014. 1

Combinatorial Regularization of Descriptor Matching for Optical Flow Estimation

Benjamin Drayer
drayer@cs.uni-freiburg.de
Thomas Brox
brox@cs.uni-freiburg.de

Department of Computer Science,
Centre of Biological Signalling Studies (BIOSS),
University of Freiburg, Germany

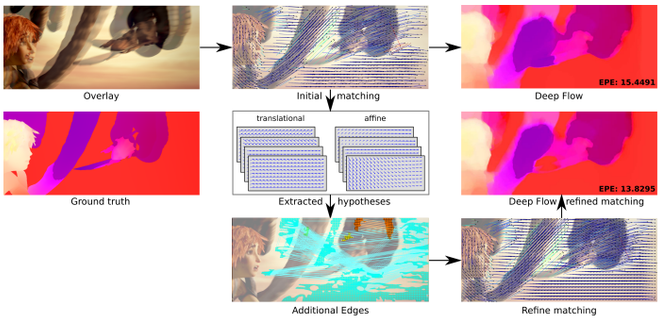


Figure 1: Outline of our method. The first row shows the overlaid input frames, the matching step (red arrows=high error, blue=low error) and the estimated optical flow using Deep Flow [5]. Starting from the initial matching, we extract translational and affine hypotheses. The 4-connected grid in the optimization/refinement step is enhanced by edges connecting similar regions. Our refinement improves at several position (note the fewer erroneous flow vectors and the denser field).



Figure 2: Benefit of additional edges. When matching the first frame to the second (b), we compute regions with little texture, highlighted as black and colored areas in (c). Noise and small regions are discarded (marked as black). The remaining regions are connected if their appearance is similar. Connected regions have the same color. The bottom row shows the improvement from the initial LDOF [1] correspondences (d), over the refined matches (e) to and the refinement with the extended set of edges (f). The arrows are colored from blue to red (small to high error).

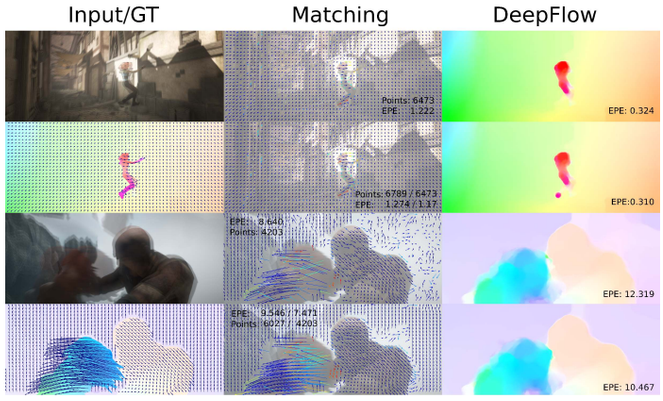


Figure 3: The upper line of each example shows the overlay of the input frames, the initial matching and the results of DeepFlow. The bottom line contains, groundtruth, our refined matching and the respective optical flow.

Introduction Dealing with large displacements has been the main research focus in the field of optical flow estimation in recent years [1, 5]. To estimate the large motion of small, detailed structures, a popular strategy is to pre-compute correspondences based on discriminative descriptors, such as HOG. While [1] relies on simple nearest neighbor matching to establish such correspondences, subsequent works have proposed more involved strategies to avoid wrong correspondences. The difference in performance between LDOF [1] and DeepFlow [5] is due to a more involved hierarchical matching strategy.

Refinement The contribution of the present paper is a refinement algorithm that further improves a given set of correspondences in several ways: (1) The refinement is computed densely (on a grid), which increases the number of matches. (2) Due to the explicit regularization, ambiguities in homogeneous regions or areas with repetitive texture are resolved. (3) Initially wrongly estimated outliers can be corrected. See Figure 1 for an outline of the algorithm.

We cast the refinement problem as a combinatorial energy minimization problem. The energy consists of three terms:

$$E(\mathcal{L}) = \underbrace{E_A(\mathcal{L}) + E_M(\mathcal{L})}_{\text{Data terms}} + \underbrace{E_S(\mathcal{L})}_{\text{Smoothness}} \quad (1)$$

The task is to find a labeling $\mathcal{L}$ that yields the minimal energy. The individual labels are affine motion hypotheses, derived with an alternating optimization scheme from the initial matching. For each hypothesis we assign a matching cost E_M on how similar the displacement is to the initial guess (if there is one) and an appearance term E_A that measures the similarity of the matched features. We regularize in the space of affine motion with the so-called flow continuity term [3]. While it is well-established to use a 4- or 8-connected neighborhood for discrete regularization, we propose to extend the set of edges by connecting homogeneous regions which likely belong together. In this way, the matching in enclosed homogeneous areas without sufficient characteristics is improved. Figure 2 gives an overview of the extended regularization.

Results The proposed refinement is a self-contained procedure and can be employed in any matching based optical flow algorithm. We show the improvement of our matches on the challenging Sintel dataset [2], where we evaluate the proposed refinement in the framework of LDOF, DeepFlow and EpicFlow (see Table 1). Qualitative results are shown in Figure 3.

[1] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012.

[3] Y. Jiaolong and L. Hongdong. Dense, accurate optical flow estimation with piecewise parametric model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1385–1392. IEEE, 2013.

LDOF [1]	LDOF+R	Deep [5]	Deep+R	Epic [4]	Epic+R
6.026	**5.616**	4.022	**3.852**	3.566	**3.497**

Table 1: Comparison of the different optical flow methods with (+R) and without our regularization on the final pass of the Sintel training dataset. Throughout the methods, we get better results with the refined matches.

Discrete Light Source Estimation from Light Probes for Photorealistic Rendering

Farshad Einabadi
einabadi@intel-vci.uni-saarland.de
Intel Visual Computing Institute
Saarbrücken, Germany
Oliver Grau
oliver.grau@intel.com
Intel Corporation

Applications like rendering of images using computer graphics methods are usually requiring more sophisticated light models to give better control. Complex scenes in computer generated images are requiring very differentiated light models to give a realistic rendering of the scene. That usually includes a high number of (virtual) light sources to model a scene to reproduce accurate shadows and shadings. In particular in the production of visual effects for movies and TV the real scene lighting needs to be captured very accurately to give a realistic rendering of virtual objects into that scene. In this context the light modeling is usually done manually by skilled artists in a time consuming process.

This contribution describes a new technique for estimation of discrete spot light sources. The method uses a consumer grade DSLR camera equipped with a fisheye lens to capture light probe images registered to the scene. From these probe images the geometric and radiometric properties of the dominant light sources in the scene are estimated. The first step is a robust approach to identify light sources in the light probes and to find exact positions by triangulation. Then the light direction and radiometric fall-off properties are formulated and estimated in a least square minimization approach.

There are a number of advantages in our approach. First, the probing camera is registered using a multi-camera setup which requires the minimum amendments to the studio. Second, we are not limited to any specific probing object since the properties of each light are estimated based on processing the probe images. In addition, since the probing camera can move freely in the area of interest, there are no limits in terms of the covered space. Large field of view of the fisheye lens is also beneficial in this matter.

Calibration and Registration of Cameras. We propose a two-step calibration and registration approach. In the first step, a planar asymmetric calibration pattern is used for simultaneous calibration of the intrinsics and the pose of all the witness cameras and the principal camera using a bundle adjustment module. In the next step, parameters of witness cameras are kept fixed and the probing camera is registered in the same coordinate system by using color features of an attached calibration rig.

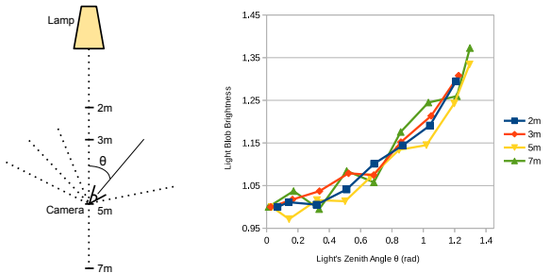
Position Estimation. To estimate the 3D position vectors of the light sources, one needs to shoot rays from every detected light blob in all probe images and triangulate the corresponding rays from at least two probe positions for each source. Figure 1 summarizes the required steps.

Inputs:
A couple of registered light probe images
Intrinsic calibration parameters of the probing camera
Output:
3D position vectors of light sources
Steps:
Detect light blobs in all probe images
Match light blobs to their corresponding light source
For all detected light sources
Shoot rays from corresponding light blobs
Triangulate computed rays
Return the estimated 3D position of the light source

Figure 1: Pseudocode for estimating positions of the light sources

Dominant Direction and Fall-off Curve Estimation Observed intensity of a light source at each probe position is considered to be related to the brightness of its corresponding blob in the thresholded probe image. The brightness of each blob is also intuitively defined as summation over its pixel values. However, the probing device must be radiometrically calibrated to find this relation and count for all the present effects of the camera-lens system (Figure 2).

After this calibration, for a light source j with 3D position L_j, the goal is to find its dominant direction, $\overrightarrow{l_j}$, and parameters of its intensity fall-off



(a) (b)

Figure 2: Radiometric calibration of the probing camera-lens system: (a) sampling process (b) distance-normalized blob brightness values w.r.t. zenith angle θ of the lamp

function, $I_j(\phi)$, that minimize

$$\sum_{i=0}^{m}\left(I_j(\phi_{ij})-\frac{B_{ij}\cdot\|P_i-L_j\|^2}{C(\theta)}\right)^2 \quad (1)$$

with m the number of probes where the light source j is visible in them and P_is their corresponding 3D positions. B_{ij} is brightness of light blob j in probe i and the terms $\|P_i-L_j\|^2$ and $C(\theta)=\alpha\theta^2+1$ compensate for its quadratic attenuation w.r.t. probing distance and its brightness factor based on its place on camera's sensor plane, respectively. The angle ϕ_{ij} is also defined as $\phi_{ij}=\angle(\overrightarrow{l_j},P_i-L_j)$.

In our formulation we assume a quadratic intensity fall-off for spot lights $I_j(\phi_{ij})=k_1\phi_{ij}^2+k_2$ where the maximum intensity, k_2, happens at the dominant direction $\phi=0$ and $-k_2/k_1$ is the beam angle. The chosen I_j can model the coarse fall-off characteristics of a real studio spot lamp with a limited number of probe observations.

Experiments. We verify our proposed approach on a real dataset recorded in our studio of a scene illuminated with two tungsten halogen lamps. Table 1 shows the quantitative estimation results. Figure 3 shows renders of a reference fish 3D model side by side to its real appearance at two different positions using these estimations.

#		Position L (m)	Direction $\overrightarrow{l}$	Intensity k_2	Beam Angle $-k_2/k_1$
1	Estimated	(2.76, 3.75, 1.95)	(-0.8, -0.35, -0.46)	2.75	56.22°
	Measured	(-, -, 1.84)	(-0.80, -0.41, -0.43)	4.0e3 lux	58.0°
	Error	(-, -, 0.11)	3.86°	-	1.78°
2	Estimated	(-1.11, 1.15, 1.78)	(0.47, 0.52, -0.71)	8.98	50.50°
	Measured	(-, -, 1.73)	(0.49, 0.51, -0.71)	1.59e4 lux	50.3°
	Error	(-, -, 0.04)	0.88°	-	0.2°

Table 1: Estimation results for the lamps in our dataset assuming a quadratic fall-off curve

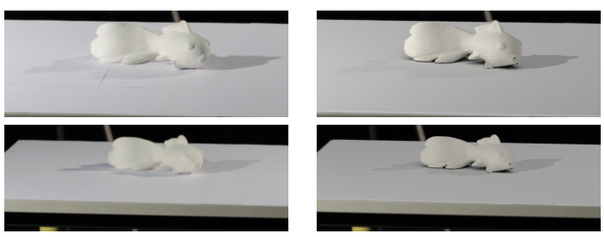


Figure 3: Visual comparison of renders of the reference object (right) to the real ones (left)

Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification

Elyor Kodirov
e.kodirov@qmul.ac.uk
Tao Xiang
t.xiang@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

School of Electronic Engineering and
Computer Science,
Queen Mary University of London
London E1 4NS, UK

1 Introduction

Person re-identification (Re-ID) is the problem of matching people across non-overlapping camera views. Despite the best efforts from the computer vision researchers, it remains an unsolved problem. This is because a person's appearance often changes dramatically across camera views due to changes in pose, occlusion, lighting, and illumination conditions.

Many existing approaches to person re-identification (Re-ID) are based on supervised learning [1, 2, 3, 4] which requires hundreds of matching pairs to be labelled for each pair of cameras. This severely limits their scalability for real-world applications. This work aims to overcome this limitation by developing a novel unsupervised Re-ID approach. The approach is based on a new dictionary learning for sparse coding formulation with a graph Laplacian regularisation term whose value is set iteratively. As an unsupervised model, the dictionary learning model is well-suited to the unsupervised task, whilst the regularisation term enables the exploitation of cross-view identity-discriminative information ignored by existing unsupervised Re-ID methods. Importantly this model is also flexible in utilising any labelled data if available. Experiments on VIPeR and PRID benchmark datasets demonstrate that the proposed approach significantly outperforms the state-of-the-art.

2 Methodology

Problem Definition. Suppose a set of training person images are collected from a pair of camera views denoted as A and B respectively. An n-dimensional feature vector is computed from each person's image to represent ones appearance. Let's denote the training data matrix as $X = [X^a,\ X^b] \in \mathbb{R}^{n\times m}$ where $X^a = [x^a_1,\ \ldots\ ,x^a_{m_1}] \in \mathbb{R}^{n\times m_1}$ contains the feature vectors of m_1 images in view A as columns, while $X^b = [x^b_1,\ \ldots\ ,x^b_{m_2}] \in \mathbb{R}^{n\times m_2}$ does the same for the m_2 images in view B. We thus have $m = m_1 + m_2$. Note, the training data are *unlabelled* therefore it is unknown which person observed in view A corresponds to a given person in view B and vice versa. The objective of unsupervised person Re-ID is to learn a matching function f from X, so that given x^a and x^b representing two test person images from A and B respectively, $f(x^a,x^b)$ can be used for matching their identities.

Dictionary Learning with Graph Laplacian Regularisation. Our solution to the problem defined above is to learn a shared dictionary $D \in \mathbb{R}^{k\times m}$ for the two camera views using X. With this dictionary, each n-dimensional feature vector, regardless which view it comes from, is projected into a lower k-dimensional subspace spanned by the k dictionary atoms (columns of D) so that they can be matched by the cosine distance in this subspace. The underpinning idea is that each atom or the dimension of the subspace corresponds to a latent appearance attribute which is invariant to the camera view changes, thus useful for cross-view matching. Formally, we aim to learn the optimal dictionary D, such that the sparse code of X, denoted as $Y = [Y^a,\ Y^b] \in \mathbb{R}^{k\times m}$, where $Y^a = [y^a_1,\ \ldots\ ,y^a_{m_1}] \in \mathbb{R}^{k\times m_1}$ and $Y^b = [y^b_1,\ \ldots\ ,y^b_{m_2}] \in \mathbb{R}^{k\times m_2}$, can be used for matching the training data; and we wish the same D can be generalised to match unseen test image pairs from the two views.

Using a conventional dictionary learning formulation, D and Y can be estimated as:

$$(D^*,Y^*) = \underset{D,Y}{\operatorname{argmin}} \|X - DY\|_F^2 + \alpha||Y||_1 \qquad (1)$$

where $\|X - DY\|_F^2$ is the reconstruction error term evaluating how well a linear combination of the learned atoms can approximate the input data, and $||.||_F$ denotes the matrix Frobenious norm; $||Y||_1$ is the sparsity term favouring small number of atoms to be used for reconstruction; this term is weighted by α. It is clear from this formulation that the conventional dictionary learning model only cares about how to best reconstruct X using D and Y, without giving any consideration to whether the sparse code is meaningful for matching people cross camera views. In order to make the learned dictionary discriminative for cross-view matching, one must exploit cross-view identity discriminative information. With cross-view labels, this can be achieved by forcing the two matched images to have identical sparse codes [5]. However, without any labels available under our unsupervised setting, it is not possible to use this conventional formulation for person Re-ID.

To overcome this problem, we introduce a graph Laplacian regularisation term in the dictionary learning formulation, and rewrite Eq. (1) as

$$(D^*,Y^*) = \underset{D,Y}{\operatorname{argmin}} \|X - DY\|_F^2 + \alpha||Y||_1 + \beta\sum_{i,j}^{m} ||y^a_i - y^b_j||_2^2 W_{ij} \qquad (2)$$

where β is the weight of the new regularisation term, and $W \in \mathbb{R}^{m\times m}$ is a cross-view correspondence matrix capturing the identity relationship between the people in X^a and X^b which needs to be preserved after they are projected and become Y^a and Y^b. Note, since the training data are unlabelled, the true cross-view correspondence relationship is unknown. We therefore use W to represent a soft cross-view correspondence relationship. That is, each person in A can correspond to multiple people in B depending on their visual similarity.

3 Experimental Results

Experimental results on VIPeR and PRID benchmark datasets show that our model significantly outperforms the state of the art.

Table 1: Unsupervised Re-ID results on VIPeR and PRID

Dataset	VIPeR				PRID			
Ranks	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
eSDC	26.7	50.7	62.4	76.4	-	-	-	-
SDALF	19.9	38.9	49.4	65.7	16.3	29.6	38.0	48.7
ISR	27.0	49.8	61.2	73.0	17.0	34.4	42.0	54.3
CPS	22.0	44.7	57.0	71.0	-	-	-	-
GTS	25.2	50.0	62.5	75.8	-	-	-	-
Ours	**29.6**	**54.8**	**64.8**	**77.3**	**21.1**	**43.7**	**55.8**	**64.8**

Table 2: Semi-supervised Re-ID results on VIPeR and PRID

Dataset	VIPeR				PRID			
Ranks	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
RankSVM	20.7	41.8	54.6	68.1	-	-	-	-
KISSME	18.5	43.7	57.9	74.5	5.1	15.2	24.1	40.1
kLFDA	27.5	56.0	69.6	82.6	14.1	33.7	44.0	56.2
KCCA	24.6	56.2	71.7	**85.6**	5.3	15.7	25.8	37.0
MFA	25.3	53.6	66.7	78.8	13.3	32.5	43.3	56.4
SSCDL	25.6	53.7	68.2	83.6	-	-	-	-
Ours	**32.5**	**61.8**	**74.3**	84.1	**22.1**	**45.3**	**56.5**	**66.3**

[1] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Proc. ACCV*, 2011.

[2] L. Giuseppe, M. Iacopo, and D. B. Alberto. Matching people across camera views using kernel canonical correlation analysis. In *Proc. ICDSC*, 2014.

[3] M. Hirzer, M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *Proc. AVSS*, 2012.

[4] M. Hirzer, M. Roth, M. Koestinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proc. ECCV*, 2012.

[5] X. Liu, M. Song, D. Tao, X. Zhou, Ch. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proc. CVPR*, 2014.

Camera Pose Estimation from Lines using Plücker Coordinates

Bronislav Přibyl
ipribyl@fit.vutbr.cz
Pavel Zemčík
zemcik@fit.vutbr.cz
Martin Čadík
cadik@fit.vutbr.cz

Graph@FIT, CPhoto@FIT
Department of Computer Graphics and Multimedia
Faculty of Information Technology
Brno University of Technology
Božetěchova 2
612 66 Brno
Czech Republic

Camera pose estimation is the task of determining the position and orientation of a camera in 3D space, based on correspondences between 3D features and their 2D images. When these features are *lines*, the task is called the *Perspective-n-Line* (PnL) problem. A remarkable progress in solving PnL has been achieved in the last years, particularly thanks to the work of Mirzaei and Roumeliotis [5] and more recently to the work of Zhang *et al*. [6]. Both of the methods are accurate, cope well with noisy data, and they are more efficient than the previously known methods.

In this paper, we propose an algebraic solution to the PnL problem which **(i)** is more than the order of magnitude faster than the state-of-the-art [5, 6], **(ii)** yields only one solution of the PnL problem, and **(iii)** similarly to the state-of-the-art, copes well with image noise and is initialization-free. As an alternative to the commonly used RANSAC, Algebraic Outlier Rejection scheme [1] is used to deal with mismatched line correspondences. The proposed method requires at least 9 lines, but it is particularly suitable for large scale and noisy scenarios, where it can be reliably used for 25 and more lines.

Given two distinct 3D points $\mathbf{A} = (a_x\ a_y\ a_z\ a_w)^\top$ and $\mathbf{B} = (b_x\ b_y\ b_z\ b_w)^\top$ in homogeneous coordinates, a line joining them can be represented using Plücker coordinates [3] as a homogeneous 6-vector $\mathbf{L} = (\mathbf{u}^\top\ \mathbf{v}^\top)^\top = (L_1\ L_2\ L_3\ L_4\ L_5\ L_6)^\top$, where

$$\begin{aligned} \mathbf{u}^\top &= (L_1\ L_2\ L_3) = (a_x\ a_y\ a_z) \times (b_x\ b_y\ b_z) \\ \mathbf{v}^\top &= (L_4\ L_5\ L_6) = a_w(b_x\ b_y\ b_z) - b_w(a_x\ a_y\ a_z)\ , \end{aligned} \tag{1}$$

'$\times$' denotes a vector cross product. The $\mathbf{v}$ part encodes direction of the line while the $\mathbf{u}$ part encodes position of the line in space. In fact, $\mathbf{u}$ is a normal vector of an interpretation plane – a plane passing through the line and the origin. As a consequence, $\mathbf{L}$ must satisfy a bilinear constraint $\mathbf{u}^\top\mathbf{v} = 0$.

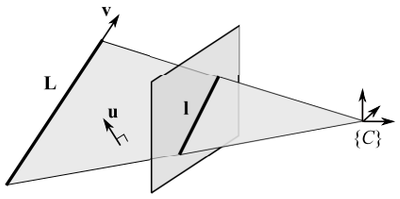


Figure 1: 3D line projection. The 3D line $\mathbf{L}$ is parameterized by its direction vector $\mathbf{v}$ and a normal $\mathbf{u}$ of its interpretation plane, which passes through the origin of the camera coordinate system $\{C\}$. Since the projected 2D line $\mathbf{l}$ lies at the intersection of the interpretation plane and the image plane, it is fully defined just by the normal $\mathbf{u}$.

A 3D line $\mathbf{L}$ is projected onto the normalized image plane using the 3×6 line projection matrix $\mathbf{P}$ as

$$\mathbf{l} \approx \mathbf{P}\mathbf{L}\ , \tag{2}$$

where $\mathbf{l} = (l_x\ l_y\ l_w)^\top$ is a homogeneous 2D line, '$\approx$' denotes an equivalence of homogeneous coordinates and

$$\mathbf{P} = \begin{pmatrix} \mathbf{R} & \mathbf{R}[-\mathbf{t}]_\times \end{pmatrix}\ . \tag{3}$$

$\mathbf{R}$ is a 3×3 rotation matrix and $[\mathbf{t}]_\times$ is a 3×3 skew-symmetric matrix constructed from the translation vector $\mathbf{t}$.

As the projection of 3D lines is defined by Eq. (2), the problem of camera pose estimation resides in estimation of the line projection matrix $\mathbf{P}$, which encodes all the six camera pose parameters t_x, t_y, t_z, α, β, γ. We solve this problem using the Direct Linear Transformation (DLT) algorithm similarly to Hartley [2], who works with points. The system of linear equations (2) can be transformed into a homogeneous system

$$\mathbf{M}\mathbf{p} = \mathbf{0}\ . \tag{4}$$

This forms a $2n \times 18$ measurement matrix $\mathbf{M}$ which contains coefficients of equations generated by correspondences between 3D lines and their projections $\mathbf{L}_i \leftrightarrow \mathbf{l}_i$ $(i = 1 \ldots n,\ n \geq 9)$. The DLT then solves (4) for $\mathbf{p}$ which is a 18-vector containing the entries of the line projection matrix $\mathbf{P}$. Eq. (4), however, holds only in the noise-free case. If a noise is present in the measurements, an inconsistent system is obtained.

$$\mathbf{M}\hat{\mathbf{p}} = \boldsymbol{\varepsilon} \tag{5}$$

Only an approximate solution $\hat{\mathbf{p}}$ may be found minimizing a $2n$-vector of measurement residuals $\boldsymbol{\varepsilon}$. Since DLT algorithm is sensitive to the choice of coordinate system, it is crucial to prenormalize the data to get properly conditioned $\mathbf{M}$.

Once the system of linear equations given by (5) is solved in the least squares sense, *e.g.* by SVD of $\mathbf{M}$, the estimate $\hat{\mathbf{P}}$ of the 3×6 line projection matrix can be recovered from the 18-vector $\hat{\mathbf{p}}$. $\hat{\mathbf{P}}$ is obtained as a least squares solution of Eq. (5) so it does not satisfy the constraints imposed on $\mathbf{P}$ by Eq. (3). We propose a method to extract the camera pose parameters from the estimate $\hat{\mathbf{P}}$, which is based on an observation that the right 3×3 submatrix of $\mathbf{P}$ has exactly the same structure as the essential matrix [4] used in multi-view computer vision. Details of the method are described in the full paper.

Simulations and experiments with real data show that computational efficiency of the proposed method is superior to the state-of-the-art, achieving speed-ups of more than one order of magnitude for high number of lines. At the same time, accuracy and robustness of the method are on par with the state-of-the-art.

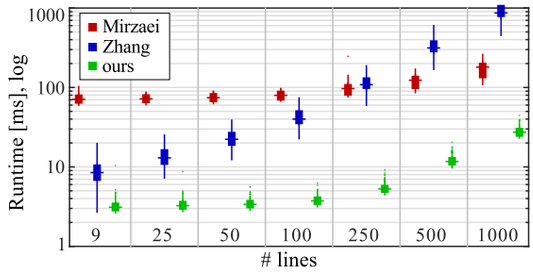


Figure 2: The distribution of runtimes as a function of the number of lines.

[1] L. Ferraz, X. Binefa, and F. Moreno-Noguer. Very fast solution to the pnp problem with algebraic outlier rejection. In *IEEE Conference on Computer Vision and Pattern Recognition 2014*, pages 501–508. IEEE, 2014.

[2] R. I. Hartley. Minimizing algebraic error. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 356(1740):1175–1192, 1998.

[3] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. ISBN 0521540518.

[4] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[5] F. M. Mirzaei and S. I. Roumeliotis. Globally optimal pose estimation from line correspondences. In *IEEE International Conference on Robotics and Automation 2011*, pages 5581–5588. IEEE, 2011.

[6] L. Zhang, C. Xu, K.-M. Lee, and R. Koch. Robust and efficient pose estimation from line correspondences. In *Asian Conference on Computer Vision 2012*, pages 217–230. Springer, 2013.

Linear Global Translation Estimation with Feature Tracks

Zhaopeng Cui[1]
zhpcui@gmail.com
Nianjuan Jiang[2]
nianjuan.jiang@adsc.com.sg
Chengzhou Tang[1]
chengzhout@gmail.com
Ping Tan[1]
pingtan@sfu.ca

[1] GrUVi Lab
Simon Fraser University
Burnaby, Canada

[2] Advanced Digital Sciences Center of Illinois
Singapore

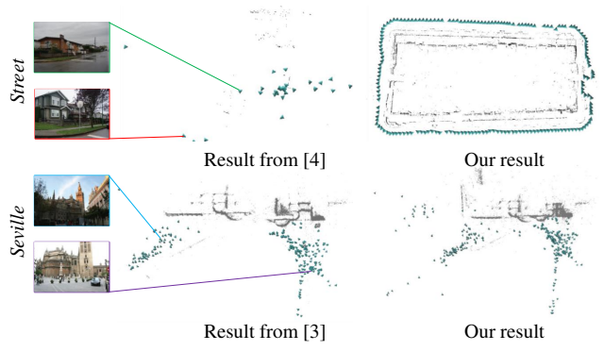


Figure 1: 1DSFM [4] and triplet-based methods (*e.g.* [3]) require strong association among images. As shown in the left, they fail for images with weak association. In comparison, as shown in the right, the results from our method do not suffer from such problems.

Global structure-from-motion (SfM) algorithms register all cameras simultaneously, which are potentially more efficient and less prone to drifting than incremental SfM methods. Global SfM methods often solve the camera orientations and positions separately. This paper focuses on the problem of global position (*i.e.* translation) estimation.

Essential matrix based global translation estimation methods (*e.g.* [1]) usually degenerate at collinear camera motion because the translation scale is not determined by an essential matrix. Trifocal tensor based methods (*e.g.* [3]) usually rely on a strongly connected camera-triplet graph, where two triplets are connected by their common edge. The 3D reconstruction will distort or break into disconnected components when such strong association among images does not exist. The recent 1DSfM method [4] designs a smart filter to discard outlier essential matrices and solves scene points and cameras together by enforcing orientation consistency. However, this method requires abundant association between input images, *e.g.* $\sim O(n^2)$ essential matrices for n cameras, which is more suitable for Internet images and often fails on sequentially captured data.

The data association problem of [4] and [3] is exemplified in Figure 1. The *Street* example on the top is a sequential data where each image is only matched upto 4 neighbors. 1DSfM fails on this example due to insufficient image association. In the *Seville* example on the bottom, those Internet images are mostly captured from two viewpoints (see the two representative sample images) with weak affinity between images at different viewpoints. This weak data association causes seriously distorted reconstruction for the triplet-based method in [3].

This paper introduces a direct linear algorithm to address the presented challenges. It avoids degeneracy at collinear motion and deals with weakly associated data. Our method capitalizes on constraints from essential matrices and feature tracks. As shown in Figure 2 (a), the location of a scene point $\mathbf{p}$ can be computed as the middle point of the mutual perpendicular line segment AB of the two rays passing through $\mathbf{p}$'s image projections:

$$\mathbf{p} = \frac{1}{2}(A+B) = \frac{1}{2}(\mathbf{c}_i + s_i\mathbf{m}_i + \mathbf{c}_j + s_j\mathbf{m}_j). \quad (1)$$

Here, $\mathbf{c}_i$ and $\mathbf{c}_j$ are the two camera centers. The two unit vectors $\mathbf{m}_i$ and $\mathbf{m}_j$ origin from the camera centers and point toward the image projections of $\mathbf{p}$. s_i and s_j are the distances from the points A, B to $\mathbf{c}_i, \mathbf{c}_j$ respectively.

We use the rotation trick in [3] to compute $\mathbf{m}_i$ and $\mathbf{m}_j$ by rotating the relative translation direction $\mathbf{c}_{ij}$ between $\mathbf{c}_i$ and $\mathbf{c}_j$, *i.e.* $\mathbf{m}_i = \mathbf{R}(\theta_i)\mathbf{c}_{ij}$ and $\mathbf{m}_j = -\mathbf{R}(\theta_j)\mathbf{c}_{ij}$. Then Equation 1 becomes

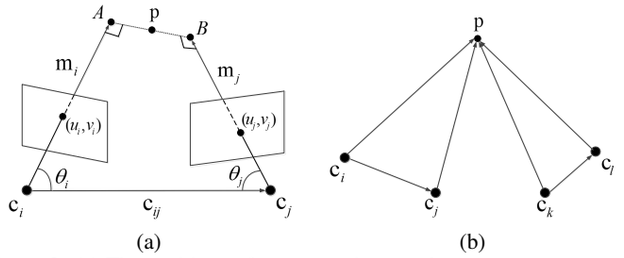


Figure 2: (a) The positions of a scene point $\mathbf{p}$ and two camera centers $\mathbf{c}_i$ and $\mathbf{c}_j$ satisfy a linear constraint. (b) The positions of four cameras seeing the same scene point satisfy a linear constraint.

$$\mathbf{p} = \frac{1}{2}\left(\mathbf{c}_i + s_i\mathbf{R}(\theta_i)\frac{\mathbf{c}_j - \mathbf{c}_i}{\|\mathbf{c}_j - \mathbf{c}_i\|} + \mathbf{c}_j + s_j\mathbf{R}(\theta_j)\frac{\mathbf{c}_i - \mathbf{c}_j}{\|\mathbf{c}_i - \mathbf{c}_j\|}\right). \quad (2)$$

The two 3D rotation matrices $\mathbf{R}(\theta_i)$ and $\mathbf{R}(\theta_j)$ rotate the relative translation direction $\mathbf{c}_{ij}$ to the directions $\mathbf{m}_i$ and $\mathbf{m}_j$. Both rotations can be computed easily in the local pairwise reconstruction. In addition, the two ratios $s_i/\|\mathbf{c}_j - \mathbf{c}_i\|$ and $s_j/\|\mathbf{c}_j - \mathbf{c}_i\|$ can be computed by the middle-point algorithm [2]. Thus, Equation 2 is reduced to,

$$\mathbf{p} = \frac{1}{2}\left((\mathbf{A}_j^{ij} - \mathbf{A}_i^{ij})(\mathbf{c}_i - \mathbf{c}_j) + \mathbf{c}_i + \mathbf{c}_j\right) \quad (3)$$

where $\mathbf{A}_i^{ij} = s_i/||\mathbf{c}_j - \mathbf{c}_i||\mathbf{R}(\theta_i)$ and $\mathbf{A}_j^{ij} = s_j/||\mathbf{c}_i - \mathbf{c}_j||\mathbf{R}(\theta_j)$ are known matrices. This equation provides a linear constraint among positions of two camera centers and a scene point.

If the same scene point $\mathbf{p}$ is visible in two image pairs $\mathbf{c}_i, \mathbf{c}_j$ and $\mathbf{c}_k, \mathbf{c}_l$ as shown in Figure 2 (b), we obtain two linear equations about $\mathbf{p}$'s position according to Equation 3. We can eliminate $\mathbf{p}$ from these equations to obtain a linear constraint among four camera centers as the following,

$$(\mathbf{A}_j^{ij} - \mathbf{A}_i^{ij})(\mathbf{c}_i - \mathbf{c}_j) + \mathbf{c}_i + \mathbf{c}_j = (\mathbf{A}_l^{kl} - \mathbf{A}_k^{kl})(\mathbf{c}_k - \mathbf{c}_l) + \mathbf{c}_k + \mathbf{c}_l. \quad (4)$$

Given a set of images, we build feature tracks and collect such linear equations from camera pairs on the same feature track. Solving these equations will provide a linear global solution of camera positions. To resolve the gauge ambiguity, we set the orthocenter of all cameras at origin when solving these equations.

This direct linear method minimizes a geometric error, which is the Euclidean distance between the scene point to its corresponding rays of projection. A key finding in this paper is that, a direct linear solution (without involving scene points) exists by minimizing the point-to-ray error instead of the reprojection error. Since the point-to-ray error approximates the reprojection error well when cameras are calibrated, our method is a good linear initialization for the final nonlinear BA. At the same time, we minimize the L_1 norm when solving the linear equation of camera positions. We derive a linearization of the alternating direction method of multipliers algorithm to address the L_1 optimization problem.

[1] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *Proc. 3DPVT*, 2012.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[3] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *Proc. ICCV*, 2013.

[4] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *Proc. ECCV (3)*, pages 61–75, 2014.

Learning Discriminative Visual N-grams from Mid-level Image Features

Raj Kumar Gupta
gupta-rk@ihpc.a-star.edu.sg
Institute of High Performance Computing (A*STAR)
Singapore

Megha Pandey
pandeym@i2r.a-star.edu.sg
Institute of Infocomm Research (A*STAR)
Singapore

Alex YS Chia
alex.a.chia@rakuten.com
Rakuten Institute of Technology
Singapore

Mid-level image features have been shown to be helpful to bridge the semantic gap between low-level and high-level image representations. Many existing methods to learn mid-level visual elements consider each mid-level feature individually, and do not take their *mutual relationships* into account. We follow the intuitive idea that learning discriminative combinations of visual elements can help us deal with ambiguities better, and propose the concept of visual n-grams to effectively represent combinations of visual elements along with their relative spatial configuration and co-occurrence relationships.

An overview of our approach is shown in Figure 1. Figure 1 (a) shows the process of learning discriminative visual n-grams based on relative spatial position, orientation and co-occurrence relationships of mid-level image patches. Figure 1 (b) further shows how these visual n-grams are used to finally learn a feature vector representing test and training images.

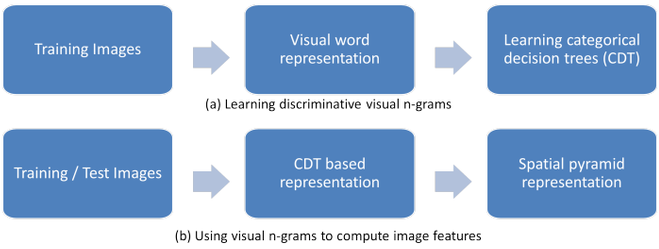


Figure 1: An overview of our approach. (a) Learning discriminative visual n-grams based on relative spatial position, orientation and co-occurrence relationships of mid level patches. (b) Using visual n-grams to compute feature representation for images.

We begin by densely extracting mid-size patches at different scales from the training images. Each patch is represented by a SIFT descriptor. We then learn a codebook by applying standard k-means algorithm. Each patch is then quantized to the nearest codeword representation in SIFT space. It is noteworthy, that while we use SIFT features for all the experiments in this work, our framework is generic and can be used with any other image descriptors as well.

The information about co-occurrence and relative positions and scales of different codewords is implicitly encoded by means of a *spatial co-occurrence vector*. For a given mid-level patch in the training set, we define a neighborhood over nearby grid locations and adjacent scales. The codeword indices of the patches in the neighborhood are concatenated to form the spatial co-occurrence vector for the given patch. Each dimension index of this vector refers to a particular position and scale relative to the current patch, while the value of the corresponding vector element captures the visual appearance of the respective neighboring patch.

Next, our goal is to learn combinations of mid-level elements that can best discriminate one image class from others. We employ categorical decision trees to represent and learn such combinations. Figure 2 shows a toy example for learning a categorical decision tree. For a given codeword c_0, we first locate its occurrences in the training images (Figure 2(a)), and extract spatial co-occurrence vectors for each of these locations. Figure 2(b) shows 4-dimensional vectors for illustrative purposes. In practice, these vectors are extracted over a larger neighborhood and multiple scales. Each of these vectors is weakly labeled with the label of the training image from which it was extracted.

Next, we learn a set of decision rules that can best separate this set of vectors into positive and negative instances, as illustrated in Figure 2(c). This set of decision rules can be encoded in the form of categorical decision tree, which is said to be anchored at the codeword c_0. *Each path from the root node to a leaf node of this tree is a visual n-gram*. In this manner, we can learn a categorical decision tree anchored at each of the codewords in

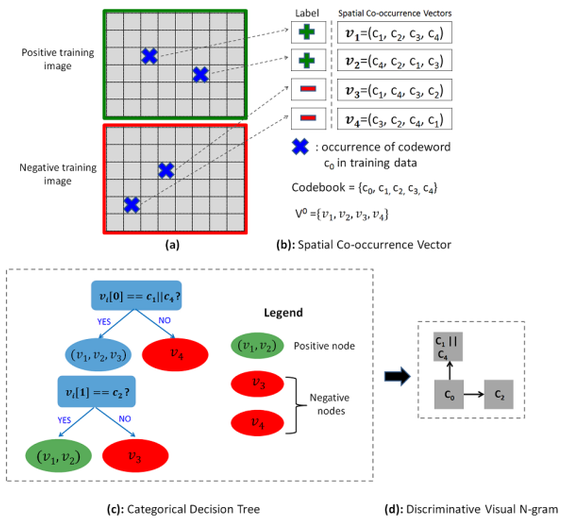


Figure 2: Illustrative example for learning a discriminative visual n-gram. Best viewed in color.

the codebook. We further employ multiple boosting iterations to ensure we effectively capture the diversity in the given set of images, and learn multiple categorical decision trees anchored at each codeword.

Having learnt a series of discriminative visual n-grams in this manner, we can now use them to compute feature representation for a given image. Given an image, we locate all patches represented by a particular codeword c_i, and extract the corresponding set of spatial co-occurrence vectors. These vectors are then classified using the categorical decision trees anchored at c_i, and the classification so obtained is used to compute feature values. Each visual n-gram i.e. each path from the root node to a leaf node in a tree contributes one feature value to the overall image representation. Once the feature values have been computed for all codeword occurrences in the image, we further use spatial pyramid representation to obtain the final image feature vector. The set of feature vectors so obtained is used in a standard SVM framework to perform image classification.

Dataset	Ours	Ours + IFV
Graz-01 (Average EER %)	94.0	95.5
INRIA horses (Classification rate %)	91.76 ±0.33	94.71 ±0.31
UIUC sports (Classification rate %)	83.54 ±0.41	93.12 ±0.28
Land-Use (Classification rate %)	79.52	87.24

Table 1: Classification results on four datasets using our image representation.

We have conducted experiments on four publicly available datasets: Graz-01, INRIA horse images, UIUC 8-sports events and Land-Use dataset. Table 1 lists the classification performance obtained using our feature representation on these datasets. We further combine our representation with global Improved Fisher Vector (IFV) features by concatenation, and include the resulting performance in Table 1 as well. Our method achieves high classification accuracy on each of these datasets. Our features also demonstrate excellent complementarity to global IFV features, in combination with which we outperform the state-of-the-art results on all four datasets.

To conclude, we have proposed an approach to learn discriminative combinations of mid-level elements by exploiting their spatial configuration and co-occurrences. Our method is, by nature, flexible to automatically learn a variety of combinations with different configurations and different number of visual elements. Our experiments demonstrate the effectiveness of the image representation so achieved.

Minimizing the Number of Keypoint Matching Queries for Object Retrieval

Johannes Niedermayer
niedermayer@dbs.ifi.lmu.de
Peer Kröger
kroeger@dbs.ifi.lmu.de

Institute for Informatics
LMU Munich
Munich, Germany

While the development of the SIFT-Descriptor made effective object retrieval on a large scale feasible, its initial use of nearest neighbor queries lead to slow runtimes even on relatively small data sets. These slow runtimes were first compensated by rough quantization using the Bag of Visual Words (BoVW) technique. In recent years, the focus turned back more and more to approximate kNN queries (e.g. [3]) due to their possible gain in matching accuracy [2]: kNN queries provide an accurate ranking of the match candidates and a measure of proximity between database features and query vectors. However, with the vast amount of features that have to be matched during recognition (up to a few thousand), even very fast kNN indexing techniques that can provide approximate query results in under ten milliseconds (e.g. [3]), would yield recognition runtimes of several seconds. We argue that the use of kNN queries for object recognition in large-scale systems cannot be achieved by developing efficient indexing techniques alone. The problem of efficiency has to be approached from additional research directions as well, such as the *number* of kNN queries posed on the system. In this paper, we evaluate an alternative recognition pipeline that ranks query features by assessing their matchability. In order to reduce the number of kNN queries, only the most promising features in this ranking are matched against the database. However, despite gaining efficiency, the enforced reduction of kNN queries causes a reduction of feature matches, decreasing the quality of the query result. While recall can be increased by increasing k, to increase Mean Average Precision (MAP) we expand matches on the image level, aiming at bypassing the indexing step in order to reduce runtimes. This approach stands in contrast to research in the area of BoVW-based retrieval: Research involving the BoVW pipeline assumes that the matching step is relatively cheap, especially if approximate cluster assignment techniques are used. In contrast, this paper aims at maximizing MAP for a small number of processed features, assuming that feature matching is expensive. To summarize, let n denote the upper bound on the number of matching queries, constraining the number of kNN queries. We aim at modifying the standard interest-point based image recognition pipeline such that a given performance measure (in our case MAP) is maximized for a fixed (and low) n. Therefore the problem setting is similar to BoVW-based approaches, however in such a context it is usually assumed that $n = n_{max}$. In this paper we address the opposite case where $n << n_{max}$.

The retrieval pipeline considered in this publication is shown in Figure 1. During feature extraction (Step 1), a set of keypoints and descriptors is extracted from the query. Then (Step 2), features are ranked based on their matchability. Feature ranking is based on the idea that some features in an image contain more information than other features. We aim at ordering the extracted features by a given quality measure and query only the features with the highest chance of providing good match hypotheses. Feature matching (Step 3) aims at finding match hypotheses for the features with highest expected matchability. For each of the first n features in the ranking, a kNN query is posed on the database, where k can be seen as a way to tweak *recall* at a given number of query features, as the number of images returned by the query is at most $n * k$. To increase the efficiency of feature matching, we rely on fast state-of-the-art (approximate) indexing techniques optimized for high-dimensional data, for example Locally Optimized Product Quantization [3]. Then (Step 4), match expansion is performed on the resulting correspondences. In our scenario where we want to pose a small number of kNN queries on the system, we face the problem that even if we find some correspondences between the query and a database image, their number will be relatively low, increasing the probability that a good match is outranked by an image containing common random matches only. Expansion exploits the keypoint information of the seed matches that provide scale, rotation, and possibly affine information. These properties can be used to identify spatially close keypoints such as for example in [4]: Given that a match hypothesis is correct, not only the corresponding feature pair should match, but also its spatial neighborhood. The similarity of a match's neighborhood is evaluated using the

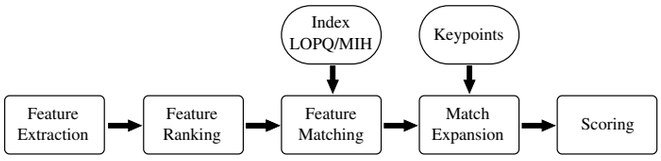


Figure 1: Recognition pipeline.

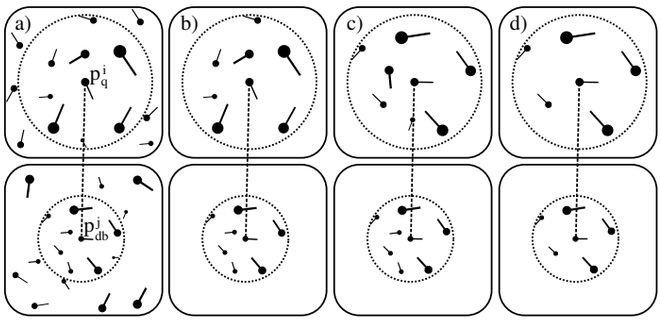


Figure 2: Generation of additional match hypotheses.

procedure shown in Figure 2. The figure shows an initial seed match, i.e. a kNN of a query feature and keypoints surrounding the seed match. The size of each keypoint is represented by the icon diameter, and the gradient direction is represented by a line anchored in the icon's center. The top row of this figure visualizes the features of the query image I_q, while the bottom row visualizes the image features of a tentative match image I_{db}. Starting point is an initial correspondence pair (p^i_q, p^j_{db}) established by kNN-search in feature space, see Figure 2 a). In a first step, features in a given spatial range are retrieved in the image I_q for p^i_q and in Image I_{db} for p^j_{db}, see Figure 2 b); the spatial range is visualized by a dotted circle. Spatially close keypoints with a significantly different scale than their reference feature are discarded (see the small features in the figure) similar to [1], resulting in two sets of features. These remaining features are rotation-normalized using the reference keypoint's gradient orientation information, rotating the set of keypoints and their corresponding gradient orientations, see Figure 2 c). Then the two lists of keypoints are traversed in parallel. If two features correspond sufficiently well in their appearance, orientation and position, the corresponding features are accepted as a matching pair (see Figure 2 d)). Based on the expanded list of matches, a score is computed for every database image (Step 5); we adapt the technique from [2], weighting scores based on their distance to the query feature and the number of features in the image. To increase the memory efficiency of the feature database, we compress the feature vectors used during match expansion using Product Quantization.

In our experimental evaluation we evaluate the effect of the number n of features queried, the effect of k in relation to the number of keypoints queried, and the pipeline's behaviour on different feature descriptors including real-valued and binary features.

[1] Ondrej Chum, Michal Perdoch, and Jiri Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *Proc. CVPR*, pages 17–24, 2009.

[2] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Exploiting descriptor distances for precise image search. Technical Report 7656, INRIA, 2011.

[3] Yannis Kalantidis and Yannis Avrithis. Locally optimized product quantization for approximate nearest neighbor search. In *Proc. CVPR*, 2014.

[4] Cordelia Schmid and Roger Mohr. Combining greyvalue invariants with local constraints for object recognition. In *Proc. CVPR*, pages 872–877, 1996.

JH2R: Joint Homography Estimation for Highlight Removal

Sungmin Eum
smeum@umiacs.umd.edu
Hyungtae Lee
htlee@umiacs.umd.edu
David Doermann
doermann@umiacs.umd.edu

University of Maryland Institute of Advanced Computer Studies (UMIACS)
University of Maryland College Park
College Park, MD, USA

Imagine being in an art museum where there are paintings or pictures held inside glass-frames for protection. There are pieces which you wish to capture using a camera, but you experience difficulties avoiding highlights which are generated by indoor lighting reflected off the glossy surfaces. Similar problems occur when capturing contents off of whiteboards, documents printed on glossy surfaces, objects such as books or CDs with plastic covers.

In this work, we address the problem of removing unwanted highlight regions in images generated by reflections of light sources on glossy surfaces. Although there have been efforts made to synthetically fill in the missing regions using the neighboring patterns by applying methods like inpainting [3, 4], it is impossible to recover the missing information in completely saturated regions. Therefore, we need to use multiple images where corresponding regions are not covered by the saturated highlights. Unlike other methods, our method uses the relationship between the highlight regions resulting in more robust removal of saturated highlights.

Our method - Overview Our method was motivated by a widely acknowledged physical phenomenon referred to as the 'motion parallax'. Without loss of generality, we can similarly view the relationship between the desired content (e.g., a painting) and the highlights. Since the highlights caused by the light source are the result of the reflection on the glossy surface before they reach the camera, the light source can be modeled to virtually exist on the other side of the content. Note that, the distance from the light source is always larger than the distance from the content ($D > d$, in Figure 1).

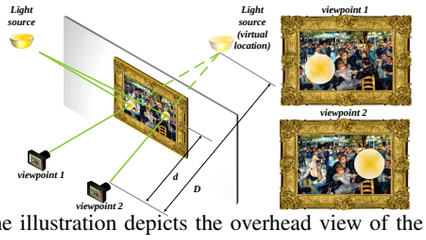


Figure 1: The illustration depicts the overhead view of the camera, the desired content, and the light source.

In order to distinguish the movements of the highlights, we need at least two images captured from different views. We then detect where the highlights are by searching for the two separate homography matrices: one for the content ($\mathbf{H}_C$) and the other for the highlights ($\mathbf{H}_H$). We exploit the fact that the homography ($\mathbf{H}_C$) which can properly overlay the desired contents in the two images will generate an erroneous overlap between the corresponding highlight regions. Similarly, the desired contents will display incorrect overlap when $\mathbf{H}_H$ is employed. This is shown in the second step of Figure 2(b).

Unlike the intrinsic layer separation problems, removing the saturated highlights from the images requires another image which can provide the corresponding non-highlight pixels. To perform such "pixel-transfer", it is necessary to have the pixel-level detection results of the highlights. In our approach, we first detect the highlight regions at the feature level by jointly estimating the two homographies using the proposed Joint Homography Estimation for Highlight Removal (JH2R) algorithm. However, the feature-level detection of the highlight regions is insufficient to properly eliminate the highlights. Thus, $\mathbf{H}_H$ is used to estimate the highlight regions at the pixel-level. Finally, we remove the highlights in both of the images by transferring the corresponding pixels from the complementary image using Poisson blending [4]. Figure 2 shows the schematic overview of our method. Details of the algorithm are explained in the original manuscript.

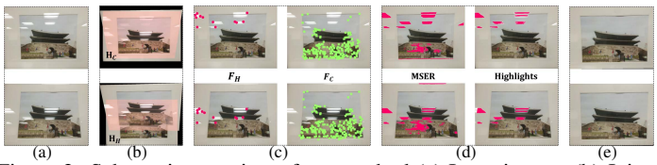


Figure 2: Schematic overview of our method (a) Input images (b) Joint homography estimation (c) Feature-level labeling (d) Pixel-level labeling (e) Final results

Comparison with state-of-the-art We have compared our method with four state-of-the-art algorithms [1, 2, 3, 5]. They are chosen to represent three different approaches to solve the given problem : 1) highlight removal, 2) single image-based reflection removal, and 3) multiple image-based reflection removal.

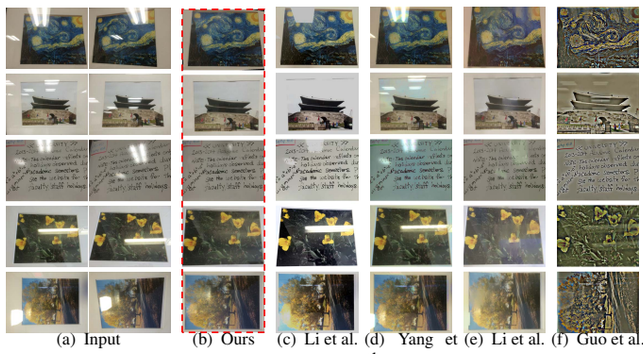


Figure 3: Five examples of highlight removal results using (b) our method compared with those produced by (c) Li et al. [3], (d) Yang et al. [5], (e) Li et al. [2], (f) Guo et al. [1]

Figure 3 shows five sample results of real world images. As can be observed in Figure 3(c) and 3(d), both [3] and [5] are incapable of removing the highlights due to the lack of information within the regions. Multiple-image based approaches (Figure 3(e) and 3(f)) by [1, 2] produce results where the highlights are only partially removed. Figure 3(e). Our method, unlike others, specifically uses the relationship between the "highlight" regions resulting in more precise detection and removal. One may observe from Figure 3 that our method can also handle dim highlights as there still exist geometrical distinction between desired contents and dim highlights in terms of homography. In overall, our method produces the most visually pleasing results.

[1] X. Guo, X. Cao, and Y. Ma. Robust separation of reflection from multiple images. In *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, pages 2195–2202, 2014.

[2] Y. Li and M.S. Brown. Exploiting reflection change for automatic reflection removal. In *Computer Vision (ICCV), IEEE International Conference on*, pages 2432–2439, 2013.

[3] Y. Li and M.S. Brown. Single image separation using relative smoothness. In *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on.*, pages 2752–2759, 2014.

[4] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, Jul 2003.

[5] Q. Yang, J. Tang, and N. Ahuja. Efficient and robust specular highlight removal. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1(1):1, 1 2014.

Action Recognition based on Subdivision-Fusion Model

Zongbo Hao[1]
zbhao@uestc.edu.cn
Linlin Lu[1]
fountainous@gmail.com
Qianni Zhang[2]
qianni.zhang@qmul.ac.uk
Jie Wu[1]
jiewu.uestc@gmail.com
Ebroul Izquierdo[2]
ebroul.izquierdo@qmul.ac.uk
Juanyu Yang[1]
yangjuanyu2008@163.com
Jun Zhao[1]
zhaojun.hjh@gmail.com

[1] School of Information and Software Engineering
University of Electronic Science and Technology of China
[2] Multimedia and Vision Group
School of Electronic Engineering and Computer Science
Queen Mary, University of London

This paper proposes a novel Subdivision-Fusion Model (SFM) to recognize human actions. In most action recognition tasks, overlapping feature distribution is a common problem leading to overfitting. In the subdivision stage of the proposed SFM, samples in each category are clustered. Boundaries for the more concentrated subcategories are easier to find and as consequence overfitting is avoided. In the subsequent fusion stage, we convert the multi-subcategories classification back to the original classification problem. Two methods to determine the number of clusters are provided. We evaluated our model SFM on four popular datasets to demonstrate the enhancement in performance.

Motivation

In most recognition tasks, the overlapping feature distribution is popular, as presented in Figure 1(a). Category 3 and 4 are both distributed in two regions and overlapped, instead of distributing in separating regions as supposed, like category 1 and 2 do. In this case it is not wise to find the boundary that enclose each category into one region and not overlapped with others, as shown in Figure 1(b). Trying to find that kind of boundaries, overfitting will be obviously resulted in. Hence, category 3 and 4 can be divided into two subcategories respectively, and the smooth boundaries can be found as shown in Figure 1(c).

Subdivision-Fusion Model (SFM)

Suppose there are n samples in the dataset, which are grouped into L categories. The features are extracted by some kind of models:

$$F(X)=[f_1,f_2,...,f_n]=h([x_1,x_2,...,x_n]) \quad (1)$$

where f_i is the feature vector for each sample x_i, h is the transformation by the network from the sample space $\mathbb{R}^s$ to feature space $\mathbb{R}^F$.

As presented in the motivation part, we can subdivide each category using clustering algorithm. The data of the ith category in the feature space are grouped into K_i clusters, and the labels are updated by the results of clustering. Let M be the total number of updated subcategories. Sparse Subspace Clustering (SSC) [1] is employed in clustering in this paper. By clustering and updating the labels, we made a transformation from the feature space to the sample space:

$$Y'=[y_1',y_2',...,y_n']=c([f_1,f_2,...,f_n]) \quad (2)$$

where c is the clustering operation, which transforms features to labels.

In Figure 2, Layer V with M nodes acts as an M-subcategory classifier. We connect an output Layer O with L nodes to Layer V. Let $W(i,k)$ be the weight between node k in Layer V and node i in Layer O:

$$W(i,k)=\begin{cases}1, V_k\in O_i\\ 0, V_k\notin O_i\end{cases} \quad (3)$$

where V_k is the kth subcategory output in Layer V, O_i is the ith category output in Layer O. If subcategory k belongs to category i, $W(i,k)=1$, otherwise $W(i,k)=0$. The final classification result R is calculated by:

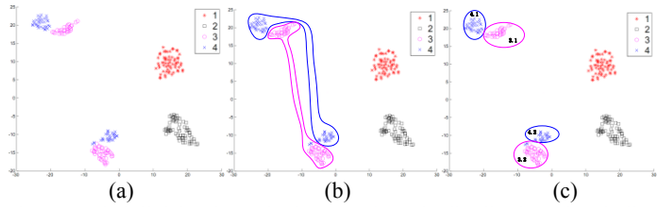
$$R=\max(O_i), O_i=\sum_{k=1}^{M}W(i,k)*V_k, 1\le i\le L, 1\le k\le M \quad (4)$$

According to (4), the M-subcategory classification is converted back to the L-category classification which is identical to the original problem.

The rule of determining cluster number K_i

1) Considering the overlapping feature distribution:

Our first method to determine the proper number K_i of subcategories is based on observing the t-Distributed Stochastic Neighbor Embedding (t-SNE) [2] 2-dimensional visualization directly.



(a) (b) (c)

Figure 1: Subdivision in each class. (a) Sample distribution. (b) Finding boundaries to distinguish category 3 and 4 in a usual way, it is prone to be overfitting. (c) Clustering in each class makes it easier to distinguish the classes and avoid overfitting.

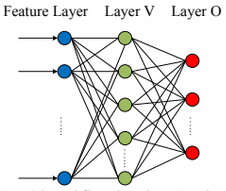


Figure 2: Classifier in the Fusion Stage

2) Considering the class imbalance problem:

Our second method to determine the proper number K_i of subcategories is based on the ratio of the number of each category samples to the number of the minority category samples.

Experimental results

SFM has been thoroughly tested with four popular datasets:

1. Hollywood2 dataset: the accuracy has been improved from 53.3% with Convolutional ISA [4] to 79.4%, significantly outperforming the state-of-the-art accuracy of 64.3% of Improved Trajectories [3].
2. YouTube Action dataset: the performance has been improved from 75.8% with Convolutional ISA [4] to 82.5%.
3. KTH dataset: the accuracy has been improved from 90.2% with 3D CNN [5] to 94.0%.
4. UCF50 dataset: the performance has been slightly improved from 76.4% with Action Bank features [6] to 76.9%.

[1] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(11), 2765-2781, 2013.

[2] L. Maaten and G.E. Hinton. Visualizing data using t-SNE. Journal of Machine Learning, 85(9), 2579-2605, 2008.

[3] H. Wang and C. Schmid. Action recognition with improved trajectories. In Computer Vision, IEEE International Conference on, 3551-3558, 2013.

[4] Q. V. Le, W. Zou, S. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal feature for action recognition with independent subspace analysis. In Computer Vision and Pattern Recognition, IEEE Conference on, 3361-3368, 2011.

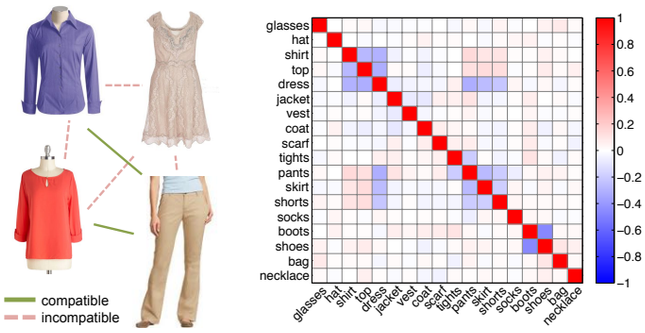
[5] S. Ji, W. Xu, M. Yang, and K.Yu. 3D convolutional neural networks for human action recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 221-231, 2013.

[6] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In Computer Vision and Pattern Recognition, IEEE Conference on, 1234-1241, 2012.

Mix and Match: Joint Model for Clothing and Attribute Recognition

Kota Yamaguchi[1]
http://vision.is.tohoku.ac.jp/~kyamagu
Takayuki Okatani[1]
okatani@vision.is.tohoku.ac.jp
Kyoko Sudo[2]
sudo.kyoko@lab.ntt.co.jp
Kazuhiko Murasaki[2]
murasaki.kazuhiko@lab.ntt.co.jp
Yukinobu Taniguchi[3]
ytaniguti@ms.kagu.tus.ac.jp

[1] Tohoku University
Sendai, Japan
[2] NTT
Yokosuka, Japan
[3] Tokyo University of Science
Tokyo, Japan



(a) Matching clothing pairs (b) Pearson correlation between clothing items

Figure 1: We consider the compatibility of items and attributes in detection. In b), notice the exclusive blocks, e.g., *shirt*, *top*, and *dress*.

Abstract This paper studies clothing and attribute recognition in the fashion domain. Specifically, in this paper, we turn our attention to the compatibility of clothing items and attributes (Fig 1). For example, people do not wear a skirt and a dress at the same time, yet a jacket and a shirt are a preferred combination. We consider such inter-object or inter-attribute compatibility and formulate a Conditional Random Field (CRF) that seeks the most probable combination in the given picture. The model takes into account the location-specific appearance with respect to a human body and the semantic correlation between clothing items and attributes, which we learn using the max-margin framework. Fig 2 illustrates our pipeline. We evaluate our model using two datasets that resemble realistic application scenarios: on-line social networks and shopping sites. The empirical evaluation indicates that our model effectively improves the recognition performance over various baselines including the state-of-the-art feature designed exclusively for clothing recognition. The results also suggest that our model generalizes well to different fashion-related applications.

Joint detection Let us denote a set of labels by $Y \equiv \{y_i\},\ y_i \in \{0,1\}$, where i is one of the clothing items or attributes, such as *shirt* or *skirt:plain*. Given a feature $X \equiv \{\mathbf{x}_i\}$, we define our joint probability distribution over labels by a log-linear model:

$$\ln P(Y|X) \equiv \sum_i w_i \phi(\mathbf{x}_i, y_i) + \sum_{i,j \in V} w_{i,j} \psi(y_i, y_j) - \ln Z, \tag{1}$$

We use logistic regression of each label for the unary term, expressed by:

$$\phi(\mathbf{x}_i, y_i) \equiv \ln p(y_i|\mathbf{x}_i), \tag{2}$$

$$p(y_i = 1|\mathbf{x}_i) \equiv \sigma(\mathbf{a}_i^{\mathrm{T}} \mathbf{x}_i + b_i), \tag{3}$$

where $\mathbf{a}_i$ and b_i are the regression parameters for each item. For the binary term, we use the normalized Pearson correlation:

$$\psi(y_i, y_j) \equiv \begin{cases} \ln \frac{1}{2}\left(1 + c_{i,j}\right), & \text{if } y_i = y_j \\ \ln \frac{1}{2}\left(1 - c_{i,j}\right), & \text{otherwise} \end{cases} \tag{4}$$

where $c_{i,j}$ is the Pearson correlation between item i and j in the training examples.

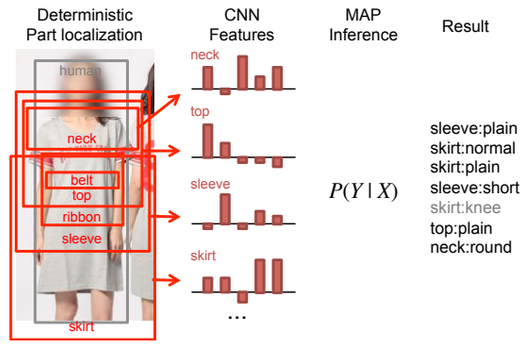
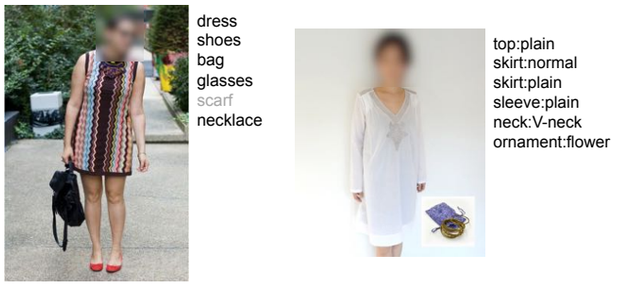


Figure 2: Our joint detection pipeline.



(a) Clothing detection (b) Attribute recognition

Figure 3: Results in a) clothing detection and b) attribute recognition.

Deterministically localized feature To reduce the dependency to very accurate pose estimation, we apply a simple yet effective approach for part localization. Given a bounding-box around the full human body, we extract an image patch from a specific location relative to the body bounding-box. Fig 2 illustrates an example of part bounding boxes for attribute detection. This deterministic approach only takes constant time yet produces surprisingly good results for fashion problems.

Once the region of interest is determined, we extract the CNN feature (`fc7` of AlexNet, 4096 dimensions) learned from ImageNet [1], and use as a feature-input to the logistic regression in Eq 2.

Experimental results We used two datasets: a) Chictopia dataset that considers automatic clothing tagging in fashion blogs, and b) Dress dataset that considers attribute recognition in e-commerce. Through the comparison between various baselines including the state-of-the-art image feature [2], we confirmed that the proposed joint model successfully improves the precision of prediction in both two scenarios. Fig 3 shows a few qualitative results.

[1] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014.

[2] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Retrieving similar styles to parse clothing. *TPAMI*, 2014.

Describing Common Human Visual Actions in Images

Matteo Ruggero Ronchi
http://vision.caltech.edu/~mronchi/
Pietro Perona
perona@caltech.edu

Computational Vision Lab
California Institute of Technology
Pasadena, CA, USA

Tuesday Poster 1

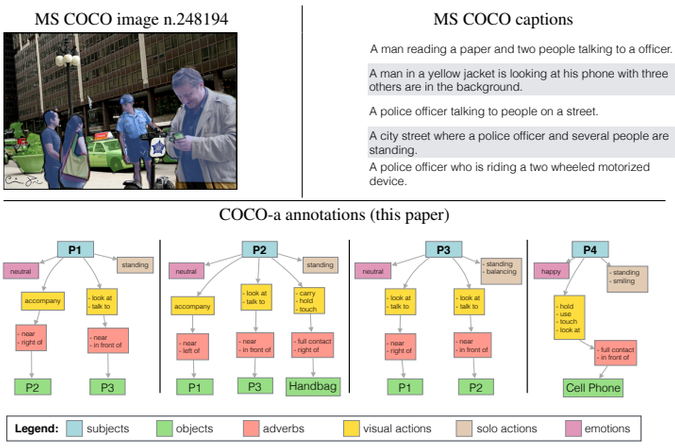


Figure 1: **COCO-a annotations**. (Top) MS COCO image and corresponding captions. (Bottom) COCO-a annotations. Each person (denoted by P1–P4, left to right in the image) is in turn a subject (blue) and an object (green). Annotations are organized by subject. Each subject and each subject-object pair is associated to states and actions. Each action is associated to one of the 140 visual actions in our dataset.

Which common human actions and interactions are recognizable in monocular still images? Which involve objects and/or other people? How many is a person performing at a time? We address these questions by exploring the actions and interactions that are detectable in the images of the MS COCO dataset. We make two main contributions. First, a list of 140 common 'visual actions', obtained by analyzing the largest on-line verb lexicon currently available for English (VerbNet [4]) and human sentences used to describe images in MS COCO. Second, a complete set of annotations for those 'visual actions', composed of subject-object and associated verb, which we call COCO-a (a for 'actions'). COCO-a is larger than existing action datasets in terms of number instances of actions, and is unique because it is data-driven, rather than experimenter-biased. Other unique features are that it is exhaustive, and that all subjects and objects are localized. A statistical analysis of the accuracy of our annotations and of each action, interaction and subject-object combination is provided.

In order to detect actions alongside objects the relationships between those objects needs to be discovered. For each action the roles of 'subject' (active agent) and 'object' (passive - whether thing or person) have to be identified. This information may be expressed as a 'semantic network' [5], which is the first useful output of a vision system for scene understanding.

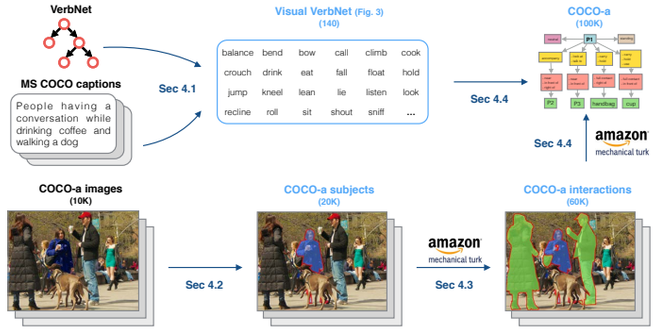


Figure 2: **Steps in the collection of COCO-a.** From VerbNet and MS COCO captions we extracted a list of visual actions. All the persons that are annotated in the MS COCO images were considered as potential 'subjects' of actions, and Amazon Mechanical Turk workers annotated all the objects they interact with, and assigned the corresponding visual actions. Titles in light blue indicate the components of the dataset.

accompany	chew	exchange	jump	pay	punch	sing	swim
avoid	clap	fall	kick	perch	push	sit	talk
balance	clear	feed	kill	pet	put	**skate**	taste
bend (pose)	climb	fight	kiss	photograph	reach	**ski**	teach
bend (something)	cook	fill	kneel	pinch	read	slap	**throw**
be with	crouch	float	laugh	play	recline	sleep	tickle
bite	cry	fly	lay	play baseball	remove	smile	touch
blow	cut	follow	lean	play basketball	repair	sniff	use
bow	dance	get	lick	play frisbee	ride	snowboard	walk
break	devour	give	lie	play instrument	roll	spill	wash
brush	dine	groan	lift	play soccer	row	spray	wear
build	disassemble	groom	**light**	play tennis	run	spread	whistle
bump	draw	hang	listen	poke	sail	squat	wink
call	dress	help	look	pose	separate	squeeze	write
caress	drink	hit	massage	pour	shake hands	stand	
carry	drive	hold	meet	precede	shout	steal	
catch	drop	hug	mix	prepare	show	straddle	
chase	eat	hunt	**paint**	pull	signal	surf	

Figure 3: **Visual VerbNet (VVN).** (Top-Left) The list of 140 visual actions that constitute VVN – bold ones added after the comparison with MS COCO captions. Of the total 2321 verbs in MS COCO captions, there is 60% overlap with the 66 verbs in VVN with > 500 occurrences.

Our goal is to collect an unbiased dataset with a large amount of meaningful and detectable interactions involving human agents as subjects. We put together a process, exemplified in Fig. 2, consisting of four steps: **(1)** Obtain the list of common visual actions that are observed in everyday images, by a combined analysis of VerbNet and MS COCO captions. Our list, which we call Visual VerbNet (Fig. 3) attempts to include all actions that are visually discriminable. It avoids verb synonyms, actions that are specific to particular domains, and fine-grained actions. Unlike previous work, Visual VerbNet is not the result of experimenter's idiosyncratic choices; rather, it is derived from linguistic analysis (VerbNet) and an existing large dataset of descriptions of everyday scenes (MS COCO captions). **(2)** Identify who is carrying out actions (the subjects), as all the people in an image whose pixel area is larger than 1600 pixels. All the people, regardless of size, are still considered as possible objects of an interaction. **(3)** For each subject identify the objects that he/she is interacting with, based on the agreement of 3 out of 5 Amazon Mechanical Turk annotators asked to evaluate each image. **(4)** For each subject-object pair (and each single agent) label all the possible actions and interactions involving that pair, along with high level visual cues such as emotion and posture, spatial relationship and distance.

Our novel dataset, COCO-a, consists of the 'semantic networks' extracted, by following the above process, from $10,000$ MS COCO images – examples of the obtained annotations are shown in Fig. 1. MS COCO images are representative of a wide variety of scenes and situations; 81 common objects are annotated in all images with pixel precision segmentations. A key aspect of our annotations is that they provide a complete description of an image (unlike captions which typically focus only on a limited part of it) and unambiguous, as they are based on Visual VerbNet, aiming to eliminate natural language ambiguities and non-visual actions.

We hope that our dataset will provide researchers with a starting point for conceptualizing about actions in images: which representations are most suitable, which algorithms should be used. We also hope that it will provide an ambitious benchmark on which to train and test algorithms. Amongst applications that are enabled by this dataset are building visual Q&A systems [1, 2], more sophisticated image retrieval systems [3], and automated analysis of actions in images of social media.

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. URL `http://arxiv.org/abs/1505.00468`.

[2] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A visual turing test for computer vision system. *Proceedings of the National Academy of Sciences (PNAS)*, 2015.

[3] J. Johnson, R. Krishna, M Stark, Li-Jia Li, D.A. Shamma, M.S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008.

[5] Stuart Russell and Peter Norvig. *Artificial intelligence, a modern approach*. Prentice-Hall, Egnlewood Cliffs, 1995.

Searching for Objects using Structure in Indoor Scenes

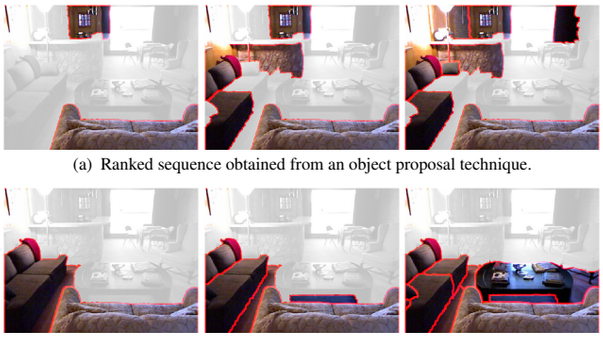
Varun K. Nagaraja
varun@umiacs.umd.edu
Vlad I. Morariu
morariu@umiacs.umd.edu
Larry S. Davis
lsd@umiacs.umd.edu

University of Maryland
College Park, MD. USA.

Consider the situation where a computer vision system needs to identify the presence or location of a particular object in an image. In a passive computer vision system, if we ask a specific question like "Where is the chair in this room?", it would process all the regions in the image to detect a chair instance. Such a vision system does not exploit the structure in the scene to efficiently process the image. In this work, our goal is to locate objects of interest in an image by processing as few image regions as possible using scene structure. Given a set of region proposals, we propose a search technique that sequentially processes the regions such that the regions that are more likely to correspond to the query class object are explored earlier. At each step during the search, we use the labels of the explored regions and spatial context to predict the likelihood that each unexplored region is an instance of the target class. We then select a few regions with highest likelihood, obtain the class label from the region classification module and add them to the explored set. The process is repeated with the updated set of explored regions.

We frame our sequential exploration problem as a Markov Decision Process (MDP) and use a reinforcement learning technique to learn an optimal search policy. Let R be the set of indices of the regions in the image and t correspond to a step index. Let R_e^t be the set of indices of the explored image regions and $R_u^t = R \setminus R_e^t$ be the set of indices of the unexplored image regions at a step t. To state our problem in the reinforcement learning setting, a state s_t is the set of all the image regions (r) explored until that step $s_t = \{r_i | i \in R_e^t\}$. An action corresponds to selecting the next image region to explore, $a_t = r_j$ where $j \in R_u^t$. A policy π is a function that maps states to actions $\pi(s)$. The goal is to find a policy that will maximize a cumulative function of the reward. However, it is challenging to manually specify a reward function for the search policy. The true reward function is unknown for our sequential exploration problem since the underlying distribution from which a spatial arrangement of objects in an image is generated is unknown, analogous to a game generated by a hidden emulator. But we have access to an oracle's actions in the individual images. Learning an optimal policy in such situations is known as imitation learning [3] where an oracle predicts the actions it would take at a state and the search policy learns to imitate the oracle and predict similar actions. The oracle in our image exploration problem selects the next set of regions to explore based on the groundtruth labels.

We use the DAgger (Dataset Aggregation) algorithm of Ross *et al*. [3] that trains a classifier as the search policy which takes in features extracted at a state as input and predicts an action as the output. In typical imitation learning algorithms, the policy is learned by training a classifier on the dataset of state features and actions obtained by sampling sequences produced by an oracle policy. They make an i.i.d assumption about the states encountered during the execution of a learned policy which does not hold for our problem since the policy's prediction affects future states. During the test stage, if the policy encounters a state that was not generated by the oracle policy, it could predict an incorrect action that can lead to compounding of errors. But DAgger does not make the i.i.d assumptions about states. During training, it starts with an initial classifier and runs through the states, predicting labels for each state. Based on its predictions, it is assigned a loss value at each state. At the end of an iteration, all the features, the predicted labels and the loss values for all states are collected. The aggregate of all the collected datasets until the current iteration is used to train a cost sensitive classifier, which becomes the policy for the next iteration. Since the training states are generated by the policy being learned, it is a non i.i.d supervised learning problem. DAgger is available in the Vowpal Wabbit library through a programming abstraction proposed by Daumé III *et al*. [1] where a developer writes a single *predict* function that encodes the algorithm for the testing stage and the training is done by making repeated calls to this predict function.



(a) Ranked sequence obtained from an object proposal technique.

(b) Sequence obtained from a search strategy that uses structure in the scene.

Figure 1: **Searching for a table**. Each step in the above sequence shows exploration of three additional regions in the image. The search strategy learned using our method utilizes the room structure and the presence of other objects in the image to discover the table region much earlier than using the ranked sequence from an object proposal technique.

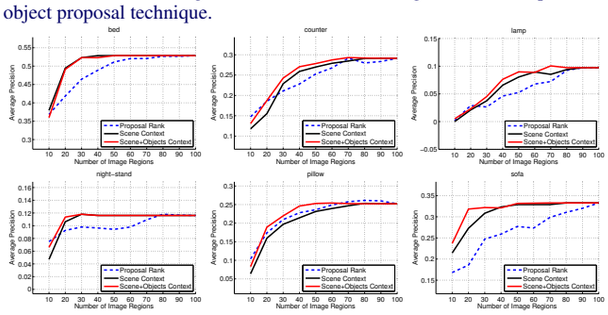


Figure 2: **Average Precision (AP) vs. number of processed regions.** A classifier trained for a query class with unary scene context features alone can achieve a significantly high average precision by processing very few regions. Classes like *bed, nightstand* and *sofa* need only 20-25% of the regions when compared to the proposal ranking sequence. A search strategy trained for a query class using both object-object context and scene-context features further improves the performance for classes like *counter, lamp, pillow* and *sofa*.

Since structure in the scene is essential for search, we work with indoor scene images as they contain both unary scene context information and object-object context in the scene. We demonstrate our approach on the NYU depth v2 dataset [4]. We use the RCNN-Depth module of Gupta *et al*. [2] for the region classification. Their region proposal module is a modified Multiscale Combinatorial Grouping (MCG) technique that incorporates depth features. Their feature extraction module is RCNN which includes CNNs fine-tuned on the depth images. We perform experiments on the NYU-depth v2 dataset and show that the unary scene context features alone can achieve a significantly high average precision while processing only 20-25% of the regions for classes like *bed* and *sofa*. By considering object-object context along with the scene context features, the performance is further improved for classes like *counter*, *lamp*, *pillow* and *sofa*. Our sequential search process adds a negligible overhead when compared to the time spent on extracting CNN features, hence the reduction in number of regions leads directly to a gain in computation speed of the object detection process.

[1] Hal Daumé III, John Langford, and Stephane Ross. Efficient programmable learning to search. *arXiv preprint arXiv:1406.1837*, 2014. URL `http://arxiv.org/abs/1406.1837`.

[2] Saurabh Gupta, Ross Girshick, P Arbeláez, and J Malik. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In *ECCV*, 2014.

[3] Stéphane Ross, Geoffrey J Gordon, and J Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction. In *AISTATS*, 2011.

[4] Nathan Silberman, Pushmeet Kohli, Derek Hoiem, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 2012.

Fast Inverse Compositional Image Alignment with Missing Data and Re-weighting

Vincent Lui
vincent.lui@monash.edu
Dinesh Gamage
dinesh.gamage@monash.edu
Tom Drummond
tom.drummond@monash.edu

ARC Centre of Excellence for Robotic Vision,
Monash University,
Clayton, Australia

Image alignment is often performed using variants of the Lucas-Kanade algorithm [3]. Among these variants, the inverse compositional (IC) [1] method and the efficient second-order minimization (ESM) [2] method are the most efficient variants. While ESM is computationally more efficient for 2D image alignment problems such as homographies, IC is more computationally efficient for image alignment problems using RGB-D data where it is expensive to re-compute the Jacobian.

In this paper, we look at methods to accelerate the convergence of IC and ESM algorithms while remaining robust to outliers. We propose a preconditioning strategy to perform inverse composition with re-weighting and missing data which avoids the need to re-compute the Jacobian and its Hessian at every iteration. We also consider how the effects of image noise and/or spectral aliasing over the scale of the deformation can be used to speed up the convergence of all these methods.

The inverse compositional Lucas-Kanade algorithm minimizes the following cost function:

$$C_{\text{IC}} = \sum_{\mathbf{x}} [T(\mathbf{M}(\mathbf{x};\Delta\mathbf{p})) - I(\mathbf{M}(\mathbf{x};\mathbf{p}))]^2 \tag{1}$$

with respect to $\Delta\mathbf{p}$. $\mathbf{p}$ is the parameter vector and $\mathbf{x}$ is the pixel location in the template. $\mathbf{M}(\mathbf{x};\mathbf{p}))$ is a warping function which maps the pixel $\mathbf{x}$ in the template T to a sub-pixel location $I(\mathbf{M}(\mathbf{x};\mathbf{p}))$ in the reference image I. As $J_{\mathbf{x}}$ is computed using the template image T which does not depend on $\mathbf{p}$, it can be pre-computed as $J(0)$. Equation (1) is minimized iteratively in a non-linear least squares manner. The parameter vector $\mathbf{p}$ is related to the error vector $\mathbf{e}$ as

$$J\Delta\mathbf{p} = \mathbf{e}. \tag{2}$$

where J is the Jacobian. At each iteration, we form the normal equations

$$H\Delta\mathbf{p} = J^T\mathbf{e}, \tag{3}$$

where $H = J^TJ$ is the Hessian. $\Delta\mathbf{p}$ is then computed as $\Delta\mathbf{p} = H^{-1}J^T\mathbf{e}$. The re-weighted version of the problem can be posed as $(J^TWJ)\Delta\mathbf{p} = J^TW\mathbf{e}$, where W is usually a diagonal matrix. Inverse composition with re-weighting requires the Jacobian and Hessian to be re-computed at every iteration.

Conventionally, the pseudo-inverse of the Hessian is computed using Cholesky factorization. Here, we propose to perform a QR factorization on the Jacobian J instead. This allows us to re-formulate Equation (2) as

$$Q\begin{bmatrix} R \\ 0 \end{bmatrix}\Delta\mathbf{p} = \mathbf{e}, \tag{4}$$

where Q is an $m \times m$ orthogonal matrix where $Q^T = Q^{-1}$, and R is an $m \times n$ upper triangular matrix. The parameter updates $\Delta\mathbf{p}$ can now be computed as $\Delta\mathbf{p} = R^{-1}Q^T\mathbf{e}$. Without going into details here, the re-weighting problem can be formulated as

$$\Delta\mathbf{p} = R^{-1}(Q^TWQ)^{-1}QW\mathbf{e}. \tag{5}$$

The factors Q and R can be pre-computed. This leaves the weighting matrix W and the term $(Q^TWQ)^{-1}$ to be computed at every iteration. Here, we propose to use a preconditioning approach to approximate $(Q^TWQ)^{-1}$, where the goal is to construct a matrix P such that the product $P^{-1}(Q^TWQ)$ has a smaller condition number to Q^TWQ. We propose three preconditioners, each with increasing complexity (see Figure 1). The first preconditioner is a *scaled identity matrix* where every element in the diagonal represents the inverse of the average weight. The second preconditioner is the *Jacobi preconditioner*, where the matrix P is a diagonal matrix. Every element in the diagonal is computed by accumulating the sum of weighted, squared elements in the corresponding column of the pre-computed factor Q. The last preconditioner extends the Jacobi preconditioner to accumulate off-diagonal elements as well. Assuming that the off-diagonal elements are close to zero, we approximate the inverse of this preconditioner using a first-order Neumann series.

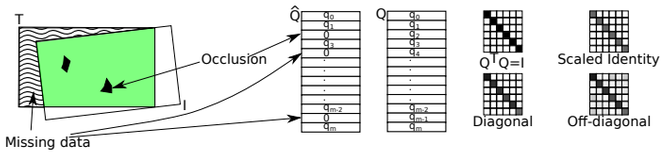


Figure 1: Left: Missing data and re-weighting causes the factor Q to turn into a matrix $\hat{Q}$ that is no longer orthogonal. Right: The proposed preconditioners P. The product $P^{-1}(\hat{Q}^T\hat{Q})$ should be as close to I as possible. Here, a darker shade represents a value closer to 1.0.

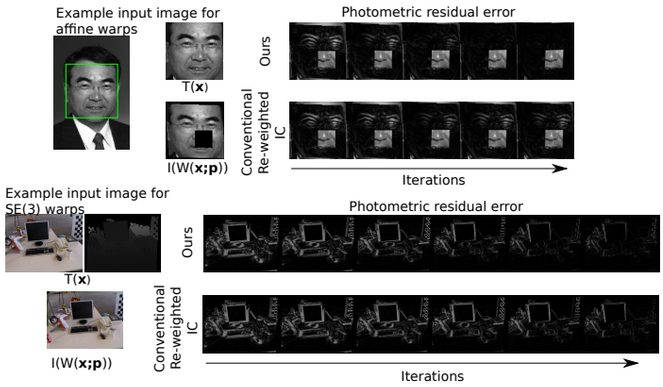


Figure 2: Qualitative results showing the how the residual photometric error changes with the number of iterations.

Finally, we also investigated the effects of noise and/or spectral aliasing over the scale of deformation affects the convergence of image alignment algorithms. We found that noise and/or spectral aliasing acts as a regularizer term that damps the system and causes the parameter updates to be underestimated. We showed that a moderate increase in the step size of the parameter updates can further improve the convergence of image alignment algorithms.

We have evaluated our proposed methods using different motion models such as affine warps, homographies, and SE(3) warps (see Figure 2), and our experimental results show that our method is equally as robust as the conventional re-weighted inverse composition while exhibiting faster convergence rates. We found that the scaled identity and diagonal preconditioners offer the best trade-off between computational efficiency and robustness. For further details of our proposed method, including implementation details, experimental results, and a discussion, please refer to the full paper.

[1] S. Baker, R. Gross, T. Ishikawa, and I. Matthews. Lucas-Kanade 20 Years On : A Unifying Framework : Part 2. Technical report, 2003.

[2] S. Benhimane and E. Malis. Real-time image-based tracking of planes using Efficient Second-order Minimization. In *International Conference on Intelligent Robot and Systems*, volume 1, pages 943–948, 2004.

[3] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

Automatic Age Estimation from Face Images via Deep Ranking

Huei-Fang Yang[1]
hfyang@citi.sinica.edu.tw
Bo-Yao Lin[2]
boyaolin@iis.sinica.edu.tw
Kuang-Yu Chang[2]
kuangyu@iis.sinica.edu.tw
Chu-Song Chen[2]
song@iis.sinica.edu.tw

[1] Research Center for Information Technology Innovation
Academia Sinica
Taipei, Taiwan

[2] Institute of Information Science
Academia Sinica
Taipei, Taiwan

This paper focuses on automatic age estimation (AAE) from face images, which amounts to determining the exact age or age group of a face image according to features from faces. Although great effort has been devoted to AAE [1, 4, 6], it remains a challenging problem. The difficulties are due to large facial appearance variations resulting from a number of factors, *e.g.*, aging and facial expressions. AAE algorithms need to overcome heterogeneity in facial appearance changes to provide accurate age estimates.

To this end, we propose a generic, deep network model for AAE (see Figure 1). Given a face image, our network first extracts features from the face through a 3-layer scattering network (ScatNet) [2], then reduces the feature dimension by principal component analysis (PCA), and finally predicts the age via category-wise rankers constructed as a 3-layer fully-connected network. The contributions are: (1) Our ranking method is point-wised and thus is easily scaled up to large-scale datasets; (2) our deep ranking model is general and can be applied to age estimation from faces with large facial appearance variations as a result of aging or facial expression changes; and (3) we show that the high-level concepts learned from large-scale neutral faces can be transferred to estimating ages from faces under expression changes, leading to improved performance.

Our model is with the following characteristics:

(1) **The scattering features are invariant to translation and small deformations.** ScatNet is a deep convolutional network of specific characteristics. It uses predefined wavelets and computes scattering representations via a cascade of wavelet transforms and modulus pooling operators from shallow to deep layers. With the nonlinear modulus and averaging operators, ScatNet can produce representations that are discriminative as well as invariant to translation and small deformations. As ScatNet provides fundamentally invariant representations for discriminating feature extraction, only the weights of the fully-connected layers are learned in our network model, which considerably reduces the training time.

(2) **The rank labels encoded in the network exploit the ordering relation among labels.** Each category-wise ranker is an ordinal regression ranker. We encode the age rank based on the reduction framework [5]. Given a set of training samples $X = \{(x_i, y_i), i = 1 \cdots N\}$, let $x_i \in R^D$ be the input image and y_i be a rank label ($y_i \in \{1, \ldots, K\}$), respectively, where K is the number of age ranks. For rank k, we separate X into two subsets, X_k^+ and X_k^-, as follows:

$$\begin{aligned} X_k^+ &= \{(x_i, +1) | y_i > k\} \\ X_k^- &= \{(x_i, -1) | y_i \leq k\}. \end{aligned} \quad (1)$$

Next, we use the two subsets to learn a binary classifier from the network, which then conducts an answer to the query: "Is the image with a ranking score higher than k?" To each query, we simply make a binary decision between the positive and negative sides. Hence, each query reduces the age rank estimation task to a binary classification problem. A series of query results imply the ordinal relationship between the rank labels, where each query identifies the preferred classes.

Because the k-th binary classifier focuses on determining whether the age rank of an image is greater than k, $K-1$ such binary classifiers are required for K age ranks. Therefore, for a face image with true age k, the teaching vector with length $K-1$ is designed as $[1, \ldots, 1, -1, \ldots, -1]$, where the first $k-1$ values are 1 and the remaining -1.

(3) **The category-wise rankers perform age estimation within the same group.** Category-wise age estimation is commonly carried out in a hierarchical manner, first performing between-category classification and then within-category age estimation [3, 4]. Employing a hierarchical strategy in our network involves introducing more layers, which may result in more computational costs. We instead concatenate the encoding of multiple category-wise rankers altogether in the output layer. This design allows the category-wise rankers to learn age estimation according to the shared high-level representations.

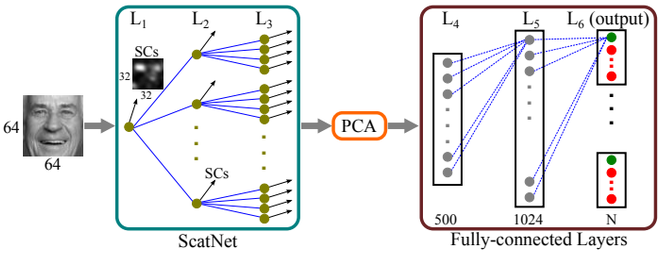


Figure 1: Our network model for human age prediction from face images. Our DeepRank+ comprises a 3-layer ScatNet, a dimensionality reduction component by principal component analysis (PCA) that reduces the feature dimension to 500, and a 3-layer fully-connected network with an architecture of 500-1024-N. The number of nodes N is application dependent. We use ReLUs in layer L_5 and sigmoids in layer L_6. The output layer is constructed with the category labels (the green nodes) and ranking ones (the red). When the category information is not used in training, the output layer can be constructed simply with the rank labels; such a network is coined as DeepRank.

Assume there are C category-wise rankers. The encoding for each ranker consists of two constituents: the category label(s) and the age rank. That is, for category j, its encoding is given as $e_j = [g_j, r_j]$, where g_j denotes the desired output of the category and r_j denotes the teaching vector of the rank. Concatenating the C encoding sets forms a final encoding: $E = [g_0, r_0, \ldots, g_j, r_j, \ldots, g_C, r_C]$.

We conduct two sets of sexperiments on three datasets. One is the experiments on age estimation from faces obtained from different gender and races on a large-scale MORPH dataset. The other is the experiments on age estimation from faces under expression changes from the Lifespan and FACES datasets. Our approach yields mean absolute errors (MAEs) of 3.49, 5.19 and 7.04 years on MORPH, Lifespan, and FACES, respectively, performing favorably against state-of-the-arts.

[1] Fares Alnajar, Zhongyu Lou, José Manuel Álvarez, and Theo Gevers. Expression-invariant age estimation. In *Proc. BMVC*, 2014.

[2] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[3] Guodong Guo and Guowang Mu. Human age estimation: What is the influence across race and gender? In *Proc. CVPRW*, 2010.

[4] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014.

[5] Hsuan-Tien Lin and Ling Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24 (5):1329–1367, 2012.

[6] Dong Yi, Zhen Lei, and Stan Z. Li. Age estimation by multi-scale convolutional network. In *Proc. ACCV*, 2014.

Feature Encoding of Spectral Signatures for 3D Non-Rigid Shape Retrieval

Frederico A. Limberger
http://www.cs.york.ac.uk/~fal504
Richard C. Wilson
http://www.cs.york.ac.uk/~wilson
University of York
York, UK

A large number of local signatures have been created to represent local characteristics of geometric 3D shapes for the purpose of many computer vision, geometry processing and shape analysis tasks. A local signature is a compact representation that characterizes a small region of a shape. They usually capture information about the neighbourhood of a vertex and so they can be directly applied to some important tasks like point correspondence and shape segmentation. For this purpose, it is desirable to compute signatures that are invariant under rigid, non-rigid and isometric deformations, the typical deformations that 3D models undergo. However, local descriptors cannot be immediately applied to the problem of shape retrieval, because this task is not addressed by comparing local signatures but by comparing global descriptors (signatures that represent the shape as a whole). Creating a global descriptor is not a simple task since shapes can have arbitrary number of vertices, edges and faces. To create a generic representation of a shape all important characteristics should be preserved during the encoding process. Even so, the global representation must compress local characteristics using the same basis for all shapes in order to facilitate comparisons.

In this paper, we propose an efficient and discriminative encoding framework to address the problem of creating global signatures for 3D models from local descriptors based on the spectrum of the shape, for the purpose of shape retrieval and classification. In this way, we propose the use of Fisher Vector (FV) to describe the entire representation of a shape. Differently from [1, 2], our approach uses a Gaussian Mixture Model (GMM) as a dictionary of probabilistic visual words, and encodes the global signature using three orders statistics (0-th, 1-st, 2-nd) rather than using only the first order. Further, while the classical Bag-of-Features (BoF) generates a K-dimensional histogram, where K is the vocabulary size, Fisher Vector encoding generates a high-dimensional vector with $2KD$ dimensions, where D is the size of each local descriptor, being more discriminative but still simple to compare, as all shapes are encoded in the same basis. We also propose some improvements to the Wave Kernel Signature (WKS), since it overtakes the Heat Kernel Signature (HKS) and Scale Invariant Heat Kernel Signature (SI-HKS) in retrieval performance. A better scaling is proposed for the eigenvalues of the Laplace-Beltrami operator which captures more information about the shape and we also propose the use of principal curvatures to increase the efficiency of the encoding method.

The Fisher Vector encoding characterizes a large set of vectors by their three-order deviation from a vocabulary, creating a high-dimensional gradient vector representation. The gradient of the log-likelihood, also called *Fisher score*, describes the contribution of each parameter to the generation process. To compute the FV encoding, we write the local shape descriptors wrt. the probabilistic model (Gaussian Mixture Model). To properly compare shape signatures we have also applied L2 Normalization and Power Normalization to the Fisher Vector, also known as the Improved FV [3].

When it comes to improving the WKS (1), we have analysed the influence of the logarithmic scaling to the computation of the signature and we concluded that it loses high frequency information about the spectrum of the shape (even important high frequencies that do not originate from noise are blurred). Then, we analysed the distributions of the differences between eigenvalues of 24 articulated woman shapes (Figure 1) and we propose to use the cubic root scaling (2) rather than the classical logarithmic scale (1) since it fits much better a normal distribution, which is used to handle the differences between same-class shapes.

$$\mathrm{WKS}(x,e) = C_e \sum_{k=1}^{\infty} \phi_k(x)^2 f_E(\Lambda_k)^2 \qquad f_E(\Lambda_k)^2 = e^{\frac{-(e-\log(\Lambda_k))^2}{2\sigma^2}} \tag{1}$$

$$f_C(\Lambda_k)^2 = e^{\frac{-(e-\sqrt[3]{\Lambda_k})^2}{2\sigma^2}} \qquad e \in [\sqrt[3]{\lambda_{min}}, \sqrt[3]{\lambda_{max}}] \tag{2}$$

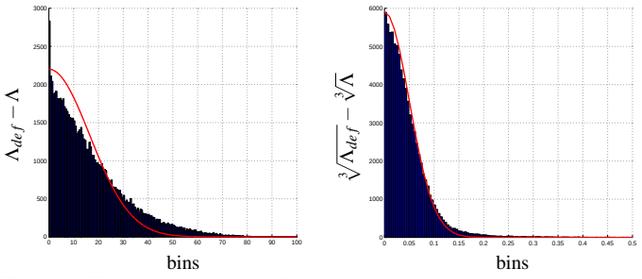


Figure 1: Histograms of the differences between the eigenvalues Λ of the Laplace-Beltrami operator of 24 articulated woman shapes. (left) Histogram of the differences not scaled. (right) Histogram of the differences scaled by the cubic root. On each graph, the red line is a reference normal distribution fitted to the respective histogram.

We also analyze the WKS when combined with encoding methods. These methods are based on the differences between descriptor histograms and a probabilistic model. The more discriminative these histograms are the more the encoding process will be able to distinguish between shapes of different classes. We have noticed that when shapes undergo pose changes, the maximum principal curvature of the vertices remains stable. This is a local feature of the surface, which is isometry invariant and stable at most points under object articulation but not well coded in the WKS. Therefore, as another improvement to the shape retrieval task, we propose the use of principal curvatures to increase the separation of features leading to a more discriminative histogram. Practically, we shift the WKS individually by the maximum principal curvature c for each vertex x as stated in Equation (3). We smooth the maximum principal curvature by taking the mean of the respective neighbour vertices to diminish the influence of noise. In (3), α is a weight that normalizes c accordingly to the signature values. We use $\alpha = 0.015$ in all our experiments.

$$\mathrm{IWKS}(x,e) = C_e \sum_{k=1}^{\infty} \phi_k(x)^2 f_C(\Lambda_k)^2 + c_x\alpha \tag{3}$$

We compare our method against the competitors of the non-rigid tracks of SHREC'11 and SHREC'15. By analyzing our retrieval accuracy in the SHREC'11 track it is clear that our approach improves spectral descriptors to tackle shape retrieval problems since it achieves excellent retrieval measures (DCG very close to 1) and comparable results with best retrieval methods (very similar e-Measure to the top tier methods). Furthermore, our method presents a much better performance when compared to other spectral descriptors, showing its potential.

Our conclusion is that although our method does not beat all other groups in the SHREC'11 and SHREC'15 benchmarks it proves that spectral methods are a good choice to retrieve shapes when combined with informative encoding methods, since our method is close to the best techniques in performance and achieves excellent results in most benchmark classes. We observed that the our worst retrieval performances happen when shapes undergo huge topology changes. Specially in these cases, spectral signatures still need to be improved. In future works, we plan to create a spectral signature that is less variant to major topology changes.

[1] Alexander M. Bronstein, Michael M. Bronstein, Leonidas J. Guibas, and Maks Ovsjanikov. Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Trans. Graph.*, 30(1):1:1–1:20, February 2011.

[2] M. Ovsjanikov, A.M. Bronstein, M.M. Bronstein, and L.J. Guibas. Shape google: a computer vision approach to isometry invariant shape retrieval. In *ICCV Workshops*, pages 320–327, Sept 2009.

[3] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. ECCV*, pages 143–156, Berlin, 2010.

A BoW-equivalent Recurrent Neural Network for Action Recognition

Alexander Richard
richard@iai.uni-bonn.de
Juergen Gall
gall@iai.uni-bonn.de

Institute of Computer Science III
University of Bonn
Bonn, Germany

Bag-of-words (BoW) models are widely used in the field of computer vision and find application in texture and object classification, object discovery, and action recognition. In the traditional BoW model, kMeans or a Gaussian mixture model is used to generate a visual vocabulary based on which a histogram of visual words is computed. This histogram is then used for classification, *e.g.* in combination with a support vector machine (SVM). In our paper, we show an equivalent formulation of BoW as a recurrent neural network that allows to train the visual words discriminatively and directly on video level rather than on frame level only. We apply our method to four action recognition benchmarks as well as two small image recognition datasets and show its superiority over the classical model and some sparse coding methods such as LLC [3] and ScSPM [4].

In the classical approach, a set $\mathbf{x}$ of T feature vectors is quantized into a histogram of K visual words $v_1, \dots, v_K$,

$$\mathcal{H}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} h(x_t), \quad h(x_t) = \begin{pmatrix} p(v_1|x_t) \\ \vdots \\ p(v_K|x_t) \end{pmatrix}, \tag{1}$$

where $p(v_k|x_t)$ is 1 if v_k is the visual word being closest to x_t and 0 otherwise. As an alternative to this hard assignment, soft assignment is frequently used, *i.e.* $p(v_k|x_t)$ is defined as a posterior distribution of a Gaussian mixture model or the kMeans model. Utilizing that kMeans is a special case of Gaussian mixture models, where the visual word prior is uniform and all variances are set to 1, we formulate an equivalent representation as a single-layer neural network with softmax output,

$$p_{\mathrm{NN}}(v_k|x) := \mathrm{softmax}_k(\mathbf{W}^\intercal x + b), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{D\times K}$ is a weight matrix and $b \in \mathbb{R}^K$ the bias. Setting these to

$$\mathbf{W} = (v_1 \dots v_K), \quad b = -\frac{1}{2}(v_1^\intercal v_1 \dots v_K^\intercal v_K)^\intercal \tag{3}$$

results in a neural network that generates the same posterior distribution as kMeans with soft assignment.

In order to realize the summation over all x_t that is required in Equation (1), we add a recurrent layer with the unit matrix as weights to the network. The normalization over the sequence length is realized by the activation function

$$\sigma_t(z) = \begin{cases} z & \text{if } t < T, \\ \frac{1}{T} z & \text{if } t = T. \end{cases} \tag{4}$$

An additional softmax layer is finally added to model a posterior distribution of the action classes given the video, *c.f.* Figure 1.

Standard neural network training algorithms can now be applied to optimize the weights and biases in the model. The softmax layer we added on top of the network is only a linear classifier. In the classical BoW model, usually a SVM with non-linear kernel is applied. Thus, we use the softmax classifier only during training in order to obtain visual words that are optimized

1. discriminatively, *i.e.* separating the different classes as good as possible, and
2. on video level, *i.e.* the visual vocabulary is learned to best differentiate between collections of feature vectors that form the video rather than between single feature vectors or video frames only.

Once the network is trained, the output layer is discarded. The neural network histograms of an input sequence $\mathbf{x} = x_1, \dots, x_T$ are then used in combination with an SVM after application of a non-linear kernel.

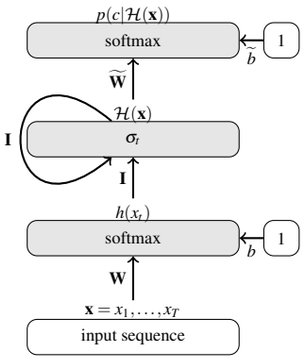


Figure 1: Neural network encoding the BoW model. The output layer is discarded after training and the histograms from the recurrent layer are used for classification in combination with an SVM.

Method	HMDB-51	UCF101
Traditional models		
Improved DT + bag of words	52.2%	73.3%
Improved DT + fisher vectors [2]	57.2%	85.9%
Improved DT + LLC	50.8%	71.9%
Neural networks		
Composite LSTM [1]	44.0%	75.8%
Ours	**54.0%**	**76.9%**

Table 1: Comparison of our model to published results on HMDB-51 and UCF101.

Note that the result of the neural network training can be transformed back into a traditional kMeans model with non-uniform prior. The corresponding visual words and the prior can be computed as

$$v_k = (\mathbf{W}_{1,k} \dots \mathbf{W}_{D,k})^\intercal, \tag{5}$$

$$p_{\mathrm{NN}}(v_k) = \frac{\exp\left(b_k + \frac{1}{2} v_k^\intercal v_k\right)}{\sum_{\tilde{k}} \exp\left(b_{\tilde{k}} + \frac{1}{2} v_{\tilde{k}}^\intercal v_{\tilde{k}}\right)}. \tag{6}$$

Using improved dense trajectories [2] as video features, we conduct experiments on four different action recognition benchmarks. We observe a consistent improvement between two and five percent compared to the traditional BoW model and also outperform other methods that learn video representations discriminatively, such as LLC or composite LSTMs [1], see Table 1. An evaluation of our method on two image datasets led to similar results.

[1] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. *arXiv preprint arXiv:1502.04681*, 2015.

[2] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Int. Conf. on Computer Vision*, pages 3551–3558, 2013.

[3] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.

[4] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.

Depth Restoration via Joint Training of a Global Regression Model and CNNs

Gernot Riegler*[1]
riegler@icg.tugraz.at
René Ranftl*[1]
ranftl@icg.tugraz.at
Matthias Rüther[1]
ruether@icg.tugraz.at
Thomas Pock[1,2]
pock@icg.tugraz.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision
Graz University of Technology
Graz, Austria

[2] Digital Safety & Security Department
Austrian Institute of Technology
Vienna, Austria

Depth sensors have become increasingly popular in the recent years for a wide range of applications. Among these are video games, gesture control, and applications in the automotive industry. However, the noise and the resolution of depth sensors are problematic. To generate useful depth estimates, the data provided by the sensor is typically subject to post-processing steps. In this work we propose a method that combines a global regression model with a convolutional neural network (CNN) to tackle those problems. Global models, such as TGV-L2 [1, 3], are well suited for the restoration of depth maps, since their prior assumption modeled in the regularization term fit the piecewise affine nature of the data. The data term in these models is designed based on fixed assumptions about the underlying sensor noise, however. Instead of fixed a-priori assumption about the data, we propose to parametrize a global model using a CNN and learn the parameters of the complete system end-to-end in a data-driven way. In this way, our model can automatically adapt to the underlying noise characteristics of the data.

To estimate accurate depth from an input I_k, we use a Global Regression Model (GRM) of the following general form

$$E(u; f(w, I_k)) = R(u, h(w_h, I_k)) + \frac{\exp(w_\lambda)}{2} \|u - g(w_g, I_k))\|^2. \quad (1)$$

Here, $R(u, h(w_h, I_k))$ is a regularization term that introduces prior knowledge and is parameterized by the function $h(w_h, I_k)$. Similarly, the function $g(w_g, I_k)$ is used to transform the input data such that the GRM can make reliable estimates. We use for both parameterization functions h, g a CNN. As regularizer we utilize a non-local pairwise model

$$R(u) = \sum_{i=1}^{N} \sum_{j>i}^{N} n_\varepsilon(h_{ij}(w_h, I_k) \cdot (u_i - u_j)), \quad (2)$$

where n_ε is a twice-differentiable approximation of the Huber function [4]

$$n_\varepsilon(t) = [|t| \leq \varepsilon] \cdot \left(-\tfrac{1}{8\varepsilon^3} t^4 + \tfrac{3}{4\varepsilon} t^2 + \tfrac{3\varepsilon}{8} \right) + [|t| > \varepsilon] \cdot |t| \,. \quad (3)$$

To train the weights $w = [w_h, w_\lambda, w_g]^T$, we assume that K pairs of sensor images I_k together with their ground-truth v_k are given. We formulate the training task as a bi-level problem [6]:

$$\min_{w \in W} \tfrac{1}{K} \sum_{k=1}^{K} L(u^*(f(w, I_k)), v_k) \quad \text{(HL)}$$

$$\text{s.t. } u^*(f(w, I_k)) = \arg\min_{u \in \mathbb{R}^N} E(u; f(w, I_k)). \quad \text{(LL)}$$

The training procedure can be interpreted as follows: Find parameters w such that the minimizer u^* of $E(\cdot)$ yields a small training loss $L(\cdot)$. We proof in this work necessary conditions for the energy $E(\cdot)$ and the loss $L(\cdot)$ that allows us to efficiently compute the gradient of the higher-level (HL) problem with respect the weights w. If this conditions are satisfied, we can compute the gradient as follows:

$$\Delta E = - \left(\left[\left(\nabla_u^2 E \right)^{-1} \frac{\partial L}{\partial u_k} \right]^T \frac{\partial^2 E}{\partial u_k \partial f} \right) \Bigg|_{u_k = u_k^*}. \quad (4)$$

This formulation fits nicely into the backpropagating scheme to train neural networks and enables us to train the GRM and the CNN in an end-to-end fashion. Further, we can use stochastic gradient descent, or any variation of it, to train the complete model.

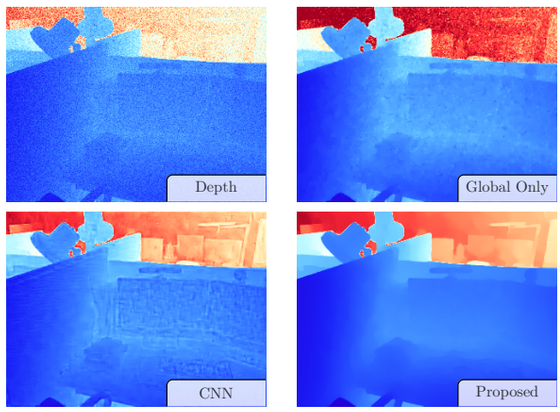


Figure 1: Our proposed method that combines a GRM with a CNN achieves better results than the individual components alone.

		Ours (NL)	CNN	K-SVD	BM3D	TGV-L2
Local Variance	$\sigma_d = 0.5d$	**2.730**	3.913	5.741	3.178	3.042
Poisson	$\lambda = 10^{-3}$	**3.4016**	8.660	6.695	4.129	10.595
Salt & Pepper	$p = 0.35$	**10.484**	18.880	81.685	77.010	80.764

Table 1: Quantitative results on different noise distributions. The error is measured as RMSE in *cm*.

	Ours (NL)	Ours (L)	CNN	Ferstl *et al.*
×2	**2.940**	3.042	6.427	3.834
×4	**4.530**	4.813	8.411	5.506

Table 2: Upscaling results of our method with non-local (NL) and local (L) regularization compared to the SRCNN [2] and the method by Ferstl *et al.* [3]. The error is measured as RMSE in *cm*.

We evaluated our approach for depth map denoising and upscaling on the New Tsukuba dataset [5] for different noise types and levels. An excerp of our quantitative results are presented in Table 1 and 2. Figure 1 depicts an exemplar qualitative result for the denoising task.

The evaluations demonstrate that the proposed GRM parameterized with a CNN is indeed able to adapt to different noise characteristics and performs better than the individual parts independently. Finally, we see potential of our method in computer vision tasks that can benefit from joint learning of a global model and its parameterization.

Acknowledgments This work was supported by *Infineon Technologies Austria AG* and the Austrian Research Promotion Agency (FFG) under the *FIT-IT Bridge* program, project #838513 (TOFUSION).

[1] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total Generalized Variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a Deep Convolutional Network for Image Super-Resolution. In *ECCV*, 2014.

[3] David Ferstl, Christian Reinbacher, René Ranftl, Matthias Rüther, and Horst Bischof. Image Guided Depth Upsampling using Anisotropic Total Generalized Variation. In *ICCV*, 2013.

[4] Karl Kunisch and Thomas Pock. A Bilevel Optimization Approach for Parameter Learning in Variational Models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

[5] Sarah Martull, Martin Peris, and Kazuhiro Fukui. Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps. In *ICPR Workshops*, 2012.

[6] René Ranftl and Thomas Pock. A Deep Variational Model for Image Segmentation. In *GCPR*, 2014.

*Authors contributed equally

Shape Detection with Nearest Neighbour Contour Fragments

Kasim Terzić
w3.ualg.pt/~kterzic
Hussein Adnan Mohammed
a48025@ualg.pt
J.M.H. du Buf
w3.ualg.pt/~dubuf

Vision Lab (LARSyS)
University of the Algarve
Faculdade de Ciências e Tecnologia
Gambelas, Faro, Portugal

Shape is probably the single most important feature for object detection and much research has gone into developing deformable shape models. However, contours extracted by bottom-up edge detectors are notoriously unreliable, especially in natural images. In this paper, we present a very simple, yet powerful method for model creation, hypothesis generation, and hypothesis verification, which is competitive with much more complex methods.

Let $\mathbf{s}$ be a descriptor in some high-dimensional space of an edge fragment representing part of an object contour. In Bayesian terms, this fragment was likely to be generated by some class $c \in C$ if the conditional likelihood $P(\mathbf{s}|c)$ is greater for c than that for any other class, $P(\mathbf{s}|c' \neq c)$, including the background class. $P(\mathbf{s}|c)$ can be estimated non-parametrically, for example by using Gaussian kernels associated with some nearby samples in the feature space. In [2] it was shown that a good approximation can be obtained by using only the nearest sample. A fragment is *discriminative* if it is much more likely to belong to a class c than any other class $c' \neq c$. We thus define the discriminative power of a fragment by:

$$d(\mathbf{s}) := \log \frac{P(\mathbf{s}|c)}{P(\mathbf{s}|c' \neq c)} \,. \tag{1}$$

As proposed in [3], we can approximate the second likelihood using the distance to the nearest sample of any other class:

$$d(\mathbf{s}) := \|\mathbf{s} - NN_{c'}(\mathbf{s})\|^2 - \|\mathbf{s} - NN_c(\mathbf{s})\|^2; \qquad c' \neq c \,, \tag{2}$$

which is a simple, distance-based criterion. We further define the *relevance* $r(\mathbf{s},c)$ of a fragment for class c as the probability that a similar fragment $\mathbf{s}'$ appears in an annotated sub-image I_c containing c. We consider a fragment $\mathbf{s}'$ similar to $\mathbf{s}$ if the distance between the two is less than some threshold T:

$$r(\mathbf{s},c) = P(\|\mathbf{s} - \mathbf{s}'\| < T|c); \qquad \text{where } \mathbf{s}' = NN_c(\mathbf{s}). \tag{3}$$

We estimate T by using the nearest neighbour of $\mathbf{s}$ in all training images *not* containing c. The probability in Eqn. 3 can thus be approximated by counting the number of different annotated sub-images of c in the training set, in which there is at least one fragment closer to $\mathbf{s}$ than T.

We begin by extracting all maximally long contiguous edge fragments from the image, by linking edges produced by the Berkeley detector. We extend this set by adding subsets of all fragments with one half and one quarter length shifted to cover different parts of each fragment, which strikes a good balance between the number of fragments and sufficient statistics. Each fragment obtained is stored as a vector of 2D coordinates, and a shape context $\mathbf{s}$ is computed for each vector, using standard parameters. [1]

Next, we build a canonical model for each shape. Relevance $r(\mathbf{s},c)$ is calculated for each fragment belonging to each class c and they are sorted in descending order. We keep the most relevant 5% of the fragments and discard the rest. Thus an ensemble of fragments is used to represent each shape.

We then generate object hypotheses by determining the discriminative power $d(\mathbf{s})$ of each fragment $\mathbf{s}$ extracted from a test image. We keep the top 20% of most discriminative fragments and discard the rest. For each of the remaining fragments (called "trigger fragments"), we generate a hypothesis by projecting the bounding box of the nearest neighbour of $\mathbf{s}$ into a new image, scaling it to fit the position and scale of $\mathbf{s}$. Hypotheses are clustered around few promising spots, and after removing obvious overlaps, we are left with about 20 hypotheses per class per image. Each hypothesis is generated by one "trigger" fragment. We refine the detections by collecting fragments inside each hypothesis and fitting a bounding box which jointly minimises the bounding box error for all fragments using least square minimisation.

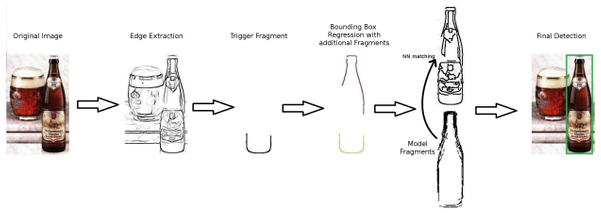


Figure 1: Overview of our method.

Apple	Bottles	Giraffes	Mugs	Swans	**Mean**
95/95	**96.4/100**	**81.3/85.4**	**87.1/87.1**	**88.2/94.1**	**89.6/92.3**

Table 1: Results on ETHZ (recall at 0.3/0.4 FPPI, respectively).

The final step in our algorithm is the verification of each hypothesis. A score for a hypothesis can be obtained from the average distance between fragments $\mathbf{s}_i$ in the hypothesis, and fragments $\mathbf{q}_m$ in the model of the corresponding class, but we have to be careful. Matching a hypothesis to the model will include the influence of clutter fragments and ignore missing parts of the shape. Both issues can be avoided by matching the fragments $\mathbf{q}_m$ from the model to the segments in the hypothesis in order to obtain the distance. Clutter fragments in the hypothesis will tend to be ignored because they do not resemble the models, and parts of the model contour that do not have a close match will increase the overall distance, thus penalising incomplete contours. We define the final scoring function of a hypothesis h in the following way:

$$\text{score}(h) = \frac{1}{M} \sum_m (\mathbf{q}_m - NN_h(\mathbf{q}_m))^2 \,, \tag{4}$$

where $\mathbf{q}_m$ is the position-enhanced shape context descriptor of the mth fragment in the model and $NN_h(\mathbf{q}_m)$ is its nearest neighbour in the hypothesis.

We evaluated our algorithm on the popular ETHZ shape dataset, with results shown in Table 1. We note that there are several difficult aspects in shape detection in natural images: i) model creation, ii) hypothesis generation, and iii) hypothesis verification/classification. Our method tackles all these problems consistently based on Bayesian criteria and nearest neighbours.

We are currently investigating a combination of our approach with advanced contour extraction methods and more powerful classifiers. We believe that contour linking and a more powerful verification step can significantly boost the results. There is also potential for further optimisation, making the algorithm more suitable for real-time applications. Finally, we would like to explore using prior information from scene models to improve detection.

Acknowledgements This work was supported by the EU under the FP-7 grant ICT-2009.2.1-270247 *NeuralDynamics* and by the FCT under the grant UID/EEA/5009/2013.

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE T-PAMI*, 24(4):509–522, 2002.

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, Anchorage, 2008.

[3] S. McCann and D.G. Lowe. Local naive bayes nearest neighbor for image classification. In *CVPR*, pages 3650–3656, Providence, 2012.

Exploiting Image-trained CNN Architectures for Unconstrained Video Classification

Shengxin Zha[1]
szha@u.northwestern.edu
Florian Luisier[2]
fluisier@bbn.com
Walter Andrews[2]
wandrews@bbn.com
Nitish Srivastava[3]
nitish@cs.toronto.edu
Ruslan Salakhutdinov[3]
rsalakhu@cs.toronto.edu

[1] Northwestern University
Evanston IL USA
[2] Raytheon BBN Technologies
Cambridge, MA USA
[3] University of Toronto
Toronto, Ontario, Canada

Tuesday Poster 1

In this paper, we propose an efficient approach to exploit off-the-shelf *image-trained* CNN architectures for video classification and evaluate on the challenging TRECVID MED'14 dataset and UCF-101 dataset. Our work is closely related to other research efforts towards the efficient use of CNN for video classification. While it is now clear that CNN-based approaches outperform most state-of-the-art handcrafted features for image classification, it is not yet obvious that this holds true for video classification. Moreover, there seems to be mixed conclusions regarding the benefit of training a spatiotemporal vs. applying an image-trained CNN architecture on videos. Although the specificity of the considered video datasets might play a role, the way the 2D CNN architecture is exploited for video classification is certainly the main reason behind these contradictory observations. The additional computational cost of training on videos is also an element that should be taken into account when comparing the two options. Prior to training a spatiotemporal CNN architecture, it thus seems legitimate to fully exploit the potential of image-trained CNN architectures. Obtained on a highly heterogeneous video dataset, we believe that our results can serve as a strong 2D CNN baseline against which to compare CNN architectures specifically trained on videos.

We conduct an in-depth exploration of different strategies for doing event detection in videos using convolutional neural networks (CNNs) trained for image classification (Figure 1, 2). We study different ways of performing spatial and temporal pooling, feature normalization, choice of CNN layers as well as choice of classifiers. Making judicious choices along these dimensions led to a very significant increase in performance over more naive approaches that have been used till now. The modality fusion of image-trained CNN features and motion-based Fisher vectors shows considerably improvement in classification performance. On TRECVID MED'14 dataset, our methods, based entirely on image-trained CNN features, can outperform several state-of-the-art non-CNN models. Our proposed late fusion of CNN- and motion-based features can further increase the mean average precision (mAP) on MED'14 from 34.95% to 38.74%, achieving the state-of-the-art (Table 1). On the challenging UCF-101 dataset, the image-based approach outperforms other image-trained CNN approaches; the fusion approach yields 89.6% classification accuracy (Table 2).

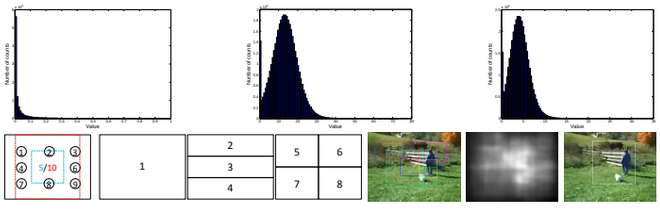


Figure 1: CNN features in output- and hidden-layer (top); spatial pyramid pooling and objectness-based pooling (bottom)

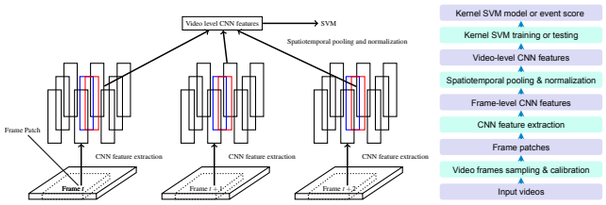


Figure 2: Overview of the proposed video classification pipeline.

Table 1: Result comparison on TRECVID MED'14 100Ex

Method	CNN	mAP
D-SIFT [3]+FV [5]	no	24.84%
IDT [7]+FV [5]	no	28.45%
D-SIFT+FV, IDT+FV (fusion)	no	33.09%
MIFS [2]	no	29.0 %
CNN-LCD$_{VLAD}$ with multi-layer fusion [8]	yes	36.8 %
proposed: CNN-hidden6	yes	33.54%
proposed: CNN-hidden7	yes	34.95%
proposed: CNN-hidden7, IDT+FV	yes	**38.74%**

Table 2: Result comparison on UCF-101

Method	Mean acc.
Spatial stream ConvNet [6]	73.0%
Temporal stream ConvNet [6]	83.7%
Two-stream ConvNet fusion by avg [6]	86.9%
Two-stream ConvNet fusion by SVM [6]	88.0%
Slow-fusion spatiotemporal ConvNet [1]	65.4%
Single-frame model [4]	73.3%
LSTM (image + optical flow) [4]	88.6%
MIFS [2]	89.1%
proposed: CNN-hidden6 only	79.34%
proposed: CNN-hidden6, IDT+FV (avg. fusion)	**89.62%**
proposed: CNN-hidden7, IDT+FV (avg. fusion)	89.30%

[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] Zhen-Zhong Lan, Ming Lin, Xuanchong Li, Alexander G. Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[4] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[7] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[8] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation

Ken Sakurada
sakurada@vision.is.tohoku.ac.jp
Tohoku Univeristy
Miyagi, Japan

Takayuki Okatani
okatani@vision.is.tohoku.ac.jp
Tohoku University/JST CREST
Miyagi, Japan

This paper proposes a method for detecting changes of a scene using a pair of its vehicular, omnidirectional images. The top images of figure 1 show an example of such image pairs taken at different times. Apparently, there are temporal differences in illumination and photographing conditions. Moreover, there has to exist visual difference in camera viewpoints, although they were captured from a vehicle running on the same street and were matched using GPS data. This is due to differences in vehicle paths and shutter timing. The type of scene changes targeted here includes 3D (e.g. vanishing/emergence of buildings, cars etc.) as well as 2D changes (e.g. changes of textures on building walls). To precisely detect these changes from such an image pair, it is necessary to overcome these unwanted visual differences.

We tackle the change detection problem in the 2D domain. That is, we consider detecting changes based on the direct comparison of a pair of images. The major issue is then how to deal with the unwanted visual differences (i.e., viewpoint differences etc.) To cope with this, we propose to use the features extracted by convolutional neural networks (CNNs). To be specific, we use a fully trained CNN for large-scale object recognition task [4] in a transfer learning setting. It was reported in the literature that using activation of the upper layers of a CNN trained for a specific task can be reused for other visual classification tasks. Several recent researches imply that the upper layers of CNNs represent and encode highly-abstract information about the input image [1, 2, 5]. We conjecture that highly-abstract (or object-level) changes can be detected by using the upper layers, whereas low-level visual changes (e.g. edge, texture etc.) will be detected using the lower layers. We show that this conjecture is true through several experimental results.

The proposed method consists of the three components: i) extraction of grid features, ii) superpixel segmentation, and iii) estimation of sky and ground areas by Geometric Context. These are described below.

(i) Extraction of grid features We denote two input images by I^t and $I^{t'}$, where t and t' are the times at which they were captured. First, I^t and $I^{t'}$ are divided into grid cells $g(=1,...,N_g)$. A feature is extracted from each grid cell g, yielding $\mathbf{x}_g{}^t$ and $\mathbf{x}_g{}^{t'}$.

The changes that we want to detect are object-level changes (e.g, the emergence/vanishing of buildings and cars) and not low-level, appearance changes due to changes in viewpoints, illumination or photographing conditions. To distinguish these two, the proposed method uses the activation of a upper layer of a deep CNN for the grid features $\mathbf{x}_g{}^t$ and $\mathbf{x}_g{}^{t'}$. To be specific, we use a pooling layer of the CNN. Each feature (e.g., $\mathbf{x}_g{}^t$) is the activation of all the units in the same location across the maps of the pooling layer. Thus $\mathbf{x}_g{}^t$ has the same number of elements as the maps of the pooling layer.

Next, these features are normalized so that $|\mathbf{x}_g{}^t| = 1$, and then their dissimilarity is calculated at each grid cell g as

$$d_g = |\mathbf{x}_g{}^t - \mathbf{x}_g{}^{t'}|. \tag{1}$$

Then, the dissimilarity d_g is projected onto the input images I^t and $I^{t'}$, determining the pixel-level dissimilarity $d_p(p=1,...,N_p)$; N_p is the number of pixels. This is done by simply setting $d_p = d_g$ for any pixel p contained in the grid cell g.

(ii) Superpixel segmentation The difference in viewpoint is arguably the major source of difficulties for 2D change detection methods. The use of the CNN features is expected to help mitigate this difficulty, owing to the property of the CNN features invariant to geometric transformation such as translation, 3D rotation, and even more complicated ones. However, the resolution of the dissimilarity map d_p's is basically very low. We use superpixel segmentation to refine the dissimilarity map to hope for obtaining precise boundaries of the detected changes.

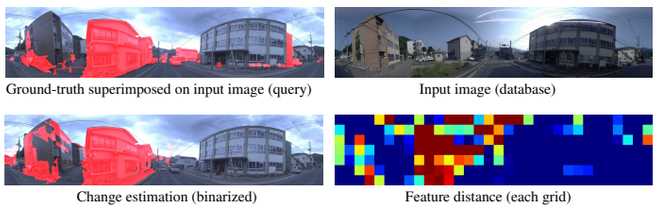


Figure 1: Results of change detection using pool-5 feature of CNN.

This starts with computing superpixel segmentation of I^t and $I^{t'}$. Let s^t be a superpixel and S^t be the set of superpixels. We define the dissimilarity d_{s^t} at a superpixel $s^t \in S^t$ to be the average of all the pixels in s^t as

$$d_{s^t} = \frac{1}{|s^t|} \sum_{p \in s^t} d_p. \tag{2}$$

We denote the maximum value of d_{s^t} and $d_{s^{t'}}$ by $d_{\max}$, i.e., $d_{\max} = \max(d_{s^t}, d_{s^{t'}}$

(iii) Estimation of sky and ground areas by Geometric Context In the last step of the proposed method, Geometric Context [3] is used to remove the segments of sky and ground from the images. Geometric Context is a segmentation method that is known to be robust to changes in illumination and photographing conditions. It estimates probabilities of the sky and the ground at each pixel $({p_{\text{sky}}}^t, {p_{\text{ground}}}^t)$ in the input image I^t. Using these, we remove these areas from the images, converting the dissimilarity at each pixel into $\overline{d_p}$ as

$$\overline{d_p} = \begin{cases} 0 & (({p_{\text{sky}}}^t > a) \wedge ({p_{\text{sky}}}^{t'} > a)) \\ & \vee (({p_{\text{ground}}}^t > b) \wedge ({p_{\text{ground}}}^{t'} > b))), \\ d_{\max} & (\text{otherwise}) \end{cases} \tag{3}$$

where $a = t_{\text{sky}}$ and $b = t_{\text{ground}}$ are constant values within the range of $0 \le t_{\text{sky}}, t_{\text{ground}} \le 1$.

We have created *Panoramic Change Detection Dataset* and used it for the experiments. The data used in this study (the pairs of the omnidirectional panoramic images taken at different time points and the hand-labeled ground-truth of change detection) can be downloaded from our website. The dataset consists of two subsets, named "TSUNAMI" and "GSV." "TSUNAMI" consists of one hundred panoramic image pairs of scenes in tsunami-damaged areas of Japan. "GSV" consists of fifty panoramic image pairs of Google Street View. The size of these images is 224×1024 pixels.

Figure 1 shows an example of the results of change detection. It is observed from them that the proposed method was able to correctly detect the scene changes, for example, demolished and new buildings and cars. In some cases, Geometric Context could not accurately estimate sky due to electrical wire and pole, or could not distinguish between the ground and low height object (e.g., debris and car). In contrast, the proposed method was able to accurately detect object-level scene changes.

[1] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599. Springer, 2014.

[2] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[3] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

Robust Multiview Registration of 3D Surfaces via ℓ_1-norm Minimization

Anil C. Raghuramu
cranil@ee.iisc.ernet.in

Department of Electrical Engineering
Indian Institute of Science
Bangalore, India

Registration (or alignment) of surfaces is an important task in many areas such as robotics, computer aided modeling, virtual reality and surface reconstruction. Building a full model of a object requires registering multiple surfaces observed from different viewpoints. In this paper we present an approach robust to outliers, for registering multiple surfaces simultaneously when the transformation is rigid.

Multiview registration simultaneously registers multiple scans. Given scans $\mathcal{S}_1 \dots \mathcal{S}_N$ and a view-graph (i.e. a graph with edges between scans which have common points), $\mathcal{G} = (\{1 \dots N\}, E)$. This problem can be generalized as

$$\min_{\mathbf{M}_i \in SE(3)} \mathcal{E}(\mathbf{M}_1 \dots \mathbf{M}_N) \tag{1}$$

Where

$$\mathcal{E}(\mathbf{M}_1 \dots \mathbf{M}_N) = \sum_{(i,j)\in E} \sum_{\mathbf{x}\in\mathcal{S}_i} d(\mathbf{M}_i\mathbf{x}, \mathbf{M}_j\mathcal{S}_j)^2 \tag{2}$$

and $\mathbf{M}\mathcal{S} = \{\mathbf{M}\mathbf{x} \,|\, \mathbf{x} \in \mathcal{S}\}$.

The iterative closest point algorithm [1, 3] (ICP) is a pairwise alignment method which iteratively solves the optimization problem (2) by splitting the problem into a correspondence estimation step and a transformation estimation step.

Correspondence Estimation: The correspondence estimation problem is to find a point in $\mathbf{y}_i \in \mathcal{S}_2$ for each $\mathbf{x}_i \in \mathcal{S}_1$. When the scans are close enough the correspondences can be calculated using closest point queries [1] as above. If the scans are far apart correspondences can be found using feature matching. In either case the matches are not always accurate. For instance when the scans are partially overlapping, some points in the scans will not have matching points. Hence, heuristics are applied for better alignment *e.g.* removing correspondences which are too far away; see Rusinkiewicz and Levoy [6] for more details.

Transformation Estimation: Once the correspondences are known, the transformation can be estimated by minimising a sum of squared distance function

$$\{\mathbf{M}_i^{(s)}\} = \underset{\{\mathbf{M}_i\}\subset SE(3)^N}{\arg\min} \sum_{(i,j)\in E} \sum_{k=1}^{n_{ij}} d(\mathbf{M}_i\mathbf{x}_{ij}^k, \mathbf{M}_j\mathbf{x}_{ji}^k)^2 \tag{3}$$

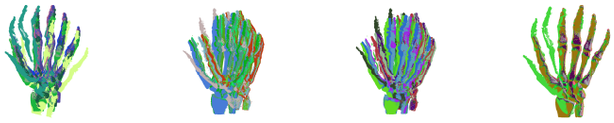
The distance function $(\mathbf{x}, \mathbf{y})$ varies based on the approximation of $d(\mathbf{x}, \mathcal{S})$ used. Commonly used are the euclidean distance (or point-to-point distance) and the distance between the point $\mathbf{x}$ and the tangent plane of $\mathcal{S}$ containing $\mathbf{y}$.

Recent works include Krishnan et al. [5], Govindu and Pooja [4] and Torsello et al. [7]. Krishnan et al. [5] solves (3), with d being the euclidean distance, using a manifold based Newton's method. While Govindu and Pooja [4] and Torsello et al. [7] find the ICP estimates for the pairwise transformation and obtain global transformations by global averaging. They use the consistency constraint $\mathbf{M}_{ij} = \mathbf{M}_j\mathbf{M}_i$ and find the global motions by solving the minimization problem for some distance d

$$\min \sum_{i,j} d(\mathbf{M}_{ij}, \mathbf{M}_j\mathbf{M}_i)^2 \tag{4}$$

These algorithms are dependent on finding matching points between scans, but a significant number of them are spurious, and it is necessary to clean up the matches obtained. This requires a substantial amount of tuning of parameters and the final result might still contain outliers. Since the number of outliers are sparse we formulate the registration optimization using the ℓ_1-norm. We present experimental results to show that the performance of our algorithm is comparable to state of the art algorithms.

If we assume that the residual noise in the scans are Gaussian we can model the transformation estimation step of multiview registration problem as



ICP initialization Torsello et al. ℓ_2-MVICP ℓ_1-MVICP

Figure 1: The results here show the performance of different algorithms when the initialization (left) is bad. Torsello et al. [7] and ℓ_2-MVICP Krishnan et al. [5] both fail while ℓ_1 MVICP ignores the outlier scan and registers the rest of them correctly.

$$\mathbf{e}_{ij}^k = \mathbf{M}_i\mathbf{p}_{ij}^k - \mathbf{M}_j\mathbf{p}_{ji}^k \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

Thus, the maximum likelihood estimate gives us ℓ_2 optimization problem which was solved by Krishnan et al. [5]. We instead formulate the problem in terms of outliers and assume that the noise is small (and not necessarily Gaussian)

$$\mathbf{M}_i\mathbf{p}_{ij}^k - \mathbf{M}_j\mathbf{p}_{ji}^k = \mathbf{O}_{ij}^k \tag{6}$$

Since the vector $\mathbf{O}_{ij} = [\mathbf{O}_{ij}^{k\top} \dots \mathbf{O}_{ij}^{n_{ij}\top}]^\top$ is sparse we can use the ℓ_1-norm to impose this constraint. So our cost function is

$$\mathcal{E}_{\ell_1}(\mathbf{M}_1 \dots \mathbf{M}_N) = \sum_{(i,j)\in E} ||\mathbf{O}_{ij}||_1 \tag{7}$$

By using the approximation $\mathbf{M}_i = \mathbf{I} + \mathbf{m}_i$ where $\mathbf{m}_i$ is of the form

$$\mathbf{m} = \begin{bmatrix} 0 & -\omega_z & \omega_y & u_x \\ \omega_z & 0 & -\omega_x & u_y \\ -\omega_y & \omega_z & 0 & u_z \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

we can reduce (7) to

$$\mathcal{E}_{\ell_1}(\mathbf{x}) = ||\boldsymbol{\delta} - \mathbf{D}\mathbf{x}||_1 \tag{9}$$

where, $\mathbf{x} = [\boldsymbol{\omega}_1 \cdots \boldsymbol{\omega}_N \, \mathbf{u}_1 \cdots \mathbf{u}_N]^T$ which can be solved iteratively (for an available implementation see [2]).

It must be noted that the approximation (8) is valid only if $\boldsymbol{\omega}_i$'s and $\mathbf{u}_i$'s are small. Hence, the optimization cannot start from an arbitrary point. For most cases, a good starting point can be obtained by an initial pairwise ICP step. The implementation details are described in the paper. We also show that when there are outliers, our method does significantly better than the state-of-the-art algorithms.

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[2] Emmanuel Candes and Justin Romberg. l1-magic: Recovery of sparse signals via convex programming. 2005. URL `http://statweb.stanford.edu/~candes/l1magic/downloads/l1magic.pdf`.

[3] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image Vision Comput.*, 10(3):145–155, April 1992. ISSN 0262-8856.

[4] Venu Madhav Govindu and A. Pooja. On averaging multiview relations for 3d scan registration. *IEEE Transactions on Image Processing*, 23(3):1289–1302, March 2014. ISSN 1057-7149.

[5] Shankar Krishnan, Pei Yean Lee, John B Moore, and Suresh Venkatasubramanian. Global registration of multiple 3d point sets via optimization-on-a-manifold. In *Symposium on Geometry Processing*, pages 187–196, 2005.

[6] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001.

[7] Andrea Torsello, Emanuele Rodola, and Andrea Albarelli. Multiview registration via graph diffusion of dual quaternions. In *CVPR*, pages 2441–2448. IEEE, 2011.

Because better detections are still possible: Multi-aspect object detection with Boosted Hough Forest

Carolina Redondo-Cabrera
carolina.redondoc@edu.uah.es
Roberto López-Sastre
robertoj.lopez@uah.es

University of Alcalá
GRAM
Alcalá de Henares, ES

In the last few years, many adaptations of the Hough Forest (HF) [2] approach for object detection have been proposed. In this work, we proceed to deconstruct the HF learning model to investigate whether a considerable better performance can be obtained detecting multi-aspect objects.

Inspired by [4], we introduce the Boosted Hough Forest (BHF). Essentially, the BHF is a HF where the decision trees are trained in a stage-wise fashion, by optimizing a global loss function.

While the tress are learned, any image patch $\mathcal{P}_t$ can be propagated through them, following the path from the root to the leaves. Leveraging this fact, in a BHF we consider each depth d of the forest as a weak object detector, and like in the ARF model [4], the gradient of a loss, for each training sample, can be calculated and exploited to optimize a global loss function over the whole forest in the next stage $d+1$.

Following a Gradient Boosting formulation, at each depth d of the forest, the BHF improves the regressor of the previous iteration $F_{d-1}(\mathcal{P}_t)$, by learning a new model $F_d(\mathcal{P}_t) = F_{d-1}(\mathcal{P}_t) + h_d(\mathcal{P}_t)$, in which the regressor $h_d(\mathcal{P}_t)$ is added.

In our BHF, the forest prediction corresponds to an object center, which lets us compute a relative offset for each patch, *i.e.* $\hat{\mathbf{d}}_\mathbf{t}$. So, following an iterative training procedure, we start with an initial regressor F_0, which corresponds to the N root nodes of the trees. Each iteration d adds a new level of depth to the forest. A regressor $F_{d-1}(\mathcal{P}_t, \phi)$ trained up to $d-1$ gives a prediction for each training patch $\mathcal{P}_t$, and these predictions are use to compute the residuals,

$$r_{td} = d_t - F_{d-1}(\mathcal{P}_t), \tag{1}$$

which will be used to train the base learner $h_d(\mathcal{P}_t)$, *i.e.* the depth d of the forest, according to the original HF formulation [2].

The question is now, how does $F_{d-1}(\mathcal{P}_t, \phi)$ obtain the *weak* prediction $\hat{\mathbf{d}}_\mathbf{t}$ for each patch $\mathcal{P}_t$? Here, in contrast to the ARFs [4], we introduce the concept of *intermediate Hough space*. The BHF interprets each depth of the forest as a stage-wise HF weak object detector.

During training, each training patch $\mathcal{P}_t$ traverses the trees and casts votes to the intermediate Hough space $\mathcal{H} \in \mathbb{R}^2$ based on the location stored in the "leaves" at depth $d-1$. The current object center prediction $\hat{\mathbf{h}}_\mathbf{t}$ of the training patch centered at position y, *i.e.* $\mathcal{P}_t(y)$, can be obtained by finding the local maximum on its corresponding intermediate Hough space $\mathcal{H}$. Using $\hat{\mathbf{h}}_\mathbf{t}$, we can calculate the estimated offset $\hat{\mathbf{d}}_\mathbf{t}$ as $\hat{\mathbf{d}}_\mathbf{t} = y - \hat{\mathbf{h}}_\mathbf{t}$. Finally, this estimation is used to compute the residual using Eq. 1. In Algorithm 1, we summarize the complete training procedure of the BHF.

Algorithm 1 Training a *Boosted Hough Forest*

Require: Labeled training set $\{\mathcal{P}_t, c_t, d_t\}_{t=1}^{T}$
Require: Number of trees N, maximum tree depth D_{max}
1: INIT F_0 using the N root nodes
2: **for** d from 1 to D_{max} **do**
3: Check stopping criteria for all nodes in depth d
4: **for** $\mathcal{P}_t(y)$ from $t = 1$ to T **do**
5: Cast votes in the intermediate Hough space $\mathcal{H} \in \mathbb{R}^2$
6: Find the object center prediction $\hat{\mathbf{h}}_\mathbf{t}$
7: Calculate the estimated offset: $\hat{\mathbf{d}}_\mathbf{t} = y - \hat{\mathbf{h}}_\mathbf{t}$
8: Update the residual r_{td} following Eq. 1
9: **end for**
10: Learn $h_d(\mathcal{P}_t, \varphi_d)$ using the set $\{\mathcal{P}_t, c_t, r_{td}\}_{t=1}^{T}$, and build the level d of the forest
11: **for** $\mathcal{P}_t(y)$ from $t = 1$ to T **do**
12: Propagate $\mathcal{H} \in \mathbb{R}^2$ from parent node to child node in each tree
13: **end for**
14: **end for**

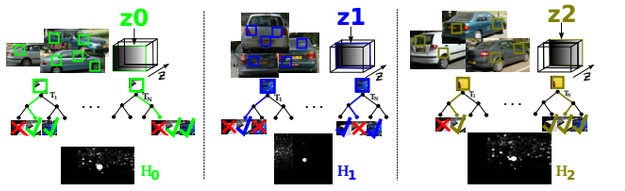
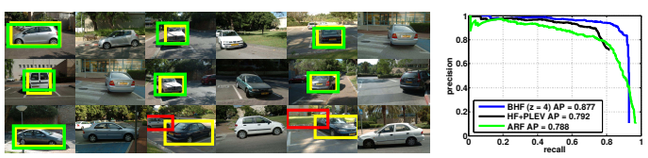


Figure 1: Toy example for 3 aspects of the car category. Our BHF model augments the Hough spaces by adding a dimension $\mathcal{Z}$, which encodes the object aspect. The training patches are passed through the trees to determine a leaf node. Only the patches with the same aspect z in the reached leaf cast probabilistic votes in the corresponding voting space $\mathcal{H}_z$.

In order to further improve the detection performance for multi-aspect objects, we show how the BHF can be naturally extended to deal with this problem. The solution is simple: augment the dimensionality of the Hough voting spaces, with one dimension to encode the different aspects. This allows us to enforce consistency of the votes for each object category aspect separately. For instance, a BHF for detecting cars can be trained to deal with two views (frontal/rear vs. left/right) simultaneously, having a separate Hough voting space per aspect. Nothing changes during the training of the BHF when multiple aspects are integrated: the residual of each training sample is computed considering only its corresponding dimension in the augmented intermediate Hough voting space (see Fig. 1).

We also extend our BHF model for object detection to simultaneously predict the continuous pose of the object. The pose is recovered from the training sample which contributes the most to the final hypothesis.

In our experimental validation, we have observed that the detection performance of our BHF increases with respect to both the traditional HF and the ARF models. We also show how augmenting the dimensionality of the Hough voting spaces, the BHF is able to deal with the problem of multi-aspect object detection and pose estimation. Figure 2 shows qualitative and quantitative results obtained by the BHF on the Weizmann Cars Viewpoint dataset [3]. Additional results using the PASCAL3D+ [5] and some TUD Pedestrian [1] datasets can be found in the paper.



(a) (b)

Figure 2: Results on the Weizmann Cars Viewpoint dataset [3]. (a) Qualitative results. Columns 2,4 and 6 show the training images selected to estimate the pose. Ground truth in yellow, estimations in green and wrong detections in red. (b) Precision-Recall curves for HF, ARF and BHF.

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010.

[2] J. Gall, A. Yao, N. Razavi, L. van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. In *PAMI*, 2011.

[3] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and continuous pose estimation. *IVC*, 2012.

[4] S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof. Alternating regression forests for object detection and pose estimation. In *ICCV*, 2013.

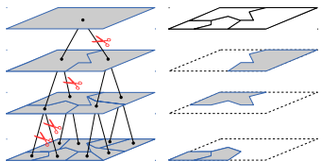
[5] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *WACV*, 2014.

Optimizing Partition Trees for Multi-Object Segmentation with Shape Prior

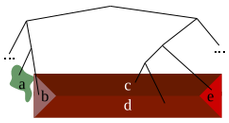
Emmanuel Maggiori[1]
emmanuel.maggiori@inria.fr
Yuliya Tarabalka[1]
yuliya.tarabalka@inria.fr
Guillaume Charpiat[2]
guillaume.charpiat@inria.fr

[1] Inria Sophia Antipolis Méditerranée
Titane team
Sophia Antipolis, France

[2] Inria Saclay
TAO team
Orsay, France



(a) Segmentation as a BPT cut.



(b) Faulty BPT, an object (*bcde*) not represented in a single node.

Figure 1: Binary partition trees.

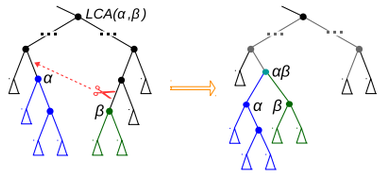
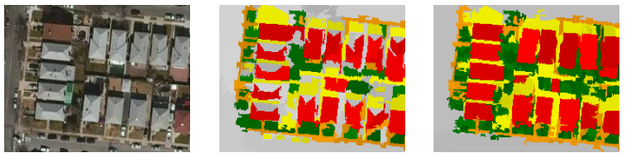
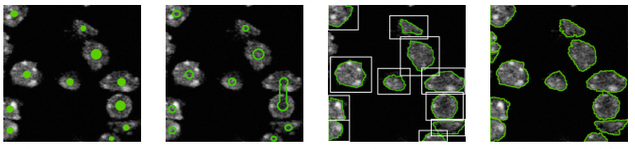


Figure 2: *Prune-and-paste* move to optimize BPTs.



(a) Color image. (b) Unoptimized BPT. (b) Optimized BPT.

Figure 3: Cuts on a BPT with trained color and shape priors.



(a) Input with markers. (b) Method [1]. (c) [1], on user-supplied windows. (d) Optimized BPT.

Figure 4: Convexity prior with [1] and optimized BPTs. [1] must be applied on individual user-supplied windows to account for multiple objects in the scene (b-c). BPT optimization(d) handles multiple objects directly.

The multi-object segmentation of images is one of the great challenges in computer vision. It contemplates the existence of many objects of possibly different classes in the same image. The introduction of **shape descriptors** into segmentation significantly improves its quality. However, it is difficult to optimize energies that involve shape priors because of their non-local nature.

In the last years, hierarchical partitions have been intensively used in different domains. The overall idea is to represent an image as a tree of hierarchical partitions, which is later processed to extract meaningful information. **Binary partition trees** (BPTs) [2] are a particularly efficient structure for this purpose. In our context, multi-object segmentations can be extracted from such trees by performing a cut (see Fig. 1a). Given an energy functional, the optimal cut on a BPT is found in linear time by means of a dynamic programming algorithm.

BPTs can be easily augmented to include shape information, but their traditional bottom-up construction cannot make use of shape constraints, and the tree structure restricts the possible cuts that can be done on them (see e.g., Fig.1b). In this work we exploit the idea of iteratively optimizing the structure of BPTs to produce better partitions with shape priors.

Let $I = (I_j)_{1 \leqslant j \leqslant n}$ be an image containing n pixels. We suppose we are given a set of possible object classes, as well as priors for each class. Multi-label segmentation consists in the partitioning of the pixels into a non-overlapping set of regions $\mathcal{R} = (R_i)$, together with associated class labels $\mathcal{L} = (L_i)$. It can be stated as an optimization problem: minimize

$$E(\mathcal{R},\mathcal{L}) = E_C(I,\mathcal{R},\mathcal{L}) + \sum_{i=1}^{|\mathcal{R}|} E_S(R_i,L_i), \tag{1}$$

where E_C expresses the color prior, and E_S, the shape prior. The color prior is as follows:

$$E_C(I,\mathcal{R},\mathcal{L}) = \sum_{j=1}^{n} -\log P(L_{R(j)}|I_j), \tag{2}$$

where the posteriors $P(L_{R(j)}|I_j)$ can be obtained by training classifiers on pixel color samples. The shape prior term is defined as follows:

$$E_S(R_i,L_i) = -|R_i| \log P(L_i|\mathbf{S}_i), \tag{3}$$

$|R_i|$ being the area of region R_i, and $P(L_i|\mathbf{S}_i)$ being the probability of assigning the label L_i to the region R_i, given a vector $\mathbf{S}_i$ of shape features of that region. The distributions $P(L_i|\mathbf{S}_i)$ are trained from observed data by smoothing feature's histograms.

We wish to enrich the nodes of BPTs by including shape information of the regions. Given that the optimization of the trees will involve recomputing region descriptors, we must design a pool of features that can be computed efficiently from children nodes. We propose to store the **convex hull** of the region at every node. When two regions are merged, the convex hull of the new region can be computed from the children efficiently by using *rotating calipers* [3]. From the convex hull several descriptors can be derived, such as solidity, rectangularity and elongatedness.

To **optimize the BPTs** we follow a local search approach, in which a solution is iteratively modified by performing local transformations on the trees. We propose a move that performs a *prune-and-paste* of a branch into another part of the tree. Fig. 2 illustrates such a move: α is the paste place and β is the pruning place. We denote by $\mathrm{LCA}(\alpha,\beta)$ their lowest common ancestor in the tree. The move creates a new node $\alpha\beta$ in the paste side that comprises α and β. In the pruned side, the tree is collapsed after β is removed. Among the ancestry, only the nodes below $\mathrm{LCA}(\alpha,\beta)$ require to recompute their models (shape and color features), given that further up the regions represented by the nodes do not change.

In our method we simulate all possible moves to compute the impact of each move (ΔC) to the energy in Eq. 1, and construct a priority queue of moves on the values ΔC. In each iteration, we apply the best move, or a number k of independent best moves, prior to reconstructing a new up-to-date queue. When we apply k moves at once, we must verify that the ΔC of a move has not been invalidated by the application of other moves. We have proved theoretical guarantees that allow to reduce the search space of moves, turning the algorithm efficient.

Our experiments validated the effectiveness of our method to make use of shape priors to enhance segmentation, as we show on the satellite image of Fig. 3. We can include different priors simultaneously and the optimization does not depend on which priors were selected. In the case of convexity prior, we compared to [1], as shown on a microscopic image of cell nuclei (Fig. 4). Our method proved then to be an efficient way of introducing shape descriptors in multi-object segmentation.

[1] Lena Gorelick, Olga Veksler, Yuri Boykov, and Claudia Nieuwenhuis. Convexity shape prior for segmentation. In *ECCV*, pages 675–690, 2014.

[2] Philippe Salembier and Luis Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE TIP*, 9(4):561–576, 2000.

[3] Godfried T Toussaint. The rotating calipers: An efficient, multipurpose, computational tool. In *ICCTIM*, pages 215–225, 2014.

Face Painting: querying art with photos

Elliot J. Crowley
elliot@robots.ox.ac.uk
Omkar M. Parkhi
omkar@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

We study the problem of matching photographs of people to similar looking paintings in a large corpus.

Of course, the extent to which a person in a photo resembles a different person in a painting is subjective, and very difficult to quantify. So, instead, we consider the question: given *photographs* of a person, can we retrieve *paintings* of that same person? If we can, we should then be able to find different, but similar looking, people in paintings, starting from a photo.

Initially, one might be skeptical over whether retrieving paintings of a person starting from a photo is achievable. Photographs and paintings have very different low level statistics and to make matters worse, painted portraits are prone to large variations in style: politicians are often highly caricatured, Hollywood icons of the past frequently get the Andy Warhol treatment and are transformed into pop art. This problem is essentially one of domain adaptation from faces in photos to those in paintings; learning how to overcome both the low-level and stylistic differences.

Contributions. (i) we introduce new datasets for learning and evaluating this adaptation problem (figure 1); (ii) we show that, depending on the face representation used, performance can be improved substantially by learning – either by a linear projection matrix common across identities, or by a per-identity classifier (figure 2); (iii) we compare Fisher Vector [6] and Convolutional Neural Network (CNN) representations [5] for this task – some successful painting retrievals from photos using CNN features are provided in figure 3. Finally, (iv) using the learnt descriptors, we show that, given a photo of a person, we are able to find their doppelgänger in a large dataset of oil paintings (figure 4).

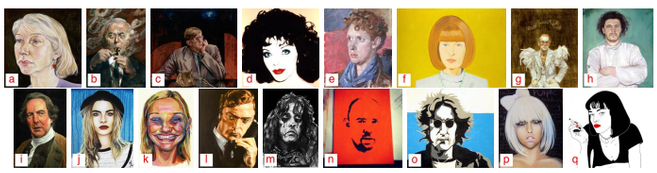


Figure 1: **Painting Datasets.** Top row – Paintings from the National Portrait Gallery, these are largely photo-realistic in nature. Bottom row – Paintings crawled from DeviantArt [2], which are much more varied in style. The contrast between these datasets allows us to observe what effect large variations in style have on retrieval. Identities (a)-(q) are given in the paper.



Figure 2: **Photo-painting pairs used for learning.** In each case the photo is on the left and the painting is on the right. The photos are obtained from Google Image Search and the paintings from DeviantArt. These photos and paintings are used to learn either (i) a linear projection on the features using discriminative dimensionality reduction (an idea first explored for faces in [3]) or (ii) exemplar-like [4] face-specific classifiers to boost retrieval performance.



Figure 3: **Successful retrievals using a CNN representation.** In each case, five photos are used to make a query. These are shown beside the top retrieved painting.



Figure 4: **Celebrities and their doppelgängers.** Each photo is matched to the closest portrait in the 'Your Paintings' dataset of oil paintings [1] using Euclidean distance between CNN features.

[1] BBC – Your Paintings. http://www.bbc.co.uk/arts/yourpaintings/.

[2] DeviantArt. http://www.deviantart.com/.

[3] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *Proc. ICCV*, 2009.

[4] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *Proc. ICCV*, 2011.

[5] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC*, 2015.

[6] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

Cross-Domain Object Recognition Using Object Alignment

Pengcheng Liu
pengcheng.liu@ia.ac.cn
Chong Wang
cwang@nlpr.ia.ac.cn
Peipei Yang
ppyang@nlpr.ia.ac.cn
Kaiqi Huang
kqhuang@nlpr.ia.ac.cn
Tieniu Tan
tnt@nlpr.ia.ac.cn

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences,
Beijing, China

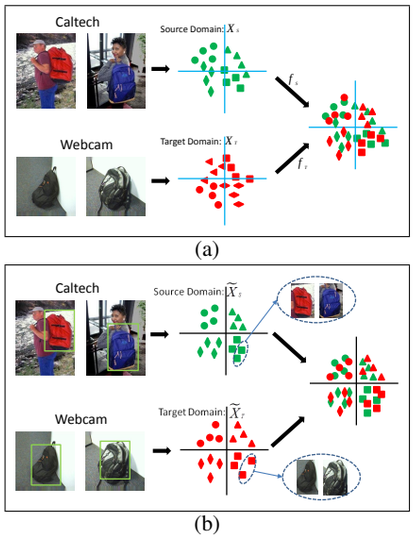


Figure 1: Comparison of methods for a four-class problem on domain adaptation. (a) Existing methods minimize the domain difference based on two ambiguous image feature spaces: X_S and X_T. (b) We minimize the domain difference based on two more similar feature spaces consist of aligned objects: $\widetilde{X}_S$ and $\widetilde{X}_T$. Best viewed in color.

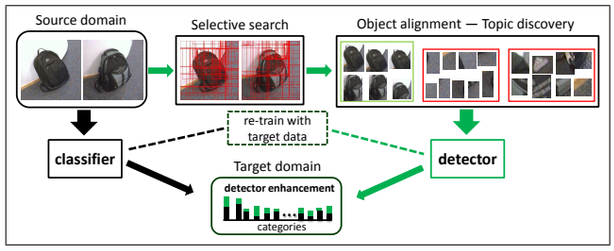


Figure 2: Pipeline of our method. Best viewed in color.

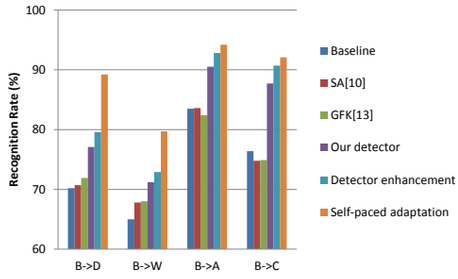


Figure 3: Performance of different methods on Bing-Caltech based cross-domain object recognition. This figure is best viewed in color.

In this paper, we focus on the problem of cross-domain object recognition [4], which has long been one of the challenging problems in computer vision. This problem typically arises when training (source domain) and test (target domain) samples are drawn from different distributions. In the problem of object recognition, this case is usually caused by the situation that training and test samples are acquired under different sets of background, lighting, view point, resolution conditions, etc.

One popular solution to the problem of cross-domain object recognition is minimizing the difference between the source and target distributions. Existing methods are devoted to minimizing that domain difference in a complex image space, which makes the problem hard to solve because of background influence, as shown in Figure 1 (a). Since the object and background are twisted in that image feature space, the discrepancy caused by background is difficult to eliminate, which makes it hard to learn optimal f_S and f_T for minimizing $D(f_S(X_S), f_T(X_T))$. To discount the influence of the background, we propose to minimize that difference using object alignment. As shown in Figure 1 (b), we minimize the domain difference by transferring to the feature space of aligned objects $\widetilde{X}_S$ and $\widetilde{X}_T$, but not the image feature space having background influence.

The key insight of our approach is that the difference between the source and target distributions can be reduced by discounting the influence from the ambiguous background. We define the semantic object as the object that occurs in all the images of one class. To discount the background influence, our primary goal is to automatically localize the semantic object so that the irrelevant background can be eliminated. Then based on the semantic object regions, we can learn an object detector that is robust to the influence of the irrelevant background and makes the cross-domain object recognition much easier than before. In addition, since our detectors are learned in a weakly supervised way, we utilize the classification results from the image feature space to enhance our object detectors to avoid performance degradation. Finally, based on the detectors, a self-paced adaptation method is introduced to link the source and target data to further reduce the domain difference. The pipeline of our method is summarized in Figure 2.

Extensive experiments are implemented to evaluate the effectiveness of our method on cross-domain object recognition. Here, we simply present our experimental results on the data from the 10 common classes among Bing-Caltech [1] (denoted by **B**) and the four domains (denoted by **A**, **D**, **W** and **C**) in Office-Caltech [2, 4]. For each image, we extract deep convolutional features based on an ImageNet pre-trained CNN [3]. We denote a cross-domain image recognition problem by the notation S→ T, where S and T denote the source and target domains respectively. Figure 3 shows a comparison of the results of different methods on the 4 domain pairs. As can be seen, our method (Self-paced adaptation) and its two variants (Our detector and Detector enhancement) show significant improvement compared to other related domain adaptation method, which validates the effectiveness of our method.

[1] Lorenzo Torresani Alessandro Bergamo. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, 2010.

[2] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. *Technical Report*, 2007.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

Dynamical Regularity for Action Analysis

Vinay Venkataraman[1]
http://www.public.asu.edu/~vvenka18/
Ioannis Vlachos[2]
ivlachos@latech.edu
Pavan Turaga[1]
http://www.public.asu.edu/~pturaga/

[1] School of Arts, Media + Engineering,
School of Electrical, Computer and Energy Engineering,
Arizona State University,
Arizona, USA

[2] Center for Biomedical Engineering and Rehabilitation Science,
Louisiana Tech University,
Louisiana, USA

An activity can be seen as a resultant of coordinated movement of body joints and their respective interdependencies to achieve a goal-directed task. This idea is further supported by Johansson's demonstrations that visual perception of the entire human body motion can be represented by a few bright spots which holistically describe the motion of important joints. Traditional dynamical modeling approaches usually operate on the level of individual joints of the human body, lacking any information about the interdependencies between joints. We propose a novel approach for dynamical modeling by extending conventional ideas to quantify the interdependencies between body joints. Towards this end, we propose a new approach – approximate entropy-based feature representation to model the dynamics in human movement by quantifying dynamical *regularity*. In this paper, we utilize the algorithmic framework of [3] for estimating approximate entropy from time series data and extend it to model the dynamics in human activities for applications such as temporal segmentation and fine-grained quality assessment of actions.

Approximate entropy is a statistical tool proposed by Pincus [3, 4] for quantification of regularity of time series data and system complexity, based on the log-likelihood of repetitions of patterns of length m being close within a defined tolerance window that will exhibit similar characteristics as patterns of length $(m+1)$ [2, 3]. It assigns a non-negative number to time series data, with lower values for predictable (ordered) signals and higher values for signals with increased irregularity (or randomness). It is defined using three parameters: embedding dimension (m), radius (r), and time delay (τ). Here, m represents the length of pattern (also called as embedding vector) in the data which is checked for repeatability, τ is selected so that the components of the embedding vector are sufficiently independent, and r is used for the estimation of local probabilities. Given N data samples $\{x_1, x_2, \ldots, x_N\}$, we can define embedding vector $\mathbf{x}(i)$ as,

$$\mathbf{x}(i) = [x_i, x_{i+\tau}, \ldots, x_{i+(m-1)\tau}]^T; \quad for \quad 1 \leq i \leq N-(m-1)\tau. \tag{1a}$$

The frequency of repeatable patterns of the embedding vector within a tolerance r is given by $\mathbf{C}_i^m(r)$ as

$$\mathbf{C}_i^m(r) = \frac{1}{N-(m-1)\tau} \sum_{<j>} \Theta(r - d(\mathbf{x}(i), \mathbf{x}(j))). \tag{1b}$$

Approximate Entropy is given by

$$ApEn(m, r, \tau) = \Phi^m(r) - \Phi^{m+1}(r). \tag{1c}$$

where:

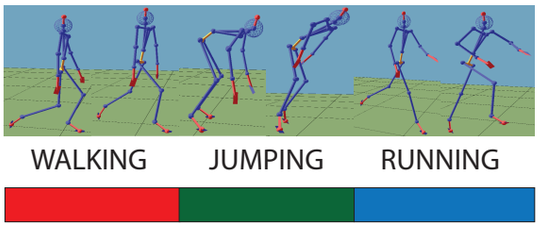
$$\Phi^m(r) = \frac{1}{N-(m-1)\tau} \sum_{i=1}^{N-(m-1)\tau} \ln \mathbf{C}_i^m(r). \tag{1d}$$

$\mathbf{C}_i^m(r)$ represents the frequency of repeatable patterns in the embedding vector $\mathbf{x}(i)$, $\Theta(a)$ is the Heaviside step function, and $\Phi^m(r)$ represents the conditional frequency estimates.

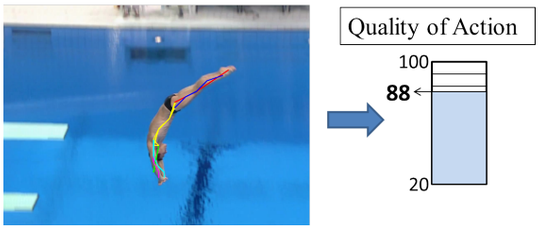
Multivariate Cross Approximate Entropy (XAPEN): Recent theoretical and empirical findings have demonstrated that multivariate embedding of time series data by simple concatenation of individual univariate embedding vectors achieves good state space reconstruction as evaluated by the shape and dynamics distortion measures. In this work, we propose to use the multivariate embedding procedure as described by Cao *et al.* [1] per body joint and estimate the approximate entropy feature representation. In addition, natural human movement involves multiple body joints interacting with each other to together accomplish a particular action task. Hence, it would be beneficial to utilize the cross-coupling information between these joint trajectories. We use cross approximate entropy to quantify the cross-coupling information between body joints, with frequency of repeatable patterns within the embedding vectors $\mathbf{x}_1(i)$ and $\mathbf{x}_2(i)$ generated from two time series data given by $\mathbf{C}_i^m(r)(v||u)$ as

$$\mathbf{C}_i^m(r)(v||u) = \frac{1}{N-(m-1)\tau} \sum_{<j>} \Theta(r - d(\mathbf{x}_1(i), \mathbf{x}_2(j))). \tag{2}$$

We estimate the cross approximate entropy feature across all pairs of body joints after performing multivariate embedding using data available from each body joint. We then concatenate ApEn and XApEn values to form our final approximate entropy-based feature vector to model actions. The principle herein is to quantify regularity (frequency of typical patterns) in the dynamical space computed from trajectories of action data. Our experimental evaluation on theoretical models and two publicly available databases show that the proposed features can achieve state-of-the-art results on applications such as temporal segmentation and quality assessment of actions.



(a) Temporal segmentation of actions using motion capture data.



(b) Quality assessment of diving actions using videos.

Figure 1: A visual representation of our applications of interest in this work. In (a), our aim is to achieve temporal segmentation of actions from continuous untrimmed motion capture data in an unsupervised manner. In (b), we use a supervised learning framework to assess the quality of diving actions from videos.

Acknowledgments: This work was supported in part by National Science Foundation (NSF) CAREER grant 1452163 and NSF 1320267.

[1] Liangyue Cao, Alistair Mees, and Kevin Judd. Dynamics from multivariate time series. *Physica D: Nonlinear Phenomena*, 121(1):75–88, 1998.

[2] Steve Pincus and Burton H Singer. Randomness and degrees of irregularity. *Proceedings of the National Academy of Sciences*, 93(5): 2083–2088, 1996.

[3] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6): 2297–2301, 1991.

[4] Steven M Pincus. Irregularity and asynchrony in biologic network signals. *Methods in enzymology*, 321:149–182, 2000.

Efficient Spatio-Temporal Data Association Using Multidimensional Assignment for Multi-Camera Multi-Target Tracking

Moonsub Byeon
msbyeon@snu.ac.kr
Songhwai Oh
songhwai@snu.ac.kr
Kikyung Kim
koreaton@snu.ac.kr
Haan-Ju Yoo
neohanju@snu.ac.kr
Jin Young Choi
jychoi@snu.ac.kr

Perception and Intelligence Laboratory,
Department of Electrical and Computer Engineering,
Seoul National University, Seoul, Korea

This paper proposes a novel multi-target tracking method which solves a data association problem using images from multi-cameras. In multi-cameras, two combinatorial problems should be solved at the same time: spatial data association between cameras and temporal data association between frames. The *spatio-temporal data association* problem is a well known NP-hard problem even in a small number of cameras or frames (more than 3). Current existing methods [1, 2] simplify the spatio-temporal data association problem through assumption of simple motion model (shortest path) and 3D location estimation. However, the complexity grows exponentially with the number of cameras.

In this work, the spatio-temporal data association problem is formulated as a multidimensional assignment problem (MDA). To achieve a fast, efficient, and easily implementable algorithm, we solve the MDA problem iteratively by solving a sequence of bipartite matching problems using random splitting and merging operations. Hence, the proposed algorithm can be considered as a guided random search to find the global optimum through repeated random local searches (bipartite matchings). In addition, we design a new cost function considering 3D reconstruction accuracy, motion smoothness, visibility from cameras, starting/ending at entrance and exit zone, and false positive. Our approach reconstructs 3D trajectories that represent people's movement as 3D cylinders whose locations are estimated considering all adjacent frames (See Figure 2).

The observations during $1,...,T$ frames and at cameras $1,...K$ form a KT-partite (hyper) graph $G = (V,E) = (\mathbf{I}_{11} \cup ... \cup \mathbf{I}_{KT}, E)$, where vertices V are partitioned into $K \times T$ different independent sets $\mathbf{I}_{11},...,\mathbf{I}_{KT}$ and each hyperedge in E contains at least one vertex in each partite set. Trajectory hypotheses set $\mathbf{T}$ can be defined as a set of all hyperedges E. We represent each trajectory hypothesis $\mathcal{T}_n \in \mathbf{T}$ as a matrix whose entry in the k-th row and t-th column corresponds to an observation index at the t-th frame of the k-the camera. The problem of finding a set of disjoint trajectory hypotheses with a minimum sum of costs can be formulated as the MDA problem which is equivalent to the problem of minimizing the sum of costs of hyperedges containing one element per partite set in the hypergraph G. With binary decision variables $x_{\mathcal{T}_n} \in \{0,1\}$ deciding whether the trajectory $\mathcal{T}_n$ is in the association hypothesis $\mathbf{H}$, and cost function $c : \mathbf{T} \rightarrow \mathbb{R}$, the objective function and disjointness constraints are given by

$$\min \sum_{\mathcal{T}_n \in \mathbf{T}} c(\mathcal{T}_n) x_{\mathcal{T}_n} \quad s.t. \sum_{\mathcal{T}_n \in \mathbf{T}_{[kt],i}} x_{\mathcal{T}_n} = 1 \quad \begin{cases} \mathbf{T}_{[kt],i} = \{\mathcal{T}_n \in \mathbf{T} | [\mathcal{T}_n]_{k,t} = i\} \\ k = 1,2,...,K \\ t = 1,2,...T \\ i = 1,2,...,m_{kt} \end{cases} \tag{1}$$

where $\mathbf{T}_{[kt],i}$ is a subset of all trajectory hypothesis set $\mathbf{T}$ whose value in the k-th row and t-th column has the detection index i and cost $c(\mathcal{T}_n)$ denotes the cost of the n-th trajectory $\mathcal{T}_n$.

Denoting an association hypothesis at $iter$-th iteration as $\mathbf{H}^{iter}$, we can get a new association hypothesis $\mathbf{H}^{iter}$ by improving the previous association hypothesis $\mathbf{H}^{iter-1}$. The key idea is that we randomly split the previous association hypothesis $\mathbf{H}^{iter-1}$ and optimally re-merging by solving the approximated matching problem. Our split/merge strategy is expected to find the solution of which quality is better than or equal to the previous one, because it splits the previous association hypothesis $\mathbf{H}^{iter-1}$ while maintaining the disjointness of the trajectory and re-merges the split trajectories to form a new association hypothesis $\mathbf{H}^{iter}$ with a less than or equal to the previous cost (See an example in Figure 1).

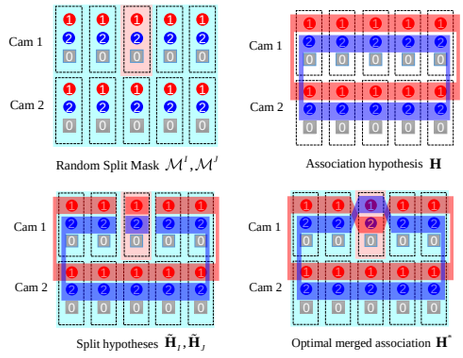


Figure 1: An example of 5 frames and 2 cameras. Our data association strategy is to split and re-merge the previous association hypothesis $\mathbf{H}$ iteratively. Trajectory splitting is done by the Random Split Mask determined randomly at each iteration and merging is done to search the optimal bipartite matching between association hypotheses splits $\tilde{\mathbf{H}}^I$ and $\tilde{\mathbf{H}}^J$. The resulting optimal merged association hypotheses is $\hat{\mathbf{H}}^*$.

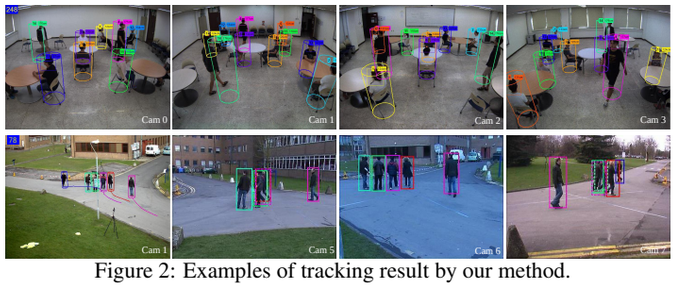


Figure 2: Examples of tracking result by our method.

Our cost function is a summation of five individual terms:

$$c(\mathcal{T}_n) = c_{\mathcal{T}_n} = c_{rec} + c_{mot} + c_{vis} + c_{tse} + c_{fpt}, \tag{2}$$

c_{rec}: measures 3D reconstruction error between a 3D cylinder model and its observations from multi-cameras,
c_{mot}: evaluates how well trajectory describes real motion of a person,
c_{vis}: designed for modeling whether a person is visible at each camera,
c_{tse}: prevents a trivial solution which includes too many track fragments,
c_{fpt}: prevents a trivial association hypothesis where a trajectory hypothesis is considered as a false positive.

The experiments illustrate the proposed method shows the state-of-the-art performance in challenging multi-camera datasets and the computational efficiency with 8 times faster computation than the existing BIP approach. In particular, the proposed high-order motion model and 3D trajectory estimation with 3D cylinder model can reduce the possibility of ID switches.

[1] M. Hofmann, D. Wolf, and G. Rigoll. Hypergraphs for Joint Multi-view Reconstruction and Multi-object Tracking. In *CVPR*, 2013.

[2] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Branch-and-price global optimization for multi-view multi-target tracking. In *CVPR*, 2012.

A New Face Recognition Algorithm based on Dictionary Learning for a Single Training Sample per Person

Yang Liu
yl504@cam.ac.uk
Ian Wassell
ijw24@cam.ac.uk
Computer Laboratory,
University of Cambridge

The number of the training samples per person has a significant impact on face recognition (FR) performance. For the single training sample per person (STSPP) problem, most traditional FR algorithms exhibit performance degradation owing to the limited information available to predict the variance of the query sample. This paper proposes a new method for the STSPP problem in FR, namely the Learn-Generate-Classify (LGC) method. The overall framework of the LGC method is presented in Fig.1.

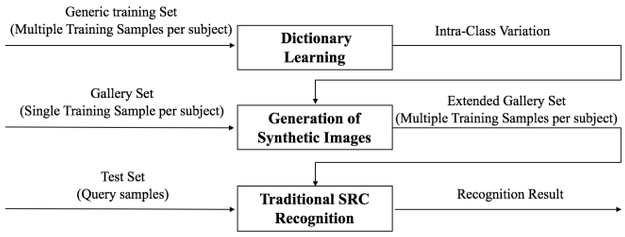


Figure 1: The overall framework of LGC method

The first step is using the dictionary learning method to capture the intra-class variance from the generic set. We assume that the faces of the same subject in different views can be represented as the same one in the discriminant space. For a specific subject, the representation of its samples from different views with respect to the corresponding view-dictionary will be the same. Consequently, for the ith subject, we have the following set of equations

$$\begin{cases} x_{i1} = D_1\alpha_i + e_1 \\ \quad\cdots \\ x_{ip} = D_p\alpha_i + e_p \\ \quad\cdots \\ x_{iN} = D_N\alpha_i + e_N \end{cases}, \tag{1}$$

where the sparse representation α_i is shared among the different view conditions of the subject i. The Dictionary D_p is the corresponding dictionary in the pth view condition. There are a total of N different view conditions and e_p is the residual for the recovered image based on the dictionary D_p and sparse representation α_i . In order to make the learned dictionary both reconstructive and discriminat, the dictionary learning process should add another constraint to encourage the images from the same subject to have similar sparse coefficients and those from different subject to have dissimilar sparse coefficients representations. More specifically, we need to consider a new label consistency constraint, called the 'discriminant sparse-code error' and combine it with the reconstruction error to form a unified objective function, i.e.,

$$arg\min_{D,W,\alpha} \|X - D\alpha\|_2 + \lambda_1 \|Q - W\alpha\|_2 + \lambda_2 \|\alpha\|_1 , \tag{2}$$

where λ_1 controls the relative contribution of the reconstructive term and the discriminant term. The consistently label Q is the ground truth for which dictionary columns should contribute to each of the training images. The matrix W is a linear transformation matrix and the linear transformation $g(W,\alpha) = W\alpha$ transforms the original sparse codes α to the most discriminate sparse feature domain. The objective function Eq.(2) is not jointly convex to D, W, α. Therefore, we solve this problem by breaking it into two sub problems, and alternately update the unknown variables. It involves a sparse coding stage using a pursuit algorithm, followed by an update of the dictionary.

The second step is generation of synthesized images. To generate multiple synthetic samples from the single training sample, two steps are required. The first step is to recover the discriminant representation α_i over the generic training set, which aims to represent the query sample as a linear combination of the images from the generic set. The second step is to synthesize a series of virtual images using α_i and the corresponding view-basis dictionary, which represents an information flow from the discriminant representation back to the observation space.

The third step is using the Sparse Representation Classifier (SRC) framework for classification. The pseudo code of SRC algorithm can be found in [4]. The overall LGC algorithm is summarized in Algorithm 1.

Algorithm 1 Learn-Generate-Classify (LGC) Method

Initialization: D and Q
1: Variation Dictionary Learning
2: **while** not converge **do**
3: Update α in Eq.(2) with fixed Dictionary D and Q
4: Update dictionary D atom by atom via K-SVD Algorithm [1]
5: **end while**
6: Generate Synthetic images
7: Classify the test sample via SRC Algorithm [4]
Output: Classification result

Finally, we verified the effectiveness of the new LGC method on the CMU Multi-pie database, with different illumination, expression and pose variation factors. Table 1 and Table 2 show the recognition rates achieved by the different methods for illumination and pose variation respectively. According to the experimental results, it shows that the LGC method demonstrates a promising FR performance with only a STSPP.

Table 1: Recognition Rates for illumination variation

Session	S2	S3	S4
NN	43.64%	40.23%	38.99%
SVM	43.64%	40.53%	42.50%
SRC	44.66%	38.80%	43.21%
AGL [3]	84.37%	79.51%	79.33%
ESRC [2]	89.25%	84.06%	87.40%
LGC	**90.72%**	**88.42%**	**89.68%**

Table 2: Recognition Rates for pose variation

Session	P05-0-S2	P04-1-S3	P04-1-S4
NN	19.01%	8.99%	6.72%
SVM	18.54%	8.73%	6.72%
SRC	18.64%	9.21%	6.99%
AGL	50.71%	23.58%	19.88%
ESRC	53.95%	29.74%	22.58%
LGC	**56.77%**	**33.74%**	**29.25%**

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[2] Weihong Deng, Jiani Hu, and Jun Guo. Extended SRC: Undersampled face recognition via intraclass variant dictionary. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1864–1870, 2012.

[3] Yu Su, Shiguang Shan, Xilin Chen, and Wen Gao. Adaptive generic learning for face recognition from a single sample per person. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2699–2706. IEEE, 2010.

[4] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2): 210–227, 2009.

Robust Direct Visual Localisation using Normalised Information Distance

Geoffrey Pascoe
gmp@robots.ox.ac.uk
Will Maddern
wm@robots.ox.ac.uk
Paul Newman
pnewman@robots.ox.ac.uk

Mobile Robotics Group
University of Oxford
Oxford, UK

Real-time visual localisation is a key technology enabling mobile location applications [7], virtual and augmented reality [1] and robotics [3]. The recent availablity of low-cost GPU hardware and GPGPU programming has enabled a new class of 'direct' visual localisation methods that make use of every pixel from an input image for tracking and matching [6], in contrast to traditional feature-based methods that only use a subset of the input image. The additional information available to direct methods localising against a dense 3D map increases robustness against typical failure modes for feature-based methods, such as motion blur and viewpoint change [2]. In this paper we present a visual localisation system which utilises the entropy-based Normalised Information Distance (NID) metric for image registration.

For computational reasons, existing direct methods generally minimise a cost function based on photometric error on a per-pixel basis (*e.g.* [6]), which assumes both the live image and the reference map are embedded in the same space. Equation (1) defines such a metric based on the sum of squared differences between corresponding pixels in a reference image (I_r) and a synthetic image (I_s).

$$\mathrm{SSD}(I_r, I_s) = \sum_{\mathbf{x} \in I_r} \| I_r(\mathbf{x}) - I_s(\mathbf{x}) \|^2, \tag{1}$$

where $\mathbf{x} = (u, v)^\mathsf{T}$ is a pixel location within the image.

Although photoconsistency is efficient to compute and find derivatives for (in order to use in an optimisation framework), as mentioned in [6] it suffers from a number of limitations. Principally, it requires I_s provide a photorealistic rendering of the scene, such that the resulting synthetic image matches the reference image I_r on a pixel-by-pixel basis. A true match under significant appearance changes would require modelling of the material and illumination properties of the real-world environment, along with the response of the sensor. This limitation restricts photoconsistency to applications involving frame-to-frame tracking, where the synthetic image I_s can be derived from a warping of the previous reference image I_r, or applications in small indoor environments with controlled illumination where the scene does not change over time [6].

In this paper we instead make use of the Normalised Information Distance (NID) metric, given by Equation (2) [4].

$$\mathrm{NID}(I_r, I_s) = \frac{\mathrm{H}(I_r, I_s) - \mathrm{MI}(I_r; I_s)}{\mathrm{H}(I_r, I_s)} \tag{2}$$

where $\mathrm{H}(I_r, I_s)$ and $\mathrm{MI}(I_r, I_s)$ are the joint entropy and mutual information of the two images respectively, defined as follows:

$$\mathrm{MI}(I_r; I_s) = \mathrm{H}(I_r) + \mathrm{H}(I_s) - \mathrm{H}(I_r, I_s) \tag{3}$$

$$\mathrm{H}(I_s) = -\sum_{b=1}^{n} p_s(b) \log(p_s(b)); \; \mathrm{H}(I_r, I_s) = -\sum_{a,b=1}^{n} p_{r,s}(a,b) \log(p_{r,s}(a,b))$$

where $\mathrm{H}(I_r)$ is defined similarly to $\mathrm{H}(I_s)$. p_s and $p_{r,s}$ are the marginal and joint discrete distributions of the images I_r and I_s, represented by n-bin discrete histograms where a and b are individual bin indices.

As NID is not a function of the actual values of the pixels in the image, but of their distribution, NID provides robustness to illumination change and sensor modality. Unlike mutual information, NID is a true metric, which satisfies the triangle inequality, and is normalised over the number of pixels used in the calculation, thus allowing comparisons between image pairs with differing amounts of overlap.

In order to use NID as a cost function in a gradient-based optimisation framework, we modify the construction of our image histograms to allow the calculation of analytic derivatives. The values in each histogram bin are defined as follows:

$$p_s(b) = \frac{1}{|I_s|} \sum_{\mathbf{x} \in I_s} \beta_s(b, \mathbf{x}), \quad p_{r,s}(a,b) = \frac{1}{|I_s|} \sum_{\mathbf{x} \in I_s} \beta_r(a, \mathbf{x}) \beta_s(b, \mathbf{x}) \tag{4}$$

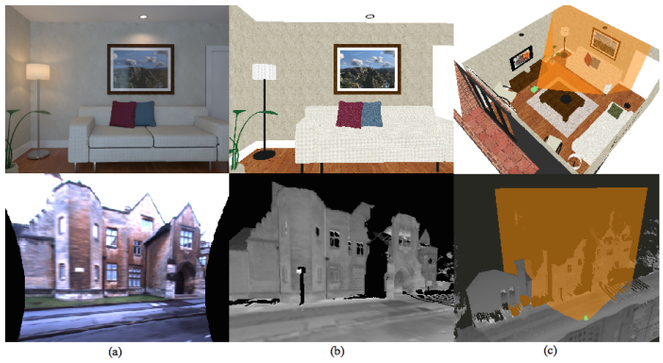


Figure 1: We localise the pose of a camera by registering a rendered image of our prior against the live camera image. Here we show an example from both a synthetic indoor scene and a real-world outdoor scene. (a) Camera image; (b) a render of the prior as used for localisation; and (c) the localised pose and viewing frustrum of the camera within the prior. Our information-theoretic metric is robust to changes in illumination, motion blur, and sensing modality between the live image and prior map.

In a standard histogram, the weighting functions β would take a value of 1 in a single bin, and 0 everywhere else. We instead use the coefficients from a cubic B-spline function as the weighting function, yielding a continuous, twice differentiable function for the value of each histogram bin.

Using a low-fidelity 3D appearance prior of the environment, e.g. from a dense reconstruction [5], CAD model or LIDAR scanner as shown in Figure 1, our method is able to localise a camera under a wide range of conditions, including image under/overexposure, outdoor lighting changes, significant occlusions, motion blur, colour space changes, and differences between image and map modality. We present results showing successful online visual localisation under significant appearance change both in a synthetic indoor environment and outdoors with real-world data from a vehicle-mounted camera.

[1] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, 2007.

[2] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Computer Vision–ECCV 2014*, pages 834–849. Springer, 2014.

[3] Paul Furgale and Timothy D Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.

[4] Ming Li, Xin Chen, Xin Li, Bin Ma, and P M B Vitanyi. The similarity metric. *Information Theory, IEEE Transactions on*, 50(12): 3250–3264, 2004.

[5] Richard A Newcombe and Andrew J Davison. Live dense reconstruction with a single moving camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1498–1505. IEEE, 2010.

[6] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011.

[7] Duncan P Robertson and Roberto Cipolla. An image-based system for urban navigation. In *BMVC*, pages 1–10, 2004.

Joint Clustering and Classification for Multiple Instance Learning

Karan Sikka
ksikka@ucsd.edu
Ritwik Giri
rgiri@ucsd.edu
Marian Bartlett
mbartlett@ucsd.edu

University of California San Diego
La Jolla, California
USA

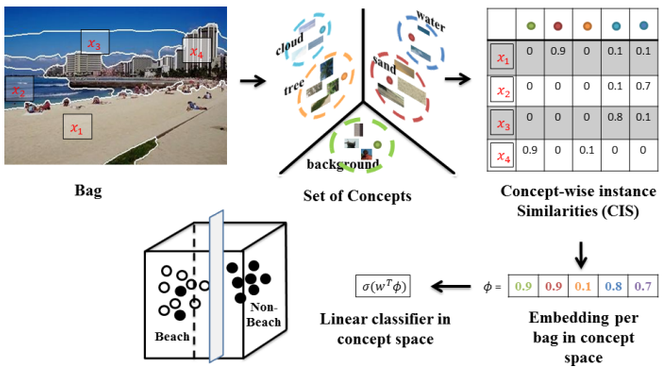


Figure 1: Illustrates the underlying idea in Embedding Space based MIL approaches. A beach scene, segmented into regions, is represented as a bag of instances, where each instance represents the corresponding region. A set of concepts is then used to calculate a similarity between each instance and the concept, referred to as concept-wise instance similarity (CIS). The *max* MIL assumption is used to embed each bag into the concept space using the CIS. Classification can then be performed in the embedding space using standard classifiers (best viewed in color). This work proposes to learn both the concept space and the classifier in a joint fashion.

1 Overview

The **Multiple Instance Learning** (MIL) has been extensively used to solve weakly supervised visual classification problems, where visual data is represented as a bag of instances. Previous MIL algorithms mostly extended standard supervised learning algorithms to the MIL settings by defining the classification score of a bag as the maximum score of each of its instances. Although these algorithms (referred to as Instance Space (IS) based methods) are popularly used, they do not account for the possibility that the instances may have multiple intermediate concepts. On the other hand, Embedding-space (ES) based MIL approaches are able to tackle this issue by defining a set of concepts, and then embedding each bag into a concept space, followed by training a standard classifier in the embedding space.

Contribution. In previous ES based approaches, the concepts were discovered separately from the classifier, and thus were not optimized for the final classification task. This work proposes a novel algorithm to estimate concepts and classifier parameters by jointly optimizing a classification loss. This approach discovers a set of discriminative concepts, which yield superior classification performance. The proposed algorithm is referred to as **Joint Clustering Classification for MIL** (JC^{2}MIL) because the discovered concepts induce clusters of data instances. Moreover, we show that the proposed algorithm achieves state-of-the-art results on several MIL datasets, by discovering fewer number of concepts compared to previous ES-based methods.

2 Methodology

Our target is to jointly learn (1) a set of concepts that are used to embed each bag into a concept space, and (2) a classifier that combines the embedding to produce a classification score. We achieve this by posing the problem as joint minimization of the classification loss with respect to both the set of concepts and the classifier parameters. The model makes two assumptions (1) the probability of a concept lying in a bag is maximum over the probability of each of its instances (similar to IS based methods), and (2) the similarity between k^{th} concept and an instance is defined using the rbf kernel (similar to [2, 3, 5] . The classification loss includes the mean negative log-likelihood and a regularization term, and is written as:

$$\mathcal{L}(B) = -\frac{1}{N}\sum_i (y_i \log p_i + (1-y_i)\log(1-p_i)) + \frac{\lambda}{2} w^T w \quad (1)$$

$$\{C^*, w^*\} = \arg\min_{C,w} \mathcal{L}(B) \quad (2)$$

where C and w are the concept and classifier parameters, and p_i is the classifier score for each bag. The above optimization is solved using coordinate descent.

Method	Corel-2000
MI-SVM [1]	54.6: [53.1 63.1]
MILES [3]	68.7: [67.3 70.1]
DD-SVM [2]	67.5: [66.1 68.9]
k-means-SVM [4]	52.3: [51.6 52.9]
DMIL [5]	70.2: [68.3 72.1]
JC^{2}MIL (Ours)	**73.2**: [71.2 74.8]

Table 1: Evaluation (multiclass % accuracy) of Instance Space and Embedding Space based MIL algorithms on the Corel-2000 dataset [2] along with 95% confidence interval.

3 Evaluation

Experiments: The algorithm was evaluated on five MIL datasets- Tiger, Fox, Elephant, Corel-2000 and UCSB Breast Cancer dataset. The results for the Corel-2000 dataset are shown in Table 1 and extensive results on other datasets can be found in the paper.

Results: The proposed algorithm achieved state-of-the-art results on the MIL datasets. In addition, we showed that, for the Corel-2000 dataset, JC^{2}MIL outperformed its unsupervised variant (similar to BoW) by discovering a (relatively) small number of concepts. The best performance of JC^{2}MIL on Corel-2000 dataset was 73.2% with a concept size of 20, while the best performance for the unsupervised counterpart was 70.1% with a concept size of 3000. Thus joint learning of concepts and classifier leads to excellent results with a smaller number of discriminative concepts.

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 561–568, 2002.

[2] Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004.

[3] Yixin Chen, Jinbo Bi, and James Ze Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[4] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision Workshops*, volume 1, pages 1–2. Springer, 2004.

[5] A. Shrivastava, J.K. Pillai, V.M. Patel, and R. Chellappa. Dictionary-based multiple instance learning. In *IEEE International Conference on Image Processing*, pages 160–164, Oct 2014.

Multi-Modality Feature Transform: An Interactive Image Segmentation Approach

Moustafa Meshry
moustafa.meshry@alexu.edu.eg
Ahmed Taha
ahmed.taha@alexu.edu.eg
Marwan Torki
http://www.eng.alexu.edu.eg/~mtorki/

CSE Dept.
Alexandria University
Alexandria, Egypt

In this paper, we tackle the interactive image segmentation problem. Unlike the regular image segmentation problem, the user provides additional constraints that guide the segmentation process. In some algorithms, like [1, 4], the user provides scribbles on foreground/background (Fg/Bg) regions. In other algorithms, like [6, 8], the user is required to provide a bounding box or an enclosing contour to surround the Fg object, other outside pixels are constrained to be Bg. In our problem, we consider scribbles as the form of user-provided annotation.

Introducing suitable features in the scribble-based Fg/Bg segmentation problem is crucial. In many cases, the object of interest has different regions with different color modalities. The same applies to a non-uniform background. Fg/Bg color modalities can even overlap when the appearance is solely modeled using color spaces like RGB or Lab. Therefore, in this paper, we purposefully discriminate Fg scribbles from Bg scribbles for a better representation. This is achieved by learning a discriminative embedding space from user-provided scribbles. The transformation between the original features and the embedded features is calculated. This transformation is used to project unlabeled features onto the same embedding space. The transformed features are then used in a supervised classification manner to solve the Fg/Bg segmentation problem. We further refine the results using a self-learning strategy, by expanding scribbles and recomputing the embedding and transformations.

Figure 1 illustrates the motivation for this paper. Color features usually cannot capture different modalities available in the scribbles and successfully distinguish Fg from Bg at the same time. As we can see in figure 1(b), the RGB color space will eventually mix Fg/Bg scribbles. On the other hand, figure 1(c) shows that a well-defined embedding space can clearly distinguish between Fg and Bg scribbles, while preserving different color modalities within each scribble.

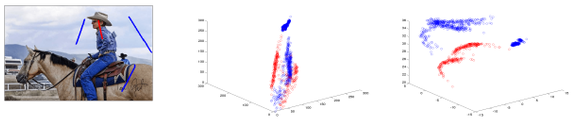


Figure 1: The effect of discriminative embedding. **Left (a)**: Image with provided user scribbles; red for Fg and blue for Bg. **Middle (b)**: 3D plot of the RGB channels for the provided scribbles. The scribbles are mixed in the RGB color space. **Right (c)**: 3D plot of the first 3 dimensions of our discriminative embedding. Color modalities present in the scribbles are preserved. Note that the Fg has two modalities, namely skin color and jeans. Also, the Bg has two modalities: the sky and horse body.

Our contributions in this paper are multifold; **First**, we present a novel representation of image features in the scribble-based Fg/Bg segmentation problem. **Second**, we utilize this representation in two novel interactive segmentation algorithms: ***(i)*** One-pass supervised algorithm, which we extend to ***(ii)*** a self-learning semi-supervised algorithm. **Third**, we present an extensive evaluation on a standard dataset with clear improvements over state-of-the-art algorithms.

The proposed segmentation algorithm learns a discriminative embedding space for the scribbles using a supervised dimensionality reduction technique, like LDA [2, 3] or LFDA [7]. LDA seeks to maximize the between-class separation while minimizing the within-class proximity. LFDA extends LDA by preserving the locality of features that belong to the same class. This is illustrated in figure 1, where Fg has two modalities (skin color and jeans) and Bg also has two modalities (sky and horse body). We then use the learned transformation matrix to transform pixels' color features by projecting them onto the new embedding space. Finally, we classify every pixel as Fg or Bg based on its embedding coordinates. To enhance the classification, we use an iterative version which expands the original scribbles and recomputes the whole pipeline until a stopping criterion is met. A final post processing step is used to remove small islands as done in [5]. Our methods are proved to outperform state-of-the-art algorithms on the standard ISEG dataset [4]. Table 1 shows the segmentation result of different feature embeddings. It is clear that careful embedding can elevate the results significantly. Figure 2 shows qualitative results of our approach.

Transformation Method	Jaccard Index
No transformation	0.549 ± 0.260
1-pass LDA	0.627 ± 0.179
Iterative LDA	0.636 ± 0.180
1-pass LFDA	0.664 ± 0.184
Iterative LFDA	0.678 ± 0.180

Table 1: Segmentation results on ISEG dataset



Figure 2: Qualitative results for 6 out of 151 images. First and third columns show the original image with user scribble annotation. Second and fourth columns show our output.

[1] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *IEEE International Conference on Computer Vision, ICCV*, 2001.

[2] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[3] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.

[4] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] Jiangyu Liu, Jian Sun, and Heung-Yeung Shum. Paint selection. In *ACM Transactions on Graphics (ToG)*, 2009.

[6] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, 2004.

[7] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 905–912. ACM, 2006.

[8] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

Multi-Shot Human Re-Identification Using Adaptive Fisher Discriminant Analysis

Yang Li
yangli625@gmail.com
Ziyan Wu
ziyan@alum.rpi.edu
Srikrishna Karanam
karans3@rpi.edu
Richard J. Radke
rjradke@ecse.rpi.edu

Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY USA

Human re-identification (re-id) in surveillance videos has been widely studied over the past few years. The research effort was mainly focused on building appearance models [1, 3] and learning a suitable metric to determine whether or not an image pair belongs to the same person[5, 8]. In terms of the available number of images per person in each camera view, the re-id problem can be categorized into single- [7] and multi-shot [4] scenarios. In this paper, we propose a novel algorithm based on the Fisher criterion to learn a representative and discriminative feature subspace from image sequences to perform re-id.

From the practical perspective of video surveillance applications, each person should have an image sequence available from tracking algorithms. To re-identify a person, we want to learn a feature space where images belonging to the same person stay close while images belonging to different people are far apart, which can be achieved by Fisher Discriminant Analysis (FDA) [2], defined as

$$J = \mathrm{Tr}\left((\boldsymbol{T}^\top \boldsymbol{S}^w \boldsymbol{T})^{-1} \boldsymbol{T}^\top \boldsymbol{S}^b \boldsymbol{T}\right) \quad (1)$$

where $\boldsymbol{T}$ is the linear tranformation matrix that projects data samples onto a low-dimensional subspace. $\boldsymbol{S}^w$ and $\boldsymbol{S}^b$ are the within-class scatter matrix and between-class scatter matrix respectively.

If we want to describe a person using a number of images, intuitively we need to capture as many different states (i.e., visual aspects) of a person as possible. However, it may be that in an image sequence, a majority of frames of the person are in a single state, which may cause a biased result in the FDA. That is, features reflecting other important states will be ignored. Hence, it is necessary to select representative samples which can cover the diversity of the person by clustering, as shown in Figure 1, before conducting FDA. On the other hand, the performance of clustering depends on the feature space. Only important and discriminative features should be used in clustering. Either conducting clustering before or after FDA, the results will be suboptimal.

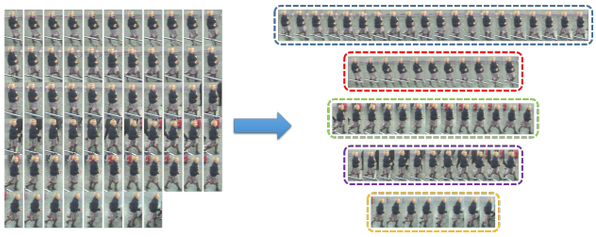


Figure 1: A tracking image sequence is clustered into 5 groups.

In this paper we propose an Adaptive Fisher Discriminant Analysis (AFDA) algorithm to mitigate this issue. We apply the Fisher criterion in two different ways. First, considering the inherent difference between camera views, Local Fisher Discriminant Analysis (LFDA) [6] is adapted to learn a feature subspace that preserves the local structure of each person, while maximizing inter-class distance and minimizing intra-class difference. Then, we hierarchically cluster the sample data from each image sequence under this subspace, again based on the Fisher criterion. New sample points are selected from the clustering results and used to learn a new LFDA subspace. Thus, by iteratively updating the representative samples and the discriminative feature space, the algorithm can determine a better feature subspace before feeding into a metric learning method to establish the relationship between cameras. Finally, since LFDA preserves local structures, a metric learning step is added to compensate for the difference caused by viewpoint variation.

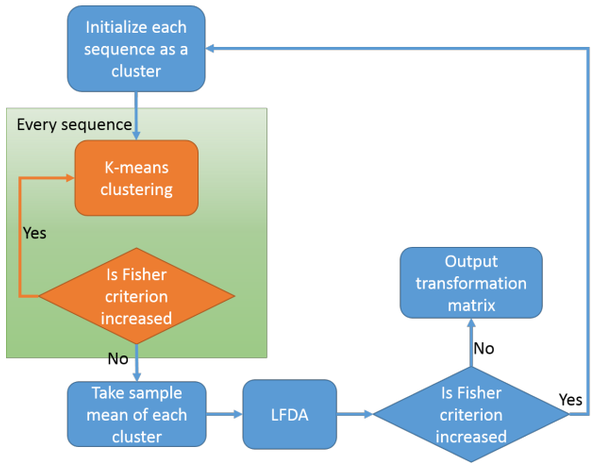


Figure 2: The algorithm flowchart

A flowchart of the proposed algorithm is shown in Figure 2. The proposed algorithm effectively mitigates the problem of biased sample distributions; thus, more discriminative features can be selected and more robust classifiers can be trained. It is tailored for the multi-shot scenario and shows significant improvement compared with state-of-the-art techniques.

[1] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[2] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[3] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[4] Yang Li, Ziyan Wu, and Richard J. Radke. Multi-shot re-identification with random-projection-based random forests. In *WACV*, 2015.

[5] Alexis Mignon and Frédéric Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[6] Masashi Sugiyama. Local Fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, 2006.

[7] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[8] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

Hierarchical-Hybrid Shape Representation for Medical Shapes

Abhishek Kolagunda

Guoyu Lu

Chandra Kambhamettu
https://www.eecis.udel.edu/wiki/vims/

Department of Computer and Information Science
University of Delaware
DE, USA

Recently, shape analysis has become of increasing interest in the medical community due to its potential in capturing the morphological variations across a population. The high quality 3D images captured can be used to extract 3D shape of the organs. 3D models of organs can also be used for training personnel, for visualization during image guided interventions and in simulations. A compact shape model that has implicit and explicit forms will aid in some of these medical use-cases.

We propose a compact hybrid shape model as a combination of Extended Superquadrics (ESQ) [1] and Radial basis interpolation function (RBF). The hybrid shape model in its parametric form is given as $(f(\theta,\phi) = h(\theta,\phi)+g(\theta,\phi))$. h is the extended superquadric function and g is radial basis interpolation function. The points on the surface of the shape are given by

$$(X,Y,Z) = f(\theta,\phi)$$

$$\begin{aligned} X &= a \times sign(cos\theta cos\phi)|cos\theta|^{\varepsilon_2}|cos\phi|^{\varepsilon_1} + r.cos\theta cos\phi \\ Y &= b \times sign(sin\theta cos\phi)|sin\theta|^{\varepsilon_2}|cos\phi|^{\varepsilon_1} + r.sin\theta cos\phi \\ Z &= c \times sign(sin\phi)|sin\phi|^{\varepsilon_1} + r.sin\phi \end{aligned} \quad (1)$$

$$-\pi/2 \le \theta \le \pi/2, -\pi \le \phi \le \pi.$$

Constants a, b and c give the extent of the ESQ. Exponents ε_1 and ε_2 are the shape parameters of the ESQ. The exponents are functions of θ(azimuth) and ϕ(elevation) respectively, modeled using cubic splines. Offset r is a function of both θ and ϕ, modeled by RBF. An ESQ at an arbitrary position also has translation and rotation parameters. The implicit representation of the shape model is given as

$$\begin{aligned} F(x,y,z)^{\frac{\varepsilon_1}{2}} + \sum_{j=1}^{N} w_j \lambda(d_j) = 1 \\ F(x,y,z) = \left[\left(\left|\frac{x}{a}\right|^{2/\varepsilon_2} + \left|\frac{y}{b}\right|^{2/e2} \right]^{\varepsilon_2/\varepsilon_1} + \left|\frac{z}{c}\right|^{2/\varepsilon_1} \end{aligned} \quad (2)$$

λ is a Gaussian with compact support, d_j is the cosine distance from the jth RBF center, w_j is a weight associated with j^{th} RBF center.

To handle concavities in the shape we propose to hierarchically divide the shape into approximately convex parts [2, 3]. The two step process involves first, finding pairs of points, called *mutex pairs*, which cannot belong to the same convex part. Then, finding cutting planes that divide the shape into parts that have no *mutex pairs*. Given a shape S, points $x \in S, y \in S$ are said to be mutually exclusive if there exists a plane p such that x, y lie on p and are disconnected in the contour map produced by the projection of shape S onto p. A weight is associated with each *mutex pair* which is a measure of concavity of the shape between the pair of points. For each *mutex pair* a candidate cutting plane is defined as a plane that bisects the pair and has its normal parallel to the pair. The following heuristics are defined to find the minimum number of cutting planes required to remove all *mutex pairs*: 1) The cut reduces the total non-convexity of the shape; 2) The cut results in parts where at least one is likely to contain no mutex-pairs.

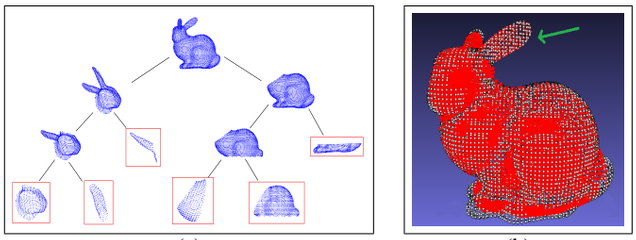
Given a shape as point cloud or a mesh, the method to build the hierarchical-hybrid shape representation is as follows. The input point cloud is hierarchically decomposed into approximately convex parts. The hybrid shape model is then fit to each of parts. LM method for non-linear least square fit is used to minimize the error of fit function derived from the implicit form of the shape model. The error of fit for a point P is defined as the distance from point P to a point Q on the ESQ surface such that $P - O = \beta(Q - O)$ where O is the center of the ESQ. The EoF is given as,

$$EOF = \sum_{i=1}^{Ndata} ||(x_i,y_i,z_i)|| \times |1 - F(x_i,y_i,z_i)^{\varepsilon_1/2}| + C_1 + C_2 + C_3 \quad (3)$$

Additional constraints C_1, C_2 and C_3 on the parameters of ESQ are also added to the EoF.

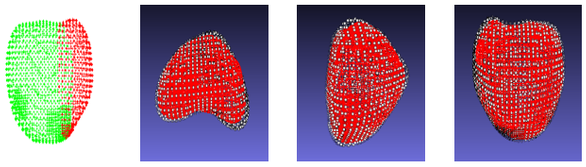
A binary tree is used for the shape representation. Each node in the tree represents a part of the shape. The intermediate nodes contain links to its constituent parts and holds the 2D shape of intersection between its constituent parts. The leaf nodes contain the hybrid shape representation of the part.

To reconstruct the shape, a bottom up approach is employed. The explicit form of the shape model is used to generate points/meshes for each part. The 2D intersection shape is then used to blend the parts together. Fig. 1 and Fig. 2 show fitting of the shape representation to data.



(a) (b)

Figure 1: Fitting our shape model to Stanford Bunny (8454 vertices): (a) Approximate convex decomposition, parts shown inside red box; (b) Shown in *Red* is HSSR fit, ground truth boundary points are shown in *White* and *Green* arrow points to the part that was retained as points due to large error of fit of the shape model.



(a) (b) (c) (d)

Figure 2: Shape representation for Prostate shape. (a) Prostate split into approximately convex parts; (b), (c) and (d) are axial, sagittal and coronal views of prostate. Red represents the mesh generated form HSSR. Shown in White are the ground truth organ boundary points

[1] Lin Zhou, and Chandra. Kambhamettu. "Extending superquadrics with exponent functions: modeling and reconstruction", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999.

[2] Hairong Liu, Wenyu Liu, Latecki, L.J., "Convex shape decomposition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[3] Zhou Ren, Junsong Yuan, Chunyuan Li, Wenyu Liu, "Minimum near-convex decomposition for robust shape representation," IEEE International Conference on Computer Vision (ICCV), 2011.

Data Separation of ℓ_1-minimization for Real-time Motion Detection

Yu Liu[1]
jasonyuliu@nudt.edu.cn
Huaxin Xiao[1]
xiaohuaxin@nudt.edu.cn
Zheng Zhang[1]
zhangzheng@nudt.edu.cn
Wei Xu[1]
xuwei@nudt.edu.cn
Maojun Zhang[1]
mjzhang@nudt.edu.cn
Jianguo Zhang[2]
jgzhang@computing.dundee.ac.uk

[1] College of Information System and Mangement
National University of Defense Technology
Changsha, China

[2] School of Computing
University of Dundee
Dundee DD1 4HN, U.K.

The ℓ_1-minimization (ℓ_1-min) used to seek the sparse solution restricts the applicability of compressed sensing. In this study, We use the existing ℓ_1-min algorithms as a pre-process step that converts the iterative optimization into linear addition and multiplication operations. This paper then proposes a data separation algorithm with computationally efficient strategies to achieve real-time performance of sparse-based motion detection.

For a given signal $y \in \mathbb{R}^m$, the proposed approach separates y as the linear combination of the basis vectors e_i, which is defined as atoms in this paper:

$$y = \gamma_1 e_1 + \cdots + \gamma_i e_i + \cdots \gamma_m e_m, \tag{1}$$

where γ_i is the coefficient of y over the basis vectors e_i, and it could be computed easily by simply projecting y on a basic vector e_i, i.e., the i-th coordinate of y. For example, the image vector in this work is separated to the smallest atoms e_i which contains one non-zero standard element as:

$$e_i = \left[0, 0, \cdots, \underset{i}{1}, \cdots, 0, 0\right]^T. \tag{2}$$

The signal is not limited to the atom used in this work, but can be separated into a variety of patterns of atoms. Each e_i can be considered as the observed signal and convert original ℓ_1-min problem P_1 as follows:

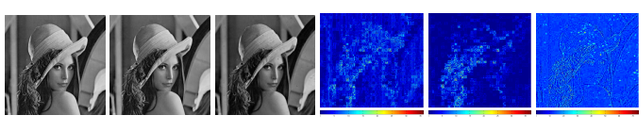
$$P_1^{e_i}: \quad \hat{\boldsymbol{\beta}}_i = \arg\min \|\boldsymbol{\beta}_i\|_1 \qquad s.j. \;\; D\boldsymbol{\beta}_i = e_i, \tag{3}$$

where $\boldsymbol{\beta}_i$ is defined as the children sparse vector in this paper.

Inspired by [2], we assume that the sparse solution α of y can be separated into the linear combination of its children sparse vector $\boldsymbol{\beta}_i$ as follows:

$$\alpha \approx \gamma_1 \boldsymbol{\beta}_1 + \cdots + \gamma_i \boldsymbol{\beta}_i + \cdots \gamma_m \boldsymbol{\beta}_m. \tag{4}$$

For the motion detection problem, we can pre-solve the children sparse vector $\boldsymbol{\beta}_i$ with the existing ℓ_1-min algorithms. Then, the sparse solution of a new signal or image y can be rapidly obtained by Eq. (4). Then, the iterative process in the existing ℓ_1 algorithms is replaced by addition and multiplication operations.



(a) (b) (c) (d) (e) (f)

Figure 1: (a)-(c): The recovered Lena image (256×256) by proposed Data separation, Bregman iterative [3], and Lasso [1]. Their execution time of recovery are 0.12s, 100.16s, and 4.18s, respectively. (d)-(e): The recovered error by proposed Data separation, Bregman iterative [3], and Lasso [1]. The percentage of recovered error in pixel level is 0.0331, 0.0176 and 0.1317 respectively.

Another important question is about the numerical distance of the sparse solution between the use of the classic ℓ_1-min and data separation algorithm. The distance is acceptable for many applications, (e.g., motion detection), but not for others, (e.g., image deblurring). If tolerable in a specific work, distance can be used as acceleration engine, which can dramatically improve applicability. The numerical distance can be visually represented as the recovered error in Fig. 1. Although the recovered results are not the best, the proposed data separation approach can significantly accelerate the ℓ_1-min.

We construct the image atoms e_i in Eq. (2) as the same size as the background I_B. The Bregman iterative algorithm [3] is employed to calculate the children sparse vector $\boldsymbol{\beta}_i$ of e_i. Consequently, the background model based on sparse representation can be rewritten as follows:

$$I_B = D \times \alpha \approx D \times \sum_i \gamma_i \boldsymbol{\beta}_i, \tag{5}$$

where γ_i is linear coefficients of the background model I_B over the atom e_i.

Detection requires a high level of consideration to enhance the robustness because the dynamic textures or complicated environment can affect the corrupted frame I. In this work, we project the current frame I over the pre-learned dictionary D and compute the sparse codes α' with the data separation algorithm. Then, the formula is converted as follows:

$$\begin{aligned} P_F^i &= D\alpha' - D\alpha \approx D \times \sum_j \gamma_j'^i \boldsymbol{\beta}_j - D \times \sum_j \gamma_j^i \boldsymbol{\beta}_j \\ &= D \times \sum_j (\gamma_j'^i - \gamma_j^i) \boldsymbol{\beta}_j, \end{aligned} \tag{6}$$

where γ_i' is linear coefficients of current frame I over the atom e_i.

Accordingly, we compare each patch to decide whether the frame belongs to the foreground through the distribution of the sparse coefficients.

$$\begin{cases} \Delta_1 = \left\| \sum_i \gamma_i' \beta_i - \sum_i \gamma_i \beta_i \right\|_1, \\ \Delta_2 = \left| \left\| \sum_i \gamma_i' \beta_i \right\|_0 - \left\| \sum_i \gamma_i \beta_i \right\|_0 \right|, \end{cases} \tag{7}$$

where Δ_1 and Δ_2 are the differences of sparse coefficients distributions and values between the background model and the current frame.

To obtain a more precision result, we post-process the differences of the sparse coefficients as shown in follow:

$$\Delta = \lambda_1 \Delta_1 + \lambda_2 \Delta_2, \tag{8}$$

where λ_1 and λ_2 are the unitary parameters which determine the weight of Δ_1 and Δ_2 respectively.

This work is at very preliminary stages. How the signal separated into basic atoms has remained an open question. A satisfactory result can be obtained in separating the signal even with the use of the simplest method as demonstrated by this work. Another future work is to measure the numerical difference of the sparse solution with or without using data separation. The difference is acceptable for motion detection, but it does not mean it can be used for other applications. Thus, mathematically defining this difference is able to decide the potential of the data separation algorithm.

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[2] Y. C. Eldar and G. Kutyniok (Eds.). *Compressed sensing: theory and applications. Chapter 11*. Cambridge University Press, ISBN: 9781107005587, 2012.

[3] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for compressed sensingand related problems. *SIAM J. Imaging Sciences*, 1(1):143–168, 2008.

Color Constancy by Deep Learning

Zhongyu Lou
z.lou@uva.nl
Theo Gevers
Th.Gevers@uva.nl
Ninghang Hu
n.hu@uva.nl
Marcel P Lucassen
marcel@lucr.nl

Intelligent System Laboratory Amsterdam, University of Amsterdam,
Science Park 904, 1098 XH Amsterdam, The Netherlands

We propose a new framework using Deep Neural Networks (DNNs) to obtain accurate light source estimation. Color constancy is formulated as a DNN-based regression approach to estimate the color of the light source.

Traditional color constancy algorithms are constrained to a number of imaging assumptions or shallow learning models based on hand-crafted features. In this paper, instead of using traditional feature representations, we exploit deep learning architectures by means of Convolutional Neural Networks (CNN). Different from existing methods which rely on predefined low-level features, we propose to use CNNs to learn feature hierarchies to achieve robust color constancy. The deep CNN model used in this work consists of eight (hidden) layers. Such a deep model will yield multi-scale image features composed of pixels, edges, object parts and object models. The proposed deep learning approach needs large amounts of data with ground-truth for training. Unfortunately, there are no such datasets available for color constancy. Therefore, we propose a different training approach. Firstly, we propose a new training strategy which contains three steps to learn hierarchical features for color constancy. Secondly, we propose a method to generate more training images with ground-truth labels.

The model used in this work consists of eight layers as defined in [3]. The first five layers are convolutional layers. The last three layers are fully collected layers. Combining all the layers, the total number of parameters in this model is very large (around 60M). Therefore, large scale datasets with ground-truth light source labels are required to directly apply this model to the color constancy problem. However, such large scale dataset are not available. To this end, we propose an alternative training procedure consisting of different training steps in the following section. Further, we proposed a data augmentation method to generate more training images with ground truth labels.

The training contains three steps. In the first step, we train the model on ImageNet to derive features for object description. In this way, a rich and generic feature hierarchy is learnt to capture the complex visual patterns in generic, real-world images. The last layer of the model is replaced by a 1000 dimensional vector. The soft-max loss function is used for training. The aim of the first training step is to obtain a pre-trained feature model representing general images. Since the ImageNet dataset contains 1000 object categories, it provides abundant back-propagation information for training. We denote the parameters obtained by training on ImageNet as the $Net1$ network.

In the second step of training, the aim is to retrain and adjust the parameters of $Net1$ for the purpose of color constancy. We perform retraining of the (initial) parameters of $Net1$ based on light source estimation obtained by existing color constancy algorithms as labels. Although any other or combination of color constancy algorithms can be used to generate the labels for the ImageNet dataset, the gray-shades algorithm is used due to its efficiency and good performance. Specifically, we use $Net1$ as the initial weights and retrain the network on ImageNet dataset with light source labels generated by gray-shades. The resulting sets of parameters are denoted by $Net2$. The obtained feature representations $Net2$ are merely adopted (color constancy) versions of $Net1$. It is hypothesized that the obtained models $Net2$ replicate the performance of the color constancy methods used to provide the labels i.e. the gray-shades algorithm.

In the third step of training, the parameters of the model obtained in the previous step are retrained using existing (publicly available) datasets with (real) ground-truth label sets (e.g. Grayball [1] and ColorChecker [2]). In these datasets, the ground-truth color of the light source is given for each image under which it has been recorded. We use $Net2$ as initial parameters to retrain the model with a Euclidian loss. The parameters of this network obtained after this step is the final model, denoted by $Net3$.

Deep learning approaches greatly benefit from large training datasets. Since ground-truth labels are expensive to get (especially for the color constancy problem), data augmentation is widely exploited by different deep learning approaches. For example, in image denoising, extra training data is generated by applying simulated noise. In deep learning-based image classification, Krizhevsky et.al [3] use image translations and horizontal reflections to generate more training images.

Inspired by these data augmentation methods, in this paper, we use data augmentation to obtain more training data for color constancy. Specifically, for each training image, we correct for the color of the light source using the diagonal model of the ground-truth. Using Eq. 1 ,

$$\begin{pmatrix} R_u \\ G_u \\ B_u \end{pmatrix} = \begin{pmatrix} e_R & 0 & 0 \\ 0 & e_G & 0 \\ 0 & 0 & e_B \end{pmatrix} \begin{pmatrix} R_c \\ G_c \\ B_c \end{pmatrix}, \quad (1)$$

the canonical image is obtained. Then, simulated light sources can be applied to the corrected image using the diagonal model in Eq. 1. Any simulated light source color can be used i.e. the Spectral Power Distribution (SPD) of different light sources such as tungsten halogen, fluorescent lamp, high pressure sodium, or daylight. However, many of them are less frequently present than others depending on the scenes from which the images are recorded. Therefore, in this paper, simulated light sources are derived from the training dataset. By clustering the ground-truth light color of the training set into k clusters, we obtain k simulated light sources by collecting the means of each cluster. In this way, per image, k additional training images are obtained with different ground-truth illuminant ($k = 10$ in our experiments). Note that data augmentation is only performed on the training images. We test the algorithm on the original testing images. We denote the proposed model trained on the extended Grayball dataset as $NetFinal$. Fig. 1 shows a number of images obtained by the proposed data augmentation method. Experiments show that the proposed algorithm ($NetFinal$) outperforms the state-of-the-art by 9%. Especially in cross dataset validation, our approach reduces the median angular error by 35%.



(a) GT(0°) (b) 4.14° (c) 6.80° (d) 11.22° (e) 22.78°

Figure 1: Data augmentation. (a) canonical image obtained by removing the effect of the light source color using Eq. 1. (b-e) images generated by applying simulated light sources using Eq. 1. The numbers are the angles between the simulated light sources and the canonical (white) light source.

[1] F. Ciurea and B. Funt. A large image database for color constancy research. *Proceedings of the Eleventh Color Imaging Conference*, 2003.

[2] P.V. Gehler, C. Rother, A. Blake, T. Sharp, and T. Minka. Bayesian color constancy revisited. *CVPR*, 2008.

[3] A. Krizhevsky, I. Krizhevsky, and G. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

Model-based 3D Hand Tracking with on-line Hand Shape Adaptation

Alexandros Makris
amakris@ics.forth.gr
Antonis Argyros
argyros@ics.forth.gr

Institute of Computer Science,
FORTH, Heraklion, Greece
Computer Science Department,
University of Crete,
Heraklion, Greece

3D hand tracking is an interesting problem with high complexity due to the high dimensionality of the human hand and its frequent and often severe self occlusions. Most of the curent methods, track a human hand under the assumption that its parameters (e.g., finger lengths, palm dimensions, e.t.c.) are already known. This assumption limits the applicability of tracking methods. Recently, a few approaches that attempt to solve the hand shape estimation problem have been proposed [5, 6].

In this work we present an on-line method that solves simultaneously the hand tracking and hand shape estimation problems. Let $\mathbf{x}_t$ and $\mathbf{y}_t$ be the pose and shape state at time step t respectively. For the pose estimation the Bayesian Hierarchical Model Framework (HMF) is employed [3, 4]. The framework uses six *auxiliary models* that lie in lower dimensional spaces as proposals for the 26-DOF *main model* of the hand. The shape estimate at each time step is provided by per-frame shape parameters optimization, followed by a robust fitting framework. The per-frame optimizer generates possible shape proposals $\mathbf{y}_t^{pso}$ by optimizing the shape parameters at each frame given fixed (i.e., already estimated) pose parameters. Since the actual shape parameters are constant, the robust fitting cross-validates the shape proposals over a frame history. The output of the fitting is the best estimate given the considered history of the shape parameters $\bar{\mathbf{y}}_t$ that is used in the subsequent frame by the pose tracker.

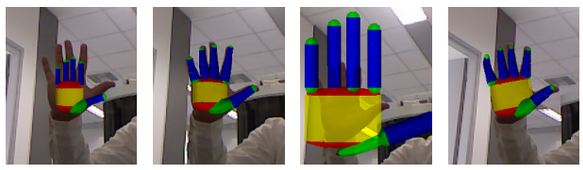
Hand Model The shape of the hand $\mathbf{y}_t$ is parametrized by an 11D vector that controls finger lengths and widths and the width and height of the palm. The pose of the hand $\mathbf{x}_t$ is parametrized by a 27D vector. The kinematics of each finger are modeled using four parameters, two for the base angles and two for the remaining joints. The global position of the hand is represented by a fixed point on the palm and the global orientation.

Pose Tracking The HMF tracking framework [3, 4] that is used to track the hand pose updates at each frame t the pose parameters $\mathbf{x}_t$ given the estimate of the shape parameters $\bar{\mathbf{y}}_{t-1}$. The HMF uses several auxiliary models that are able to provide information for the state of the main model which is to be estimated. Each of the auxiliary models tracks a distinct part of the hand; we use one for the palm with 6-DOF for its 3D position and orientation and one for each finger with 4-DOF for the joint angles. A particle filter is used to sequentially updates the sub-states.

Shape Optimization At each time step t the particle filter described above maintains a set of N weighted particles for the main model. An optimization of the shape parameters using the PSO algorithm is performed independently for the $N^{pso} << N$ particles with the higher weights resulting in N^{pso} updated estimates for the shape parameters paired with the corresponding pose parameters. The likelihood of these pairs is calculated and the shape parameters with the max-likelihood $\mathbf{y}_t^{pso}$ are retained as the current shape estimate.

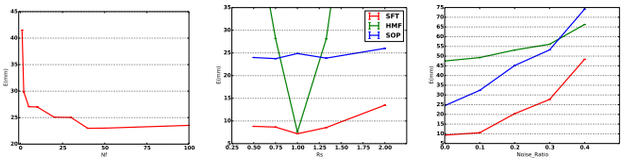
Shape Fitting The per-frame shape estimates up to the current frame are processed by a robust fitting framework. The framework stores a history of N_f frames along with their corresponding poses $H_F = \{\mathbf{z}_f, \bar{\mathbf{x}}_f\}_{f=1}^{N_f}$, and a history of N_s shape parameters $H_S = \{\mathbf{y}_s^{pso}\}_{s=1}^{N_s}$. Every shape $\mathbf{y}_s$ in history H_S is paired with every pose $\bar{\mathbf{x}}_f$ in history H_f. The likelihood $L([\bar{\mathbf{x}}_f, \mathbf{y}_s^{pso}], \mathbf{z}_f)$ of each pair is evaluated and the shape parameters are ranked according to that likelihood. The per-frame ranks $R_f(\mathbf{x}_f, \mathbf{y}_s)$ of each shape parameter set $\mathbf{y}_s^{pso}$ are then averaged to obtain the global rank for the set $R(\mathbf{y}_s)$. The new estimate for the shape parameters is selected by choosing the estimate with the best average rank among the history frames.

Experiments We used real data obtained by RGB-D sensors to qualitatively evaluate the methods and synthetic data for quantitative evaluations. The methods that have been included in our comparative evaluation are: (i) **HMF**: The method of [4] that tracks a hand without estimating its shape. (ii) **SOP**: Tracking the hand through HMF and perform only shape optimization per frame. (iii) **SFT**: The full proposed method. The synthetic dataset that we used for the evaluation consists of 1400 frames of free hand movement. For the shape initialization we test different parameter sets that are scaled with respect to the groundtruth shape by a ratio R_s. We test values for R_s from 0.5 to 2. The pose error E_p measures the average distance between corresponding phalanx endpoints over a sequence similarly to [2]. We are also interested in assessing the performance of the proposed method with respect to noise in the observations. We simulate the imperfect data that come from a real depth sensor, the imperfect foreground detection and the possible occlusions from unknown objects. The results show that the shape estimate converges fast, and this significantly improves the overall tracking accuracy. Test runs on real data are provided in: `https://youtu.be/4dgwoKkDSn8`.



(a) (b) (c) (d)

Figure 1: **SFT** Tracked sequences examples. Two sets of two frames of the same sequence tracked with two different shape initializations. Figures (a) and (c) show the initialization while (b) and (d) the pose/shape estimation several frames later.



(a) (b) (c)

Figure 2: **Quantitative Experiments (a)** Pose error for various maximum history frame values N_f. **(b)** Pose error for various shape initializations. The initialization ratios (x-axis) express the ratio between the shape parameters values that were used for initialization and the ground truth shape parameters. **(c)** Pose error for various sequence noise levels.

Acknowledgments: This work was supported by the EU IST-FP7-IP-288533 project RoboHow.Cog.

[1] I. Albrecht, J. Haber, and H.P. Seidel. Construction and animation of anatomically based human hand models. In SCA, 2003.

[2] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009.

[3] A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. A hierarchical feature fusion framework for adaptive visual tracking. *Image and Vision Computing*, 29(9):594–606, 2011.

[4] A. Makris, N. Kyriazis, and A. Argyros. Hierarchical Particle Filtering for 3d Hand Tracking. In *CVPR, HANDS Workshop*, 2015.

[5] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, Robust, and Flexible Real-time Hand Tracking. In CHI, NY, USA, 2015. ACM.

[6] J. Taylor, R. Stebbing, V. Ramakrishna, C. Keskin, J. Shotton, S. Izadi, A. Hertzmann, and A. Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *CVPR*, 2014.

Globally Optimal DLS Method for PnP Problem with Cayley parameterization

Gaku Nakano
g-nakano@cq.jp.nec.com
Information and Media Processing Labs, NEC Corporation
Kawasaki, Japan

The perspective-n-point (PnP) problem, which estimates 3D rotation and translation of a calibrated camera from n pairs of known 3D points and corresponding 2D image points, is a classical problem but still fundamental in the computer vision community. It is well studied that the PnP problem can be solved by at least three points [1]. If $n \geq 4$, the PnP problem becomes a nonlinear problem where the number of the solutions depend on n and the shape of the scene. This paper proposes an efficient, scalable, and globally optimal DLS method parameterized by Cayley representation, which has been regarded as a unsuitable parametrization due to its singularity.

First we derive a new optimality condition without Lagrange multipliers. Letting $\mathbf{p}_i = [x_i, y_i, z_i]^T$ be an i-th 3D point and $\mathbf{m}_i = [u_i, v_i, 1]^T$ be the corresponding calibrated image point in homogeneous coordinates, the PnP problem can be formulated as a nonlinear optimization

$$\begin{aligned} \min_{\mathbf{R},\mathbf{t}} \quad & \sum_{i=1}^{n} \left\| [\mathbf{m}_i]_\times (\mathbf{R}\mathbf{p}_i + \mathbf{t}) \right\|^2 \\ \text{s.t.} \quad & \mathbf{R}^T\mathbf{R} = \mathbf{I}, \quad \det(\mathbf{R}) = 1 \end{aligned} \tag{1}$$

where $[\ \ ]_\times$ denotes a matrix representation of the vector cross product. The optimal translation $\mathbf{t}$ can be expressed as a function of $\mathbf{R}$ since Eq. (1) is a linear least squares problem of $\mathbf{t}$. If we define $\mathbf{M} \in \mathbb{R}^{9\times 9}$ by a symmetric coefficient matrix computed from $\mathbf{p}_i$ and $\mathbf{m}_i$, Lagrange function of Eq. (1) can be written as

$$L(\mathbf{R},\mathbf{S},\lambda) = \frac{1}{2}\mathbf{r}^T\mathbf{M}\mathbf{r} - \frac{1}{2}\mathrm{trace}\left(\mathbf{S}(\mathbf{R}^T\mathbf{R} - \mathbf{I})\right) - \lambda(\det(\mathbf{R}) - 1). \tag{2}$$

Here, $\mathbf{r}$ is a vector form of $\mathbf{R}$, λ is a Lagrange multiplier, and $\mathbf{S} \in \mathbb{R}^{3\times 3}$ is a symmetric matrix of Lagrange multipliers. Then, the first-order optimality condition is given by

$$\frac{\partial L}{\partial \mathbf{R}} = \mathrm{mat}(\mathbf{M}\mathbf{r}) - \mathbf{R}\mathbf{S} - \lambda\mathbf{R} = \mathbf{0}, \tag{3}$$

$$\frac{\partial L}{\partial \mathbf{S}} = \mathbf{R}^T\mathbf{R} - \mathbf{I} = \mathbf{0}, \tag{4}$$

$$\frac{\partial L}{\partial \lambda} = \det(\mathbf{R}) - 1 = 0, \tag{5}$$

where $\mathrm{mat}(\ \ )$ is a reshaping operator from a 9×1 vector to a 3×3 square matrix. Multiplying $\mathbf{R}^T$ from the left-hand and the right-hand side of Eq. (3), we have

$$\mathbf{R}^T\mathrm{mat}(\mathbf{M}\mathbf{r}) = \mathbf{S} + \lambda\mathbf{I}, \tag{6}$$

$$\mathrm{mat}(\mathbf{M}\mathbf{r})\mathbf{R}^T = \mathbf{R}(\mathbf{S} + \lambda\mathbf{I})\mathbf{R}^T. \tag{7}$$

Since $\mathbf{S} + \lambda\mathbf{I}$ is a symmetric matrix, the left-hand side of Eqs. (6) and (7) must be symmetric matrices. Hence, we obtain a new optimality condition where the Lagrange multipliers are eliminated:

$$\mathbf{P} = \mathbf{R}^T\mathrm{mat}(\mathbf{M}\mathbf{r}) - \mathrm{mat}(\mathbf{M}\mathbf{r})^T\mathbf{R} = \mathbf{0}, \tag{8}$$

$$\mathbf{Q} = \mathrm{mat}(\mathbf{M}\mathbf{r})\mathbf{R}^T - \mathbf{R}\mathrm{mat}(\mathbf{M}\mathbf{r})^T = \mathbf{0}. \tag{9}$$

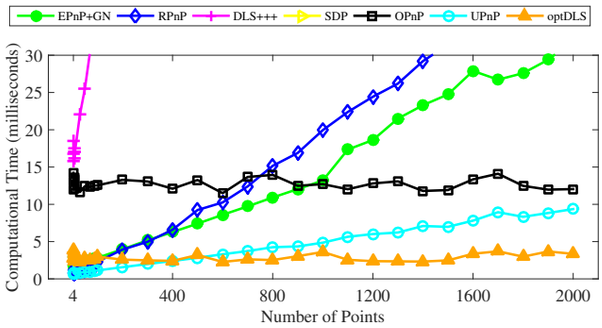


Figure 1: Computational time for varying $4 \leq n \leq 2000$ and fixed $\sigma = 2.0$

Let $P_{j,k}$ and $Q_{j,k}$ be the (j,k) element of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Obviously the diagonal elements are zeros, $P_{j,j} = Q_{j,j} = 0$. On the other hand, the non-diagonal elements are second degree polynomials in $\mathbf{R}$. Due to the symmetry, $P_{j,k} = P_{k,j}$ and $Q_{j,k} = Q_{k,j}$, we have six polynomials in total:

$$\begin{aligned} P_{1,2} = 0, \quad P_{1,3} = 0, \quad P_{2,3} = 0, \\ Q_{1,2} = 0, \quad Q_{1,3} = 0, \quad Q_{2,3} = 0. \end{aligned} \tag{10}$$

Although Eq. (10) is derived from a general rotation parameterization, any parameterizations satisfying Eqs. (4), (5), and (10) are usable for solving the PnP problem. By using Kukelova *et al.*'s automatic Gröbner basis solver, we compared three types of parameterizations: general rotation matrix, quaternion, and Cayley parameterization. Table 1 is a comparison of the above parameterizations with existing methods. For efficiency and stability, this paper selects Cayley parameterization whose elimination template and action matrix are the smallest among the three representations.

The proposed method was evaluated on synthetic data with existing methods in terms of robustness against image noise and computational time. While the proposed method has the same robustness as the state-of-the-art, OPnP [5], the computational time is less than 3 msec for almost all cases. As shown in Fig. 1, it is the fastest especially for $n \geq 400$ points.

The conclusion is that the new optimality condition without the Lagrange multipliers can be solved by any types of rotation parameterizations. Furthermore, Cayley parameterization is suitable for realtime applications, such as augmented reality and visual SLAM, where hundreds or thousands of the points is not a rare situation.

[1] B. M. Haralick *et al.* Review and analysis of solutions of the three point perspective pose estimation problem, *INT J COMPUT VISION*, 13(3):331–356, 1994.

[2] J. A. Hesch and S. I. Roumeliotis. A direct least-squares (DLS) method for PnP. In *ICCV*, 2011.

[3] L. Kneip *et al.* UPnP: An Optimal O (n) Solution to the Absolute Pose Problem with Universal Applicability. In *ECCV*, 2014.

[4] Z. Kukelova *et al.* Automatic generator of minimal problem solvers. In *ECCV*, 2008.

[5] Y. Zheng *et al.* Revisiting the pnp problem: A fast, general and optimal solution. In *ICCV*, 2013.

Table 1: Comparison of rotation parameterizations

	Existing methods			Proposed params		
	DLS [2] (Cayley)	OPnP [5] (Non-unit Quat.)	UPnP [3] (Quaternion)	Rotation Matrix	Quaternion	Cayley
# of unknowns	3	4	4	9	4	3
# of equations	3	4	8	27	7	6
# of solutions	27	40	8	40	80	40
singularity	yes	no	no	no	no	yes
elim. templ.	120×120	348×376	141×149	1936×1976	630×710	124×164
action matrix	27×27	40×40	8×8	40×40	80×80	40×40

Joint Tracking and Event Analysis for Carried Object Detection

Aryana Tavanai
fy06at@leeds.ac.uk
Muralikrishna Sridhar
scms@leeds.ac.uk
Eris Chinellato
e.chinellato@leeds.ac.uk
Anthony G. Cohn
a.g.cohn@leeds.ac.uk
David C. Hogg
d.c.hogg@leeds.ac.uk

School of Computing
University of Leeds
Leeds, UK

Motivation This paper investigates whether reasoning about events that objects may participate in can facilitate tracking. For example, in domains where objects are frequently carried, dropped and exchanged, the trajectories of objects and the people carrying them have prototypical spatio-temporal relationships. In this work, we exploit this relationship introduced by events in order to improve the outcome of object tracking. We particularly focus on event-based tracking for domains which tend to contain generic objects, for which it is not straightforward to train class specific object detectors. The task of object tracking becomes especially challenging in such domains, as false and missing detections are highly prevalent [4]. This leads to false tracks, and also tracks that are heavily fragmented. We observe this phenomenon when we apply state-of-the-art trackers [1, 3, 4]. Furthermore, the prevalence of false and fragmented tracks makes the goal of incorporating events particularly challenging as event based tracking is a circular problem. This problem arises as it involves inferring events using reasonable tracks, and then using the events to subsequently improve the tracks.

Approach Given a set of tracklets, the aim is to exploit the spatio-temporal structure of tracks induced by events in order to improve object tracking. More specifically, we exploit learned domain-specific temporal transitions between events in order to target true positive tracklets, which we then use to form meaningful whole tracks. Our approach is illustrated in Fig. 1 and is performed iteratively in an optimisation similar to [2]. Our framework is not constrained to using any particular tracker to build the set of tracklets and, in principle, can be applied to any tracker. To illustrate this point, we have applied our framework to three state-of-the-art trackers [1, 3, 4].

Evaluation We have evaluated our approach in terms of both tracking and event recognition which is performed on a newly created MINDSEYE2015 dataset (Fig. 2). This dataset contains a large number of events representing the changing relation between objects and people, captured from three different viewpoints. We have defined seven events (Carry, Static, Pickup, Putdown, Drop, Raise, Roll) to allow for a full description of the scene with regards to the state of the carried object, from the start of its appearance to its disappearance. Ground truth for person tracks, carried object tracks and events are fully annotated. This dataset has been made publicly available with all ground truth annotations, carried object detections, tracklets and final tracks at http://doi.org/10.5518/9. Relevant code including our carried object detector [4] can be found at:
http://www.engineering.leeds.ac.uk/joint-tracking-and-event-analysis.
We are currently extending our approach to include multi-person, multi-object events such as giving, exchanging or replacing objects.

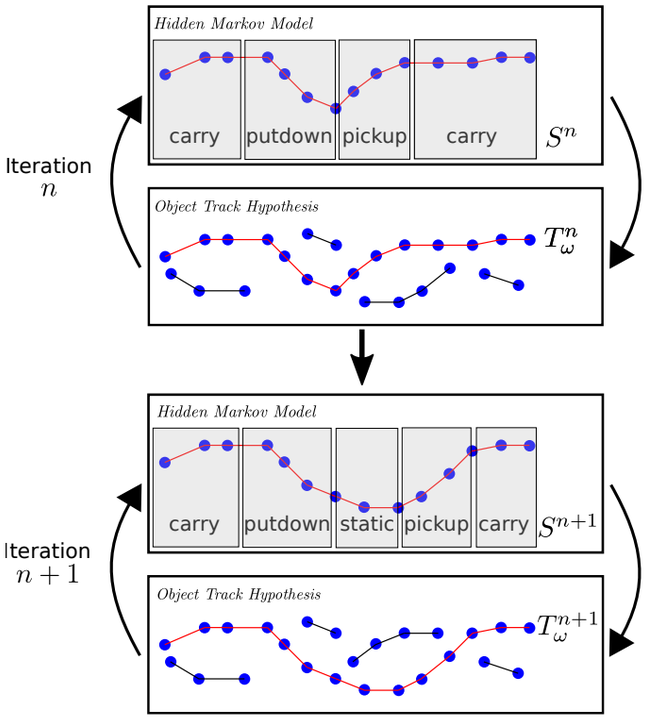


Figure 1: We illustrate two consecutive iterations within the Joint Tracking and Event Analysis optimisation. Given a set of tracklets $\mathcal{T}$ (on the left), at each iteration, a temporally-disjoint subset ω is selected and a contiguous track T_ω is produced by linearly interpolating across any gaps. The Viterbi path S^* of event labels in the HMM is inferred from T_ω (arrow up), leading to an HMM measure (arrow down) and combined with the spatio-temporal factors to give an overall probability. In the next iteration, a change to the subset ω is made and the overall probability re-computed. In this case, the new configuration is accepted since the probability is increased.



(a)

Figure 2: Sample images from the MINDSEYE2015 dataset illustrating various viewpoints, people (cyan), objects (red) and events (yellow text).

[1] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR*, pages 1926–1933, 2012.

[2] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for multi-target tracking. *Transactions on Automatic Control,*, 54(3):481–497, March 2009.

[3] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, pages 1201–1208, 2011.

[4] A. Tavanai, M. Sridhar, F. Gu, A. G. Cohn, and D. C. Hogg. Carried object detection and tracking using geometric shape models and spatio-temporal consistency. In *Computer Vision Systems*, volume 7963 of *LNCS*, pages 223–233. Springer, 2013.

Occlusion-Aware Object Localization, Segmentation and Pose Estimation

Samarth Brahmbhatt
http://www.cc.gatech.edu/~sbrahmbh
Heni Ben Amor
http://henibenamor.weebly.com
Henrik Christensen
http://www.cc.gatech.edu/~hic

School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA



Figure 1: Object localization and segmentation example. Left: Image, Right: Refined mask from SD-HOP

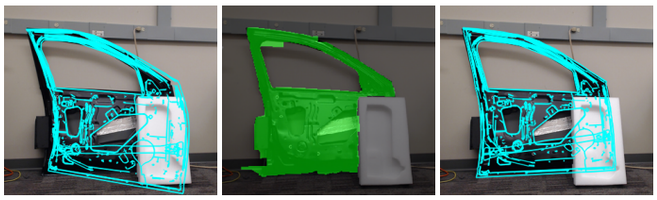


Figure 2: 3D pose estimation. Left to right: Pose estimation with IRLS, SD-HOP refined segmentation, Pose estimation with OR-IRLS.

We present a learning approach for localization and segmentation of objects in an image in a manner that is robust to partial occlusion. Our algorithm, *Segmentation and Detection using Higher-Order Potentials* (SD-HOP) produces a bounding box around the full extent of the object and labels pixels in its interior that belong to the object. This is different from semantic segmentation, which does not provide information about the spatial position of labelled pixels inside the object.

A common theme in the literature is to model occlusion geometrically or appearance-wise, thereby allowing it to contribute to the detection process. The former often make simplifying assumptions about occluder and scene geometry. Our appearance-based approach avoids these assumptions and performs better than existing appearance-based approaches due to the use of higher-order potentials for modelling neighbour influence and a loss function that targets both localization and segmentation.

SD-HOP discriminatively learns HOG templates for objects and occlusion. Whereas the object templates model the objects of interest, the occlusion templates provide discriminative support and do not model a specific occluder. Segmentation is done by considering the response of patches to these templates, and influence of neighbouring patches through a CRF with higher-order connections. The training phase requires a set of images with different occlusions of the object(s) of interest. Each training sample is (1) over-segmented and (2) annotated with a bounding box around the full extent of the object and a binary segmentation of the area inside the box into object vs. non-object pixels. Given these, we train a structured Support Vector Machine (SVM) that learns the HOG templates and CRF weights. Object segmentation is done by assigning binary labels to HOG cells within the bounding box, 1 for visible and 0 for occluded. Neighbour influence for segmentation can take two forms: (1) pairwise terms that impose a cost for 4-connected neighbours to have different labels and (2) higher-order potentials that impose a cost for cells to have a different label than the dominant label in their segment of the image. These segments are produced separately by an unsupervised segmentation algorithm.

The label for an object in an image $\mathbf{x}$ is represented as $\mathbf{y} = (\mathbf{p}, \mathbf{v}, a)$, where $\mathbf{p}$ is the bounding box, $\mathbf{v}$ is a vector of binary variables indicating the visibility of HOG cells within $\mathbf{p}$ and $a \in [1, A]$ indexes the discrete viewpoint. $\mathbf{p} = (p_x, p_y, p_\sigma)$ indicates the position of the top left corner and the level in a scale-space pyramid. The width and height of the box are fixed per viewpoint as w_a and h_a HOG cells respectively. Hence $\mathbf{v}$ has $w_a \cdot h_a$ elements. Given a labelled image, a sparse joint feature vector $\Psi(\mathbf{x}, \mathbf{y})$ is formed by stacking A vectors, each corresponding to a different discretized viewpoint. These vectors consist of vectorized HOG features and visibility labels of cells, count of cells in $\mathbf{p}$ that lie outside the image boundary, statistics of visibility agreement between 4-connected neighbouring cells and cells in the same unsupervised segment, and a constant bias. All vectors except for the one corresponding to viewpoint a are zeroed out.

Learning involves determining linear weights $\mathbf{w}$ such that the score $\mathbf{w}^T\Psi(\mathbf{x}_i, \mathbf{y}_i)$ of any ground truth labelled image $\mathbf{x}_i$ must be smaller than the score $\mathbf{w}^T\Psi(\mathbf{x}_i, \hat{\mathbf{y}}_i)$ of any other labelling $\hat{\mathbf{y}}_i$ by the distance between the two labellings $\Delta(\mathbf{y}_i, \hat{\mathbf{y}}_i)$ minus the slack variable ξ_i, where $\|\mathbf{w}\|_2$ and ξ_i are minimized. Hence we learn $\mathbf{w}$ by solving the following constrained Quadratic Program

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|_2 + C\sum_{i=1}^{N} \xi_i \tag{1}$$

$$\text{s.t.} \quad \mathbf{w}^T(\Psi(\mathbf{x}_i, \hat{\mathbf{y}}_i) - \Psi(\mathbf{x}_i, \mathbf{y}_i)) + \xi_i \geq \Delta(\mathbf{y}_i, \hat{\mathbf{y}}_i) \ \forall i, \hat{\mathbf{y}} \in Y_i$$
$$\xi_i \geq 0 \ \forall i$$
$$\mathbf{D}^2\mathbf{w} \geq \mathbf{0}$$

$\mathbf{D}^2$ is a second order curvature constraint on the K + 1 weights for the higher-order potentials, which forces them to make a concave lower envelope. Training is performed by using the cutting plane training algorithm of [3], with adaptation for training higher-order potentials as described in [2]. The loss function between two labels $\mathbf{y}$ and $\hat{\mathbf{y}}$ depends on the amount of overlap between the two bounding boxes and the Hamming distance between the visibility labellings

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = \left(1 - \frac{\text{area}(\mathbf{p} \bigcap \hat{\mathbf{p}})}{\text{area}(\mathbf{p} \bigcup \hat{\mathbf{p}})}\right) + \frac{\text{area}(\mathbf{p} \bigcap \hat{\mathbf{p}})}{\text{area}(\mathbf{p} \bigcup \hat{\mathbf{p}})} \cdot H(\mathbf{v}, \hat{\mathbf{v}}) \tag{2}$$

Inference is performed by finding the labelling that minimizes the dot-product energy: $\mathbf{y}^* = \text{argmin}_{\mathbf{y}}\, \mathbf{w}^T\Psi(\mathbf{x}, \mathbf{y})$. Due to the linear parametrization of energy and decomposability of the loss function over the unary terms, inference is efficient. At every bounding box location in a pyramid, it is performed by a single $s-t$ mincut on a graph constructed as described in [1] and [2].

We implemented SD-HOP in Matlab, with MVC search and inference implemented in CUDA since they are massively parallel problems. Inference on a 640x480 image with 11 scales takes 3s for a single object with a single viewpoint on our 3.4 GHz CPU and NVIDIA GT-730 GPU. SD-HOP achieves 13.52% segmentation error and 0.81 area under the false-positive per image vs. recall curve on average over the challenging CMU Kitchen Occlusion Dataset. This is a 42.44% decrease in segmentation error and a 16.13% increase in localization performance compared to the state-of-the-art. Figure 1 shows a sample output on this dataset.

We demonstrate that the segmentation output of SD-HOP can be used to ignore edges produced by occlusion, thereby making model-based 3D pose estimation robust to partial occlusion as shown in Figure 2

[1] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 2004.

[2] Stephen Gould. Max-margin learning for lower linear envelope potentials in binary markov random fields. In *ICML*, 2011.

[3] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 2009.

VESICLE: Volumetric Evaluation of Synaptic Interfaces using Computer Vision at Large Scale

William Gray Roncal[1,2] · wgr@jhu.edu
Michael Pekala[2] · mike.pekala@jhuapl.edu
Verena Kaynig-Fittkau[3] · vkaynig@seas.harvard.edu
Dean M Kleissas[2] · dean.kleissas@jhuapl.edu
Joshua T Vogelstein[4,5] · jovo@jhu.edu
Hanspeter Pfister[3] · pfister@seas.harvard.edu
Randal Burns[1] · randal@cs.jhu.edu
R Jacob Vogelstein[2] · jacobv@gmail.com
Mark A Chevillet[2] · mark.chevillet@jhuapl.edu
Gregory D Hager[1] · hager@cs.jhu.edu

[1] Johns Hopkins University, Baltimore, Maryland
Department of Computer Science
[2] JHU Applied Physics Laboratory, Laurel, Maryland
Research and Exploratory Development
[3] Harvard University, Cambridge, Massachusetts
School of Engineering and Applied Sciences
[4] Johns Hopkins University, Baltimore, Maryland
Department of Biomedical Engineering,
[5] Johns Hopkins University, Baltimore, Maryland
Institute for Computational Medicine

An open challenge at the forefront of modern neuroscience is to obtain a comprehensive mapping of the neural pathways that underlie human brain function; an enhanced understanding of the wiring diagram of the brain promises to lead to new breakthroughs in diagnosing and treating neurological disorders. Inferring brain structure from image data, such as that obtained via electron microscopy (EM), entails solving the problem of identifying biological structures in large data volumes. Synapses, which are a key communication structure in the brain, are particularly difficult to detect due to their small size and limited contrast. Prior work in automated synapse detection has relied upon time-intensive, error-prone biological preparations (isotropic slicing, post-staining) in order to simplify the problem.

This paper presents VESICLE, the first known approach designed for mammalian synapse detection in anisotropic, non-poststained data. Our methods explicitly leverage biological context, and the results exceed existing synapse detection methods in terms of accuracy and scalability. We provide two different approaches - a deep learning classifier (VESICLE-CNN) and a lightweight Random Forest approach (VESICLE-RF), to offer alternatives in the performance-scalability space. Addressing this synapse detection challenge enables the analysis of high-throughput imaging that is soon expected to produce petabytes of data, and provides tools for more rapid estimation of brain-graphs. Finally, to facilitate community efforts, we developed tools for large-scale object detection, and demonstrated this framework to find $\approx$ 50,000 synapses in 60,000 μm^3 (220 GB on disk) of electron microscopy data.

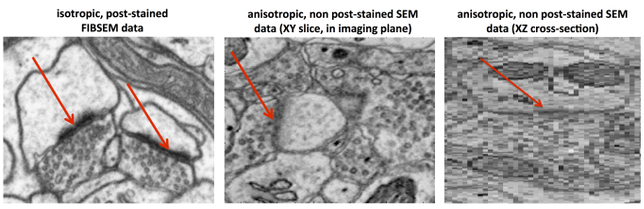


Figure 1: Previous work on synapse detection has focused on isotropic post-stained data (left), which shows crisp membranes and dark fuzzy post synaptic densities (arrows) from all orientations. The alternative imaging technique of non post-stained, anisotropic data (middle, right) promises higher throughput, lack of staining artifacts, reduction in lost slices, and less demanding data storage requirements - all critically important for high-throughput connectomics. The XZ plane of a synapse in anisotropic data is shown (right), illustrating the effect of lower resolution. We address this more challenging environment, in which membranes appear fuzzier and are harder to distinguish from synaptic contacts. Data courtesy of Graham Knott (left) and Jeff Lichtman (middle, right).

In Figure 1, we provide examples demonstrating the challenging setting of our detection task. Figure 2 illustrates the importance of leveraging contextual information, especially neurotransmitter-containing vesicles and cell membranes. Finally, Figure 3 presents a summary of classifier performance; visualizations of the results are not shown here, but are available in the full paper and on our website. Our code and data are open source and available at: `openconnecto.me/vesicle`.

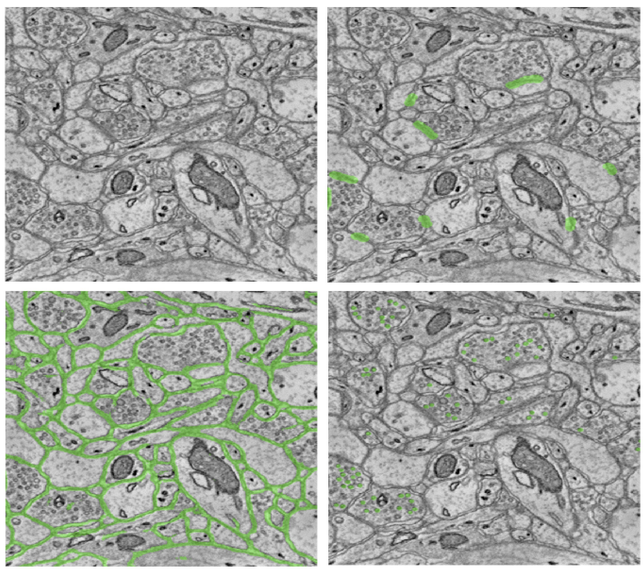


Figure 2: A single cross-section of EM data is shown (upper left). The detection task is to identify synapses shown in green (upper right). These synapses are known to exist at the interface of two neurons; these boundaries can be approximated by previously computed membranes, allowing us to restrict the evaluation regions to the green pixels (lower left). Clusters of vesicles are a good indicator of an axonal bouton, suggesting that one or more synaptic sites is likely nearby. Vesicles found by our automated detection step are highlighted in green (lower right).

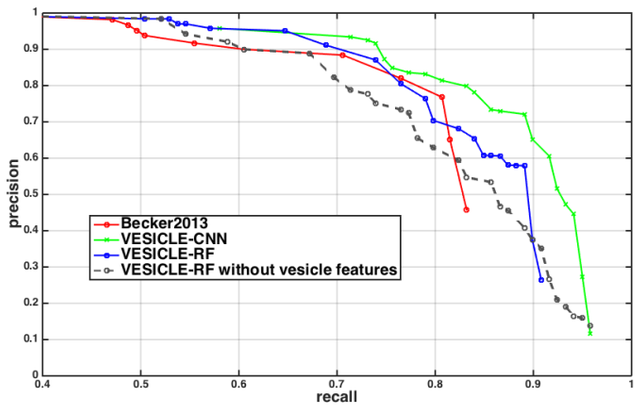


Figure 3: VESICLE-RF and VESICLE-CNN significantly outperform prior state-of-the art, particularly at high recall rates. The relatively abrupt endpoint of the Becker2013 method occurs because beyond this point, thresholded probabilities group into large detected regions rather than individual synapses, which are disallowed.

Epipolar Consistency in Fluoroscopy for Image-Based Tracking

André Aichert[1]
andre.aichert@cs.fau.de
Jian Wang[1]
jian.wang@cs.fau.de
Roman Schaffert[1]
roman.schaffert@fau.de
Arnd Dörfler[2]
arnd.doerfler@uk-erlangen.de
Joachim Hornegger[1]
joachim.hornegger@cs.fau.de
Andreas Maier[13]
andreas.maier@cs.fau.de

[1] Pattern Recognition Lab
Friedrich-Alexander Universität
Erlangen-Nürnberg, Germany
[2] Department of Neuroradiology
Universitätsklinikum Erlangen
Erlangen, Germany
[3] Graduate School in Advanced Optical Technologies (SAOT),
Erlangen, Germany

Geometry and physics of absorption imaging impose certain constraints on X-ray projections. Various consistency conditions have been derived in the Computed Tomography (CT) community under the assumption of specific scanner trajectories. They can be exploited, for example, in truncation and motion correction. Recently, the Epipolar Consistency Conditions (ECC) [1] have been derived from the epipolar geometry between two arbitrary flat-panel projections. The fact that they make no assumption on trajectory enables us to apply the ECC to patient tracking in fluoroscopy for interventional radiology. For the first time, we perform 3D tracking of an object solely based on the consistency between 2D X-ray shots. The core idea is that before an intervention under fluoroscopy, between two and five reference images of the region of interest are acquired from different angles. We suggest, that an optimization of the ECC of an unseen image with respect to a few reference images enables us determine a rigid 3D pose of an object or patient.

The ECC are based on the observation, that certain lines in X-ray images contain redundant information. Intuitively, integrating over a line in a projection image is *roughly* the same as integrating over the corresponding plane $\mathbf{E}$ of absorption coefficients through the object. The plane $\mathbf{E}$ is the plane through the source position which intersects the detector in the respective line. For the projection matrices $\mathbf{P}_0, \mathbf{P}_1 \in \mathbb{R}^{3\times4}$, the epipolar lines $\mathbf{l}_0, \mathbf{l}_1 \in \mathbb{P}^2$ are special lines in the two images, whose corresponding epipolar plane $\mathbf{E}$ is the same [2]:

$$\mathbf{E} \cong \mathbf{P}_0^\top \mathbf{l}_0 \cong \mathbf{P}_1^\top \mathbf{l}_1. \tag{1}$$

There are two redundant ways to compute the plane integral over $\mathbf{E}$ from either $\mathbf{l}_0$ or $\mathbf{l}_1$, respectively. Epipolar Consistency can thus be quantified by taking several epipolar planes and measuring the difference between these redundant line integrals. Let $\rho_I(\mathbf{l}) = \rho_I(\alpha,t)$ denote the Radon transform of the projection image I at line $\mathbf{l}$ of angle α to the image u-axis and distance t to the origin. The Epipolar Consistency Conditions state

$$\frac{d}{dt}\rho_{I_0}(\mathbf{l}_0) - \frac{d}{dt}\rho_{I_1}(\mathbf{l}_1) \approx 0 \tag{2}$$

where the derivative $\frac{d}{dt}$ in direction of line normals accounts for non-parallel ray geometries [1].

In this work, we assume to have a fluoroscopic video of a patient and a set of reference projections of known projection geometry. Patient movement results in inconsistencies of the fluoroscopic image I_0 with respect to the reference projections. Our goal is to estimate patient motion by minimizing these inconsistencies. A 6-DOF rigid motion is represented by the parameter vector ϕ. For each reference image I_i we definite a metric over a set of epipolar planes $\mathcal{E} = \{\mathbf{E}_1, \ldots, \mathbf{E}_N\}$

$$M_{0/1} = \frac{1}{N}\sum_{\mathbf{E}_k\in\mathcal{E}} \left(\frac{d}{dt}\rho_{I_0}\left(\mathbf{P}_0^{+\top}\mathbf{E}_k\right) - \frac{d}{dt}\rho_{I_i}\left(\mathbf{P}_i^{+\top}\mathbf{E}_k\right)\right)^2, \tag{3}$$

which should be minimized over patient motion ϕ.

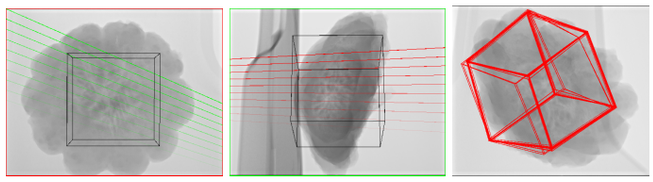


Figure 1: Pumpkin phantom. Left, Center: Two reference X-ray images. Epipolar lines between the images are shown in the color corresponding to the frame. Right: Visualization of 25 iterations of a random study of ±10° and ±25 mm showing two outliers.

There are several assumptions involved in the ECC. Among those, first, an absorption-only model for X-ray physics is assumed when computing the "ray-sums", which neglects scatter and other non-linear effects of intensity. Second, the method is based on line-integrals in the projection images. Information is therefore obtained only in an orthogonal direction to these lines. It is essential, that reference views are chosen, which ideally produce sets of orthogonal epipolar lines in the input image. Third, and most notably, we present an image-based method. This means, that the object seen by the input and reference views must actually show the same part of the same object in all views.

In our experiments we show, that in spite of these restrictions, it is possible to determine 3D motion based solely on a small set of 2D reference images. We evaluate the proposed method against 2D-3D registration with a previously acquired CT. Our experiments on synthetic data based on a patient CT and pumpkin phantom data from an interventional C-arm (compare Fig. 1) demonstrate that our method is able to compensate for rotations of up to ±10° and translations of ±25 mm between consecutive frames in real time. We successfully track rotations of as much as 45° over 45 images. We show, that small rotations of about 3° in space and translations of about 50 mm can be tracked based on just two reference X-ray images.

The outstanding property of the approach is that no 3D scan is required for tracking a 3D object in space. Since the proposed approach works directly on X-ray images, it exceeds regular 2D-3D registration with a CT in an order of magnitude in computational speed. We conclude that ECC are a simple and effective new tool for pre-aligment and online patient tracking for fluoroscopic sequences.

[1] André Aichert, Martin Berger, Jian Wang, Nicole Maass, Arnd Doerfler, Joachim Hornegger, and Andreas Maier. Epipolar Consistency in Transmission Imaging. *IEEE Trans Med Imaging*, Apr 2015. (`http://dx.doi.org/10.1109/TMI.2015.2426417`).

[2] Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

Automatic Aortic Root Segmentation with Shape Constraints and Mesh Regularisation

Robert Ieuan Palmer

Xianghua Xie

Gary Tam

Department of Computer Science
Swansea University
Swansea, UK

csvision.swan.ac.uk

A non-invasive procedure called Transcatheter Aortic Valve Implantation (TAVI) has emerged as an alternative procedure for patients suffering with aortic stenosis, but cannot undergo standard open-heart surgery. A full segmentation of the aortic root is important to the success of the procedure, and is essential for patient selection, procedural planning, and post-evaluation [1].

We propose a fully-automatic, deformable model-based method to segment the aortic root in 3D cardiac CT images. This consists of aligning an initial mesh with an initial aortic root pose estimation, before deforming the mesh towards the object boundary in the image. The estimation of the aortic root pose can be considered as an object detection problem, and a marginal space learning (MSL) method is adopted from [3] for this purpose. Once the initial mesh is aligned with the estimated pose, we implement a two-stage mesh deformation method: non-iterative boundary segmentation followed by iterative boundary refinement with mesh smoothing. Figure 1 outlines the steps taken at the testing stage for our automatic segmentation method.

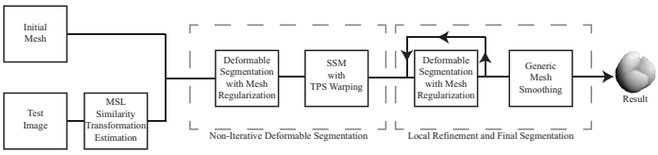


Figure 1: Overview of the proposed automatic aortic root segmentation at the testing stage.

The initial non-iterative process consists of deformable segmentation with mesh regularisation followed by a statistical-shape-model (SSM) based regularisation with thin-plate-spline (TPS) warping. The deformable segmentation consists of a learning-based boundary detector, where each vertex of the mesh template is adjusted to fit the object boundary along the orthogonal direction. After boundary detection, we have a set of original mesh vertices V, and a new set of vertex positions V' which may result in tangled and overlapping mesh faces. To address this problem, we propose a *B-spline based mesh regularisation* method which estimates a non-rigid transformation between V and V' before performing a free-form-deformation (FFD) on V to fit V'. This is estimated by warping an underlying voxel lattice controlled by a set of control points $\phi^h_{i,j,k}$, which act as parameters of a B-spline.

$$T(x,y,z) = \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_l(u)B_m(v)B_n(w)\phi_{i+l,j+m,k+n} \qquad (1)$$

where B_l represents the lth basis function of the B-spline, i, j and k are the corresponding lattice positions and u, v and w are the fractional positions along the lattice [2]. In addition, this non-rigid transformation is estimated in a multi-resolution procedure which is expressed as a summation of FFDs at multiple resolutions H:

$$T^H(x,y,z) = \sum_{h=1}^{H} T^h(x,y,z) \qquad (2)$$

This is followed by an SSM based regularisation with TPS warping, which makes the process capable of large vertex deformations without introducing mesh irregularities.

Finally, a boundary refinement iterating a local deformable segmentation with mesh regularisation process is applied. This process again combines the learning-based boundary detector and B-spline based mesh regularisation. Finally, generic mesh smoothing is used to remove any faceted mesh faces.

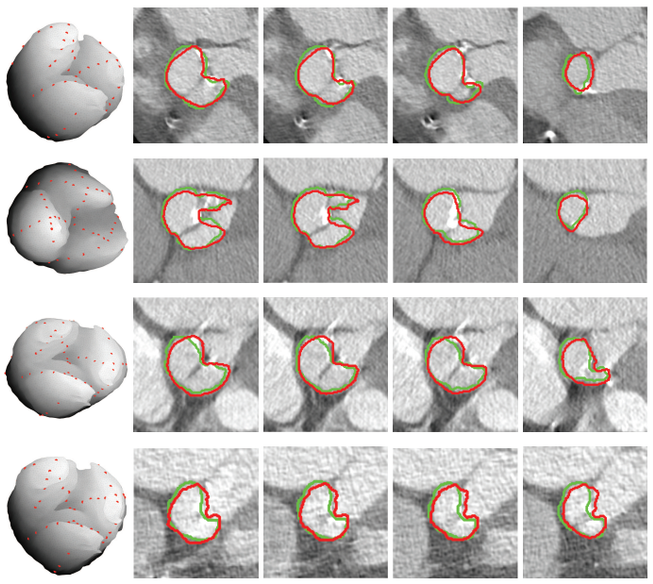


Figure 2: Results of the proposed method. Each row is a result of a different test image. The first column shows the resulting mesh. The following columns show a selection of image slices with the segmentation results. Green contours are the ground truth, while red contours are the segmentation results.

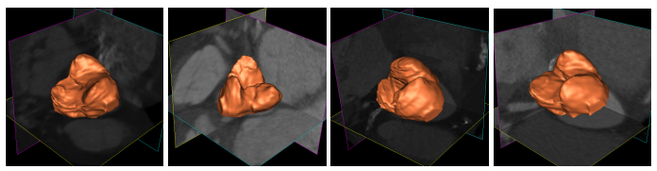


Figure 3: Further example segmentation results with our proposed method.

Quantitative comparisons were carried out against a state-of-the-art deformable model-based approach [3] and an active-shape-model based segmentation. Example segmentation results of the proposed method are shown in Figures 2 and 3. The proposed method achieves and average mesh error of $1.39 \pm 0.29mm$, and Hausdorff distance of $6.75 \pm 2.05mm$. These errors are lower than the two comparison approaches, and also results in much more regularised mesh surfaces with no tangled mesh faces. The proposed method is not only more efficient but also more accurate in segmenting complex anatomical structures with ambiguous image appearance.

[1] M. Gessat, T. Frauenfelder, L. Altwegg, J. Grunenfelder, and V. Falk. Transcatheter aortic valve implantation: Role of imaging. *Aswan Heart Centre Science and Practice Series*, 1(3), 2011.

[2] D. Rueckert, L. I. Sonoda, C. Hayes, D. Mill, O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: Application to breast mr images. *Trans. Med. Imaging*, 18(8):712–721, 1999.

[3] Y. Zheng, M. John, R. Liao, A. Nottling, J. Boese, J. Kempfert, T. Walther, G. Brockmann, and D. Comaniciu. Automatic aorta segmentation and valve landmark detection in c-arm ct for transcatheter aortic valve implantation. *Trans. Med. Imaging*, 31:2307–2321, 2012.

Anatomical triangulation: from sparse landmarks to dense annotation of the skeleton in CT images

Marie Bieth[12]
marie.bieth@tum.de
Rene Donner[3]
Georg Langs[3]
Markus Schwaiger[1]
Bjoern Menze[2]

[1] Nuclear Medicin
Klinikum Rechts der Isar
Munich, Germany

[2] Department of Computer Science
TU Munich
Munich, Germany

[3] Department of Biomedical Imaging and Image-guided Therapy, CIR Lab
Medical University of Vienna
Vienna, Austria

The automated annotation of bones that are visible in CT images of the skeleton is a challenging task which has, so far, been approached for only certain subregions of the skeleton, such as the spine. For example, in [4] and [2], methods are proposed for the annotation of the spine. In [5], a method is proposed to segment cartilage in the knee. In [6], a method is proposed to annotate the ribs. However, these methods are specific to a particular region of the skeleton and generalizing them to the whole skeleton, with a much larger field of view, a higher variability of the anatomy, and also variations in the positioning of the patient, remains an open task.

In this paper, we propose an approach for fully automatic dense annotation of substructures of the whole skeleton in CT images for staging bone tumour patients, where we want to localize anatomical substructures in order to estimate the local tumour load as visible from PET/CT images.

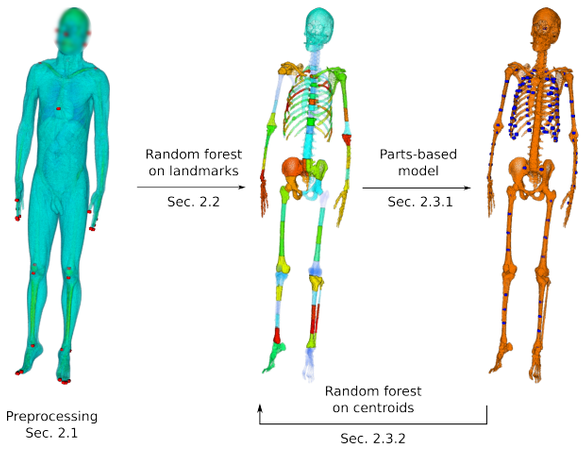


Figure 1: Description of the workflow. On the left, landmarks generated by [1] are shown. A first labelling (middle) is obtained using a Random Forest classifier. Centroids of each segment (right) are disambiguated by the parts-based model. They serve as new landmarks for another random forest classifier. The last two steps can be iterated.

As shown in Fig. 1, our method contains four main steps: the preprocessing, a first Random Forest classifier, a parts-based model and a second Random Forest classifier. The last two steps can be iterated.

Preprocessing: The bones are segmented from the CT image by adapting [3]. The intensities in the image are modelled as a mixture of two Gaussians and the resulting thresholding is regularized using graphcut. Landmarks distributed on the whole body are pre-filtered by a classifier and refined through a Hough regression model and a parts-based model of the global landmark topology as explained in [1].

Anatomical triangulation: A Random Forest classifier provides a first annotation of the skeleton. It uses as features: the triangulation features (signed distance to each landmarks in each direction), geodesic distances to selected landmarks, mean image intensity and bone proportion around the center voxel, position in the image.

Centroid regularization: The position of the centroid of each substructure of the skeleton is disambiguated by a Conditional Random Field (CRF), whose unary and binary potentials use global anatomical knowledge. The best candidate centroid for each substructure is computed by minimizing the energy of the CRF.

Iterative triangulation: The centroids computed in the previous step are used as new, evenly distributed landmarks for a new Random Forest classifier based on distance features. The resulting probability distribution can either be converted to a hard labeling or used as input for a new centroid regularization step.

We have evaluated our method on 18 whole body CT scans. 136 bones or bone segments were annotated manually (example with shuffled labels in Fig. 1 middle). 57 landmarks were computed for each dataset (examples in Fig. 1 left). The test were run using six-fold cross-validation. Our method achieves a mean overall weighted Dice Score of 90.54.

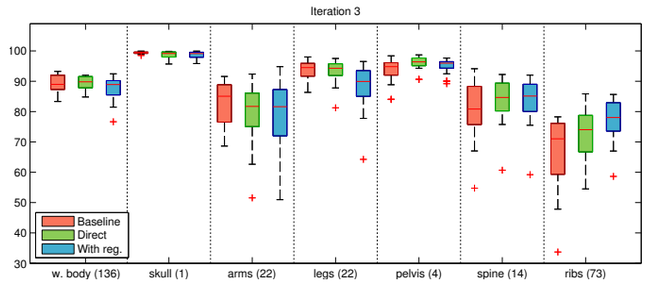


Figure 2: Weighted Dice scores for the segmentation of different groups of substructures by the baseline-random forest (red), after 3 iterations without centroid regularization (green), after three iterations with centroid regularization (blue). The total number of substructures of each group is given in brackets.

[1] René Donner et al. Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization. *Medical Image Analysis*, 17(8):1304–1314, 2013.

[2] Ben Glocker et al. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In *Proc. MICCAI 2013*, pages 262–270. Springer, 2013.

[3] Yan Kang et al. A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data. *IEEE Transactions on Medical Imaging*, 22(5):586–598, 2003.

[4] Stefan Schmidt et al. Spine detection and labeling using a parts-based graphical model. In *Information Processing in Medical Imaging*, pages 122–133. Springer, 2007.

[5] Quan Wang et al. Semantic context forests for learning-based knee cartilage segmentation in 3D MR images. In *Medical Computer Vision. Large Data in Medical Imaging*, pages 105–115. Springer, 2014.

[6] Dijia Wu et al. A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images. In *Proc. CVPR 2012*, pages 980–987. IEEE, 2012.

Primal-Dual convex optimization in large deformation diffeomorphic registration with robust regularizers

Monica Hernandez
mhg@unizar.es

Aragon Institute on Engineering Research (I3A),
University of Zaragoza

Non-rigid image registration is a highly ill-posed problem. This means that a number of qualitatively different transformations can achieve the same image similarity after registration. This justifies the vast literature on image registration methods with differences on transformation characterization, regularizers, image similarity metrics, optimization methods, and additional constraints [3].

In the last decade, diffeomorphic registration has arisen as a powerful paradigm for non-rigid image registration, with application to Computational Anatomy. Large Deformation Diffeomorphic Metric Mapping (LDDMM) [1] and Diffeomorphic Demons [4] are among the most widespread methods for diffeomorphic registration. In both methods, transformations are characterized to belong to an infinite dimensional Riemannian manifold of diffeomorphisms, parameterized by flows of smooth vector fields in the tangent space. The invertibility of the transformations is numerically guaranteed by the use of sufficiently smooth regularizers. Customarily, these methods regularize the problem with the L^2-norm of some physically meaningful differential expression of the vector fields, or with the Gaussian smoothing of the vector fields.

Simultaneously to the development of the diffeomorphic registration paradigm, the computer vision community has shown a growing interest in robust regularizers based on the Total Variation (TV) norm. The popularity of these regularizers has increased thanks to the availability of optimization methods for solving this challenging problem. The ability of TV based regularizers to preserve discontinuities has led these methods to occupy top positions in optical flow benchmark studies and non-rigid image registration evaluations.

The purpose of this article is to propose a method for primal-dual optimization of convexified LDDMM problems, formulated with robust regularizers and image similarity norms related to the TV norm. The method is based on Chambolle and Pock algorithm with diagonal preconditioning [2].

Let $Diff(\Omega)$ be the manifold of diffeomorphisms. Let V be the corresponding tangent space at the identity. Let $\mathcal{L} = Id - \gamma\Delta$ be the autoadjoint Laplacian operator associated to the scalar product in V, providing the Riemannian metric in $Diff(\Omega)$. The LDDMM variational problem is given by the minimization of the energy functional

$$E(v) = E_{\text{reg}}(v) + \alpha E_{\text{img}}(v) = \langle \mathcal{L}v, \mathcal{L}v \rangle_{L^2} + \alpha \| I_0 \circ \phi^v_{1,0} - I_1 \|^2_{L^2}. \quad (1)$$

As a result of the composition of the image with the diffeomorphism $\phi^v_{1,0}$, this energy functional is non-convex. In order to apply Fenchel-Duality principles in the primal-dual optimization of the problem, the functional needs to be transformed into a convex energy. This is done using the stationary parameterization of diffeomorphisms and approximating the Gateaux derivative $\partial_v \phi^v_{1,0}$ by $-D\phi^v_{1,0} \cdot v$ using the normal coordinate representation, yielding

$$E_{conv}(v) = E_{\text{reg}}(v) + \alpha E_{\text{img}}(I_0 \circ \phi^{v_0}_{1,0} - I_1 + \nabla (I_0 \circ \phi^{v_0}_{1,0})^T v_0 - \nabla (I_0 \circ \phi^{v_0}_{1,0})^T v). \quad (2)$$

In this work, we study three robust regularizers liable to provide acceptable results in diffeomorphic registration: Huber, V-Huber and Total Generalized Variation (TGV). The general algorithm that solves the different variational problems is given by

Data: $v^0 \in V$, $\overline{v}_0 \in V$, $p^0 \in P$, $q^0 \in Q$,
Σ_p, Σ_q, T preconditioning matrices, $\theta \in [0,1]$
Result: $v \in V$, $p \in P$, $q \in Q$ solutions of the primal-dual problem
for $n \leftarrow 0$ **to** *maxits* **do**
$p^{n+1} = (Id + \Sigma_p \partial F^*)^{-1}(p^n + \Sigma_p K \overline{v}^n)$
$q^{n+1} = (Id + \Sigma_q \partial G^*)^{-1}(q^n + \Sigma_q A \overline{v}^n)$
$v^{n+1} = v^n - TK^* p^{n+1} - TA^* q^{n+1}$
$\overline{v}^{n+1} = v^{n+1} + \theta(v^{n+1} - v^n)$
end

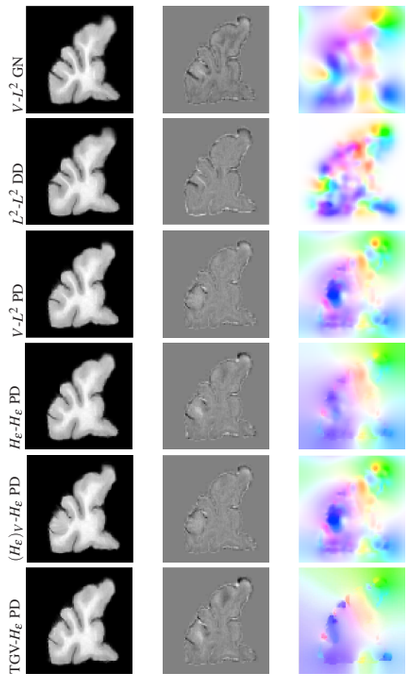


Figure 1: **2D MRI experiment.** Image registration results with the state of the art methods, and the proposed method and the considered regularizers. From left to right, warped sources ($I_0 \circ \phi^v_{1,0}$), residuals ($I_0 \circ \phi^v_{1,0} - I_1$) and velocity fields for the methods considered in the comparison.

The method is compared in a complex geometry 2D MRI data set with state of the art optical flow and diffeomorphic registration methods. In addition, the 3D version of the method is evaluated with the manual segmentations of the Non Rigid Image Registration Evaluation Project (NIREP) database.

Results in the 2D MRI data set have demonstrated that diffeomorphic solutions can be obtained for Huber, V-Huber and TGV regularizers, despite the preservation of discontinuities favored by the robust regularizers. The method has shown to be able to perform similarly to state of the art diffeomorphic registration methods in terms of the image similarity after registration.

The evaluation in the NIREP database has shown a comparable performance for V-Huber regularizer with respect to the original V-L^2 variational formulation and L^2-L^2 log-domain diffeomorphic Demons. For each region, the results obtained by V-Huber regularizer were with the best scoring methods. The performance of TGV regularizer was usually located above the worst performing methods and slightly below the best performing methods. Moreover, primal-dual optimization methods (with the exception of the TGV regularized) were more efficient than diffeomorphic Demons, widely used because of its computational efficiency.

[1] M. F. Beg, M. I. Miller, A. Trouve, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vision*, 61 (2):139–157, 2005.

[2] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. *Proc. of the 13th IEEE International Conference on Computer Vision (ICCV'11)*, 2011.

[3] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE Trans. Med. Imaging*, 32(7):1153 – 1190, 2013.

[4] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *Neuroimage*, 45(1):S61 – S72, 2009.

Global Minimum for Curvature Penalized Minimal Path Method

Da CHEN
chenda@ceremade.dauphine.fr
Jean-Marie MIREBEAU
mirebeau@ceremade.dauphine.fr
Laurent D. COHEN
cohen@ceremade.dauphine.fr

Université Paris Dauphine,
CNRS, CEREMADE, UMR7534
F-75016, Paris, France

Minimal path problem [1] is posed on a bounded domain Ω and a metric $\mathcal{F}$ prescribing norm $\mathcal{F}_x$ (potentially asymmetric) at each point $x \in \Omega$. Let $\Im$ denote the collection of Lipschitz paths $\gamma : [0,1] \to \Omega$, and the path length $\ell_{\mathcal{F}}$ can be measured through the metric $\mathcal{F}$ as:

$$\ell_{\mathcal{F}}(\gamma) = \int_0^1 \mathcal{F}_{\gamma(t)}(\dot\gamma(t))\, dt, \tag{1}$$

where $\dot\gamma(t) = \frac{d}{dt}\gamma(t)$. The minimal action map $\mathcal{U}$ from a set of initial source points $\mathcal{W}$, is the minimal energy of any path joining $x \in \Omega$ to $\mathcal{W}$:

$$\mathcal{U}(x) := \min\left\{\ell_{\mathcal{F}}(\gamma); \gamma \in \Im,\, \gamma(1) = x,\, \gamma(0) \in \mathcal{W}\right\}. \tag{2}$$

The minimal action map $\mathcal{U}$ is the unique viscosity solution to an Eikonal PDE, defined in terms of the dual metric $\mathcal{F}^*$. For all $x \in \Omega$ one has $\mathcal{U}(x) = 0$ if $x \in \mathcal{W}$ and otherwise

$$\mathcal{F}^*_x(-\nabla \mathcal{U}(x)) = 1, \quad \text{where } \mathcal{F}^*_x(u) = \sup_{v \neq 0} \frac{\langle u, v\rangle}{\mathcal{F}_x(v)}. \tag{3}$$

The metric $\mathcal{F}$ considered in this paper combine a symmetric part, defined in terms of a symmetric positive tensor field $\mathcal{M}$, and an asymmetric part involving a vector field ω:

$$\mathcal{F}_x(u) = \sqrt{\langle u, \mathcal{M}(x)\, u\rangle} - \langle \omega(x), u\rangle. \tag{4}$$

We require $\langle \omega(x), \mathcal{M}^1 u\rangle < 1$ to ensure the metric positivity. In this work, we show that the Euler's elastic bending energy can be approximated by the Finsler metric with the form of (4).

Following Mumford [4], the weighted Euler's elasticas, minimizing the following bending energy:

$$\mathcal{L}(\Gamma) = \int_0^L \left(\frac{1}{\alpha(s)} + \frac{1}{\beta(s)}\kappa^2(s)\right) ds \tag{5}$$

L denotes the classical curve length, s is the arc-length parameter, and $\Gamma : [0, L] \to \Omega$ is a curve with non-vanishing velocity vector. κ is the curvature and α, β are two positively weighted functions computed by the optimally oriented flux filter [2].

Our first step is to cast the elastica energy (5) in the form of path length with respect to a degenerate Finsler metric. For that purpose, let $S^1 = [0, 2\pi[$ be the space of angles, with periodic boundary conditions, and for each angle θ let $\vec v_\theta = (\cos\theta, \sin\theta)$ be the corresponding unit vector. For $\gamma = (\Gamma, \theta) \in \Omega \times S^1$ and $\dot\gamma = (\dot\Gamma, \dot\theta) \in \mathbb{R}^2 \times \mathbb{R}^1$ we can define

$$\mathcal{F}^\infty_\gamma := \begin{cases} \|\dot\Gamma\| + |\dot\theta|^2/\|\dot\Gamma\| & \text{if } \dot\Gamma \text{ is positively proportional with } \vec v_\theta, \\ \infty & \text{otherwise.} \end{cases} \tag{6}$$

Consider an curve Γ parametrized by arc-length, and its orientation lifting $\gamma = (\Gamma, \theta)$, with $\dot\Gamma(s) := \vec v_\theta$. By assumption one has $\|\dot\Gamma(s)\| = 1$, hence $\ddot\Gamma = \kappa(s)\dot\Gamma(s)^\perp$, so that $\kappa(s) = \dot\theta(s)$ and

$$\int_0^L (1 + \kappa^2(s))\, ds = \int_0^L \mathcal{F}^\infty_{\gamma(s)}(\dot\gamma(s))\, ds. \tag{7}$$

The weights α, β of (5) are easily taken into account, by modifying (6) approximately. Note that thanks to the homogeneity of $\mathcal{F}^\infty_\gamma(\dot{)}$, the right hand side of (7) is invariant under reparametrizations of the lifted path γ. However, the degenerate metric $\mathcal{F}^\infty$ is too singular to apply the fast marching algorithm directly, hence we introduce a family tamings depending on a penalization parameter $\lambda \gg 1$:

$$\mathcal{F}^\lambda_\gamma(\dot\gamma) := \sqrt{\lambda^2\|\dot\Gamma\|^2 + 2\lambda|\dot\theta|^2} - (\lambda - 1)\langle \vec v_\theta, \dot\Gamma\rangle. \tag{8}$$

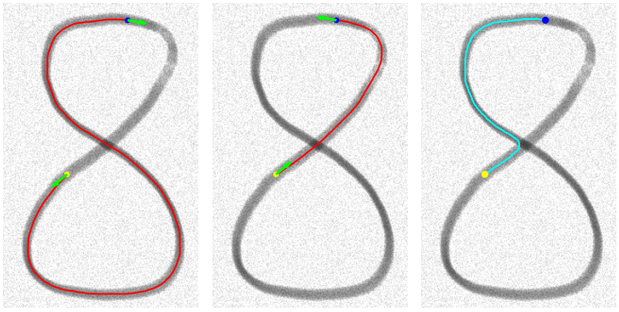
Again, the weights α, β can easily be incorporated into the metric $\mathcal{F}^\lambda$. As $\lambda \to \infty$ one has:

$$\begin{aligned} \mathcal{F}^\lambda_\gamma(\dot\gamma) &= \lambda\|\dot\Gamma\|\sqrt{1 + \frac{2|\dot\theta|^2}{\lambda\|\dot\Gamma\|^2}} - (\lambda - 1)\langle \vec v_\theta, \dot\Gamma\rangle \\ &= \lambda\|\dot\Gamma\|\big(1 + \frac{|\dot\theta|^2}{\lambda\|\dot\Gamma\|^2} + \mathcal{O}(\frac{1}{\lambda^2})\big) - (\lambda - 1)\langle \vec v_\theta \dot\Gamma\rangle \\ &= \|\dot\Gamma\| + \frac{|\dot\theta|^2}{\|\dot\Gamma\|} + (\lambda - 1)(\|\dot\Gamma\| - \langle \vec v_\theta, \dot\Gamma\rangle) + \mathcal{O}(\frac{1}{\lambda}) \end{aligned}$$

which tends to $\mathcal{F}^\infty_\lambda(\dot\gamma)$ as $\lambda \to \infty$. The metric $\mathcal{F}^\lambda$ has precisely the required form (4), with a diagonal tensor field $\mathcal{M}$:

$$\mathcal{M} = \begin{pmatrix} \lambda^2 & 0 & 0 \\ 0 & \lambda^2 & 0 \\ 0 & 0 & 2\lambda \end{pmatrix},$$

and a non-zero vector field $\omega(\gamma, \theta) := \vec v_\theta$. Then the minimal action map $\mathcal{U}^\lambda$ associated to the metric $\mathcal{F}^\lambda$ can be efficiently computed with the fast marching method [3].



(a) (b) (c)

Figure 1: Comparison results on synthetic images. Yellow points denote the initial source point and blue points denote the endpoints. Green arrows in (a) and (b) demonstrate the initial and end tangent vectors for the proposed algorithm.

In Fig. 1, we show the extraction results (Fig. 1(a) and (b)) from the proposed minimal path model and the Cohen-Kimmel model [1](Fig. 1(c)). In (Fig. 1(a) and (b)), given different initial and end tangent vector as indicated by the green arrows, one could get different extraction results (shown as red solid curves) due to the asymmetric Finsler metric. Especially in Fig. 1(a), the curve length of the extracted geodesic is much longer than that in Fig. 1(b). With the proposed Finsler metric based curvature-penalty minimal path model, the proposed model can choose a correct path. For the Cohen-Kimmel model, the extracted path only follows the way which owns the largest speed values instead of choosing a smooth path.

[1] L. D. Cohen and R. Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24(1):57–78, 1997.

[2] Max W. K. Law and Albert C. S. Chung. Three Dimensional Curvilinear Structure Detection Using Optimally Oriented Flux. *Computer Vision-ECCV 2008. Springer Berlin Heidelberg*, pages 368–382, 2008.

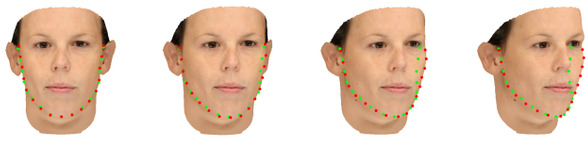
[3] Jean-Marie Mirebeau. Efficient fast marching with Finsler metrics. *Numerische Mathematik*, 126(3):515–557, 2014.

[4] David Mumford. *Elastica and computer vision*. Springer, 1994.

Adaptive Contour Fitting for Pose-Invariant 3D Face Shape Reconstruction

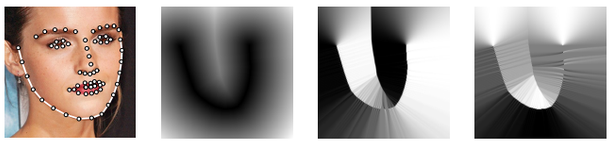
Chengchao Qu[12]
chengchao.qu@iosb.fraunhofer.de
Eduardo Monari[2]
eduardo.monari@iosb.fraunhofer.de
Tobias Schuchert[2]
tobias.schuchert@iosb.fraunhofer.de
Jürgen Beyerer[21]
juergen.beyerer@iosb.fraunhofer.de

[1] Vision and Fusion Laboratory
Karlsruhe Institute of Technology
Karlsruhe, Germany
[2] Fraunhofer IOSB
Karlsruhe, Germany



(a) 0° (b) 10° (c) 20° (d) 30°

Figure 1: Correspondence errors of 2D (red) and 3D (green) facial contour landmarks w.r.t. yaw angles of (a) 0°, (b) 10°, (c) 20° and (d) 30°.



(a) 2D features (b) $\mathbf{D}$ (c) $\mathbf{D}_x$ (d) $\mathbf{D}_y$

Figure 3: (a) Improved 2D features with connected contour lines, (b) their Distance Transform and its derivatives in (c) x and (d) y directions.

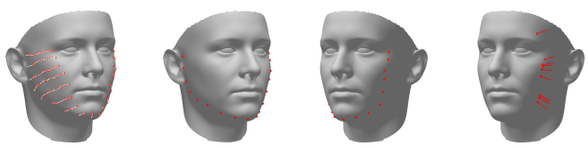
Motivation Direct reconstruction of 3D face shape—solely based on a sparse set of 2D feature points localized by a facial landmark detector—offers an automatic, efficient and illumination-invariant alternative to the widely known analysis-by-synthesis framework, which is extremely time-consuming considering the enormous parameter space for both shape and photometric properties. Given 2D landmarks $\mathbf{y}$ and their correspondence on the 3D Morphable Model (3DMM), the 3D shape can be recovered by minimizing the distance between 2D and the projected 3D landmarks

$$\hat{\alpha} = \arg\min_{\alpha} E(\alpha) = \arg\min_{\alpha} \|\mathbf{Q}\alpha - \mathbf{y}\|_2^2 + \eta \|\alpha\|_2^2, \tag{1}$$

where α is the target 3DMM shape parameter. $\mathbf{Q}$ is a simplification of the projected 3D landmarks $\mathbf{Q} = \Pi\Phi\mathbf{S}\,\mathrm{diag}(\sigma)$ containing the affine camera projection matrix Π, sparse landmark selection Φ, as well as 3DMM eigenvectors $\mathbf{S}$ and eigenvalues σ. The term $\eta\|\alpha\|_2^2$ prevents overfitting and regulates the plausibility of the reconstructed faces.

The empirical assumption of a fixed mapping between 2D and 3D landmarks gives rise to a major flaw of the landmark-based methods. Fig. 1 demonstrates that with increasing yaw angle, remarkable deviation in the self-occluded face half can be observed. Moreover, we argue that not only the invisible contour landmarks, but also the visible ones are susceptible to 2D–3D correspondence discrepancy. Even in frontal view, a tight correspondence cannot be necessarily guaranteed (see Fig. 1a). In this paper, we propose a novel algorithm to address these two problems.

Fast Detection of Silhouette Vertices Due to the fact that 2D contour landmarks are always located on the face silhouette, the tangent planes of such 3D vertices are perpendicular to the image plane, meaning that their normal vectors projected onto the z-axis are close to zero. It is then an intuitive idea to treat those with the absolute z-projection values of the normals $|\mathbf{n}_z| < t$ as silhouette points (see Fig. 2d). However, a universally valid threshold t is hard to find and the spatial distribution is uncontrollable, too. Instead, we impose geometric constraints by labeling extended vertices starting from the original 3D contour landmarks on the 3DMM towards the center of the face (see Fig. 2a). During fitting, the ones with the smallest $|\mathbf{n}_z|$ on each line are chosen (see Figs. 2b and 2c). The number of the evaluated normals is thereby reduced by two orders of magnitude, allowing for fast and robust online update during the iterations.



(a) (b) (c) (d)

Figure 2: Fast detection of silhouette vertices using (a) annotated candidates. (b) and (c) show our result compared to (d) the direct approach.

Fitting To facilitate flexible fitting of 3D contour vertices, the continuous curve formed by the corresponding discrete 2D landmarks is exploited when reconstructing the shape (see Fig. 3a). Separating all landmarks into two disjoint subsets of fixed and contour ones in Eq. (1) leads to

$$E(\alpha) = \left\| \begin{bmatrix} \mathbf{Q}_{\text{contour}} \\ \mathbf{Q}_{\text{fixed}} \end{bmatrix} \alpha - \begin{bmatrix} \mathbf{y}_{\text{contour}} \\ \mathbf{y}_{\text{fixed}} \end{bmatrix} \right\|_2^2 + \eta \|\alpha\|_2^2. \tag{2}$$

An unknown mapping denoted $\phi(i) = j$ which selects, for each 3D contour vertex i, the corresponding 2D contour pixel j, is now also a part of the minimization process

$$E(\alpha, \phi) = \sum_i \left\| \mathbf{Q}_i\alpha - \mathbf{y}_{\phi(i)} \right\|_2^2 + \|\mathbf{Q}_{\text{fixed}}\alpha - \mathbf{y}_{\text{fixed}}\|_2^2 + \eta \|\alpha\|_2^2. \tag{3}$$

Estimation of α is then a "minimization of minimization" problem

$$\hat{\alpha} = \arg\min_{\alpha} \sum_i \min_j \left\|\mathbf{Q}_i\alpha - \mathbf{y}_j\right\|_2^2 + \|\mathbf{Q}_{\text{fixed}}\alpha - \mathbf{y}_{\text{fixed}}\|_2^2 + \eta \|\alpha\|_2^2. \tag{4}$$

The trick to circumvent this difficulty is to apply Distance Transform (DT) to the discrete 2D features with the Levenberg-Marquardt Iterative Closest Point (LM-ICP) algorithm. On the 2D image raster $\mathbf{x}$, DT assigns each image pixel with the distance to its closest point on the contour lines $D(\mathbf{x}) = \min_j \left\|\mathbf{x} - \mathbf{y}_j\right\|_2$, which is reusable for the entire reconstruction procedure by virtue of its independence of the model parameter α (see Fig. 3b). The merit of DT lies in that the mapping function $\phi(i)$, or the minimization over j in the contour cost of Eq. (3), then vanishes and is thereby simply replaced with $D(\mathbf{Q}_i\alpha) = \min_j \left\|\mathbf{Q}_i\alpha - \mathbf{y}_j\right\|_2$. Integrating DT into Eq. (3) and vectorizing over all contour vertices i yields

$$E(\alpha) = \|\mathbf{D}(\mathbf{Q}_{\text{contour}}\alpha)\|_2^2 + \|\mathbf{Q}_{\text{fixed}}\alpha - \mathbf{y}_{\text{fixed}}\|_2^2 + \eta \|\alpha\|_2^2. \tag{5}$$

Rather than the sum of squares, LM-ICP demands the vector of residuals

$$\mathbf{e}(\alpha) = \begin{bmatrix} \mathbf{D}(\mathbf{Q}_{\text{contour}}\alpha) \\ \mathbf{Q}_{\text{fixed}}\alpha - \mathbf{y}_{\text{fixed}} \\ \sqrt{\eta}\alpha \end{bmatrix}. \tag{6}$$

Differentiating Eq. (6) analytically subject to the shape parameter α is possible when the chain rule $\frac{\partial D_i}{\partial \alpha_j} = \frac{\partial D}{\partial x}\mathbf{f}_i \cdot \frac{\partial \mathbf{f}_i^x}{\partial \alpha_j} + \frac{\partial D}{\partial y}\mathbf{f}_i \cdot \frac{\partial \mathbf{f}_i^y}{\partial \alpha_j}$ and the precomputed gradient images in x and y directions (Figs. 3c and 3d) are applied to calculate the derivatives of the contour cost $\nabla_\alpha \mathbf{D}(\mathbf{Q}_{\text{contour}}\alpha)$. The target Jacobian matrix $\mathbf{J}_{ij} = \frac{\partial \mathbf{e}_i}{\partial \alpha_j}$ is then

$$\mathbf{J} = \begin{bmatrix} \mathbf{D}_x(\mathbf{Q}_{\text{contour}}\alpha) \cdot \mathbf{Q}^x_{\text{contour}} + \mathbf{D}_y(\mathbf{Q}_{\text{contour}}\alpha) \cdot \mathbf{Q}^y_{\text{contour}} \\ \mathbf{Q}_{\text{fixed}} \\ \sqrt{\eta}\mathbf{I} \end{bmatrix}. \tag{7}$$

Results On the Basel Face Model (BFM), promising results outperforming state-of-the-art methods demonstrate the robustness of the proposed 3D shape reconstruction framework against pose variation. The effectiveness and efficiency justify our theoretical and practical advances.

Freehand Laser Scanning Using Mobile Phone

Ron Slossberg
http://www.cs.technion.ac.il/~ronslos/
Aaron Wetzler
http://www.cs.technion.ac.il/~twerd/
Ron Kimmel
http://www.cs.technion.ac.il/~ron/

Technion Israel Institute of Technology
Haifa, Israel

3D scanners are growing in their popularity as many new applications and products are becoming a commodity. These applications are often tethered to a computer and/or require expensive and specialized hardware. In this note we demonstrate that it is possible to achieve good 3D reconstruction on a mobile device. We describe a novel approach for mobile phone scanning which utilizes a smart-phone and a cheap laser pointer with a cylindrical lens which produces a line pattern attached to the phone using a 3D printed adapter as demonstrated in Figure 1. Our 3D reconstruction

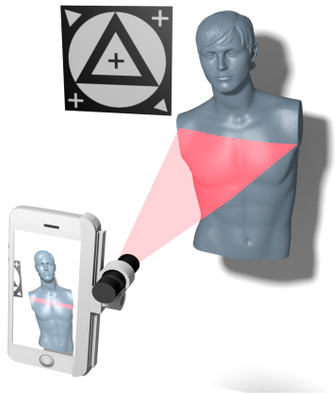


Figure 1: Illustration of the scanning process with a mobile device.

method is built upon six steps that are preformed online on the mobile device.

- Calibration
  As a precursor for reconstruction the system must be calibrated. We use openCV [3] for camera intrinsic calibration, and a simple user friendly method for laser plane calibration. This procedure provides all the knowledge about the system configuration that is necessary for 3D reconstruction.

- Pose estimation
  We estimate the camera pose for each captured frame. We use [1] which is able to track a physical marker combined with sparse points in the scene combined with cues from onboard inertial sensors. This library provides an estimated camera position in real time. Another example for camera pose estimation that we tried is the OpenCV checkerboard calibration tool.

- Laser line extraction
  The laser line position is extracted from each frame using an non-linear filteration scheme described by Koller *et al.* [5] and applied by Matiukas and Miniotas [6]. This non-linear combination of filter kernels produces a strong response in the region of the laser line. We estimate the maxima of the response with sub-pixel accuracy using polynomial fitting.

- 3D point reconstruction
  For each image point where the laser line is detected we reconstruct a point in the point cloud. Initially the point is reconstructed relative to the camera using the camera intrinsic calibration and the laser plane calibration. We are then able to reconstruct translate the points into world coordinates using per-frame camera position estimation obtained by the tracking system, by a simple matrix multiplication. Each reconstructed point is added to the final point cloud.

- Per point color recovery
  Due to the fact that each reconstructed point is taken from an area of the image which is colored by the laser, the color cannot be directly inferred. To resolve this issue, we reproject our reconstructed 3D points into an image of the scene captured after a small time interval. This interval allows the laser to pass the reconstructed point and for it's true color to be recovered.

- Surface reconstruction
  Once a dense point cloud is reconstructed, we are able to estimate the surface of the scanned object. This can be done by various reconstruction methods such as Poisson surface reconstruction as demonstrated in Figure 2.

We validate the proposed method by comparing the reconstruction error to the ground truth obtained from an industrial laser scanner. This validation is carried out by scanning the same objects with our scanner and with a highly accurate commercial scanner. We then use the Iterative Closest Point (ICP) [2, 4] to align the scans, allowing a precise measurement of the distortion between the point clouds

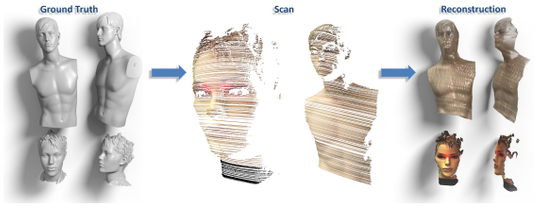


Figure 2: Left: Ground truth models obtained by a professional Artec scanner. Center: Point clouds obtained by our scanning process. Right: Triangle mesh reconstructed from point cloud.

Acknowledgment

This research was supported by European Community's FP7- ERC program, grant agreement no. 267414.



[1] 13th lab. Pointcloud SDK. `http://developer.pointcloud.io/sdk`, 2014.

[2] P.J. Besl and N.D McKay. Method for registration of 3D shapes. In *Robotics-DL*, pages 586–606. International Society for Optics and Photonics, 1992.

[3] G. Bradski. Opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. In *Proceedings of International Conference on Robotics and Automation*. IEEE, 1991.

[5] T.M. Koller, G. Gerig, G. Szekely, and D. Dettwiler. Multiscale detection of curvilinear structures in 2D and 3D image data. In *Fifth International Conference on Computer Vision. (ICCV)*. IEEE, 1995.

[6] V. Matiukas and D. Miniotas. Detection of laser beam's center-line in 2D images. *Electronics & Electrical Engineering*, 2009.

Stochastic visibility in point-sampled scenes

Miles Hansard
http://www.eecs.qmul.ac.uk/~milesh/

School of EECS
Queen Mary, University of London
London E1 4NS

This work introduces a new statistical model of visibility in point-sampled scenes, such as those constructed from multiple lidar or depth-camera scans. A *visibility density* is defined along each optical ray, which gives the probability that a non-occluded scene-point exists at any particular location, with respect to a given camera. The new approach avoids any commitment to a surface-mesh, in order to develop a more data-driven probabilistic model. Furthermore, this approach naturally allows for the existence of gaps and uncertainty in the point-cloud (see fig. 2).

The new model has potential applications to multi-view stereo problems, in which visibility is an essential component of the photometric re-projection error [3]. There are other potential applications to the graphical rendering of point-cloud data [2].



Figure 2: A typical 3-D point-cloud, constructed from multi-view RGB-D scans [1], as used in the experimental evaluation.

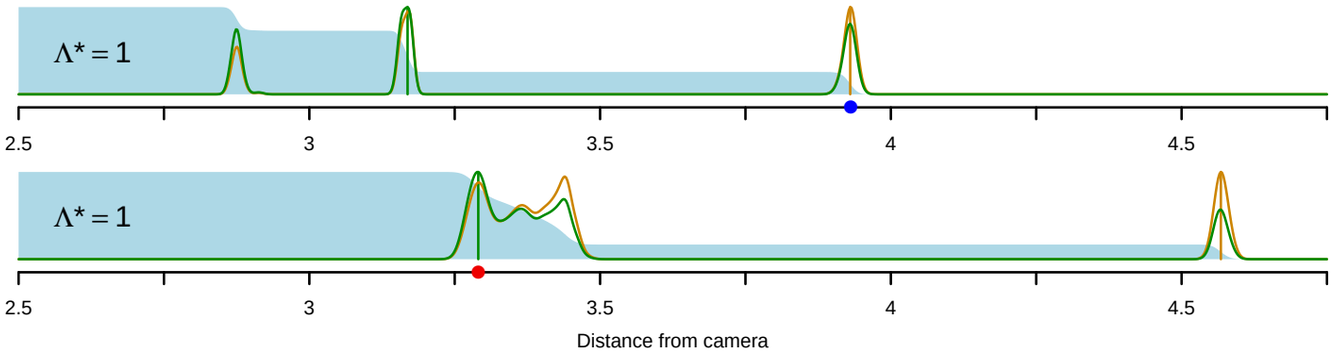


Figure 1: The visibility density (3 & 7) shown in green, for two rays in the evaluation data-set. The target-point of each ray is indicated by a dot, colour-coded by its true state; blue means occluded, and red means visible. Orange curves show the occupancy density (2 & 4), while the blue polygon indicates the attenuating vacancy density (1 & 5). Vertical lines indicate global maxima along each ray. **Top**: the distant occluded target generates the occupancy maximum (orange), but not the visibility maximum (green). **Bottom**: the nearby visible target generates the visibility maximum (green), but not the occupancy maximum (orange).

It is essential to make a careful definition of visibility, for point-sampled scenes. Let $\mathbf{c}$ be the camera-centre, and let $\mathbf{u}$ be the direction of visual ray $\mathcal{R}$. Then the *event* that point $\mathbf{p}(t) = \mathbf{c} + t\,\mathbf{u}$ is visible will be denoted $(\multimap t \,|\, \mathcal{R})$, where the dot is intended to suggest the end of a visual ray. Visibility will now be defined as the conjunction of two events. Firstly, the ray segment $\mathbf{c} + r\,\mathbf{u}$, where $0 \leq r < t$, should be *vacant*, i.e. free of occluders. Secondly, the point $\mathbf{c} + t\,\mathbf{u}$ should be *occupied*. The following notation will be used for these events:

$$\big(\varnothing_t \,|\, \mathcal{R}\big) \;\Leftrightarrow\; \text{The ray-segment from } \mathbf{c} \text{ to } \mathbf{c} + t\,\mathbf{u} \text{ is vacant} \qquad (1)$$

$$\big(\mathbb{1}_t \,|\, \mathcal{R}\big) \;\Leftrightarrow\; \text{The point } \mathbf{p} = \mathbf{c} + t\,\mathbf{u} \text{ is occupied} \qquad (2)$$

Hence the *probability* of the point $\mathbf{p}(t) = \mathbf{c} + t\,\mathbf{u}$ on ray $\mathcal{R}$ being visible in scene $\mathcal{S}$ is the *product* of the vacancy and occupancy probabilities:

$$\mathrm{pr}\big(\multimap t \,|\, \mathcal{R}, \mathcal{S}\big) \propto \mathrm{pr}\big(\varnothing_t \,|\, \mathcal{R}, \mathcal{S}\big) \times \mathrm{pr}\big(\mathbb{1}_t \,|\, \mathcal{R}, \mathcal{S}\big). \qquad (3)$$

This density, which is a function of distance t, can now be developed in relation to the point-sampled scene model.

The scene $\mathcal{S}$ is represented as a mixture of N Gaussian surface-patches, the positions and orientations of which are readily estimated from the data. The 'intersection' of ray $\mathcal{R}$ with 3-D patch k defines a 1-D Gaussian, and so the **occupancy density** (2) is a 1-D mixture, corresponding to the **orange curves** in fig. 1:

$$\mathrm{pr}\big(\mathbb{1}_t \,|\, \mathcal{R}, \mathcal{S}\big) = \frac{1}{N} \sum_k^N w_k\, G\big((t - \mu_k)^2 / \sigma_k^2\big). \qquad (4)$$

Here $G(x) = (2\pi)^{-\frac{1}{2}} \exp(-x/2)$, and the parameters (μ_k, σ_k) and weights w_k can all be computed directly from the known geometry.

If the scene consisted of *randomly* distributed particles, of radius ε, then it could be modelled as a volumetric Poisson distribution. It would follow that the probability of point $\mathbf{p}(t)$ being non-occluded would be $\mathrm{pr}(\varnothing_{\mathcal{C}}) \propto \exp(-\lambda|\mathcal{C}|)$, where $|\mathcal{C}| = \pi\varepsilon^2 t$ is the volume of a cylinder, which must be *empty*, between $\mathbf{p}(t)$ and the optical centre. This can be generalized to the case of non-uniformly distributed points, leading to a **vacancy density** corresponding to (1) and the **blue polygons** in fig. 1:

$$\mathrm{pr}\big(\varnothing_t \,|\, \mathcal{R}, \mathcal{S}\big) = \exp\big(-\eta\,\Lambda(t)\big) \qquad (5)$$

where η is a free parameter, and $\Lambda(t)$ is a generalized volume, which increases in dense regions of the point cloud. This can be defined as:

$$\Lambda(t) = \int_0^t \mathrm{pr}\big(\mathbb{1}_r \,|\, \mathcal{R}, \mathcal{S}\big)\,\mathrm{d}r. \qquad (6)$$

It can be shown that $\Lambda(t)$ is a *product* of generalized step-functions. These steps are probabilistic representations of potential occluders, lying between $\mathbf{p}(t)$ and the optical centre.

Substituting (4) and (5) into (3) gives the final **visibility density**, corresponding to the **green curves** in fig. 1, along ray $\mathcal{R}$:

$$\mathrm{pr}\big(\multimap t \,|\, \mathcal{R}, \mathcal{S}\big) = \frac{\exp\big(-\eta\,\Lambda(t)\big)}{|\mathcal{R} \cap \mathcal{S}|} \sum_k^N w_k\, G\big((t - \mu_k)^2 / \sigma_k^2\big). \qquad (7)$$

The scalar $|\mathcal{R} \cap \mathcal{S}|$ is a normalizing constant, representing the 'total intersection' of the ray with the probabilistic scene-model, which can be computed numerically.

The model was evaluated by computing reference visibilities in high-resolution point-clouds [2], then decimating these clouds, and re-estimating the visibility of a large number of test-points. An ROC analysis was performed on the estimated vs. true states (visible/occluded) of the test points. The results indicate that the new model outperforms naive visibility tests, in addition to the theoretical contributions outlined here.

[1] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3D scene labelling. In *Proc. ICRA*, 2014.

[2] G. Schaufler and H.W. Jensen. Ray-tracing point-sampled geometry. In *Proc. Eurographics Workshop on Rendering*, 2000.

[3] A. Yezzi and S. Soatto. Stereoscopic segmentation. *IJCV*, 53(1), 2003.

MGM: A Significantly More Global Matching for Stereovision

Gabriele Facciolo[1]
facciolo@cmla.ens-cachan.fr
Carlo de Franchis[2]
carlo.de-franchis@ens-cachan.fr
Enric Meinhardt[2]
enric.meinhardt@cmla.ens-cachan.fr

[1] IMAGINE/LIGM,
École Nationale des Ponts et Chaussées, France
[2] CMLA, École Normale Supérieure de Cachan, France

Semi-global matching [3] (SGM) is a stereovision algorithm that approximately minimizes a global energy composed of pixel-wise matching cost and pair-wise smoothness terms. The accuracy and speed of SGM are the main reasons for its widespread adoption, even for applications beyond stereovision. In SGM the two-dimensional smoothness constraint is approximated as the average of one-dimensional line optimization problems, which amounts to solving the problem on a star-shaped graph (usually with 8 cardinal directions) centered at each pixel. However, since two adjacent scan lines share little information, this approximation also produces characteristic streaks in the final disparity image (see fig.1).

Based on a recently proposed interpretation of SGM as a min-sum Belief Propagation algorithm [1], we propose in this paper a new algorithm that improves the energy gap of SGM with respect to more comprehensive optimization algorithms. The proposed method comes with no compromises with respect to the baseline SGM, no parameters and virtually no computational overhead. At the same time it yields higher quality results by removing the streaking artifacts of SGM.

SGM formulates stereo matching as finding the disparity map D that minimizes the global energy defined on the graph $G = (I, \mathcal{E})$

$$E(D) = \sum_{\mathbf{p}\in I} C_{\mathbf{p}}(D_{\mathbf{p}}) + \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{E}} V(D_{\mathbf{p}}, D_{\mathbf{q}}), \tag{1}$$

where the unary terms $C_{\mathbf{p}}(d)$ represent the pixel-wise cost of matching $\mathbf{p}$ for disparity $d \in \mathcal{D}$ ($\mathcal{D} = \{d_{min}, \cdots, d_{max}\}$). The pairwise terms $V(\cdot,\cdot)$ enforce smoothness of the solution by penalizing changes of neighboring disparities on the edge set $\mathcal{E}$ (usually the 8-connected image graph). SGM considers truncated pairwise terms of the form (with $P2 > P1$)

$$V(d,d') = \begin{cases} 0 & \text{if } d = d' \\ P1 & \text{if } |d-d'| = 1 \\ P2 & \text{otherwise} \end{cases} . \tag{2}$$

In SGM the 2D problem (1) is splitted into 1D sub-problems defined on scan lines that run through the image in the 8 cardinal directions. For each direction $\mathbf{r}$ SGM recursively computes the costs $L_{\mathbf{r}}$ from the edges of the image along the path in the direction $\mathbf{r}$:

$$L_{\mathbf{r}}(\mathbf{p},d) = C_{\mathbf{p}}(d) + \underbrace{\min_{d'\in\mathcal{D}} (L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d') + V(d,d'))}_{m_{(\mathbf{p}-\mathbf{r})\to(\mathbf{p})}(d)} . \tag{3}$$

The form of the smoothness potential (2) permits to compute $L_{\mathbf{r}}(\mathbf{p},\cdot)$ with just 7 instructions per disparity [2]. The costs $L_{\mathbf{r}}$ computed for all directions $\mathbf{r}$ are then added to obtain the aggregated cost volume from which the final disparity is selected with a Winner-Take-All (WTA) strategy.

The SGM algorithm amounts to the min-sum Belief Propagation algorithm on a star-shaped graph centered at each pixel [1]. That is, the recursive formula (3) is actually computing the state belief of the node $\mathbf{p}$ for each $\mathbf{r}$-oriented path. And the aggregate of state beliefs for the 8 directions ($N_{dir} = 8$)

$$S_{oc}(\mathbf{p},d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p},d) - (N_{dir}-1) C_{\mathbf{p}}(d), \tag{4}$$

corresponds to the min-marginals for the star-shaped graph centered at $\mathbf{p}$.

Min-sum Belief Propagation (BP) [4] can be used as an approximate energy minimization algorithm on a graph. On a generic graph, BP computes each node's belief by sending messages along the edges of the graph. A message from node $\mathbf{q}$ to node $\mathbf{p}$ is defined recursively as

$$m_{\mathbf{q}\to\mathbf{p}}(d) = \min_{d'\in\mathcal{D}} (C_{\mathbf{q}}(d') + \sum_{(\mathbf{q},\mathbf{k})\in\mathcal{E}, \mathbf{k}\neq\mathbf{p}} m_{\mathbf{k}\to\mathbf{q}}(d') + V(d,d')). \tag{5}$$

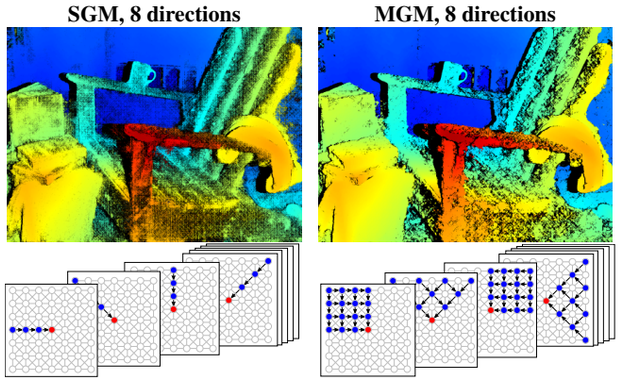


Figure 1: Disparity results of our baseline implementation of SGM and the proposed method MGM for the *Adirondack* pair (Middlebury 2014). The diagrams depict for each method the information used by the recursive update of the red pixel in each of the 8 scans of the algorithms.

The state belief of a node is then computed from the messages as

$$B(\mathbf{p},d) = C_{\mathbf{p}}(d) + \sum_{(\mathbf{q},\mathbf{p})\in\mathcal{E}} m_{\mathbf{q}\to\mathbf{p}}(d). \tag{6}$$

Messages are iteratively updated according to some schedule [5] and upon convergence $\arg\min_d B(\mathbf{p},d)$ yields the estimated solution. For the star-shaped graph associated with SGM [1] equations (3) and (5) are related by $L_{\mathbf{r}}(\mathbf{p},d) = C_{\mathbf{p}}(d) + m_{(\mathbf{p}-\mathbf{r})\to(\mathbf{p})}(d)$ and $S_{oc}(\mathbf{p},d) = B(\mathbf{p},d)$.

More Global Matching[1] (MGM). Our contribution consists in changing the recursive update formula (3). During the left-to-right pass of SGM the image is traversed in raster order (left-right, top-down), but node $\mathbf{p}$ is updated using only the cost of the node on its left $L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},\cdot)$. In MGM we propose (in the spirit of BP) to also consider the costs from the node directly above $\mathbf{p}$ (indicated by $\mathbf{r}^{\perp}$). Because of the raster traversal $L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}^{\perp},\cdot)$ is up-to-date, so the recursion becomes:

$$L_{\mathbf{r}}(\mathbf{p},d) = C_{\mathbf{p}}(d) + \sum_{\mathbf{x}\in\{\mathbf{r},\mathbf{r}^{\perp}\}} \frac{1}{2} \min_{d'\in\mathcal{D}} (L_{\mathbf{r}}(\mathbf{p}-\mathbf{x},d') + V(d,d')). \tag{7}$$

This recursion gathers information from an entire quadrant of the graph, instead of a segment as in SGM (illustrated in fig.1). The costs $L_{\mathbf{r}}$ computed in all traversals are combined using the eq. (4) and the disparity is estimated by WTA. Compared to SGM, MGM only requires a few extra operations per pixel and parallelization is also possible (along diagonals). Since messages from non visited nodes are initialized to 0, each pass of MGM can be seen as the first iteration of a sequential BP algorithm [5].

In summary, MGM produces qualitatively and quantitatively denser results than the baseline SGM with little computation overhead. MGM also yields lower energies than SGM for problem (1).

[1] A. Drory, C. Haubold, S. Avidan, and F. A. Hamprecht. Semi-global matching: a principled derivation in terms of message passing. *Pattern Recogn.*, 2014.

[2] P. Felzenszwalb and D. Huttenlocher. Efficient Belief Propagation for Early Vision. *IJCV*, 2006.

[3] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 2008.

[4] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*, 1988.

[5] M.F. Tappen and W.T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. *ICCV*, 2003.

[1] **On-line demo of MGM is available at: `http://dev.ipol.im/~facciolo/mgm`**

Automated Identification of Individual Great White Sharks from Unrestricted Fin Imagery

Benjamin Hughes
benjamin.hughes@bristol.ac.uk
Tilo Burghardt
tilo@cs.bris.ac.uk

Department of Computer Science
University of Bristol
Bristol, UK

Introduction. Recognising individuals repeatedly over time is a basic requirement for field-based ecology and related life sciences [5, 6]. In this paper we propose a visual identification approach for great white shark fins as outlined in Figure 1, one that is applicable to unconstrained fin imagery and fully automates the pipeline from feature extraction to matching of identities. We pose the associated vision task as a fine-grained, multi-instance classification problem for flexible, smooth and partly occluded object parts. To the best of our knowledge, the proposed system is the first fully automated contour-based animal biometrics system.

Contour Stroke Object Model. A key technical contribution of the paper is a contour stroke model trained for fin detection. It combines a partitioning of ultrametric contour maps (UCM; generated via [3, 4]) with normal descriptors, dense local features and a random-forest regressor. UCM region boundaries are partitioned at keypoints detected as local maxima of the Difference of Gaussian (DoG) function $D(u,\sigma)$ detailed by Zhang et al. [10]. We generate fin candidates as contour strokes by sampling the region contour between every permutation of keypoint pairs. Each stroke is then described by a 160 dimensional feature vector consisting of two components: the first is a bag of opponentSIFT [9] visual words, the second describes boundary shape using a histogram of boundary normals. A random forest regressor is finally trained to predict the fin-like quality of contour stroke candidates. Figure 2 illustrates fins detected by this strategy. In common with the segmentation stage of Arandjelovic and Zisserman's 'Bag of Boundaries' pipeline [2], our contour stroke model separates objects in natural images and against cluttered backgrounds, but differs in that it exploits species characteristic shape encoded in the open contours, to specifically and explicitly detect fin *parts*.

Biometric Encoding: [1, 2] perform smooth object recognition using semi-local boundary normal and HoG+occupancy representations, respectively. These allow efficient and robust matching, but their encoding of inter-class variance will always be sub-maximal. Meanwhile, the semi-automated fin recognition system DARWIN [8] uses a global 2D Cartesian representation. This encoding maximises inter-class variance, but is sensitive to partial occlusions and detection errors. By contrast we utilise semi-local *and* global shape descriptions in a vector-based combinatorial encoding strategy, one that enables efficient and accurate individual recognition while being robust to noisy, partially occluded input.

As with generating contour strokes, we combinatorially sample fin boundaries between pairs of keypoints detected using $D(u,\sigma)$, reparameterised for fine-grained feature detection. We then scale-normalise the resulting contour subsections and encode their shape over various spectral scales. Denoting the set of descriptors for a query object at scale j as D_Q^j, for each query descriptor $d_i^j \in D_Q^j$, we find the two nearest neighbours $(NN_C(d_i^j), NN_{\bar{C}}(d_i^j))$ where C is the class of the nearest neighbour and $\bar{C}$ is the set of all other classes. Subsection information is then combined using a multi-scale decision rule based on the Local Naive Bayes Nearest Neighbour (LNBNN) classifier [7]:

$$\hat{C} = \arg\max_{C} \sum_{j=1}^{v} w_j \cdot \sum_{i=1}^{|D_Q^j|} ||d_i^j - NN_{\bar{C}}(d_i^j)||^2 - ||d_i^j - NN_C(d_i^j)||^2 \qquad (1)$$

Results: We benchmark system performance using a dataset presenting 85 individuals and consisting of 2456 images (one labelled reference image per shark), exhibiting significant variability in waterline and splash occlusion, viewpoint, orientation and scale (see various Figures). Overall, 72% of queries are correctly identified with a mAP of 0.79. Discounting detection errors and partial occlusions, this rises to 82% accuracy with a mAP of 0.84. Fin detection performance depends on the quality threshold t for detected fin candidates, as measured by F-scores. For unoccluded instances we obtain: AP=0.98 for $t > 0.7$ and AP=0.85 for $t > 0.9$.

Conclusion: We have combined a contour stroke model with a spectral and combinatorial contour encoder. The method shows promising individual shark fin identification performance. We conclude applicability at accuracy levels ready to assist human identification without labelling.

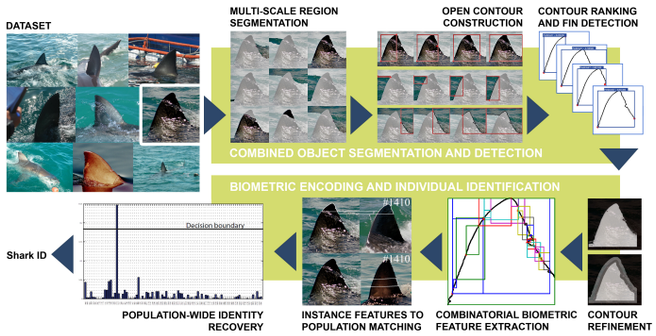


Figure 1: SYSTEM OVERVIEW: The system performs a coarse and a fine-grained recognition task. The first is to simultaneously segment and detect shark fins, and the second is to recognise individuals.

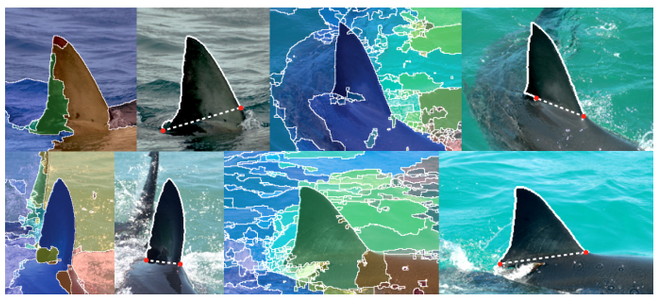


Figure 2: FIN DETECTION AS OPEN CONTOUR STROKES: Segmentation algorithms [5] (left images) alone often fail to detect fins. Our stroke model returns the fin section more robustly as shown in solid white.

[1] Ognjen Arandjelovic. Object matching using boundary descriptors. BMVC 2012.
[2] Relja Arandjelovic and Andrew Zisserman. Smooth object retrieval using a bag of boundaries. ICCV 2011.
[3] Pablo Arbeláez et al. Multiscale combinatorial grouping. CVPR 2014.
[4] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. ICCV 2013.
[5] Hjalmar S Kühl and Tilo Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. Trends in ecology & evolution 2013.
[6] AD Marshall and SJ Pierce. The use and abuse of photographic identification in sharks and rays. Journal of Fish Biology 2012.
[7] Sancho McCann and David G Lowe. Local naive bayes nearest neighbor for image classification. CVPR 2012.
[8] R Stanley. Darwin: Identifying dolphins from dorsal fin images. Eckerd College 1995.
[9] Koen EA Van De Sande et al. Evaluating color descriptors for object and scene recognition. PAMI 2010.
[10] Xiaohong Zhang et al. Robust image corner detection based on scale evolution difference of planar curves. Pattern Recognition Letters 2009.



Figure 3: INDIVIDUAL IDENTIFICATION EXAMPLES: left images are queries and right ones are predicted individuals. Coloured lines indicate ten sections contributing most evidence for the matched individual.

Generating Multi-sentence Natural Language Descriptions of Indoor Scenes

Dahua Lin[1]
dhlin@ie.cuhk.edu.hk
Sanja Fidler[2]
fidler@cs.toronto.edu
Chen Kong[3]
chenk@cs.cmu.edu
Raquel Urtasun[2]
urtasun@cs.toronto.edu

[1] Department of Information Engineering,
The Chinese University of Hong Kong.
[2] Department of Computer Science,
University of Toronto.
[3] Robotics Institute,
Carnegie Mellon University.

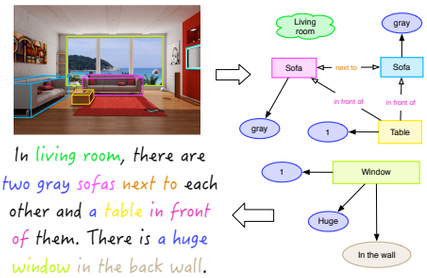


Figure 1: Our method visually parses an RGB-D image to get a *scene graph* that represents objects, their attributes and relations between objects. Based on the scene graph we then generate a multi-sentence textual description via a learned grammar. The description generation takes into account co-reference and saliency of how people describe scenes.

This paper proposes a novel framework for generating lingual descriptions of indoor scenes. This is an important problem, as an effective solution to this problem can enable many exciting real-world applications, such as human robot interaction, image/video synopsis, and automatic caption generation. Whereas substantial efforts have been made to tackle this problem, previous approaches focusing primarily on generating a single sentence for each image, which is not sufficient for describing complex scenes. We attempt to go beyond this, by generating coherent descriptions with multiple sentences.

Particularly, we are interested in generating multi-sentence descriptions of cluttered indoor scenes. Complex, multi-sentence output requires us to deal with challenging problems such as consistent co-referrals to visual entities across sentences. Furthermore, the sequence of sentences needs to be as natural as possible, mimicking how humans describe the scene. This is especially important for example in the context of social robotics to enable realistic communications.

Towards this goal, we develop a framework with three major components: (1) a *holistic visual parser* based on [3] that couples the inference of objects, attributes, and relations to produce a semantic representation of a 3D scene (Fig. 1); (2) a *generative grammar* automatically learned from training text; and (3) a *text generation algorithm* that takes into account subtle *dependencies across sentences*, such as logical order, diversity, saliency of objects, and co-reference resolution.

From RGB-D Images to Semantics. Given an RGB-D image, we extract semantics, such as objects of interest, their attributes, and their physical relations, via visual parsing, and thereon construct a *scene graph*. The detailed procedure is as follows. First, a set of *"objectness" regions* are generated following [1], which are encouraged to respect intensity as well as occlusion boundaries in 3D. These regions are projected to 3D via depth and then cuboids are fit tightly around them, under the constraint that they are parallel to the ground floor.

A *holistic CRF model* is then constructed to jointly reason about the cuboid classes as well as the scene class (*e.g.* kitchen, bathroom). The model exploits various geometric and semantic relations, including *scene appearance*, *cuboid appearance*, *object geometry*, *co-occurrence relations*, and *spatial relations*. These features and relations are incorporated into the CRF formulation as *potentials*. The CRF weights to combine the potentials are learned with a primal dual learning framework [2], and inference of class labels is done with an approximated algorithm [4].

Based on the extracted visual information, we construct a *scene graph*, with *nodes* representing objects and their attributes, and *edges* representing relations between nodes. There are three kinds of edges: *attribute edges* that link objects to their attributes, *position edges* that represent the positions of objects relative to the scene, (*e.g. corner-of-room*), and *pairwise edges* that characterize the relative positions between objects (*e.g. on-top-of* and *next-to*).

Generating Lingual Descriptions. Given a *scene graph*, we generate a descriptive paragraph in two steps. First, we transform the scene graph into a sequence of *semantic trees*. Then, we produce sentences, one from each semantic tree, following a *generative grammar*. A *semantic tree* contains a set of *terminal nodes* corresponding to individual entities or their attributes and *relational nodes* that express relations among them. Such a tree can capture information like *what entities are being described* and *what are the relationships between them*. To generate sentences in a coherent manner, we further devise a method that transforms a *scene graph* into a sequence of *semantic trees*, with multiple kinds of dependencies among sentences taken into account, including *logical order*, *diversity*, *saliency*, *co-reference*, and *richness vs. conciseness*.

For each *semantic tree*, we produce a sentence via a *generative grammar*, *i.e.* a map from each semantic relation to a set of templates (derivation rules). The *grammar* for generating sentences are often specified manually in previous work. This way, however, is time consuming, unreliable, and tends to oversimplify the language. Here, we explore a new approach, that is, to learn the grammar from data. The basic idea is to construct a semantic tree from each sentence through linguistic parsing, and then derive the templates by matching nodes of the semantic tree to parts of the sentence.

Experimental Evaluation. We tested the proposed framework on the NYU-v2 dataset [5] augmented with an additional set of textual descriptions, one for each image. Below are two representative examples. Please refer to the paper for more details about the experiments.



Figure 2: Examples of descriptions generated using our framework.

[1] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012.

[2] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010.

[3] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

[4] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011.

[5] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Overlapping Domain Cover for Scalable and Accurate Regression Kernel Machines

Mohamed Elhoseiny
m.elhoseiny@cs.rutgers.edu
Ahmed Elgammal
elgammal@cs.rutgers.edu

Computer Science Department
Rutgers University
New Jersey, USA

Table 1: Contrast against most relevant methods

	[5]	FIC/PIC [7]	NN [1]	ODC (ours)
Accurate	No for high input dimension	Limited	Yes	Yes
Efficient	No	Yes	No	Yes
Scalable to arbitrary input dimension	No (2D)	Yes	Yes	Yes
Consistent on Boundaries	Yes	No	Yes	Yes
supported kernel machines	GPR	GPR	TGP	GPR, TGP, IWTGP and others
Easy to parallelize	No	No	Yes	Yes

Motivation. Recent advances in structure regression encouraged researchers to adopt it for formulating various problems with high dimensional output spaces, such as segmentation, detection, and image reconstruction, as regression problems. However, the computational complexity of the state-of-the-art regression algorithms limits their applicability for big data. In particular, kernel-based regression algorithms such as Ridge Regression [3], Gaussian Process Regression (GPR) [6], and the Twin Gaussian Processes (TGP) [1] require inversion of kernel matrices ($O(N^3)$), where N is the number of the training points), which limits their applicability for big data. We refer to these non-scalable versions of GPR and TGP as full-GPR and full-TGP, respectively.

The problems of the existing approximation approaches, detailed and justified in the paper, motivated us to develop an approach that satisfies the properties listed in table 1. The table also shows which of these properties are satisfied for the relevant methods. Khandekar et. al. [4] discussed properties and benefits of overlapping clusters for minimizing the conductance from spectral perspective. These properties of overlapping clusters also motivate studying scalable local prediction based on overlapping kernel machines; see figure 1.

Our Contribution. In summary, the main question, we address in this paper, is how local kernel machines with overlapping training data could help speedup the computations and gain accurate predictions. We achieved considerable speedup and good performance on GPR, TGP, and IWTGP (Importance Weighted TGP) applied to 3D pose estimation datasets. To the best of our knowledge, our framework is the first to achieve quadratic prediction complexity for TGP. The ODC concept is also novel in the context of kernel machines and is shown here to be successfully applicable to multiple kernel-machines. We also theoretically justified the idea behind our method and build on it to propose an ODC framework that reduces the complexity of TGP regression from cubic to quadratic. As a part of the framework, we proposed Assign&Balance K-Means algorithm, a version of K-means clustering that generates equal size clusters and we showed that it better than RPC used previously for GPR; see details in the main paper. We validated and analyzed our method on three human pose estimation datasets and interesting findings are discussed.

ODC Framework Overview We define the ODC as a collection of overlapping subsets of the training points, denoted by subdomains, such that they are as spatially coherent as possible. During training, an ODC is computed such that each subdomain overlaps with the neighboring subdomains. Then, a local prediction model (kernel machine) is created for each subdomain and the computations that does not depend on the test data are factored out and precomputed (e.g. inversion of matrices). The nature of the ODC generation makes these kernel machines consistent in the overlapped regions, which are the boundaries since we constraint the subdomains to be coherent. On prediction, the output is calculated as a

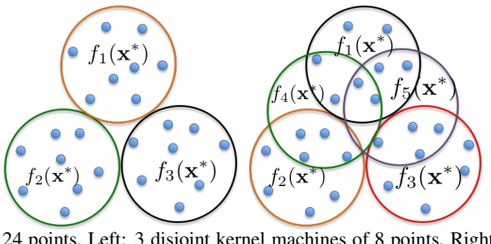


Figure 1: 24 points, Left: 3 disjoint kernel machines of 8 points, Right: 5 Overlapping kernel machines of 8 points. $f_i(\mathbf{x}^*)$ is the ith kernel machine prediction for $\mathbf{x}^*$ test point.

Table 2: Error & Time on Poser and Human Eva datasets (Intel core-i7 2.6GHZ), M = 800

		Poser			HumanEva		
		Error (deg)	Training Time	Prediction Time	Error (mm)	Training Time	Prediction Time
TGP	NN [1]	5.43	-	188.99 sec	**38.1**	-	6364 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-**Ekmeans**	**5.40**	(3.7 +25.1) sec	**16.5** sec	**38.9**	(2001 + 45.4) sec	**298** sec
	ODC ($p = 0, t = 1, K' = 1$)-**Ekmeans**	7.60	(3.9 + 1.33) sec	14.8 sec	41.87	(240 + 4.9) sec	257 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-**RPC**	5.60	(0.23 +41.6) sec	15.8 sec	39.9	(0.45 + 49.1) sec	277 sec
	ODC ($p = 0, t = 1, K' = 1$)-**RPC**	7.70	(0.15 + 1.7) sec	13.89 sec	42.32	(0.19 + 5.2) sec	242 sec
GPR	NN	6.77	-	24 sec	54.8	-	618 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-**Ekmeans**	**6.27**	(3.7 +11.1) sec	**0.56** sec	**49.3**	(2001 + 42.85)sec	**79** sec
	ODC($p = 0.0, t = 1, K' = 1$)-**Ekmeans**	7.54	(3.9 + 1.38 sec)	0.35 sec	49.6	(240 + 6.4) sec	48 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-**RPC**	6.45	(0.23 +17.3) sec	0.52 sec	52.8	(0.49 + 46.06) sec	64 sec
	ODC ($p = 0.0, t = 1, K' = 1$)-**RPC** = [2]	7.46	(0.15 + 1.5) sec	0.27 sec	54.6	(0.26 + 4.6) sec	44 sec
	FIC [7]	7.63	(- + 20.63)	0.3106	68.36	-	102 sec

reduction function of the predictions on the closed subdomain(s).

Given a set of input data $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, our prediction framework firstly generates a set of non-overlapping equal-size partitions, $C = \{C_1, \cdots, C_K\}$, such that $\cup_i C_i = X$, $|C_i| = N/K$. Then, the ODC is defined based on them as $\mathcal{D} = \{D_1, \cdots, D_K\}$, such that $|D_i| = M \forall i$, $D_i = C_i \cup O_i, \forall i$. O_i a the set of points that overlaps with the other partitions, i.e., $O_i = \{x : x \in \{\cup_{j \neq i} C_j\}\}$, such that $|O_i| = p \cdot M$, $|C_i| = (1-p) \cdot M$, $0 \leq p \leq 1$ is the ratio of points in each overlapping subdomain, D_i, that belongs to/overlaps with partitions, other than its own, C_i.

An ODC could be specified by two parameters, M and p, which are the number of points in each subdomain and the ratio of overlap respectively; this is since $K = N/(1-p)M$. As p goes to 0, the generated ODC reduces to the set of non-overlapping clusters. Similarly, as p approaches $1 - 1/M$, the ODC reduces to generating a cluster at each point with maximum overlap with other clusters, i.e., $K = N$, $|C_i| = 1$, and $|O_i| = M - 1$. *Our main claim is two fold. First, precomputing local kernel machines (e.g. GPR, TGP, IWTGP) during training on the ODC significantly increase the speedup on prediction time. Second, given a fixed M and N, as p increases, local prediction performance increases, theoretically supported by Lemma 4.1.* Detailed about training and prediction could be found in the main paper.

Lemma 4.1. Under ODC notion, as the overlap p increases, the closer the nearest model to an arbitrary test point and the more likely that model get trained on a big neighborhood of the test point; see the proof in the Supplementary Materials (SM).

Experiments. We validated our framework on Poser, HumanEva, and Human3.6M datasets for human pose estimation task. Table 2 shows comparison between our method and the baseline approximation methods on Poser and HumanEva datasets; details could be found in the paper. We also tried full TGP and GPR on Poser and Human Eva Datasets. Full TGP error is 5.35 for Poser and 40.3 for Human Eva. Full GPR error is 6.10 for Poser and 59.62 for Human Eva. The results indicate that ODC achieves either better or competitive to the full models. Based on our comprehensive experiments on HumanEva and Poser datasets, we conducted an experiment on Human3.6M dataset with TGP kernel machine, where $M = 1390$, $t = 1$, $p = 0.6, K' = 1$, Ekmeans for clustering. We achieved a speedup of $41.7X$ on prediction time using our ODC framework compared with NN-scheme, i.e., 7 days if NN-scheme is used versus 4.03 hours in our case. More experiments and details are in the paper.

[1] Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *Int. J. Comput. Vision*, 87(1-2):28–52, March 2010. ISSN 0920-5691.

[2] Krzysztof Chalupka, Christopher K. I. Williams, and Iain Murray. A framework for evaluating approximation methods for gaussian process regression. *JMLR*, 14(1), February 2013.

[3] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1970.

[4] Rohit Khandekar, Guy Kortsarz, and Vahab Mirrokni. On the advantage of overlapping clusters for minimizing conductance. *Algorithmica*, 69(4):844–863, 2014.

[5] Chiwoo Park, Jianhua Z. Huang, and Yu Ding. Domain decomposition approach for fast gaussian process regression of large spatial data sets. *Journal of Machine Learning Research*, 12:1697–1728, 2011.

[6] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

[7] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *NIPS*, 2006.

Visual Comparison of Images Using Multiple Kernel Learning for Ranking

Amr Sharaf[1]
amr.sharaf@alexu.edu.eg
Mohamed E. Hussein[2,1]
mohamed.e.hussein@ejust.edu.eg
Mohamed A. Ismail[1]
maismail@alexu.edu.eg

[1] Alexandria University
El-Shatby, Alexandria 21526, Egypt
[2] Egypt-Japan University of Science and Technology (E-JUST)
New Borg Al-Arab, Alexandria 21934, Egypt

Abstract

Ranking is the central problem for many applications such as web search, recommendation systems, and visual comparison of images. In this paper, the multiple kernel learning framework is generalized for the learning to rank problem. This approach extends the existing learning to rank algorithms by considering multiple kernel learning and consequently improves their effectiveness. The proposed approach provides the convenience of fusing different features for describing the underlying data. As an application to our approach, the problem of visual image comparison is studied. Several visual features are used for describing the images and multiple kernel learning is adopted to find an optimal feature fusion. Experimental results on three challenging datasets show that our approach outperforms the state-of-the art and is significantly more efficient in runtime.

Proposed Approach (RankMKL)

Given two images, it is required to learn which image exhibits a particular visual attribute more than the other. Our approach works on a per attribute basis, thus a separate model is learned for each visual attribute. Figure 1 demonstrates the outline of our approach. The first step is to extract a set of features from each image. Several feature sets are selected to capture different visual cues in the image. To capture the image texture, we extract Local Binary Patterns (LBP) [3] and compute the response from a set of Gabor filters. For capturing the shape and appearance of the images, GIST [4] and HoG [1] descriptors are used. Finally, a color histogram is computed in the LAB color space to capture the color information.

The second step is to fuse the different feature sets and learn the ranking model. For this task, a separate kernel function is computed for each set of features (i.e. we compute five different kernels). The computed kernels are considered as base kernels for our multiple kernel learning module. Using the multiple kernel learning algorithm, we learn the optimal weights for creating a linear combination from the base kernels together with the optimal parameters for the ranking model.

Instead of using a single kernel matrix (K) for learning the ranking model, an optimal combination from several base kernels is learned, and the combination of the base kernels matrix $(K_{\mathbf{d}})$ is used for training the ranking model, where $k_{\mathbf{d}}(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T_{\mathbf{d}}\phi(\mathbf{x}_j)_{\mathbf{d}}$ represents the dot product in feature space ϕ and is parametrized by $\mathbf{d}$ such that:

$$k_{\mathbf{d}}(\mathbf{x}_i,\mathbf{x}_j) = f_{\mathbf{d}}(\{k_i(\mathbf{x}_i,\mathbf{x}_j)\}_{i=1}^{t}), \tag{1}$$

where t is the number of base kernels, $\mathbf{d} \in \mathbb{R}^t$ is the optimal kernel weights to be learned, and the combination function $f_{\mathbf{d}}$ can be a linear or a non-linear function for combining the base kernels . Our goal is to learn the optimal values for $(\mathbf{d})$ together with the optimal values for the Lagrange multipliers (α) representing the learned ranking model. Accordingly, the standard rankSVM [2] objective function is updated as follows:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \{\mathbf{1}^{\mathbf{T}}\alpha - \frac{1}{2}\alpha^T Q_{\mathbf{d}}\alpha + r(\mathbf{d})\} \\ \text{subject to} \quad & 0 \le \alpha_{i,j} \le C, \forall (i,j) \in P, \\ & \mathbf{d} \ge 0, \end{aligned} \tag{2}$$

$$Q_{\mathbf{d},(i,j),(u,v)} = k_{\mathbf{d}}(\mathbf{x}_i,\mathbf{x}_u) + k_{\mathbf{d}}(\mathbf{x}_j,\mathbf{x}_v) - k_{\mathbf{d}}(\mathbf{x}_i,\mathbf{x}_v) - k_{\mathbf{d}}(\mathbf{x}_j,\mathbf{x}_u), \tag{3}$$

where both the regularizer r and the kernel $k_{\mathbf{d}}$ can be any general differentiable functions of $\mathbf{d}$ with continuous derivatives and P represents the set of preference paris such that: $P = \{(i,j)|\mathbf{x}_i \succ \mathbf{x}_j\}$. In our approach, five base kernels are used, one for each of the five feature sets (LBP, HoG, Gabor, GIST, and Color). The kernel function $k_{\mathbf{d}}$ is selected as a linear combination from the five base kernels: $k_{\mathbf{d}}(\mathbf{x}_u,\mathbf{x}_v) = \sum_{i=1}^{5} d_i k_i(\mathbf{x}_u,\mathbf{x}_v)$ and L2 regularization function is used for $r(\mathbf{d})$. Gradient descent is used for solving Eq. (2) using the same algorithm in [5].

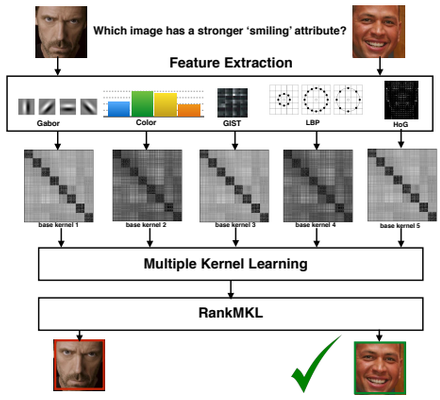


Figure 1: Illustration of our proposed approach. Given two images, it is required to detect which image has a stronger visual attribute than the other. Different features are extracted and Multiple Kernel Learning (MKL) is used for fusing the kernels from each feature set. RankMKL is used for ranking the images.

Conclusion

In this paper, the standard multiple kernel learning formulation is extended to the learning to rank problem. Effectiveness of the proposed approach is demonstrated on the visual image comparison task. Although MKL has been extensively used for object recognition and image categorization, this is the first time it has been used for image comparison. Through extensive experiments, the advantage of our approach is clearly demonstrated both in terms of accuracy and runtime efficiency. Future work includes exploring more applications of multiple kernel learning for ranking, such as web search and recommendation systems.

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[2] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[3] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[4] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[5] Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.

Bridging the Domain Shift by Domain Adaptive Dictionary Learning

Hongyu Xu
hyxu@umiacs.umd.edu
Jingjing Zheng
zjngjng@umiacs.umd.edu
Rama Chellappa
rama@umiacs.umd.edu

Department of Electrical and Computer Engineering and the
Center for Automatic Research, UMIACS
University of Maryland
College Park, Maryland, USA

Domain adaptation (DA) tackles the problem where data from the training set (**source** domain) and test set (**target** domain) have different underlying distributions. For instance, training and testing images may be acquired under different environments, viewpoints and illumination conditions. In this paper, we focus on the more challenging unsupervised DA problem where the samples in the target domain are unlabeled. It is noticed that dictionary learning has gained a lot of popularity due to the fact that images of interest could be reconstructed sparsely in an appropriately learned dictionary [1].

Specifically, we propose a novel domain-adaptive dictionary learning approach to generate a set of intermediate domains which bridge the gap between source and target domains. Our approach defines two types of dictionaries: a common dictionary and a domain-specific dictionary. The overall learning process illustrated in Figure 1 consists of three steps: (1) At the beginning, we first learn the common dictionary D^C, domain-specific dictionaries D^0 and D^t for source and target domains. (2) At the k-th step, we enforce the recovered feature representations of target data in all available domains to have the same sparse codes, while adapting the most recently obtained dictionary D^k to better represent the target domain. Then we multiply dictionaries in the k-th domain with the corresponding sparse codes to recover feature representations of target data X_t^k in this domain. (3) We update D^k to find the next domain-specific dictionary D^{k+1} by further minimizing the reconstruction error in representing the target data. Then we alternate between the steps of sparse coding and dictionary updating until the stopping criteria is satisfied.

Notations: Let $X_s \in \mathbb{R}^{d\times N_s}$, $X_t \in \mathbb{R}^{d\times N_t}$ be the feature representations of source and target data respectively, where d is the feature dimension, N_s and N_t are the number of samples in the two domains. The feature representations of recovered source and target data in the k-th intermediate domain are denoted as $X_s^k \in \mathbb{R}^{d\times N_s}$ and $X_t^k \in \mathbb{R}^{d\times N_t}$ respectively. The common dictionary is denoted as D^C, whereas source-specific and target-specific dictionaries are denoted as D^0, D^t respectively. Similarly, we use $D^k, k = 1...N$ to denote the domain-specific dictionary for the k-th domain, where N is the number of intermediate domains. We set all the dictionaries to be of the same size $\in \mathbb{R}^{d\times n}$. At the beginning, we learn the common dictionary D^C by minimizing the reconstruction error of both source and target data as follows:

$$\min_{D^C,Z^0,Z^t} ||X_s - D^C Z^0||_F^2 + ||X_t - D^C Z^t||_F^2 \quad s.t.\ \forall i, \quad \|z_i^0\|_0 \le T, \|z_i^t\|_0 \le T \tag{1}$$

where $Z^0 = [z_1^0 ... z_{N_s}^0] \in \mathbb{R}^{n\times N_s}, Z^t = [z_1^t ... z_{N_t}^t] \in \mathbb{R}^{n\times N_t}$ are sparse representations of X_s and X_t respectively, T is the sparsity level.

Then we learn domain-specific dictionaries D^0 and D^t by further reducing the reconstruction error of the source and target data as follows:

$$\min_{D^0,\Gamma^0} \|X_s - D^C Z^0 - D^0\Gamma^0\|_F^2 + \lambda \|D^0 D^{C^T}\|_F^2 \quad s.t.\ \forall i, \quad \|z_i^0\|_0 + \|\alpha_i^0\|_0 \le T \tag{2}$$

$$\min_{D^t,\Gamma^t} \|X_t - D^C Z^t - D^t\Gamma^t\|_F^2 + \lambda \|D^t D^{C^T}\|_F^2 \quad s.t.\ \forall i, \quad \|z_i^t\|_0 + \|\alpha_i^t\|_0 \le T \tag{3}$$

where $\Gamma^0 = [\alpha_1^0 ... \alpha_{N_s}^0] \in \mathbb{R}^{n\times N_s}$ and $\Gamma^t = [\alpha_1^t ... \alpha_{N_t}^t] \in \mathbb{R}^{n\times N_t}$ are sparse representations of X_s and X_t with respect to D^0 and D^t, and λ is the regularization parameter.

Next, we alternate between domain-adaptive sparse coding and dictionary updating to learn a set of domain-specific dictionaries for generating intermediate domains. We regularize the sparse representation of X_s, X_t and X_t^i to be the same as follows:

$$\begin{aligned} Z^k, \Gamma^k = &\arg\min_{Z,\Gamma} \|X_t - D^C Z - D^k\Gamma\|_F^2 + \sum_{i=0}^{k-1} \|X_t^i - D^C Z - D^i\Gamma\|_F^2 \\ &+ \|X_t - D^C Z - D^t\Gamma\|_F^2 \quad s.t. \quad \forall i, \|z_i\|_0 + \|\alpha_i\|_0 \le T \end{aligned} \tag{4}$$

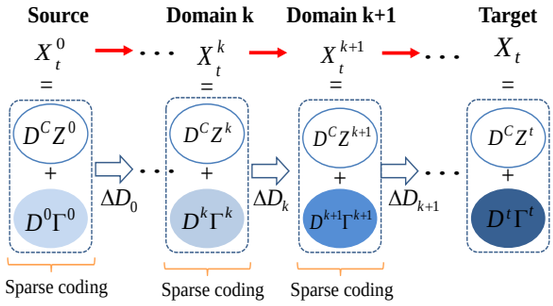


Figure 1: The overall learning process in domain-adaptive dictionary learning framework.

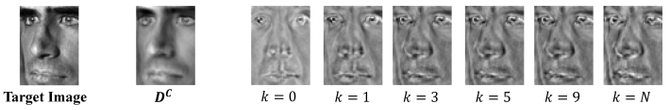


Figure 2: The first image is the original target face image, the second image is the recovered face image corresponding to the common dictionary. The remaining six images are the recovered face images corresponding to domain-specific dictionaries.

Then after sparse coding at the k-th step, we update D^k to find the next domain-specific dictionary D^{k+1} by further reducing the reconstruction error of target data in the k-th domain as in [2].

$$\min_{\Delta D^k} \|J^k - \Delta D^k \Gamma^k\|_F^2 + \eta \|\Delta D^k\|_F^2 \tag{5}$$

where $J^k = X_t - D^C Z^k - D^k\Gamma^k$ is the target reconstruction residue in the k-th domain. The next domain-specific dictionary D^{k+1} is obtained as: $D^{k+1} = D^k + \Delta D^k$.

We use the sparse codes obtained in the last iterative step to derive new feature representations for the source and target data. The new augmented feature representation of source and target data are $\tilde{X}_s = [\tilde{X}_s^0, ..., \tilde{X}_s^N]$ and $\tilde{X}_t = [\tilde{X}_t^0, ..., \tilde{X}_t^N]$ respectively, where $\tilde{X}_s^i = D^C Z_s^N + D^i \Gamma_s^N$ and $\tilde{X}_t^i = D^C Z_t^N + D^i \Gamma_t^N$ and Z_s^N, Z_t^N, Γ_s^N, Γ_t^N are the sparse codes obtained in the last iterative step where $k = N$. The final stage of recognition across all the domains is performed using an SVM classifier trained on new feature vectors after dimension reduction via the Principal Component Analysis (PCA).

Visualization: We demonstrate the benefits of the separation of the common dictionary from domain-specific dictionaries. We select frontal face images as the source domain and face images at 45° yaw as the target domain. We visualize the recovered face images along the intermediate domains in Figure 2. Each recovered face image has two components corresponding to the learned dictionaries respectively.It can be seen that the components corresponding to domain-specific dictionaries in intermediate domains gradually adapt from the frontal face to non-frontal faces. This demonstrates that the domain-specific dictionaries have the ability to encode the domain shift caused by different yaw angles.

We also extensively evaluate our approach using the DA benchmark dataset [3] and experimental results clearly demonstrate the effectiveness of our approach.

[1] M Aharon, M Elad, and A Bruckstein. K-SVD : An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] Jie Ni, Qiang Qiu, and Rama Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*, pages 692–699, 2013.

[3] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

Joint Feature Selection with Low-rank Dictionary Learning

Homa Foroughi
homa@ualberta.ca
Moein Shakeri
shakeri@ualberta.ca
Nilanjan Ray
nray1@ualberta.ca
Hong Zhang
hzhang@ualberta.ca

Department of Computing Science
University of Alberta
Edmonton, Canada

In many areas, such as computer vision and pattern recognition, data are characterized by high dimensional feature vectors. If these vectors are processed directly, it usually leads to difficult pattern recognition task. However, in practice, only a small subset of features is really important and discriminative. Feature selection (FS) is one of the well known dimensionality reduction methods that efficiently describes the input data by removing irrelevant variables and reduces the effects of noise to provide good prediction results.

Recently, Yan *et al*. [2] introduced the sparse representation-based classification (SRC) [1] measurement criterion into FS and designed a joint sparse discriminative FS method. Based on the assumption of SRC, their method selects a subset of features which minimize the within-class reconstruction residual and simultaneously maximize the between-class reconstruction residual in the subset of selected features. Although they can achieve promising results compared to other FS methods, it is well-known that SRC suffers from major drawbacks such as high computational complexity and more importantly, low discriminativity of sparse coefficients and naive dictionary. Consequently, the reconstruction scatter matrices would not preserve the reconstructive relationship of data well, which means the selected features are not discriminant enough.

To overcome the drawbacks associated with the SRC algorithm, in this paper we propose a new FS method by learning a smaller-sized dictionary while maintaining the sparse reconstruction relationship among samples. In order to generate discriminative dictionary and sparse coefficients, we utilize the power of a supervised dictionary learning (DL) method called Fisher discrimination dictionary leaning (FDDL) [3]. The discrimination capability of FDDL originates from two facts. First, each sub-dictionary is trained to have good representation power to the samples from the corresponding class, but have poor representation power to the samples from other classes. Second, the sparse coefficients are made discriminative through minimizing the within-class scatter and maximizing the between-class scatter of them. Both of these properties make FDDL a good choice for finding the discriminative sparse coefficients of training samples. Nevertheless, all the DL algorithms, including FDDL, work well when the input images are clean or corrupted by small noise and their performance deteriorates when the training data is contaminated because of occlusion, disguise or lighting variations [4]. Low-rank (LR) matrix recovery, which determines a LR data matrix from corrupted data, has been successfully applied to different tasks including image classification. To improve the performance of FDDL with noises, we integrate LR matrix approximation into sparse representation for DL. Specifically, we formulate FS problem under LR dictionary learning with Fisher discrimination regularization and propose a Joint Feature Selection method using Low-rank Dictionary Learning (JFS-LDL) as follows:

(a) **Low-rank Approximation**: Given a set of training data vectors $X = [X_1, X_2, \ldots, X_K]$, the samples from i^{th} class, X_i, are linearly correlated in many situations. LR matrix recovery seeks to decompose a data matrix X_i into $L_i + E_i$ by minimizing the rank of matrix L_i, while reducing $\|E_i\|_0$, the associated sparse noise [4]. Clearly, LR reduces the diversity across items within each class and consequently dissimilarity between different classes is increased, which means sub-dictionaries are more discriminant toward each other.

(b) **Low-rank Dictionary Learning using Fisher Discrimination**: To improve the performance of FDDL with noises and contamination, we use a more robust representation of X_i, which is basically its LR representation, L_i in the FDDL objective function. Furthermore, when the standard LR matrix recovery is combined with Fisher discrimination, the images tend to be more similar to each other for the same class, which means more compactness exists within the same class and dissimilarity between different classes. Therefore, the learned sub-dictionaries would have better discrimination and reconstruction capabilities compared to FDDL model. Accordingly, the quality of structured dictionary will influence the discriminativeness of the sparse coefficients.

(c) **Joint Feature Selection**: We aim to select a subset of features that preserve the sparse reconstructive relationship of the training samples. This is achieved by minimizing the within-class reconstruction residual error and simultaneously maximizing the between-class reconstruction residual in the subset of selected features. We exploit the sparse coefficients obtained by objective function of LR dictionary learning using Fisher discrimination to obtain the within-class and between-class reconstructive scatter matrices. Simultaneously, the $l_{2,1}$-norm minimization on projection matrix is applied to jointly select the most relevant and discriminative features. As a result, the projected samples are more discriminative and simultaneously retain the important properties for classification, e.g., intraclass compactness and interclass separability, as well as the reconstructive relationship.

We conduct extensive experiments on benchmark datasets including handwritten digits, face and sport action to verify the effectiveness of the proposed JFS-LDL in comparison with other FS methods and validate its capability for image classification task. Experiments show that JFS-LDL consistently outperforms all the other evaluated FS methods, especially in lower dimensions in image classification task. JFS-LDL can maintain a relatively stable performance under different dimensions, and as the number of selected features decreases, its advantage becomes more obvious. The combination of LR approximation and Fisher discrimination, leads in more compactness within the same class and dissimilarity between different classes. Consequently, a simple and fast classifier like KNN or SVM would perform well for classification. In contrasts to most of DL methods which use l_1-optimization to find the representation of test images and use the reconstruction error for classification, our classification schema is very efficient and fast. Furthermore, we observe that JFS-LDL with a selected subset of features can achieve superior or competitive recognition rate compared to the recently proposed DL methods, with much higher feature dimensions. This implies the effectiveness of our method in capturing the discriminative information for classification. The experimental results together with the theoretical analysis validate the effectiveness of our method for feature selection, and its efficacy for image classification.

[1] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2): 210–227, 2009.

[2] Hui Yan and Jian Yang. Sparse discriminative feature selection. *Pattern Recognition*, 2014.

[3] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3): 209–232, 2014.

[4] Yangmuzi Zhang, Zhuolin Jiang, and Larry S Davis. Learning structured low-rank representations for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 676–683. IEEE, 2013.

Scalable Visual Instance Mining with Instance Graph

Wei Li[1]
liwei2658@gmail.com
Changhu Wang[2]
chw@microsoft.com
Lei Zhang[3]
leizhang@microsoft.com
Yong Rui[2]
yongrui@microsoft.com
Bo Zhang[1]
dcszb@mail.tsinghua.edu.cn

[1] State Key Lab of Intelligent Technology and Systems, TNList, Department of Computer Science and Technology, Tsinghua University Beijing 100084, China

[2] Microsoft Research No. 5 Danling Street, Haidian District, Beijing 100080, China

[3] Microsoft Research One Microsoft Way, Redmond WA 98052, USA

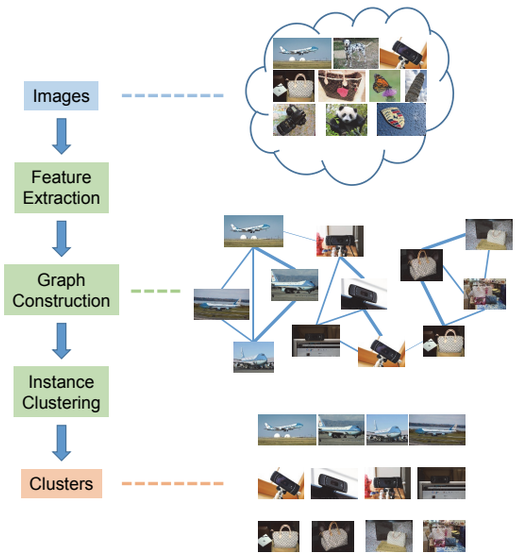


Figure 1: The proposed framework for visual instance mining.

Visual instances are the basic re-occurring information in the visual world. In this paper we address the problem of visual instance mining. Specifically, the goal is to automatically discover frequent visual instances from a collection of images. We are interested in mining specific instances, for example, Air Force One and Monarch Butterfly, which are different from high-level visual categories like plane and butterfly. This task plays a fundamental role in many applications such as multimedia summarization [7] and image annotation [6].

The problem of visual instance mining brings two challenges. The first one is the mining of small instances that have large variations. In many cases, visual instances may only cover very limited image areas, which sometimes can hardly be noticed even by humans. Different types of variations, such as scale, rotation, and occlusion, especially the variations resulted from out-of-plane rotation and non-rigid objects, require highly robust algorithms. The second challenge is the scale of the dataset. A large-scale database is essential for discovering practically useful visual instances. However, large-scale databases usually contain more noises that significantly affect the performance of the methods designed on small databases. Also, the need of high efficiency rules out complex and non-parallelizable methods.

Most existing work focused on large instance mining, or image clustering. An inspiring work by Philbin and Zisserman [4] constructed a matching graph by searching with every image as a query in the dataset and then finds its dense sub-graphs. Their matching graph is a powerful way to represent image-level context properties, but it suffers from a low recall due to the whole-image search with spatial verification. Among a few methods that target for scalable small instance mining, geometric min-hashing [1] and thread of features [9] are two representative works. These methods are capable of discovering small instances via instance-level local matches, but they do not consider the structure of connections between images.

To tackle these two challenges, we propose a novel method that is robust and scalable from the graph perspective for mining both large and small instances. We argue that a graph structure constructed by local matches not only helps to capture the similarities between instances, but also reflects image-level context properties via indirect paths, thus combining the advantages of [4] and [9]. Figure 1 shows the framework.

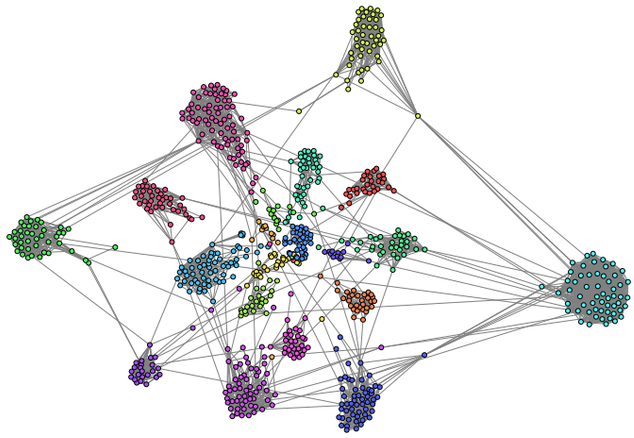


Figure 2: The instance graph of PartialDup dataset.

First of all we extract local SIFT [3] features, quantize them [5], and augment each local feature with the Hamming Embedding [2] binary code of its feature vector and neighbor features. Then we build a weighted and undirected instance graph with images as vertices. The similarity scores between the augmented local features, which combine the HE weighting scheme [2] and the Jaccard similarity of neighboring visual word sets, contribute to the weights of edges between images. As an example, Figure 2 demonstrates an instance graph of PartialDup [8] dataset. From the sparse instance graph we are able to efficiently discover instance clusters by the proposed greedy breadth-first search algorithm. Experiments show that our method outperforms the state-of-the-art ToF method on both clustering performance and running speed.

[1] O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: finding a (thick) needle in a haystack. In *CVPR*, 2009.

[2] H. Jegou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 2010.

[3] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[4] J. Philbin and A. Zisserman. Object mining using a matching graph on very large image collections. In *ICVGIP*, 2008.

[5] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, 2003.

[6] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma. Arista - image search to annotation on billions of web photos. In *CVPR*, 2010.

[7] X.-J. Wang, Z. Xu, L. Zhang, C. Liu, and Y. Rui. Towards indexing representative images on the web. In *ACM MM*, 2012.

[8] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, 2009.

[9] W. Zhang, H. Li, C.-W. Ngo, and S.-F. Chang. Scalable visual instance mining with threads of features. In *ACM MM*, 2014.

Perceptually motivated benchmark for video matting

Mikhail Erofeev[1]
merofeev@graphics.cs.msu.ru
Yury Gitman[1]
ygitman@graphics.cs.msu.ru
Dmitriy Vatolin[1]
dmitriy@graphics.cs.msu.ru
Alexey Fedorov[1]
afedorov@graphics.cs.msu.ru
Jue Wang[2]
juewang@adobe.com

[1] Lomonosov Moscow State University
Moscow, Russia
[2] Adobe Systems
Seattle, WA, United States



Figure 1: Preview frames of our test sequences.

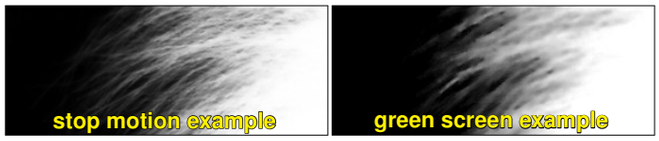


Figure 2: Alpha mattes from chroma keying and stop-motion capture for the same image region. The stop-motion result is significantly better at preserving details.

Despite recent progress in the field of video matting, neither public data sets nor even a generally accepted method of measuring quality has yet emerged. In this paper we present an online benchmark for video-matting methods. Using chroma keying and a reflection-aware stop-motion capturing procedure, we prepared 12 test sequences. Then, using subjective data, we performed extensive comparative analysis of different quality metrics. The goal of our benchmark is to enable better understanding of current progress in the field of video matting and to aid in developing new methods.

Formally, matting is an inverse alpha-compositing problem: i.e., given pixel I, we want to find transparency value $\alpha \in [0;1]$, foreground pixel F, and background pixel B so that

$$I = \alpha F + (1-\alpha)B. \quad (1)$$

The problem is ill posed yet solvable by considering the affinity of pixels in natural images. Matting of natural images is well studied, and according to [3], natural-image matting algorithms are continuously improving.

Video matting is a relatively new research direction that arose recently as available processing power increased. Applied to video, matting has two special requirements: tolerance of sparse user input and temporal coherence of the resulting transparency values.

Despite the rising interest, research in the field of video matting is still weakly organized. In fact, many developers estimate the quality of their methods by visual comparison [1, 2, 4].

The two main challenges facing an effective comparison are preparation of the data set and choice of a quality metric. In this paper we address both challenges and describe a benchmark, available at http://videomatting.com, that provides a comparison, two training sequences with ground-truth transparency, and multiple visualizations for convenient analysis of the comparison results.

To prepare the data set (see Figure 1), we imposed four requirements on our test sequences: high quality for the ground-truth transparency, natural appearance, complexity, and diversity. To satisfy the first two requirements, we used two different techniques of foreground-object capture: namely, capture in front of a green screen and sequential photography (stop motion) against different backgrounds. Chroma keying enabled us to obtain alpha mattes of natural-looking objects with arbitrary motion. Nevertheless, this technique cannot guarantee that the alpha maps are natural, because it assumes the screen color is absent from the foreground object (see Figure 2). To get alpha maps that have a very natural appearance, we used the stop-motion method.

We designed the following procedure to perform stop-motion capture: an object with a fuzzy edge sits on the platform in front of an LCD monitor. The object rotates in small, discrete steps along a predefined 3D trajectory, controlled by two servomotors connected to a computer. After each step, the digital camera in front of the setup captures the motionless object against a set of background images. At the end of the process, we remove the object, and the camera again captures all of the background images.

Following [3], we can solve for transparency values in a system of alpha-compositing equations (see Equation 1). Instead, however, we add reflectance term to Equation 1 to allow for lighting variations caused by background-image changes thus ensuring correct transparency value for object's parts that reflect light emitted by the monitor.

Finally, we composed the extracted objects with ground-truth transparency map over a set of challenging backgrounds and prepared several trimap-width gradations.

Having obtained our test sequences, we conducted extensive subjective comparison of 12 matting methods. Using the collected data, we then compared different quality metrics. The results showed that alpha temporal coherence is significantly more important to human perception of video-matting quality than accuracy and that despite the imperfection of optical-flow techniques, use of these techniques can improve estimation of matte temporal consistency.

To simplify access to the benchmark and enable addition of new methods, we created the http://videomatting.com website, which contains scatterplots and rating tables for different quality metrics. At the time of publication benchmark includes comparison of 12 methods by accuracy and temporal coherence. Besides the comparison, the website includes instructions for new participants.

[1] Xue Bai, Jue Wang, and David Simons. Towards temporally-coherent video matting. In *International Conference on Computer Vision (ICCV)*, pages 63–74, 2011. doi: 10.1007/978-3-642-24136-9_6.

[2] Inchang Choi, Minhaeng Lee, and Yu-Wing Tai. Video matting using multi-frame nonlocal matting laplacian. In *European Conference on Computer Vision (ECCV)*, pages 540–553, 2012. doi: 10.1007/978-3-642-33783-3_39.

[3] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *Computer Vision Pattern Recognition (CVPR)*, pages 1826–1833, 2009. URL www.alphamatting.com.

[4] Mikhail Sindeev, Anton Konushin, and Carsten Rother. Alpha-flow for video matting. In *Asian Conference on Computer Vision (ACCV)*, pages 438–452, 2013. doi: 10.1007/978-3-642-37431-9_34.

SDICP: Semi-Dense Tracking based on Iterative Closest Points

Laurent Kneip[1,2]
laurent.kneip@anu.edu.au
Zhou Yi[1,2]
yi.zhou@anu.edu.au
Hongdong Li[1,2,3]
hongdong.li@anu.edu.au

[1)]Research School of Engineering, Australian National University
[2)]ARC Centre of Excellence for Robotic Vision
[3)]NICTA Canberra Labs

The paper addresses the problem of *camera tracking*, which denotes the continuous image-based computation of a camera's position and orientation with respect to a reference frame. The method aims at regular cameras, which means that 3D-3D registration methods applicable to RGB-D cameras are not an option. The tracked frame contains only 2D information, thus requiring a solution to the *absolute pose* or *2D-3D registration* problem.

While traditional solutions to camera tracking [3] rely on sparse feature correspondences, the community has recently seen a number of direct photometric registration methods such as Newcombe et al. [8] and Engel et al. [1]. [1] is conceptually similar to [8], however gains computational efficiency by reducing the computation from dense to semi-dense regions that correspond to a thresholded edge-map of the image.

Photometric methods have the more general advantage of compensating for appearance variations caused by perspective view-point changes, whereas classical sparse methods often rely on static feature descriptors only (providing at most rotation and scale invariant properties [5, 6]). However, photometric registration techniques inherently suffer from the disability to overcome large disparities, where *large* sometimes means even just a couple of pixels [7]. Many photometric registration techniques therefore depend on pyramidal subsampling schemes in order to alleviate this problem.

The goal of the present paper is a novel 2D-3D registration paradigm for semi-dense depth maps that relies on the Iterative Closest Point (ICP) technique, and thus a reintroduction of geometric error minimization as a valid alternative for real-time monocular camera tracking in the case of semi-dense features. An example semi-dense depth map is indicated in Figure 1. In comparison to photometric registration techniques, our ICP technique has the conceptual advantage of requiring neither isotropic enlargement of the employed semi-dense regions, nor pyramidal subsampling. The work is in line with Feldmar et al. [2], Tomono [9], and Klein and Murray [4], which already attempt curve or edge registration in 2D using ICP.

Based on a hypothesized relative pose, the basic idea consists of *warping* a reference curve into the tracked image based on a prior 3D model or depth (in our case semi-dense) inside a reference frame. From a mathematical point of view, our idea may be formulated as follows. Let

$$\mathcal{S}^{\mathcal{F}_k} = \{\mathbf{s}_i^{\mathcal{F}_k}\} = \{(d_i^{\mathcal{F}_k})^{-1}\pi(\mathbf{p}_i^{\mathcal{F}_k})\}$$

denote the semi-dense depth map, where $\mathcal{P} = \{\mathbf{p}_i\}$ is the set of pixel locations that defines the semi-dense region in the reference frame $\mathcal{F}_k$, d_i the inverse depth of a pixel, and $\pi(\mathbf{p}_i) = \mathbf{f}_i$ is a known function that transforms points in the image plane into unit direction vectors located on the unit sphere around the camera center. The warped semi-dense region is easily obtained by

$$\mathcal{O}^{\mathcal{F}_{k+1}} = \{\mathbf{o}_i^{\mathcal{F}_{k+1}}\} = \{\pi^{-1}\left(\mathbf{R}^T\left(\mathbf{s}_i^{\mathcal{F}_k} - \mathbf{t}\right)\right)\},$$

where $\mathbf{t}$ and $\mathbf{R}$ denote the seeked position and orientation of the current frame. The final objective results in

$$\{\hat{\mathbf{t}}, \hat{\mathbf{R}}\} = \underset{\mathbf{t},\mathbf{R}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\{ \| n(\mathbf{o}_i^{\mathcal{F}_{k+1}}) - \mathbf{o}_i^{\mathcal{F}_{k+1}} \| \right\},$$

where $n(\mathbf{o}_i^{\mathcal{F}_{k+1}})$ is a function that returns the pixel from $\mathcal{P}^{\mathcal{F}_{k+1}}$ that is closest to $\mathbf{o}_i^{\mathcal{F}_{k+1}}$ under the Euclidean distance metric. We propose iterative minimization of the sum of all distances to the closest points (i.e. ICP) using gradient descent.

The core of our contribution extends ICP-based camera tracking to the case of noisy, outlier-affected semi-dense depth maps. We achieve large convergence basins and robustness with respect to outliers by prior evaluation of sparse hypotheses and a robust evaluation of distances in the image plane, respectively. The computation is speeded up by employing a distance transformation and the bin sort technique.

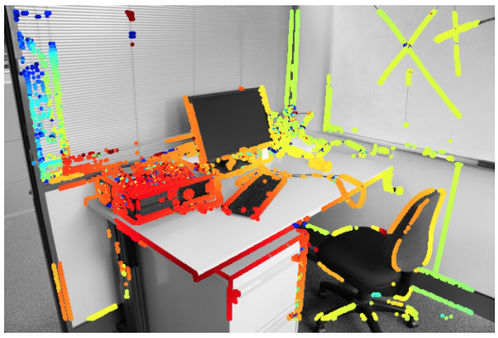


Figure 1: Example of a semi-dense depth map (hot colors=close, cold colors=far). The goal of this work is to track a monocular camera with respect to such a map. As can be observed, the depth map is typically affected by outliers, especially in new regions of the image (left part) and on self-similar background.

Our results demonstrate successful application to real data, showcasing real-time tracking over large disparities, very large view point changes, and severly corrupted depth maps on a publically available dataset. In addition to an improvement over classical sparse methods, we also show how our approach leads to a larger convergence basin in comparison to a photometric error minimization strategy.

[1] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.

[2] J. Feldmar, N. Ayache, and F. Betting. 3D-2D projective registration of free-form curves and surfaces. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Cambridge, MA, 1995.

[3] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[4] G. Klein and D. Murray. Full-3D edge tracking with a particle filter. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1119–1128, Edinburgh, UK, 2006.

[5] Stephan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: Binary Robust Invariant Scalable Keypoints. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[6] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[7] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Vancouver, Canada, 1981.

[8] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[9] M. Tomono. Robust 3D SLAM based on an edge-point ICP algorithm. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 2009.

Human Pose as Context for Object Detection

Abhilash Srikantha[1,2]
abhilash.srikantha@tue.mpg.de
Juergen Gall[1]
gall@iai.uni-bonn.de

[1] University of Bonn, Germany
[2] Max Planck Institute for Intelligent Systems, Tuebingen, Germany

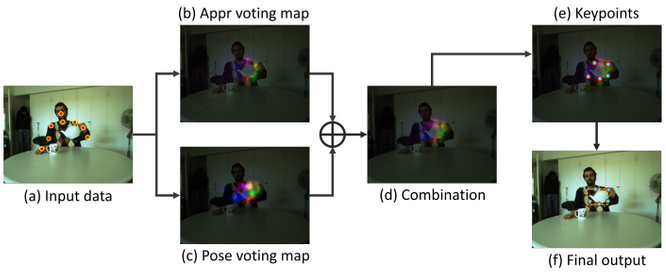


Figure 1: Detecting teapots: (a) Input is an image and automatically extracted human pose. (b) Object keypoint unaries based on appearance features and (c) using human pose features. (d) Linear combination of unaries. (e) Inferring keypoints (f) Regressing bounding box.

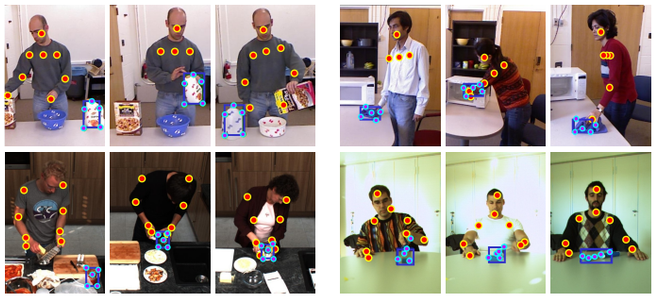


Figure 2: Qualitative results showing input human pose and most confident inferred bounding boxes as per Eqn (2). Successful detections are shown for classes *Milkbox, Cloth* from CAD-120; *Tin* from MPII-Cooking and *Roller* from ETHZ-Activity.

Object detection has seen considerable success, but the case of medium and small sized everyday objects still remains an open problem. Although such objects appear at low image resolutions, they often occur in the context of human interactions. Human context has been exploited in [1] which extends a deformable part model (DPM) to model spatial relations between body parts and parts of objects. This approach, however, only works well for images showing the instant of human-object interaction, i.e., when a human is closely in contact with an object. For images without an interaction, pose and objects are independently modelled, e.g., by having several models including either object or pose, or both together thereby leaving the human context unutilised.

In this work, we propose an approach that includes human pose as an additional context for object detection. Our approach is not limited to images showing explicit human-object interactions, but also works for general images where pose can be inferred. To this end, we model objects by a part based model and predict locations of parts from both image and pose data using regression forests. An outline of the approach is presented in Figure 1.

As illustrated in Figure 1(f), we represent an object by a set of descriptive keypoints $\mathcal{K} = \{\mathbf{k}_i\}$ where $\mathbf{k}_i$ encodes the image location of the ith keypoint. Following pictorial structures model, an optimal keypoint configuration given an observation $\mathcal{D}$ is given by

$$p(\mathcal{K}|\mathcal{D}) \propto \prod_i \phi_i(\mathbf{k}_i) \cdot \prod_{i,j \in E} \psi_{ij}(\mathbf{k}_i, \mathbf{k}_j) \tag{1}$$

While we retain binary potentials to model relative keypoint offsets in the tree structured graph E as in [1], our work focuses on extracting more discriminative unary potentials $\phi_i(\mathbf{k}_i)$ derived from observations in appearance $\mathcal{D}_A$ and human pose $\mathcal{D}_P$ and is given by

$$\phi_i(\mathbf{k}_i) = p(\mathbf{k}_i|\mathcal{D}_A, \mathcal{D}_P) \tag{2}$$

$$\sim \left(K(\sigma_A) * \phi_i^A(\mathbf{k}_i)\right) + \alpha \left(K(\sigma_P) * \phi_i^P(\mathbf{k}_i)\right) \tag{3}$$

where $*$ represents the convolution operation and σ is the standard deviation for the Gaussian blur kernel K. Since the human pose can only provide a rough prior for the location of an object class but is insufficient for accurate object localization, $\sigma_P > \sigma_A$. The probabilities are estimated by random forests either trained on image patches [3] or on joint features [6]. The unary potential for any modality is defined by

Table 1: Average AUC measures for various datasets.

Dataset	Appr.	Pose	Gall [3]	Desai [1]	Concat.	Comb.
MPII-Cooking	0.38	0.22	0.37	0.19	0.25	**0.41**
ETHZ-Action	0.46	0.24	0.42	0.50	0.23	**0.51**
CAD-120	0.31	0.09	0.29	0.21	0.20	**0.32**

$$\phi_i(\mathbf{k}_i) = \sum_{m=1}^{M} \frac{1}{|\mathcal{T}_i|} \sum_{T \in \mathcal{T}_i} p_m(\mathbf{k}_i - \mathbf{y}|c, L_T) \cdot p(c|L_T), \tag{4}$$

where $\mathcal{T}_i$ is the forest for the keypoint $\mathbf{k}_i$, anchor $\mathbf{y}$ is either the patch location for appearance or joint location for pose observations, L_T is the leaf resulting from the observation and $p(c|L_T)$ is the class probability stored at the leaf.

We evaluate the proposed approach on three datasets: ETHZ-Activity [2], CAD-120 [4] and MPII-Cooking [5]. Human pose is automatically inferred in all three datasets. We use the PASCAL-VOC measure for object detection. We compare our approach in various settings in Table 1. It can be seen that the appearance (Appr.) only features significantly outperform the pose (Pose) only features.

Comparing with state of the art methods, method [3] which uses appearance feature for object detection by voting for the center of the object performs slightly worse than the appearance only setup. The method [1] combines human pose estimation and object detection. The approach performs better than pose only features in ETHZ-Action and CAD-120 datasets, but significantly worse in the MPII-Cooking dataset.

As for combining appearance and pose features, we compare to an approach where a single forest is trained on a concatenation of both features (Concat). The accuracy of this approach, however, drops sharply in contrast to appearance only features. Finally, combining both modalities (Comb.) as per Eqn (2) yields the best results in all three datasets with gains ranging from 1% to 5%.

[1] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*. 2012.

[2] Juergen Gall, Andrea Fossati, and Luc Van Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011.

[3] Juergen Gall, Angela Yao, Nima Razavi, Luc Van Gool, and Victor Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 2011.

[4] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.

[5] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[6] Angela Yao, Juergen Gall, and Luc Van Gool. Coupled action recognition and pose estimation from multiple views. *IJCV*, 2012.

Learning Depth Calibration of Time-of-Flight Cameras

David Ferstl, Christian Reinbacher, Gernot Riegler,
Matthias Rüther, Horst Bischof
{ferstl,reinbacher,riegler,ruether,bischof}@icg.tugraz.at

Graz University of Technology
Institute for Computer Graphics and Vision
Inffeldgasse 16, 8010 Graz, AUSTRIA

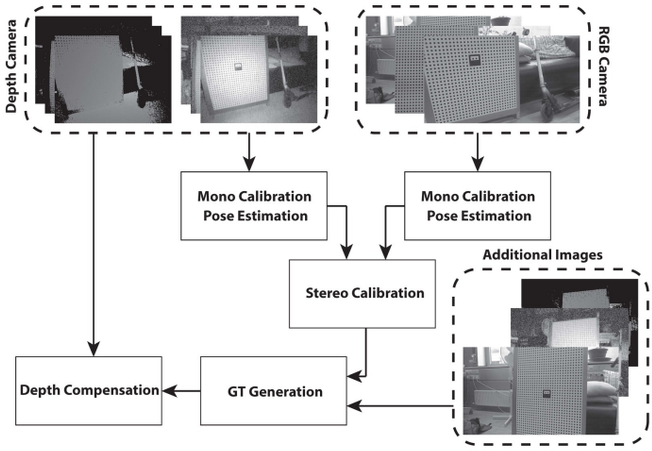
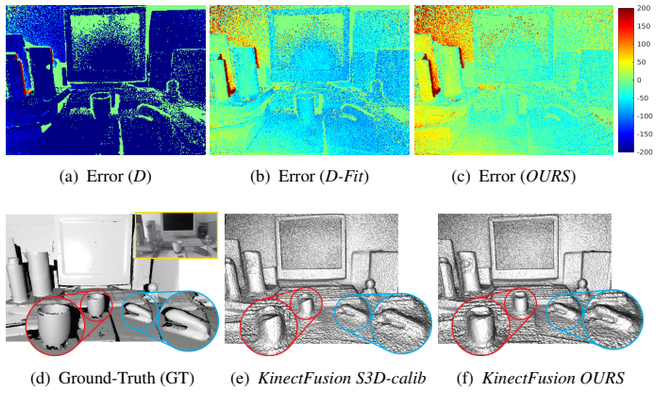


Figure 1: System overview. Using a planar target with known dimensions the system automatically detects feature points in both the RGB and the depth image. These points are used for a full stereo calibration and serve as ground truth 3D points for depth calibration.



(a) Error (*D*) (b) Error (*D-Fit*) (c) Error (*OURS*)

(d) Ground-Truth (GT) (e) *KinectFusion S3D-calib* (f) *KinectFusion OURS*

Figure 2: Evaluation of the influence of depth camera calibration to dense 3D reconstruction. The first row shows the color-coded error to GT(d) for one depth acquisition. It compares the uncalibrated camera output *D* (a), the camera output corrected with depth dependent polynomial fitting *D-Fit* (b) to our method *OURS* (c). Regions where no depth data is available are marked with zero. The second row shows the rendered *KinectFusion* output of an image series of the *Desk Scene* with magnified regions of interest.

Time-of-Flight (ToF) sensors are widely used in many applications such as autonomous navigation, 3D reconstruction and human-computer interaction. Thanks to an adoption by the gaming industry, ToF cameras have become reliable and affordable and are present in many living rooms as an accessory to gaming consoles. These modern devices combine a depth sensor with a traditional RGB sensor to form a so-called RGB-D camera.

Unlike passive 3D sensors based on stereo triangulation, ToF cameras emit light and therefore deliver acquisitions mostly independent from lighting conditions resulting in dense, real-time depth-maps of nearly arbitrary surfaces. The quality of the depth-maps depends on the scene geometry and surface properties in the scene. The two main error sources are 1) intrinsic calibration errors and 2) a spatially varying, context sensitive error in measured depth. This paper tackles both error sources of a RGB-D camera system using a novel calibration framework. The main goal of this work is to provide a highly accurate calibration and error compensation with a minimum of user interaction and without relying on a complex, specifically machined 2.5D calibration target.

According to the two error sources, our method is divided into two main parts: 1) The geometric camera calibration, to accurately estimate the intrinsic and extrinsic parameters of the RGB-D system. Features are automatically detected on a calibration target with sub-pixel accuracy which are used in both single-camera and stereo calibration. 2) The compensation of the measured depth errors learned from ground-truth depth data acquired during geometric camera calibration. Both parts of our method require a minimum of user interaction besides providing the calibration images. An overview of our RGB-D calibration framework is shown in Fig. 1.

We treat the intensity image of the depth camera as a regular intensity camera which allows us to parameterize the depth camera using a standard pinhole model with non-linear lens distortion. After estimating the extrinsic and intrinsic parameters of both depth and RGB camera we aim at minimizing the spatially varying depth bias which is inherently present in ToF-based depth measurements [1]. It has been shown that this error bias depends on scene geometry, reflection properties of the scene and position of the pixel in image space [2, 3]. Instead of directly modeling each error distribution with its own heuristics. Unfortunately, the different error sources highly correlate, which makes it hard to calibrate each error on its own. In our work we solve the problem of depth-bias compensation directly by using machine learning. We use a Random Regression Forest (RRF) to directly infer the mapping from depth and intensity features to a depth offset, which eliminates all error sources.

	Desk Scene				
	D	S3D-calib	I-Fit	D-Fit	OURS
RMSE(E)[mm]	168.48	158.23	47.45	46.09	**39.83**
$\bar{E}$[mm]	-141.19	-139.33	-23.40	-18.27	**2.65**
σ(E)[mm]	91.92	74.99	41.67	42.32	**39.74**

Table 1: Accuracy evaluation. The depth error is measured as RMSE(E), mean error $\bar{E}$ and standard deviation $\sigma(E)$ to GT in mm. The accuracy is evaluated for the recorded *Desk Scene*. The error of our method is compared to the uncalibrated camera output (*D*), the camera output calibrated with the values provided by the manufacturer (*S3D-calib*), to error fitting to ToF intensity (*I-Fit*) and to error fitting to depth (*D-Fit*) according to [1, 2].

The quality of our method is shown for a *Desk Scene* which is selected to incorporate a high texture and structure variation. Since the utilized depth camera has a depth range of ≈ 300 to 1000 mm the scene size is chosen appropriately. The GT measurement is generated using a Structured Light (SL) scanner which consists of two 4MP intensity cameras and one projector. The depth uncertainty of the SL system at the given baseline is ≤ 1.2 mm. We compare the different methods on single ToF acquisitions in Tab. 1 and show an improvement over separate depth or intensity fitting methods currently used in practice ([1, 2]). Further, we show that our method outperforms existing correction techniques and improves the quality of dense 3D reconstruction in Fig. 2. A more detailed evaluation can be found in the paper.

The MATLAB code for the automatic feature detection and calibration is available on our website[1] since we believe that it can not only benefit RGB-D camera calibration but also a variety of other camera calibration tasks.

[1] S. Fuchs and G. Hirzinger. Extrinsic and depth calibration of tof-cameras. In *CVPR*, 2008.

[2] J. Jung, J.-Y. Lee, Y. Jeong, and I.S. Kweon. Time-of-flight sensor calibration for a color and depth camera pair. *TPAMI*, PP(99):1–1, 2014.

[3] Y. M. Kim, D. Chan, C. Theobalt, and S. Thrun. Design and calibration of a multi-view tof sensor fusion system. In *CVPR Workshops*, 2008.

[1] http://rvlab.icg.tugraz.at/calibration

Low-Rank Spatio-Temporal Video Segmentation

Alasdair Newson
alasdairnewson@gmail.com
Mariano Tepper
http://www.marianotepper.com.ar
Guillermo Sapiro
http://www.ee.duke.edu/faculty/guillermo-sapiro

Duke University
Durham, NC, USA

Recently, a great deal of interest has been generated by the technique known as *Robust Principle Component Analysis* (RPCA) of Candès et al. [1], which addresses the problem of separating a matrix into a low-rank and a sparse component. This very general formulation can be used for tasks such as background estimation in videos and face recognition. In the case of background estimation, the low-rank matrix models the background, and the sparse matrix corresponds to the foreground. A considerable drawback of this approach is its poor robustness to local lighting conditions. If lighting conditions vary locally, one of two things may happen. Either the method incorporates the lighting variation into the foreground, which is clearly undesirable, or the rank of the background model is allowed to increase. Unfortunately, this second option means that the *true* foreground is likely to become included in the background, especially for objects which are static for a short while. Here, we propose to model the background as a *piece-wise low-rank* matrix. In this manner, it will be possible to extract several localised models which correspond to coherent lighting conditions. However, for this we need to *segment* the input video into such coherent regions.

We refer to this problem as a *low-rank spatio-temporal video segmentation*. We present an algorithm to address this segmentation problem, based on region merging and spectral clustering techniques. We show that by carrying out a local RPCA in each region, the results of foreground/background separation are greatly improved, in comparison with both the standard RPCA and several other well-known background estimation techniques.

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ represent an input video, in matrix form. Each frame contains m pixels, and there are a total of n frames in our video. The goal of RPCA is to decompose $\mathbf{X}$ as $\mathbf{X} \approx \mathbf{L} + \mathbf{S}$, where $\mathbf{L}$ is the low-rank matrix and $\mathbf{S}$ is the sparse matrix. Unfortunately, the rank of a matrix is a non-convex function, so a surrogate function, the *nuclear norm* is used. Thus, the background/foreground separation problem may be formulated as follows:

$$\min_{\mathbf{L},\mathbf{S} \in \mathbb{R}^{m \times n}} \tfrac{1}{2}\|\mathbf{X} - \mathbf{L} - \mathbf{S}\|_F^2 + \lambda_* \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1, \tag{1}$$

where $\|\mathbf{L}\|_* = \sum_i \sigma_i(\mathbf{L})$ is the nuclear norm of $\mathbf{L}$ and $\sigma_i(\mathbf{L})$ is the ith singular value of $\mathbf{L}$. The scalars λ_* and λ are optimisation parameters, $\|\cdot\|_F$ is the Frobenius matrix norm and $\|\cdot\|_1$ is the ℓ^1 matrix norm, which induces sparsity in the foreground matrix.

To segment $\mathbf{X}$ into different regions where the low-rank requirement is respected, we start by creating a regular 3D grid, which we denote with Ω, on the video domain. Each Ω_i corresponds to a rectangular cuboid of video information. We then create an undirected, weighted graph where each node represents a region Ω_i, and a node is connected with a 6-connectivity to the regions around it. Our goal will be to cluster this graph using spectral clustering techniques. The main challenge here is to design a cost function which shows how "coherent" two regions are in terms of their low-rank background representation.

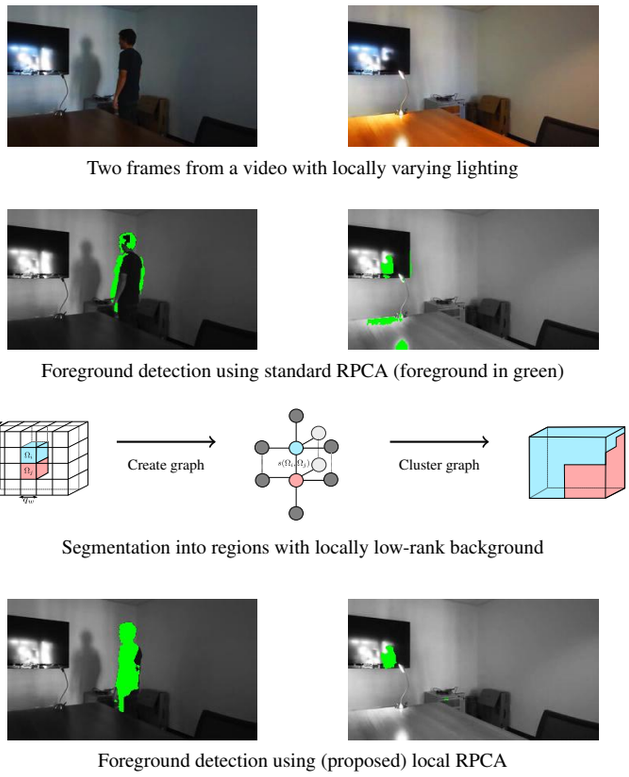
More formally, consider two regions to merge, Ω_i and Ω_j. We wish to see whether it is better to decompose the regions separately or jointly. The decomposition of Ω_i will be $\Omega_i \approx \mathbf{L}_i + \mathbf{S}_j$, and similarly for Ω_j. Our first observation is that it is easier to compare the coherence of the decompositions resulting from a *rank-constrained* version of Equation (1):

$$\{\mathbf{L}_i, \mathbf{S}_i\} = \underset{\mathbf{L},\mathbf{S}}{\operatorname{argmin}} \tfrac{1}{2}\|\mathbf{X}_i - \mathbf{L} - \mathbf{S}\|_F^2 + \lambda \|\mathbf{S}\|_1 \tag{2}$$
$$\text{subject to} \quad \operatorname{rank}(\mathbf{L}) \leq r.$$

The comparisons are made clearer because the λ_* parameter is removed and replaced with one which is more easily interpretable, the maximum rank of each local model, r. Once the decompositions of Ω_i, Ω_j and $\Omega_{i\cup j}$ are obtained in this manner, we can calculate the cost of merging the two regions. Let $e_i = \|\mathbf{X}_i - \mathbf{L}_i - \mathbf{S}_i\|_F^2$ be the quadratic error of the decomposition of Ω_i, and similarly for Ω_j and $\Omega_{i\cup j}$. Our cost function is:

$$d(\Omega_i, \Omega_j) = \frac{|e_i + e_j - e_{i\cup j}|}{\phi_{i\cup j}}. \tag{3}$$

where $\phi_{i\cup j}$ is a scalar. Once we have established the cost of merging two regions, we convert it into a *similarity* cost, and cluster the resulting graph using robust spectral clustering techniques [5].



Two frames from a video with locally varying lighting

Foreground detection using standard RPCA (foreground in green)

Segmentation into regions with locally low-rank background

Foreground detection using (proposed) local RPCA

Figure 1: Illustration and results of the algorithm

Figure 1 illustrates the problems caused by locally varying lighting conditions: either the foreground is merged into the background (second row, left), or the global (standard) RPCA is not able to represent local lighting changes (second row, right). This is corrected by segmenting the video, and carrying out a local RPCA in each region. We compare our algorithm qualitatively and quantitatively with respect to several algorithms of the literature [2, 3, 4] and find greatly improved performance in challenging situations.

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the ACM*, 2011.

[2] L. Maddalena and A. Petrosino. A Fuzzy Spatial Coherence-based Approach to Background/Foreground Separation for Moving Object Detection. *Neural Computing & Applications*, 2010.

[3] C. Stauffer and W. E. L. Grimson. Adaptive Background Mixture Models for Real-Time Tracking. In *CVPR*, 1999.

[4] J. Yao and J-M. Odobez. Multi-Layer Background Subtraction Based on Color and Texture. In *CVPR*, 2007.

[5] L. Zelnik-Manor and P. Perona. Self-tuning Spectral Clustering. In *Advances in Neural Information Processing Systems*, 2004.

Fast Online Upper Body Pose Estimation from Video

Ming-Ching Chang[1,2]
changm@ge.com
Honggang Qi[1,3]
hgqi@ucas.ac.cn
Xin Wang[1]
xwang26@albany.edu
Hong Cheng[4]
hcheng@uestc.edu.cn
Siwei Lyu[1]
slyu@albany.edu

[1] Computer Science Department
University at Albany, State University of New York
Albany, USA
[2] Computer Vision Lab
GE Global Research Center
Niskayuna, USA
[3] University of Chinese Academy of Sciences
Beijing, China
[4] Center for Robotics
University of Electronic Science and Technology of China
Chengdu, China

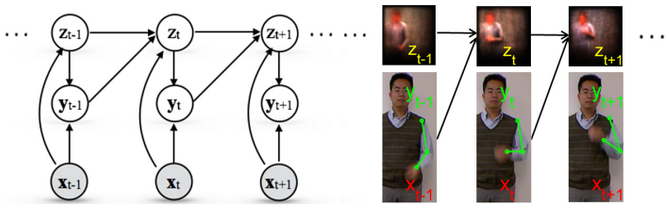


Figure 1: (**Left**) *The CDBN model structure and* (**Right**) *an example of the CDBN-MODEC model applied to online video pose estimation. Variable* $\mathbf{x}_t$ *corresponds to observations at time* t, i.e., *image features in individual video frames;* $\mathbf{y}_t$ *corresponds to poses,* i.e., *joint locations, and* z_t *is the latent pose modes. See text for detailed explanations.*

Estimation of human body poses, represented as the ensemble of joint locations, is an important problem in computer vision. In this work, we describe a fast *online* method for video pose estimation. We aim to extend existing single frame methods for *online* use, where latent pose modes (or "poselets") can be directly leveraged to improve motion consistency, in a paradigm similar to detect-and-track for object tracking.

Our method is based on a general *conditional dynamic Bayesian network* (CDBN) model, which is a combination of two widely used probabilistic graphical models, namely the dynamic Bayesian network (DBN) [3] and conditional random field (CRF) [2]. The DBN aspect of our model captures the temporal correlations between variables in a sequence, and the CRF aspect incorporates the complex relations between the observations and latent variables.

Specifically, we consider the following dynamic model that involves three time series variables: (1) an input sequence of observation variables $\mathbf{x}_{0:t}$, (2) an output temporal sequence of latent state variables $\mathbf{y}_{0:t}$, and (3) the sequence of latent mode selection variables $z_{0:t}$ with $z_t \in \{1,\cdots,M\}$ indicating one of the M input/output relation modes is active at time t. The CDBN model represents the dependencies of these variables with a dynamic probabilistic model, which corresponds to a factorization of probability distribution $p(z_{0:t},\mathbf{y}_{0:t}|\mathbf{x}_{0:t})$ according to the graphical structure in Fig.1(a), as:

$$p(z_{0:t},\mathbf{y}_{0:t}|\mathbf{x}_{0:t}) = p(z_0|\mathbf{x}_0)\prod_{\tau=0}^{t} p(\mathbf{y}_\tau|z_\tau,\mathbf{x}_\tau)\times\prod_{\tau=0}^{t-1} p(z_{\tau+1}|z_\tau,\mathbf{y}_\tau,\mathbf{x}_{\tau+1}). \quad (1)$$

The joint model in Eq.(1) can be used for dynamic Bayesian inference, implemented with particle filtering.

A key characteristic of the CDBN model is that it is an "open architecture", as it can incorporate different underlying CRF models (including future ones) into the DBN structure. When applying to *online* video pose estimation, this becomes an advantage, as it allows the resulting algorithm to incorporate effective single frame pose estimation method into the dynamic framework that models intra-frame correlations. In this work, we adopt part of the efficient CRF pipeline from the MODEC single pose estimation of Sapp and Taskar [4] as the CRF model in our CDBN framework. We term our method CDBN-MODEC for video pose estimation.

Fig.1 illustrates the structure of CDBN as a graphical model. To better evaluate *online* pose estimation from practical video streams, we also create a new high frame rate labeled dataset that calls for run time efficiency. When evaluated on this dataset and the VideoPose2 benchmark dataset, CDBN-MODEC achieves considerable improvements in both performance and efficiency over several state-of-the-art *online* video pose estimation methods. Qualitative results of using CDBN-MODEC estimating poses in several frames of videos from the HFR and VideoPose2 datasets are presented in Fig.2. In Fig.3 we compare the running time versus accuracy of MODEC [4], MODEC+S [5], CDBN-MODEC, and Cherian *et.al.* [1].



Figure 2: *Estimated upper body poses from frames of two videos in the HFR (top) and VideoPose2 (bottom) using CDBN-MODEC.*

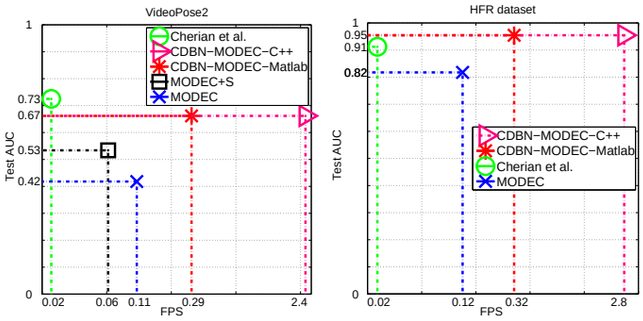


Figure 3: *Comparison of running time versus accuracy. See text for details.*

[1] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing Body-Part Sequences for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2361–2368, 2014.

[2] John Lafferty. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

[3] Kevin Patrick Murphy. Dynamic Bayesian Networks: Representation, Inference and Learning. In *PhD Thesis*, 2002.

[4] Benjamin Sapp and Ben Taskar. Multimodal Decomposable Models for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3681, 2013.

[5] D. Weiss, B. Sapp, and B. Taskar. Dynamic Structured Model Selection. In *International Conference on Computer Vision (ICCV)*, pages 2656–2663, 2013.

Simultaneous Inpainting and Super-resolution Using Self-learning

Milind G. Padalkar
milind_padalkar@daiict.ac.in
Manjunath V. Joshi
mv_joshi@daiict.ac.in
Nilay Khatri
nilay_space@yahoo.co.in

Dhirubhai Ambani Institute of Information & Communication Technology (DA-IICT),
Gandhinagar, Gujarat,
India – 382007

In applications like creating immersive walkthrough systems or digital reconstruction of invaluable artwork, both inpainting and super-resolution of the given images are the preliminary steps in order to provide better visual experience. The usual practice is to solve these problems independently in a pipelined manner. In this paper we propose a unified framework to perform simultaneous inpainting and super-resolution (SR). The main focus of this paper is inpainting, i.e. to remove objects in photographs and replace them with visually plausible backgrounds. The super-resolved version is obtained as a by-product in the process of using an additional constraint that helps in finding a better source for inpainting.

The proposed approach starts with a given test image I_0 having a region Ω_0 to be inpainted. We obtain the coarser resolution image I_{-1} by blurring and downsampling I_0 as done in [2]. We then construct dictionaries of image-representative low and high resolution (LR-HR) patch pairs from the known regions in the test image I_0 and its coarser resolution I_{-1}. The inpainting of the missing pixels is performed using exemplars found by comparing patch details at a finer resolution. Here, self-learning [4] is used to obtain the finer resolution patches by making use of the constructed dictionaries. The obtained finer resolution patches represent the super-resolved patches in the missing regions. Advantage of our approach when compared to other exemplar based inpainting techniques are (1) additional constraint in the form of finer resolution matching results in better inpainting and (2) inpainting is obtained not only in the given spatial resolution but also at higher resolution leading to super-resolution inpainting. The proposed approach is summarized in table 1 and one result is shown in figure 2.

1:	Construct LR-HR pair dictionaries using the known regions in I_0 and I_{-1}.
2:	Select highest priority patch $y_p = y_p^k \cup y_p^u$ for inpainting using method in [1]. Here, $y_p^k \in I_0$ and $y_p^u \in \Omega_0$.
3:	Search for candidate sources (exemplars) $y_{q_1}, \ldots, y_{q_K}$ in I_0.
4:	Self-learn HR patches $Y_{q_1}, \ldots, Y_{q_K}$ and Y_p using the constructed dictionaries: (a) Obtain $Y_{q_1}, \ldots, Y_{q_K}$ corresponding to $y_{q_1}, \ldots, y_{q_K}$. (b) Estimate Y_p corresponding to y_p.
5:	Find best exemplar Y_q in HR by comparing Y_p with $Y_{q_1}, \ldots, Y_{q_K}$.
6:	Obtain final inpainted HR patch H_p using Y_p and Y_q.
7:	Obtain inpainted LR patch L_p from H_p using transformation estimated from the constructed dictionaries and update Ω_0.
8:	Repeat steps 2–7 till all patches in Ω_0 are inpainted.

Table 1: Summary of the proposed approach

Consider an LR patch of size $m \times m$ in the known region $I_0 - \Omega_0$. By searching for a similar $m \times m$ sized patch in the coarser resolution I_{-1} we can obtain the corresponding $2m \times 2m$ sized HR patch as illustrated in figure 1. Although not all LR patches can find a good match in the coarser resolution, we use this methodology to create dictionaries D_{LR} and D_{HR} of image-representative LR-HR patch pairs.

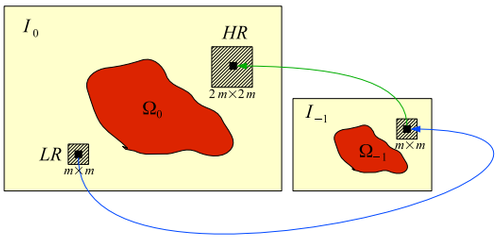


Figure 1: Finding LR-HR patch pairs using the given image I_0 and its coarser resolution I_{-1}.

For every $m \times m$ sized patch on the boundary of Ω_0, a data term and a confidence term denoting the presence of structure and proportion of known pixels, respectively, are calculated using the method in [1]. The patch y_p around a pixel p for which the product of the data and confidence terms is the highest is then selected as the highest priority patch for inpainting. The patch y_p is compared with every $m \times m$ sized patch in the known region $I_0 - \Omega_0$ to obtain K best matches denoted as $y_{q_1}, \ldots, y_{q_K}$ representing the candidate exemplars. The key idea is to compare the HR details of y_p with those of $y_{q_1}, \ldots, y_{q_K}$ to identify the best source for inpainting. Here, the HR patch Y corresponding to an LR patch y is self-learnt using the LR-HR patch pair dictionaries as follows:

$$Y = D_{HR} * \alpha, \tag{1}$$

where, α is the sparse representation obtained by optimizing

$$\min ||\alpha||_{l_1}, \quad \text{subject to} \quad y = D_{LR} * \alpha. \tag{2}$$

For an LR patch y_p with missing pixels y_p^u and known pixels y_p^k, the corresponding sparse representation is obtained by using $D_{LR} = D_{LR_p}^k$ and $y_p = y_p^k$ in equation (2). Here $D_{LR_p}^k$ consists of only those rows in D_{LR} that correspond to the known pixels y_p^k in y_p. The complete HR dictionary D_{HR} and the obtained sparse representation are then used in equation (1) to get the HR details Y_p. Similarly HR $Y_{q_1}, \ldots, Y_{q_K}$ for candidates are obtained. We then compare each of the HR patches $Y_{q_1}, \ldots, Y_{q_K}$ with Y_p and choose the one having minimum sum of squared distance as Y_q. The pixels Y_p^u in the HR patch Y_p (that correspond to pixels y_p^u of the LR patch y_p) are now replaced by corresponding pixels in Y_q to obtain the final inpainted HR patch H_p. The corresponding LR patch L_p which is the inpainted version of y_p is obtained by using the HR to LR transformation, which in our case is blurring and downsampling or can be estimated from the constructed dictionaries. This process is repeated to inpaint the entire missing region Ω_0 such that in every iteration only the missing pixels y_p^u in the selected patch y_p are inpainted and Ω_0 is updated accordingly.

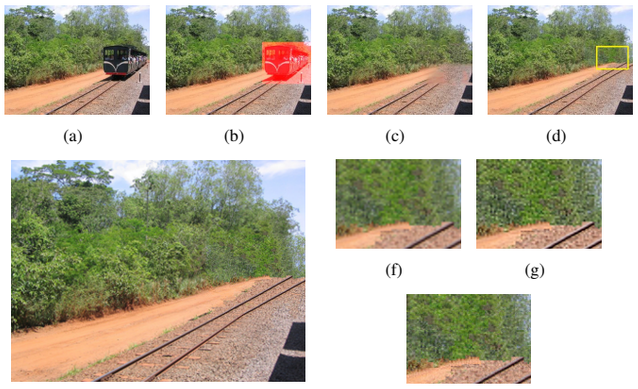


Figure 2: Simultaneous inpainting and super-resolution: (a) input; (b) region to be inpainted; (c) inpainting using planar structure guidance [3]; (d) inpainting using proposed method showing yellow box inside the inpainted region; (e) simultaneously inpainted and super-resolved image (by a factor of 2) using the proposed method with known regions upsampled using bicubic interpolation; (f)–(h) expanded versions after upsampling (the region marked by the yellow box in (d)) using various approaches viz. (f) bicubic interpolation, (g) Glasner *et al.*'s method [2] and (h) proposed method for super-resolution.

[1] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based inpainting. *Trans. Image Proc.*, 13(9): 1200–1212, Jan. 2004.

[2] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV 2009*, pages 349–356.

[3] J. Huang, S. Kang, N. Ahuja, and J. Kopf. Image completion using planar structure guidance. *ACM Trans. Graph.*, 33(4):129:1–129:10, July 2014.

[4] N. Khatri and M. V. Joshi. Image super-resolution: Use of self-learning and Gabor prior. In *ACCV 2012*, pages 413–424.

Manitest: Are classifiers really invariant?

Alhussein Fawzi
alhussein.fawzi@epfl.ch
Pascal Frossard
pascal.frossard@epfl.ch

Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

Invariance to geometric transformations is a highly desirable property of classifiers in image recognition tasks. Nevertheless, it is unclear to which extent state-of-the-art classifiers are invariant to transformations such as rotations and translations. This is mainly due to the lack of general methods that properly measure such an invariance. We propose *Manitest*, a rigorous and systematic approach for quantifying the invariance to geometric transformations of an arbitrary classifier. The source code of Manitest is available at the project website `http://sites.google.com/site/invmanitest/`.

Definition of *Manitest* measure. For a fixed image I, we measure the robustness of a classifier f relatively to the transformation group $\mathcal{T}$ as the minimal normalized *distance* between the identity transformation e and a transformation $\tau \in \mathcal{T}$ that changes the classification label when applied to the image.

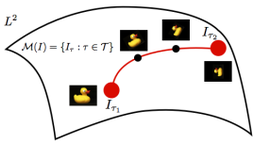
$$\Delta_{\mathcal{T}}(I;f) = \min_{\tau \in \mathcal{T}} \frac{d(e,\tau)}{\|I\|_{L^2}} \text{ subject to } f(I_\tau) \neq f(I),$$

where
$\mathcal{T}$: transformation group,
I_τ : I transformed by τ,
d : distance on $\mathcal{T}$,
e : identity transformation.

For a distribution of samples μ, the global invariance measure of f is obtained by averaging $\Delta_{\mathcal{T}}(I;f)$

$$\rho_{\mathcal{T}}(f) = \mathbb{E}_{I\sim\mu}\Delta_{\mathcal{T}}(I;f).$$

Which distance to use on $\mathcal{T}$? A crucial element in the definition of our Manitest invariance measure $\rho_{\mathcal{T}}$ is the choice of the distance measure d. In order to define d, our novel key idea is to represent the set of transformed versions of an image as a manifold; the transformation metric is naturally captured by the geodesic distance on the manifold. For a given image, the invariance measure $\Delta_{\mathcal{T}}(I;f)$ therefore corresponds to the minimal normalized *geodesic* distance on the manifold that leads to a point where the classifier's decision is changed.



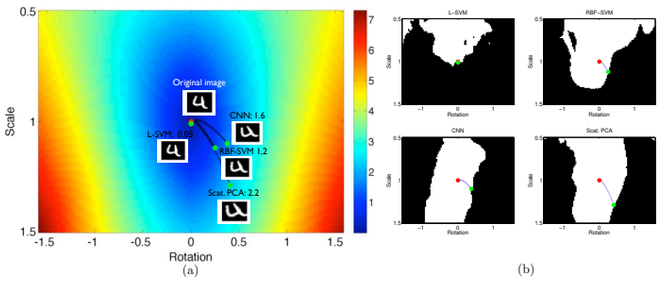
Algorithm for computing $\Delta_{\mathcal{T}}(I;f)$. The manifold $\mathcal{T}$ is sampled using a regular grid, and the geodesic distances are estimated using the *Fast Marching* algorithm [2, 3]. The algorithm is terminated whenever a node that changes the classifier's decision is visited.

Experimental results. A comparison of different classifiers in terms of their Manitest invariance scores on the MNIST digit classification dataset is shown in Table 1. The classifier based on scattering transforms outperforms other classifiers in terms of robustness to translations and global similarity transformations. We visualize the invariance scores of the different tested classifiers on an example test image shown in Fig. 1.

Group	Lin. SVM	RBF-SVM	CNN	Scat. [1]
Test error (%)	8.4	1.4	**0.7**	0.8
Translations	0.8	1.3	1.7	**2.1**
Scale + Rotation	0.8	1.5	**1.9**	1.8
Similarity	0.6	1.1	1.5	**1.6**

Table 1: Accuracy and invariance scores of different classifers on the MNIST dataset.

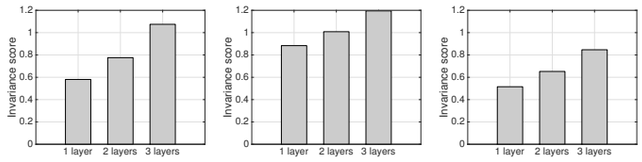
Using Manitest, we also quantify the effect of *data augmentation* and *CNN depth* on the invariance of a classifier. Our result shows that the invariance score of an RBF-SVM increases by 50% on the similarity group



(a) (b)

Figure 1: Visualization of invariance to the Scale+Rotation group for an example image of digit "4". **(a):** Distance map, where the color code indicates the geodesic distance from the identity transformation (shown by red dot at the center). For each classifier, the minimal transformation for which the output of the classifier is not correct (i.e., not "4") is indicated, along with the corresponding transformed image and geodesic path. **(b):** The region where the classifier correctly outputs the label "4" is shown in white. Geodesic paths are also shown.

by merely adding transformed samples in the training set. Quite surprisingly, the invariance score of the RBF-SVM trained with augmented samples surpasses that of a CNN (without augmentation), which shows the merits of data augmentation in terms of increasing invariance. Moreover, we show in Fig. 2 the increasing invariance of Manitest scores with the number of layers of a CNN on the CIFAR-10 dataset.



(a) Translations (b) Dilation + Rotation (c) Similarity

Figure 2: Invariance scores of CNNs on $\mathcal{T}_{\text{trans}}$, $\mathcal{T}_{\text{dil+rot}}$ and $\mathcal{T}_{\text{sim}}$, for the CIFAR-10 dataset.

Fig. 3 shows a ranking of the images in terms of their invariance scores $\Delta_{\mathcal{T}}(I;f)$. Despite the high accuracy of the three layer CNN on the CIFAR-10 task, note that a slight transformation of the original images can change the classification label of the classifier (see worst 10 images). More emphasis should therefore be put in order to achieve higher levels of invariance.



(a) Worst 10 (b) Top 10

Figure 3: Illustration of images having (a) worst, (b) top invariance to similarity transformations, for the three-layer CNN. The odd rows show the original images, and the even rows the minimally transformed images changing the prediction of the CNN. The Manitest invariance score is indicated on each transformed image.

[1] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

[2] Ron Kimmel and James A Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences*, 95(15):8431–8435, 1998.

[3] John N Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE Transactions on Automatic Control*, 40(9):1528–1538, 1995.

MCSLAM : a Multiple Constrained SLAM

Datta Ramadasan[1]
datta.ramadasan@gmail.com
Marc Chevaldonné[2]
marc.chevaldonne@udamail.fr
Thierry Chateau[1]
thierry.chateau@univ-bpclermont.fr

[1] Pascal Institute
Blaise Pascal University
Clermont-Ferrand, FR
[2] ISIT
Auvergne University
Clermont-Ferrand, FR

In this paper, we propose a new algorithm, named MCSLAM (Multiple Constrained SLAM), designed to dynamically adapt each optimization to the variable number of parameters families (sensor pose, roughly known objects poses and dimensions, delay between sensors, ...) and heterogeneous constraints (reprojection error, distance from points or edges to object surface, acceleration...). The proposed algorithm is based on three contributions: 1) a new Levenberg-Marquardt C++ library named LMA and freely available [1] 2) an architecture allowing a high level of flexibility and performances 3) a real-time usage of a temporal spline curve as the parametric trajectory model that provides an efficient way to add heterogeneous constraints within the optimization.

The main idea of LMA is to provide a simple interface with a non-intrusive mechanism of adaptation to a problem while maintaining good performances. LMA works as a meta-program using C++ template to analyse at compile-time (CT) the problem to optimize from a list of C++ functors. The parameters are deduced from the functors arguments and the degree of freedom (*dof*) of each parameter is defined by the user. Then LMA generates a data structure to store the functors and the parameters in *tuple* according to the number of parameters families and constraints. The resolution of the normal equations is written efficiently using a sparse representation constructed with a set of small matrices of static sizes. The LMA library solves the normal equations using dense Cholesky or a sparse PCG specially designed to manage little matrices of static size. It also implements the classical optimization tricks to be effective on little, medium, and big size problems. LMA also implements common features as automatic differentiation and robust cost functions.

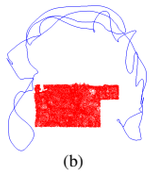
The continuous-time representation of the trajectory is used to deal with constraints on the motion. This allows to mix data from many unsynchronised sensors and evolution model. To apply constraints on the 3D structure of the environment, 3D models of coarsely knowns shapes are used. To represent the motion, we use the uniform cumulative b-spline described by Lovegrove *et al.* [2] but we separate position and orientation in two different splines and we use the Rodriguez formula to compute the exponential and the logarithm of $\mathbb{SO}3$ group. Moreover, we adapt the key-frame based SLAM to deal with the spline: key-frames are constrained to be on the spline. We also use every inter-key-frame poses computed by the localization process to apply a weak constraint on the spline. A temporal sliding window of 3 seconds is used to select, from the SLAM and the IMU, the more recent data used to constrain the spline.

The MCSLAM algorithm is based on a graph composed by constraints, parameters and dependencies. First, the problem configuration is analysed at compile-time to generate a specified LMA solver, whose cost function to minimize is the sum of the constraints C dynamically added: $\mathcal{E} = \sum_{i=0}^{C} \sum_{k=0}^{K_i} \left\| \frac{\rho_{i,k}}{\sigma_i} \right\|^2$ with K_i the number of observations corresponding to the constraint i, $\rho_{i,k}$ the error associated to the observation k of the constraint i and σ_i the estimation of the measurement error. The σ_i value is specific to each constraint. Then the graph is skimmed according to the dependencies to feed the solver with the parameters and the constraints. Each constraint has a list of functors and each functor corresponds to one error term of the cost function to minimize. Moreover, one function has to be written for each constraint: this function has two input, the dependencies graph and the solver. Thus, each constraint has a full access to the graph and any required data in order to fill the solver. To achieve real-time performances, we extend the incremental reconstruction scheme described by [3] to manage the constraints. This produces an anchorage of the optimized parameters according to the deprecated parameters. Consequently, the constraints are applied only on parameters corresponding at the end of the reconstruction (spatially and temporally).

[1] git.univ-bpclermont.fr/datta.ramadasan/lma

Circle (2 *dof.*)	g2o	Ceres	LMA
Dense	2.26	2.49	**0.51**
BAL$_1$ (66k *dof.*)	g2o	Ceres	LMA
Sparse	x	2.63	**1.06**
Dense Schur	3.60	2.36	**1.01**
Sparse Schur	3.49	2.42	**1.10**
Implicit Schur	x	2.97	**1.12**
BAL$_2$ (4M *dof.*)	g2o	Ceres	LMA
Implicit Schur	x	298	**131**

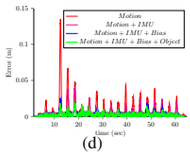
(a)



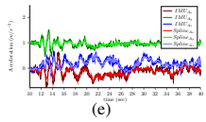
(b)



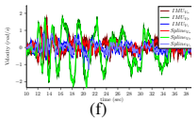
(c)



(d)



(e)



(f)

Figure 1: (a) LMA compared to Ceres and g2o (results are in seconds). (b) Ground truth of the second experiment. (c) Image from the sequence of the second experiment. (d) Error in position of the SLAM using different configuration of constraints. (e,f) IMU's accelerometer and gyrometer compared to the spline acceleration and orientation velocity.

Three comparatives summarized in table 1(a) are performed on problems of different sizes using the solvers g2o, Ceres-1.10 and LMA. The first comparative is a basic problem designed to optimize the equation of a circle. In this experiment, LMA is 4.8 times faster than Ceres and 4.3 times faster than g2o. The second experiment is performed on an instance of the dataset used in the Ceres benchmark [1]. Each solver is limited to 10 iterations of Levenberg-Marquardt, derivatives are numeric and centred, and the PCG is limited to 20 iterations when it's used. On the first instance, LMA is 2 to 3 times faster than Ceres and g2o, with a similar accuracy, using the 4 different linear solvers. The third experiment is performed on the 35 instances of the dataset [1] with the Implicit Schur method and shows that LMA is, on average, 2.5 times faster than Ceres, and more accurate on 21 datasets out of 35.

Four different configurations of the MCSLAM are tested : 1) SLAM with a constant velocity model constraint, 2) SLAM with a constant velocity model and IMU constraints, 3) SLAM with a constant velocity model and IMU constraints and IMU's bias optimization, 4) SLAM with spline, constant velocity model, IMU, IMU's bias optimization, and 3D object constraints. The ground truth is an indoor trajectory computed from a structure from motion algorithm. The ground truth is visible on the figure 1(b) in blue, and the edge model in red and an image from the camera is shown on the figure 1(c). Moreover, to highlight the robustness of the system, we artificially stop the camera poses constraints on the the spline during 1 second every 2 seconds. The errors in position are shown in the figure 1(d). Figures 1(e) and 1(f) shows the acceleration and the orientation velocity of the trajectory and the IMU data using all constraints.

[1] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *Computer Vision–ECCV 2010*, pages 29–42. Springer, 2010.

[2] Steven Lovegrove, Alonso Patron-Perez, and Gabe Sibley. Spline fusion: A continuous-time representation for visual-inertial fusion with application to rolling shutter cameras. In *Proceedings of the British machine vision conference*, pages 93–1, 2013.

[3] Etienne Mouragnon, Maxime Lhuillier, Michel Dhome, Fabien Dekeyser, and Patrick Sayd. Real time localization and 3d reconstruction. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 363–370. IEEE, 2006.

Incremental Dictionary Learning for Unsupervised Domain Adaptation

Boyu Lu[1]
bylu@umiacs.umd.edu
Rama Chellappa[1]
rama@umiacs.umd.edu
Nasser M. Nasrabadi[2]
nasser.m.nasrabadi.civ@mail.mil

[1] Department of Electrical and Computer Engineering
University of Maryland
College Park, MD, USA

[2] U.S. Army Research Lab.
Adelphi, MD, USA

Domain adaptation (DA) methods attempt to solve the domain mismatch problem between source and target data. In this paper, we propose an incremental dictionary learning method where some target data called supportive samples are selected to assist adaptation. The idea is partially inspired by the bootstrapping-based methods [1, 3], which choose from the target domain some samples and add them into source domain for retraining the classifier. However, the suitable sample selection and stopping criteria for DA setting is a tricky problem. For the sample selection criteria, we choose supportive samples that are close to the source domain, so that they act as a bridge to connect the two domains and reduce the domain mismatch. More sepecifically, given the source dictionary D, we select the target samples that minimize the reconstruction error when represented by D. Then we augment the source domain by adding supportive samples and retrain the dictionary. For the stopping criteria, we guarantee that the domain mismatch decreases monotonically during adaptation. This is realized by checking whether adding new supportive samples will reduce the domain dissimilarity after each iteration. The proposed approach is shown in Fig. 1.

Supportive Samples Selection: We select the supportive samples using $W^{(k+1)}$ by solving the following optimization problem:

$$\begin{aligned} W_j^{(k+1)} &= \underset{W_j}{\operatorname{argmax}} \quad tr(W_j P_j^{(k+1)}) \\ s.t. \quad & W_j^{(k+1)} \cdot \sum_{l=1}^{k} W_j^{(l)} = 0, \quad ||W_j^{(k+1)}||_0 = Q, \quad j = 1,...,C \end{aligned} \tag{1}$$

where $W_j \in R^{N_t \times N_t}$ are diagonal matrices with each element in the jth column of W on the diagonal, *e.g.*, $W_j = diag\{w_{1j}, w_{2j}...\}$ and similarly $P_j = diag\{p_{1j}, p_{2j}...\}$. $p_{ij} \in [0,1]$ represents the probability that target sample x_i^t belongs to the class j and $w_{ij} \in \{0,1\}$ indicates whether the target sample x_i^t is selected as supportive samples for class j. Q is the number of supportive samples for each class. This objective function (1) maximizes the confidence of the selected supportive samples. The first constraint requires that we keep adding new supportive samples to the source domain. The second constraint ensures that the number of supportive samples for each class is balanced.

Augmented Source Domain Update: After selecting the supportive samples, we augment source data by adding weighted supportive samples to existing source data:

$$X_j^{(k+1)} = [X_j^{(k)} | X^t W_j^{(k+1)} P_j^{(k+1)}] \quad j = 1,...,C \tag{2}$$

Since the labels of the supportive samples may be erreneous, each selected supportive sample is weighted by its confidence. The weights indicate the reliability of the labels of the supportive samples and highly confident supportive samples will contribute more to the model.

Dictionary Update: Dictionary is updated by solving the following dictionary learning problem:

$$D_j^{(k+1)} = \underset{D_j, Z_j}{\operatorname{argmin}} ||X_j^{(k+1)} - D_j \cdot Z_j||_F^2 + \lambda ||Z_j||_1 \quad j = 1,...,C. \tag{3}$$

We solve (3) using the online dictionary learning method [2]. The dictionary obtained in the previous iteration is used as the initial dictionary in the next iteration. In this way, the computational cost is relatively low.

Stopping criterion: One trivial stopping criterion is to stop when there is no new supportive samples for one of the classes. But our goal is to guarantee that the adaptation monotonically reduce the domain divergence. So we design a domain similarity measure and perform adaptation only when the domain similarity increases after each iteration.

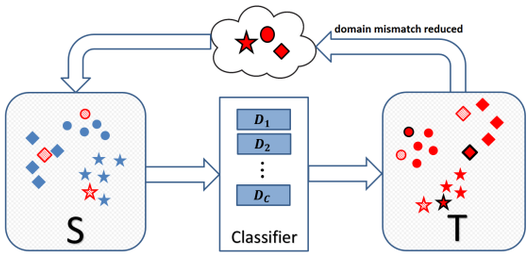


Figure 1: Scheme of the proposed method. The original source data is colored in *blue* and the target data is colored in *red*. Different shapes represent different classes. The red samples with shadow indicate the previously selected supportive samples that have been added to the source domain. The red samples with black border represent the supportive samples selected in the current iteration.

In order to quantify domain similarity, we introduce a simple domain similarity measure for X^s and X^t:

$$\rho(X^s, X^t) = \sqrt{\frac{1}{N_s N_t} \sum_i \sum_j (x_i^{sT} x_j^t)^2} = \sqrt{\frac{tr((X^s)^T X^t (X^t)^T X^s)}{N_s N_t}}. \tag{4}$$

Since the classification accuracy on supportive samples is good, the main reason for the performance to drop in the target domain is that the source classifier behaves poorly on the non-supportive samples. It indicates that domain mismatch mainly lies between the source samples and the non-supportive samples. If the distance between supportive samples and non-supportive samples is smaller than the distance between the source domain and the non-supportive samples, selecting supportive samples can help reduce the domain mismatch.

Theorem 1. *We divide the target samples into two part, supportive samples X_f and non-supportive samples X_n with N_f and N_n samples, respectively. With the definition of ρ above, and if $\rho(X_f, X_n) > \rho(X^s, X_n)$, then the domain similarity (or mismatch) will increase(or decrease) when we add some supportive samples to the source domain:*

$$\rho(X_{new}^s, X^t) > \rho(X_{old}^s, X^t) \tag{5}$$

where $X_{old}^s = X^s$ and $X_{new}^s = [X^s | X_f]$.

We evaluate the proposed method for object classification and face recognition and compare with several state-of-the-art unsupervised DA methods. Experimental results show that our method outperforms other approaches significantly in most cases.

[1] Lorenzo Bruzzone and Mattia Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):770–787, 2010.

[2] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[3] Tatiana Tommasi and Barbara Caputo. Frustratingly easy nbnn domain adaptation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 897–904. IEEE, 2013.

StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation

Dan Levi[1]
http://sites.google.com/site/danmlevi
Noa Garnett[1]
noagarnett@gmail.com
Ethan Fetaya[2]
ethan.fetaya@weizmann.ac.il

[1] Advanced Technical Center Israel
General Motors R&D
Herzlyia, Israel

[2] Weizmann Institute of Science
Rehovot , Israel

Obstacle detection is a fundamental technological enabler for autonomous driving and vehicle active safety applications. While dense laser scanners are best suitable for the task (e.g. Google's self driving car), camera-based systems, which are significantly less expensive, continue to improve. Stereo-based commercial solutions such as Daimler's "intelligent drive" are good at general obstacle detection while monocular-based systems such as Mobileye's are usually designed to detect specific categories of objects (cars, pedestrians, etc.). The problem of general obstacle detection remains a difficult task for monocular camera based systems. Such systems have clear advantages over stereo-based ones in terms of cost and packaging size.

Another related task commonly performed by camera-based systems is scene labeling, in which a label (e.g. road, car, sidewalk) is assigned to each pixel in the image. As a result full detection and segmentation of all the obstacles and of the road is obtained, but scene labeling is generally a difficult task. Instead, we propose in this paper to solve a more constrained task: detecting in each image column the image contact point (pixel) between the closest obstacle and the ground as depicted in Figure 1(Left). The idea is borrowed from the "Stixel-World" obstacle representation [1] in which the obstacle in each column is represented by a so called "Stixel", and our goal is to find the bottom pixel of each such "Stixel". Note that since we don't consider each non-road object (e.g. sidewalk, grass) as an obstacle, the task of road segmentation is different from obstacle detection. Notice also that free-space detection task is ambiguously used in the literature to describe the above mentioned obstacle detection task [1] and the road segmentation task [4].

Current "Stixel-based" methods [1] for general obstacle detection use stereo vision while our method is monocular-based. A different approach for monocular based obstacle detection relies the host vehicle motion and uses Structure-from-Motion (SfM) from sequences of frames in the video [3]. In contrast our method uses a single image as input and therefore operates also when the host vehicle is stationary. In addition, the SfM approach is orthogonal to ours and can therefore be later combined to improve performance. For the task of road segmentation, the common approach is to perform pixel or patch level [4]. In contrast, we propose to solve the problem using the same column-based regression approach as for obstacle detection. Our approach is novel in providing a unified framework for both the obstacle detection and road-segmentation tasks, and in using the first to facilitate the second in the training phase.

We propose solving the obstacle detection task using a two stage approach. In the first stage we divide the image into columns and solve the detection as a regression problem using a convolutional neural network, which we call "StixelNet". Figure 1(Right) shows an example network input and output. In the second stage we improve the results using interactions between neighboring columns by imposing smoothness constrains via a Conditional Random Field (CRF) over consecutive columns. The flowchart of the obstacle detection algorithm is presented in Figure 2(top). To train the network we introduce a new loss function based on a semi-discrete representation of the obstacle position probability. In this approach we model the probability $P(y)$ of the obstacle position as a piecewise-linear probability distribution.

The road segmentation is done is three stages. The first two, StixelNet (trained on the road segmentation task) followed by a CRF, are the same as in obstacle detection. The final stage performs a graph-cut segmentation on the image to achieve higher accuracy by enforcing road boundaries to coincide with image contours. The flowchart of the road segmentation algorithm is presented in Figure 2(bottom).

It is well known that having large quantities of labeled data is crucial for training deep CNNs. A major advantage of our unique task formulation is the ability to use laser-scanners, which are excellent at the given task, for labeling, thus eliminating the need for manual annotation. In addition, we further leverage this by fine-tuning StixelNet to the road segmentation task using a smaller amount of hand labeled data. Our experiments use the KITTI dataset [2]. On obstacle detection our approach is the state-of-the-art camera-based method even when compared to the stereo-based "Stixel" approach [1]. On the KITTI road segmentation challenge, our fine-tuned network, although not suitable to model all cases, is ranked second among all methods.



Figure 1: (Left) Obstacle detection example. (Right) Input to StixelNet and output example

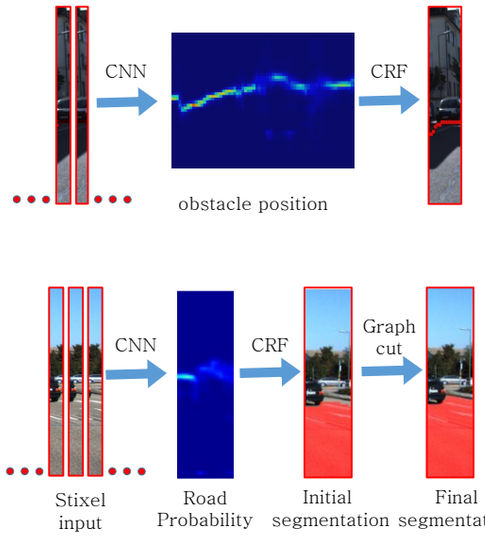


Figure 2: (Top) Obstacle detection algorithm flowchart. (Bottom) Road segmentation algorithm flowchart.

[1] Hernán Badino, Uwe Franke, and David Pfeiffer. The stixel world - a compact medium level representation of the 3d-world. In *Pattern Recognition*, 2009.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[3] Qifa Ke and T. Kanade. Transforming camera geometry to a virtual downward-looking camera: robust ego-motion estimation and ground-layer detection. In *CVPR*, 2003.

[4] G. B. Vitor, A. C. Victorino, and J. V. Ferreira. A probabilistic distribution approach for the classification of urban roads in complex environments. In *Workshop on Modelling, Estimation, Perception and Control of All Terrain Mobile Robots on IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

Segmenting natural images with the least effort as humans

Qiyang Zhao
http://gait.buaa.edu.cn/~zqy

State key lab of software development environment,
Beihang University

Given that natural image segmentation is well-known as an ill-posed problem, then how can we design an algorithm to obtain good performance as human subjects? A choice is to learn from human segmentations as MCG [3] which obtains the state-of-the-art performance and fairly high computational efficiency. Then a question arises here: what should we learn? The way all existing literatures exploit human segmentations ignores a basic fact that human segmentations are produced by human operations. The human segmentation process would inevitably fulfill the human behavior's principles including **the least effort principle (LEP)**: a human *will strive to solve his problem in such a way as to minimize the total work that he must expend* [1].

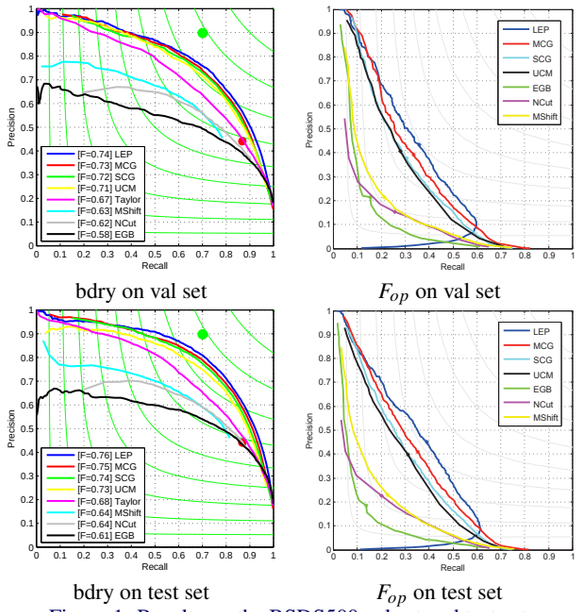
The principle gives us a new insight into our problem. Suppose you are a human subject in the BSDS segmentation experiment, and are required to segment images into pieces under the **ending instruction**: *each piece contains only one single distinguished thing*. Then your total effort F of segmenting an image should including understanding images (U) to guide following operations, and tracing boundaries (T) by hands with a mouse: $F(I,S) = U(I) + T(S)$. Then from the viewpoint of LEP, human subjects would like to find an acceptable segmentation S by minimizing their effort, or formally,

$$\min_S F(I,S), \text{s.t. each piece contains only one thing.} \quad (1)$$

We can describe a region with features/concepts with colors, textures and semantic concepts. If you can describe a region easily, then there tend to be only one 'thing'. But the description effort is not the only factor we should take into account when deciding whether to partition a region or not, instead we must balance the description effort and the tracing effort. In order to estimate the tracing effort T, we estimate the tracing loads in an empirical way. We collect several human subjects to trace the boundaries in the BSDS human segmentations, and tell them to do it as precisely as possible. The tracing style is to use many small clicks as recommended by the BSDS tools. We record the time consumptions as the measure of tracing loads, then analyze boundaries to find a way to predict them. Finally we find the tracing loads of long boundaries are linearly related to the corner amounts of different angle bins, while it is different for short boundaries as the loads seem to be only linearly related to lengths. We mix these two cases softly with a Logistic function to get a unified model. The term U is hard to tackle, because analyzing it with either the psychology or the neural science is far beyond the scope of the paper. Instead, we simply set it as a constant value.

The BSDS human segmentations are produced hierarchically with the BSDS tool. In order to mimic human operations, we sort it into an **Hierarchy Constraint** on our model, and prove that the optimal solutions on larger unit costs can be found by merging the region group of the minimum *muc* in the optimal solutions on smaller unit costs. However, the amount of all region groups is too huge to be handled. A feasible way is to consider adjacent region pairs only, so to optimize F approximately with a naive merging strategy. The key idea is to maintain a dynamically sorted search tree *BST*, but all adjacencies should be updated and re-inserted in *BST* after each mergence. By profiling we find about 80% updates are wasted as they are overwritten by new updates. Fortunately, we find almost all mergences are subject to a **Monotonicity Constraint**, with which we prove that the output of the naive merging strategy is the exact solution minimizing F, and the *muc*'s are monotonically increased. In light of the monotonicity, we replace the search tree *BST* in naive merging with a direct access table to improve the computational efficiency considerably.

There are several popular measures to evaluate the segmentation algorithms independently [2]. However in existing literatures, the best performances of a single method on these measures are usually reported on dramatically different thresholds. We integrate them to get a more objective measure which can reflect both boundary and region performance synchronously. We train our new segmentation method on the BSDS train set and evaluate our new segmentation method (LEP for abbr.) on the test set and val set. We calculate the color entropy in the Lab space and the texture entropy based on a variant of C-LBP, to measure the independent descriptions, and use normalized Euclidean Lab- distances and the odds being separated in the structured predictions of a random forest to measure the regularity description. We evaluate all involved methods with their public source codes or pre-computed results. The experiment on BSDS500 shows that our method obtains the state-of-the-art performance, as shown in figures and tables. Especially on the test set, our method is the only one with the precision 0.90+ on the recall 0.50 (see fig.1.a). It runs on all images in less than 1s in average, and is much more efficient than other method which has 0.70+ scores on the boundary measure.



bdry on val set F_{op} on val set

bdry on test set F_{op} on test set

Figure 1: Results on the BSDS500 val set and test set.

Table 1: Results on the BSDS500 val set. The best values are in bold.

	Boundary			F_{op}		Covering		PRI		VOI		Time
	ODS	OIS	AP	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	
Human	0.79	0.79	-	-	-	0.73	0.73	0.87	0.87	1.16	1.16	-
Ours	**0.74**	**0.76**	**0.79**	**0.39**	**0.46**	**0.62**	**0.68**	**0.82**	**0.86**	**1.51**	**1.31**	1s
MCG	0.73	0.76	0.74	0.36	0.42	0.61	0.67	0.81	0.86	1.55	1.37	20s+
SCG	0.72	0.73	0.63	0.34	0.40	0.59	0.65	0.80	0.85	1.62	1.44	3.5s+

Table 2: Results on the BSDS500 test set. The best values are in bold.

	Boundary			F_{op}		Covering		PRI		VOI		Time
	ODS	OIS	AP	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	
Human	0.80	0.80	-	0.56	0.56	0.72	0.72	0.88	0.88	1.17	1.17	-
Ours	**0.76**	**0.79**	**0.82**	**0.42**	**0.47**	**0.63**	**0.69**	**0.84**	**0.87**	**1.47**	**1.29**	1s
MCG	0.75	0.78	0.76	0.38	0.43	0.61	0.66	0.83	0.86	1.57	1.39	20s+
SCG	0.74	0.77	0.65	0.35	0.42	0.60	0.65	0.83	0.86	1.63	1.43	3.5s+

The success of our new algorithm shows, when learning human segmentations, it is feasible and necessary to take human behaviors into account to find a good solution. We also observe that the performance of our algorithm on the new integrating measure is much worse than that of human subjects. How to introduce semantic concepts into our algorithm is our future work. All source codes and evaluation results are available on http://gait.buaa.edu.cn/~zqy/lep.

[1] Zipf G. *Human behaviour and the principle of least effort*. Addison-Wesley Press, 1949.

[2] Pont-Tuset J. and Marques F. Measures and meta-measures for the supervised evaluation of image segmentation. In *CVPR*, 2013.

[3] Arbelaez P., Pont-Tuset J., and Barron J. et al. Multiscale combinatorial grouping. In *CVPR*, 2014.

Deep Fishing: Gradient Features from Deep Nets

Albert Gordo[1]
albert.gordo@xrce.xerox.com
Adrien Gaidon[1]
adrien.gaidon@xrce.xerox.com
Florent Perronnin[2]
perronnin@fb.com

[1] Computer Vision Group, Xerox Research Centre Europe
[2] Facebook AI Research*

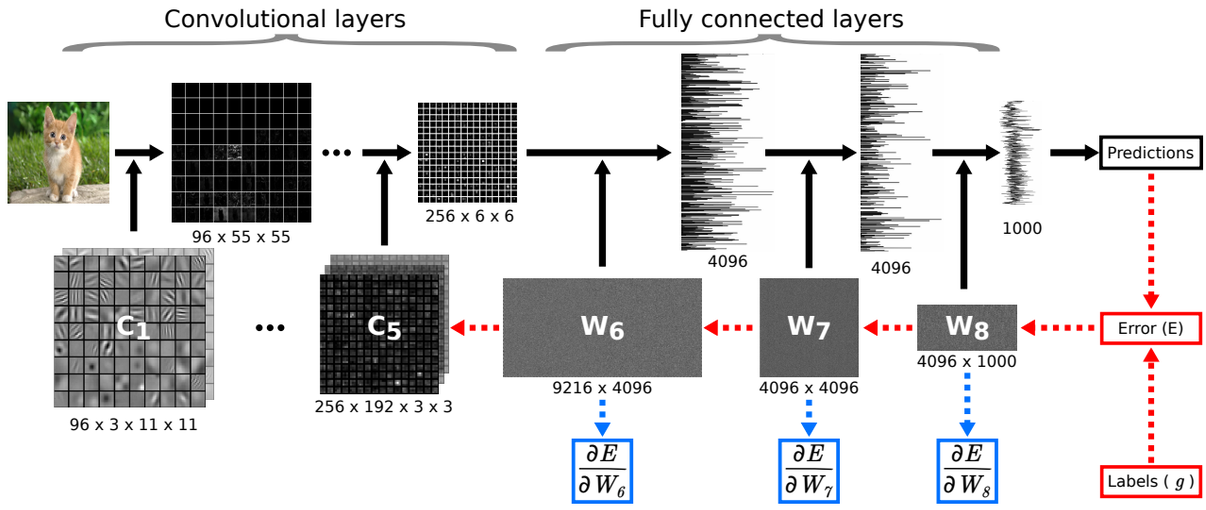


Figure 1: AlexNet architecture. C_k are the parameters (4D tensors) of the convolutional layers. W_k are the parameters (matrices) of the fully connected layers. Black (resp. red) arrows represent the information flow during the forward (resp. backward) pass. **Inspired by the Fisher Kernel, we study the use of gradient-related information (the blue matrices) as transferable representations.**

Convolutional Networks (ConvNets) have recently improved image recognition performance thanks to end-to-end learning of deep feed-forward models from raw pixels. Deep learning is a marked departure from the previous state of the art, the Fisher Vector (FV), which relied on gradient-based encoding of local hand-crafted features. In this paper, we discuss a novel connection between these two approaches. First, we show that one can derive gradient representations from ConvNets in a similar fashion to the FV. Second, we show that this gradient representation actually corresponds to a structured matrix that allows for efficient similarity computation. In particular, for the fully-connected layers, the gradient of the error w.r.t. the weights can be computed as an outer product,

$$\frac{\partial E}{\partial W_k} = x_{k-1}\left[\frac{\partial E}{\partial y_k}\right]^T, \tag{1}$$

where x_{k-1} is the input of layer k, y_k is the output of layer k *before* the non-linearity, and the gradient w.r.t. y_k is easily computed during the back-propagation stage. To compute the back-propagation at test time, where the true label vector g is unknown, we use a uniform vector with values $1/P$, where P is the number of classes.

We show that the trace kernel between gradient representations of two images A and B can be efficiently decomposed in two dot-product subkernels, one on forward features, and one on backward features, that are combined multiplicatively:

$$\begin{aligned} K_{tr}\left(\frac{\partial E}{\partial W_k}(A), \frac{\partial E}{\partial W_k}(B)\right) &= (x_{k-1}^A)^T x_{k-1}^B \cdot \left[\frac{\partial E}{\partial y_k}(A)\right]^T \frac{\partial E}{\partial y_k}(B) \\ &= K_{fw}(A,B) \cdot K_{bw}(A,B). \end{aligned} \tag{2}$$

We experimentally study the benefits of transferring this representation and kernel over the outputs of ConvNet layers, and find consistent improvements on the Pascal VOC 2007 and 2012 datasets with respect to using only the representations based on the forward activations on two popular network architectures (AlexNet and VGG16) pretrained on ImageNet ILSVRC2012 (*cf*. Tables 1 and 2 for results).

Table 1: Left: Results on Pascal VOC 2007 and VOC 2012 with AlexNet (A) and VGG16 (V). Results on VOC 2012 are on the validation set (mAP in %)

Features	VOC2007 (A)	VOC2007 (V)	VOC2012 (A)	VOC2012 (V)
x_5 (Pool5)	71.0	86.7	66.1	81.4
x_6 (FC6)	77.1	89.3	72.6	84.4
x_7 (FC7)	79.4	89.4	74.9	84.6
y_8 (FC8)	79.1	88.3	74.3	84.1
x_8 (Prob)	76.2	86.0	71.9	81.3
x_5; x_6	76.4	89.2	71.6	84.0
$\frac{\partial E}{\partial W_6} = x_5\left[\frac{\partial E}{\partial y_6}\right]^T$	80.2	89.3	75.1	84.6
x_6; x_7	79.1	89.5	74.3	84.6
$\frac{\partial E}{\partial W_7} = x_6\left[\frac{\partial E}{\partial y_7}\right]^T$	**80.9**	**90.0**	**76.3**	**85.2**
x_7; y_8	79.7	89.2	75.3	84.6
$\frac{\partial E}{\partial W_8} = x_7\left[\frac{\partial E}{\partial y_8}\right]^T$	79.7	88.2	75.0	83.4

Table 2: Comparison with other ConvNet results (mAP in %).

	VOC'07	VOC'12
Proposed - AlexNet	80.9	76.5
Proposed - VGG16	**90.0**	85.3
DeCAF	73.4	-
Razavian *et al.*	77.2	-
Oquab *et al.*	77.7	78.7
Zeiler *et al.*	-	79.0
Chatfield *et al.*	82.4	83.2
He *et al.*	80.1	-
Wei *et al.*	81.5	81.7
Simonyan *et al.*	89.7	**89.3**

* Work done while FP was at the Computer Vision Group of the Xerox Research Centre Europe.

Multimodal Stereo Vision For Reconstruction In The Presence Of Reflection

Scott Sorensen
sorensen@udel.edu
Philip Saponaro
saponaro@udel.edu
Stephen Rhein
smrhein@udel.edu
Chandra Kambhamettu
chandrak@udel.edu

Video/Image Modeling and Synthesis Lab,
University of Delaware
Newark DE, USA



Figure 1: The four camera system

Traditional stereo approaches assume a lambertian scene, an assumption which is violated in the presence of specular reflections. A variety of techniques have been developed to detect and reconstruct these surfaces [1, 4] using a variety of constraints, however in this work we attempt to reconstruct a reflecting surface and a reflected scene using different imaging modalities. Using a four camera system shown in Fig. 1, operating as two stereo pairs we reconstruct a reflecting surface as well as a reflected scene. The camera system consists of a pair of visible band cameras and a pair of Long Wave InfraRed (LWIR) cameras. We leverage these different imaging modalities by taking advantage of the fact that reflectivity is wavelength dependent and reflective materials in one modality can appear non reflective in complementing modalities.

We calibrate the system by extending the work of [3] to handle cross modality calibration. This gives us a common coordinate system. To reconstruct a reflected scene we first extract the reflecting surface. This is done by adding texture to the surface in only one modality, and reconstructing correspondences, and fitting a plane to these points using principle component analysis. The implicit representation of this plane

$$P \cdot n + d = 0 \tag{1}$$

can be used to intersect rays for reconstruction. Where P is the origin, n is the plane normal and d is a constant.

Reconstructing the scene is accomplished by stereo matching the reflected images using uncalibrated rectification and disparity matching using [2]. Each point in the disparity map can be viewed as a ray going from the camera center through the image plane defined by,

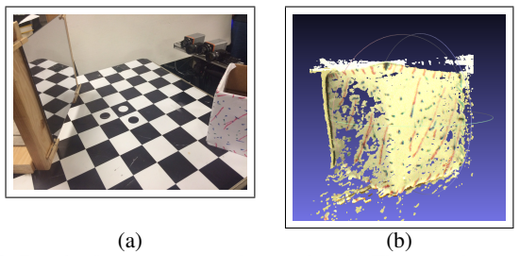
$$V_i = C_0 + t \cdot \frac{\beta}{norm(\beta)} \tag{2}$$

where C_0 is the camera center, and

$$\beta = R' \cdot A^{-1} \cdot [x_i, y_i, 1] \tag{3}$$

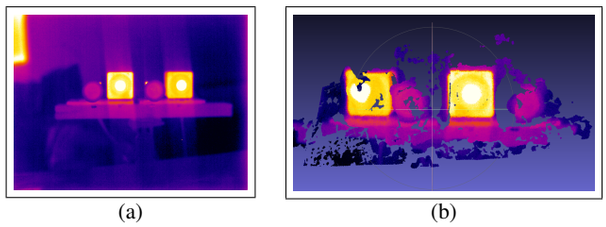
where A is the camera matrix, and R is the camera rotation matrix, and $[x_i, y_i]$ is the pixel location. We intersect these rays with the reflecting surface by substituting P from equation 1 with V_i from equation 2, and the reflected ray direction is defined as,

$$V_{reflected} = V_i + (2 \cdot n \cdot C_i) \tag{4}$$



(a) (b)

Figure 2: Results from reconstructing a cube: (a) The experimental setup; (b) The reconstructed model.



(a) (b)

Figure 3: Results from reconstructing a self portrait of our camera system: (a) The camera system reflected in LWIR; (b) The reconstructed model.

where

$$C_i = -(n \cdot V_i) \tag{5}$$

$V_{reflected}$ is the direction of the reflected ray. The reflected ray is defined as $I_i + t_i * V_{reflected}$. For each set of corresponding rays the closest point of intersection is calculated using least squares method resulting in a dense 3D point cloud.

We test our approach with a reflection setup that allows us to swap materials in place relative to the camera system. Using this setup we demonstrate texture can be added to the surfaces that is visible in just one imaging modality, and that the reflecting surface can be accurately modeled. Furthermore we show our reconstruction technique can be applied in both directions, by reconstructing a visible scene using the LWIR cameras to extract the reflecting surface, and reconstructing a LWIR scene by extracting the reflecting surface using the visible band cameras as shown in figures 2 and 3 respectively.

[1] Thomas Bonfort, Peter Sturm, and Pau Gargallo. General specular surface triangulation. In *Asian Conference on Computer Vision 2006*, 2006.

[2] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *CVPR (2)*, pages 807–814. IEEE Computer Society, 2005. ISBN 0-7695-2372-2. URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2005-2.html#Hirschmuller05`.

[3] Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Stephen Rhein, and Chandra Kambhamettu. Material classification with thermal imagery. In *CVPR*, 2015.

[4] Silvio Savarese and Pietro Perona. Local analysis for 3d reconstruction of specular surfaces. In *CVPR (2)*, pages 738–745. IEEE Computer Society, 2001. ISBN 0-7695-1272-0. URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2001-2.html#SavaresePP01`.

Real-time Human Detection based on Personness Estimation

Kyuwon Kim[1,2]
q1.kim@samsung.com

Kwanghoon Sohn[1]
khsohn@yonsei.ac.kr

[1] School of Electrical and Electronic Engineering
Yonsei University
Republic of Korea

[2] Mobile Division
Samsung Electronics
Republic of Korea

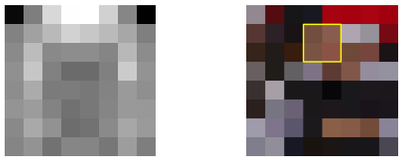
Even though deformable part model (DPM) [3] is one of the most popular human detection methods [1], it is difficult to see a practical application of DPM human detection on mobile devices due to its computational overhead. On the other hand, many *objectness* (a.k.a. *detection proposals*) measure methods have been recently proposed [4]. An objectness measure generates object windows that are likely to contain generic objects, and allows avoiding an exhaustive search. Its intention to improve detection speed seems to be perfectly matched with DPM human detection in mobile devices. However, most objectness measure methods reviewed and evaluated in [4] are not well suited for real-time detection because all objectness measure methods (except for the binarized normed gradients (BING) method [2]) spend more than 250 milliseconds.

For this reason, we propose a more efficient and accurate method for estimating human windows using the normed gradients with color feature in order to enhance detection rate within a short period of time. In this paper, we are interested in humans, not generic objects. So we name objectness estimation for humans simply as *personness* estimation.

To rapidly determine the priority order of each feature vector of an image with a supervised approach, we make use of the skin color model and the fast NG feature (or simply *Bing*) proposed by Cheng *et al*. [2]. An NG feature is a 64 dimensional vector describing the magnitude (a.k.a. *norm*) of the gradients of an 8×8 downsampled image. To examine every image window for the objectness measure, a magnitude map of gradients is resized to 36 predefined *quantized shapes* (or *quantized sizes* in [2]). The results are called *NG maps*. By calculating the correlation values between the NG maps and $\mathbf{w} \in \mathbb{R}^{8\times 8}$, we can obtain a filter score for each window. A quantized shape is ignored in the predicting stage if the quantized shape has less than or equal to 50 positive samples in the training stage. The final objectness score is calculated as follows:

$$o_i = v_i \cdot s_{(i,x,y)} + t_i \tag{1}$$

where $v_i, t_i \in \mathbb{R}$ and $s_{(i,x,y)}$ are learned coefficient, a bias term and the filter score at position (x,y) of each quantized shape i. $s_{(i,x,y)}$ is obtained using the first stage linear SVM, and v_i and t_i are obtained using the second stage linear SVM.



(a) (b)

Figure 1: NG features for personness and a 8×8 downsampled example: (a) Personness filter $\mathbf{w}_p$ for human detection. (b) The 8×8 downsampled image. The yellow box emphasizes the head part

Even though SVM training on various object categories makes the original linear model $\mathbf{w} \in \mathbb{R}^{8\times 8}$ work for generic objects, it degrades the performance of the SVM classifier for single-category object detection. Therefore, we generate the new linear model $\mathbf{w}_p \in \mathbb{R}^{8\times 8}$ shown in Fig. 1(a) by training the linear SVM only on humans that our DPM detector can detect. This restricted training technique also reduces the number of quantized shapes to consider. Fig. 1(a) illustrates that $\mathbf{w}_p \in \mathbb{R}^{8\times 8}$ surprisingly places more confidence in the shoulder and head regions. After training the linear SVM on VOC 2007 dataset, we choose four points $(3,1),(3,2),(4,1),(4,2)$ to extract skin color information. Head and neck are usually located in these four points as shown in the Fig. 1(b). When function $\text{skin}(x,y)$ returns a binary value indicating whether the point (x,y) is skin color according to a skin color model, the skin score at



(a) (b)

Figure 2: An example of human detection based on personness estimation: (a) The best matched windows generated by personness estimation: Total 723 candidate windows are generated, and 6th, 14th, and 17th windows are matched. (b) Human bounding boxes (BB) detected with the aid of candidate windows.

position (x,y) is

$$c_{(i,x,y)} = \frac{1}{4} \sum_{j=1}^{2} \sum_{k=3}^{4} \text{skin}(x+k, y+j) \tag{2}$$

And the new personness score is as follows:

$$p_i = v_i \cdot s_{(i,x,y)} + u_i \cdot c_{(i,x,y)} + t_i \tag{3}$$

where $u_i \in \mathbb{R}$ is learned coefficient for skin score using the second stage linear SVM.

To make good use of personness estimation, we divide DPM into three stages: setup, detection, and evaluation. In the setup stage, DPM builds a feature pyramid, and initializes three pyramids: filter response pyramid, probe flag pyramid and convolution flag pyramid. In the detection stage, filter responses are computed between root/part filters and some selected feature vectors. With width w, height h, and center position $(x+w/2, y+h/2)$ of a window, we can pick the corresponding $(l, \bar{x}, \bar{y})$ where l is the scale level of the filter response pyramid, and $(\bar{x}, \bar{y})$ is location of the filter response map to select feature vectors to be convolved. We call the exact response-map coordinates $(l, \bar{x}, \bar{y})$, which correspond to a given window, *pinpoints*. At this conversion from windows to pinpoints, one should keep in mind that a window for personness was quantized, so it might not be completely fitted to a real human window. Therefore, two important factors are considered for the window-pinpoints conversion: 1) *aspect-ratio threshold* 2) *patch size* for pinpoint. After DPM performs convolutions between the chosen feature vectors and a root filter, it calculates part filter responses of feature vectors corresponding to the root positions. When time runs out, the detector finishes computing filter responses, and moves to the evaluation stage. In the evaluation stage, the detector computes the overall root scores as discussed in [3].

In conclusion, we proposed personness estimation that generates promising human windows within a short period of time for real-time human detection. To show the efficiency and practicality of personness estimation, we designed a real-time DPM human detection that makes effective use of personness estimation.

[1] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV*, 2014.

[2] M. Cheng, Z. Zhang, W. Lin, and P. Torr. BING: Binarized normed gradient for objectness estimation at 300fps. In *CVPR*, pages 3286–3293, 2014.

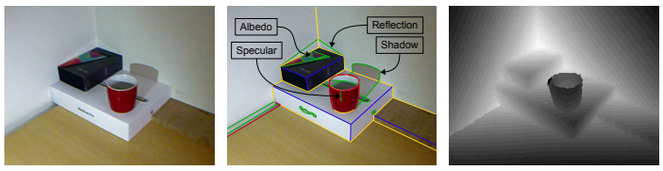
[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32:1627–1645, Sep. 2010.

[4] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *arXiv:1502.05082*, 2015.

Semantic Classification of Boundaries of an RGBD Image

Nishit Soni[1]
nishit.soni@research.iiit.ac.in
Anoop M. Namboodiri[1]
anoop@iiit.ac.in
C. V. Jawahar[1]
jawahar@iiit.ac.in
Srikumar Ramalingam[2]
ramalingam@merl.com

[1] International Institute of Information Technology, Hyderabad, India
[2] Mitsubishi Electric Research Lab (MERL), Cambridge, USA



(a) RGB Image (b) Edge Types (c) Depth Map

Figure 1: *The edges in (a) are marked in (b) with the color code: red (occluding), green (planar), blue (convex), and yellow (concave). Planar edges are caused by different phenomena as marked. (c) shows the Kinect depth map (note the depth quantization artifacts).*

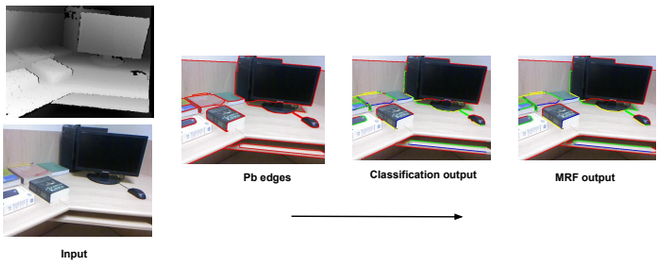


Figure 2: *This figure summarizes the pipeline of our approach. It shows RGB and depth maps as input (1st image set), with Pb edge detection [2] (2nd image). The classification and MRF outputs are shown in the last two images respectively. Color code: red (occ), green (pln), blue (cvx), yellow (ccv).*



(a) (b) (c)

Figure 3: *Ground truths (above) and the corresponding results from our approach (2nd row) and Gupta* et al. *[1] (3rd row). Color code: red (occ), green (pln), blue (cvx), yellow (ccv).*

Abstract: The problem of labeling the edges present in a single color image as convex, concave, and occluding entities is one of the fundamental problems in computer vision [4]. It has been shown that this information can contribute to segmentation, reconstruction and recognition problems. Recently, it has been shown that this classification is not straightforward even using RGBD data. This makes us wonder whether this apparent simple cue has more information than a depth map? In this paper, we propose a novel algorithm using random forest for classifying edges into convex, concave and occluding entities. We release a data set with more than 500 RGBD images with pixel-wise ground labels. Our method produces promising results and achieves an F-score of 0.84 on the data set.

Introduction: Edges in an image often correspond to depth discontinuities at object boundaries (occlusion edges) or normal discontinuities (convex or concave edges). In addition, there could be planar edges that are within planar regions. Figure 1 shows an image containing different types of edges. Note that planar edges may result from shadows, reflection, specularities and albedo variations. In classical line labeling with synthetic line drawings, we do not have any planar edges as the purpose of edge labeling has always been to classify the depth edges as occluding, convex and concave. However, in real images planar edges occur more frequently than others. This paper studies the problem of classifying boundaries from RGBD data. In many 3D models obtained using RGBD sensors or multi-view reconstruction techniques, we typically have very noisy 3D point cloud near the boundaries. This is because most stereo reconstruction algorithms and structured light techniques are known to provide noisy reconstruction near the boundaries. This makes the labeling problem challenging.

Algorithm: We use both image and depth cues to infer the labels of edge pixels. We start with a set of edge pixels obtained from an edge detection algorithm and the goal is to assign one of the four labels to each of these edge pixels. Each edge pixel is uniquely mapped to one of the contour segments. Contour segments are sets of linked edge pixels. We formulate the problem as an optimization on a graph constructed using contour segments. We obtain unary features using pixel classifier based on Random forest. We design a feature vector with simple geometric depth comparison features. We use a simple Potts model for pairwise potentials. The individual steps in the algorithm is shown in Figure 2.

Experiments: For quantitative evaluation of the method, we have created an annotated dataset of 500 RGBD images of varying complexity. Train to test ratio is 3:2. Our dataset consists of objects such as tables, chairs, cupboard shelves, boxes and household objects in addition to walls and floors. We also annotate 100 images from NYU [3] dataset, which include varying scenes from bed-room, living-room, kitchen, bathroom and so on with different complexities.

We compare our approach with Gupta *et al.* [1] and show that our approach provides better results. Figure 3 shows a qualitative comparison between our approaches. The approach that we present here provides good labels for most pixels with high precision. We get an average F-score of 0.82 on the classification results for our data set. The use of smoothness constraints in the MRF achieves an F-score of 0.84. The NYU dataset contains complex scenes containing glass windows and table heads. We achieve an average F-score of 0.74 for the NYU dataset.

[1] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013.

[2] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 2004.

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[4] K. Sugihara. *Machine Interpretation of Line Drawings*. MIT Press, 1986.

Elliptical ASIFT Agglomeration in Class Prototype for Logo Detection

Raluca Boia
rboia@imag.pub.ro;
Corneliu Florea
corneliu.florea@upb.ro;
Laura Florea
laura.florea@upb.ro;

Image Processing and Analysis Laboratory
University "Politehnica" of Bucharest
Bucharest, Romania

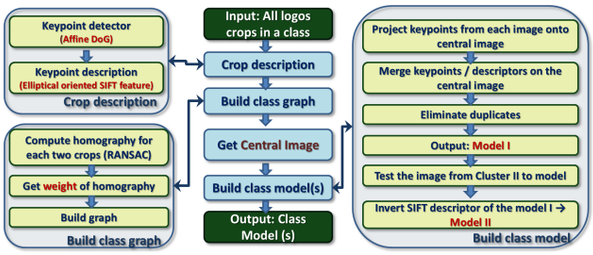


Figure 1: The schematic of the proposed system while describing the classes. With red are specific choices leading to greater performance.

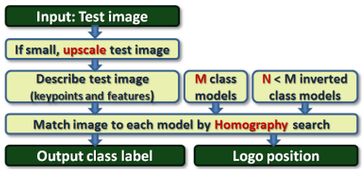


Figure 2: The schematic of the system used to locate and classify a logo in a testing image.

Logo (graphic entity that contains colors, shapes, textures and identifies organizations, goods, etc.) localization and recognition is a subproblem of object detection and recognition and a challenging pattern recognition task. Applications are in the automotive industry, sports transmissions, legal or feedback for advertising.

Logos in natural images are approached within *retrieval* systems [1] [6], [4], etc. or by *integrated detection* (localization and recognition) [2], [7], etc. Our contribution falls in the second category and consists in: (1) a new *class prototyping* method based on a central image extracted by analyzing the homographies graph and re-projecting the relevant keypoints on that image (2) a logo detection system that exhibits great performance. The main conceptual difference to previous systems is that they manually branched their process to deal with corner-cases, while we perform the branching automatically, proposing a compact and self-adjusting system.

Class Description. In the training phase (detailed in Fig. 1) we construct models (prototypes) to describe classes. In the testing phase (Fig. 2) the query image is compared with class prototypes and if they are enough similar we count a detection.

Feature extraction. The logo images are described by the Affine Difference of Gausssians (ADoG) [5] followed by the description with oriented SIFT elliptical local features. ADoG provides more keypoints on logos than other choices, while elliptical features are able to provide correctly the orientation even for circular logos.

Class Graph. All the logo crops from the same class are grouped in a weighted graph: the nodes are the logos, while an edge is created if a homography is found between that pair of logos-nodes. The edge has an weight equal to the inverse of the number of keypoints pairs matched.

The homography between two logos is found with the direct linear transformation (DLT) and 4 keypoint pairs are needed for this determination. The matching between the reference logo and the subject logo is found with RANSAC. Yet, to provide the best match, one needs to iterate more subsets than usual [3].

Next, the quality of the homographic fit is evaluated using an *error map* built as the Hellinger distance between the reference logo and the back-projected logo described with Dense SIFT.

Given the class graph, the central image is the node with the most connections.

Class Model. The class model is built by agglomerating onto the central image the suitable keypoints and their SIFT description. This information is taken from all the logo images from the main cluster of the class graph, by projecting them on the plane of the central image.

Keypoints in images directly connected to the central image are back-projected (by inverting the matching homography) on the central one. The equivalent homography between images that are not directly connected to the central image is determined by composing the homographies placed on the path between that image and the central one. The chosen path is the one that ensures minimum cumulative weight. The corresponding SIFT descriptors are merged on the central image, as they are not affected by the projection into another plane.

Color Inverted Logos: Secondary Prototype. To deal with classes containing inverted colors, a secondary model is build. If the graph contains a secondary cluster not connected to the first, the existence of color inverted case is detected and a secondary model is built by inverting the SIFT feature orientation of the main prototype.

Results. In the testing phase, we locate logos and classify them. Given at least one model for each class, we use RANSAC to match the keypoints from the query image against all the class models as shown in figure 2. To decide whether the match is correct, the average of the error map based on Dense SIFT is computed for each matching result. If, after being confronted to all the class models, no score is large enough, then the test image will be classified as "no-logo". To be able to detect small size logos, we upscale the original image 4 times.

Evaluation Procedure. Given a query image, we count a true detection if the found logo is indeed in the image and if the intersection-over-union is above 50%.

The classification rates on the tested databases are: on FlickrLogos-32: 93.33% (9.27% improvement to state of the art), BelgaLogos: 78.09% (11.76% improvement) and FlickrLogos-27: 72.8% (1.2% improvement). The method does not produces any false positives on the testing datasets, but only miss detections.

The failures are associated with logo classes with few keypoints, small logos or with extreme warping or blurring. Another situation of poor performance is for logos that are rather different from the training set.

The algorithm is also able to detect multiple logos in an image by testing the query image against all class models and retaining all homography matches that produced an average of the error map smaller than chosen threshold. Creating additional virtual classes for cases which present high variation in color (including both original and negative version) permits simultaneous detection of both cases, even in the same image.

[1] A. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo. Trademark matching and retrieval in sports video databases. In *ACM MIR*, 2007.

[2] R. Boia and C. Florea. Homographic class template for logo localization and recognition. In *Proc. of IbPRIA*, pages 487–495, 2015.

[3] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *IJCV*, 74(1):59–73, 2007.

[4] J. Krapac, F. Perronnin, T. Furon, and H. Jegou. Instance classification with prototype selection. In *ACM ICMR*, pages 431 – 4, 2014.

[5] J.M. Morel and G.Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIIMS*, 2(2):438–469, 2009.

[6] C. Ries, F. Richter, S. Romberg, and R. Lienhart. Automatic object annotation from weakly labeled data with latent structured SVM. In *CBMI*, pages 1–4, 2014.

[7] H. Sahbi, L. Ballan, G. Serra, and A. Del Bimbo. Context-dependent logo matching and recognition. *IEEE TIP*, 22:1018–1031, 2013.

Entire Reflective Object Surface Structure Understanding

Qinglin Lu[1]
qinglin.lu@u-bourgogne.fr
Olivier Laligant[1]
http://le2i.cnrs.fr/-Olivier-Laligant-?lang=fr
Eric Fauvet[1]
http://le2i.cnrs.fr/-Eric-Fauvet-?lang=fr
Anastasia Zakharova[2]
http://lmi.insa-rouen.fr/membres/10-membres/maitres-de-conference/29-zakharova-anastasia.html

[1] University of Burgundy
Le2i UMR 6306 CNRS
12,Rue de la Fonderie,71200,France
[2] INSA Rouen LMI EA3226
Avenue de l'Université,76800,France

The presence of reflection on a surface has been a long-standing problem for object recognition since it brings negative effects on object's color, texture and structural information. Because of that, it is not a trivial task to recognize the surface structure affected by the reflection, especially when the object is entirely reflective. Most of the cases, reflection is considered as noise. In this paper, our proposed approach extracts reflection motion features in the image sequence as spatiotemporal information, then sub-segment object by taking these features in order to understand the OSS. The setup of our method is straightforward, the positions of camera and object are fixed, however the light source is moving around the object in order to produce *reflection particles* (RP) on the object surface. While the RP are moving on the object surfaces, their positions, directions, and velocities are extracted in each frame as reflection motion features. These features are matched in all the frames for tracking RP in the whole sequence. We assume that the RP move smoothly along an elementary continuous surface and irregularly while passing from one surface to another. Thus, we break tracking when the motion features are irregularly compare to that in the previous frames. This guarantees to keep the trajectory of a moving RP stay on one elementary continuous surface. Later, one elementary continuous surface is segmented by employing flood fill method [5] which takes the positions in the trajectory as seeds. As this process iteratively covers all the trajectories, different surfaces of the object could be respectively labeled.

The RP are extracted by the motion history image [1, 3] (*MHI*). The *MHI* $H_\tau(x,y,t)$ can be computed from an update function $\Psi_\tau(x,y,t)$:

$$H_\tau(x,y,t) = \begin{cases} \tau & if \quad \Psi_\tau(x,y,t)=1 \\ max(0, H_\tau(x,y,t-1)-\delta) & if \quad \Psi_\tau(x,y,t)=0 \end{cases} \qquad (1)$$

Precisely, if $\Psi_\tau(x,y,t) = 1$, then the pixel at position (x,y) in t-th frame has moved. The duration τ decides the temporal extent of the movement, and δ is the decay parameter. This leads to a static scalar valued image where the more recently moving pixels are brighter.

The tracking of RP suffers from several problems: the high frequency of appearance and disappearance of the RP, the shape evolution of the RP, as well as multiple reference RP need to be tracked in the same time. Our tracker is initialized for each detection of a reference RP (C_i^t). During tracking RP in frames, positions of all tracking results are saved as the moving trajectory. The trajectory of C_i is denoted as $T(C_i) = \{p_i^1, p_i^2, ...p_i^t, \}$. One trajectory is considered as one label for a continuous surface on the object. As the RP could go through one surface in different directions, we save trajectories respectively for each direction. In this case, it ensures that one trajectory labels only one surface. On the other hand, some trajectories label the same surface.

Segmentation of the elementary continuous surfaces is to describe the surface structure of the object. As some trajectories are labeling the same surface, an iterative flood fill function is applied to merge the segmentation results of different trajectories on the same surface. The seeds which need to be flood filled are systematic sampled positions with a skip of 5 in the trajectory. Since the surface is elementary continuous, a trajectory can cover the surface regions with different brightness levels, the flood fill produces only one surface and the reflection don't produce additional sebsegments. The flood fill method which we used during the segmentation is the same for the reflection particle detection.

Since the trajectories do not have the same length, we put trajectories in an order by increasing lengths and systematic sample the positions by a skip of 5, flood fill from sampled seeds in shorter trajectories to the sampled seeds in longer trajectories. In this case, the parts of one surface which has already been labeled by seeds in shorter trajectories could be merged into other parts of this surface by the labeling of seeds in longer trajectories. As the reflection on the surface is highly variable, the segmentation phase might not cover the whole surfaces. In consequence, the final processing fills the holes which are surrounded by segmented regions with the surrounding color.

The experiments are conducted in using the camera with the resolution of 5 Mpx. A LED grow light is used to produce reflections on the object. For the outdoor experiments, two projectors are used. The number of acquired frames is depending on the complexity of the object surfaces and the number of light sources. In order to keep the number of acquired image not expanded, our LED grow light contains 30 light spots. The ground-truth images are manually labeled according to the 3D models of the objects which are obtained by a non-contact 3D digitizer VI-910.

Our primary contributions are: (1)We introduce an effective sub-segmentation method for the reflective surface structure understanding (on both specular and transparent surfaces). (2) Instead of removing reflection, we study the reflection motion and we consider it as additional information for sub-segmentation. (3) We use the reflection motion features as spatiotemporal coherence for video segmentation and fine-attributes for OSS understanding.

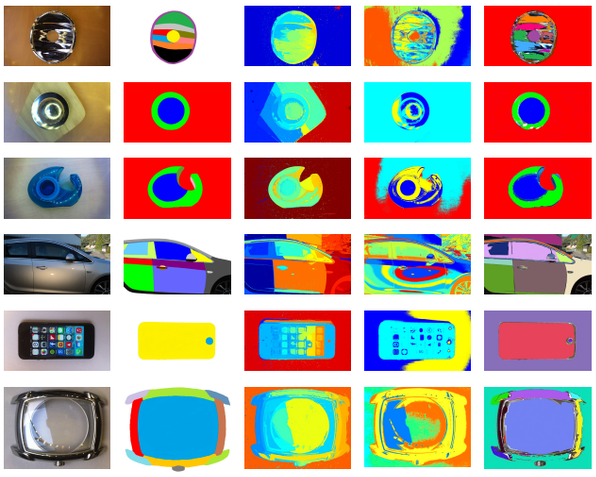


Figure 1: First column: original images. Second column: ground-truth segmentation. Third column: k nearest neighborhood graph-based segmentation [4]. Forth column: EM segmentation [2]. Last column: Segmentation by our proposed method. (better see in color)

[1] Md. Atiqur Rahman Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: Its variants and applications. *Machine Vision and Applications*, 23:255–281, 2012.

[2] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1026–1038, 1999.

[3] J. Davis. Hierarchical motion history images for recognizing human motion. *IEEE workshop DREV*, 2001.

[4] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59, 2004.

[5] A. Treuenfels. An efficient flood visit algorithm. *C/C++ Users Journal*, 12, 1994.

TennisVid2Text: Fine-grained Descriptions for Domain Specific Videos

Mohak Sukhwani
http://researchweb.iiit.ac.in/~mohak.sukhwani/
C.V. Jawahar
http://www.iiit.ac.in/~jawahar/

CVIT
IIIT Hyderabad,
India.
http://cvit.iiit.ac.in

Annotating visual content with text has attracted significant attention in recent years [1, 2, 3, 4]. While the focus has been mostly on images [2, 3, 4], of late few methods have also been proposed for describing videos [1]. The descriptions produced by such methods capture the video content at certain level of semantics. However, richer and more meaningful descriptions may be required for such techniques to be useful in real-life applications. We make an attempt towards this goal by focusing on a domain specific setting – lawn tennis videos. Given a video shot from a tennis match, we intend to predict detailed (commentary-like) descriptions rather than small captions. Figure 1 depicts the problem of interest and steps involved in our method. Rich descriptions are generated by leveraging a large corpus of human created descriptions harvested from Internet. We evaluate our method on a newly created tennis video data set comprising of broadcast video recordings of matches from London Olympics 2012. Extensive analysis demonstrate that our approach addresses both semantic correctness as well as readability aspects involved in the task.

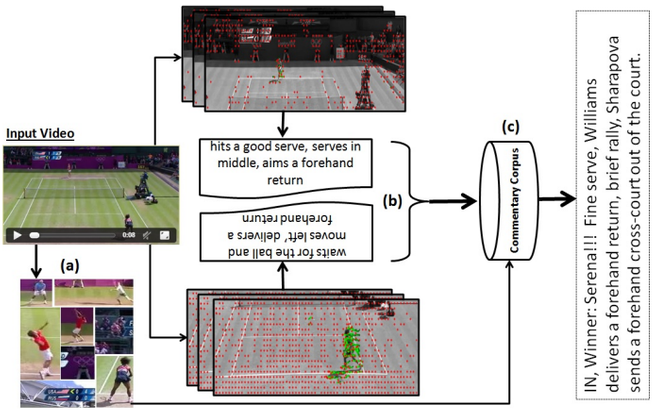


Figure 1: For a test video, the *description* predicted using our approach is dense and human-like. Above figure demonstrates our approach – a) Player Identification b) Verb Phrase Prediction c) Description Generation

Given a test video, we predict a set of action/verb phrases individually for each frame using the features computed from its neighbourhood. The identified phrases along with additional meta-data are used to find the best matching description from the commentary corpus. We begin by identifying two players on the tennis court, Figure 1(a). Regions obtained after isolating playing court regions assist us in segmenting out the candidate player regions through background subtraction using thresholding and connected component analysis. Each candidate foreground region thus obtained is represented using HOG descriptors over which a SVM classifier is trained to discard non-player foreground regions. The candidate player regions thus obtained are used to recognize players using using CEDD descriptors and Tanimoto distance.

Verb phrases are recognized, Figure 1(b), by extracting features from each frame of input video using sliding window. Since this typically results into multiple firings, non-maximal suppression (NMS) is applied. This removes low-scored responses that are in the neighbourhood of responses with locally maximal confidence scores. Once we get potential phrases for all windows along with their scores, we remove the independence assumption and smooth the predictions using an energy minimization framework. For this, a Markov Random Field (MRF) based model is used which captures dependencies among nearby phrases. We add one node for each window sequentially from left to right and connect these by edges. Each node takes a label from the set of action phrases. The energy function for nodes v, neighbourhood $\mathcal{N}$ and labels $\mathcal{L}$ is:

$$E = \sum_{p \in v} D_p(f_p) + \sum_{p,q \in \mathcal{N}} V_{pq}(f_p, f_q) \quad (1)$$

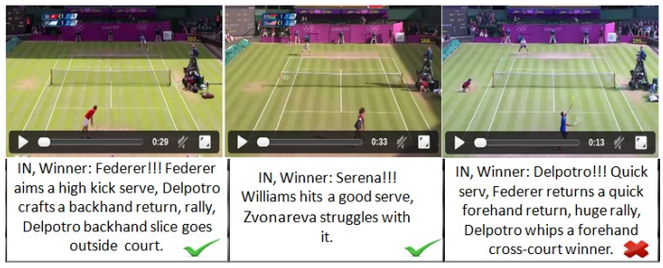


Figure 2: Sample outputs: Example videos along with their descriptions. The 'ticked' descriptions match with the ground truth, while the 'crossed' ones do not.

Here, D_p denotes *unary phrase selection cost* and V_{pq} denotes *pairwise phrase cohesion cost* associated with two neighbouring nodes. The solution of this energy minimization returns a set of actions phrases that describe the input video. We formulate the task of predicting the final description,Figure 1(c), as an optimization problem of selecting the best sentence among the set of commentary sentences in corpus which covers most number of unique words in obtained phrase set. We even employ Latent Semantic Indexing (LSI) technique while matching predicted phrases with descriptions and demonstrate its effectiveness over naïve lexical matching.

The proposed system is benchmarked against state-of-the-art methods. We compare our performance with recent methods in Table 1(right). Caption generation based approaches [1, 2] achieve significantly low score owing to their generic nature. Compared to all the competing methods, our approach consistently provides better performance. Table 1(left) demonstrates the effect of variations in corpus size on BLEU scores. It can be observed that the scores saturate soon, which validates that in domain specific settings, rich descriptions can be produced even with small corpus size.

Corp#	Voc#	B1	B2	B3	B4
100	85	0.379	0.235	0.154	0.095
500	118	0.428	0.251	0.168	0.107
5K	128	0.458	0.265	0.178	0.111
30K	140	0.460	0.277	0.182	0.113
50K	144	0.461	0.276	0.183	0.114

Method	B1	B2	B3	B4
Guadarrama [1]	0.119	0.021	0.009	0.002
Karpathy [2]	0.135	0.009	0.001	0.001
Rasiwasia [3]	0.409	0.222	0.132	0.070
Verma [4]	0.422	0.233	0.142	0.075
This work	0.461	0.276	0.183	0.114

Table 1: **Left:** Variation in BLEU score with corpus size. 'Corp#' refers to number of commentary lines, 'Voc#' is vocab dimensionality. **Right:** Performance comparison with present state-of-the-art methods. .

Our approach demonstrates the utility of the simultaneous use of vision, language and machine learning techniques in a domain specific environment to produce semantically rich and human-like descriptions. The proposed method can be well adopted to situations where activities are in a limited context and the linguistic diversity is confined.

[1] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-vemantic alignments for generating image descriptions. In *CVPR*, 2015.

[3] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[4] Yashaswi Verma and C. V. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014.

Human activity recognition in the semantic simplex of elementary actions

Beaudry Cyrille
cyrille.beaudry@univ-lr.fr
Péteri Renaud
renaud.peteri@univ-lr.fr
Mascarilla Laurent
laurent.mascarilla@univ-lr.fr

MIA lab.
University of La Rochelle
La Rochelle, France

Analyzing and recognizing human actions in videos has received considerable attention for many years in the computer vision community. Works on this topic are motivated by several potential applications (video monitoring, automatic video indexing, crowd analysis, human-machine interaction, etc). The wide variability of human actions makes it difficult to design generic methods. Many proposed approaches are based on discriminative supervised models [3, 4, 5, 6]. Other studies are focused on generative probabilistic models, and are based on LDA [2] or semi-supervised LDA [7]. However, generative models fail to match already known actions occurring in videos and it is moreover difficult to semantically analyze discovered topics.

In this paper, we present an original approach for human activities recognition in videos. It relies on a semantic representation of videos rather than a Bag of visual features approach, allowing better generalization. We characterize activities as temporal sequences of elementary actions by estimating their probabilities over time. Elementary actions are not discovered as in generative probabilistic models but learned via a robust action recognition method developed in our previous works [1]. Video sequences are characterized by critical points of optical flow and by their temporal trajectories. These features are computed at different spatio-temporal scales, using a dyadic subdivision of the sequence. A robust and generic representation of the elementary actions is then extracted using a cross-dataset learning process. A complex action is then decomposed as a sequence of elementary action proportions by transforming the decision boundaries of the classifier into probabilities. The classifier takes into account the case when there is no movement in the sequence.

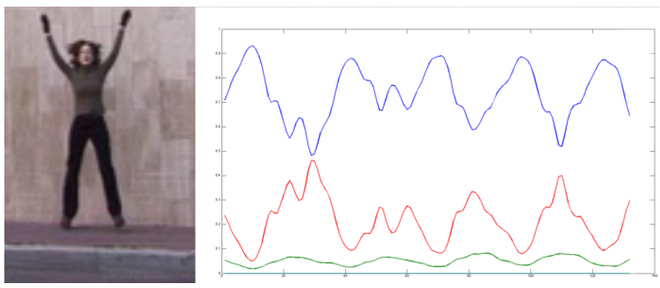


Figure 1: `Jack` action with its representation as a sequence of elementary action probabilities. The `Jack` action is composed of `Jump` and `Handwave` elementary actions. The graph represents the evolution of elementary action probabilities over time. Red curve is for `Handwave` action, blue curve for `Jump` action. The periodicity and alternation between the two elementary actions is well noticeable on the graph.

Once a frame is characterized by its L elementary action probabilities, its feature vector lies in a simplex $\mathcal{P}_L$ defined such as:
$\mathcal{P}_L = \{\pi \in R^{L+1} | \sum_{i=1}^{L+1} \pi_i = 1, \pi > 0\}$. $\mathcal{P}_L$ endowed with the Fisher information metric is a submanifold of R^{L+1}.
Figure 2 shows the global scheme for projecting activities in $\mathcal{P}_L$.

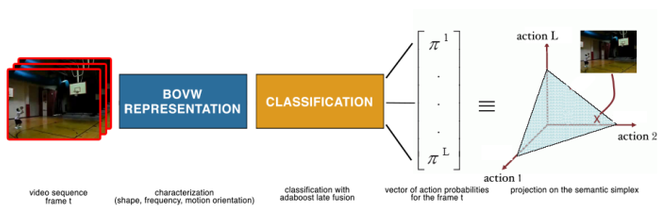


Figure 2: Global scheme for characterizing activities in the semantic simplex.

Activities are then represented as trajectories on the semantic simplex $\mathcal{P}_L$. By using a diffeomorphism F between P_L and the L-hypersphere $\mathcal{S}_L^+$, geodesic distances in P_L can be computed as shortest curves on $\mathcal{S}_L^+$. The geodesic distance between two points (π_{k1}, π_{k2}) of P_L is simply the great circle arc linking $(F(\pi_{k1}), F(\pi_{k2}))$ on S_L^+.
A robust descriptor of the trajectory shape in the simplex is obtained using the Fourier transform coefficients of the cumulative curvature in angular coordinates.
To test the discriminative performance of the proposed method, three complex actions from `UCF11`, `UCF50` datasets and `Olympic Sport` dataset are considered: `High-Jump`, `Basket-ball` and `Base-ball`. Figure 3 shows some trajectories in $\mathcal{S}_L^+$ for each activity class.

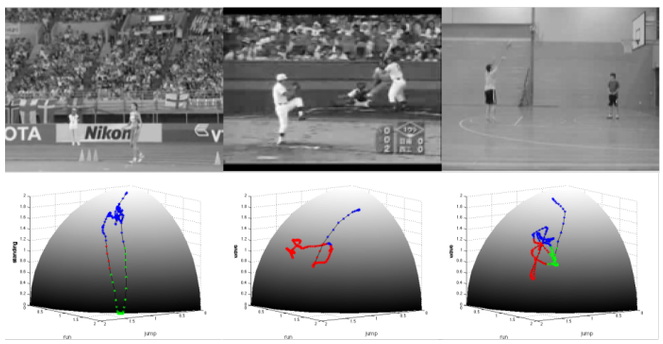


Figure 3: Examples of activity trajectories in $\mathcal{S}_L^+$: `High-Jump` activity, `Base-ball` activity, and `Basket-ball` activity.

A Leave-One-Out cross-validation test using a SVM with a RBF kernel gives a recognition rate of 96.6%, to be compared with the recognition rate of 86.6% using the STIP method [4]. The semantic aspect of our method also allows a better generalization of human activities.
Considering human actions as trajectories on the semantic manifold opens the way to different applications, such as video summary by computing a mean shape on the manifold.

[1] Cyrille Beaudry, Renaud Péteri, and Laurent Mascarilla. Action recognition in videos using frequency analysis of critical point trajectories. In *Proc. IEEE International Conference on Image Processing*, pages 1445–1449, Paris, France, 2014.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65 – 72, October 2005.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 1 –8, June 2008.

[5] M. Vrigkas, V. Karavasilis, C. Nikou, and A. Kakadiaris. Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, 119(0):27 – 40, 2014.

[6] H. Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 3169 –3176, June 2011.

[7] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Proceedings of the 2nd Conference on Human Motion*, pages 240–254, Berlin, Heidelberg, 2007. Springer-Verlag.

Character Identification in TV-series via Non-local Cost Aggregation

Ching-Hui Chen
ching@umiacs.umd.edu
Rama Chellappa
rama@umiacs.umd.edu

Department of Electrical and Computer Engineering
University of Maryland
College Park, USA

Character identification is an important task for preparing the metadata for a TV-series, since several applications, such as video summarization [5], analysis of character interactions [7], and shot retrieval [3, 8], require knowing the identities of humans in the scenes. Nevertheless, character identification in TV-series remains a challenging task since the video is usually unconstrained and the human pose varies. Existing works use the names provided in the screenplay, speech identification [6], and attributes (e.g., gender) [1], to assist person identification. Nevertheless, several prior efforts [9, 10] based on face clustering and tracking are not suitable for consistently identifying the characters in a TV-series due to shot variations and the occlusion of faces. Since the human body is more perceivable even when the face is occluded, person tracking [2] can provide additional advantages for person identification.

Recently, non-local cost aggregation methods have been shown to yield good results in establishing dense stereo correspondence [4, 11]. This framework ensures that information is effectively utilized via non-local cost aggregation on minimum spanning tree. Motivated by these works, we propose a non-local identification framework to recognize the identity of each face track and person track such that identities in the venue are consistently reported. Extending the non-local framework to solve the identification problem in a TV-series is not straightforward. Unlike pixels with identical modalities which line up in a planar graph structure, face tracks and person tracks own different modalities in the timeline. Besides, contextual information (e.g., unique presence of a subject) should be utilized to improve the identification performance.

In this paper, we propose a unified approach for identifying face and person tracks in a TV-series. We construct the track nodes to multiplex the modalities of face and clothing feature from face and person track, respectively. Our method explores the video structure by constructing the minimum-distance spanning tree (MST) from the track nodes such that track nodes that are likely to have the same identity are adjacent to each other. A typical identification task assigns the identity such that the cost of each track node is minimized. By performing the non-local cost aggregation on MST, the identity assignment becomes more reliable by minimizing the aggregated cost, which allows the information from adjacent track nodes to be utilized as a whole. Furthermore, the unique presence of a subject in a venue is taken into account by minimizing the total aggregated cost of track nodes with temporal dependency. Experimental results on TV-series datasets demonstrate the effectiveness of the proposed method.

The block diagram of our approach is illustrated in Figure 1. First, the face and person tracks form track nodes if their bounding boxes co-occur with reasonable relative positions. Track nodes can be categorized into three types: Face-body, face-only, and body-only track node. Face-body track nodes consist of both modalities from face and person tracks (See Figure 2). Some track nodes only have a single modality, either from face or person track. Face-only track nodes appear when the human body cannot be detected. On the other hand, the body-only track nodes commonly appear when actors turn their bodies around. It is clear that face-body track nodes possess more information as compared to track nodes of a single modality. Track nodes with temporal dependency are then grouped into a knot. We construct the MST from the k-knot graph such that information can be conveyed across the track nodes in the video sequence. The cost of each track node is aggregated on the MST such that track nodes with similar appearance and temporal adjacency are more likely to have the same identity. The identification problem is thus cast as a cost minimization problem per knot such that the uniqueness constraint is incorporated. In the end, face and person tracks inherit the identity of the track node they are associated with.

We conduct experiments on two TV-series datasets. Experimental results confirm that the proposed method can effectively utilize the face and person tracks for character identification.

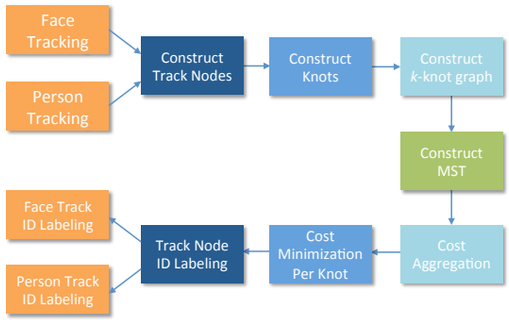


Figure 1: Block diagram of our proposed framework.



Figure 2: Face tracks (blue and green) and a person track (red) are merged into one track node.

[1] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[3] M. Everingham J. Sivic and A. Zisserman. Person spotting: Video shot retrieval for face sets. In *Image and Video Retrieval*, 2005.

[4] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang. Segment-tree based cost aggregation for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] J. Sang and C. Xu. Character-based movie summarization. In *ACM Multimedia*, 2010.

[6] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" probabilistic person identification in TV-series. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Storygraphs: Visualizing character interactions as a timeline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *ACM International Conference on Multimedia Retrieval*, 2014.

[9] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[10] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji. Constrained clustering and its application to face clustering in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[11] Q. Yang. A non-local cost aggregation method for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Prototypical Priors: From Improving Classification to Zero-Shot Learning

Saumya Jetley
sjetley@robots.ox.ac.uk
Bernardino Romera-Paredes
bernard@robots.ox.ac.uk
Sadeep Jayasumana
sadeep@robots.ox.ac.uk
Philip Torr
phst@robots.ox.ac.uk

University of Oxford
Oxford, UK

Automatic object recognition has witnessed a huge improvement in recent years due to the successful application of convolutional neural networks (CNN). This boost in performance can be explained by the replacement of heuristic parts in the previous feature representation approaches by a methodology [2, 3] based on learning the features straight from the data. The learned feature representation, which is tailored to the given learning scenario, generally outperforms heuristic approaches provided the training data is big enough. When learned over a significant sample variety, this representation captures regularities across samples of a class that help distinguish it from all the other classes.

In an alternative setup, the object recognition problem can be posed as one in which objects in real images are identified by treating them as imperfect and corrupted copies of prototypical concepts. This assumption provides an additional premise that the different samples of a class are not only similar to each other but also resemble a unique prototype. These prototypical concepts are in many cases not available, for example there does not exist a chair that contains only the essence of *chair* and nothing else. However, there are many scenarios where such prototypical instances do exist. An example of this is traffic sign recognition, in which each traffic sign class has its canonical template.

In the present work, we focus on adding this prototypical prior information into convolutional neural networks, as illustrated in Figure 1. The underlying idea is that the high-level representation learned by a CNN should be comparable to the information extracted from the prototypes. An interpretation of this is that layer-by-layer the CNN is able to learn a representation that is invariant to real world factors such as light variation, view point distortion, as described in [1], so that the representation obtained at the end of the network is invariant to all factors appearing in real images, and thus comparable to the prototype.

Prototypical information is introduced by wedging a layer before the output layer, fully connected to the C output neurons using the fixed weights $\phi(p_c) \in \mathbb{R}^k$ for all $c \in \{1,\ldots,C\}$. The new layer and its connections are shown in *blue* (*dark* for grayscale) in Figure 1. Thus, the $k \times C$ weight matrix for the last fully connected layer f_L is defined as a set of $k \times 1$ vectors $\phi(p_c)$ one for each $c \in \{1,\ldots,C\}$. In Figure 1, we use $\phi_1(p_c), \phi_2(p_c), \ldots, \phi_k(p_c)$ to represent the elements of the k-dimensional vector $\phi(p_c)$.

The modified network can now be described using the following formula:

$$\hat{y} = \underset{c\in\{1,\ldots,C\}}{\operatorname{argmax}} \; s\left(f_L\left(f_{L-1}\left(\ldots\left(f_1\left(x\right)\right)\ldots\right)\right)\right)_c = \underset{c\in\{1,\ldots,C\}}{\operatorname{argmax}} \; \langle \phi(p_c), \psi(x) \rangle, \tag{1}$$

where $\psi(.)$ and $\phi(.)$ represent the projections of input images and output labels into the joint feature space, respectively. An interpretation of this approach is that the learnable part of the network, $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^k : \psi = f_{L-1} \circ \ldots \circ f_1$, learns a non-linear mapping from the original images to a k-dimensional latent space, which in this case is defined by the prototypes.

Thus, the traditional CNN pipeline is augmented to map both the input and prototypes to a common feature space with the end goal of minimizing the final recognition error. The use of a joint embedding space, as shown in Figure 2 lends the proposed model an interesting possibility of applying it to recognize new classes not present at the training stage. This aligns the approach within the areas of zero and one-shot learning.

Conclusively, this paper makes the following contributions - (a) development of a CNN that is able to use prototypical information to guide its learning process, (b) its application to classification tasks presenting a boost in overall performance, (c) establishment of a new benchmark in logo recognition (on Belga logo dataset), and (d) the seamless application of the proposed model in zero-shot learning scenarios, given the prototypical information of new classes at run time.

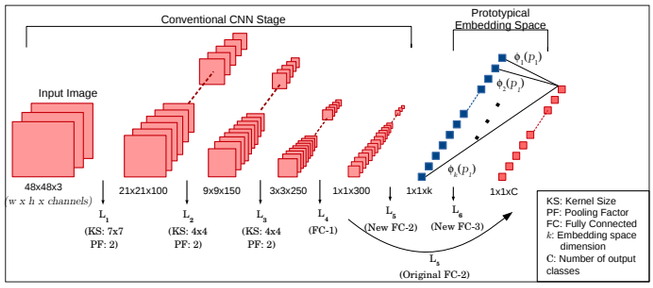


Figure 1: Network architecture with the introduction of prototypical priors. In the current experiments, k-dimensional HoG features extracted over the prototypical templates are used to define the common embedding space.

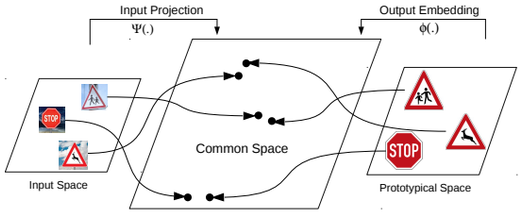


Figure 2: A joint embedding space defined by the prototypes

As observed on two different datasets of traffic signs and brand logos, results of the proposed approach are highly promising. Incorporating the given prototypes improves the classification performance. With regard to zero-shot learning, our model shows better results than a state-of-the-art competitor [4] and we show that it can be more flexibly trained for the required trade-off between seen and unseen class performance at test time.

[1] Ian Goodfellow, Honglak Lee, Quoc V Le, Andrew Saxe, and Andrew Y Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[3] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[4] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

Fast Affine Template Matching over Galois Field

Chao Zhang
zhang@scv.cis.iwate-u.ac.jp
Takuya Akashi
akashi@iwate-u.ac.jp

Graduate School of Engineering
Iwate University, Japan
Faculty of Engineering
Iwate University, Japan

In this paper, we address the problem of template matching under affine transformations with general images. Our approach is to search an approximate affine transformation over a binary Galois field. The benefit is that we can avoid matching with huge amount of potential transformations, because they are discretely sampled. However, a Galois field of affine transformation can still be impractical for exhaustive searching. To approach the optimum solution efficiently, we introduce a level-wise adaptive sampling (LAS) method under genetic algorithm framework. In LAS, individuals converge to the global optimum according to a level-wise selection and crossover while the population number is decreased by a population bounding scheme. Specifically, our paper is to infer an approximate affine transformation $\hat{T}$ from a given candidate set. In the best case, $\hat{T}$ equals to transformation $\overline{T}$. $\overline{T}$ is the closest transformation to ground truth T' among all the candidate transformations. An natural way to estimate $\hat{T}$ is to minimize SAD. Formally, our purpose can be denoted as:

$$\hat{T} = \underset{T \in \mathbb{F}_{2^{6n}}}{\arg\min} S(I_1, I_2, T). \tag{1}$$

According to [1], a general affine transformation matrix can be decomposed into $T = TrR_2SR_1$, where R_1 represents matrix operation of 1st rotation, S is scale operation in x-axis and y-axis, R_2 is 2nd rotation, Tr is translation operation in x-axis and y-axis. Transformations over each decomposed DF can be modeled by a Galois field $\mathbb{F}_{2^n}$, n is a positive integer denoting the length of binary code and 2^n is the field's size. Elements in $\mathbb{F}_{2^n}$ are expressed as binary codes. For clarity, we assume n of each decomposed DF is the same. Each DF's range is then divided into 2^n discrete segments. $T \in \mathbb{F}_{2^{6n}}$ denotes a general affine transformation in 6 DFs. Acceptable margin of error can then be guaranteed on this Galois field. To quantify the error between two transformations T_1 and T_2, following formula is defined:

$$E(T_1, T_2) = |S(I_1, I_2, T_1) - S(I_1, I_2, T_2)|. \tag{2}$$

Level-wise Adaptive Sampling (LAS): LAS aims to achieve a satisfactory error rate instead of testing the complete $\mathbb{F}_{2^n}$. From the perspective of GA, our problem can be defined as a minimization problem of SAD. In crossover operation of GA, two coded individuals swap certain portions with each other. It is a good method to span search space around a sample point in multiple directions. However, in order to optimize $\hat{T}$ in such a broad search space, two major problems should be faced: 1) how to escape from local optimum. 2) how to control the optimization response time.

It has been argued in [2] that in order to prevent GA from falling into local optimum, genetic variety should be preserved somehow. Although mutation operation can surely increase the genetic variety randomly, it can also destroy individuals which are potentially to be close to $\hat{T}$. In a broad search space, the probability to create a "suitable" diversity is very low and mutation can contrarily slow down the speed of convergence. It is worth noting that in our problem, a large enough number of randomly initialized population keeps sufficient genetic variety for converging to $\hat{T}$. During the evolution, selection operation such as roulette wheel selection is likely to select individuals which hold larger fitness for crossover operation. With the combination of selection and crossover, genetic variety decreases and the whole population converges to an optimum solution. However, if an individual happened to hold small SAD (*e.g.* a candidate area is flat) in the early stage of evolution, the whole population will easily fall into a local optimum especially when the search space is very broad. To preserve genetic variety, we select individuals from each SAD level uniformly. Each SAD level is a discrete interval which is occupied by a part of individuals. With maximum SAD in m th generation defined as S^m_{max}, minimum SAD in m th generation defined as S^m_{min} and the number of SAD level defined as σ, we can define i th SAD level as $[S^m_{min} + (i-1)(S^m_{max} - S^m_{min})/\sigma, S^m_{min} + i(S^m_{max} - S^m_{min})/\sigma]$. Each individual which is assigned to i th SAD level should have a fitness within this range. Individuals of next generation are then randomly selected from each SAD level. The number of individuals selected from each SAD level is the same. With the increase of σ, distribution of SAD in $m+1$ generation approximates to uniform distribution. Figure 1 shows an example of matching iteration.

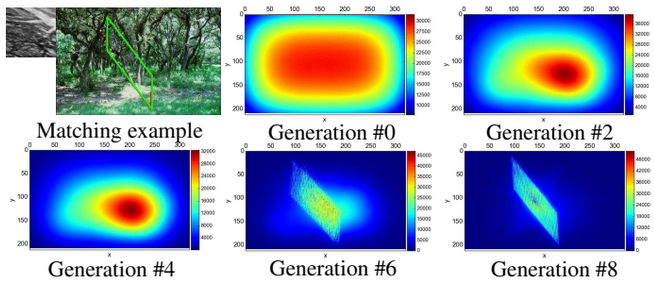


Figure 1: Heat map of matching frequency. This figure shows the frequency that each pixel has been used for calculating SAD. With the decrease of population number, the total matching frequency reduces while a more accurate candidate area can be localized.

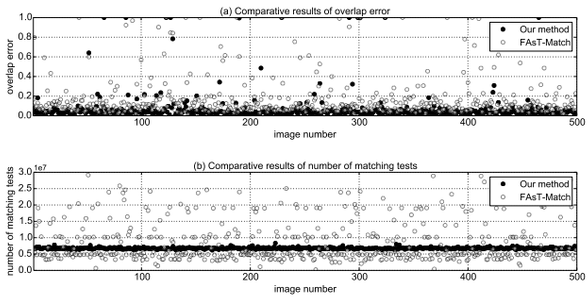


Figure 2: Comparative results with 500 images. (a) Overlap rate error of each test image. (b) Number of matching tests of each test image.

Fitness uniform selection scheme (FUSS) is proposed in [2], which selects a fitness value uniformly at first and then randomly select the nearest individual. The difference is, LAS can control the degree of uniform approximation by σ, which can directly affect the convergence speed. FUSS will take a longer time to converge, because the individuals with high fitness in FUSS make up only a small percentage of overall individuals.

Experiments: We randomly generate a ground truth affine transformation matrix for each test image, and make sure that the four corers of parallelogram generated by according matrix are all in the image. Pixels in the parallelogram are then warped to generate the square template. In our experiment, each template has a size of 100×100 pixels. We compare our method [3] against the state-of-the-art method on an evaluation data set and the results are shown in Figure 2. For more detailed discussion, please read our paper. The benchmark used and our result data can be downloaded from the following URL.
`http://cvhost.scv.cis.iwate-u.ac.jp/research/projects/affine_matching.html`

[1] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[2] Marcus Hutter and Shane Legg. Fitness uniform optimization. *IEEE Transactions on Evolutionary Computation (TEVC)*, 10(5):568–589, 2006.

[3] Simon Korman, Daniel Reichman, Gilad Tsur, and Shai Avidan. Fastmatch: Fast affine template matching. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2331–2338. IEEE, 2013.

Universal Hough dictionaries for object tracking

Fausto Milletari[1]
fausto.milletari@tum.de
Wadim Kehl[1]
kehl@in.tum.de
Federico Tombari[12]
tombari@in.tum.de
Slobodan Ilic[14]
slobodan.ilic@in.tum.de
Seyed-Ahmad Ahmadi[3]
ahmadi@cs.tum.edu
Nassir Navab[1]
navab@cs.tum.edu

[1] Computer Aided Medical Procedures
Technische Universität München
München, Germany
[2] Computer Science Department (DISI)
University of Bologna
Bologna, Italy
[3] Department of Neurology
Ludwig-Maximilians-Universität
München, Germany
[4] Siemens AG
München, Germany

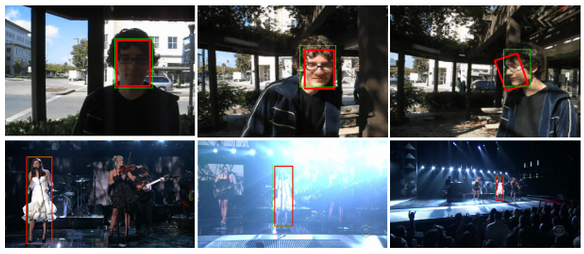


Figure 1: Few qualitative results evaluated on the benchmark dataset [2].

We propose a novel approach to online visual tracking that combines the robustness of sparse coding with the flexibility of voting based methods. Our algorithms is trained offline from a large set of patches extracted from images unrelated to the test sequences. In this way we obtain basis functions, also known as atoms, that can be sparsely combined to reconstruct local image content. During online tracking we adapt the generic knowledge learned by the dictionary to the specific object being tracked, by associating a set of votes and local object appearances to each atom. In each frame of the sequence the object's bounding box position is retrieved through a voting strategy.

1 Method

In the following we describe the three main steps of our approach.
— *Offline Dictionary learning* — we learn a dictionary of visual words from a large set of randomly sampled image patches, with the goal of obtaining a set of basis functions (i.e., *atoms*) capable of reconstructing a large variety of local image appearances. We collect a large set $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_n\}$ of image patches from generic images downloaded from the Internet and we obtain the dictionary $\mathbf{D} = \{\mathbf{d}_1, ..., \mathbf{d}_k\}$ containing k atoms by optimising the following problem with respect to $\mathbf{D}$:

$$\arg\min_{\mathbf{D}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} ||\mathbf{t}_i - \mathbf{D}\alpha_i||_2^2 + \lambda ||\alpha_i||_1. \quad (1)$$

Such *universal* dictionary encodes knowledge acquired from an amount of data that is well beyond what is available to other tracking approaches, which usually rely on single frame initialisation. As a result, our method is capable of reconstructing portions of the target object using a sparse combination of visual words while taking into account the large range of appearances that can be found in real-world situations, as supported by the findings of [1].
— *Tracker Initialisation* — we adapt the generic knowledge captured in the dictionary to the target object. We achieve this by storing votes to the bounding box centroid $\mathbf{c}$ and associated local object appearances, in correspondence to each dictionary atom. The content of the initial, manually placed, bounding box is reconstructed, patch-wise, through the dictionary $\mathbf{D}$. Image patches $\mathbf{p}_i$ are collected from the first frame of the sequence at hand in correspondence of the coordinates $\mathbf{x}_i$ and are reconstructed through $\mathbf{D}$ by solving the l_1-sparse optimisation problem

$$\arg\min_{\alpha_i} \frac{1}{2} ||\mathbf{p}_i - \mathbf{D}\alpha_i||_2^2 + \lambda ||\alpha_i||_1 \quad (2)$$

yielding the sparse coefficients α_i, and the reconstructions $\hat{\mathbf{p}}_i = \mathbf{D}\alpha_i$. Each dictionary atom is associated with a list of votes and appearances that is updated during online tracking. Using the non-zero coefficients of the vector α_i, we identify the atoms that contribute to each reconstruction and we add the votes $\mathbf{v} = \mathbf{c} - \mathbf{x}_i$ and the appearances $\hat{\mathbf{p}}$ to the relative lists.
— *Online Tracking* — We track the object across the sequence by retrieving the position of the bounding box centroid in each frame through the voting strategy. We extract image patches $\mathbf{p}_i$ from the area surrounding the last known position of the bounding box and we reconstruct them using the dictionary $\mathbf{D}$ and solving equation 2. The obtained sparse codes α_i are employed to identify the atoms involved in each reconstruction and retrieve the associated votes $\mathbf{V}_k = [\mathbf{v}_1, ..., \mathbf{v}_N]$ and appearances $\mathbf{A}_i = [\mathbf{a}_1, ..., \mathbf{a}_N]$. Each vote $\mathbf{v}_i$ contributes to the vote map after being weighted by the reciprocal of the sum of squared difference (SSD) between the reconstructions $\hat{\mathbf{p}}_i = \mathbf{D}\alpha_i$ and the appearances $\mathbf{a}_i$. The new bounding box centroid is found by identifying the location of the highest peak in the vote map. The votes and appearances associated with the dictionary atoms are therefore updated.

1.1 Results

Our method exhibits robustness towards occlusions, sudden local and global illumination changes as well as shape changes (Figure 1). We test our method on 50 standard sequences obtaining results comparable or superior to the state of the art, as shown in Figure 2.

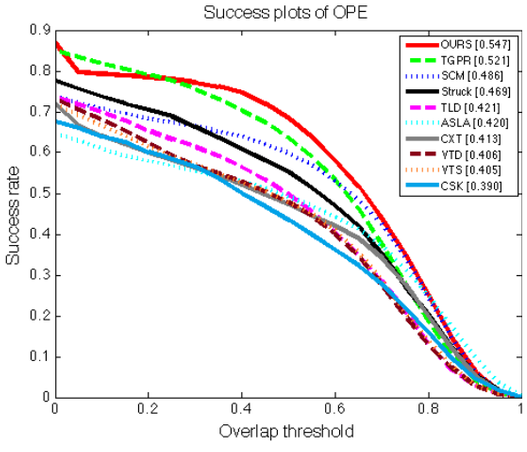


Figure 2: Success plot of our method in comparison with top performing algorithms on the 50 sequences from the CVPR13 Visual Tracking Benchmark [2]. Area under curve (AUC) is reported in brackets.

[1] Ivana Tosic and Pascal Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27–38, 2011.

[2] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

3D Tracking of Human Hands in Interaction with Unknown Objects

Paschalis Panteleris[1]
padeler@ics.forth.gr
Nikolaos Kyriazis[1]
kyriazis@ics.forth.gr
Antonis A. Argyros[21]
argyros@ics.forth.gr

[1] Insitute of Computer Science, FORTH,
N. Plastira 100, Vassilika Vouton,
GR70013, Heraklion, Crete, Greece
[2] Computer Science Department, University of Crete,
Heraklion, Crete, Greece



Figure 1: Method overview. Left: input depth and color frames. Middle: Object depth segmented using the fingertip 3D positions. Partially scanned object model and hand models. Right: 3D Rendering of the scene and final scanned model.

Experiment	Proposed mean/median error	[2], GT model mean/median error	[2], Scanned model mean/median error
Single hand, cat	0.42 / 0.39	0.47 / 0.43	0.45 / 0.43
Single hand, spray	0.65 / 0.63	0.70 / 0.53	0.63 / 0.47
Two hands, cat	0.38 / 0.34	0.33 / 0.31	0.44 / 0.39
Two hands, spray	0.59 / 0.44	0.51 / 0.38	0.62 / 0.41

Table 1: Hand tracking accuracy (in cm) measured on the synthetic datasets. The accuracy of the method is close to that of [2], although the latter assumes that the object model is known a priori.

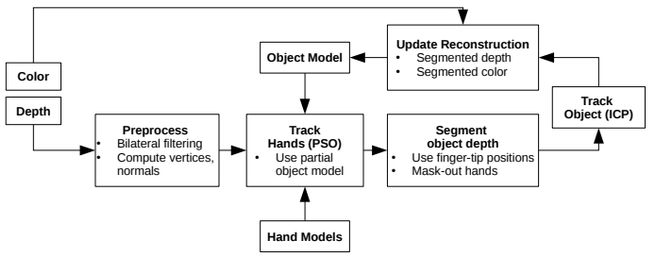


Figure 2: Work flow of the proposed method.

The analysis and the understanding of object manipulation scenarios based on computer vision techniques can be greatly facilitated if we can gain access to the full articulation of the manipulating hands and the 3D pose of the manipulated objects. Currently, there exist methods for tracking hands in interaction with objects whose 3D models are known [2]. There are also methods that can reconstruct 3D models of objects that are partially observable in each frame of a sequence [3]. However, no method can track hands in interaction with unknown objects, ie objects whose 3D model is not known a priori.

In this paper we propose a novel approach that can track human hands in interaction with unknown objects. As illustrated in Fig.1, the input to the method is a sequence of RGBD frames showing the interaction of one or two hands with an unknown object. Starting with the raw depth map (left) we perform a pre-processing step and compute the scene point cloud. We employ an appropriately modified model based hand tracker [4] and temporal information to track the hand 3D positions and posture (middle bottom). In this process, a progressively built object model is also taken into account to cope with hand-object occlusions. We use the estimated fingertip positions of the hand to segment the manipulated object from the rest of the scene (middle top). The segmented object points are used to update the object position and orientation in the current frame and are integrated into the object 3D representation (right).

More specifically, the work flow of the proposed approach consists of five main components linked together as shown in Fig. 2. At a first, preprocessing stage, the raw depth information from the sensor is prepared to enter the pipeline. A point cloud is computed along with the normals for each vertex. Then, the user's hands are tracked in the scene. An articulated model for the left and right hands, with 26 degrees of freedom each, is fit to the pre-processed depth input. The current, possibly incomplete (or even empty, for the first frame) object model is incorporated to hand tracking to assist in handling hand/object occlusions.

Using the computed 3D location of the user's hands as well as the last position of the (possibly incomplete) object model, the region of the object is segmented in the input depth map. The hands are masked-out from the observation, by comparing it to the rendered hand models.

Object tracking is achieved using a mutli-scale ICP [1]. The segmented object depth is used for a coarse to fine alignment with the (partially reconstructed) object model.

Finally, the segmented and aligned depth data of the object with the current, partial 3D model are merged. The object's 3D model is maintained in a voxel grid with a Truncated Signed Distance Function (TSDF) [3] representation.

After having processed all the frame of a given sequence, the 3D model of the object is reconstructed, provided that every part of the object was observed at at least one frame. Besides the accurate tracking of the hands, the proposed method provides an accurate 3D model of the object in the form of texture-mapped 3D mesh.

The proposed method was tested quantitatively and qualitatively in sequences where a person manipulates objects of different sizes, with either one or two hands. Table 1 shows the mean and median hand tracking error over a sequence with known ground truth. The first column shows that for the proposed method (object model is not known). The second shows that for [2] (perfectly accurate object model - ground truth - a priori known). The obtained results demonstrate that the hand tracking accuracy of our method is comparable to that o[2], although our method is not aware of the object model. Moreover, the comparison of the reconstructed object models to the actual ones shows only minor 3D reconstruction errors. Qualitative results obtained from a number of experiments are available at `http://youtu.be/9r43PtJ0Fwg`.

This work was partially supported by the EU FP7-ICT-288533 project ROBOHOW.COG.

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[2] Nikolaos Kyriazis and Antonis Argyros. Scalable 3d tracking of multiple interacting objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437, 2014.

[3] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[4] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, Dundee, UK, Aug. 2011.

Handling Data Imbalance in Automatic Facial Action Intensity Estimation

Philipp Werner
Philipp.Werner@ovgu.de
Frerk Saxen
Frerk.Saxen@ovgu.de
Ayoub Al-Hamadi
Ayoub.Al-Hamadi@ovgu.de

Otto von Guericke University
Magdeburg, Germany
www.iikt.ovgu.de/nit

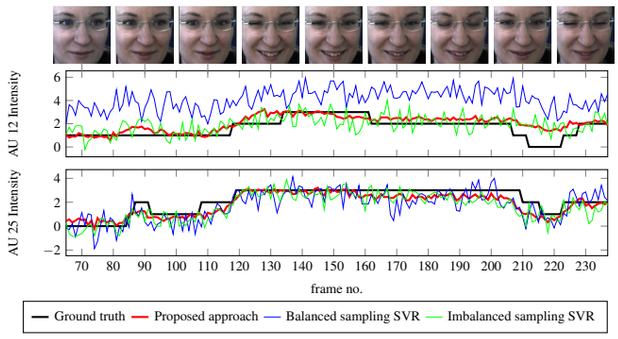


Figure 1: Intensity estimation on test set for AU 12 (lip corner pulling) and 25 (lip parting). Our approach, the MIDRUS SVR ensemble (red), is compared with balanced sampling SVR (green), imbalanced sampling SVR (blue), and ground truth (black).

Automatic Action Unit (AU) intensity estimation is a key problem in facial expression analysis. But limited research attention has been paid to the inherent class imbalance, which usually leads to suboptimal performance. To handle the imbalance, we propose (1) a novel multiclass under-sampling method and (2) its use in an ensemble. We compare our approach with state of the art sampling methods used for AU intensity estimation. Multiple datasets and widely varying performance measures are used in the literature, making direct comparison difficult. To address these shortcomings, we compare different performance measures for AU intensity estimation and evaluate our proposed approach on three publicly available datasets, with a comparison to state of the art methods along with a cross dataset evaluation.

Sampling Strategy On the one hand strong imbalance decreases performance on the minority class(es), and on the other hand under-sampling may drop relevant information about the majority class(es). We propose to choose a compromise. Instead of removing the imbalance or ignoring it, we reduce it with a method that we call *Multiclass Imbalance Damping Random Under-Sampling (MIDRUS)*. It is an algorithm with two steps: (1) calculating the number of samples to select from each class, and (2) randomly under-sample the classes without repetition according to the counts calculated in step (1).

Given that we have M classes $i = 1,...,M$ and n_i is the absolute frequency of class i in the dataset, then the number of samples $n_i^\star$ to select from class i is calculated as follows.

$$n_i^- = \lceil s \cdot (n_i)^{1-\alpha} \rceil, \text{with } s = \beta \frac{n_{f(k)}}{(n_{f(k)})^{1-\alpha}}, \quad (1)$$

$$n_i^\star = \min\{n_i, n_i^-\}. \quad (2)$$

In (1), $\alpha \in [0,1]$ is the imbalance damping parameter. It controls to which extend the imbalance is reduced, i.e. $\alpha = 1$ aims at total balancing of classes, $\alpha = 0$ keeps the imbalance, and an α in between reduces it to a certain degree. With $\alpha > 0$, the term $(n_i)^{1-\alpha}$ calculates new and more balanced class ratios. Next, these are scaled by a common factor s, which controls the total number of samples to be selected. Fig. 2 illustrates a typical use-case of MIDRUS with $\alpha = 0.5$; the imbalance is damped by taking the square root of sample counts and scaling the results in a way that 75% of the second most frequent class's samples are selected.

Ensemble It is state of the art to handle data imbalance by combining ensemble methods with sampling strategies. We propose to combine MIDRUS with an ensemble to further improve predictive performance.

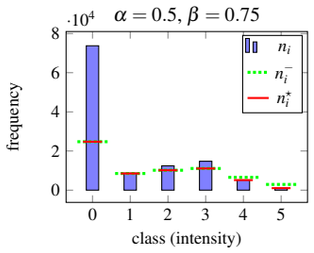


Figure 2: Multiclass Imbalance Damping Random Under-Sampling (MIDRUS) example. MIDRUS damps imbalance, which improves performance (see Fig. 3). β adjusts the number of selected samples.

Dataset	**Bosphorus**	**DISFA**	**UNBC**
Measure	ICC(3,1)$_d$	ICC(3,1)$_d$	PCC$_c$
Mavadati [3]		0.235	
Kaltwang [1]			0.306
EasyEnsemble	0.340	0.362	0.301
SVR Ensemble	0.553	0.346	0.286
	0.533	0.412	0.301
	0.603	**0.439**	**0.311**

imbalanced ‘balanced’ proposed: MIDRUS

Figure 3: Cross validated performances (mean across AUs).

We use *bagging* and apply MIDRUS T times to train T prediction models, each with one of the selected training subsets. For aggregation of the model outputs we train a fusion model by subsampling the training set with MIDRUS again.

Due to the benefit of continuous output we use Support Vector Regression (SVR) models, but the MIDRUS ensemble can also be trained with other models, including classification models.

Experiments In Fig. 3 we compare several methods on three databases. On the Bosphorus dataset, EasyEnsemble [2] is clearly outperformed by our proposed SVR Ensemble. MIDRUS improves performance compared to using the originally imbalanced data and to balancing it with the second most frequent class (‘balanced’). On the DISFA database all ensemble methods outperform the person-independent modeling results reported by Mavadati and Mahoor [3]. MIDRUS also performs best on DISFA, but the advantage over EasyEnsemble is lower than for Bosphorus. On the UNBC-McMaster dataset we compare to Kaltwang *et al.* [1] (Relevance Vector Regression on imbalanced data), but do not observe clear benefits of the ensemble methods. In contrast to Kaltwang *et al.*, we use a fully automatic landmark detector and a much simpler alignment, which is less suited for out-of-plane head poses that occur frequently in UNBC-McMaster. A better face alignment would probably improve the results obtained with ensemble methods. Nevertheless, MIDRUS still slightly outperforms the results of Kaltwang.

[1] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous Pain Intensity Estimation from Facial Expressions. In *Advances in Visual Computing*, LNCS, pages 368–377. Springer Berlin Heidelberg, 2012.

[2] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on System, Man and Cybernetics*, 39(2):539–550, 2009.

[3] S.M. Mavadati and M.H. Mahoor. Temporal Facial Expression Modeling for Automated Action Unit Intensity Measurement. In *22nd International Conference on Pattern Recognition (ICPR)*, pages 4648–4653, August 2014. doi: 10.1109/ICPR.2014.795.

Hierarchical rank-based veiling light estimation for underwater dehazing

Simon Emberton
s.emberton@qmul.ac.uk
Lars Chittka
l.chittka@qmul.ac.uk
Andrea Cavallaro
a.cavallaro@qmul.ac.uk

Centre for Intelligent Sensing,
Queen Mary University of London

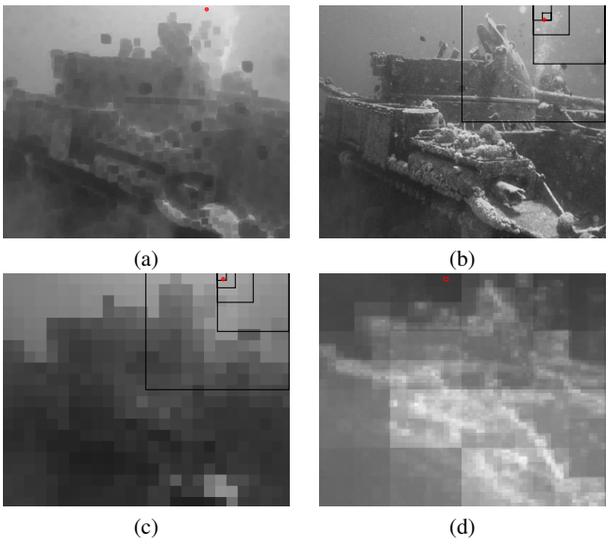


Figure 1: Comparison of veiling light estimation methods. Veiling light estimates are taken from locations highlighted in red. Current methods are biased toward areas of the Shipwreck image with bright objects (e.g. bubbles) whereas our approach is able to avoid these regions. (a) Underwater Dark Channel Prior [3]. (b) Quad-tree subdivision methods on the input image converted to greyscale [7] and (c) into blocks that are filled with their minimum value [9]. (d) Proposed method.

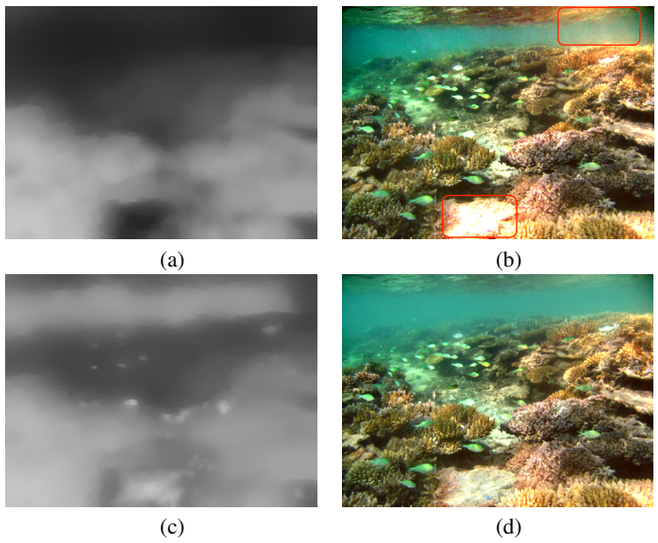


Figure 2: Transmission map (a) generated with green-blue dark channel (refined with weighted least squares (WLS) filter [4]) and (b) the resulting dehazed image. Red boxes highlight an area which is oversaturated due to a bright object being underestimated in the transmission map and a background region containing noise and artefacts. (c) Transmission map with adaptations for truncated and background regions (refined with WLS filter) (d) and the proposed resulting dehazed image.

In this paper we introduce a single image dehazing approach for underwater scenes. We employ a hierarchical rank-based method to estimate the veiling light and an adapted transmission estimation step which prevents oversaturation and artefacts in the dehazed image.

Veiling light is the atmospheric light that is scattered from particles in a hazy scene into the line of sight of an imaging device, contributing to image degradation and resulting in images with low contrast and colour cast [8]. Veiling light estimation is an important step in single image dehazing because it identifies the colour of the atmospheric light which the dehazing process aims to remove. The veiling light can be best estimated from the region of an image that is the most haze-opaque [6]. The *transmission* is the part of the light that is not scattered by the haze and reaches the camera [8].

Single image dehazing methods improve the visibility of scenes which have been degraded due to scattered light such as fog, haze and smoke by removing the haze in a scene so that objects can be seen more clearly [6]. Dehazing methods can also be used to improve the visibility in underwater images [1, 2, 3, 5].

One of the major challenges for underwater dehazing approaches is to avoid selecting pixels from bright objects during the estimation of the veiling light. We propose a method which aims to avoid objects in the veiling light detection step. We find that regions with objects contain more edges and show more variation within each colour channel than pure veiling light regions. We also consider all regions of an image in the final decision of which region is the most heavily veiled. Fig. 1 compares state-of-the-art veiling light estimation methods together with our proposed approach.

Another challenge is to estimate a transmission map which avoids output dehazed images suffering from oversaturation and artefacts in the background regions. A region-based approach is used to find optimal transmission values for areas that suffer from oversaturation. We also locate background regions through superpixel segmentation and clustering, and adapt the transmission values in these regions so to avoid artefacts (Fig. 2).

The dehazed images produced by our method show good performance in comparison to the state of the art across a range of underwater images when evaluated both subjectively and on a number of commonly used quantitative measures.

[1] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice. Initial results in underwater single image dehazing. In *Proc. of IEEE OCEANS*, Sept. 2010.

[2] J. Y. Chiang and Y. C Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. on Image Processing*, 21(4):1756–1769, Apr. 2012.

[3] P. Drews-Jr, E. Nascimento, F. Moraes, S. Botelho, M. Campos, R. Grande-Brazil, and B. Horizonte-Brazil. Transmission estimation in underwater single images. In *Proc. of IEEE ICCV*, Dec. 2013.

[4] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. on Graphics*, 27(3):67, Aug. 2008.

[5] A. Galdran, D. Pardo, A. Picon, and A. Alvarez-Gila. Automatic red-channel underwater image restoration. *Journal of Visual Communication and Image Representation*, 26:132–145, Jan. 2015.

[6] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Trans. on PAMI*, 33(12):2341–2353, Dec. 2011.

[7] J. H. Kim, J. Y. Sim, and C. S. Kim. Single image dehazing based on contrast enhancement. In *Proc. of IEEE ICASSP*, May 2011.

[8] H. Koschmieder. *Theorie der horizontalen Sichtweite: Kontrast und Sichtweite*. Keim & Nemnich, 1925.

[9] D. Park, H. Park, D. K. Han, and H. Ko. Single image dehazing with image entropy and information fidelity. In *Proc. of IEEE ICIP*, Oct. 2014.

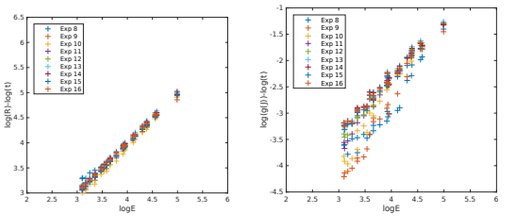
The intrinsic error of exposure fusion for HDR imaging, and a way to reduce it

Raquel Gil Rodríguez
raquel.gil(at)upf.edu
Javier Vazquez-Corral
javier.vazquez(at)upf.edu
Marcelo Bertalmío
marcelo.bertalmio(at)upf.edu

Department of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain

Human visual system is capable to capture details in dark and bright areas at the same time despite the dynamic range of the scene. Digital cameras, however, have a limited dynamic range, therefore only capturing dark or bright areas in a single shot. Considering that, Mann and Picard [6] presented a method for the creation of high dynamic range (HDR) radiance map by exposure fusion of a stack of low dynamic range (LDR) images. There is a wide range of HDR literature: assuming static scenes and no camera motions [2, 6], considering motion on camera and/or in the scene [4], and most recently video creation [5].

Camera sensor provides the irradiance, RAW values (which are proportional to the radiance scene), but then cameras apply linear and non-linear transformations, and a quantization step are applied resulting in a non-RAW image (such as JPEG). All approaches based on exposure fusion applied on these non-RAW images share a set of assumptions: 1) different color channels are independent, 2) in-camera non-linear correction curves are monotonic and smooth, but arbitrary in shape, and 3) the camera response function (CRF) remains constant while changing the exposure. Our work shows that these assumptions are not preserved when working with non-RAW formats, because camera may introduce changes from one exposure to the next. Thus, the exposure fusion of non-linear images will introduce more error in the final HDR radiance map, than computing it directly from RAW linear data.



(a) (b)

Figure 1: Plots computed for points in the colored squares in the dark color checker in *LuxoDoubleChecker* scene from Fairchild dataset [3], using (a) the RAW values, and (b) the JPEG values. While in both cases the plots should theoretically be a single line of slope one, it can be seen that in the JPEG case the points are more dispersed than in the RAW case.

Our algorithm follows the approach in [8, 9], which performs color stabilization based on camera color processing pipeline described in [1]. A set of LDR images with different exposure times are acquired $I_j,\ j = 1,\ldots,N$ of the same HDR static scene. From these input images, we first select the best exposed image $I_{N'}$ as the one with less over and under-exposed pixels. For one-to-one pixel correspondences we perform image registration, then for each $(I_j,\ I_{N'})$ pair we undo the gamma correction and finally we compute the color correction matrix H_j that transforms colors from gamma corrected I_j to gamma corrected $I_{N'}$. In the end, we obtain the new transformed linear I'_j images.

Algorithm 1 HDR radiance map generation

Input: set of N images
Result: *HDR* radiance map
1. Select the best exposed image $I_{N'}$,
2. Register the images,
3. For each pair $(I_j,\ I_{N'}),\ j = 1,\ldots,P$, [8]:
  3.1. Undo γ_j correction, and
  3.2. Estimate H_j color correction matrix.
4. Average the new linear images $I'_j = H_j \cdot I_j^{1/\gamma_j}$.

In order to evaluate our results, we use the Fairchild dataset [3] which provides linear data (RAW), and its correspondent non-linear data (JPEG). A ground truth (GT) is created using the linear data, thus we computed the PSNR of Debevec and Malik results, because most of recent methods follow its same assumptions, and our approach with the GT. In table 1, we show that in terms of PSNR our method improves the average by $\sim$ 2dB. We present qualitative results in Fig. 2, where our results are more contrast, and therefore, more details are perceived.

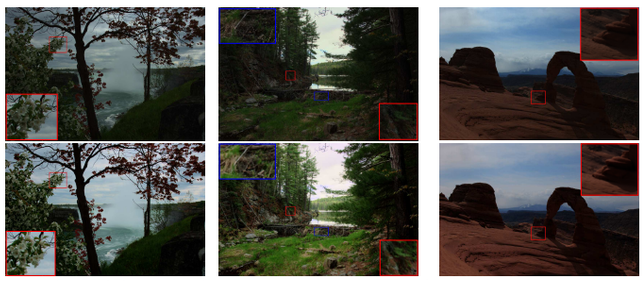


Figure 2: HDR results (tone mapped with [7]) obtained with the method of Debevec and Malik (1st row) and with our approach (2nd row).

Table 1: PSNR on the color channels (R,G,B) and the luminance channel (L), for the method of Devebec and Malik [2] and the proposed method.

	$PSNR_R$	$PSNR_G$	$PSNR_B$	$PSNR_L$
Devebec and Malik [2]	25.70	26.15	26.80	26.30
Proposed	**28.47**	**28.73**	**28.37**	**28.86**

In conclusion, we present an algorithm to obtain an HDR map from a set of LDR images which is based upon the digital camera imaging pipeline, and useful when working with non-linear data as JPEG. Results show that our algorithm quantitatively outperforms the work of Debevec and Malik in most of the cases considered.

[1] S. Bianco, A. Bruna, F. Naccari, and R. Schettini. Color space transformations for digital photography exploiting information about the illuminant estimation process. *JOSA A*, 29(3):374–384, 2012.

[2] P. E. Debevec and J. Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, pages 369–378, New York, NY, USA, 1997.

[3] M. D Fairchild. The HDR photographic survey. *Color and Imaging Conference*, 2007(1):233–238, 2007.

[4] J. Hu, O. Gallo, K. Pulli, and X. Sun. HDR Deghosting: How to deal with Saturation ? In *CVPR*, 2013.

[5] J. Kronander, S. Gustavson, and J. Unger. Real-time HDR Video Reconstruction for Multi-sensor Systems. In *ACM SIGGRAPH 2012 Posters*, SIGGRAPH '12, pages 65:1–65:1, New York, NY, USA, 2012. ACM.

[6] S. Mann and R. W. Picard. On Being 'undigital' With Digital Cameras: Extending Dynamic Range By Combining Differently Exposed Pictures. In *Proceedings of IS&T*, pages 442–448, 1995.

[7] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, July 2002.

[8] J. Vazquez-Corral and M. Bertalmío. Color Stabilization Along Time and Across Shots of the Same Scene, for One or Several Cameras of Unknown Specifications. *Image Processing, IEEE Transactions on*, 23(10):4564–4575, Oct 2014.

[9] J. Vazquez-Corral and M. Bertalmío. Simultaneous blind gamma estimation. *IEEE Signal Process. Lett.*, 22(9):1316–1320, 2015.

Detecting Change for Multi-View, Long-Term Surface Inspection

Simon Stent[1]
sistent@cantab.net
Riccardo Gherardi[2]
riccardo.gherardi@crl.toshiba.co.uk
Björn Stenger[2]
bjorn.stenger@crl.toshiba.co.uk
Roberto Cipolla[1]
rc10001@cam.ac.uk

[1] Machine Intelligence Lab
Department of Engineering
University of Cambridge
Cambridge, U.K.

[2] Toshiba Research Europe Ltd.
Cambridge, U.K.

We describe a system for the detection of changes in multiple views of a textured surface taken at different times by a moving camera. Our motivation is the development of a non-contact inspection system, summarised in fig. 1, to be used for detecting anomalous visual changes on surfaces - in this case on concrete tunnel linings. This application is of increasing social importance as tunnels and other large-scale infrastructure age and more efficient methods for structural inspection are required to allow their continued safe operation.

The problem is challenging for several reasons: (i) Size and nature of changes. Changes of interest are often small and subtle – e.g. a fattening in the width of a hairline crack or a patch of discolouration caused by organic growth or surface damage. (ii) Nuisance factors. A sizeable proportion of the observed change over time is caused by nuisance factors, either internal to the acquisition system (such as different image sensors, capture settings or lighting setup) or due to external causes (for example, seasonal changes of temperature and humidity). (iii) Registration error. Achieving the pixel-accurate registration typically required for change detection is challenging because neither the sensor positions nor the tunnel geometry can be reliably determined. Parallax errors are common.

We address these challenges by first using a structure-from-motion pipeline to approximately register images from our robotic inspection rig to the reconstructed surface of interest. Given a pair of registered image patches from different times, our main contribution is a novel approach to detect changes between the patches using a two-channel convolutional neural network (CNN). CNNs have recently been shown to be very effective at learning invariance to certain modes of image variability, but require large amounts of labelled image data to train. We create an unlimited source of negative pairs (i.e. patches where no abnormal change has occurred) by taking registered viewpoints from different cameras from the same time. We supplement this with a smaller dataset of negative pairs across the different test times from regions where no changes of interest have occurred. This requires a limited effort in coarsely labelling a small subset of the test data. Together, these negative pairs capture much of the natural variance from nuisance factors and registration error. For the positive (changed) pair generation, we provide randomly sampled pairs as well as synthetically generated changes using a crack model (fig. 2). The homogeneity of the tunnel environment allows a network to generalize well from a manageable amount of labelled ground-truth data.

Our approach is similar to [1], who learn a two-channel CNN for the inverse problem of similarity measurement between image patch pairs. A key difference is that we train directly on a mixture of synthetic data and task data generated by our own pipeline, allowing us to learn task-specific invariances for the improved detection of changes.

We evaluate our system using three datasets captured from a live tunnel over two months (fig. 3). A trained inspector was tasked with simulating real changes in the tunnel between captures and a set of ground truth change images were generated for testing. We compare our method against our existing probabilistic change detection approach [2] and against the results of a manual inspection carried out by a second trained inspector in the field (table. 1). This comparison is of particular importance to industry, since manual inspection is still in common practice. Finally, we visualise our system's output using the Google Maps API (fig. 4).

[1] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.

[2] S. Stent, R. Gherardi, B. Stenger, K. Soga, and R. Cipolla. Visual change detection on tunnel linings. *Machine Vision and Applications*, pages 1–12, 2014.

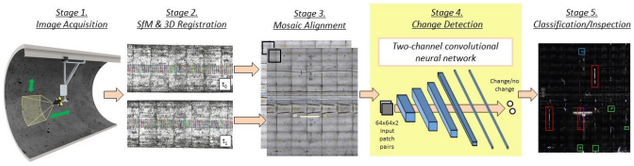
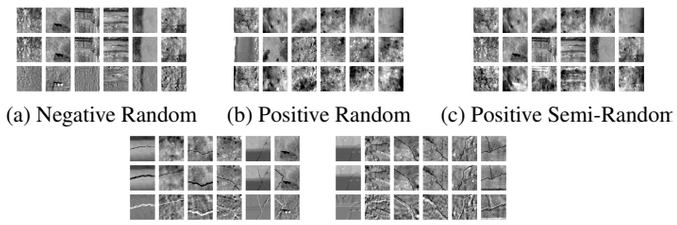


Figure 1: System overview. The main novelty is in stage 4, in which changes are detected between registered sets of image mosaics captured at different times. We propose and evaluate a new approach using a two-channel CNN. The network learns a model for normal modes of image variation, so as to detect abnormal changes with fewer false positives.



(a) Negative Random (b) Positive Random (c) Positive Semi-Random

(d) Positive Crack (e) Negative Crack

Figure 2: Sample training pairs (rows 1+2) and their difference images (row 3) from different training sets: (a) negative (unchanged) pairs; (b) positive (changed) random pairs, with both members chosen randomly; (c) semi-random positive pairs, combining (a) and (b); (d) positive crack pairs, including crack appearance/disappearance, extension and widening; (e) negative crack pairs.

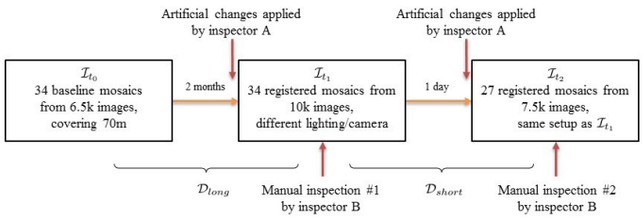


Figure 3: Timeline and datasets gathered for the inspection experiments.

Inspection dataset:	$\mathcal{D}_{short}$ (1 day)			$\mathcal{D}_{long}$ (2 months)		
FPR per pixel:	0.01	0.05	0.10	0.01	0.05	0.10
Manual inspection	0.29	0.29	0.29	**0.58**	0.58	0.58
Prior method [2]	0.20	0.55	0.64	0.00	0.00	0.32
Proposed method	**0.73**	**0.87**	**0.87**	0.26	**0.65**	**0.84**

Table 1: Percentage of artificial changes detected by the compared systems at different false positive rates. Changes are considered detected if they are greater than >50% positively labelled.

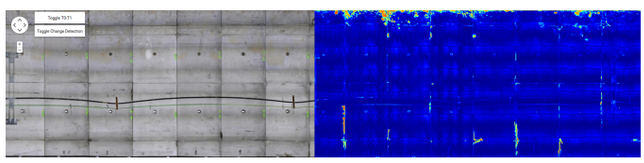


Figure 4: Visualisation using Google Maps API with maximum resolution of 0.3mm/px. Left: mosaic image of an 8m section; right: change mask.

Part Localization using Multi-Proposal Consensus for Fine-Grained Categorization

Kevin J. Shih
kjshih2@illinois.edu
Arun Mallya
amallya2@illinois.edu
Saurabh Singh
ss1@illinois.edu
Derek Hoiem
dhoiem@illinois.edu

University of Illinois
Urbana-Champaign
IL, US

The most common approach to keypoint localization is to learn a set of keypoints detectors to model appearance and an associated spatial model [3, 4, 5, 9] to capture their spatial relations. Individual keypoint detectors typically model local appearance and thus rely on expressive spatial models to capture long range dependencies. Alternatively, the keypoint detectors could condition their predictions on larger spatial support and jointly predict several keypoints [2], then the need for expressive spatial models could be eliminated, leading to simpler models.

For effective fine-grained category detection, the keypoint localization methods must have high accuracy, low false positive rates, and low false negative rates. Missed or poorly localized predictions make it impossible to extract the relevant features for the task at hand. If a keypoint is falsely determined to be present within a region, it is hard to guarantee that it will appear at a reasonable location. In the case of localizing keypoint-defined regions of an image, such as head or torso of a bird, a single outlier in the keypoint predictions can significantly distort the predicted area. This specific case is noteworthy, as several of the current best-performing methods on the CUB 200-2011 birds dataset [8] rely on deep-network based features extracted from localized part regions [1, 3, 4, 9].

In this work, we tackle the problem of learning a keypoint localization model that relies on larger spatial support to jointly localize several keypoints and predict their respective visibilities. Leveraging recent developments in Convolutional Neural Networks (CNNs), we introduce a framework that outperforms the state-of-the-art on the CUB dataset. Further, while CNN-based methods suffer from a loss of image resolution due to the fixed-sized inputs of the networks, we introduce a simple sampling with outlier rejection scheme that allows us to work around the issue without the need to train cascades of coarse-to-fine localization networks [6, 7]. Finally, we test our predicted keypoints on the fine-grained recognition task. Our keypoint predictions are able to significantly boost the performance of current top-performing methods on the CUB dataset.

We design our model to simultaneously predict keypoint locations and their visibilities for a given image patch. Given N keypoints of interest, we train a network to output an N dimensional vector $\hat{v}$ and a $2N$ dimensional vector $\hat{l}$ corresponding to the visibility and location estimates of each of the keypoints k_i, $i \in \{1,N\}$, respectively. The corresponding groundtruth targets during training are v and l. We define v to consist of indicator variables $v_i \in \{0,1\}$ such that $v_i = 1$ if keypoint k_i is visible in the given Edge Box image before padding is performed, and 0 otherwise. The groundtruth location vector l is of length $2N$ and consists of pairs (l_{x_i}, l_{y_i}) which are the normalized $(\tilde{x},\tilde{y})$ coordinates of keypoint k_i with respect to the un-padded Edge Box image. Output predicted from the network, $\hat{v}_i \in [0,1]$, acts as a measure of confidence of keypoint visibility.

To share the information across categories, our model is trained in a category agnostic manner. At test time, we efficiently sample each image with Edge Boxes, make predictions from each Edge Box, and reach a consensus by thresholding for visibility and reporting the medoid. Our method is illustrated in Fig. 1.

Results We evaluate our prediction model on the Caltech-UCSD Birds dataset [8]. This dataset contains 200 bird categories with 15 keypoint location and visibility labels for each of the total of 11788 images. We first evaluate our keypoint localization and visibility predictions against other top-performing methods and demonstrate state-of-the-art resutls in both keypoint localization and visibility. Next, we demonstrate their effectiveness in the fine-grained categorization task by using the predicted keypoints to align head and torso regions, then extracting finetuned AlexNet features from the localized regions to classify with a linear SVM. While this essentially re-creates the classification step of Zhang et al. [9], substituting in our better localized parts improves their accuracy by over 4% when ground truth bounding boxes are not provided.

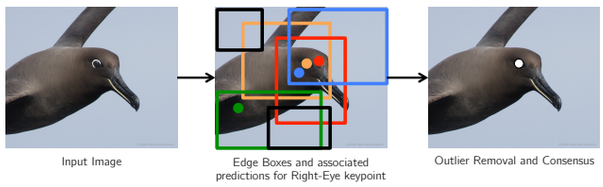


Figure 1: The pipeline of our keypoint localization process: Given an input image, we extract multiple Edge Boxes. Using each edge box, we make predictions for the location of each of the 15 keypoints, along with their visibility confidences. We then find the best predicted location by performing confidence thresholding and finding the medoid. The process is illustrated for the right eye keypoint (Black edge boxes without associated dots make predictions with confidences below the set threshold, and green is an outlier with a high confidence score).

[1] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[3] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014.

[4] Jiongxin Liu and Peter N Belhumeur. Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2520–2527. IEEE, 2013.

[5] Jiongxin Liu, Yinxiao Li, and Peter N Belhumeur. Part-pair representation for part localization. In *Computer Vision–ECCV 2014*, pages 456–471. Springer, 2014.

[6] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[7] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014.

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

[9] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision–ECCV 2014*, pages 834–849. Springer, 2014.

Gesture and Action Recognition by Evolved Dynamic Subgestures

Víctor Ponce-López [1,2,3]
http://victorponce.org
Hugo Jair Escalante [4]
http://hugojair.org/
Sergio Escalera [2,3]
http://www.maia.ub.es/~sergio/
Xavier Baró [1,2,3]
https://xbaro.wordpress.com/

[1] IN3, Open University of Catalonia
Rambla Poblenou, 156, 08018 Barcelona, Spain
[2] Dept. MAiA, University of Barcelona
Gran Via de Les Corts Catalanes, 585, 08007, Barcelona, Spain
[3] Computer Vision Center
Building O, Campus UAB, 08193, Bellaterra, Spain
[4] Computational Sciences Dept. INAOE
Luis Enrique Erro, 1, 72840, Tonantzintla, Puebla, Mexico

Gesture and human *action* recognition are two widely studied topics in computer vision and machine learning. In this work we perform gesture-action recognition base on the evolution of *temporal* gesture *primitives*, or *subgestures*. It is inspired on the principle of producing *genetic* variations within a population of gesture subsequences, with the goal of obtaining a set of gesture units that enhance the generalization capability of standard gesture recognition approaches. The underlying assumption is that whole gestures are composed by primitives (that can be shared or not among gestures from different categories), and the hypothesis is that learning with primitives leads to better recognition performance.

Very recently, evolutionary algorithms have been also developed for keyframe extraction [1, 2]. In these works, a bag-of-key-poses representation is adopted and an evolutionary algorithm was used to select the number of key-poses for the vocabulary (using $k-$means for clustering), the training set, features and parameters of the model (using DTW for recognition). These methods look for a subset of frames, whereas in subgesture modeling we aim at learning spatio-temporal units (subgestures) [3, 4, 5, 6]. On the other hand, many works of the literature assume and demonstrate that class-specific key poses/subgestures give a good performance. Nevertheless, we include the fact that some classes may contain or share similar subgestures [5]. Under this additional assumption, our method also reaches the state of the art performance and provides considerable improvements in gesture and action recognition domains.

In this work, a genetic algorithm is used to evolve gesture primitives integrated into a gesture recognition framework coupled with either DTW or HMMs. Different from most of the reviewed work, our approach obtains dynamic subgestures (*i.e.*, sequences of frames of different lengths) and simultaneously learns the parameters of the recognition model (either DTW or HMM). To consider an appropriate measure as DTW so as to treat temporal deformations and obtain subgestures, we implement a temporal clustering algorithm based on an extension of k-means that incorporates the temporal dimension both to represent and to cluster samples on the space and time. Then, we design each class model by means of representing each class sequence-samples in terms of subgestures, as illustrated in Figure 1. For the evaluation, we compute the mean score of classifying each sequence given the learned model parameters either for the DTW or HMM approaches.

In few generations, the proposed subgesture-based representation of actions and gestures achieves the state of the art performance on the MSR-Daily3D and MSRAction3D data sets, and outperforms previous methods for the different validation settings.

[1] A.A. Chaaraoui and F. Florez-Revuelta. Adaptive human action recognition with an evolving bag of key poses. *IEEE Transactions on Autonomous Mental Development*, 6(2):139–152, 2014.

[2] H.J. Escalante, J. Martinez, S. Escalera, V. Ponce-López, and X. Baró. Improving the bag of visual words with genetic programming. In *IJCNN*, 2015.

[3] K. Li, J. Hu, and Y. Fu. Modeling complex temporal composition of actionlets for activity prediction. In *ECCV*, volume 7572, pages 286–299, 2012.

[4] M. R. Malgireddy, I. Nwogu, S. Ghosh, and V. Govindaraju. A shared parameter model for gesture and sub-gesture analysis. In *Combinatorial Image Analysis*, volume 6636, pages 483–493, 2011.

[5] V. Ponce, M. Gorga, X. Baro, and S. Escalera. Human behavior analysis from video data using bag-of-gestures. In *IJCAI*, 2011.

[6] L. Wang, Y. Qiao, , and X. Tang. Video action detection with relational dynamic-poselets. In *ECCV*, 2014.

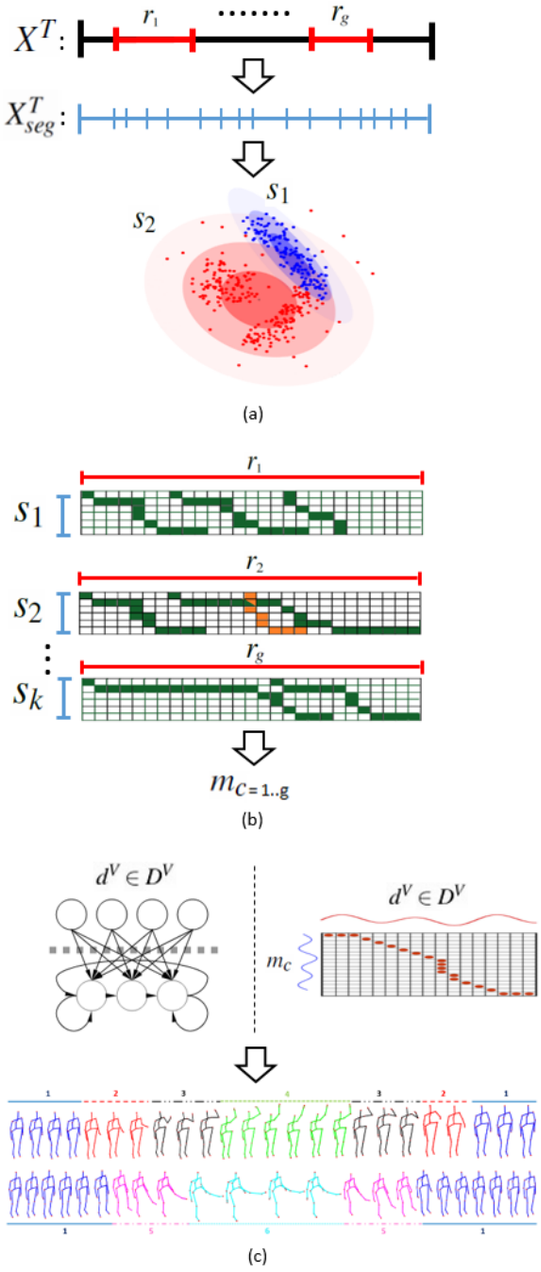


Figure 1: General diagram of the subgesture framework within the evolutionary procedure. In image (a), first we obtain the representants for each class from the whole training data X^T, and besides X^T is split in random segments to obtain subgestures by means of the aligned temporal clustering method. Image (b) shows the backward loop algorithm used both to design each class model $mc \in M$ for the DTW version, where M is the set of g class models $C = \{c_1, c_2, ..., c_g\}$, and to discretize the input sequences for the HMM version. Image (c) shows the recognition of an action/gesture test sequence given the models trained either with HMM or DTW, so that it can be represented as a set of subgesture models (best seen in color).

Face alignment assisted by Head Pose Estimation

Heng Yang[1]
heng.yang@cl.cam.ac.uk
Wenxuan Mou[2]
w.mou@qmul.ac.uk
Yichi Zhang[3]
yichizhang@fas.harvard.edu
Ioannis Patras[2]
i.patras@qmul.ac.uk
Hatice Gunes[2]
h.gunes@qmul.ac.uk
Peter Robinson[1]
peter.robinson@cl.cam.ac.uk

[1] Computer Laboratory
University of Cambridge
Cambridge, UK

[2] School of EECS
Queen Mary University of London
London, UK

[3] Faculty of Arts & Sciences
Harvard University
Cambridge, MA, US

Both head pose estimation and face alignment have been well studied in recent years given their wide application in human computer interaction, avatar animation, and face recognition/verification. However, even the most sophisticated face alignment methods also show some failures when they are applied on face images collected in the wild. In this paper, we show how face alignment can be improved by explicit head pose estimation. In summary, we make the following contributions:

- We investigate the failure cases of several state of the art face alignment approaches and find that the head pose variation is a common issue across those methods. See Fig. 1.
- Based on the above observation, we propose a ConvNet framework for explicit head pose estimation (Fig. 2). It is able to achieve an accuracy of 4° absolute mean error of head pose estimation for face images acquired in unconstrained environment.
- We propose two initialisation schemes based on reliable head pose estimation. They enable baseline face alignment method (RCPR [1]) perform better and reduce large head pose failures by 50% when using only one initialisation.

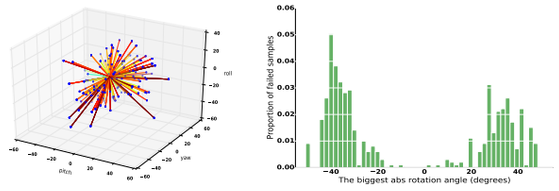


Figure 1: Distribution of the most erroneous samples of various state of the art face alignment methods.

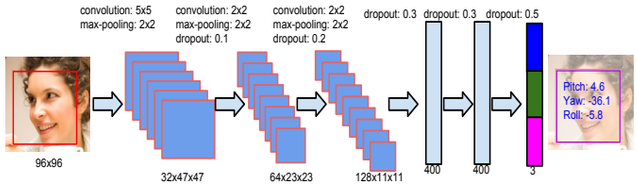


Figure 2: ConvNet model for head pose estimation.

3D face shape based initialisation scheme Given a 3D mean face shape, represented by 68 3D facial landmark locations, as shown in Fig. , we first project this shape under the estimated head pose to a set of canonical 2D locations. More specifically we use constant translation and focus length in order to get a reasonable projection for all images. Then we re-scale the canonical 2D projection by the face bounding box scale of the test image to get the initialisation. We can represent the initialisation process by function $\mathcal{F}$ as follows.

$$S_0 = \mathcal{F}(\theta, bb, \bar{S}^{3D}) \tag{1}$$

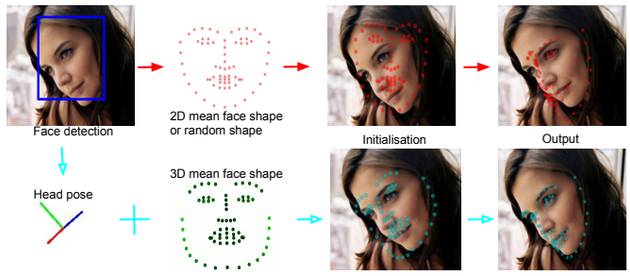


Figure 3: Our proposed head pose based cascaded face alignment procedure (path in cyan color) vs. conventional cascaded face alignment procedure (path in red color).

with bb the face bounding box, $\bar{S}^{3D}$, the 3D mean face shape, θ, the estimated head pose, which can be represented by:

$$\theta = \mathcal{G}(I, bb) \tag{2}$$

where $\mathcal{G}$ is the deep convolutional model.

Results

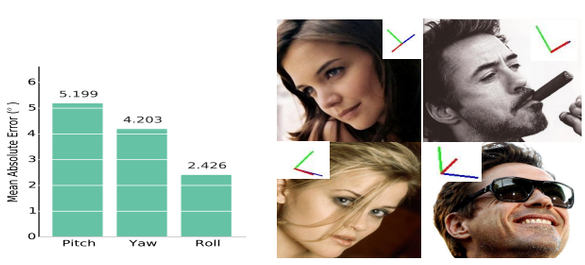


Figure 4: Head pose estimation result. Left, absolute mean error on test set; right, example results of head pose estimation.

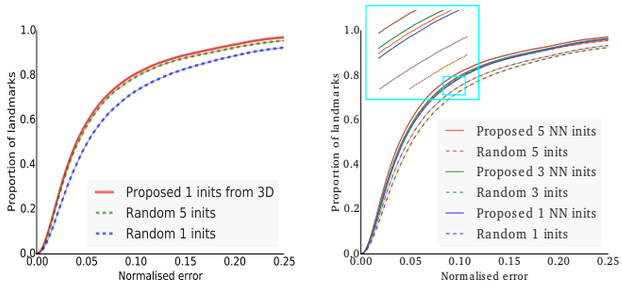


Figure 5: Comparison to baseline method with random initialisation.

[1] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1513–1520, 2013.

Score Normalization in Multimodal Systems using Generalized Extreme Value Distribution

Renu Sharma[1,2]
renu@cdac.in
Sukhendu Das[2]
sdas@iitm.ac.in
Padmaja Joshi[1]
padmaja@cdac.in

[1]Centre for Development of Advanced Computing, Mumbai, India
[2]Indian Institute of Technology, Madras, India

In multimodal biometric systems, human identification is performed by fusing information at different levels like sensor-level, feature-level, score-level, rank-level and decision-level. Score level fusion is preferred over other levels of fusion because of its low complexity and sufficient availability of information for fusion. However, the scores obtained from different unimodal systems are heterogeneous in nature. For instance, one classifier gives a similarity measure while another gives a dissimilarity measure. Some classifiers give scores indicating the probability of the input pattern to be a genuine subject. Further, distributions of scores obtained from the individual matchers need not be on the same numerical scale (variance) and location (mean). Hence there is a need of normalization before score-level fusion.

Jain et *al.* [1] described various score normalization techniques for multimodal systems. Shi et *al.* [2] introduced Extreme Value Theory (EVT) distribution (based on Generalized Pareto Distribution) in unimodal biometric systems for score normalization. They have used a non-parametric method for modelling the significant part of the genuine distribution and a parametric Generalized Pareto Distribution for modelling the tail part of the genuine distribution. Later on, Scheirer et *al.*[3] also proposed an EVT-based adaptive score normalization method (W-Score) using the Weibull distribution. The methods proposed in [2, 3] focused on the modelling the tail of the impostor (or genuine [2]) distribution. We assume that genuine scores form the tail of a complete (genuine and imposter combined) distribution, and hence we analyze the tail of the complete distribution by considering only the genuine scores. Generalized Extreme Value (GEV) distribution is used for modelling the genuine scores, and parameters of GEV are estimated using the maximum likelihood estimation (MLE). Score vectors used for the estimation of parameters (training) are termed as probe score vectors (Figure 1). Score normalization of the test score vector (query) is performed using the cumulative density function (CDF) of the GEV distribution formed with the learned parameters. Figure 1 represents the entire architecture of the proposed method for score normalization.

There are two extreme value analysis based approaches [4]: (i) Block Maxima and (ii) Peak over Threshold. Shi et *al.* [2] method is based on peak over threshold approach. Our proposed method and W-Score [3] technique are based on the block maximum approach for extreme value analysis. W-Score method is an adaptive impostor-centric technique, which uses a single score vector obtained by comparing the input test template (during query or testing) to the enrolled templates. From the score vector, the W-Score method uses only the top impostor scores (excluding the topmost score) to fit a Weibull distribution. In contrast to the W-Score, our method is client-centric, and for modelling the GEV distribution the number of probe samples and their corresponding score vectors are utilized. This process occurs offline. From each score vector, one genuine score and N-1 impostor scores are obtained. A single genuine score is considered as an extreme value in the score vector. A collection of genuine scores form a set of extreme values with respect to the whole set of probe score vectors. If a single probe score vector is considered as a block, genuine scores form a sequence of minimum (or maximum) values. According to the EVT theory [4], minima (or maxima) of sequences is characterized by the GEV distribution, whose CDF is given as:

$$G(x,\mu,\sigma,k) = \begin{cases} exp\left(-\left(1+k\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/k}\right) & if\ k \neq 0 \\ exp\left(-exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right) & if\ k = 0 \end{cases} \quad (1)$$

where, $1 + k\,(x-\mu)/\sigma > 0$ is such that $1 + kx > 0$. μ, σ and k correspond to the mean, standard deviation and shape parameters of the distribution respectively. So, the genuine data from the probe score vectors are modelled by the GEV distribution and the parameter set (mean, scale and location) is computed by the maximum likelihood estimation method. If $S_1, \ldots, S_M$ are the M genuine scores, the log-likelihood function to be maximized, is formulated as:

$$LL(\mu,\sigma,k) = -M\log\sigma - \sum_j \left[1+\frac{k(S_j-\mu)}{\sigma}\right]^{-1/k} - \left(\frac{1}{k}+1\right)\sum_j \log\left[1+\frac{k(S_j-\mu)}{\sigma}\right]. \quad (2)$$

As GEV is a parametric distribution and requires the estimation of only three parameters, few genuine values are sufficient to model the GEV distribution as opposed to the non-parametric techniques which require a much higher number of genuine scores to estimate the distribution reliably. Hence, the parametric techniques have lower computational complexity than the non-parametric techniques.

Given a GEV distribution, estimating the probability that a given score is an outlier is computed from the value of CDF of the GEV distribution. So, the normalization test score vector is computed using the CDF of the GEV distribution, as follows:

$$S'_i = G(S_i,\mu,\sigma,k) \quad (3)$$

where, S'_i is the ith class normalized score. After normalization, scores from the different unimodal systems are fused by using a score-level fusion technique. In identification mode, the user is identified if the enrolled subject corresponds to the top score from the fused score vector. Efficiency of the proposed method is compared (see Table 1) with GPD [2] and W-Score [3] methods, using Identification (IR) and Verification (VR) rates. Results are shown over four multimodal biometric datasets: Test Sets I and II are obtained from NIST-BSSR1; while Test Sets III and IV are formed (chimeric) using samples from NIST-BSSR1 and 'FRGC v2.0 + LG4000' datasets respectively.

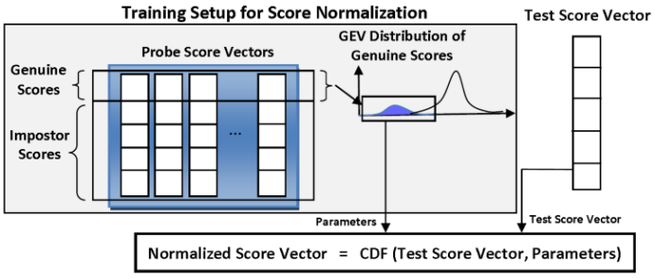


Figure 1: Architecture of the proposed method for score normalization.

	Test Set I		Test Set II		Test Set III		Test Set IV	
Methods	VR	IR	VR	IR	VR	IR	VR	IR
GPD[2]	94.71	81.11	95.74	89.55	94.86	82.50	98.91	97.49
W-Score[3]	96.75	81.75	99.80	98.06	98.23	87.03	99.35	92.73
Our Method	**96.89**	**85.11**	**99.99**	**100**	**99.86**	**99.43**	**99.59**	**98.00**

Table 1: Identification and verification rate of Test Sets I, II, III, and IV.

[1] A. Jain, K. Nandakumar, A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270-2285, 2005.

[2] Z. Shi, F. Kiefer, J. Schneider, and V. Govindaraju. Modeling Biometric Systems Using the General Pareto Distribution (GPD). In *SPIE*, 6944, 2008.

[3] W. J. Scheirer, A. Rocha, R.J. Micheals, and T. E. Boult. Meta-Recognition: The Theory and Practice of Recognition Score Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1689-1695, 2011.

[4] S. Kotz and S. Nadarajah. Extreme Value Distributions: Theory and Applications. 1 Edn. *World Scientific Publishing Co*., 2001.

Manifold-Regularized Selectable Factor Extraction for Semi-supervised Image Classification

Xin Shi[1]
shixinnk@gmail.com
Chao Zhang[1]
chzhang@cis.pku.edu.cn
Fangyun Wei[1]
weifangyun@pku.edu.cn
Hongyang Zhang[1]
hy_zh@pku.edu.cn
Yiyuan She[2]
yshe@stat.fsu.edu

[1] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, PR China
[2] Department of Statistics, Florida State University, USA

Feature selection methods are efficient in modern computer vision applications to reduce the computational cost and the chance of over-fitting. Recently, a novel selectable factor extraction (SFE[3]) framework is proposed to simultaneously perform feature selection and extraction, and is theoretically and practically proved to be effective for high-dimensional data. Although it is advantageous in several aspects, SFE is only designed for either supervised or unsupervised learning, and is not suitable when there are limited labeled samples and a large number of unlabeled samples. To tackle this problem, we propose a novel manifold regularized SFE (MRSFE) framework for semi-supervised image classification.

We use a low rank penalized regression model to explore the label information. A low rank matrix of the regression coefficients, together with the $\ell_{2,1}$ or $\ell_{2,0}$ norm penalty is learned for joint feature selection and extraction. In addition, all the labeled and unlabeled samples are utilized in MRSFE to construct the data adjacency graph to approximate the underlying data manifold, which the data distribution is assumed to be supported on. The graph Laplacian is then incorporated as a regularization term to smooth the coefficients matrix. In this way, the local structures of the whole dataset are preserved, and the data distribution is well exploited.

To derive our model, we begin with the reduced rank regression (RRR) model[1]:

$$\min_B \|Y_L - X_L B\|_F^2, \quad \text{s.t. } \operatorname{rank}(B) \le r. \tag{1}$$

where $B = [b_1, b_2, \cdots, b_d]^T \in \mathbb{R}^{d\times c}$ is the representation matrix, $X_L \in \mathbb{R}^{l\times d}$ is the design matrix of the labeled samples, l is the number of the labeled samples and d is the number of the features, $Y_L \in \mathbb{R}^{l\times c}$ is the response matrix where c is the number of the classes, i.e., $Y_{ij} = 1$ if and only if the ith sample belongs to the jth class.

Although successful in practice, the plain RRR has however certain drawbacks — the model (1) typically involves all input features of X. To this end, we implement feature selection for RRR,

$$\min_B \|Y_L - X_L B\|_F^2 + \alpha\|B\|_{2,1}, \quad \text{s.t. } \operatorname{rank}(B) \le r. \tag{2}$$

where $\|B\|_{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^c B_{ij}^2}$, which promotes row sparsity of B. Note that we select features from X_L according to the non-zero rows of the representation matrix B, i.e., if the i-th row of B is non-zero, we conclude that the ith feature of X_L(i-th column) is significant. We call (2) the selectable factor extraction (SFE) method.

In many applications where we have only a small number of labeled samples, the representation matrix B learned from model (2) is often unreliable. We use the manifold regularization (MR[2]) to help learn the manifold structure of large number of unlabeled samples. Combining model (2) with MR, we have the following optimization problem:

$$\min_B \|Y_L - X_L B\|_F^2 + \alpha\|B\|_{2,1} + \beta\operatorname{tr}(B^T X_{LU}^T L X_{LU} B), \text{s.t. } \operatorname{rank}(B) \le r. \tag{3}$$

where L is the Laplacian matrix, X_{LU} is the data matrix containing both the labeled and unlabeled samples.

To deal with the row sparsity and the rank constraint on B, we write $B = SV^T$, where $S \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{c\times r}$ is an orthogonal matrix.Thus model (3) is equivalent to

$$\min_{S,V^TV=I_r} \|Y_L - X_L S V^T\|_F^2 + \alpha\|S\|_{2,1} + \beta\operatorname{tr}(VS^T X_{LU}^T L X_{LU} S V^T). \tag{4}$$

The full-rank factorization of B enables us to tackle the sparsity regularization and the low-rank constraint separately. In other words, S selects

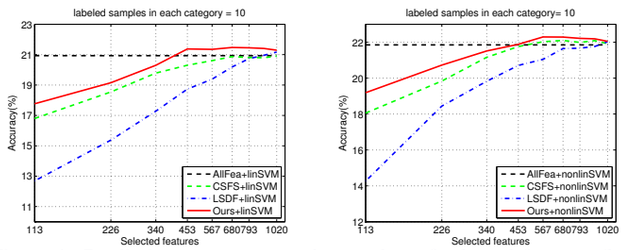


Figure 1: Prediction accuracy vs. the number of selected features (left) using linear SVM (right) using nonlinear SVM

significant features from X_L, while the orthogonal matrix V determines the subspace after dimension reduction. So we can implement feature selection and extraction simultaneously.

Considering the ℓ_1 penalty cannot handle the collinearity and may lead to inconsistent and biased estimation([4]), similar to $\ell_{2,1}$, we advocate to use non-convex constraint such as $\ell_{2,0}$ instead of widely-used $\ell_{2,1}$ penalty for S in (4)

$$\min_{S,V^TV=I_r} \|Y_L - X_L S V^T\|_F^2 + \beta\operatorname{tr}(VS^T X_{LU}^T L X_{LU} S V^T), \text{s.t. } \|S\|_{2,0} \le q_s \tag{5}$$

where q_s is a parameter to control the number of selected features. Using the constraint form instead of the penalty is intuitive, because we can directly control the number of features we need. Although the $\ell_{2,0}$ penalty is nonconvex and hard to optimization in classical methods, it is doable in our algorithm. Once S is obtained from (4) or (5), we can select the significant features according to top-k index of the row-norms in descending order or nonzero rows of S. We call both of the models (4) and (5) manifold-regularized semi-supervised selectable factor extraction method(MRSFE).

We use alternating optimization method to solve model (4) and (5), an efficient and easy-to-implement algorithm is designed to find the solutions. Our algorithm only consists of SVD decomposition of small-scale W, together with some thresholding operations which are at low cost. We evaluate the effectiveness of our MRSFE by applying it to a challenge web image dataset, NUS-WIDE-OBJECT. Experiments on a this dataset demonstrate the superiority of the proposed method.

[1] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *JMA*, 5(2):248–264, 1975.

[2] Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE TIP*, 22(2):523–536, 2013.

[3] Yiyuan She. Selectable factor extraction in high dimensions. *arXiv preprint arXiv:1403.6212*, 2014.

[4] Peng Zhao and Bin Yu. On model selection consistency of lasso. *JMLR*, 7:2541–2563, 2006.

Leaf Segmentation under Loosely Controlled Conditions

Simone Buoncompagni[1]
simone.buoncompagni2@unibo.it
Dario Maio[1]
dario.maio@unibo.it
Vincent Lepetit[2]
lepetit@icg.tugraz.at

[1] Department of Computer Science and Engineering
University of Bologna
Bologna, Italy

[2] Institute for Computer Graphics and Vision
Graz University of Technology
Graz, Austria

Extracting accurately the shape of a leaf is a crucial step in image-based plant identification systems. The partial or total absence of textures on leaf surface and the high color variability of leaves belonging to same species make shape as the main recognition element. For such reason, leaf segmentation plays a decisive role in the leaf recognition process.

Even though many general segmentation methods have been proposed in the last decades, leaf segmentation presents specific challenges. In particular, a pixel-level precision is required in order to highlight fine scale boundary structures and discriminate similar global shapes. Moreover, even if the input image can typically be taken in controlled conditions, where the leaf is the only visible object over a white background, the user taking the picture is not necessarily an expert and the conditions are often not ideal: the leaf exhibits specular reflections, casts shadows, the background is never exactly white and is usually non-uniform, and the image can be blurry.

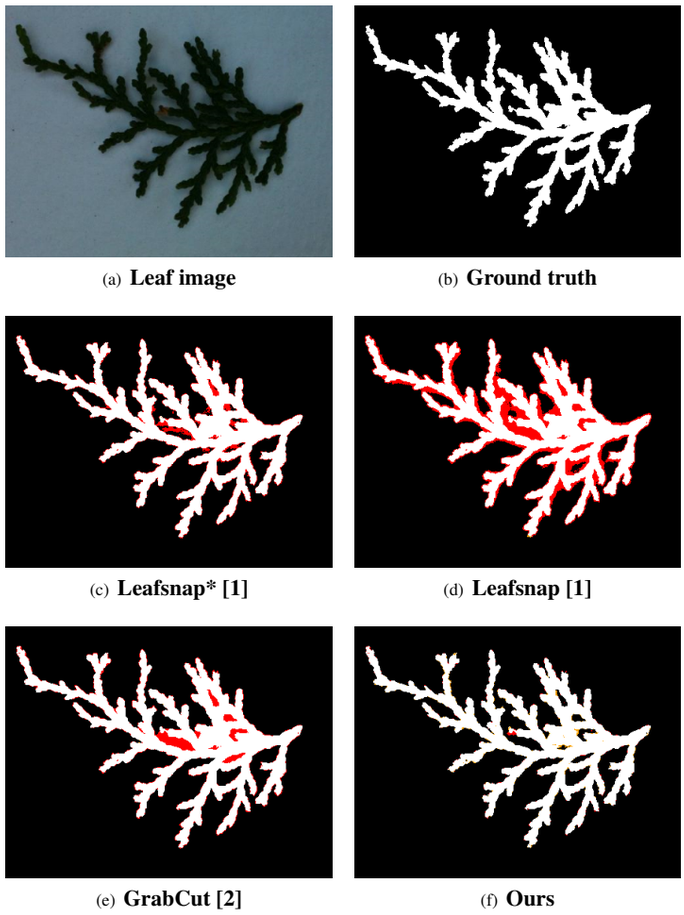
Recently, a solution for the problem at hand has been proposed in [1] where leaf segmentation is carried out by estimating the probability distribution of foreground and background pixels. However, several drawbacks appear in this formulation due to challenging leaves like pine needles, false positives detection related to shadows and false negatives detection related to specularities. Prior distributions and post-processing operations are employed to tackle such problem, with the risk of hurting the final leaf shape.

In this paper we introduce a new solution by training a pixel-wise classifier [3] that learns filter responses associated to background and foreground regions in images of leaves. Our classifier is trained by selecting positive (leaf) and negative (non-leaf) feature samples that lie on the neighboorhod of the leaf boundary thus focusing learning only on "sensitive" pixels. Such classifier is then applied to each pixel location of a given unknown test image. This provides a score map that we then threshold using two different thresholds to detect pixels that belong to foreground and background with a high level of probability. With these pixels at hand we initialize an EM algorithm with a good initial estimate of foreground and background cluster parameters in the saturation-value color space, differently from [1] which has to initialize the EM segmentation with the same values for all the images. The other difference with [1] is that we can consider as unlabeled data only the pixels that are in the neighborhood of the detected leaf boundary. This allows to keep focusing on segmenting correctly the pixels around the leaf boundary, and in practice it is enough to get a good segmentation of the other pixels, which are easier to classify.

For evaluation we use the Leafsnap *Field* image dataset publicly available online [1] where different leaves of different species are acquired against solid background and variable light conditions thus simulating typical images that a user could provide for plant recognition.

To train our pixel-wise classifier we randomly select one image for each species and we manually produce segmentation and thicker contours to discriminate between positive and negative training samples placed in the neighborhood of boundary. Since segmentation ground truth is not available and its manual production for thousands of images would require an inestimable amount of time, we considered a subset of the original *Field* dataset. Our testing set is made of 300 images: 150 images for which the EM approach of [1] performs already well thus producing faithful segmentation in accordance with the leaf shape plus 150 more challenging images for which EM partially or totally fails.

The general behavior of different methods can be qualitatively appreciated looking at Fig. 1 where results returned by Leafsnap, Leafsnap without post-processing (marked with *), GrabCut and our method are reported. As the reader can see comparing ground truth details with real segmentations, it is confirmed that post-processing hurts quality of boundaries and should be avoided. On the other hand, GrabCut tends to return round contours. With our method some errors still remain, due to those background pixels that look strongly similar to leaf and vice-versa. However, our method represents a good trade-off since we do not use post-processing but at the same time we assure a good robustness to false positives. Furthermore, our contours are much closer to the ground truth.



(a) **Leaf image** (b) **Ground truth**

(c) **Leafsnap* [1]** (d) **Leafsnap [1]**

(e) **GrabCut [2]** (f) **Ours**

Figure 1: Leaves segmentation under loosely controlled conditions with different methods. False positives in red, false negatives in orange (best viewed in color).

[1] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In *ECCV*, pages 502–516. Springer, 2012.

[2] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts. *TOG*, 23(3):309–314, 2004.

[3] A. Sironi, V. Lepetit, and P. Fua. Multiscale Centerline Detection by Learning a Scale-Space Distance Transform. In *CVPR*, pages 2697–2704. 2014.

Shape from Focus with Adaptive Focus Measure and High Order Derivatives

Yuval Frommer[1]
yuvalfrommer@yahoo.com
Rami Ben-Ari[2]
ramib@il.ibm.com
Nahum Kiryati[1]
nahum.kiryati@gmail.com

[1] School of Electrical Engineering
Tel-Aviv University
Tel-Aviv, Israel

[2] IBM Research
Haifa, Israel

Shape From Focus (SFF) methods frequently use a single focus measure to obtain a depth map. Common focus measures are fixed and spatially invariant. In this paper we present a framework to create an adaptive focus measure based on ensemble of basis focus operators defined by the image derivatives in various orders and directions. The method adaptively chooses the most informative combination of derivatives at each pixel by weighting the contribution of each derivative accordingly. Weighting is achieved using the curve standard deviation (CSTD) which measures how peaked the distribution of focus measures is. The CTSD characteristic is used both during focus measure calculation and aggregation. This approach effectively generalizes some of the existing measures *e.g.* ML [2]. We assess the performance of our new approach by extensive experiments showing a significant increase in accuracy compared to state-of-the-art focus measures and with comparison to other methods which perform aggregation *e.g.* [1] and more.

Agglomeration of the focus measures in an image stack $I_z(x,y)$ builds up the focus volume $\phi : (X \times Y) \times Z \mapsto \mathbb{R}$, assigning a focus measure to each point in the image stack. We now define a linear space with the basis functions as the image derivatives, in different orders and directions:

$$\phi_{ij} = |\frac{\partial^i I}{\partial \mathbf{x}_j^i}| \tag{1}$$

where i, j index the the derivative order and the discrete directions respectively. This paper suggests a focus measure based on a linear combination of image derivatives represented by a vector in the focus-measure space:

$$\psi(x,y,z) = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_{ij}(x,y)\ \hat{\phi}_{ij}(x,y,z) \tag{2}$$

where, α_{ij} denote the coefficients of the basis focus components. Note that existing image based derivative focus measures *e.g.* ML [2] become a particular case in this framework. We call the new focus measure ψ as the Adaptive High Order (AHO) focus measure. A focus curve should ideally be a symmetric uni-modal function with a narrow peak. To this end, we introduce a reliability measure for a given focus curve as curves' standard deviation (CSTD). The coefficients $\alpha_{ij}(x,y)$ are then set locally according to CSTD.

Once the focus volume in (2) is computed, a standard procedure includes an aggregation process. Using the newly introduced CSTD we suggest a novel iterative aggregation scheme:

$$\tilde{\psi}_t(x,y,z) = \sum_{(x',y')\in W} \hat{w}(x',y')\ \psi_{t-1}(x+x', y+y', z) \tag{3}$$

where W denotes the aggregation window and $\hat{w}$ represents the associated weights set according to CSTD measure. Finally the depth at each point is computed as the focus curve *centroid* introducing a non-parametric cue for depth estimation.

To evaluate the new focus measure we conduct a quantitative comparison in terms of depth map RMSE (Root Mean Square Error) and report the results in Table 1. Notably the AHO achieves results that are more than twice as accurate as the best focus measures in the literature. Fig. 1 presents a qualitative comparison with the state-of-the-art focus measures in the literature. The results show the significant improvement obtained by our AHO focus measure, well coping with different textures as well as intensity and depth edges. Next, we present in Fig. 2 the performance of our complete pipeline for shape reconstruction using the proposed focus measure, aggregation and reconstruction approach. Note that the Antlion and the Spider depth maps at the third and fourth columns are driven from image stacks captured in the wild with a SLR camera.

Object	GLV	ML [2]	WAV [3]	Ours
Cloth	2.6	1.42	1.78	**0.76**
Synthetic-Cone	7.05	3.73	1.05	**0.33**
Middlebury-Cones	3.42	1.56	2.49	**1.17**
Cube	5.18	3.63	1.67	**0.76**
Average	4.56	2.59	1.75	**0.76**

Table 1: Comparison of focus measure operators in terms of depth map RMSE. Best results are in bold.

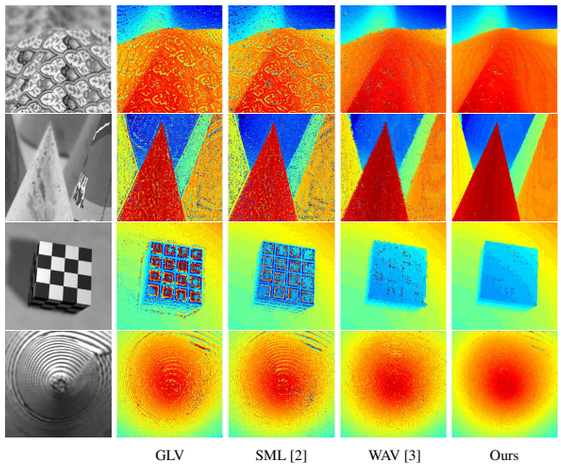


Figure 1: Comparison of focus measures. Note the significant improvement obtained using the proposed focus measure.

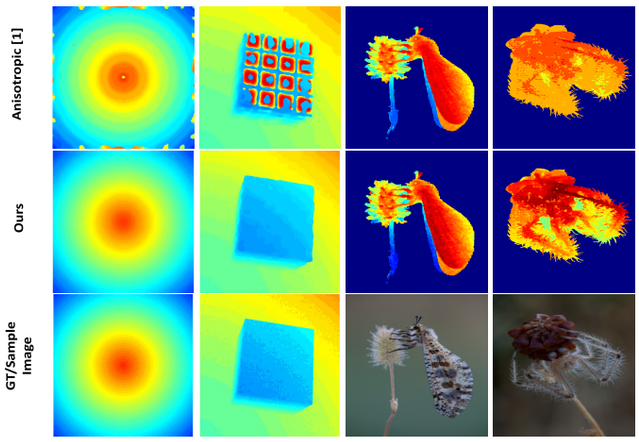


Figure 2: Comparison with a full pipeline method.

[1] M. T. Mahmood, S. O. Shim, and T. S. Choi. Depth and image focus enhancement for digital cameras. *International Symposium on Consumer Electronics*, pages 50–53, 2011.

[2] S. K. Nayar and Y. Nakagawa. Shape from focus. *TPAMI*, 16(8): 824–831, 1994.

[3] G. Yang and B. J. Nelson. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In *Int. Conference on Intelligent Robots and Systems*, volume 3, pages 2143–2148, 2003.

A Segmentation Technique for Flexible Pipes in Deep Underwater Environments

Saulo Pessoa[1]
sap@cin.ufpe.br
Vinicius Cesar[1]
vmc@cin.ufpe.br
Bernardo Reis[1]
bfrs@cin.ufpe.br
Judith Kelner[1]
jk@cin.ufpe.br
Ismael Santos[2]
ismaelh@petrobras.com.br

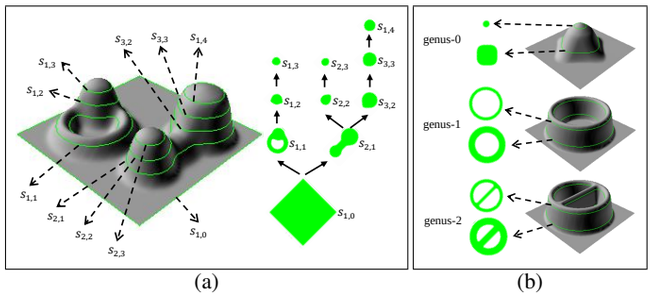
[1] Informatics Center
Federal University of Pernambuco
Recife, Brazil
[2] CENPES
Petrobras,
Rio de Janeiro, Brazil

This paper presents a novel segmentation technique for flexible pipes in deep underwater environments. It consists in finding regions of the pipe denominated *vertebrae* and has three stages. Firstly, the input image is pre-processed to reduce noise. Secondly, the pre-processed image is binarized by the proposed Multi-level Topological Binarization technique, resulting in an image which highlights (in white) the regions that are potential vertebrae (the *blobs*). Lastly, the pipe is segmented by finding the best sequence of blobs that fulfills a set of restrictions inherent to a pipe.

For improving the technique robustness, an alternating pattern of white (a vertebra) and black regions is marked on the pipe. Results show that the proposed technique can segment pipes even under harsh conditions such as: low contrast between pipe and background; uneven illumination of the pipe; and high level of noise due to floating particles. In particular, the proposed binarization technique achieves valuable results without any parameter, whereas state-of-the-art techniques did not achieve the same quality even after their parameters were fine-tuned for each condition.

Two blurring filters are used for pre-processing input image: a bilateral filter [5] removes small particles floating around, while preserving the vertebrae's edges; and a Gaussian filter uniformly smooths the images.

The proposed binarization technique finds peaks in the pre-processed image. This is done by "slicing" the image into a number of slices $s_{i,b}$ (i is its index and b is its level). A tree of slices is then constructed by hierarchically connecting nodes of consecutive levels (Figure 1a). Intuitively speaking, peaks are prominences similar to the first one exhibited in Figure 1b. More formally, a peak is a sequence of nodes denoted by $P = \{n_k\}_{k=1}^{m} = n_1, n_2, \ldots, n_m$, such that: n_m is a leaf node of the tree; n_k, such that $1 \leq k < m$, is the parent node of n_{k+1}; n_1 is the only element that has a brother node, or n_1 is the only element that has a parent with genus greater than zero, or n_1 is the root node. Nodes n_1 and n_m are respectively named the base and the top of a peak. Given a tree of slices, one has to traverse it from leaves to root in order to find its peaks. For example, Figure 1a contains three peaks: the first one is $P' = s_{1,2}, s_{1,3}$; the second is $P'' = s_{2,2}, s_{2,3}$; and the third is $P''' = s_{3,2}, s_{3,3}, s_{1,4}$.



(a) (b)

Figure 1: Surface plot of an image and the respective tree of slices in (a). Topological classification of slices obtained from different surfaces in (b).

The binarization stage produces a set of blobs $B = \{b_i : i \in \{1, \ldots, n\}\}$ (the white regions depicted in Figure 2). The blobs due to the pipe vertebrae are selected by a backtracking algorithm. It searches for the longest chain of blobs that respects restrictions inherent to a pipe. This chain is denoted by a sequence $V = \{j_k\}_{k=1}^{m} = j_1, j_2, \ldots, j_m$, where j_k is the index of a blob b_i. During the search, b_i is the current blob being tested to be inserted into V; b_l is the last blob inserted into V; and b_p is the blob inserted immediately before b_l.

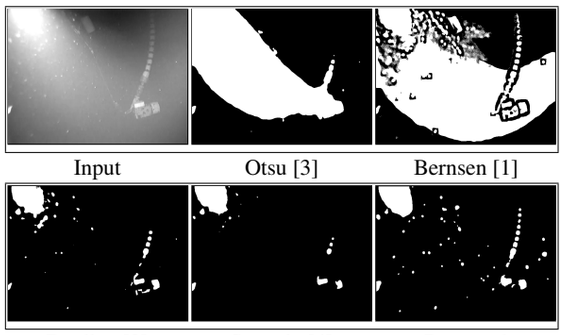


Figure 2: Input image depicting uneven contrast and noise. Results of Otsu, Bernsen, Bradley, and Sauvola binarization algorithms after parameters were fine-tuned. The proposed binarization technique extracts most of the pipe vertebrae without requiring any parameterization.

The first restriction limits the distance between the centroids of b_i and b_l, respectively c_i and c_l:

$$d(c_i, c_l) \leq d_{max}, \tag{1}$$

where d_{max} is the maximum acceptable distance. The second restriction limits the length of the gap between b_i and b_l. This gap is denoted by the line segment $\overline{o_i o_l}$, where o_i is the closest intersection point between the line segment $\overline{c_i c_l}$ and the contour of b_i (o_l is similarly defined):

$$d(o_i, o_l) \leq g_{max}, \tag{2}$$

where g_{max} is the maximum acceptable gap length. The third restriction limits the distance $d(o_i, c_i)$, which can be understood as a radius of b_i. This distance has an upper and lower limit:

$$\left(\frac{1}{1+v_{max}}\right) \times d(o_l, c_l) \leq d(o_i, c_i) \leq (1+v_{max}) \times d(o_l, c_l), \tag{3}$$

where $v_{max} \geq 0$ is a user parameter. The last restriction limits the exterior angle between $\overline{c_p c_l}$ and $\overline{c_l c_i}$, which is denoted by θ_i. Since three points are required to form an angle, this restriction only applies when there are at least two blobs already inserted into V:

$$\theta_i \leq \theta_{max}, \tag{4}$$

where θ_{max} is a parameter defined by consulting the pipe's specification.

[1] John Bernsen. Dynamic thresholding of grey-level images. In *International conference on pattern recognition*, pages 1251–1255, 1986.

[2] Derek Bradley and Gerhard Roth. Adaptive thresholding using the integral image. *Journal of graphics, gpu, and game tools*, 12(2):13–21, 2007.

[3] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[4] Jaakko Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern recognition*, 33(2):225–236, 2000.

[5] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998.

Structural Symmetries from Motion for Scene Reconstruction and Understanding

Natesh Srinivasan
http://www.cc.gatech.edu/grads/n/nsriniva/
Luca Carlone
http://www.lucacarlone.com/
Frank Dellaert
http://www.cc.gatech.edu/~dellaert/FrankDellaert/Frank_Dellaert/Frank_Dellaert.html/

Robotics and Intelligent Machines Lab,
College of Computing,
Georgia Institute of Technology

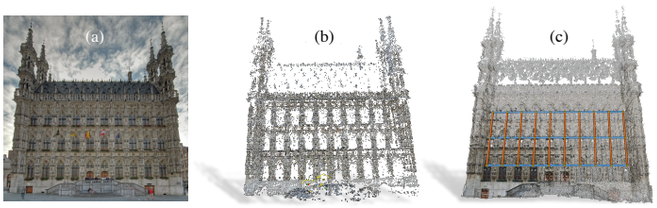


Figure 1: The proposed approach uses (a) a set of 2D images, and (b) a sparse 3D reconstruction to detect repetitive patterns (regular structure) and exploits them for SfM. (c) Generative model of the regular structure obtained from the proposed approach, for the Leuven dataset.

The identification and description of partial symmetries in man-made structures is a powerful tool to improve the quality of 3D reconstruction from unordered images and to enable high-level understanding of scene geometry. In this work we propose an approach to identify symmetries and exploit them in Structure from Motion (SfM). Our first contribution is a symmetry detection approach that uses the 3D geometry of the scene as well as 2D appearance cues. We show that a particular parametrization of the *transformation space* (space in which each point represents a candidate symmetry relation) exposes the dominant symmetries in the scene. Then, we use appearance information to prune incorrect symmetry hypotheses. The second contribution is a *constrained bundle adjustment* (CBA) scheme that jointly optimizes for the best 3D reconstruction and the symmetry *generators*. In contrast to related work on CBA, our approach models n-fold (rotational and translational) repetitions of architectural elements, and allows estimating a *generative model* of the 3D geometry. Experimental results confirm that our method can correctly identify and exploit partial symmetries in noisy and sparse SfM datasets.

We propose an approach for 2D-3D symmetry detection and we show how to leverage the presence of repetitive structure to improve SfM reconstruction. Our approach includes three building blocks. The first is a multi-hypotheses estimator for 3D symmetry generators. We borrow key insights from [4], which shows how to map putative symmetry transformations into 1D or 2D lattices. However, we skip grid fitting (which is unreliable on SfM data), and we show that a polar parametrization of the transformation space clearly exposes dominant symmetries. The second block prunes the multiple hypotheses on the generators and returns the generators that are most consistent with 2D appearance. This is similar in spirit to [3], while we avoid 3D surface fitting. Finally, the last block takes the estimate for the generators and jointly refines this estimate and the 3D reconstruction. We present an on-manifold optimization scheme that respects the regular structure and the manifold nature of the variables involved in the problem (rotations, poses). Contrarily to standard CBA we explicitly model n-fold repetitions, we avoid priors [2] and user intervention [1], and we also output a *generative model* of the scene.

We first detect 3D features on the sparse point cloud. While related work obtains these features by uniform sampling [4], we use 3D SIFT keypoint. For each feature, we associate a set of neighboring points which are within a ball of predetermined radius. Then, for each pair of features, we compute a relative transformation by applying ICP to local patches around the features. This gives a set of putative, highly noisy set of transformations similar to [4]. Our goal is to obtain a good guess for the parameters of the underlying symmetry(a.k.a generators) from these noisy transformations.

The key insight for obtaining a good guess for the generators is that the presence of the repetitive structure in the 3D model implies that many of the transformations will be (approximately) in the form $x_{ij} = n \cdot g_k$ ($k \in \{1,2\}$) i.e., will be produced by an n-fold repetition along the generator $g_k \in \mathbb{R}^2$ (this is also one of the motivations for the grid fitting of [4]).

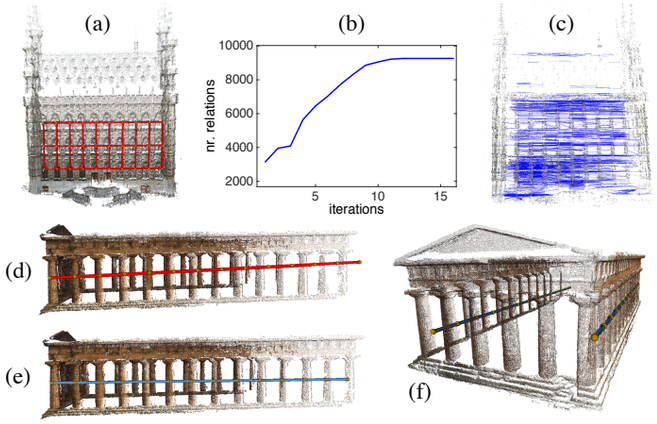


Figure 2: Leuven dataset: (a) initial estimate for the generators (before CBA); (b) number of symmetry relations discovered during the CBA iterations; (c) symmetry relations shown as lines connecting points considered n-fold repetitions of each other. Neptune dataset: (d) initial estimate of the generators, (e,f): refined generator estimate from CBA.

Moreover, the 2-vector g_k can be written as $g_k = \delta_k \cdot u_k$, where $\delta_k \in \mathbb{R}$ is the *repetition period* and the unit vector $u_k \in \mathbb{R}^2$ ($\|u_k\| = 1$) is the *repetition direction*. Therefore, all the putatives generated by g_k become $x_{ij} = (n\delta_k) \cdot u_k$, i.e., they share the same direction, while haivng different norms depending on the unknown number of repetitions n). Polar coordinates naturally adapt well to recover the underlying symmetries for the 2 parameters family of generators. We further refine the guess of the generator by using a appreerence based verification step. The intuition is that a good generator is one that maps a patch to another one having similar appearance (this is essentially the concept of regular structure). Since each point in the 3D model is obtained via SfM, it has a corresponding feature descriptor. We use this insight to devise a generator ranking scheme.

We refine the generators that is obtained in the previous step using a bundle adjustment scheme that implicitly encode n-fold repetitions in them. Contrary to existing approaches [2], that consider only translational symmetries or 1-fold repetitions, we model n-fold rotational and translational repetitions. We jointly optimize for the value of the generators and also refine the sparse point cloud and the assosiated camera poses similar to the method described in [1]. We distinguish ourselves from previous work here by dealing with all kinds of 2-parameter transforms descibed in [4] but applied to noisy point clouds reconstructed from SfM.

[1] D. Ceylan, N.J. Mitra, Y. Zheng, and M. Pauly. Coupled structure-from-motion and 3D symmetry detection for urban facades. *ACM Transactions on Graphics*, 33(2), January 2014.

[2] A. Cohen, C. Zach, S. Sinha, and M. Pollefeys. Discovering and exploiting 3d symmetries in structure from motion. In *CVPR*. Computer Vision and Patter Recognition, June 2012. URL `http://research.microsoft.com/apps/pubs/default.aspx?id=163609`.

[3] N. Jiang, P. Tan, and L.-F. Cheong. Multi-view repetitive structure detection. In *ICCV*, pages 535–542, 2011.

[4] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. Guibas. Discovering structural regularity in 3D geometry. *ACM Transactions on Graphics*, 27(3):#43, 1–11, 2008.

Stereo Tracking and 3D Reconstruction of Underwater Pipes

Vinicius Cesar[1]
vmc@cin.ufpe.br
Bernardo Reis[1]
bfrs@cin.ufpe.br
Saulo Pessoa[1]
sap@cin.ufpe.br
Judith Kelner[1]
jk@cin.ufpe.br
Ismael Santos[2]
ismaelh@petrobras.com.br

[1] Informatics Center
Federal University of Pernambuco
Recife, Brazil
[2] CENPES
Petrobras,
Rio de Janeiro, Brazil

While manipulating flexible oil pipes in deep underwater environments, it is desired to monitor the pipe geometry in order to avoid damaging the equipment due to excessive tension. In order to achieve this goal, we propose an algorithm for tracking and reconstructing the pipe medial axis (PMA) using stereo cameras. The underwater scenario involves monochrome and low-resolution images, and a moderate amount of noise from floating particles and fish, such as presented in Figure 1. The pipe is marked with an alternating pattern of contrasting colors, and is compelled to configure a plane curve. These characteristics are used by the algorithm to achieve precise and robust results.

Given the bright marks on the pipe, the tracking algorithm attempts to find a bi-dimensional curved region $\mathcal{R}$ representing the pipe, by maximizing the intensity of the pipe projection in the image I. $\mathcal{R}$ is defined by the function $\mathbf{R}(s,t,\mathbf{x}(t))$, where $\mathbf{x}(t)$ is the projection of the parametric curve describing the PMA. This problem is summarized in Eq. 1:

$$\mathbf{x}^* = \arg\max_{\hat{\mathbf{x}}} \iint I(\mathbf{R}(s,t;\hat{\mathbf{x}}(t)))\,ds\,dt. \tag{1}$$

Trying to directly solve Eq. 1 would incur in errors, such as the collapse of the curve or the degeneration of the extremities. Therefore, this paper proposes to divide the problem into two unidimensional tracking phases, which first locates the pipe extremities along the longitudinal axis (t dimension), and then adjusts the whole curve in the transversal direction (s dimension). In order to be able to do that, the input image is transformed from the spatial domain into the longitudinal and transversal parametrization based on the curve, defined as

$$T(s,u) = \sqrt{I(\mathbf{R}(s,t))}, \quad \text{where } u = \int_{t_{min}}^{t} \left\| \frac{d\mathbf{x}(\tau)}{d\tau} \right\| d\tau. \tag{2}$$

The simplification provided by this transform enhances the tracking execution time substantially, given the fewer operations executed and the reduced image size.

The first phase of the tracking, which searches for the pipe in the longitudinal dimension, considers that pipe extremities lay upon a bright region of the alternating pattern marked over the pipe. The tracking is done by maximizing the energy in a window around each pipe extremity, using a Gauss-Newton approach, such that

$$u^* = \arg\max_{\hat{u}} \int_{\hat{u}-\omega/2}^{\hat{u}+\omega/2} \int_0^1 T(s,u)^2\,ds\,du \tag{3}$$

Each extremity is handled separately and the window size is as large as the pipe thickness.

Equivalently, the transversal tracking aims at maximizing the energy inside the region $\mathcal{R}$ along the s dimension. This is done by adjusting the previously detected PMA by a smooth function $\phi(u;\mathbf{q})$, which is governed by the parameter vector $\mathbf{q}$ and modeled as a B-spline. The problem can be described as finding the parameter vector $\mathbf{q}^*$, such that

$$\mathbf{q}^* = \arg\max_{\hat{\mathbf{q}}} \int_0^1 \int_{u^*_{min}}^{u^*_{max}} w(s)\; T(s+\phi(u;\hat{\mathbf{q}}),u)^2\,du\,ds, \tag{4}$$

where $w(s)$ is a weighting function used to increase the robustness of the tracking around the border of the pipe. By transforming Eq. 4 into a least-squares problem, it can be solved using the Gauss-Newton method.

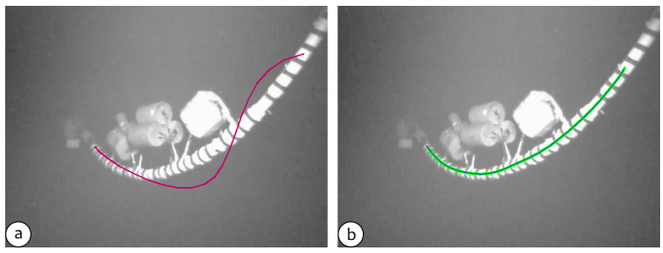


Figure 1: Underwater pipe tracking scenario. (a) shows in magenta the initialization, and (b) presents in green the tracked curve at the following frame.

Once the current position of the pipe projection in both images has been found, it is necessary to reconstruct the pipe in order to assess its geometry. The chosen approach is to reconstruct the PMA as a rational B-spline, which, according to Xiao and Li [2], requires to simultaneously fit a rational B-spline to each of the sample points. Therefore, the whole fitting can be modeled as a least-squares problem and solved minimizing the cost

$$\sum_i ||\mathbf{p}_i - \mathbf{x}(t_i;\{\mathbf{q}_k\})||^2 + \sum_j ||\mathbf{p}'_j - \mathbf{x}(t_j;\{\mathbf{q}'_k\})||^2, \tag{5}$$

where $\mathbf{p}_i$ and $\mathbf{p}'_j$ are sample points, and $\{\mathbf{q}_k\}$ and $\{\mathbf{q}'_k\}$ are the control points of the B-splines. By assuming that the pipe lies on a plane, it is possible to relate the curve projection in both images by an homography, which can only alter the x coordinates of the rectified curve points. These additional constraints are contributions of this paper. The solution, together with the centripetal parametrization [1], requires only $2\gamma+3$ values (where γ is the number of control points) to be optimized on the Levenberg-Marquardt algorithm.

The technique has been evaluated on videos from synthetic and real scenarios, by analyzing tracking and reconstruction errors. On synthetic videos, the mean tracking error was 0.011 pixels and the maximum 0.646 pixels, while 98.2% of the points had an error less than 0.1 pixel. The mean reconstruction error was 0.0066 m. On videos from a 1 : 8.89 scale experiment, the mean reconstruction error was less then 0.5 cm in 7 of 8 videos (which represents $1/6$ of the pipe's diameter). While it is impossible to assess these errors on videos from the actual operation, there is strong evidence that the reconstruction is precise and accurate. Further details on the implementation of this algorithm is described in the paper.

[1] E.T.Y. Lee. Choosing nodes in parametric curve interpolation. *Computer-Aided Design*, 21(6):363 – 370, 1989. ISSN 0010-4485. doi: http://dx.doi.org/10.1016/0010-4485(89)90003-1. URL http://www.sciencedirect.com/science/article/pii/0010448589900031.

[2] Yi Jun Xiao and YF Li. Optimized stereo reconstruction of free-form space curves based on a nonuniform rational b-spline model. *JOSA A*, 22(9):1746–1762, 2005.

Nonlinear Metric Learning for Alzheimer's Disease Diagnosis with Integration of Longitudinal Neuroimaging Features

Bibo Shi[1]
bs354409@ohio.edu
Yani Chen[1]
yc147311@ohio.edu
Kevin Hobbs[2]
hobbsk@ohio.edu
Charles D. Smith[3]
csmith@mri.uky.edu
Jundong Liu[1]
liu@cs.ohio.edu

[1] School of Electrical Engineering and Computer Science
Ohio University
Athens OH, USA

[2] Department of Biological Sciences
Ohio University
Athens OH, USA

[3] Department of Neurology
University of Kentucky
Lexington KY, USA

Identifying neuroimaging biomarkers of Alzheimer's disease (AD) is of great importance for diagnosis and prognosis of the disease. In this study, we develop a novel nonlinear metric learning method (**ML-TPS**) to improve biomarker identification for Alzheimer's disease (AD) and its early stage Mild Cognitive Impairment (MCI). Formulated under a constrained optimization framework, the proposed method learns a smooth nonlinear feature space transformation that pulls the samples of the same class closer to each other while pushing different classes further away. The thin-plate spline (TPS) is chosen as the geometric model due to its remarkable versatility and representation power in accounting for sophisticated deformations. In addition, a multi-resolution patch-based feature selection strategy is proposed to extract both cross-sectional and longitudinal features from MR brain images.

Nonlinear ML-TPS Learning a metric from the training input is equivalent to learn a feature transformation [1]. Depending on the feature space transformation to be sought, metric learning can be divided into linear and nonlinear groups. The proposed ML-TPS model is a direct generalization of linear ML through the application of a deformable geometric model – the thin-plate spline (TPS) - to transform the feature space. TPS is chosen due to its remarkable versatility and representation power in accounting for high-order deformations.

With the side information embedded in the class-equivalent constraints $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{x}_j) | \, \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class}\}$ and class-nonequivalent constraints $\mathcal{N} = \{(\mathbf{x}_i, \mathbf{x}_j) | \, \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to different classes}\}$, the goal of our ML solution is pulling the similar subjects closer while pushing dissimilar subjects apart, directly through a TPS nonlinear transformation f as described in Eqn. (2). This can be achieved through the following constrained optimization:

$$\begin{aligned} \min_{B,W} \quad & J = \sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{P}} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 + \lambda \|W\|_F^2 \\ \text{s.t.} \quad & \sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{N}} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 \geq 1; \\ & \sum_{i=1}^{p} W_i^k = 0, \quad \sum_{i=1}^{p} W_i^k \mathbf{x}_i^k = 0, \ \forall k = 1 \ldots m. \end{aligned} \tag{1}$$

where, f is the TPS transformation, as described in matrix format:

$$f(\mathbf{x}) = \mathbf{x} \cdot B + \begin{pmatrix} G(\mathbf{x}, \mathbf{x}_1) \\ \cdots \\ G(\mathbf{x}, \mathbf{x}_p) \end{pmatrix} \cdot W = \mathbf{x} \cdot B + \vec{G} \cdot W. \tag{2}$$

B is the linear transformation matrix, and W is the weight matrix for the nonlinear parts. $\mathbf{x}_p$ are the anchor points used to compute the TPS kernel. Compared with a classic liner metric learning model MMC [4], another component $\|W\|_F^2$, the squared Frobenius norm of W, is added to the objective function as a regularizer to prevent overfitting. λ is the weighting factor to control the importance of two components. Similar as in MMC, the nonequivalent constraint $\sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{N}} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 \geq 1$ is to impose a scaling control to avoid trivial solutions. The other two equivalent constraints with respect to W is to ensure that the elastic part of the transformation is zero at infinity. W^k is the kth column of W, and $\mathbf{x}^k$ is the kth component of $\mathbf{x}$.

Feature extraction Feature extraction and selection from the ADNI database [3] is also in great need of further exploration. Most of the existing works are either cross-sectional features obtained at one point in time, or "static" longitudinal volumetric information acquired at two or multiple time points but only through structural segmentation. In part due to the unavailability of deformation data under ADNI, "dynamic" longitudinal information such as the atrophy over time at various gray matter (GM) areas, which is a major hallmark in the progression of AD, has not been fully utilized in the literature. In this paper, we propose a multi-resolution patch selection strategy, with both cross-sectional (baseline) and longitudinal atrophy features extracted from MR brain images.

The neuroimage data used in this work were obtained from the ADNI database. We consider only the subjects for whom the baseline (M0) visits and 12-month follow-up (M12) T1-weighted MRIs, together with their *MIDAS Whole Brain Masks*, are all available. Fig. 1 illustrates an overview of the schematic diagram of our proposed framework. Two sets of class-discriminative patches are first extracted at three different levels of resolutions: one set is based on the gray matter (GM) densities of the baseline MRIs, and the other is selected through the longitudinal deformation magnitudes (DM) estimated between baselines (M0-MRIs) and follow-ups (M12-MRIs). Both selection procedures are conducted through statistical significance tests between groups. With p-value set to 0.05, this step returns more than 1000 patches in each set. To capture the most discriminative patches, a wrapper feature selection [2] is carried out with a greedy forward searching within the pools of GM patches, DM patches, and their combination.

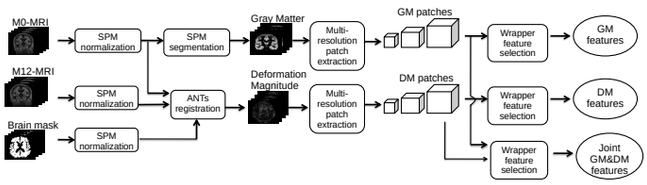


Figure 1: Flowchart of the proposed two-step patch based feature extraction and selection strategy.

Conclusion Implementations of the proposed methods, as well as the detailed experimental evaluations, are described in the paper. The two novel components in our solution – nonlinear metric learning through TPS and integration of longitudinal atrophy information – have both proven to be highly effective in improving the classification of AD/MCI versus NC.

[1] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[2] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3: 1157–1182, 2003.

[3] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[4] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. *NIPS*, pages 521–528, 2003.

Boosted Metric Learning for Efficient Identity-Based Face Retrieval

Romain Negrel
romain.negrel@unicaen.fr
Alexis Lechervy
alexis.lechervy@unicaen.fr
Frederic Jurie
frederic.jurie@unicaen.fr

GREYC, CNRS UMR 6072
Université de Caen Basse-Normandie
France

The performance of many visual recognition algorithms – *e.g.* face verification, person re-identification – heavily depends on the metric used to measure the similarity between input data. Recent works such as have shown how interesting it is to learn an optimal metric for a particular task using Metric Learning (ML). Most approaches learn a Mahalanobis metric based on an objective function whose constraints comes either from a set of labelled examples or, more frequently, from sets of positive (same class) and negative (different class) pairs.

In the recent works that have used metric learning (ML), the works of Shen *et al*. [4] and Bi *et al*. [2], introducing algorithms based on Boosting approach, deserves a particular attention due to the interesting properties they offers: i) they are efficient and scalable, as no semidefinite programming is required, at each iteration only the largest eigenvalue and corresponding eigenvectors are needed, ii) Like AdaBoost, they don't have any parameter to tune and is easy to implement as no sophisticated optimization techniques are involved. It hence contrasts with most of the commonly used ML methods for which hyper-parameters, often introduced for regularization purpose, have to be adjusted by cross-validation.

However, one strong limitation of these approaches [2, 4] is that it requires having triplets for learning the metric *i.e.* constraints expressed under the form $\mathrm{D}(x_a,x_b) < \mathrm{D}(x_a,x_c)$ where x_a, x_b, x_c are three input vector for which the label is known: positive and negative pairs within a constraint do have to share a common vector. This is a limitation as for most of the verification task only training pairs are available (*e.g.* for person re-identification on the Viper dataset only pairs of same/different persons are provided for training, and thus using such triplets is not possible).

One of the key contribution of this paper is to propose a metric learning approach based on boosting allowing to use pairs of points for training *i.e.* which can use constraints such as $\mathrm{D}(x_a,x_b) < \mathrm{D}(x_c,x_d)$ while keeping the good properties of [4] (scalability, simplicity, no parameters, *etc*.). Our approach is based on the ranking algorithm RankBoost [3], known to offer 3 particularly interesting features: i) no hyperparameter, ii) great robustness to overfitting (explained with a standard VC-dimensional analysis techniques), and iii) a computational trick for reducing the complexity of the learning step. In the following, we show how to build on RankBoost [3] to efficiently address this metric learning problem.

Moreover, we propose a method for building hierarchical face representations for efficient *identity face retrieval task*. The main idea is to use a ML algorithm for building a tree structured representation of a large set of faces, such that identical-identity faces belongs to the same cluster at each tree level. The hierarchical clustering is built recursively: at each level, a new metric is learned using the faces belonging to a particular tree node which is then used to locally cluster the faces and to create the sub-branches.

The retrieval for a new query face is done by visiting the tree from top to bottom until a leaf is reached; if the current is a non-leaf node, the face is projected into the corresponding subspace and compared with its centroids, which allows to move to the closest child node. When a leaf is reached, the face is compared with all the faces contained in the leaf using the corresponding local metric. The search cost is then equal to the cost of traverses of the tree, plus the cost of faces comparison in the final node. This cost is much lower than of the cost to compared all the faces of the database.

Our method is a extension of the method proposed by Bhattarai *et al*. [1]. We introduced a new method to select the ML training samples in the tree node in order to keep easily the information learned in the metric of parent nodes.

The proposed approach is validated on the *identity face retrieval task*, such as defined by Bhattarai *et al*. [1]. This task consists in retrieving the face(s) of the same person than a query face, from a large database of faces. The task is evaluated both in terms of scalability and accuracy. We use the Labeled Faces in the Wild (LFW) database with the evaluation protocol, the images/query and the LBP image signature than the one used by Bhattarai *et al*. [1]. The dimension of the output space to 32. The criterion used to evaluate the performance is the mean of k-call@n, with $k = 1$. We measure the performance for several values of n, *i.e.*, $n \in \{1, 10, 20\}$.

Method	Dim.	No distractor			+100.000 distractors			+1.000.000 distractors		
		n= 1	n= 10	n= 20	n= 1	n= 10	n= 20	n= 1	n= 10	n= 20
Unreduced	9860	31.9	53.7	60.5	31.4	52.7	59.3	31	50.6	57.2
MLBoost	32	30.7	53.7	62.2	26.2	43.0	48.7	21.5	34.5	38.5
PCCA [1]	32	-	32.1	41.3	-	23.6	28.3	-	13.4	18.2
MLBoost-Tree-d3-Global	32	27.9	50.4	58.6	24.1	39.5	46.1	20.1	31.4	36.2
MLBoost-Tree-d3	32	29.6	54.6	62.2	23.6	39.5	45.2	17.7	31.0	35.2
PCCA-Tree-d3 [1]	32	-	33.3	40.9	-	27.4	32.2	-	18.5	23.9
MLBoost-Tree-d4-Global	32	27.0	49.2	56.7	22.7	37.4	42.3	18.9	30.5	33.6
MLBoost-Tree-d4	32	28.6	52.0	59.3	21.0	35.5	40.9	17.3	26.7	30.5
PCCA-Tree-d4 [1]	32	-	36.2	41.6	-	32.4	36.6	-	30.3	31.7

Table 1: Performance of the proposed method and comparison with PCCA [1].

Table 1 compares the performance of the proposed MLBoost algorithm to the performance of the state-of-the art PCCA (Pairwise Constrained Component Analysis) [1] with and without distractors and with and without tree. We restrict ourselves to a binary clustering tree of 3 and 4 levels (denoted with suffix "-d3" and "-d4"). Without the tree, the performance of the proposed approach improves over PCCA by about 20% ($n = 10$).

Searching the faces in only one leaf of the tree affects the overall performance: misclassifications that can happen at any level of the tree cannot be recovered later. This loss increases with the depth of the tree. We can observe this by comparing the performance of the global metric used without and with the tree structure ("MLBoost") (methods whose names include "MLBoost-Tree-d*-Global"). The loss is of about 1-3% at $n = 1$.

We can also observe that the local metric (the lines denoted as "MLBoost-Tree-d*") improves the performance of the tree-based structure for the small values of n, compensating the loss induced by the tree (without distractors). Comparing our method with the state-of-the-art method of [1] (lines denoted as "PCCA-Tree-d*") shows that our methods can perform up to 20% better for some configurations.

In conclusion, we have proposed a ML method base on Boosting with quadrupole. Our method have show better results of the stat-of-the-are method moreover without tuning hyperparameters. Our method to build a tree for efficient *identity face retrieval task* greatly reduced the research cost while preserving the performances.

[1] Binod Bhattarai, Gaurav Sharma, Frederic Jurie, and Patrick Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *European Conference on Computer Vision (ECCV) Workshops*, pages 1–13, 2014.

[2] Jinbo Bi, Dijia Wu, Le Lu, Meizhu Liu, Yimo Tao, and Matthias Wolf. Adaboost on low-rank psd matrices for metric learning. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2617–2624. IEEE, 2011.

[3] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.

[4] Chunhua Shen, Junae Kim, Lei Wang, and Anton Van Den Hengel. Positive semidefinite metric learning using boosting-like algorithms. *The Journal of Machine Learning Research*, 13(1):1007–1036, 2012.

Enforcing Point-wise Priors on Binary Segmentation

Feng Li
http://www.fengl.org/
Fatih Porikli
http://www.porikli.com/

Light
Palo Alo, CA 94301, USA
Australian National University / NICTA
Canberra ACT 0200, Australia

Segmentation is one of the extensively studied areas in computer vision. Pioneering work [4] treats image segmentation as a graph partitioning problem and proposes a normalized cut criterion measures both the total dissimilarity between the different groups as well as the total similarity within the groups. Seminal work [3] derives a matting Laplacian matrix from multiple matte equations. In comparison with random walk [2] and normalized cuts [4], they adapt a correlation measure instead of the exponent of color distance, a local scaling instead of a global scaling, and formulate a least-squares solution with constraints from user input. Local scaling leads to better clustering especially when the data includes multiple scales and the clusters are placed within a cluttered background.

Non-negative point-wise priors such as saliency map, defocus field, foreground mask, object location window, and user given seeds, come in the form of confidence or probability values, and they are often incomplete, irregular, and noisy, which eventually makes the labelling task a challenge. In this paper, we aim to to extract image regions that are aligned on the object boundaries and also in accordance with the given point-wise priors. To this end, we define a graph Laplacian spectrum based cost function and embed it into a minimization framework. For a comprehensive understanding, we analyze five alternative formulations, and demonstrate that the robust function version produces consistently superior results.

To estimate the optimal $\mathbf{x}$ from the prior $\mathbf{x}^\star$, we compute a graph Laplacian matrix L from G, where G is a graph representation of the input image $\mathbf{y}$. In other words, the Laplacian matrix L will regularize our under-constrained optimization formulation by laying on the image structure inherent in $\mathbf{y}$. This enables us to define the binary segmentation problem as a least-squares constrained optimization

$$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{x}^\star\|^2, \quad \text{s.t.} \quad L\mathbf{x}=0. \tag{1}$$

We call the above constraint $L\mathbf{x}=0$ the *Laplacian spectrum* constraint. This is a generalization of the conventional approaches and does not require a specific numerical solver as the matting Laplacian. This constraint enforces a given image structure on the prior information (in the data fidelity term $\|\mathbf{x}-\mathbf{x}^\star\|^2$). With this constraint, the optimal $\mathbf{x}$ should lie in the null-space of L, which means $\mathbf{x}$ should be constant within each connected component of the graph G.

Depnding on the norm used for the *Laplacian spectrum* constraint, we define 5 alternative objective functions solving the optimal $\mathbf{x}$. Here we only discuss about Convex Function and Robust Function with ℓ_2 Norm on Constraint, for the rest 3 formations please refer to our paper.

Instead of solving a constrained optimization problem Eq.(1), we can transform it into an unconstrained minimization:

$$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{x}^\star\|^2+\beta\|L\mathbf{x}\|^2\,, \tag{2}$$

with a penalty β that enforces the structure in $\mathbf{y}$. Setting the derivative of the objective function Eq.(2) to 0, we obtain a closed form solution:

$$\mathbf{x}=(\beta L^\top L+I)^{-1}\mathbf{x}^\star=P\mathbf{x}^\star\,, \tag{3}$$

where I is an identity matrix, and P can be viewed as a modified projection matrix[1].

Since the residual $\delta=|\mathbf{x}-\mathbf{x}^\star|$ has many spatially continuous large outliers and the least square data fidelity term weights each sample with a quadratic norm, the final estimation of Eq.(2) can be distorted severely. Depending on its quality, the prior information $\mathbf{x}^\star$ could contain incomplete and inaccurate indicators, for instance strong responses across segment boundaries. This may confuse the segmentation algorithm and cause mislabeling. A better option is to weight large outliers less and use the structure information from the Laplacian spectrum constraint to recover the $\mathbf{x}$. Therefore, we borrow existing principles from robust statistics [1] and adapt a robust functional to replace the least square cost as

$$\min_{\mathbf{x}} \rho(\mathbf{x}-\mathbf{x}^\star)+\beta\|L\mathbf{x}\|^2\,, \tag{4}$$

where ρ is the Huber function, which is a parabola in the vicinity of 0 and increases linearly when δ is large. Thus, the effects of large outliers can be eliminated significantly.

When written in matrix form, we use a diagonal weighting matrix $W=\mathrm{diag}(w_1,\ldots,w_n)$ to represent the Huber weight function. Therefore, the data fidelity term can be simplified as $\rho(\mathbf{x}-\mathbf{x}^\star)=\|W(\mathbf{x}-\mathbf{x}^\star)\|^2$. As a result, the problem Eq.(4) can be solved efficiently in an iterative least square approach. At each iteration, the optimal $\mathbf{x}$ is updated as

$$\mathbf{x}=(\beta L^\top L+W)^{-1}W\mathbf{x}^\star\,. \tag{5}$$

To initialize the algorithm, we could set $W=I$ for the first iteration, or when a confidence measure $m^\star$ is available for $\mathbf{x}^\star$, we could directly use $m^\star$ to initialize W. For the details of this algorithm and the comparison with other algorithms, please refer to our paper.

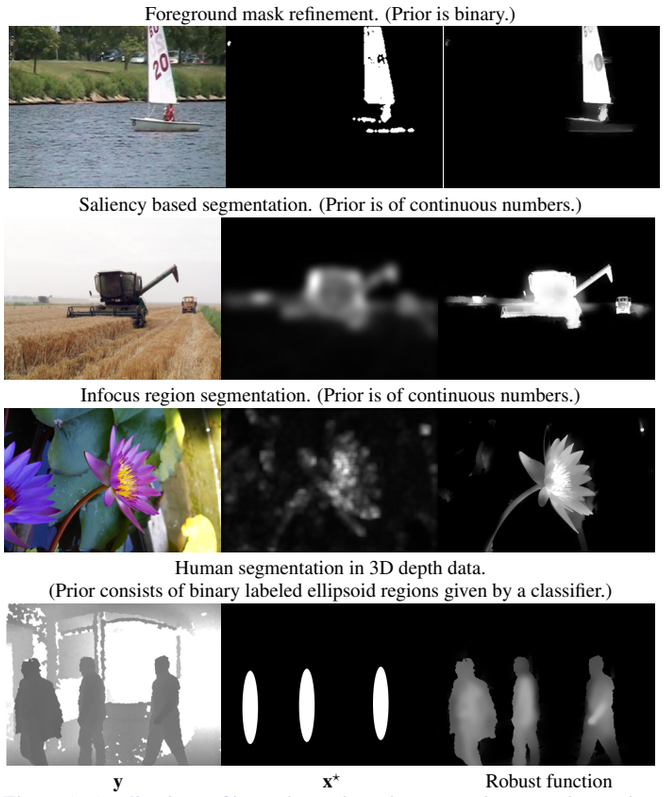


Figure 1: Applications of imposing point-wise constraints to a given prior. For each example, we show the input image structure $\mathbf{y}$, the confidence maps (labels) $\mathbf{x}^\star$, and our results. Segmentation results are obtained by simply thresholding and the threshold is fixed to 0.1 for all examples.

[1] M. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *IJCV*, 19(1):57–91, 1996.

[2] L. Grady. Random walks for image segmentation. *PAMI*, 28(11): 1768–1783, 2006.

[3] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *PAMI*, 30(2):228–242, 2008.

[4] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[1] P does not satisfy the idempotent property of the projection matrix.

Sparse Discrimination based Multiset Canonical Correlation Analysis for Multi-Feature Fusion and Recognition

Hongkun Ji
jihongkunnust@126.com
Xiaobo Shen
njust.shenxiaobo@gmail.com
Quansen Sun
qssun@126.com
Zexuan Ji
jizexuan840820@hotmail.com

School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing 210094, China

Canonical correlation analysis (CCA) [3] is a powerful tool for finding the correlation between two sets of multidimensional variables. It investigates the linear correlations between two sets of random variables and linearly projects two sets of random variables into a lower-dimensional space in which they are maximally correlated. However, most of CCA-based methods are not efficient for multiple (more than two) feature representations classification tasks [5]. As a generalized extension of CCA, multiset canonical correlation analysis (MCCA) [6, 7] was proposed to solve this problem. In recent years, with the development of sparse representation [1, 2, 4, 9], sparse representation based classification (SRC) has been proved to be robust for feature selection and efficient in handling occlusion and corruption. Sparsity preserving projections (SPP) [8] aims to preserve the sparse reconstructive relationship of the data. Extensive experiments on some common datasets show that the projections obtained by SPP are invariant to rotations, rescalings and translations of the data.

However, MCCA usually fails to discover the intrinsic sparse reconstructive relationship and discriminating structure of multiple data spaces in real-world applications. In this paper, by taking discriminative information of within-class and between-class sparse reconstruction into account, we propose a novel algorithm, called sparse discrimination based multiset canonical correlations (SDbMCCs), to explicitly consider both discriminative structure and sparse reconstructive relationship in multiple representation data. In addition to maximizing between-set cumulative correlations, SDbMCC minimizes within-class sparse reconstructive distances and maximizes between-class sparse reconstructive distances, simultaneously. The feasibility and effectiveness of the proposed method is verified on four popular databases (CMU PIE, ETH-80, AR and Extended Yale-B) with promising results.

We characterize the within-class and between-class scatters, i.e., minimizing the average within-class distance and maximizing the average between-class distance, simultaneously. Given m feature sets $\{X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}] \in \Re^{d_i \times n}\}_{i=1}^m$ extracted from n samples of c classes. Let $\Phi : \Re^n \to \Re^n$ be the characteristic function that selects the coefficients of $s_j^{(i)}$ associated with the same class of the sample $x_j^{(i)}$, where $s_j^{(i)}$ is the sparse weight vector of $x_j^{(i)}$ as in SPP. For $s_j^{(i)} \in \Re^n$, with the assumption that $x_j^{(i)}$ belongs to the kth class, $\Phi(s_j^{(i)})$ is a vector whose only nonzero entries are the entries in $s_j^{(i)}$ that are associated with class k. For each feature set $X^{(i)}$, after the projections of $\{x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}\}$ onto the projection axis $\alpha^{(i)}$, we can get the within-class scatter defined by

$$\begin{aligned} J_w^{(i)} &= \sum_{i=1}^{n} \| \alpha^{(i)T} x_j^{(i)} - \alpha^{(i)T} X^{(i)} \Phi(s_j^{(i)}) \|^2 \\ &= \| \alpha^{(i)T} X^{(i)} - \alpha^{(i)T} X^{(i)} \Phi(S^{(i)}) \|^2 \\ &= \alpha^{(i)T} X^{(i)} S_\Phi^{(i)} X^{(i)T} \alpha^{(i)} \end{aligned} \tag{1}$$

where $S^{(i)} = [s_1^{(i)}, s_2^{(i)}, \cdots, s_n^{(i)}]$ and $S_\Phi^{(i)} = [I - \Phi(S^{(i)})] \cdot [I - \Phi(S^{(i)})]^T$. Then we can construct the total within-class scatter for all m feature sets as follows:

$$J_w = \sum_{i=1}^{m} J_w^{(i)} = \sum_{i=1}^{m} \alpha^{(i)T} X^{(i)} S_\Phi^{(i)} X^{(i)T} \alpha^{(i)} \tag{2}$$

Obviously, minimizing the total within-class scatter J_w can lead to a better sparse reconstructive relationship of the samples belonging to the same class.

Similarly to the within-class scatter, let $\Psi : \Re^n \to \Re^n$ be the characteristic function that selects the coefficients of $s_j^{(i)}$ associated with the different classes of the sample $x_j^{(i)}$. We can construct the total between-class scatter for all m feature sets as follows:

$$J_b = \sum_{i=1}^{m} J_b^{(i)} = \sum_{i=1}^{m} \alpha^{(i)T} X^{(i)} S_\Psi^{(i)} X^{(i)T} \alpha^{(i)} \tag{3}$$

where $S_\Psi^{(i)} = [I - \Psi(S^{(i)})] \cdot [I - \Psi(S^{(i)})]^T$. Maximizing the total between-class scatter can significantly enhance the discriminative capability.

By considering both discriminative and sparse reconstructive relationship in multiple representation data, we tend to minimize the total within-class sample distance and maximize the total between-class sample distance in the low-dimensional embedding subspace, as well as maximizing the correlations. Combining with the original MCCA objective function, we can construct the model of SDbMCC as follows:

$$\begin{aligned} &max J(\alpha) = \Sigma_{i=1}^m \Sigma_{j=1}^m \alpha^{(i)T} S_{ij} \alpha^{(j)} - \tau[\eta J_w - (1-\eta) J_b] \\ &s.t.\ \ \Sigma_{i=1}^m \alpha^{(i)T} S_{ii} \alpha^{(i)} = 1. \end{aligned} \tag{4}$$

where $\alpha^T = [\alpha^{(1)T}, \alpha^{(2)T}, \cdots, \alpha^{(m)T}]$, S_{ii} is the within-set covariance matrix of feature set $X^{(i)}$, and S_{ij} is the between-set covariance matrix between feature sets $X^{(i)}$ and $X^{(j)}$. Because the ratio between within-class and between-class sparse representation weights suffers from a variety of noises, we further hope that the tradeoff between between-class scatters and within-class scatters, as well as the tradeoff between correlations and the sparse reconstructive relationship, can be tuned. Therefore, we construct the corresponding regularizations with two tradeoff parameters τ and η. By utilizing Lagrange multiplier, we can solve the optimization problem in Eq. (4).

[1] M. Davenport, M. Duarte, M. Wakin, D. Takhar, K. Kelly, and R. Baraniuk. The smashed filter for compressive classification and target recognition. In *In Proc. IS&T/SPIE Symposiumon Electronic Imaging: Computational Imaging*, 2006.

[2] M. Davenport, M. Wakin, and R. Baraniuk. Detection and estimation with compressive measurements. In *Technical Report*, 2007.

[3] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28: 321–377, December 1936.

[4] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[5] M. Kan, S.G. Shan, H.H. Zhang, S.H. Lao, and X.L. Chen. Multi-view discriminant analysis. In *European Conference on Computer Vision (ECCV)*, volume 7572, pages 808–821, 2012.

[6] J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

[7] A.A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on image processing*, 11:293–305, 2002.

[8] L.S. Qiao, S.C. Chen, and X.Y. Tan. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43:331–341, 2010.

[9] J. Wright, A. Yang, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, 2009.

Towards 4D Coupled Models of Conversational Facial Expression Interactions

Jason Vandeventer[1]
VandeventerJM@Cardiff.ac.uk
Lukas Gräser[3]
Lukas.Graeser@Campus.tu-berlin.de
Magdalena Rychlowska[2]
Rychlowska@Cardiff.ac.uk
Paul L. Rosin[1]
RosinPL@Cardiff.ac.uk
David Marshall[1]
MarshallAD@Cardiff.ac.uk

[1] School of Computer Science and Informatics
Visual Computing Group
Cardiff University
Cardiff, Wales, UK

[2] School of Psychology
Cardiff University
Cardiff, Wales, UK

[3] School of Electrical Engineering and Computer Sciences
Berlin Institute of Technology
Berlin, Germany

Abstract: In this paper we introduce a novel approach for building 4D coupled statistical models of conversational facial expression interactions. To build these coupled models we use 3D AAMs for feature extraction, 4D polynomial fitting for sequence representation, and concatenated feature vectors of frontchannel-backchannel interactions. Using a coupled model of conversation smile interactions, we predicted each sequence's backchannel signal. In a subsequent experiment, human observers rated predicted backchannel sequences as highly similar to the originals. Our results demonstrate the usefulness of coupled models as powerful tools to analyse and synthesise key aspects of conversational interactions, including conversation timings, backchannel responses to frontchannel signals, and the spatial and temporal dynamics of conversational facial expression interactions.

Methodology: Using a 4D database of natural, dyadic conversations [3], conversational interactions were manually annotated for conversational expressions. The sequences were tracked using a 4D sparse tracking approach, which uses 3D shape and texture.These tracked points are used as control points in a dense correspondence method. This method uses a Thin Plate Spline (TPS) based algorithm, with an additional "snapping" step, to modify the geometry of one mesh (reference mesh) so that it matches that of another mesh (target mesh). The tracking and inter-subject registration methods were developed in-lab and details for these approaches can be found in [2]. Statistical modelling of these sequences was performed using 3D Active Appearance Models (AAMs) [1] and a polynomial regression technique for sequence representation.

Experiments: In Experiment 1, individual sequences were classified as either frontchannel or backchannel. In Experiment 2, these sequences were also modified and used in a perceptual experiment that evaluated the realism of the synthesised sequences. For Experiments 3 and 4, a coupled statistical model of conversation interactions was built by concatenating the frontchannel sequence and corresponding backchannel sequence feature vectors (Table 1).

	Sequences			
FC	PC 1	PC 1	PC 1	PC 1
	PC 2	PC 2	PC 2	PC 2
	PC 3	PC 3	PC 3	PC 3
	Offset	Offset	Offset	Offset
BC	PC 1	PC 1	PC 1	PC 1
	PC 2	PC 2	PC 2	PC 2
	PC 3	PC 3	PC 3	PC 3

Table 1: Example of the coupled model's feature vectors.

The *offset* value is the number of frames between the beginning of the frontchannel expression and the beginning of the backchannel expressions. In this work, we used the coupled model to predict the characteristics of one side of the interaction given the information about the other side. In Experiment 3, the predicted sequences were used in a classification experiment which used the original sequences for training. In Experiment 4, predicted backchannel sequences were synthesised and used in a perceptual study, where human observers evaluated the similarity between the original and predicted sequences.

Experiment Results Overview:

- **Experiment 1:** Frontchannel and backchannel sequences were classified with 97.54% accuracy.
- **Experiment 2:** Modified sequences were perceived as realistic, for both expression realism and image quality.
  - Figure 1 shows the realism ratings for each modification level. Any values above the red line (realism midpoint) represent stimuli that were perceived as realistic.

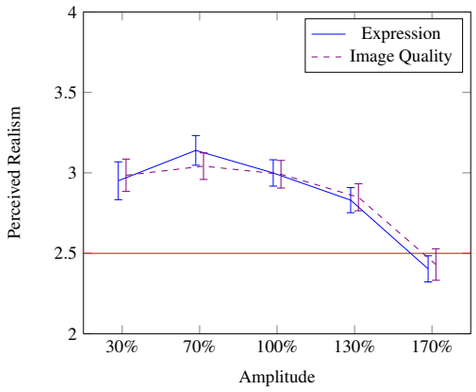


Figure 1: Experiment 2 Results

- **Experiment 3:** Predicted frontchannel and backchannel sequences were classified with 95.45% accuracy.
- **Experiment 4:** Similarity ratings, averaged within participants, were significantly higher than the scale midpoint (2.5), $M = 2.90$, $SD = 0.31$, $t(27) = 7.00$, $p < .001$, suggesting that participants consistently perceived the predicted videos as similar to the original versions.

Conclusion: We introduce a novel method for modelling 4D conversation facial expression interactions. This approach allows for the synthesis of highly-realistic expression sequences, which when modified are still perceived as realistic. Our techniques also allow for the analysis of conversation interactions, as well as the prediction and synthesis of interaction characteristics. The methods described in our paper are supported by the results of two classification experiments and two perceptual studies. Coupled statistical models of conversational interactions will allow for advances in many areas, such as behaviour analysis, perceptual psychology, and modelling and synthesising facial expressions of digital characters.

[1] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *Pattern Analysis & Machine Intelligence (PAMI), IEEE Transactions on*, 23(6):681–685, 2001.

[2] Lukas Gräser, Jason Vandeventer, Job van der Schalk, Paul L. Rosin, and David Marshall. 4D tracking and inter-subject registration for the synthesis of realistic facial expression sequences. *Under Review*, 2015.

[3] Jason Vandeventer, Andrew J. Aubrey, Paul L. Rosin, and David Marshall. 4D Cardiff Conversation Database (4D CCDb): A 4D database of natural, dyadic conversations. In *Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, 2015.

A Unified Bayesian Approach to Multi-Frame Super-Resolution and Single-Image Upsampling in Multi-Sensor Imaging

Thomas Köhler[12]
thomas.koehler@fau.de
Johannes Jordan[1]
johannes.jordan@fau.de
Andreas Maier[12]
andreas.maier@fau.de
Joachim Hornegger[12]
joachim.hornegger@fau.de

[1] Pattern Recognition Lab,
Department of Computer Science,
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

[2] Erlangen Graduate School in Advanced Optical Technologies (SAOT)
Erlangen, Germany

Multi-sensor imaging has become an emerging field of research with a vast number of applications including color and multispectral cameras or 3-D range imaging to capture RGB-D data. One common property of these systems is that they provide *multi-channel* images that are, unfortunately, often limited in terms of their spatial resolution. However, channel-wise resolution enhancement might not be optimal as it ignores dependencies across the channels. This is obvious if structures are visible in multiple channels. For this reason, tailor-made approaches for single-image upsampling and multi-frame super-resolution have been investigated in different domains. Super-resolution of color images exploits the correlation of color channels as a prior to super-resolve them [2]. In multispectral imaging, super-resolution can be guided by panchromatic data [1]. A similar concept has been established in range imaging, where high-quality color images guide single-image upsampling [3, 5] or multi-frame super-resolution [6] of range images. To avoid the need of reliable guidance data, joint resolution enhancement of all channels has been proposed [4]. However, this formulation does only consider simplified setups with two image channels.

Bayesian Multi-Sensor Model. This work introduces multi-frame super-resolution and single-image upsampling that exploit correlations across complementary channels in a unified way. Unlike prior work, this framework is neither limited to a certain setup nor a fixed number of channels and does not require guidance data. Our framework is derived from a Bayesian model to infer a high-resolution multi-channel image $\boldsymbol{x}$ assembled from n channels $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ from multiple low-resolution multi-channel images $\boldsymbol{y}_1,\dots,\boldsymbol{y}_K$. As a key idea, we formulate an image prior according to:

$$p(\boldsymbol{x}_i \,|\, \mathcal{X}_i) \propto \exp\Big\{ -\big(\lambda_i R_{\text{intra}}(\boldsymbol{x}_i) + \sum_{j=1,j\neq i}^{n} \mu_{ij} R_{\text{inter}}(\boldsymbol{x}_i,\boldsymbol{x}_j;\boldsymbol{\Phi}_{ij})\big)\Big\}, \quad (1)$$

where $\mathcal{X}_i = \{\boldsymbol{x}_1,\dots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\dots,\boldsymbol{x}_n\}$ and consider two aspects: 1) Each channel $\boldsymbol{x}_i$ is described by an intra-channel prior defined by the regularization term $R_{\text{intra}}(\boldsymbol{x}_i)$ with the regularization weight $\lambda_i \geq 0$. 2) Dependencies across channels $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are modeled by an inter-channel prior $R_{\text{inter}}(\boldsymbol{x}_i,\boldsymbol{x}_j;\boldsymbol{\Phi}_{ij})$ with pair-wise weights $\mu_{ij} \geq 0$ and hyperparameters $\boldsymbol{\Phi}_{ij}$.

Locally Linear Regression. The inter-channel prior is defined by *locally linear regression* (LLR) that explains dependencies across channels by:

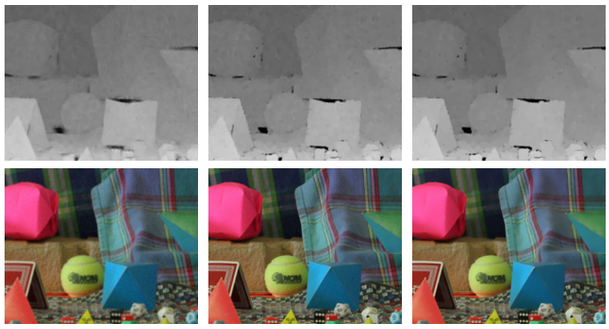
$$R_{\text{inter}}(\boldsymbol{x}_i,\boldsymbol{x}_j;\boldsymbol{\Phi}_{ij}) = \big|\big|\, \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) \odot \big(\boldsymbol{A}_{ij}\boldsymbol{x}_i + \boldsymbol{b}_{ij} - \boldsymbol{x}_j\big)\big|\big|_2^2, \quad (2)$$

where the hyperparameters $\boldsymbol{\Phi}_{ij}$ are given by filter coefficients $\boldsymbol{A}_{ij}$ and $\boldsymbol{b}_{ij}$ for each channel pair $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ and associated confidence weights $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.



(a) Channel-wise (b) Multi-channel [2] (c) Multi-channel proposed

Figure 1: Comparison of multi-frame color super-resolution (top row) and single-image upsampling of multispectral data (bottom row).



(a) Guided upsampling [5] (b) Channel-wise (c) Multi-channel proposed

Figure 2: Single-image upsampling of RGB-D data.

This prior is spatially adaptive and tolerates outliers due to image regions that violate the LLR assumption. The inference of the hyperparameters $\boldsymbol{\Phi}_{ij}$ that are treated as latent variables as well as the multi-channel image $\boldsymbol{x}$ are formulated as *maximum a-posteriori* (MAP) estimation. For this purpose, our paper presents an efficient alternating minimization scheme.

Applications. We evaluated our unified approach for color-, multispectral and range imaging. In color and multispectral imaging, we compared the proposed inter-channel prior to conventional channel-wise processing as well as the color channel regularization of Farsiu et al. [2], see Fig. 1. Our multi-channel method got rid of color artifacts present in channel-wise super-resolution, e. g. jagged edges highlighted in Fig. 1 (top). Unlike [2] that erroneously copied structure from other, original channels to the reconstructed channel, this issue was avoided by our adaptive confidence weighting as highlighted for multispectral data in Fig. 1 (bottom). In range imaging, our method was evaluated for single-image upsampling, where we compared channel-wise and guided upsampling [5], see Fig. 2. Unlike [5], our multi-channel approach does not rely on high-quality color images as guidance. In particular, it achieved better reconstructions of surfaces and edges in range images compared to channel-wise and guided upsampling.

[1] M. Aguena and N. Mascarenhas. Multispectral image data fusion using POCS and super-resolution. *Computer Vision and Image Understanding*, 102(2):178–187, 2006.

[2] S. Farsiu, M. Elad, and P. Milanfar. Multiframe demosaicing and super-resolution of color images. *IEEE Transactions on Image Processing*, 15(1):141–159, 2006.

[3] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof. Image Guided Depth Upsampling Using Anisotropic Total Generalized Variation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2013.

[4] F. C. Ghesu, T. Köhler, S. Haase, and J. Hornegger. Guided image super-resolution: A new technique for photogeometric super-resolution in hybrid 3-d range imaging. In *Pattern Recognition*, pages 227–238. Springer, 2014.

[5] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–409, 2013.

[6] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, C. Stock, J. Hornegger, and H. Feußner. Multi-sensor super-resolution for hybrid range imaging with application to 3-d endoscopy and open surgery. *Medical Image Analysis*, 24(1):220–234, 2015.

Boosting the Performance of Model-based 3D Tracking by Employing Low Level Motion Cues

Ammar Qammaz[1]
ammarkov@ics.forth.gr
Nikolaos Kyriazis[1]
kyriazis@ics.forth.gr
Antonis A. Argyros[21]
argyros@ics.forth.gr

[1] Insitute of Computer Science, FORTH,
N. Plastira 100, Vassilika Vouton,
GR70013, Heraklion, Crete, Greece
[2] Computer Science Department, University of Crete,
Heraklion, Crete, Greece

3D tracking of objects and hands in an object manipulation scenario is a very interesting computer vision problem with a wide variety of applications ranging from consumer electronics to robotics and medicine. Recent advances in this research topic allow for 3D tracking of complex scenarios involving bi-manual manipulation of several rigid objects using commodity hardware and with high accuracy. The problem with these approaches is that they treat tracking as a search problem whose dimensionality increases with the number of objects in the scene. This fact typically limits the number of the tracked objects and/or the processing framerate. In this paper we present a method that utilizes simple low level motion cues for dynamically assigning computational resources to parts of the scene where they are actually required. In a series of experiments, we show that this simple idea improves tracking performance dramatically at a cost of only a minor degradation of tracking accuracy.

The works that are most related to ours are the approaches by Kyriazis and Argyros on top-down 3D tracking of multiple active objects from RGBD input [1, 2]. The methodological part of our contribution can be briefly described as an extra processing node in the pipeline of [2] which it extends. The tracking approach in [2], the Ensemble of Collaborative Trackers (ECT), regards a set of semi-independent trackers. Each tracker is associated with a distinct object in the scene. For an object to be tracked, a separate optimization problem is solved, one for each frame and each object. Each optimization problem is numerically solved using a black box optimizer, i.e. a variant of the Particle Swarm Optimization (PSO) algorithm. PSO treats the objective function as an oracle and queries it on purposefully evolved "guesses" in the search space, in order to find the optimum. The harder the problem is, the more guesses (budget, which is the product of PSO particles and PSO generations that need to be computed) are required for adequate accuracy to be achieved. For example, tracking the pose and the articulation of the hand amounts to solving a 27-parameter optimization problem. This is much harder than solving for the 3D pose of a rigid object (6 parameters). ECT [2] assigns a fixed amount of computational resources to each tracking sub-problem which depends on this notion of complexity and which is empirically estimated.

In this work, we quantify at run time, how hard the tracking of an object should be, not only based on its intrinsic complexity, but also based on its observed dynamics. An object that appears static in the recent temporal window requires less resources compared to an object whose state is more dynamic. Thus, change detection is performed on the image space of color intensities and depth measurements of the RGBD input. In more detail, for the next tracking frame, and the ith tracker, a value $\mathfrak{m}_i$ is computed, which takes the value of 0 if the corresponding tracked object appears to be relatively static, and takes the value of 1 otherwise. For $\mathfrak{m}_i$ to be 1, any of the following need to be true:

- the mean value of the pixel-wise differences between the observations of object i and the back-projection of the last estimated configuration of object i is high enough,
- the kinetic energy of object i, as computed from so far tracked velocities, suggests a moving object,
- the amount of missing depth measurements changes substantially, from one frame to the next, which might me attributed to change in the slant of object i, or
- any of the above was satisfied in the recent past (damping).

Budget has the trivial minimum value of 0. Moreover, as it has been shown in [3, 4], a PSO budget of 64 particles and 64 generations suffices to track a hand, even in interaction with another hand. No more budget should be required to track simpler structures such as rigid objects.

The proposed dynamic budget allocation policy assigns a minimum budget B_{min} of 64 particles running for 4 generations to the objects that are static and a maximum budget B_{max} that never exceeds the aforementioned budget of 64 particles, 64 generations. thus, the budget $B_i, i = 1,\ldots,N$ for all N trackers is set to B_{max} if $m_i = 1$ or to B_{min}, otherwise.



(a) Pouring pancake mix (b) Waiting for the mix to cook

Figure 1: When making pancakes in real-life, several objects remain static for considerable time intervals. The proposed method can very efficiently skip through these idle times without sacrificing tracking accuracy.

Dataset	No of Objects	Object Complexity	Idle time	fps (ECT/F.ECT)	Error (ECT/F.ECT)
Two hands	17	Mixed	Medium	2.06 / 4.19	2.99 / 3.07
Spray	2	Mixed	Small	4.74 / 6.16	3.80 / 3.98
Cans	13	Low	Large	2.19 / 51.13	1.55 / 1.92
Pancake	4	Low	Large	5.62 / 86.22	2.30 / 3.10

Table 1: Experimental results and performance indicators.

In order to evaluate the proposed method we conducted several experiments on real (e.g. Fig. 1) and synthetic datasets. A summary of the results is presented in Table 1. In brief, several scenarios with different dynamics (number of objects, object complexity and idle time) have been considered. In all scenarios, the proposed method outperformed ECT [2] in tracking throughput, noticeably or even remarkably, while maintaining the accuracy levels (negligible error increase). Especially in scenarios with large idle times (e.g. cans, pancake, in Table 1) the increase in tracking throughput was dramatic ($15\times - 23\times$ speedup). A video showing qualitative experimental results is available at `https://youtu.be/nPru6PpWrK4`.

By focusing computational resources to active (i.e., non static) objects, tracking performance is not any more a function of the number of the involved entities but rather a function of the complexity of the actual motion and the interaction of objects present in the scene. Therefore, the proposed approach makes it possible to handle, accurately and at interactive framerates, scenes and scenarios that the current state of the art can only handle in offline mode.

Acknowledgements: This work was partially supported by projects FP7-IP-288533 Robohow and FP7-ICT-2011-9 WEARHAP.

[1] Nikolaos Kyriazis and Antonis Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 9–16. IEEE, 2013.

[2] Nikolaos Kyriazis and Antonis Argyros. Scalable 3d tracking of multiple interacting objects. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3430–3437. IEEE, 2014.

[3] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, volume 1, page 3, 2011.

[4] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1862–1869. IEEE, 2012.

Real-time RGB-D Tracking with Depth Scaling Kernelised Correlation Filters and Occlusion Handling

Massimo Camplani
massimo.camplani@bristol.ac.uk
Sion Hannuna
sh1670@bristol.ac.uk
Majid Mirmehdi
majid@cs.bris.ac.uk
Dima Damen
dima.damen@bristol.ac.uk
Adeline Paiement
csatmp@bristol.ac.uk
Lili Tao
lili.tao@bristol.ac.uk
Tilo Burghardt
tb2935@bristol.ac.uk

Visual Information Laboratory
Faculty of Engineering
University of Bristol
Bristol, UK

The recent surge in popularity of real-time RGB-D sensors has encouraged research into combining colour and depth data for tracking. The results from a few, recent works in RGB-D tracking have demonstrated that state-of-the-art RGB tracking algorithms can be outperformed by approaches that fuse colour and depth, for example [1, 3, 4, 5].

In this paper, we propose a real-time RGB-D tracker which we refer to as the Depth Scaling Kernalised Correlations Filters (DS-KCF). It is based on, and improves upon, the RGB Kernelised Correlation Filters tracker (KCF) from [2]. KCF is based on the use of the 'kernel trick' to extend correlation filters for very fast RGB tracking. The KCF tracker has important characteristics, in particular its ability to combine high accuracy and processing speed as demonstrated in [2, 6]. It is based on a simple processing chain that comprises training, detection, retraining and model update obtained by linear interpolation. The key to KCF is that it exploits the properties of circulant matrices to achieve efficient learning by implicitly encoding convolution and by allowing to operate in the Fourier domain using mainly element wise operations.

The proposed DS-KCF tracker[1] extends the RGB KCF tracker in three ways: (i) we employ an the efficient combination of colour and depth features (ii) we propose an efficient a novel management of scale changes and (iii) occlusions handling. The improvements we implement provide higher rates of accuracy while still operating at better than real-time frame rates (35fps on average). In particular, depth data in the target region is segmented with a fast K-means approach to extract relevant features for the target's depth distribution. Modelled as a Gaussian distribution, this data allows to identify scale changes and efficiently model them in the Fourier domain. The advantage of the proposed approach is that only a single target model is kept and updated. Furthermore, region depth distribution enables the detection of possible occlusions identified as sudden changes in the target region's depth histogram, and recovering lost tracks by searching for the unoccluded object in specifically identified key areas. During an occlusion, the model is not updated and the occluding object is tracked to guide the target's search space.

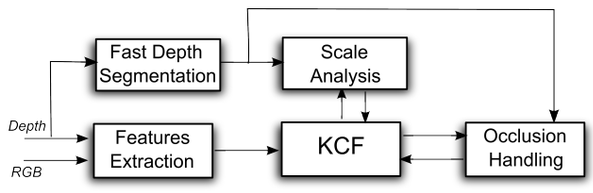


Figure 1: Block diagram of the proposed DS-KCF tracker.

Comparison on the Princeton Dataset [4] - We compare tracking performance by reporting the precision value for an error threshold equal to 20 pixels (P20), the area under the curve (AUC) of success plot measure, and the number of processed frames per second (fps). Table 1 shows that the proposed DS-KCF tracker outperforms the baseline KCF leading to better results both in terms of AUC and P20 measures. DS-KCF also outperforms the other two RGB-D trackers tested, Prin-Track [4] and OAPF [3]. Furthermore, the average processing rate in the Prin-Track (RGB-D) is 0.14fps and 0.9fps for the OAPF tracker in striking contrast to 40fps for DS-KCF. Example results of the trackers are shown in Figure 2.

Table 1: Trackers' performance on Princeton Validation Dataset [4].

	Average		
	AUC	**P20**	**fps**
KCF [2]	56.6	73.3	**103**
DS-KCF	**79.5**	94.2	40
Prin-Track(RGB-D) [4]	74.1	**96.0**	0.14
OAPF [3]	76.5	-	0.90

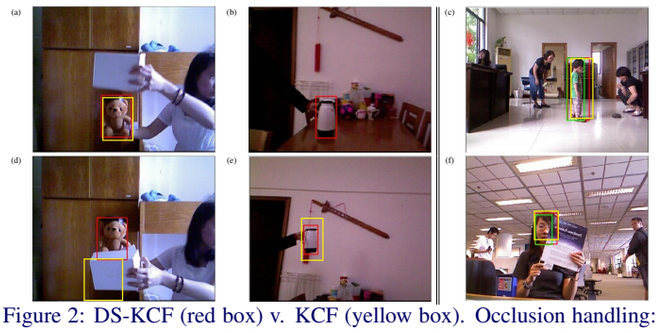


Figure 2: DS-KCF (red box) v. KCF (yellow box). Occlusion handling: (a) before occlusion, and (d) after occlusion. Change of scale: (b) initial target, and (e) target after change of scale. DS-KCF (red) v. Prin-Track(RGBD) (yellow) v. groundtruth (green) on (c) 'child_no1', and (f) 'face_occ5'.

[1] DS-KCF code is available at http://www.irc-sphere.ac.uk/work-package-2/DS-KCF

[1] G. García, D. Klein, J. Stückler, S. Frintrop, and A. Cremers. Adaptive multi-cue 3D tracking of arbitrary objects. In Pattern Recognition, pages 357–366. 2012.

[2] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2015.

[3] T. Meshgi, S. Maeda, S. Oba, H. Skibbe, Y. Li, and S. Ishii. Occlusion aware particle filter tracker to handle complex and persistent occlusions. Computer Vision and Image Understanding, 2015 to appear.

[4] S. Song and J. Xiao. Tracking revisited using RGBD camera: Unified benchmark and baselines. In ICCV, pages 233–240, 2013.

[5] Q. Wang, J. Fang, and Y. Yuan. Multi-cue based tracking. Neurocomputing, 131(0):227 – 236, 2014.

[6] Y. Wu, J. Lim, and M. Yang. Online Object Tracking: A Benchmark. In CVPR, pages 2411–2418, 2013.

Experimental Evaluation of the Bag-of-Features Model for Unsupervised Learning of Images

Mariana Afonso
marianafza@fe.up.pt
Department of Electrical Engineering,
University of Porto
Luis F. Teixeira
luisft@fe.up.pt
INESC TEC and Department of Computer Science,
University of Porto

The Bag-of-Features (BoF) [2] is a popular model that aims to represent images as an orderless collection of features without the use of any spatial information. Each image is represented by a frequency histogram of visual words from a codebook. Although the model is quite simple with regards to the implementation, there are several steps in which parameters and algorithms need to be chosen.

This work aimed to assess the performance of this model for the application of unsupervised learning for a set of images, also called image clustering. Additionally, it aims to provide valuable insight on the different steps of the model and to compare different algorithms in order to achieve the best performance for a given dataset. All the source code of this work is available open-source [1] at Github.

The applications of image clustering are endless and could include social network mining, more specifically for summarization of the huge amount of content shared everyday by millions of users.

In order to obtain the BoF representation of an image collection, many steps are required and are illustrated in Figure 1. The first one is the image description step, in which the input images are processed by first detecting keypoints or patches and then describing them using a certain algorithm. The next step is codebook learning, where a portion of the keypoints extracted from the images are clustered in order to obtain a codebook of visual words. This usually requires sampling of the total number of keypoints obtained from the images. The following step is the BoF representation of the images where each image is represented by a histogram of frequencies of visual words from the codebook obtained previously. The words are then filtered and the histograms are normalized following a chosen methodology. Finally, clustering is applied to the final representation of the images.

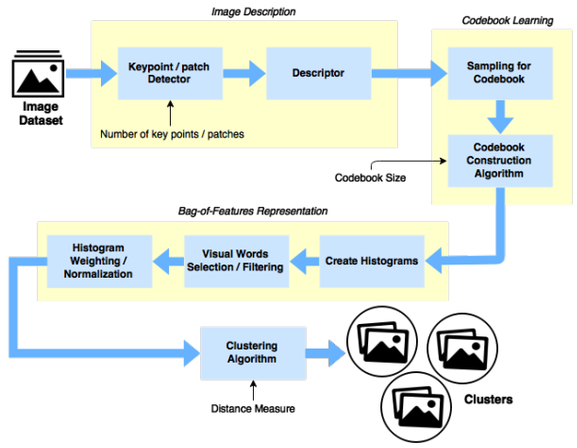


Figure 1: The process and the main steps of the Bag-of-Features model for the application of image clustering.

Due to the popularity of the BoF model, a number of works have been focused on evaluating its performance. For instance, in [1] the authors presented the results of an experimental study concerning the BoF model applied to the problem of image classification. Several key steps of the model were tested using different algorithms and parameters. Another empirical study presented in [3] evaluated the impact of applying techniques used in text categorization to the BoF model for the application of scene classification.

The main contributions of this study are: (1) the experimental analysis of the BoF model for image clustering, (2) the addition of a number of steps and algorithms (e.g. sampling the features for codebook learning and visual words selection) and (3) the proposal of a sampling technique for the features obtained from the images for the codebook learning procedure.

In terms of the experimental design, in each of the steps of the model, several parameters and algorithms were varied. Three datasets were used in order to obtain different levels of difficulty and complexity, an object dataset and two scene datasets. The performance measures used were the Normalized Mutual Information (NMI) and the Adjusted Rand Index (ARI). Both these metrics are popular choices for assessing clustering results and since they have different natures, a more complete evaluation was possible. All the tests were repeated 10 times to obtain the average and the standard deviation of the performance measures.

The results from the detectors and descriptors step indicate that the performance of the BoF model highly depends on the algorithm for the description of the images and is less influenced by the detector used (except for object datasets). Also, the choice of the best algorithms is dependent on the dataset.

Next, the influence of the average number of keypoints per image and the codebook size were tested varying those values for the three datasets. The results show that regardless of the codebook size used, the performance almost always increases with the average number of keypoints per image. Additionally, it could be observed that the ideal size of the codebook increases with the complexity of the dataset.

It was also found that the strategy for sampling the keypoints, the codebook learning algorithm, the visual words filtering and the normalization and weighting of the histograms were steps that did not influence the performance of the model significantly, and therefore, do not require much tunning of the parameters and algorithms used.

Lastly, different clustering algorithms were tested for the final clustering of the histograms and the conclusion of this last step is that, although an algorithm which does not require the number of clusters is desirable, they perform much worse and require several parameters that need to be adjusted, which can be very specific to a given set of images and settings.

The final performance of the image clustering task for the three datasets using the BoF model after the tunning of the parameters and algorithms can be found in Table 1.

Table 1: Final performance of the BoF model for the three datasets tested.

Dataset	Avg. ARI	Avg. NMI
Coil-20 (object - low difficulty)	**67,4%**	**84,8%**
Natural and Urban (scene - medium difficulty)	**30,2%**	**40,6%**
Event (scene - high difficulty)	**19,4%**	**27,4%**

From all the different experiences developed and presented in this work, it can be concluded that although the Bag-of-Features model can be successfully applied to the problem of unsupervised learning images, it provides a poor representation of the images when the datasets represent complex scenes and requires fine tunning of the algorithms and parameters used in each step. For this reason, more advanced techniques are required in order to be able to effectively extract information from large image collections in an unsupervised fashion.

[1] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Computer Vision–ECCV 2006*, pages 490–503. Springer, 2006.

[2] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[3] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.

[1] https://github.com/marianaAfonso/BOFClustering

Indoor Localisation with Regression Networks and Place Cell Models

Jose Rivera-Rubio[1]
http://www.bicv.org
Ioannis Alexiou[2]

Anil A. Bharath[1]

[1] Biologically Inspired Computer Vision Group
Imperial College London
London, UK

[2] Computer Vision Research Group
Queen Mary University of London
London, UK

One of the key behaviours found in biological place cells (BPCs) is a rate-coding effect: a neuron's rate of firing decreases with distance from some landmark location. We used visual information from wearable and hand-held cameras in order to reproduce this rate-coding effect in artificial place cells (APCs). The accuracy of localisation using these APCs was evaluated using different visual descriptors and different place cell widths. Simple localisation using APCs was feasible by noting which APC yielded the maximum response. We also propose using joint position coding using a number of automatically defined APCs. Using both these approaches, we were able to demonstrate good self-localisation from very small images taken in indoor settings. Average localisation performance is favourable when compared with ground-truth and LSD-SLAM; even without the use of a motion model, errors using a single device were as low as 2 m for some journeys and corridors.

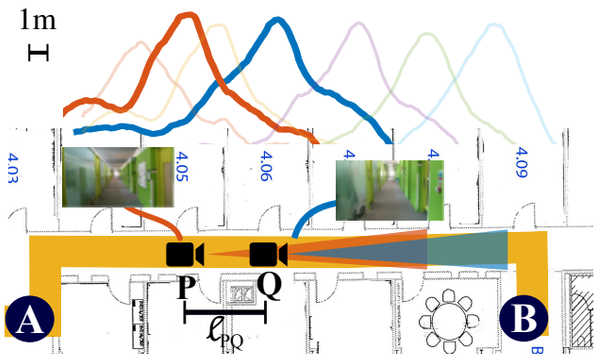


Figure 1: The concept of artificial place cells (APCs) for navigation. APC behaviour is produced by comparing FEVs from a series of frames to those at several landmark locations.

Given a series of video frames extracted from footage recorded during indoor navigation, we made use of the visual path concept [2, 3], to perform matching between locations of a physical environment being traversed and a database of views captured with two different devices, and at different moments in time. The similarity scores obtained from appearance-based comparison methods, applied between sequences of frames of a journey and these virtual landmarks, exhibit a behaviour (Fig. 1(a)) that is similar to those recorded in mammalian place cells.

We compared the accuracy of several patch-level descriptors for this purpose [3]. Patch descriptors were L_2 normalized, vector quantised and then frames were encoded into a 400-element Frame-Encoding Vector (FEV). For the comparison with a state-of-the-art SLAM method, we chose Engel's LSD-SLAM [1].

In order to model place cell behaviour, we need to map pairs of frames onto a scalar value that is analogous to a firing rate. One way to mimic the behaviour of BPCs within APCs is to introduce a *kernel* function that maps a pair of FEVs onto a positive scalar value. The following mapping between two vectors $\mathbf{v}_a$ and $\mathbf{v}_b$ maps the two FEVs onto a scalar value that takes a maximum when the two vectors are identical:

$$\kappa_{\chi^2}(\mathbf{v}_a, \mathbf{v}_b) = \sum_{j=1}^{400} \frac{v_a(j) \cdot v_b(j)}{v_a(j) + v_b(j)} \tag{1}$$

To model the response of a place cell to an image stimulus with respect to some reference location, ℓ_i, the result of the κ_{χ^2} function is thresholded. The FEV, $\mathbf{v}_{r_i}$, is first constructed when the camera is at position ℓ_i from one or more reference journeys and calculating a supra-threshold response to some frame $\mathbf{v}_\ell$ acquired at location ℓ. As ℓ is varied,

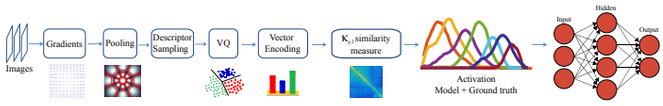


Figure 2: Overview of the training pipeline. The diagram of the neural network is merely illustrative, and does not represent the real GRNN architecture used.

$\kappa_{\chi^2}(\mathbf{v}_\ell, \mathbf{v}_{r_i})$ changes accordingly. Thresholds, T_i, relative to each response curve, are used to create a set of supra-thresholded response curves, $r_i(\ell)$:

$$r_i(\ell) = U(\kappa_{\chi^2}(\mathbf{v}_\ell, \mathbf{v}_{r_i}) - T_i) \cdot \kappa_{\chi^2}(\mathbf{v}_\ell, \mathbf{v}_{r_i}) \tag{2}$$

where $U(\cdot)$ represents the unit step function. Curves acquired by averaging responses from several journeys with respect to the same APC location may be referred to as an APC *tuning curve*.

Given a series of APC responses to visual cues of a person's location along some journey, there are two obvious ways of estimating location, ℓ. The first is simply to use the APC which displays maximum activation (firing rate) as a rough indicator of where the person is. The second technique achieves more accurate localisation of a camera from its captured visual data by using the joint distribution $p(\mathbf{r}|\ell)$, of APC responses, $\mathbf{r}$ to infer location ℓ relative to some designated ground truth. We use a single index, i, to refer to the response, r_i, of a unique place cell. Given several active cells that are a subset of all place cells in a location, sub-APC localisation is possible using APC responses from previous journeys using empirical Bayes' techniques. For example, if three cells are active, the chain rule can be used to obtain successively refined estimates of ℓ:

$$\begin{aligned} p(\ell|\mathbf{r}) &\propto p(r_3, r_4, r_5|\ell)p(\ell) \\ &\propto p(r_3|r_4, r_5, \ell) \times p(r_4|r_5, \ell) \times p(r_5|\ell) \times p(\ell) \end{aligned} \tag{3}$$

so that the responses of spatially close APCs can be used to infer sub-APC position. If the width of an APC is set to around 2 m, localisation of the order of tens of centimetres is plausible.

A Generalized Regression Neural Network (GRNN) was used to provide sub-APC position estimates, obviating the need to construct *ad-hoc* empirical estimators. This regression network consists of two-layers, and uses radial-basis functions. The responses from 16 place cells were input to the network, and ground truth of location within a section of corridor – up to 4 m long – used to train it as a regressor. In all experiments, dictionary generation was performed independently of the APC responses used in training the regression network.

In conclusion, this work demonstrates that computational models of place cells can provide effective estimates of camera location without relying on tracking or geometric models of the local environment. Such techniques, although simple, have achieved errors that range from tens of cm to few metres, matching and certainly complementing the more sophisticated position inference approaches used in computer vision.

[1] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Computer Vision–ECCV 2014*, pages 834–849. Springer, 2014.

[2] Yoshio Matsumoto, Masayuki Inaba, and Hirochika Inoue. Visual navigation using view-sequenced route representation. In *Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on*, volume 1, pages 83–88. IEEE, 1996.

[3] Jose Rivera-Rubio, Ioannis Alexiou, Luke Dickens, Riccardo Secoli, Emil Lupu, and Anil A Bharath. Associating locations from wearable cameras. In *Proceedings of the British Machine Vision Conference*, 2014.

Kinship Verification with Deep Convolutional Neural Networks

Kaihao Zhang[12]
kaihao.zhang@nlpr.ia.ac.cn
Yongzhen Huang[2]
yzhuang@nlpr.ia.ac.cn
Hong Wu[1]
hwu@uestc.edu.cn
Liang Wang[2]
wangliang@nlpr.ia.ac.cn

[1] Statistical Machine Intelligence & LEarning (SMILE)
Big Data Research Center
University of Electronic Science and Technology of China
Chengdu, China

[2] National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Kinship verification from facial images is an interesting and challenging problem. The current algorithms on this topic typically represent faces with multiple low-level features, followed by a shallow learning model. In this paper, we propose to extract high-level features for kinship verification based on deep convolutional neural networks. Our method is end-to-end, without complex pre-processing often used in traditional methods. The high-level features are produced from the neuron activations of the last hidden layer, and then fed into a soft-max classifier to verify the kinship of two persons.

We first propose a basic structure of CNN (CNN-Basic) contains three convolutional layers, followed by a fully-connected layer and a soft-max layer. As shown in Figure 1, the input is a pair of 64×64 images with three channels (RGB). Following the input, the first convolutional layer is generated after convolving the input via 16 filters with a stride of 1. Each filter is with the size $5 \times 5 \times 6$. The second convolutional layer filters the input of the previous layer with 64 kernels of size $5 \times 5 \times 16$. The third convolutional layer contains 128 kernels of the size $5 \times 5 \times 64$. After the convolutional layers, a fully-connected layer projects the extracted features into a subspace with 640 neurons. Max-pooling layers follow the first and second convolutional layers. Finally, this network is trained via a two-way soft-max classifier at the top layer.

We adopt the ReLU function [1] as the activation function of the convolution layers, which has been shown to achieve better performance than the sigmoid function. With ReLU, the convolution operation is formulated as

$$y^{j(r)} = \max\left(0, b^{j(r)} + \sum_i w^{ij(r)} * x^{i(r)}\right), \tag{1}$$

where x^i and y^j are the i-th input map and the j-th output map, respectively. w^{ij} denotes the weight between the i-th input map and the j-th output map. b^j is the bias of the j-th output map, and $\times$ denotes the convolutional operation.

We choose max-pooling with a neighboring region size of 2×2. Max pooling is helpful to increase the translation invariance and avoid overfitting, which is defined as

$$y^i_{j,k} = \max_{0\le m,n\le s}\left\{x^i_{j\cdot s+m,k\cdot s+n}\right\}, \tag{2}$$

where $y^i_{j,k}$ denotes the outputs of the i-th feature map in the location of (j,k). Similarly, $x^i_{j,k}$ denotes the value of location (j,k) in the i-th feature map.

The CNN is trained by back-propagation with logistic loss over the predicted scores using the soft-max function. To initialize weights, we use a Gaussian distribution with zero mean and a standard deviation of 0.01. The biases are initialized as zeros. In each iteration, we update all the weights after learning the mini-batch with the size of 128. In all layers, the momentum is set as 0.9 and the weight decay is set as 0.005. To expand the training set, we also randomly flip images during training.

When a subject is demanded to verify the kinship from two face images, it is highly possible that the key-points are focused, such as their eyes, mouth and nose. We consider that the facial key-points have a significant impact on kinship analysis, and thus design a key-points-based feature representation for kinship verification. In particular, we detect the centers of two eyes, the corners of the mouth and the nose with a facial point detection algorithm [2]. Then each face image is cropped and aligned according to the five key-points. To extract more complementary information, we also crop other five face regions without key-points detection. The five images are the original image and its four local regions,

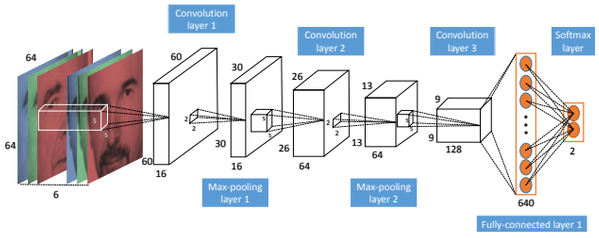


Figure 1: The proposed architecture of basic CNN for kinship verification. For all layers, the length of each cuboid is the map number, and the width and height of each cuboid are the dimension of each map.

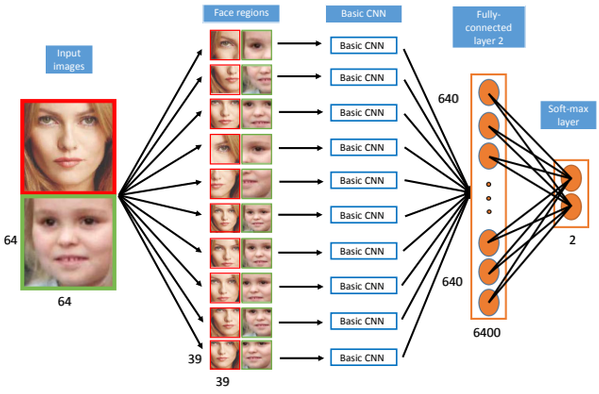


Figure 2: Overview of the proposed CNN-Points structure for kinship verification. The input is a pair of RGB images, which are cropped into ten face regions and fed into different basic CNNs.

i.e., the top-left corner, the top-right corner, bottom-left corner, bottom-right corner.

In order to improve kinship verification with these face regions, we propose a new structure (CNN-Points) which is shown in Figure 2. The new structure contains 10 basic CNNs (see Figure 1), each of which receives a pair of face regions. Ten sets of 640-dimensional features are produced from the last hidden layer of the basic CNNs. The last hidden layer of the CNN-Points is fully-connected to the ten basic CNNs, which is defined as

$$y^i = f\left(\sum_{k=1}^{10}\sum_{j=1}^{640} w^i_{j,k} * x_{j,k} + b^i\right), \tag{3}$$

where y^j is the output of the i-th neuron activation, $w^i_{j,k}$ denotes the weight between the input features and the i-th neuron, and $f(\cdot)$ is chosen to be the sigmoid function. The final representation is fed into a soft-max classifier to predict the kinship of two persons.

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[2] Y. Sun, X. L. Wang, and X. O. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

Exploiting Low-rank Structure for Discriminative Sub-categorization

Zheng Xu[1]
xuzhustc@gmail.com
Xue Li[1]
echolixue@gmail.com
Kuiyuan Yang[2]
kuyang@microsoft.com
Tom Goldstein[1]
tomg@cs.umd.edu

[1] Department of Computer Science, University of Maryland, College Park, USA
[2] Microsoft Research, Beijing, China

In visual recognition, sub-categorization, which divides a category into some sub-categories, has been proposed to deal with large intra-class variance in the real world. Recent discriminant sub-categorization approaches utilize samples that do not belong to the category under consideration as negative data for supervision, and cluster positive samples of the category into sub-categories, then simultaneously train the corresponding classifier for each sub-category [2, 4]. In the jointly clustering and classification framework of previous methods, the classifier for each sub-category is trained by using samples hard-assigned to the sub-category. However, some samples would contribute to the training of several sub-categories since the intra-variance of a category is caused by complex factors. Moreover, sub-categories are closely related since they are discovered from the same category, and the common information among these sub-categories is beneficial for classifier training.

We propose a new approach for discriminative sub-categorization, which adopts the exemplar based method to address the intra-variance in category, and exploits the low rank structure to preserve common information while discovering sub-categories. Our approach builds up the exemplar-LDAs [3], which generates a set of exemplar classifiers with each classifier trained by a single positive sample and all the negative samples. The extreme case of sub-category is to have only one positive sample, which is a compact set for training and modeling. We adopt exemplar classifiers to represent the compact sub-categories and preserve intra-variance in a category. In order to share common information among exemplar classifiers while preserving diversity, we jointly train the exemplar-LDAs for all the positive samples and introduce the trace-norm regularizer on the matrix of weights, as we assume the weights lie on a union of subspaces such that the matrix of weights is low-rank.

We formulate the proposed low-rank least squares exemplar-LDAs (LRLSE-LDAs) as follows. Let $\mathbf{X}_1 = [\mathbf{x}_1^+, \ldots, \mathbf{x}_n^+]$ and $\mathbf{X}_2 = [\mathbf{x}_1^-, \ldots, \mathbf{x}_m^-]$ denote the centered data matrix[1] for positive samples and negative samples, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$ denote the weight matrix where each w_i is the weight vector of exemplar-LDA for a positive sample. The objective function for training the exemplar-LDAs of positive samples together is

$$J_{LSE-LDAs}(\mathbf{W}) = \frac{\delta}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\mathbf{X}_2'\mathbf{W}\|_F^2 - trace(\mathbf{X}_1'\mathbf{W}) \quad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, $trace()$ represents the trace of a matrix. We minimize the least squares form in Eq. 1 instead of maximizing the Fisher criterion so that the objective function is convex, inspired by [6]. Eq. 1 has closed-form solution as

$$\mathbf{W} = (\mathbf{X}_2\mathbf{X}_2' + \delta\mathbf{I})^{-1}\mathbf{X}_1 \quad (2)$$

where $\mathbf{I}$ is the identity matrix. To discover the structure of sub-categories, we jointly learn the weight for positive samples/exemplars of the category and regularize the weight matrix with a low-rank constraint. Finally, we arrive at the objective function of LRLSE-LDAs,

$$J_{LRLSE-LDAs}(\mathbf{W}) = \xi\|\mathbf{W}\|_* + J_{LSE-LDAs}(\mathbf{W}) \quad (3)$$

$\|\cdot\|_*$ is the trace norm used to regularize the weight matrix, which is a convex approximation of the rank of a matrix

To solve the convex formulation in Eq. 3, we propose an efficient algorithm based on the scaled form of alternating direction method of multipliers (scaled ADMM) [1]. We reformulate minimizing $J_{LRLSE-LDAs}(\mathbf{W})$ in Eq. 3 as an equality-constrained convex optimization problem by introducing an intermediate variable $\mathbf{F}$,

$$\min_{\mathbf{W},\mathbf{F}} J_{LSE-LDAs}(\mathbf{W}) + \xi\|\mathbf{F}\|_* \quad \text{s.t. } \mathbf{W} = \mathbf{F} \quad (4)$$

The augmented Lagrangian for the formulation in Eq. 4 can be written as:

$$L(\mathbf{W},\mathbf{F},\mathbf{\Lambda}) = J_{LSE-LDAs}(\mathbf{W}) + \xi\|\mathbf{F}\|_* + \frac{\tau}{2}(\|\mathbf{W}-\mathbf{F}+\mathbf{\Lambda}\|_F^2 - \|\mathbf{\Lambda}\|_F^2) \quad (5)$$

where $\mathbf{\Lambda}$ is the scaled dual parameter matrix, and τ is the penalty parameter. We iteratively update variables $\mathbf{W},\mathbf{F},\mathbf{\Lambda}$ as in scaled ADMM, where $\mathbf{W},\mathbf{F}$ are updated by solving two subproblems both with closed-form solutions, and $\mathbf{\Lambda}$ is updated by dual ascent. The two subproblems are

$$\mathbf{W} = \arg\min_{\mathbf{W}} J_{LSE-LDAs}(\mathbf{W}) + \frac{\tau}{2}\|\mathbf{W}-\mathbf{F}+\mathbf{\Lambda}\|_F^2 \quad (6)$$

$$\mathbf{F} = \arg\min_{\mathbf{F}} \xi\|\mathbf{F}\|_* + \frac{\tau}{2}\|\mathbf{W}-\mathbf{F}+\mathbf{\Lambda}\|_F^2 \quad (7)$$

where Eq. 6 has a closed-form solution benefits from the least squares form and Eq. 7 can be solved by singular value thresholding method.

After training the weights of LRLSE-LDAs, we utilize those exemplar classifiers to perform sub-category discovery and visual recognition. For sub-category discovery, we adopt spectral clustering with affinity matrix defined by the prediction scores on positive samples. For visual recognition, we adopt the cross domain recognition approach in [5] by fusing the top-K prediction scores from trained exemplar classifiers.

We conduct comprehensive experiments on various datasets to validate the effectiveness and efficiency of our approach in sub-category discovery and visual recognition. We follow the experimental setting in [4] to evaluate the performance of sub-category discovery. We conduct experiments on ten public datasets from the UCI repository and MNIST, which cover a large variant types of data. LRLSE-LDAs based clustering achieves promising results measured by purity on those datasets. We follow the experimental setting in [5] to evaluate the performance of visual recognition. We use the Office-Caltech dataset for object recognition and the IXMAS dataset for action recognition. LRLSE-LDAs based classification achieves order-of-magnitude speedup with matching performance comparing with state-of-the art in [5].

[1] Data matrix is centered by subtracting the mean of training samples from each sample. We use mean of negative samples to approximate the mean of all negative sample and a positive sample for each exemplar classifier.

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[3] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*. 2012.

[4] Minh Hoai and Andrew Zisserman. Discriminative sub-categorization. In *CVPR*, 2013.

[5] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014.

[6] Jieping Ye. Least squares linear discriminant analysis. In *ICML*, 2007.

Sparse 3D convolutional neural networks

Ben Graham
b.graham@warwick.ac.uk

Department of Statistics
University of Warwick
CV4 7AL, UK

Convolutional neural networks (CNNs) are powerful tools for understanding data with spatial structure such as photos. They are most commonly used in two dimensions, but they can also be applied more generally. One-dimensional CNNs are used for processing time-series such as human speech. Three dimensional CNNs have been used to analyze movement in 2+1 dimensional space-time [2] and for helping drones find a safe place to land [3]. Three dimensional convolutional deep belief networks have been used to recognize objects in 2.5D depth maps [4].

In [1], a *sparse* two-dimensional CNN is implemented to perform Chinese handwriting recognition. When a handwritten character is rendered at moderately high resolution on a two dimensional grid, it looks like a sparse matrix. If we only calculate the hidden units of the CNN that can actually *see* some part of the input field the pen has visited, the workload decreases.

Sparsity is a useful optimization in two dimensions, and it is potentially even more useful in three or higher dimensions. This is related to the *curse of dimensionality*; an $N \times N \times N$ cubic grid contains many more points than an $N \times N$ square grid. We have adapted the algorithm from [1] to implement sparse CNNs on range of different graphs. CUDA GPU code for running sparse 2, 3 and 4 dimensional CNNs is available at:

`https://github.com/btgraham/SparseConvNet`

The world we live in is three dimensional, and time can also be thought of as an extra dimension, so there are a large number of possible applications for three and even four dimensional CNNs. Figure 1 shows what happens to sparse 3D data as it passes though a CNN. In this paper I apply CNNs to a variety of sparse 3D datasets.

When applying CNNs to sparse data, it may be better to use small convolutional filters, as they do a better job of preserving sparsity in the computationally expensive lower layers of the network. To reduce the size of the filters, we have experimented with changing the underlying graph. See Figure 2.

Figure 3 shows a variety of objects from the SHREC2015 Non-rigid 3D Shape Retrieval dataset, each stored as a mesh of triangles in the OFF-file format. The dataset contains 1200 exemplars split evenly between 50 classes (aliens, ants, armadillo, ...). The dataset was intended to be used for unsupervised learning, but as CNNs are most often used for supervised learning, we used 6-fold cross-validation to measure the ability of our 3D CNNs to learn shapes. To stop the dataset being too easy, we randomly rotated the objects during training and testing. This is to force the CNN to truly learn to recognize shape, and not rely on some classes of objects tending to have a certain orientation. We tested a variety of network architectures to explore the trade-off between speed and accuracy.

Figure 4 shows an image from the Recognizing Human Actions video dataset. Taking differences between successive frame converts the dataset to a collection of sparse 2+1 dimensional objects. We also experimented with the more complicated UCF101 video dataset.

[1] Ben Graham. Spatially-sparse convolutional neural networks. 2014. URL `http://arxiv.org/abs/1409.6070`.

[2] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, January 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.59. URL `http://dx.doi.org/10.1109/TPAMI.2012.59`.

[3] D. Maturana and S. Scherer. 3D Convolutional Neural Networks for Landing Zone Detection from LiDAR. In *ICRA*, 2015.

[4] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

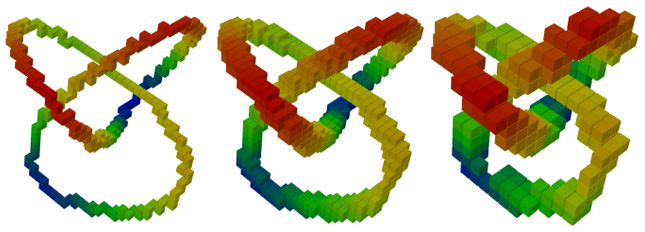


Figure 1: Left to right: A trefoil knot has been drawn in the cubic lattice; these are the input layer's active sites. Applying a $2 \times 2 \times 2$ convolution, the number of active (i.e. non-zero) sites increases. Applying a $2 \times 2 \times 2$ pooling operation reduces the scale, which tends to decrease the number of active sites.

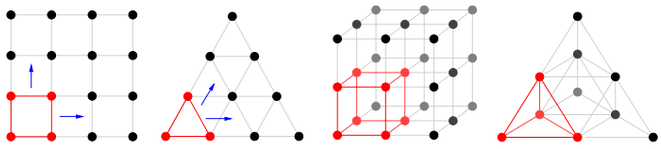


Figure 2: Convolutional filter shapes for different lattices: (i) A 4×4 square grid with a 2×2 convolutional filter. (ii) A triangular grid with size 4, and a triangular filter with size 2. (iii) A $3 \times 3 \times 3$ cubic grid, and a $2 \times 2 \times 2$ filter. (iv) A tetrahedral grid with size 3, and a filter of size 2.

In d-dimensions, the smallest practical filter for the triangular/tetrhedral/hypertetrahedral lattice contains $d+1$ vertices, while the smallest practical filter for the square/cubic/hypercubic lattice contains 2^d vertices. Thus, the higher the dimension, the greater the potential benefit in terms of efficiency from considering non-square lattices.

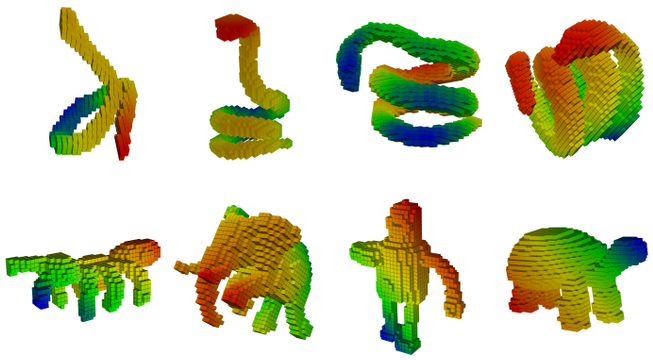


Figure 3: Items from a 3D object dataset, embedded into a $40 \times 40 \times 40$ cubic grid. Top row: four items from the snake class. Bottom row: an ant, an elephant, a robot and a tortoise.

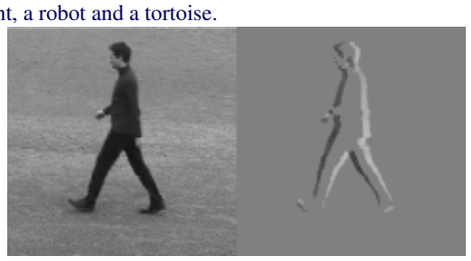


Figure 4: Left: a frame from a video from the Recognizing Human Actions video dataset. Right: the difference between that frame and the previous frame, with thresholding applied to increase sparsity. Stacking the frame differences produces a sparse 2+1 dimensional space-time object.

Multi-task Gaussian Process Regression-based Image Super Resolution

Xinwei Jiang
xinwei.jiang@ia.ac.cn
Jie Yang
yangjie@nlpr.ia.ac.cn
Lei Ma
lei.ma@ia.ac.cn
Yiping Yang
yiping.yang@ia.ac.cn

Institute of Automation
Chinese Academy of Sciences
Beijing, China

Image super resolution (SR) aims at recovering the missing high frequency details from single image or multiple images. Existing SR methods can be divided into three categories: interpolation-based, reconstruction-based and example learning-based. Our paper focuses on the third category. Example learning-based SR methods [6] utilize the LR-HR image pair to infer the missing high-frequency details in the LR image and achieve state-of-the-art performance.Recently, in the field of example learning-based SR, more and more researchers resort to learn the LR-HR relationship directly, i.e. $y = f(x)$, where x is the input LR image feature, y is the targeted HR image and f is the mapping function that transforms the LR feature into HR image. Instead of commonly used parametric models, non-parametric methods [3], especially gaussian process regression (GPR)-related methods [2, 4, 5] begin to emerge in the SR field. However, previous GPR-based SR methods simply learn all the GPR models independently and ignore the correlation between them. On the other hand, each pixel prediction can be treated as a task, so that inferring a HR patch can be regarded as a multi-task problem. In this paper, we focus on the multi-task gaussian process (MTGP) regression and apply it to super-resolution problem. We first give a brief overview of MTGP proposed in [1]. Then we study how SR problem corresponds to MTGP and propose the multi-task gaussian process super-resolution (MTGPSR) framework.

MTGP tries to solve the following problem: Given N distinct inputs $x_1,...,x_N$ we define the complete set of responses for M tasks as $\mathbf{y} = (y_{11},...,y_{N1},...,y_{12},...,y_{N2},...,y_{1M},...,y_{NM})^T$, where y_{ij} is the response for the jth task on the ith input x_i. We also denote the $N \times M$ matrix Y such that $\mathbf{y} = \text{vec}Y$. Given a set of observations $\mathbf{y_o}$, which is a subset of $\mathbf{y}$, we wish to predict the unobserved values of $\mathbf{y_u}$ of some input points for some tasks. MTGP wishes to learn M related latent functions $\{f_l\}$ by placing a GP prior over $\{f_l\}$ and directly induce correlations between tasks. Assuming that the GPs have zero mean we define

$$\langle f_l(x) f_k(x') \rangle = K^f_{lk} k^x(x,x') \tag{1}$$

$$y_{ij} \sim N(f_l(x_i), \sigma_j^2) \tag{2}$$

where K^f is a positive semi-definite (**PSD**) matrix that specifies the inter-task similarities, k^x is a covariance function over inputs, and σ_j^2 is the noise variance for the jth task. The key property of multi-task gaussian process model is the introduction of inter-task correlation matrix K^f, so that observations of one task can affect the predictions on another task.

Inference in the MTGP model can be carried out by using the standard GP formulation for the predictive mean and variance. The predictive mean on a test point x_* for task j is obtained by

$$f_j(x_*) = (k_j^f \otimes k_*^x)^T \Sigma^{-1} y, where \Sigma = K^f \otimes K^x + D \otimes I \tag{3}$$

where $\otimes$ denotes the Kronecker product, k_j^f is the jth column of K^f, k_*^x represents the vector consisting of covariances between the test point x_* and all the training points, K^x stands for the covariance matrix obtained by computing covariances between all the training points pairs, D is an $M \times M$ diagonal matrix with σ_j^2 in the $(j,j)^{th}$ position, and Σ is an $MN \times MN$ matrix.

In the learning stage, the MTGP hyper-parameters are optimized as follows: Let f be the vector of function values corresponding to y, and similarity for F for Y. Further, let $y_{\cdot j}$ denote the vector $(y_{1j},...,y_{Nj})^T$ and similarly for $f_{\cdot j}$. Given the missing data, which in this case is f, the complete-data log-likelihood is

$$\begin{aligned} L_{comp} = &-\frac{N}{2} log|K^f| - \frac{M}{2} log|K^x| - \frac{1}{2} tr\left[(K^f)^{-1} F^T (K^x)^{-1} F\right] \\ &-\frac{N}{2} \sum_{j=1}^{M} log\sigma_l^2 - \frac{1}{2} tr[(Y-F) D^{-1} (Y-F)^T] - \frac{MN}{2} log 2\pi \end{aligned} \tag{4}$$

from which we have following updates:

$$\hat{l}_x = arg\min_{l_x} \left(Nlog|\langle F^T (K^x(l_x))^{-1} F \rangle| + Mlog|K^x(l_x)| \right) \tag{5}$$

$$\hat{K}^f = N^{-1} \langle F^T (K^x(l_x))^{-1} F \rangle \tag{6}$$

$$\hat{\sigma}_j^2 = N^{-1} \langle (y_{\cdot j} - f_{\cdot j})^T (y_{\cdot j} - f_{\cdot j}) \rangle \tag{7}$$

where the expectations $\langle \cdot \rangle$ are taken with respect to $p(f|y_o, l_x, K^f)$, and $\hat{}$ denotes the updated parameters.

Processing pipeline of our method is as follows:in the training stage, we first construct HR/LR patch pairs by downsampling and then conduct K-means clustering on LR patch dataset. For each cluster, we learn one MTGP model to fit the training data. In the predicting stage, given an LR image, we first overlap sample the image getting test patch dataset and classify the data using the K-NN algorithm based on the cluster centers obtained from training stage. Next, each HR patch subspace corresponding to the LR patch subspace are recovered through learned MTGP regression. Finally, all the predicted patches are reconstructed into a HR image using average weighting scheme. Details of algorithm and experiment are presented in the paper.

In this paper we propose a super resolution framework based on the multi-task gaussian process regression and study how the models are optimized and inferred effectively. The proposed MTGPSR framework makes the pixel prediction correlation into consideration based on the image local structure. Experimental results show that the proposed algorithm achieves the comparative performance and makes the super-resolved image more accurate and natural.

[1] Edwin V Bonilla, Kian M. Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems 20*, pages 153–160. 2008.

[2] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 449–456, 2011.

[3] Li He, Hairong Qi, and Russell Zaretzki. Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 345–352, 2013.

[4] Younghee Kwon, Kwang In Kim, Jin Hyung Kim, and Christian Theobalt. Efficient learning-based image enhancement: Application to super-resolution and compression artifact removal. In *Proceedings of the British Machine Vision Conference(BMVC)*, pages 1–12, 2012.

[5] Jianmin Li, Yanyun Qu, Cuihua Li, Yuan Xie, Yang Wu, and Jianping Fan. Learning local gaussian process regression for image super-resolution. *Neurocomputing*, 154(0):284–295, 2015.

[6] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

Spatiotemporal Stereo Matching with 3D Disparity Profiles

Yongho Shin
yongho@gist.ac.kr
Kuk-Jin Yoon
kjyoon@gist.ac.kr

Computer Vision Laboratory
Gwangju Institute of Science and
Technology (GIST), South Korea

Adaptive support weights and over-parameterized disparity estimation truly improve the accuracy of stereo matching by enabling window-based similarity measures to handle depth discontinuities and non-fronto-parallel surfaces more effectively. Nevertheless, a disparity map sequence obtained in a frame-by-frame manner still tends to be inconsistent even with the use of state-of-the-art stereo matching methods. To solve this inconsistency problem, we propose a window-based spatiotemporal stereo matching method exploiting 3D disparity profiles.

Energy modeling and minimization For a rectified stereo image sequence, we employ an energy minimization scheme for each frame. Our energy function is defined by an over-parameterized disparity $\mathbf{h} = \begin{bmatrix} z & n_{xz} & n_{yz} \end{bmatrix}$ and a guidance disparity $\tilde{\mathbf{h}} = \begin{bmatrix} \tilde{z} & \tilde{n}_{xz} & \tilde{n}_{yz} \end{bmatrix}^\top$ as

$$E(\mathbf{H}, \tilde{\mathbf{H}}) = \sum_{p \in P} D_p(\mathbf{h}_p) + \sum_{p \in P} \sum_{q \in N_p} V_{p,q}(\mathbf{h}_p, \mathbf{h}_q) + \sum_{p \in P} T_p(\mathbf{h}_p, \tilde{\mathbf{h}}_p). \quad (1)$$

D_p is the data cost, and it is computed by the combination of census transform and mutual information. $V_{p,q}$ is the smoothness cost that encourages a smooth disparity map. T_p is the temporal cost that encourages consistent disparity values between frames, and it is defined with guidance disparity and normal values, $\tilde{z}_p$, $\tilde{n}_{xz,p}$ and $\tilde{n}_{yz,p}$, as

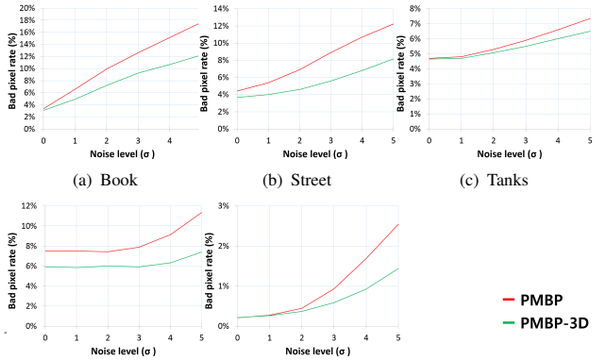
$$T_p(\mathbf{h}_p, \tilde{\mathbf{h}}_p) = \left[\min\left(\left| z_p - \tilde{z}_p \right|, \tau_z \right) + \min\left(\rho \cdot \left(\left| n_{xz,p} - \tilde{n}_{xz,p} \right| + \left| n_{yz,p} - \tilde{n}_{yz,p} \right| \right), \tau_n \right) \right]. \quad (2)$$

Ideally, $\tilde{z}_p$, $\tilde{n}_{xz,p}$ and $\tilde{n}_{yz,p}$ should be ground truth values. However, since ground truth values are not given, we determine guidance values by exploiting a 3D disparity profile which is a structure representing disparities and normal vectors of a window over multiple frames. In general, objects and cameras in a scene can be assumed to move with smooth variations in direction and velocity. In this case, the window shape and location must also be changed smoothly. It means that smoothly varying disparity and normal values between frames are more reliable than largely varying disparity and normal values. Hence, we set values that can make disparity and normal values between consecutive frames smoothly varied as guidance values. To obtain such guidance values, we compose a disparity profile, and fit the profile to a polynomial function.

We find the optimal solution to Eq.(1) in the PatchMatch Belief Propagation (PMBP) framework. PMBP starts with the initialization of particles with random sampling. Afterwards, in each iteration, particles are propagated from neighborhoods q to pixel p. Our propagation step consists of spatial, view, and temporal propagation. If the disbelief of a propagated state is lower than that of current states of p, the propagated state is selected as a new member of the particle set of p. After the propagation step, a random search around the state of p is performed for state refinement as in [2]. Before the next iteration, we update data with an intermediate disparity map sequence having the minimum energy in the current iteration. First, we update optical flows. Next, we update joint histograms for mutual information. Lastly, we update guidance values by fitting the disparity profile from the intermediate disparity map sequence. After a few iterations, the particle with the minimum disbelief is selected as the optimal solution of p.

Temporal correspondence establishment To obtain the 3D disparity profile, we must establish the temporal correspondences across frames. To this end, we employ the LDOF method [3] to handle large displacements by rapidly moving object or cameras. And, a forward-backward optical flow consistency check is used for filtering out incorrect flows. However, the consistency check makes the profile fragmentary. To resolve this problem, we propose an optical flow transfer method for establishing more reliable disparity profile by using optical flows of both views.

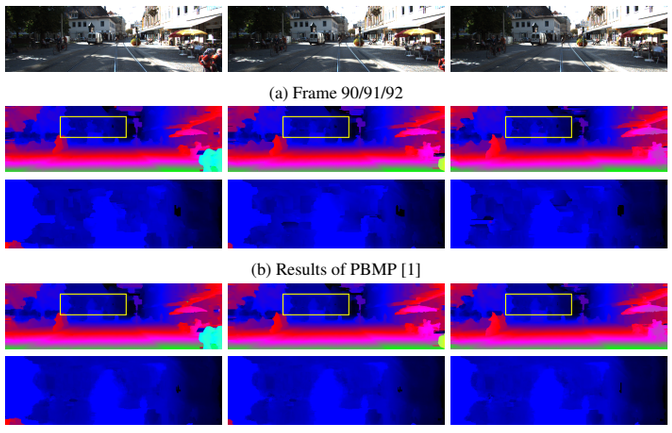
Experimental results We evaluated our method using a synthetic dataset presented in [4] and a real dataset provided by KITTI. For comparison,



(a) Book (b) Street (c) Tanks

(d) Temple (e) Tunnel

Figure 1: Average bad-pixel rates for synthetic data sets according to increasing noise level



(a) Frame 90/91/92

(b) Results of PBMP [1]

(c) Results of PMBP-3D

Figure 2: Results for real dataset.

we selected the original PMBP-based method [1] as a baseline. Here, the baseline and proposed methods are referred to as PMBP and PMBP-3D, respectively. To see clearly the consistency improvement, we analyzed the accuracy of disparity maps for synthetic image sequences with noise. Figure 1 shows average bad-pixel rates according to different noise levels. As the noise level increases, average bad-pixel rates of PMBP increase faster than those of PMBP-3D. Next, we analyzed the results for a real dataset. Figure 2 illustrates disparity maps for three successive frames. To clearly compare the performance of PMBP and PMBP-3D, we provide magnified results with contrast adjustment for a specific area of the disparity map. The specific area is marked with a yellow square in the disparity map. Because the image captured for a real scene contains ambiguous regions and inevitable noise, PMBP is likely to provide temporally inconsistent disparity maps. Whereas, PMBP-3D generates a consistent disparity map sequence.

[1] Frederic Besse, Carsten Rother, Andrew Fitzgibbon, and Jan Kautz. Pmbp: Patchmatch belief propagation for correspondence field estimation. *IJCV*, 110(1):2–13, 2014.

[2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo - stereo matching with slanted support windows. In *BMVC*, 2011.

[3] Thomas Brox and Jitendra Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE TPAMI*, 33(3):500–513, 2011.

[4] Christian Richardt, Douglas Orr, Ian Davies, Antonio Criminisi, and Neil A. Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *ECCV*, 2010.

Randomized Global Transformation Approach for Dense Correspondence

Kihong Park
khpark7727@gmail.com
Seungryong Kim
srkim89@yonsei.ac.kr
Seungchul Ryu
ryus01@yonsei.ac.kr
Kwanghoon Sohn
khsohn@yonsei.ac.kr

School of Electrical and Electronic Engineering
Yonsei University
Seoul, Korea



(a) Initial match (b) $\mathbf{T}_1$ (c) $\mathbf{T}_2$ (d) $\mathbf{T}_3$ (e) $\mathbf{T}_4$

Figure 1: Intuitions of our approach. With sparse features in (a), there exist sparse correspondence sets for global transformation which provide reliable flow on each sub-region.

Motivation Recently, many researchers have begun to attempt to solve dense correspondence problem for more challenging images which have high variability in terms of photometric and/or geometric conditions [4].

For these challenging scenarios, there exist two principal bottlenecks which make conventional methods provide limited performances; (1) photometric variations derived from different camera specifications, illumination or exposure conditions, and (2) geometric variations derived from viewpoint changes, object pose changes, and non-rigid deformation for objects [1]. For the first bottleneck, many robust feature descriptors have been proposed to alleviate photometric variations. As a pioneering work, the SIFT flow [2] has shown satisfactory results on different image pairs having semantically similar property by employing the SIFT descriptor. However, geometric variations for the second bottleneck still remain unsolved due to its large search space, including translation (or flow), rotation, and scale, which induces extremely large computational complexity.

In this paper, we propose a randomized global transformation approach to estimate reliable correspondence between challenging image pairs having photometric and geometric deformation. Our approach starts from an intuition that geometric variations between two images can be formulated as piecewise transformation model as shown in Fig. 1. Finding the optimal global transformation for each pixel enables us to infer correspondence fields without computing geometric fields.

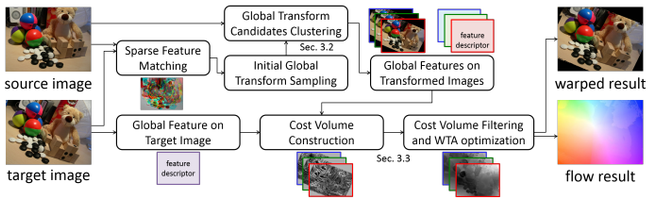


Figure 2: Framework of our randomized global transformation approach. It estimates initial global transformation candidates from sparse feature correspondences. The cost volume between global feature descriptors are then constructed. To infer reliable correspondence, it employs a cost volume filtering using fast edge-aware filtering.

Approach Fig. 2 shows overall framework of our approach. Finding the optimal global transformation for each pixel enables us to infer correspondence fields without computing geometric fields. To build reliable candidates of correspondence fields, we choose a flow hypothesis by leveraging a randomized sampling scheme from initial sparse matching fields. In order to reduce repeated transformation and provide robustness to initial outliers, we employ a clustering scheme on initial global transformation candidates. By using these global transformation candidates, dense descriptor is built on target image and transformed source images, which will be contributed to compute a cost volume between two image pairs. Furthermore, with edge-aware filtering scheme, we employ a cost filter scheme to provide a geometric smoothness.

Contributions The contribution of this paper can be summarized as follows. First of all, unlike conventional approaches that consider large search space extensively, our approach reduces search space labels by leveraging optimal global transformation, which provides accurate correspondence results while producing very low complexity. Secondly, in order to find reliable correspondences, we employ a cost filter scheme by leveraging fast edge-aware filtering. Thirdly, our approach is very simple but provides satisfactory results, and it further can be combined with any other feature descriptors and filtering approaches. Finally, we provide an intensive comparative study with state-of-the-art methods using various datasets.

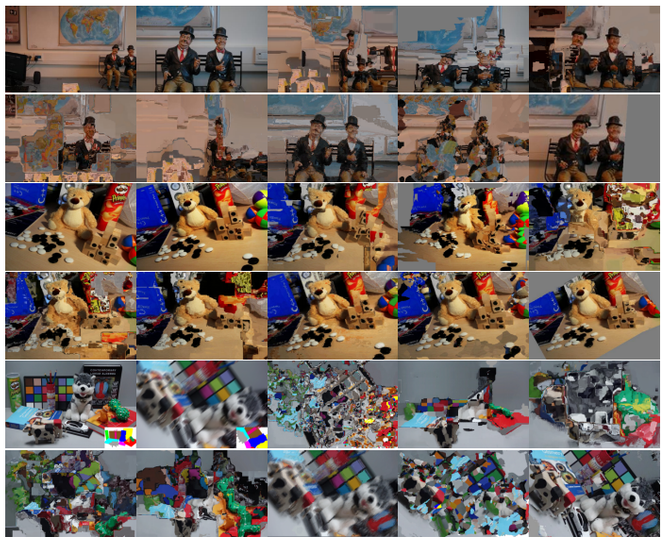


Figure 3: Comparison of dense correspondence for challenging non-rigid image pairs. (from top to bottom, from left to right) Source and target images, warping results from SIFT flow, SSF, SID, SegSID, SegSF, GPM, DFF, and proposed method.

Results Fig. 3 presents image warping results from estimated flow fields by computing dense correspondences from source to target images. These image pairs are non-rigid because their relationships cannot be expressed by just one global transformation [3]. We collected these image pairs among the geometric distorted images which were used in many previous researches, and thus can explain robustness property of our algorithm against various geometric challenging situation. Compared to conventional methods, our approach estimated reliably correspondence for challenging image pairs since even for challenging pairs the correspondence field can be modeled by piecewise parametric transformation models.

[1] J. Hur, H. Lim, C. Park, and S. C. Ahn. Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation. *In Proc. of CVPR*, 2015.

[2] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. PAMI*, 33(5):978–994, 2011.

[3] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. *In Proc. of ICCV*, 2013.

[4] H. Yang, W. Lin, and J. Lu. Daisy filter flow: A generalized discrete approach to dense correspondences. *In Proc. of CVPR*, 2014.

Achieving Turbidity Robustness on Underwater Images Local Feature Detection

Felipe Codevilla
felipe.codevilla@furg.br
Joel De O. Gaya
joelfelipe@furg.br
Nelson Duarte Filho
dmtnldf@furg.br
Silvia S. C. Botelho
silviacb@furg.br

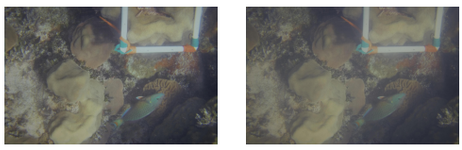
Center of Computational Sciences (C3)
Federal University of Rio Grande (FURG)
Rio Grande, Brazil

Methods to detect local features have been made to be invariant to many transformations. So far, the vast majority of feature detectors consider robustness just to over-land effects. However, when capturing pictures in underwater environments, there are media specific properties that can degrade the visual quality the captured images. Besides the shortening of information imaged, a corruption of information also happens. The main phenomena, the *Backscattering*, happens when reflected sources from outside the captured scene are scattered in a wide angle eventually reaching the image plane. This effect creates a characteristic veil on the image that reduces contrast and suppress fine structures on the image.

Little work has been made in order to study the robustness that the popular feature detectors have to underwater environment image conditions. In addition, applications benefited by finding descriptive feature points in underwater environments grows every year. They are essential for many applications like 3D reconstruction [4], visual odometry [6] and tracking [7]. Most of these applications rely on the best over-land feature detectors, without considering the water photometric properties. It is likely that some algorithms have a better behaviour than others when applied on images degraded by specific underwater conditions.

As stated before, underwater phenomena also create structural degradation. Further, it tends to eliminate all the *finer scale structures*, which is equivalent to the scale changing phenomena. Consequently, we propose the invariant points detected by some scale invariant detector can have also a good robustness to turbidity.

In this context, to evaluate feature detectors, we propose a new dataset called *TURBID*. This dataset is based on real underwater scenes photographs. The pictures are placed on the bottom of a tank filled with a milk-water solution and then are re-photographed with the degradation controlled by the amount of milk. The generated dataset for one of the printed photographs is show in Fig. 1. This dataset is an improvement in terms of visual diversity when compared to previous efforts [8] and is one of the main contributions of this work.



(a) Clean (b) 15 ml



(a) 50 ml (b) 100 ml

Figure 1: The images captured over different levels of degradation due to turbidity controlled by milk addition. We photographed three different printed pictures. (a) Clean Image, no milk. (b) *Low Turbidity* with around 15 ml of milk. (c) *Medium Turbidity* with around 50 ml of milk. (d) *High Turbidity* with around 100 ml of milk.

In order to analyze the robustness to turbidity we use the repeatability criteria on each obtained image. This criteria is proportional to the number of feature points found in the same spot given a error ε. The repeatability towards turbidity is calculated by the ratio of the number of points found in the clean image (Fig. 1(a)) and the number of points repeated in a turbid image, hence:

$$R_i = \frac{N_i}{N_0}, \tag{1}$$

where N_0 is the number of features on the clear image, and N_i is the number of features *repeated* on the image T_i.

For three different photos we computed the repeatability results. An example of this computation is show for Fig. 2.

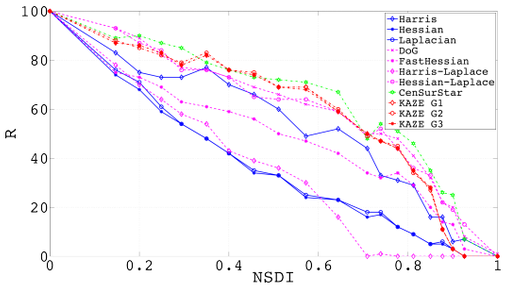


Figure 2: Results for $N = 100$ and $\varepsilon = 5$, showing the repeatability with respect to the ranging of the image degradation. For this case the *CenSurStar* [1] and KAZE [2] got the best results.

As we conjectured earlier, the *Harris* [9], the *Hessian* [5] and *Laplacian* approaches performed poorer than the scale invariant methods. *Harris* is generally used as very precise detector and is used in underwater tracking applications [7]. However, we show that, on the present scenario, the use of scale is useful also for precision.

Finally, we found that, finding scale invariant points is a useful way to find structural degradation robust points. We elected KAZE [2], *Center Surround Extremas* [1], *Difference of Gaussians* [10] and *Fast Hessian*[3] as good feature points detectors for underwater environments in all tested situations.

[1] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. Censure: Center surround extremas for realtime feature detection and matching. In *Computer Vision–ECCV 2008*, pages 102–115. Springer, 2008.

[2] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *Computer Vision–ECCV 2012*, pages 214–227. Springer, 2012.

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110, 2008.

[4] C. Beall, B.J. Lawrence, V. Ila, and F. Dellaert. 3d reconstruction of underwater structures. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4418–4423, 2010.

[5] P. R. Beaudet. Rotationally invariant image operators. In *Proceedings of the 4th International Joint Conference on Pattern Recognition*, pages 579–583, Kyoto, Japan, November 1978.

[6] S. S. C. Botelho, Paulo Drews-Jr, M. Figueiredo, C. Da Rocha, and G. L. Oliveira. Appearance-based odometry and mapping with feature descriptors for underwater robots. *Journal of the Brazilian Computer Society*, 15:47 – 54, 09 2009. ISSN 0104-6500.

[7] P. Corke, C. Detweiler, M. Dunbabin, M. Hamilton, D. Rus, and I. Vasilescu. Experiments with underwater robot localization and tracking. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 4556–4561, 2007. doi: 10.1109/ROBOT.2007.364181.

[8] Rafael Garcia and Nuno Gracias. Detection of interest points in turbid underwater images. In *OCEANS, 2011 IEEE-Spain*, pages 1–9. IEEE, 2011.

[9] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

The Self-Equalizing De Bruijn Sequence for 3D Profilometry

Tomislav Petković
tomislav.petkovic.jr@fer.hr
Tomislav Pribanić
tomislav.pribanic@fer.hr
Matea Đonlić
matea.donlic@fer.hr

University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia

Three-dimensional non-invasive surface profilometry using structured light (SL) is a widely studied topic in the field of computer vision [2]. One of the most important issues of all SL systems is the design of a SL pattern: the pattern must enable reconstruction robust to the ambient light and to the characteristics of the observed object. SL patterns may be classified as *one-shot* patterns that enable reconstruction from a single image and as *multi-shot* patterns that require more than one image. Nowadays a well-designed pattern should enable both one and multi-shot reconstruction. Such SL patterns were proposed by Li Zhang et al. [3] for color light and Yueyi Zhang et al. [4] for monochromatic light.

We propose a novel De Bruijn sequence for color or multi-channel SL which we call the *self-equalizing* sequence. The proposed SL sequence enables a dense *multi-shot* reconstruction for static objects and a sparse *one-shot* reconstruction for moving objects. Additionally, the proposed sequence has several desirable properties: (1) it enables the extraction of ambient lighting, (2) it enables the cancellation of object albedo, and (3) it enables the equalization of channel gains. The proposed sequence thus effectively removes the need for a precise color calibration of the imaging system, which is a major advantage in real-world applications.

Current approaches to SL pattern construction based on a De Bruijn sequence either use a complete sequence or impose a constraint on neighboring colors only. We propose to impose the novel *self-equalizing* constraint on the entire De Bruijn window in such way to achieve the aforementioned desirable properties. The constraint essentially requires that in every De Bruijn window all channels span the full available dynamic range, which is the sufficient condition for the removal of ambient lighting, cancellation of object albedo, and equalization of channel gains thus enabling robust identification of the projected color. A De Bruijn sequence is associated with Eulerian cycle of a De Bruijn digraph which may be used to generate it. The proposed sequence is constructed as Eulerian cycle in the pruned De Bruijn digraph where edges belonging to the invalid windows that do not satisfy the proposed *self-equalizing* constraint are removed.

The proposed SL pattern is a cyclic sequence of images whose cycle is comprised of $L \cdot S$ frames (L is sequence length, $S \geq 3$ is the number of phase shifts per symbol), of which any one frame is sufficient for one-shot reconstruction, and of which any $n \cdot S$ ($n \ll L$) consecutive frames are sufficient for multi-shot reconstruction.

One model for RGB color-space widely accepted in the literature which relates recorded color I_c to projected color I_p under the influence of object albedo K and ambient lightning I_0 is proposed by Caspi et al. [1]:

$$\underbrace{\begin{bmatrix} R \\ G \\ B \end{bmatrix}}_{I_c} = \underbrace{\begin{bmatrix} a_{RR} & a_{RG} & a_{RB} \\ a_{GR} & a_{GG} & a_{GB} \\ a_{BR} & a_{BG} & a_{BB} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} k_R & 0 & 0 \\ 0 & k_G & 0 \\ 0 & 0 & k_B \end{bmatrix}}_{K} f\Big(\underbrace{\begin{bmatrix} r \\ g \\ b \end{bmatrix}}_{I_p} \Big) + \underbrace{\begin{bmatrix} R_0 \\ G_0 \\ B_0 \end{bmatrix}}_{I_0}, \quad (1)$$

We propose a simplified model where ith channel is modeled as

$$I_c(i) = h(i) I_p(i) + I_0(i), \quad (2)$$

i.e. the channel cross-talk is disregarded making $A \cdot K$ a diagonal matrix. The cross-talk may be omitted if the color filters of the projector and the camera are well matched, which holds for standardized consumer electronics and which was also noted by Caspi et al. [1]. Combining the model (2) with the *self-equalizing* constraints imposed on the De Bruijn sequence allows the recovery of $I_p(i)$ from $I_c(i)$ independently for each channel.

We have performed three experiments: (1) the influence of channel gains, (2) the reconstruction of white and colored planar surfaces, and (3) qualitative reconstructions of several objects: fruits and vegetables, a human face, and a human hand. The first experiment demonstrates the insensitivity to artificially varying channel gains. The reconstruction of a colored planar surface of the second experiment demonstrates the insensitivity to object albedo, but only if there is no significant cross-talk between channels (see Fig. 1). The third experiment illustrates expected results in a real-world application; here we show results for a female face (Fig. 2).

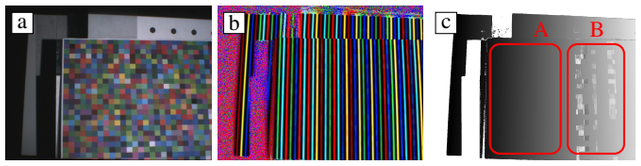


Figure 1: An extremely demanding colorful checkerboard: (a) scene seen by the camera; (b) equalized image; (c) unwrapped phase of the multi-shot method. Note the unwrapping succeeds for region A and fails for region B due to cross-talk which is disregarded in the current model.

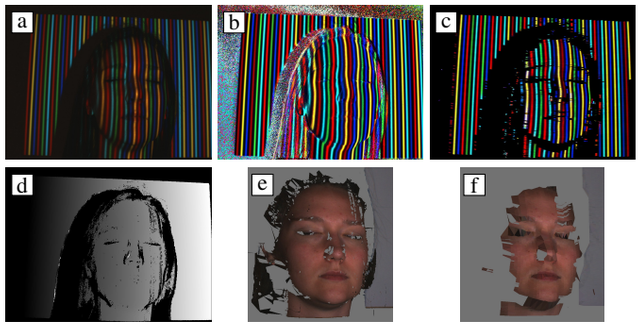


Figure 2: A female face: (a) the first of $n \cdot S$ input images; (b) temporally equalized input image (a); (c) spatially equalized input image (a); (d) the unwrapped phase; (e) textured 3D reconstruction using multi-shot method; (f) textured 3D reconstruction using one-shot method. Note that both the temporal and the spatial equalization successfully remove ambient lightning, object albedo, and varying channel gains.

[1] Dalit Caspi, Nahum Kiryati, and Joseph Shamir. Range imaging with adaptive color structured light. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(5):470–480, May 1998. ISSN 0162-8828. doi: 10.1109/34.682177.

[2] Joaquim Salvi, Sergio Fernandez, Tomislav Pribanić, and Xavier Lladó. A state of the art in structured light patterns for surface profilometry. *Pattern Recognition*, 43(8):2666–2680, 2010. doi: 10.1016/j.patcog.2010.03.004.

[3] Li Zhang, Brian Curless, and Steven M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *3D Data Processing Visualization and Transmission*, pages 24–37, 2002. doi: 10.1109/TDPVT.2002.1024035.

[4] Yueyi Zhang, Zhiwei Xiong, Zhe Yang, and Feng Wu. Real-time scalable depth sensing with hybrid structured light illumination. *Image Processing, IEEE Transactions on*, 23(1):97–109, Jan 2014. ISSN 1057-7149. doi: 10.1109/TIP.2013.2286901.

A fast and robust ellipse detector based on top-down least-square fitting

Yongtao Wang
wyt@pku.edu.cn
Zheqi He
philokeys@gmail.com
Xicheng Liu
liuxicheng@pku.edu.cn
Zhi Tang
tangzhi@pku.edu.cn
Luyuan Li
liluyuan@pku.edu.cn

Institute of Computer Science & Technology
Peking University
Beijing, China

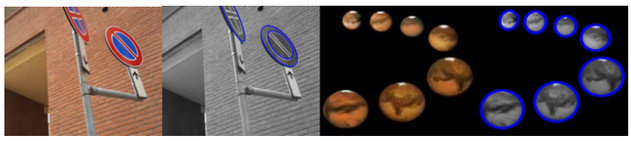


Figure 1: Detecting ellipses from images

Detecting ellipses efficiently and accurately in digital images is an fundamental problem to the field of pattern recognition and computer vision. Ellipse detection can be used in many applications. The existing algorithms such as [4] often use a bottom-up strategy to combine edge points or elliptical arcs into ellipses, which limits their robustness. In this paper, we propose a novel algorithm which can efficiently and accurately detect ellipses in digital images with a novel top-down scheme. The main idea of the proposed algorithm is to exploit a novel top-down fitting strategy to combine edge points into ellipses and use integral chain to speed up the fitting process. The proposed method is very efficient, which is faster than the most efficient method [1] which has been reported. Our experimental results also demonstrate that our algorithm is more robust than the state-of-the-art methods including [1] and ELSD [4].As shown in Fig. 2, our method consists of four major steps.

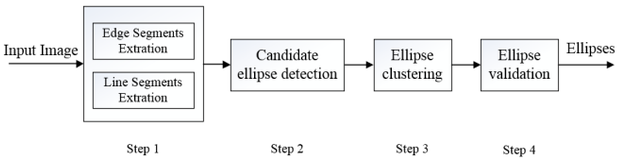


Figure 2: Workflow of the proposed ellipse detection method

In the first step, we extract edge segments from the input image and then obtain line segments within each edge segment. We directly use the edge segment extraction method and the line segment detection method proposed in [2, 3] respectively, which are very computational efficient. Secondly, we detect candidate ellipses from each edge segment by the top-down analysis. In this step, we use the 1-D case(which is called "integral chain" hereafter) of integral image technique [5] to speed up the proposed ellipse detection algorithm.Specifically, for a sequence $x_1, x_2, ..., x_n$, the corresponding integral chain is:

$$IntChain(k) = \sum_{i=1}^{k} x_i, k = 1, ..n. \quad (1)$$

The recurrent formulas for computing the integral chain are:

$$IntChain(1) = x_1, \quad IntChain(k) = IntChain(k-1) + x_k, \quad k > 1. \quad (2)$$

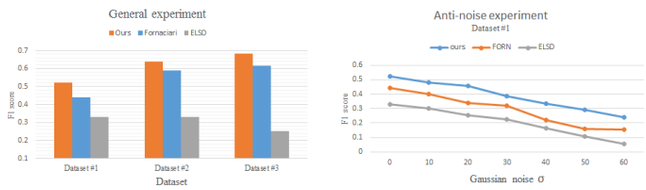
Within the integral chain, the sum of any sub-sequence can be efficiently obtained by

$$\sum_{i=1}^{k} x_i = IntChain(k), \sum_{i=s}^{k} x_i = IntChain(k) - IntChain(s-1). \quad (3)$$

After that, we cluster candidate ellipses to merge and remove the fractional detections, because a complete ellipse may be split into two or more candidate ellipses due to the incomplete edge segment extraction results. Finally, we eliminate false alarms using NFA validation. In detail, for an ellipse with l independent edge points, there are k of these points having their gradient directions aligned with the direction of this ellipse, where aligned means that a point p_i is at the direction orthogonal to the tangent of the ellipse on p_i. The expectation of such kind of ellipse, so-called "number of false alarms"(NFA), should be less than φ , that is:

$$NFA(l,k) = N_t \sum_{i=k}^{l} \binom{l}{i} p^i (1-p)^{l-i} < \varphi, \quad (4)$$

where N_t is the number of all possible ellipses; φ is set to 1 and the precision of the direction alignment p is set to $1/8$ as mentioned in [4]. We calculate the gradient direction of each edge point in the original image with Sobel operator(the aperture size is 3).



(a) (b)

Figure 3: (a) F1 scores of general experiment; (b) F1 scores of anti-noise experiment on Dateset #1

We make experiments on three public real image datasets. We compare the performance of our method with two state-of-the-art ellipse detection methods: (1) Fornaciari *et al.*'s method (FORN) which is proposed in [1], and (2) ELSD [4] which is proposed by Puatruaucean *et al.* Fig. 3 illustrate the result on the three datasets.

The general experiments indicate the proposed method achieves better performance than the other two methods: it runs faster while achieving the highest F1 score. The anti-noise experiments evaluate robustness to noise of these methods, we add Gaussian noise with different level (the variance $\sigma^2 = \{0, 10^2, 20^2, 30^2, 40^2, 50^2, 60^2\}$) to images in the three datasets. The result indicates that our method is more robust than the two other methods.

[1] Michele Fornaciari, Andrea Prati, and Rita Cucchiara. A fast and effective ellipse detector for embedded vision applications. *Pattern Recognition*, 47(11):3693–3708, 2014.

[2] Dong Liu, Yongtao Wang, Zhi Tang, and Xiaoqing Lu. A robust circle detection algorithm based on top-down least-square fitting analysis. *Computers & Electrical Engineering*, 40(4):1415–1428, 2014.

[3] Dong Liu, Yongtao Wang, Zhi Tang, and Xiaoqing Lu. A robust and fast line segment detector based on top-down smaller eigenvalue analysis. In *Fifth International Conference on Graphic and Image Processing*, pages 906916–906916. International Society for Optics and Photonics, 2014.

[4] Viorica Pătrăucean, Pierre Gurdjos, and Rafael Grompone Von Gioi. A parameterless line segment and elliptical arc detector with enhanced ellipse fitting. In *Computer Vision–ECCV 2012*, pages 572–585. Springer, 2012.

[5] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

Multiview Reconstruction of Complex Organic Shapes

Jasenko Zivanov
jasenko.zivanov@unibas.ch
Thomas Vetter
thomas.vetter@unibas.ch

Department of Mathematics and Computer Science
University of Basel
Basel, Switzerland

We propose a novel narrow baseline multiview stereo surface reconstruction method that is specifically aimed at complex shapes of biological origin that show many thin protrusions and curved occluding contours. Our method is built around fitting local quadrics to pixel-precise occluding contours and it avoids any planarity assumptions that are common to other state of the art methods. We describe a complete pipeline that begins with a sequence of calibrated noisy images and produces a final watertight surface. Although quadrics have been used before to model geometry ([1, 2, 4]), our work is to our knowledge the first to use them as part of a complete stereo reconstruction pipeline. The pipeline is sketched out in figure 1.

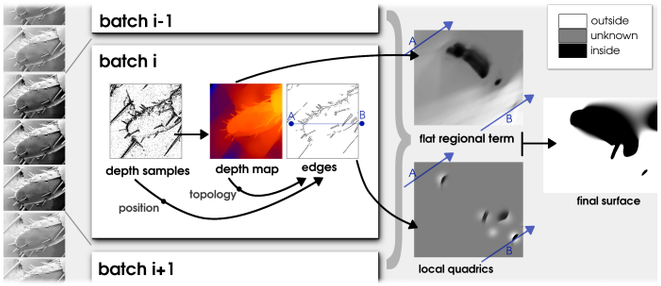


Figure 1: An overview over our pipeline. A detailed explanation is given in the text below.

Shapes of biological origin often feature curved contours that do not correspond to static curves in space - instead, those curves drift along the surface as the vantage point changes. For this reason, our method estimates the positions of occluding contours only from small subsets of the entire image sequence. By only looking at short batches of frames, we can determine the momentary positions of such contours, which later enable us to fit local quadrics to them. This separation into batches also makes our method very robust to non-Lambertian reflectiveness and to changes in illumination. Our method can even be applied to sequences where the illumination is fixed to the observer and not the scene.

For every batch, we compute a sparse depth map corresponding to the center image of the batch. The depth values are determined independently for every pixel, which allows us to reconstruct very small features. If we were to compare image windows consisting of multiple pixels, we would be making an implicit planarity assumption. In spite of considering only single pixels, our depth estimation process is still very robust to noise, since the optimal depths and the optimal colors of each pixel are determined simultaneously. This results in both a sparse depth map and a dense, denoised version of the image. An example of a denoised image is shown in figure 2.

We then apply a process similar to [3] to estimate a dense depth map that matches the sparse depth map and that conforms to the structure of the denoised image. This can be viewed as a process of reasoning that propagates information from the observed edges into the smooth areas of the image. This dense depth map is then used to detect the occluding contours of that batch while the original sparse depth map is used to estimate their locations and orientations.

Next, we use the contours detected in all the batches to estimate local quadrics in a discretized volume. Using quadrics as a surface representation offers the advantage that the smoothness of the quadric parameters does not translate into planarity of the surface. Instead, arbitrarily curved shapes can be represented by smooth quadric coefficients. At every voxel v, we determine a real, symmetrical 4×4 matrix C_v that holds the coefficients of the local quadric, where the quadric is defined as the zero level set of $f(x) = x^T C_v x$, where $x = (x_1, x_2, x_3, 1)^T$ is a homogeneous position vector. The coefficients of C_v are determined such that the quadric is tangential to the contours detected in the area of space surrounding v.

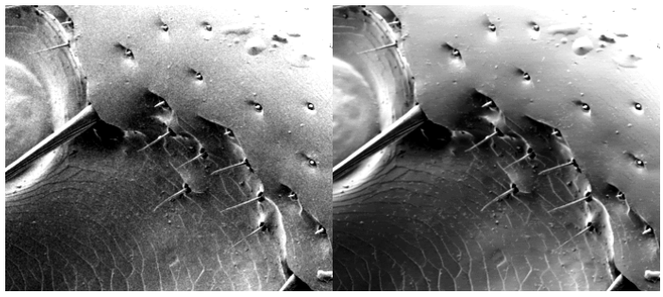


Figure 2: The denoising effect of our depth estimation process. The original image (left) and the denoised version (right). The corresponding batch contained 50 images. The contrast of the displayed images has been enhanced to emphasize the noise.

Finally, the estimated quadrics are combined with a more traditional regional term to obtain a final watertight surface using a volumetric optimization process. The final surface is required to fit the quadrics in areas close to detected contours and to be consistent with the dense depth maps otherwise.

We have tested our method on a sequence of 1400 images of a cat flea captured by a scanning electron microscope using secondary electron imaging. The results can be seen in figure 3.We show that very thin features can be reconstructed correctly by our method, even in the absence of strong contrast and even if the illumination is constant with respect to the viewer and not to the scene.



Figure 3: Reconstruction of a cat flea (ctenocephalides felis) from 1400 scanning electron microscope images. From left to right: one of the input images; enlarged area of the input image; a reconstruction of the enlarged area.

[1] Roberto Cipolla, Bjorn Stenger, Arasanathan Thayananthan, and Philip HS Torr. Hand tracking using a quadric surface model and bayesian filtering. In *Mathematics of Surfaces*, pages 129–141. Springer, 2003.

[2] Geoffrey Cross and Andrew Zisserman. Quadric reconstruction from dual-space geometry. In *Computer Vision, 1998. Sixth International Conference on*, pages 25–31. IEEE, 1998.

[3] Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Occluding contours for multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4002–4009. IEEE, 2014.

[4] Hui Xie, Jianning Wang, Jing Hua, Hong Qin, and Arie Kaufman. Piecewise c1 continuous surface reconstruction of noisy point clouds via local implicit quadric regression. In *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)*, page 13. IEEE Computer Society, 2003.

Manifold Regularized Transfer Distance Metric Learning

Haibo Shi[1]
shh@pku.edu.cn
Yong Luo[2]
yluo180@gmail.com
Chao Xu[1]
xuchao@cis.pku.edu.cn
Yonggang Wen[2]
ygwen@ntu.edu.sg

[1] Key Laboratory of Machine Perception (MOE), Cooperative Medianet Innovation Center, School of EECS, Peking University, Beijing, China
[2] School of Computer Engineering, Nanyang Technological University, Singapore

The performance of many computer vision and machine learning algorithms are heavily depend on the distance metric between samples. It is necessary to exploit abundant of side information like pairwise constraints to learn a robust and reliable distance metric[2, 3]. Let $\mathcal{D} = \{(x_i^l, x_j^l, y_{ij})\}_{i,j=1}^{n^l}$ denotes the labeled training set for the target task, wherein x_i, $x_j \in R^d$ and $y_{ij} = \pm 1$ indicates x_i^l and x_i^l are similar/dissimilar to each other.

Then, a metric is usually learned to minimize the distance between the data from the same class and maximize their distance otherwise. This leads to the following loss function for learning the metric A:

$$\begin{aligned}\Phi(A) &= \sum_{y_{ij}=1} \|x_i - x_j\|_A^2 - \mu \sum_{y_{ij}=-1} \|x_i - x_j\|_A^2 \\ &= \mathbf{tr}(\mathbf{S}\cdot A) - \mu \mathbf{tr}(\mathbf{D}\cdot A)\end{aligned} \tag{1}$$

where $d_A(x_i, x_j) = \|x_i - x_j\|_A = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$ is the distance between two data points x_i and x_j, $\mathbf{tr}(\cdot)$ is the trace of matrix, μ is a positive trade-off parameter. Here, S and D are given by $\mathbf{S} = \sum (x_i - x_j)(x_i - x_j)^T$, $(x_i, x_j) \in \mathcal{S}$, $\mathbf{D} = \sum (x_i - x_j)(x_i - x_j)^T$, $(x_i, x_j) \in \mathcal{D}$.

The loss function (1) above is widely used in distance metric learning (DML) method. A regularization term $\Omega(A) = \|A\|_F^2$ can be added in (1) to control the model complexity. However, when the number of labeled data n^l is small, such a simple regularization is often insufficient to control the model complexity. The recently proposed decomposition based TDML (DTDML) [2] algorithm is superior to the previous TMDL approaches in that much fewer variables are needed to be learned. Given the m source tasks, we assume there are large amount of n^u unlabeled data $\{x_i^u, x_j^u\}$, as well as m different but related source tasks with abundant labeled training data $\mathcal{D}_p = \{(x_{pi}, x_{pj}, y_{pij})\}_{i,j=1}^{n_p}$, $p = 1, \ldots, m$. Then we learn m corresponding metrics $A_p \in \mathbb{R}^{d\times d}, p = 1, \ldots, m$ independently. Considering that any metric A can be decomposed as $A = U\mathrm{diag}(\theta)U^T = \sum_{i=1}^d \theta_i u_i u_i^T$, DTDML proposed to learn a combination of some base metrics to approximate the optimal target metric. The base metrics can be derived from the source metrics or some randomly generated base vectors. Based on this idea, the formulation of DTDML is given by

$$\arg\min_{\beta,\theta} \Phi(\beta, \theta) + \frac{\gamma_A}{2}\|A - A_S\|_F^2 + \frac{\gamma_B}{2}\|\beta\|_2^2 + \frac{\gamma_C}{2}\|\theta\|_1^2 \tag{2}$$

where $\Phi(\cdot)$ is some pre-defined convex loss, $A = \sum_{r=1}^{m\times d} \theta_r u_r u_r^T$, and $A_S = \sum_{p=1}^m \beta_p A_p$ is an integration of the source metrics.

Although the limited labeled samples in the target task and the auxiliary source metrics are effectively utilized in problem (2) by simultaneously minimizing the losses $\Phi(\beta, \theta)$ and the divergence between A_S and A, the large amount of unlabeled data in the target task are discarded. Therefore, we propose to utilize manifold regularization [1] to take advantage of all the given labeled and unlabeled information in a unified metric learning framework.

Manifold regularization implies the geometry of the intrinsic data probability distribution is supported on the low-dimensional manifold. The Laplacian of the adjacency graph computed in an unsupervised manner using Laplacian Eigenmap with both labeled and unlabeled samples. The data manifold can be approximated with the graph Laplacian. Moreover, the distance measure is a key point for graph Laplacian construction. Since both the integrated source metric A_S and target metric A are derived from the same feature space and related tasks, these two metrics should be similar. Rather than explaining this similarity by simply minimizing the least squares difference in DTDML, we formulate it as a smoothing penalty term. Based on the obtained source metric A_p, we construct an adjacency graph W_p by using all the labeled and unlabeled data in the target task. This leads to multiple graphs W_p, $p = 1, \ldots, m$. Considering the target metric A, distance between two samples can be further written as, $d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)} = \sqrt{(x_i - x_j)^T PP^T (x_i - x_j)}$ with $P \in \mathbb{R}^{d\times d}$. As a consequence, it is equivalent to learn the target metric A and the linear mapping P. Following the manifold regularization principle, we can smooth P along the data manifold [1, 3], which is approximated by the Laplacian of the graph W_p. By summing over all the different graphs $\{W_p\}_{p=1}^m$, we obtain the following regularizer for the mapping P as well as the metric A, i.e.,

$$\begin{aligned}\Omega(A) &= \frac{1}{2}\sum_{p=1}^m \beta_p \Big(\sum_{i,j} \|Px_i - Px_j\|^2 W_p(i,j)\Big) \\ &= \mathbf{tr}(XLX^T A)\end{aligned} \tag{3}$$

where $L = \sum_{p=1}^m \beta_p L_p$, is the integrated graph Laplacian, and each $L_p = D_p - W_p$. Here, D_p is a diagonal matrix with the entity $D_{pii} = \sum_{j=1}^{n^l+n^u} W_{pij}$. In this way, target metric A is not only close to an integration of the source metrics, and also smooth along the data manifold. This leads to lower model complexity compared with DTDML, and thus better generalization ability for metric learning.

By introducing the regularizer (3) in (1), and adopting the decomposition based metric learning strategy in [2], we obtain the following optimization problem for our MTDML:

$$\begin{aligned}\arg\min_{\beta,\theta} &\mathbf{tr}(\mathbf{S}\cdot A) - \mu\mathbf{tr}(\mathbf{D}\cdot A) + \gamma_A \mathbf{tr}(XLX^T A) \\ &+ \frac{\gamma_B}{2}\|\beta\|_2^2 + \frac{\gamma_C}{2}\|\theta\|_2^2 \\ &\text{s.t.} \sum_{i=1}^m \beta_i = 1, \beta_i \geq 0, i = 1, \ldots, m\end{aligned} \tag{4}$$

where $\gamma_A, \gamma_B, \gamma_C$ are positive trade-off parameters selected empirically by grid search. With learned θ^*, we can easily construct $A^* = \sum_{r=1}^{m\times d} \theta_r^* u_r u_r^T$ as optimal distance metric for next step classification.

In the optimization, the "base metric" combination coefficients and the source graph Laplacian integration weights are learned alternatively until converge. We therefore obtain more reliable solutions given the limited side information. Experiments are conducted on NUS-WIDE, which is a challenge web image annotation dataset and USPS, a handwritten digit classification dataset. The results confirm the effectiveness of the proposed MTDML.

Acknowledgements
This research was partially supported by grants from NBRPC 2011CB302400 NSFC 61375026, 2015BAF15B00.

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7: 2399–2434, 2006.

[2] Y Luo, T Liu, D Tao, and C Xu. Decomposition-based transfer distance metric learning for image classification. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 23(9):3789–3801, 2014.

[3] Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. Robust distance metric learning with auxiliary knowledge. In *IJCAI*, pages 1327–1332, 2009.

Top-down Saliency with Locality-constrained Contextual Sparse Coding

Hisham Cholakkal
hisham002@ntu.edu.sg
Deepu Rajan
http://www3.ntu.edu.sg/home/ASDRajan
Jubin Johnson
jubin001@ntu.edu.sg

School of Computer Engineering
Nanyang Technological University
Singapore

We propose a locality-constrained contextual sparse coding (LCCSC) for top-down saliency estimation where higher saliency scores are assigned to the image locations corresponding to the target object. Three locality constraints are integrated in to this novel sparse coding. First is the spatial or contextual locality constraint in which features from adjacent regions have similar code, second is the feature-domain locality constraint in which similar features have similar code, and third is the category-domain locality constraint in which features are coded using similar atoms from each partition of the dictionary, where each partition corresponds to an object category.

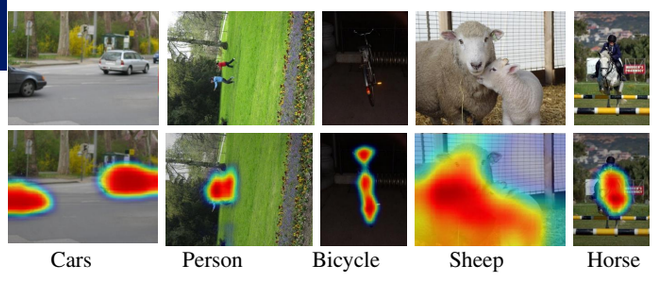


Figure 1: Our Saliency maps on Graz-02 and PASCAL VOC-07 datasets.

The pipeline of the proposed method is shown in fig. 2, which is similar to the widely used sparse coded spatial pyramid matching (ScSPM) image classifier, i.e., feature extraction, feature coding, pooling and feature classifier. Spatial neighborhood of a feature is divided into a regular grid and the codes in each cell of the grid are max-pooled individually. These max-pooled vectors are vertically concatenated to form a context max-pooled vector representing the patch. Logistic regression-based feature classifier is learnt using these context max-pooled vectors. For saliency inference on a test image, the class-conditional probability of context max-pooled vectors are estimated from the learnt logistic regression model. This probability is the saliency value of a patch in the image. The pixel-level saliency map is obtained using a novel Gaussian-weighted interpolation of the patch-level saliency map.

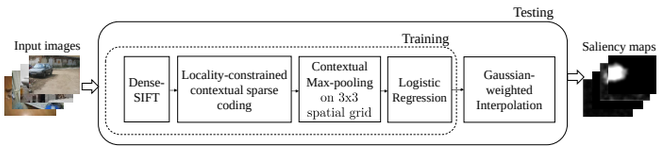


Figure 2: Training and testing of the proposed top-down saliency.

Locality-constrained contextual sparse coding (LCCSC) Various coding schemes aim for specific objectives like sparsity, feature-domain locality and spatial-domain locality. Here, all these desired properties are integrated into a single objective function. LCCSC ensures that features representative of salient regions are not ignored even if they are not discriminative.

Given a feature vector f and dictionary D, LCCSC coding searches for the codeword z that satisfies the following criteria:

$$\arg\min_{z} \|f - Dz\|_2 + \lambda_1 \|z\|_1 + \lambda_2 \|z \odot h_\pi\|_2 + \lambda_3 \sum_{j=1}^{c} \left(\left|\frac{\|z\|_0}{c} - \left\|z \odot col_j[\rho]\right\|_0\right|\right); \quad (1)$$

The first two terms are the conventional sparse coding of feature f with l_1 constraint. The third term imposes locality constraint in the feature domain as well as in the spatial domain. The fourth term imposes category-domain locality, i.e the number of atoms that contribute to the non-zero values of z are distributed among all object categories. Detailed discussion about eq. (1) and an approximate solution is available in the paper.

Table 1: Pixel-level recision rates (%) at EER on Graz-02.

	Bicycle	Car	Person	*Mean*
Objectness	53.5	48.3	43.5	48.43
Aldavert *et al.*	71.9	64.9	58.6	65.13
Khan and Tappen	72.1	-	-	-
Marszalek and Schmid	61.8	53.8	44.1	53.23
Yang and Yang [2]	62.1	60.0	62.0	61.46
Kocak *et al.* [1]	73.9	68.4	**68.2**	70.16
LCCSC-pooled (*upsampling of [2]*)	73.41	68.6	61.25	67.75
LCCSC-pooled (*Gaussian-weighted interpolation*)	**76.19**	**71.2**	64.13	**70.49**

We test our method on two challenging datasets - Graz-02 and PASCAL VOC-07. Table 1 compares the pixel-level results of our method with other top-down saliency models. The proposed method is called as LCCSC-pooled, where pooled indicates contextual max pooling of the code. It is to be noted that we achieve state-of-the-art performance by using simpler feature coding and contextual max-pooling in comparison with computationally complex dictionary learning and graph-based approaches of [1, 2].

As in [2], all models are evaluated on the entire 210 segmentation test images of PASCAL VOC-07 dataset, irrespective of the presence or absence of target. We outperform [2] in 15 out of 20 classes. On averaging across all classes, in patch-level, we achieve a mean precision rate at EER of 23.4% which is better than 16.15% of [2].

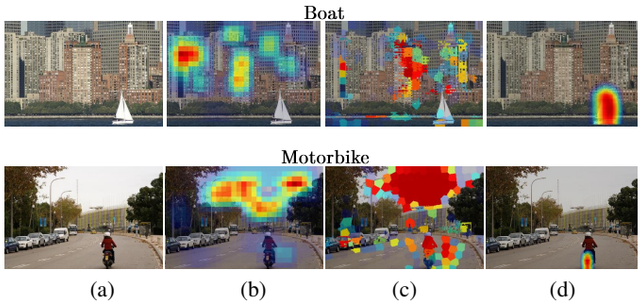


Figure 3: Comparison with state-of-the-art top-down saliency approaches on challenging test images. (a) input image, (b) Saliency maps of Yang and Yang [2], (c) Kocak *et al.* [1] and (d) the proposed method.

[1] A. Kocak, K. Cizmeciler, A. Erdem, and E. E. Top down saliency estimation via superpixel-based discriminative dictionaries. In *BMVC*, 2014.

[2] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *CVPR*, 2012.

Very Efficient Training of Convolutional Neural Networks using Fast Fourier Transform and Overlap-and-Add

Tyler Highlander*
highlander.2@wright.edu

Andres Rodriguez*,*
andres.rodriguez.8@us.af.mil

*College of Engineering and Computer Science
Wright State University
Dayton, Ohio, USA

*Air Force Research Laboratory
Dayton, Ohio, USA

Convolutional neural networks (CNNs) are currently state-of-the-art for various classification tasks, but are computationally expensive. Propagating through the convolutional layers is very slow, as each kernel in each layer must sequentially calculate many dot products for a single forward and backward propagation which equates to $\mathcal{O}(N^2n^2)$ per kernel per layer where the inputs are $N \times N$ arrays and the kernels are $n \times n$ arrays. Convolution can be efficiently performed as a Hadamard product in the frequency domain. The bottleneck is the transformation which has a cost of $\mathcal{O}(N^2 \log_2 N)$ using the fast Fourier transform (FFT). However, the increase in efficiency is less significant when $N \gg n$ as is the case in CNNs. We mitigate this by using the "overlap-and-add" technique reducing the computational complexity to $\mathcal{O}(N^2 \log_2 n)$ per kernel. This method increases the algorithm's efficiency in both the forward and backward propagation, reducing the training and testing time for CNNs. Our empirical results show our method reduces computational time by a factor of 16.3 times the traditional convolution implementation for a 8×8 kernel and a 224×224 input array.

Mathieu *et al*. [2] demonstrated that doing full convolutions as Hadamard products in the frequency domain significantly reduces the training and testing time of CNNs. In their work they efficiently calculated FFTs on a GPU and used these transforms to perform convolutions via a Hadamard product in the frequency domain.

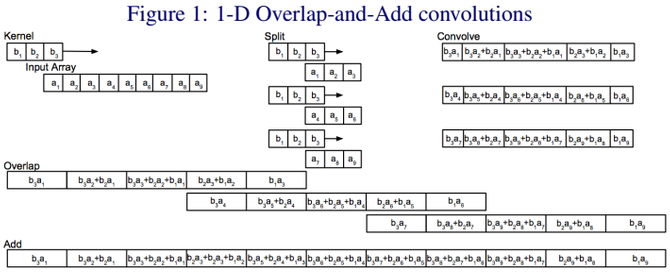
In this paper, we propose to use the overlap-and-add (OaA) technique [4] to further reduce the training and testing complexity to $\mathcal{O}(N^2 \log_2 n)$ per kernel. Note that the overlap-and-save [4] is a similar technique that may be marginally faster but has the same complexity. Table 1 compares the complexity of each method, where spaceConv refers to the traditional convolution in the space domain, FFTconv refers to convolution via a Hadamard product in the frequency domain without overlap-and-add, and OaAconv refers to convolution using overlap-and-save where each smaller convolution is efficiently computed in the frequency domain.

Method	Computational Complexity
spaceConv	$\mathcal{O}(N^2n^2)$
FFTconv	$\mathcal{O}(N^2log_2N)$
OaAconv	$\mathcal{O}(N^2log_2n)$

Table 1: Computational complexity comparison

In OaAconv, the input is broken into N^2/n^2 blocks equal to the kernel size $n \times n$. A convolution between each block and the kernel is computed and the results are overlapped and added. Figure 1 illustrates a simple 1-D overlap-and-add method for spacial convolution (that can easily generalize to 2-D). The input is first split into smaller blocks that are the size of the kernel. Smaller convolutions are computed between the kernel and the block inputs. The resulting convolutions are then added together to create the same results as a traditional spacial convolution. In practice each convolution is computed as a Hadamard product in the frequency domain. Note that even though the OaAconv method calculates more transforms than the FFTconv method, the fact that the transforms are smaller greatly outweighs the cost of having to calculate more.

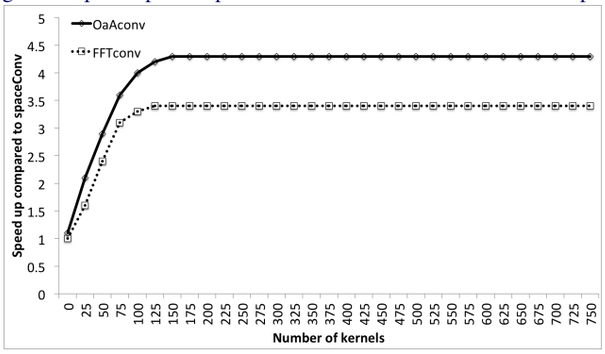
Figure 1: 1-D Overlap-and-Add convolutions



The overall time complexity of OaAconv can be further reduced by noting that all the block convolutions can be computed in parallel. If N^2 threads are available (a fair assumption for modern GPUs), the complexity on each thread is $n^2 \log_2 n$, and the overall time complexity is $\mathcal{O}(\max(N^2, n^2 \log_2 n))$ which is usually $\mathcal{O}(N^2)$. We can get additional speed up by taking advantage of the NVIDIA CUDA Fast Fourier Transform library (cuFFT) that computes each individual FFTs up to 10 times faster [3] (see also [5]). However, in order to have a fair comparison, our experiments in this paper run on single threads. As part of this work, we created a Caffe [1] fork[1] that uses a multi-thread GPU implementation of OaA for efficient convolutions.

We computed the increased in performance of FFTconv and OaAconv as we varied the number of kernels, the size of kernels, and the size of the input array for both forward and backward propagations. Figure 2 shows the speed-up factor of FFTconv and OaAconv compared to spaceConv in the forward propagation as the number of kernels is varied from 25 to 750 with a discrete step of 25. Each "number of kernels" experiment is repeated 10 times and the results are averaged. The input array is of size 32×32 and each kernel is of size 5×5. In this and most other experiments described in the full-paper, FFTconv and OaAconv outperform spaceConv, and OaAconv outperforms FFTconv at every step.

Figure 2: Speed-up over spaceConv vs. num of kernels in forward prop.



In future work we plan to optimize the FFT implementations for small size transforms, and design a CNN entirely in the frequency domain eliminating the bottleneck of the transforms. By creating a CNN that is in the frequency domain, the convolutional layers would only be the Hadamard products. The current challenge to this is to efficiently map an approximation of the non-linear transforms of CNNs (e.g., rectified linear units, sigmoid function, and/or hyperbolic tangent) to the frequency domain.

[1] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014.

[2] Michael Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through FFTs. *International Conference on Learning Representations*, 2014.

[3] CUDA Nvidia. Cufft library, 2010.

[4] Alan V Oppenheim, Ronald W Schafer, John R Buck, et al. *Discrete-time signal processing*, volume 2. Prentice-hall Englewood Cliffs, 1989.

[5] Nicolas Vasilache, Jeff Johnson, Michael Mathieu, Soumith Chintala, Serkan Piantino, and Yann LeCun. Fast convolutional nets with fbfft: A gpu performance evaluation. *International Conference on Learning Representations*, 2015.

[1] github.com/THighlander

Deep Q-learning for Active Recognition of GERMS: Baseline performance on a standardized dataset for active learning

Mohsen Malmir[1]
http://mplab.ucsd.edu/~mmalmir/
Karan Sikka[1]
http://mplab.ucsd.edu/~ksikka/
Deborah Forster[1]
dforster@ucsd.edu
Javier Movellan[2]
http://www.emotient.com/
Garrison W. Cottrell[3]
http://cseweb.ucsd.edu/~gary/

[1] Machine Perception Lab.
University of California San Diego,
San Diego, CA, USA

[2] Emotient, Inc.
4435 Eastgate Mall, Suite 320,
San Diego, CA, USA

[3] Computer Science and Engineering Dept.
University of California San Diego,
San Diego, CA, USA

Active object recognition (AOR) refers to problems in which an agent interacts with the world and controls its sensor parameters to maximize the speed and accuracy with which it recognizes objects. A wide range of approaches have been developed to re-position sensors or change the environment so that the new inputs to the system become less ambiguous [1, 2] with respect to goals such as 3D reconstruction, localization or recognition of objects. Many of the active object recognition methods are built around a specific hardware system, which makes the replication of their results very difficult. Other systems use off-the-shelf computer vision datasets, which include several views of objects captured by systematically changing object's orientation in the image. However, these datasets do not offer any active object recognition benchmark *per se*.

In this paper, we present and make publicly available the GERMS dataset (see Figure 1), that was specifically developed for active object recognition. The data collection procedure was motivated by the needs of the RUBI project, whose goal is to develop robots that interact with toddlers in early childhood education environments [4]. To collect data, we asked a set of human subjects to hand the GERM objects to RUBI in poses they considered natural. RUBI then pretends to examine the object by bringing it to its center of view and rotating the object. The background of the GERMS dataset was provided by a large screen TV displaying video scenes from the classroom in which RUBI operates, including toddlers and adults moving around.



Figure 1: Object set used in GERMS dataset. The objects represent human cell types, microbes and disease-related organisms.

We also propose an architecture (DQL) for AOR based on deep Q-learning (see Figure 2). To our knowledge, this is the first work employing deep Q-learning for active object recognition. An image is first transformed into a set of features using a DCNN borrowed from [3] which was trained on ImageNet. We add a softmax layer on top of this model to recognize GERMS objects; the output of this softmax layer is the belief over different GERMS objects given an image. This belief is combined with the accumulated belief from the previous images using Naive Bayes. This accumulated belief represents the *state* of the AOR system in each time step.

The accumulated belief is then transformed by the policy learning network into action values. This network is composed of two Rectified-Linear-Unit (ReLU) layers followed by a Linear-Unit (LU) layer. Each unit in the LU represents the action value for a given accumulated belief and one of the possible actions. In order to train this module, we implement the Q-learning iterative update:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{R(s,a) + \gamma \max_{a^*} Q(s^*,a^*) - Q(s,a)\} \qquad (1)$$

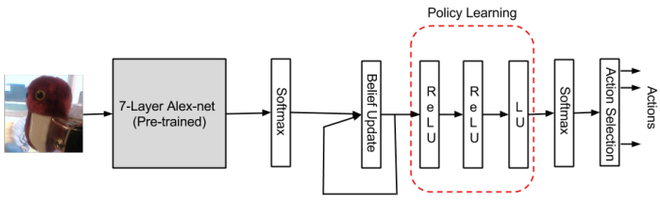


Figure 2: The proposed architecture for DQL.

into the following stochastic gradient descent weight update rule for the network:

$$W \leftarrow W - \lambda \left(R_t + \gamma \max_a Q(B_{t+1},a) - Q(B_t,a_t)\right) \frac{\partial}{\partial W} Q(B_t,a_t). \qquad (2)$$

Here, W is the weights of the policy learning network, $Q(s,a)$ is the action-value learned by the network for action a in state s, γ is the reward-discount factor and R_t is the reward at the tth time step.

The number of output units in the policy learning network is equal to the number of possible actions. Each output unit calculates the action value $Q(s,a)$ for one action a. We implemented a set of actions which rotate the robot's wrist from its *current position* by $\pm\pi/64$, $\pm\pi/32$, $\pm\pi/16$, $\pm\pi/8$, $\pm\pi/4$. The allowable range of rotation for both robot wrists is in $[0,\pi]$.

Table 1 shows the superior performance of the proposed DQL method compared to *random* and *sequential* action selection strategies. For detailed description of the dataset and the training algorithm, please refer to the paper.

Table 1: Number of steps required by the sequential, random and DQL policies to reach the same level of prediction accuracy on GERMS dataset.

Method \ Prediction Accuracy(%)	48	53	55	58	62	
Sequential	18	30	-	-	-	**Right Arm**
Random	2	4	6	10	-	
DQL	**1**	**2**	**2**	**3**	**10**	
Sequential	15	24	-	-	-	**Left Arm**
Random	3	10	18	-	-	
DQL	**1**	**3**	**3**	**7**	-	

[1] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4):333–356, 1988.

[2] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8): 966–1005, 1988.

[3] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[4] Mohsen Malmir, Deborah Forster, Kendall Youngstrom, Lydia Morrison, and Javier R Movellan. Home alone: Social robots for digital ethnography of toddler behavior. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 762–768. IEEE, 2013.

Camera Pose and Focal Length Estimation Using Regularized Distance Constraints

Ekaterina Kanaeva[1]
kanaeva.katerina6@gmail.com
Lev Gurevich[2]
lev@divisionlabs.com
Alexander Vakhitov[1]
a.vakhitov@spbu.ru

[1] Chair of Software Engineering
St. Petersburg State University
[2] Digital Vision Labs LLC
St. Petersburg

Camera pose and intrinsic parameters estimation from n 2D-to-3D point correspondences is a known problem in computer vision and photogrammetry. Depending on the set of unknown parameters, the problem is called Perspective-n-Point (PnP) when only absolute camera pose is unknown or PnPf when focal length is unknown as well. Projection error functions are highly non-convex in focal length, so before methods for PnPf were published, the only choice was to do exhaustive search not suitable for real-time applications. The EPnP method was extended to PnPf problem in [4], we refer to this method as UPnPf. RPnP inspired the authors of [5] to propose a method GPnPf+GN for PnPf problem. They use angle constraints to build specific polynomial system and solve it, then they use nonlinear refinement with Gauss-Newton algorithm. It gave superior results to [4] both in speed and accuracy in general case, and was more accurate in planar case, although UPnPf [4] was faster.

This paper is devoted to a method for PnPf problem for arbitrary amount of points, more or equal to 6. We consider both planar and non-planar cases. We fix the space of the search as a linear combination of several right singular vectors of the least squares system matrix. We use linear programming techniques to find feasible solutions faster. Then we do nonlinear refinement with Levenberg-Marquardt.

The barycentric representation of 3D points allows to express n 3D points $\mathbf{p}_i$ as a frame-independent linear combination of 4 basis points $\mathbf{c}_j$:

$$\mathbf{p}_i = \sum_{j=1}^{4} \alpha_i^{(j)} \mathbf{c}_j, \;\; i = 1 \ldots n, \;\; \sum_{j=1}^{4} \alpha_i^{(j)} = 1.$$

Each point projection as described in [2, 4] leads to two independent linear w.r.t. basis points' coordinates equations. The equations form a system with matrix M. The solution lies in a null-space (kernel) of M and can be expressed as a linear combination of kernel basis with coefficients β_i

$$\mathbf{x} = \sum_{i=1}^{N} \beta_i \mathbf{q}_i, \tag{1}$$

where $\mathbf{q}_i$ are the right-singular vectors of $\mathbf{M}$ corresponding to the N null singular values of $\mathbf{M}$. There are distance constraints which need to be satisfied:

$$\|\mathbf{c}_i^c - \mathbf{c}_j^c\|^2 = \|\mathbf{c}_i^m - \mathbf{c}_j^m\|^2 = r_{ij}^2, \tag{2}$$

where r_{ij} is a known distance between i-th and j-th basis points. Distance constraints are quadratic w.r.t. β_i and in the same time are linear w.r.t. $\mathbf{b}$:

$$\mathbf{b} = \begin{pmatrix} \beta_1^2 & \ldots & \beta_N^2 & \beta_1\beta_2 & \ldots & \beta_{N-1}\beta_N \end{pmatrix}^T. \tag{3}$$

While the EPnP method tries to solve the constraints system (2), we solve (in a set defined by (1)) a least squares problem

$$F_R(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|^2 + \gamma \|\mathbf{L}\mathbf{b} - \mathbf{r}\|^2 \to \min, \tag{4}$$

where γ is some coefficient. Distance constraints for the PnPf problem have the form:

$$r_{ij}^2 = \sum_{k=1}^{3} \|\mathbf{c}_i^{(k)} - \mathbf{c}_j^{(k)}\|^2 = \frac{1}{f^2}((\mathbf{d}_i^{(1)} - \mathbf{d}_j^{(1)})^2 + (\mathbf{d}_i^{(2)} - \mathbf{d}_j^{(2)})^2) + (\mathbf{g}_i^{(3)} - \mathbf{g}_j^{(3)})^2, \tag{5}$$

for definition of $\mathbf{d}_i, \mathbf{g}_i$ see paper. Constraint is quadratic in $\mathbf{d}_i, \mathbf{g}_i$, but if we choose the unknowns vector $\mathbf{b}$ analogously to (3), it becomes linear with $\mathbf{b^1}$ equal to $\mathbf{b}$ given in (3) and:

$$\mathbf{b^2} = f^2 \mathbf{b^1}, \;\; \mathbf{b}^T = (\mathbf{b^1}^T, \mathbf{b^2}^T)^T. \tag{6}$$

So, for the PnPf problem we get the analogous function as (4) for PnP problem as described in the paper. In the Algorithm described in the paper, for each $N = 1, 2, 3$ we find candidate solutions (1), for which we formulate additional linear constraints and solve using MATLAB's *linprog*.

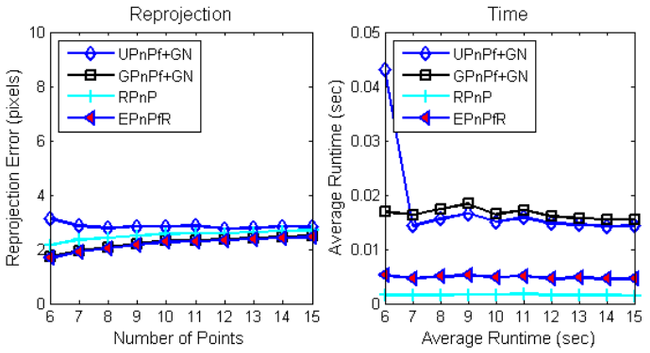


Figure 1: Mean reprojection error and time of the PnPf methods w.r.t. varying noise level, points in general 3D configuration

After a loop over N, we choose the solutions which have the least amount of points lying behind the camera, and among these solutions choose the by comparing the value of $F_R(x)$ (4). This chosen solution is subsequently refined by Levenberg-Marquardt procedure [3].

Algorithm implementation is available at `http://sites.google.com/site/alexandervakhitov/projects/epnpfr`.

We made two sets of experiments: comparative test and a test using real images. Synthetic experiments test performance of the method w.r.t. varying noise level and point number. Here we show at the figure the results for some of measured parameters for general point configuration.

The aim is to demonstrate applicability of the algorithm in a real setting. We made three shots of a non-moving scene with Nikon D3100 camera with several focal length settings. We performed standard structure-from-motion reconstruction using one initial frame pair.

We matched the SIFT points from one of the frames in the initial pair and every other frame. We ran our algorithm and the best state-of-art GPnPf+GN algorithm with the RANSAC loop, choosing 6 points [1]. When both methods returned results, they were different less than for 1% in focal length and reprojection error, except 105 mm focal length (average error GPnPf+GN 2.53 pix, EPnPfR 0.77 pix).

Proposed algorithm is as accurate and stable as the state-of-the-art methods, and more than 2 times faster.

[1] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[2] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81(2):155–166, 2009.

[3] M.I.A. Lourakis. levmar: Levenberg-marquardt non-linear least squares algorithms in C/C++. [web page] `http://www.ics.forth.gr/~lourakis/levmar/`, Jul. 2004. [Accessed on 31 Jan. 2005.].

[4] A. Penate-Sanchez, J. Andrade-Cetto, and F. Moreno-Noguer. Exhaustive linearization for robust camera pose and focal length estimation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):2387–2400, 2013.

[5] Y. Zheng, S. Sugimoto, I. Sato, and M. Okutomi. A general and simple method for camera pose and focal length determination. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 430–437. IEEE, 2014.

Classifying Global Scene Context for On-line Multiple Tracker Selection

Salma Moujtahid
salma.moujtahid@liris.cnrs.fr
Stefan Duffner
stefan.duffner@liris.cnrs.fr
Atilla Baskurt
atilla.baskurt@liris.cnrs.fr

Université de Lyon, CNRS, INSA-Lyon,
LIRIS, UMR5205,
F-69621, France

In this paper, we present a novel framework for combining independent on-line trackers using visual scene context. The aim of our method is to decide automatically at each point in time which specific tracking algorithm works best under the given scene or acquisition conditions.

In the literature, many ways of combining, fusing or selecting visual features have been presented. For example, low-level fusion of features (like motion or shape) is applied to improve the foreground-background discrimination (*e.g.* [2, 8]). Fusion is also possible at a higher level, where several trackers are run in parallel in order to select or combine their respective results (*e.g.* [1, 5]). In terms of model or feature fusion, our previous work Moujtahid *et al.* [6] concentrated on using confidence values of several individual trackers with different visual features coupled with a spatial-temporal coherence criteria to select the most suitable tracker at a given instant and enforce the continuity of tracking.

The main idea behind our framework is to use the strengths of different tracking algorithms as well as scene context information in order to improve the tracking performance. To this end, we introduce a framework that combines several independent and complementary trackers, each specialised on different scene conditions. The decision on which tracker to select is proposed by an off-line trained classifier which, in turn, is based on general scene context features that are independent from the trackers.

The general procedure of the proposed tracking framework is illustrated in Fig. 1. On a given video, N independent trackers $T_n, (n \in 1..N)$ run in parallel and, at every frame t, produce each an estimate of the object's state. This is usually a bounding box B_t^n with an associated confidence value $c_{t,n}$. The objective is to select at each frame the best tracker, *i.e.* the one that outputs the bounding box that fits best the object to track.

At the same time, the scene context features $\mathbf{f}_t$ are extracted. These features correspond to first and second order statistics of a given image-related variable such as intensity, colour and motion. They are computed on different image regions giving local, global and differential values.

The scene features $\mathbf{f}_t$ are concatenated with additional measures like the trackers confidences values $\mathbf{c}_t$ and the identifier of the last selected tracker s_{t-1} to form a large feature vector $\mathbf{i}_t$. An N-class classifier, that has been trained off-line on annotated data, is then applied on these features to estimate the best tracker for the given scene context. The classifier responds with $\mathbf{y}_t$, a probability for each class which is subsequently filtered by a Hidden Markov Model to ensure the temporal continuity of the tracker selection and reject outliers. The HMM estimates the posterior probability distribution $\mathbf{x}_t$, which is used to select the best tracker.

Finally, a Kalman Filter is applied as a post-processing step to temporally smooth the resulting object bounding box B_t^s from the selected trackers T_s. The result of the Kalman filter represents the final output of our tracking algorithm, and is further used to update the models of the individual trackers T_n.

Apart from this last update step, all the trackers are completely independent and do not cooperate or interact with each other. It is also important to mention that this approach is very generic, and in theory *any* on-line tracking algorithm can be integrated in this framework.

For our experiments, we used 3 on-line AdaBoost trackers [3] with different visual cues : Haar like features (HAAR), Histograms of Oriented Gradients (HOG) and Histograms of Colour (HOC). We also chose a fully connected Multi-Layer Perceptron (MLP) as scene context classifier.

We evaluated the performance of our framework on the Visual Object Tracking (VOT2013) benchmark [4]. First, we measured the classification rate of the proposed scene context classifier trained on the Princeton Dataset [7] and achieved a 81.80% rate of prediction. Secondly, we evaluated the overall tracking algorithm on the Visual Object Tracking (VOT2013) analysing the contribution of the different components, *i.e.* scene context classifier, HMM, and Kalman filter and comparing the proposed framework with other state-of-the-art tracking algorithms.

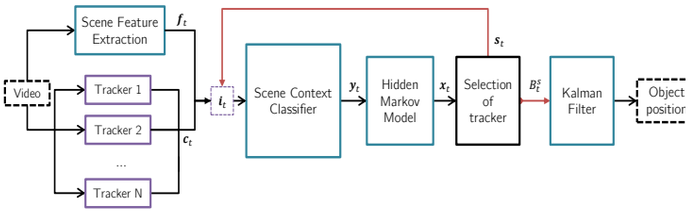


Figure 1: Overall framework of the proposed scene context-based tracking algorithm.

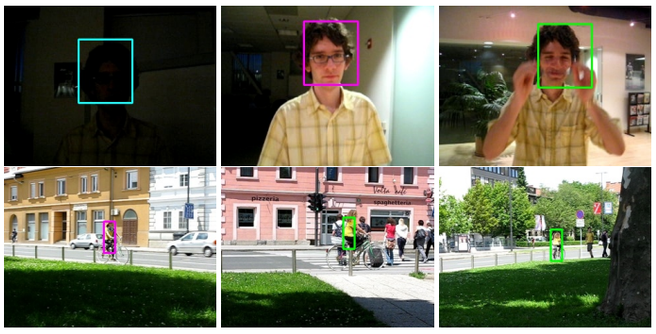


Figure 2: Illustration of our proposed framework's tracking results on the "David"(1st row) and "Bicycle"(2nd row) videos. Different scene context variations in lighting, texture or background are present throughout the videos. In the first part of the "David" video, the lighting is very poor and texture-based trackers usually work better, whereas in the second part the texture changes due to varying pose so colour-based trackers are more suitable. Our framework selects the most suitable tracker in each scenario (*pink*: HAAR, *blue*: HOG, *green*: HOC).

The proposed algorithm proved to increase the performance of our individual trackers, ranking among the top trackers of the VOT2013 challenge in terms of robustness.

[1] Christian Bailer, Alain Pagani, and Didier Stricker. A superior tracking approach: Building a strong tracker through fusion. In *Proc. of ECCV*, pages 170–185, 2014.

[2] Robert T. Collins and Yanxi Liu. On-line selection of discriminative tracking features. *IEEE Trans. on PAMI*, 27(10):1631–1643, 2005.

[3] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *Proc. of BMVC*, pages 47–56, 2006.

[4] Matej Kristan, Luka Cehovin, Roman Pflugfelder, Georg Nebehay, Gustavo Fernandez, Jiri Matas, and et al. The Visual Object Tracking VOT2013 challenge results. In *Proc. of ICCV (Workshops)*, 2013.

[5] Ido Leichter, Michael Lindenbaum, and Ehud Rivlin. A general framework for combining visual trackers – "black boxes" approach. *IJCV*, 67(3):343–363, March 2006.

[6] Salma Moujtahid, Stefan Duffner, and Atilla Baskurt. Coherent selection of independent trackers for real-time object tracking. In *VISAPP*, pages 584–592, 2015.

[7] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *Proc. of ICCV*, 2013.

[8] Alper Yilmaz, Xin Li, and Mubarak Shah. Object contour tracking using level sets. In *Proc. of ACCV*, 2004.

Sketch based Image Retrieval using Learned KeyShapes (LKS)

Jose M. Saavedra
jose.saavedra@orand.cl
Juan Manuel Barrios
juan.barrios@orand.cl

Orand S.A.
Estado 360, Of. 702-A,
Santiago, Chile

A natural alternative for querying in an image retrieval system is by simply drawing what one has in mind. Indeed, drawing was the primitive means of communication between humans. One of the goals of an image retrieval scenario is to provide users a simple modality for querying. Thereby, a drawing means a simple hand-drawn sketch composed only of strokes that users can do easily, lacking color or texture. Examples of hand-drawn sketches are shown in Figure 1.



Figure 1: Examples of hand-drawn sketches. The last two are from the Eitz's dataset [1].

This querying modality leads to the *sketch based image retrieval problem* (SBIR) which is a challenging problem because of two main reasons: (i) images that we want to retrieve are not sketches, (ii) query sketches show certain level of ambiguity by nature that may make a method get confused easily. Consequently, state-of-the-arts SBIR approaches [2, 4] still show low performance.

Therefore, taking some ideas of the human visual perception, we present a novel method for sketch based image retrieval. Our method, is based on detecting the occurrences of mid-level patterns on a sketch. To this end, we figure out a set of patterns (*learned keyshapes*) by means of an unsupervised learning process. We then build a histogram that counts the occurrences of the patterns in the underlying sketch. The histogram is built using soft-voting, spatial division and squared root normalization. We show new state-of-the-art results in two available datasets doubling the precision achieved by current methods.

Our proposal consists of two stages (Figure 2): (1) figure out a set of *keyshapes*, (2) generate the LKS descriptors based on the detected set of *keyshapes*, that will be used later for similarity search.

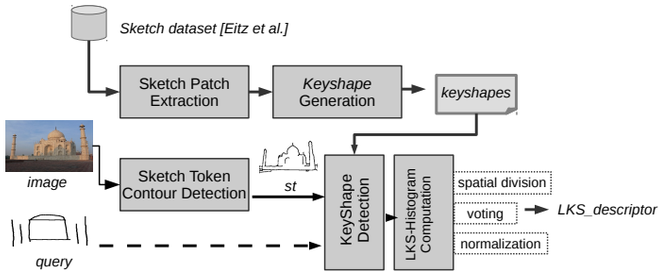


Figure 2: A scheme of our Learned KeyShapes based proposal (LKS).

Keyshape Generation: We got inspired by the work of Lim et al. [3], where reliable contour maps are obtained by detecting a set of *sketch tokens*, that are previously learned from a collection of contour images. For training, we use the Eitz's sketch dataset [1]. We extract one million of 31×31 sketch patches, each one centered in a stroke point. Sketch patches are coded by a DAISY descriptors that are then clustered by K-means. In Figure 3, we show a sample of learned *keyshapes* using $K = 150$.

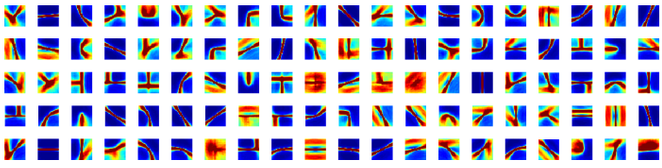


Figure 3: A sample of learned keyshapes when $K = 150$.

Keyshape Generation: Instead of using low level methods as Canny, we prefer to use the sketch token based approach proposed by Lim et al. [3].
KeyShape Detection: We extract sketch patches from the input query as well as from the contour of a test image. Each patch is centered in each stroke point. For each patch, we search for the P nearest *keyshapes* (P-NK) using DAISY descriptors.
LKS-Histogram Computation We build a K-dimensional histogram, where K is the number of keyshapes. In this process, three steps are involved: (1) Gaussian Kernel based Voting using P-NK, (2) Spatial Division, (3) Pow-based Normalization.

We evaluate our proposal in two public datasets: Saavedra's and *Flickr15k*. The performance are compared under the mean average precision metric (mAP) (Table 1) as well as the precision-recall graphic (Figure 4).

	HOG	GF-HOG[2]	SHELO[4]	**LKS**	gain
Saavedra's	0.2355	*unreported*	0.2766	**0.3251**	17.5%
Flickr15K	0.0771	0.1222	0.1236	**0.2450**	98.2%

Table 1: Mean Average Precision comparing our proposals LKS with state-of-the-art methods.

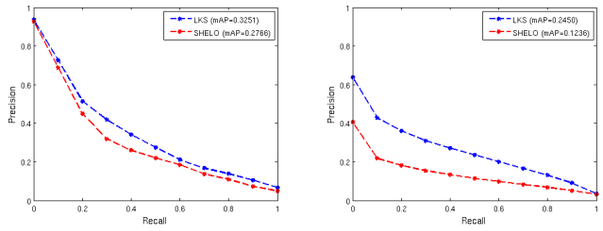


Figure 4: Precision-Recall graphic showing the performance of LKS (blue curve) and SHELO (red curve) on the Saavedra's dataset (on the left) and Flickr15k dataset (on the right).

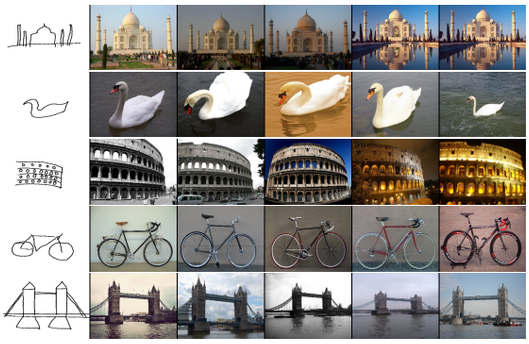


Figure 5: Examples of results using LKS on the Flickr15K dataset. Each row shows a query sketch together with the five first responses.

[1] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44:1–44:10, July 2012.

[2] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117(7):790–806, July 2013.

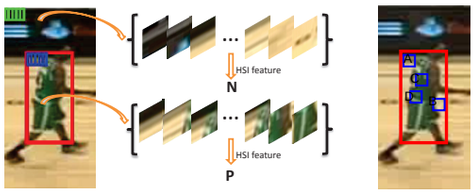
[3] J.J. Lim, C.L. Zitnick, and P. Dollar. Sketch tokens: A learned mid-level representation for contour and object detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3158–3165, June 2013.

[4] Jose M. Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *International Conference on Image Processing, ICIP'2014 (To appear)*, 2014.

Online Visual Tracking via Coupled Object-Context Dictionary

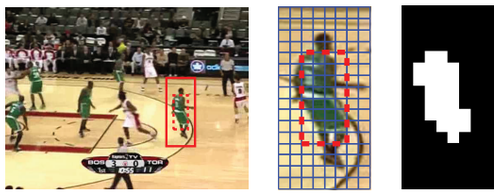
Mingquan Ye[12]
mingquan.ye@vipl.ict.ac.cn
Hong Chang[1]
hong.chang@vipl.ict.ac.cn
Xilin Chen[1]
xilin.chen@vipl.ict.ac.cn

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China



(a) (b)

Figure 1: Illustration of constructing the coupled dictionaries. The red rectangles in (a) and (b) represent the target bounding boxes. The green squares in (a), which are generated by sliding windows outside the target bounding box, correspond to basis patches involved in the background dictionary $\mathbf{N}$. The blue squares in (a) are generated inside the target bounding box in a similar way, which constitute the noisy target dictionary $\mathbf{P}$.



(a) (b) (c)

Figure 2: Illustration of computing the confidence map. (a) The t-th frame image. (b) The context window Ω is divided into nonoverlapping patches, and the red dashed line represents the target bounding box at frame $t-1$. (c) The confidence map of Ω at frame t.

Motivation: In conventional sparse coding based tracking algorithms, the dictionary is composed of holistic target templates and background templates [2]. The target templates always contain some background parts due to the non-rigidness of the tracked object. The background templates are produced around the labeled target position with big perturbations, but some target contents may still be included. Such circumstances will certainly decrease the discriminative ability of trackers. Found on the above situation, we tend to construct pure background and target dictionaries.

Constructing Pure Coupled Dictionaries: The outside of the target bounding box is pure background, while the inside includes both target and background regions [1]. Hence, we first construct the pure background dictionary with the patches in the outside context region. As shown in Figure 1, the green squares, which are generated by sliding windows in the context region, correspond to basis patches involved in the background dictionary $\mathbf{N} \in \mathbb{R}^{m\times n}$. The blue squares, as shown in Figure 1 (a), constitute a noisy target dictionary $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_l] \in \mathbb{R}^{m\times l}$. Then, we pick out real target patches from $\mathbf{P}$ based on their coding quality with respect to dictionary $\mathbf{N}$.

$$\widehat{\mathbf{C}} = \arg\min_{\mathbf{C}} \|\mathbf{P} - \mathbf{N}\mathbf{C}\|_2^2 + \lambda \sum_{i=1}^{l} \|\mathbf{c}_i\|_1, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{n\times l}$ is the coefficient matrix with the i-th column $\mathbf{c}_i$ being the sparse coding of $\mathbf{p}_i$. Then the set of reconstruction errors is expressed as $\mathcal{R} = \{\|\mathbf{p}_i - \mathbf{N}\widehat{\mathbf{c}}_i\|_2^2\}_{i=1}^{l}$. As illustrated by Figure 1 (b), the patches A and B belonging to background have small reconstruction errors. In contrast, C and D are parts of the target object, thus have big reconstruction errors. Therefore, we choose l' ($l' = \lfloor \beta \times l \rfloor$ and $\beta \in (0,1)$) basis vectors with top l' highest reconstruction errors to construct the pure target dictionary $\mathbf{P}'$, which is expressed as $\mathbf{P}' = [\mathbf{p}_{i_1}, \ldots, \mathbf{p}_{i_{l'}}]$, where $i_1, \ldots, i_{l'}$ are the indexes of the top l' values in $\mathcal{R}$.

Computing Confidence Map: The target patches should have big and small reconstruction errors when encoded by $\mathbf{N}$ and $\mathbf{P}'$, respectively, while the background patches have the reversed situations. At online tracking in t-th frame, we construct the confidence map of the context region Ω which is centered at the target position of frame $t-1$. Specifically, we normalize and divide Ω into q nonoverlapping patches of same size as those in the dictionaries (as shown in Figure 2 (b)), denoted as $\mathbf{O} \in \mathbb{R}^{m\times q}$. Then the q patches are encoded by $\mathbf{N}$ and $\mathbf{P}'$ as follows:

$$\widehat{\mathbf{C}}_1 = \arg\min_{\mathbf{C}_1} \|\mathbf{O} - \mathbf{N}\mathbf{C}_1\|_2^2 + \lambda \sum_{i=1}^{q} \|\mathbf{c}_{1i}\|_1, \tag{2}$$

$$\widehat{\mathbf{C}}_2 = \arg\min_{\mathbf{C}_2} \left\|\mathbf{O} - \mathbf{P}'\mathbf{C}_2\right\|_2^2 + \lambda \sum_{i=1}^{q} \|\mathbf{c}_{2i}\|_1, \tag{3}$$

where $\mathbf{C}_1 \in \mathbb{R}^{n\times q}$, $\mathbf{C}_2 \in \mathbb{R}^{l'\times q}$ are the coefficient matrices, $\mathbf{c}_{1i}$ and $\mathbf{c}_{2i}$ are the i-th column vectors of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively.

We define the score (confidence value) of the i-th patch as

$$s_i = \|\mathbf{o}_i - \mathbf{N}\widehat{\mathbf{c}}_{1i}\|_2^2 - \left\|\mathbf{o}_i - \mathbf{P}'\widehat{\mathbf{c}}_{2i}\right\|_2^2, i = 1, \ldots, q. \tag{4}$$

In order to alleviate the negative effects caused by outliers, we adjust the patch score as below,

$$s_i' = \begin{cases} 0, \ s_i \le 0 \text{ or } \sum_{j\in\mathcal{N}(i)} \mathbf{1}(s_j \le 0) \ge \tau \\ 1, \ \text{otherwise} \end{cases}, i = 1, \ldots, q, \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\mathcal{N}(i)$ denotes the set of 8 neighbors of the i-th patch. We assign the score of each patch to all the pixels inside it, then the confidence map of the context region Ω can be obtained, as shown in Figure 2 (c).

Bayesian Tracking Framework: Our algorithm is under the Bayesian sequential estimation framework, which performs tracking by solving the maximum a posterior (MAP) problem. The observation model $p(y_t|x_t)$ is modeled by

$$p(y_t|x_t^i) \propto \sum_{(j,k)\in B_i} s_{j,k}, \tag{6}$$

where x_t^i represents the i-th sample (particle) of state x_t, B_i is its corresponding bounding box, and $s_{j,k}$ is the pixel score at location (j,k).

Empirical Results: We provide the precision and success plots over all the 12 tested sequences, as shown in Figure 3.

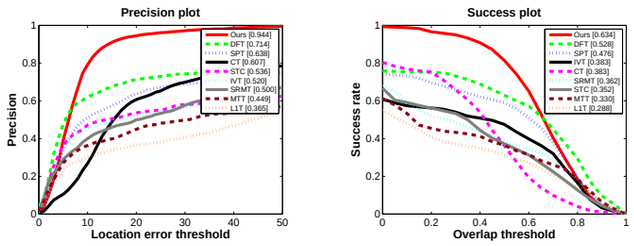


Figure 3: Precision and success plots over all the 12 tested sequences. The precision score of each tracker at 20 pixels is listed in the legend of the left plot. The right plot presents the area-under-the-curve (AUC) score for each method.

[1] Junseok Kwon, Junha Roh, Kyoung Mu Lee, and Luc Van Gool. Robust visual tracking with double bounding box model. In *ECCV*, pages 377–392. Springer, 2014.

[2] Naiyan Wang, Jingdong Wang, and Dit-Yan Yeung. Online robust non-negative dictionary learning for visual tracking. In *ICCV*, pages 657–664. IEEE, 2013.

Subspace Alignment Based Domain Adaptation for RCNN Detector

Anant Raj
anantraj@iitk.ac.in
Department of Electrical Engineering,
IIT Kanpur, India

Vinay P. Namboodiri
vinaypn@iitk.ac.in
Department of Computer Science and Engineering,
IIT Kanpur, India

Tinne Tuytelaars
tinne.tuytelaars@esat.kuleuven.be
ESAT, PSI-VISICS,
KU Leuven, Belgium

It has been an underlying assumption behind most of the machine learning algorithms that training and test instances should be sampled from a similar distribution. The difference in resolution, difference in view points, clutter and background are the main reasons which can cause the problem of domain shift. Indeed, the problem is particularly pertinent to the computer vision community due to our reliance on 'standard' challenging datasets. However most of the work in this field revolves around adapting a classifier for the task of object recognition or classification and not much effort has been put to adapt an object detector. In this paper, we propose subspace alignment based domain adaptation of the state of the art RCNN based object detector [2]. In subspace based domain adaptation for objects, we need access to source and target subspaces for the bounding box features . The absence of supervision (labels and bounding boxes are absent) makes the task challenging.

The proposed approach builds upon the previously proposed subspace alignment based method [1] for visual domain adaptation to adapt the RCNN detector [2].

1 Subspace Alignment

Subspace alignment based domain adaptation method consists of learning a transformation matrix M that maps the source subspace to the target one. Suppose, we have labelled source data S and unlabelled target data T. We normalize the data vectors and take separate PCA of the source data vectors and target data vectors. The d eigenvectors for each domain are selected corresponding to the d largest eigenvalues. We consider these eigenvectors as bases for source and target subspaces separately. They are denoted by X_s for source subspace and X_t for target subspace. We use a transformation matrix M to align the source subspace X_s to target subspace X_t. The mathematical formulation to this problem is given by

$$F(M) = \|X_S M - X_T\|_F^2 \qquad M^* = \underset{M}{\operatorname{argmin}}(F(M)). \tag{1}$$

X_s and X_t are matrices containing the d most important eigenvectors for source and target respectively and $\|.\|_F$ is the *Frobenius norm*. The solution of eqn. 1 is $M^* = X_S' X_T$ and hence for the target aligned source coordinate system we get $X_a = X_S X_S' X_T$. Once we get target aligned source co-ordinate system, we project our source data and train the classifier in this frame. While testing, target data is projected on the target subspace and classifier score is calculated.

2 RCNN-detector

Convolutional neural nets (CNN) and other deep learning based approaches have improved the object classification accuracy by a large margin. RCNN [2] uses the CNN framework and bridges the gap between object classification and object detection task. The idea of this work is to see how well the result of convolutional neural network on ImageNet task generalizes for the task of object detection on PASCAL dataset. RCNN consists of three modules. The first module generates selective search windows [3] in an image which is category independent. Second module extracts mid level convolutional neural network features for each proposed region which has been trained earlier on ImageNet dataset. In the third module, SVM classifier is trained by considering all those windows whose overlap with the ground truth bounding box are less then a threshold λ as negative examples, hard negative examples are mined from these negative examples during the training. In testing phase, again 2000 selective search windows is generated per image in fast mode. Each proposal is warped and propagated forward through pre-trained CNN to compute features. Then, for each class, the learned SVM class specific classifier is applied to those extracted features and a score is obtained corresponding to each proposal. Once we get the scores, we decide a threshold and the regions with scores greater than the decided threshold are our possible candidates for a particular object category. In the last step greedy non maximum suppression is applied to obtain desired, accurate and specific bounding box for that object category.

3 Subspace Alignment for Adapting RCNN

In this section we describe our approach to adapt the class specific RCNN -detector. On the basis of background provided in the previous section, we use subspace alignment based domain adaptation over the initial RCNN-detector. Instead of using single subspace for the full source and target data, we postulate that using class-specific different subspaces for different classes to adapt from source to target domain improves the object detection accuracy.

Algorithm 1 Subspace Alignment based Domain Adaptation for RCNN Detector

procedure SA BASED RCNN ADAPTATION(Source Data S,Target Data T)
  for each image $j \in$ *Source* and *Target Image* **do**
    $Windows(j) \leftarrow ComputeSelectiveSearchWindows(j)$
    $feat(j) \leftarrow ComputeCaffeFeat(Windows(j))$
  end for
  $InitRCNNdetector \leftarrow TrainRCNNonSource(SourceData)$
  for each class $i \in$ *Object Class* **do**
    $PosSrc(i) = ()$ and $PosTgt(i) = ()$
    for each image $j \in$ *Source* and *Target Image* **do**
      $ol(j) = ComputOverlap(gTBbox(j,i), Windows(i))$ ▷ For source images
      $PosSrc(i) = \text{Stack}(PosSrc(i), feat(i)(ol(j) \geq \gamma)$ ▷ For source images
      $score(i,j) = runInitRCNNdetector(image(j))$ ▷ For target images
      $PosTgt(i) = \text{Stack}(PosTgt(i), feat(i)(score(i,j) \geq \sigma)$ ▷ For target images
    end for
    $X_{source}(i) \leftarrow PCA(PosSrc(i))$
    $X_{target}(i) \leftarrow PCA(PosTgt(i))$
  end for
  for each class $i \in$ *Object Class* **do**
    $ProjectMat(i) \leftarrow SubspaceAlign(X_{source}(i), X_{target}(i))$
  end for
  $AdaptedRCNNdetector \leftarrow TrainRCNNonSource(ProjectedSrcDat$
  $boxes \leftarrow runAdaptedRCNNdetector(ProjectedTgtData)$
  $predictBbox \leftarrow runNonMaximumSupression(boxes)$
  return *predictBbox*
end procedure

[1] Fernando, Basura, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. "Unsupervised visual domain adaptation using subspace alignment." In ICCV, 2013 pp. 2960-2967. IEEE, 2013

[2] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014, June). Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014 (pp. 580-587). IEEE

[3] Uijlings, J. R., van de Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2), 154-171.

Time-slice Prediction of Dyadic Human Activities

Maryam Ziaeefard
http://vision.gel.ulaval.ca/en/people/Id_842/index.php
Department of Electrical Engineering, Laval University

Robert Bergevin
http://vision.gel.ulaval.ca/en/people/Id_18/index.php
Department of Electrical Engineering, Laval University

Louis-Philippe Morency
http://www.cs.cmu.edu/~morency/
School of Computer Science, Carnegie Mellon University

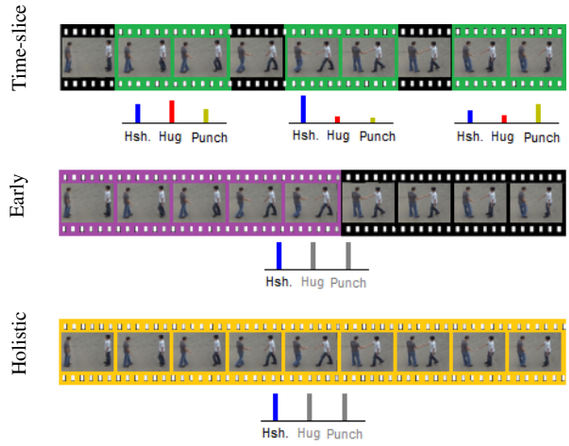


Figure 1: An illustration of human activity recognition problems: The first row illustrates "time-slice" recognition and the labels, i.e., Handshake (Hsh.), Hug, and Punch for different time-slices. The second and third rows show "early" recognition and "holistic" approaches where the label is the same for the whole sequence.

Recognizing human activities from video data is being leveraged for surveillance and human-computer interaction applications. In this paper, we introduce the problem of time-slice activity recognition which aims to explore human activity at a smaller temporal granularity. Time-slice recognition is able to infer human behaviors from a short temporal window. It has been shown that the temporal slice analysis is helpful for motion characterization and in general for video content representation. These studies motivate us to consider time-slices for activity recognition.

We present in Figure 1 an overview of our approach based on time-slice action prediction and contrast it with the conventional approaches which recognize actions based on either the whole video sequence (referred as "holistic" approach) or the first part of it (early recognition). Our time-slice approach studies not only the beginning of the action sequence but generalizes this to any short-term observation anywhere in the video sequence. Another key novelty is in the explicit modeling of the uncertainty occurring when predicting actions based on time-slices.

TAP Dataset: We introduce a new dataset, named Time-slice Action Prediction (TAP) dataset, to evaluate our proposed feature descriptors and enable future research on this topic. The dataset was created by extracting time-slices from existing public human action datasets (UT-Interaction, HMDB, TV Interaction, and Hollywood datasets) and perform a perception study with multiple annotators giving continuous ratings for each action. The continuous ratings allow to represent the uncertainty in time-slice action prediction. 3 annotators rated each time-slice on how likely a specific action is occurring. For each time-slice and for each action, the annotator was asked to pick one of 5 likelihoods from "Definitely Not Occurring" to "Definitely Occurring". Figure 3 illustrates how annotators rated for two example videos.

Methodology: **Stage 1- Discriminative segments:** When analyzing an interaction, we can definitely recognize the ongoing activity from specific time slices such as "two people are shaking each other's hands" slice in handshaking activity. To extract discriminative segments from our dataset, we used Fleiss' kappa coefficient k [2] to measure the reliability of agreement between annotators. For each interaction video, time-slices where the annotators are in complete agreement, i.e. k=1, on definitely including the interaction of interest, are selected as discriminative segments.

Stage 2- Predict-STIP: Existing STIP detectors are vulnerable to model the inherent uncertainty in partially observed action recognition and prediction, and therefore, are insufficient for time-slice recognition. We introduce Predict-STIPs which are active during the whole video. In other words, P-STIPs are the STIPs that exist in first frames of the video and still will appear in upcoming frames. Given a set of interaction video sequences $\{A_i \mid i = 1:n\}$ and their associated discriminative segments $\{S_i \mid i = 1:n\}$, we first detect a new subset of S-STIPs [1]. We then track them backward and forward to the first and last frames of the video and check whether or not they have existed during the whole video. We repeat these steps for all frames of a discriminative segment. Landmarks that are continuously observable are selected as P-STIPs

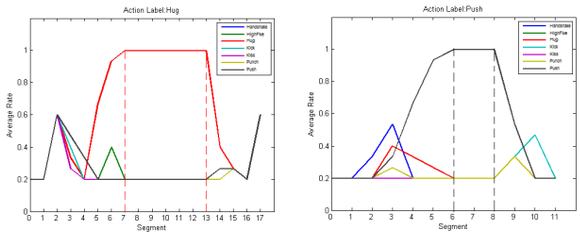


Figure 2: Human annotation: This figure shows the average rate of 3 annotators for two video examples: hug and push. The label provided by one annotator is converted to a number on a linear scale from 0 to 1 called the average rate. This average rate will be used to evaluate the performance of our method. Time-slices between dashed lines is the discriminative segment of the interaction.

	constrained set	unconstrained set
handshake	82%	76.3%
high five	–	61.4%
hug	81%	71%
kick	78%	73.7%
kiss	–	74%
punch	80%	76.2%
push	75%	–

Table 1: The average precision of Predict-STIP on constrained (UT-interaction dataset) and unconstrained sets (selected videos from HMDB, TV Interaction, and Hollywood TV show datasets).

Stage 3- Descriptors and vocabulary building: Given P-STIPs of each interaction video, we construct the descriptor vectors HOG3D over a set of gradient vectors from the cuboid neighborhood (4x4x4) around the P-STIPs. All histograms are concatenated to one descriptor vector for each video. We compute the basic Bag-of-words model and quantize the descriptor vectors into 1000 bins associated with visual words using K-means clustering. BoW features are normalized so their L1 norm is 1.

Results: At test time, a query video v_i which is a time-slice of a longer video matched to the models. To this intent, we extract S-STIPs [1] from v_i and match them to the pool of trained P-STIPs. S-STIPs of v_i that matched to P-STIPs are selected as P-STIPs of v_i (lookup table technique). Then BoW descriptors of v_i are extracted. Classification is made based on the score of interaction class-specific models applied on BoW descriptors. The average precision for all interactions (compared to human annotation) is given in the Table 1.

[1] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, and Jordi Gonzalez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396 – 410, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[2] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Adapting RANSAC-SVM to Detect Outliers for Robust Classification

Subhabrata Debnath
http://www.cse.iitk.ac.in/~sdebnath
Anjan Banerjee
http://www.cse.iitk.ac.in/~anjan
Vinay Namboodiri
http://www.cse.iitk.ac.in/~vinaypn

Department of Computer Science and Engineering
Indian Institute Of Technology Kanpur
Kanpur
UP, India

In this paper we address the problem of classifying objects where some of the labels in the training data are noisy. This is a common scenario and can be caused by the difficulty of annotation or inadvertently due to human error. In this paper, we consider the wrongly annotated examples to be outliers and try to formulate a robust outlier identification algorithm.
The task of learning a model in the presence of noise has been traditionally solved by the RANSAC algorithm[1]. RANSAC has also been adapted as RANSAC SVM by Nishida and Kurita [3]. The RANSAC-SVM method selects random subsets of the training data and trains small SVMs on them, using the rest of the training data as validation sets. It then chooses the SVM with the smallest validation error to approximate the full training set. However, if the training data is noisy, the validation sets are corrupted, and a faulty submodel may be chosen as optimal.
To address this problem, we propose a modification to RANSAC SVM thereby achieving robustness to noise. The detailed algorithm is given in Algorithm 1.

We are initially given a training set S of n examples from which we draw small random subsets of size k. For each such subset, we train a SVM to obtain a weight vector w_i using the standard support vector formulation given by:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \qquad (1)$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i; \quad \xi_i \geq 0, i = 1, \ldots, l.$$

We then classify the whole set of examples S using w_i and record the predicted labels. The misclassifications that result from this weight vector are aggregated in a vector O of dimension n. The procedure is repeated for many iterations m and examples with misclassifications beyond τ are considered to be outliers. Thus using a quantized score from many models provides us with a better estimate of outlierness.

Algorithm 1 Outlier Robust RANSAC-SVM Adaptation

procedure OUTLIER ROBUST RANSAC-SVM(SetSize k, NumIteration m,Threshold τ)
  for each instance $x_j \in$ Training Set S **do**
    OutlierScore $O(x_j) \leftarrow 0$
  end for
  for i=1 to NumIteration m **do**
    Choose a random $X_i \subset$ Training Set S s.t. $|X_i|$=SetSize k
    $w_i \leftarrow BuildSVMModel(X_i)$ using equation 1.
    $MisclassifiedSet_i \leftarrow GetMisclassifiedInstances(w_i, S)$
    for each instance $x_j \in MisclassifiedSet_i$ **do**
      $IncrementByOne$(OutlierScore $O(x_j)$)
    end for
  end for
  OutlierSet $S_o \leftarrow \{\}$
  InlierSet $S_i \leftarrow \{\}$
  for each instance $x_j \in$ Training Set S **do**
    if OutlierScore $O(x_j) >$ Threshold τ **then**
      $AddToSet$(OutlierSet S_o, x_j)
    else
      $AddToSet$(InlierSet S_i, x_j)
    end if
  end for
  InlierModel $w_{in} \leftarrow BuildSVMModel(InlierSet\ S_i)$
  AdditionalInliers $S_{ai} \leftarrow GetCorrectlyClassifiedInstances(w_{in}, S_o)$
  FinalSet $S_{fi} \leftarrow S_i \cup S_{ai}$
  FinalModel $w_{fi} \leftarrow BuildSVMModel(S_{fi})$
end procedure

We have evaluated our algorithm on the Pascal Voc 2007 dataset. In order to emulate noise, we flipped the labels of 20% of the hardest examples for each of the 20 classes in the dataset. We used distance from the hyperplane as a measure to approximate hardness. The ones closest to the hyperplane are hard in the feature space and are expected to be visually difficult to annotate. We have used linear SVM in all our experiments. To describe each image, we have used responses of the 7th layer of caffe [2] features pre-trained on the Imagenet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). This gave us a 4096-length feature vector to describe each image. Table shows that using our method we were able to achieve more than 12% improvement over both RANSAC SVM and ordinary SVM.

Our method can also be adapted to identify hard examples in the training data. This is because hard examples are also dissimilar in their feature space as compared to normal examples. Figure 1 shows some of the hard examples identified by our method.

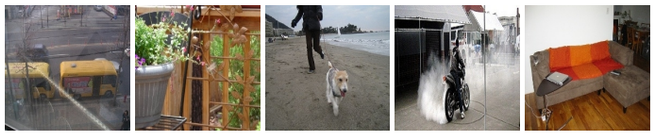


Figure 1: Few visually hard images of VOC 2007 which were detected as outliers by our method. The classes are : bicycle, bird, boat, car and cat

Voc 2007 dataset contains some images labelled as 0 which denote hard positives. We show in table 2 that even after including the 0 labels, we were able to achieve nearly the same performance as normal svm without the 0 labels. Also, our method performed better when 0 labels were not included in both normal svm and our method.

Normal SVM	RANSAC-SVM	Our method
41.2	48.6	62.7

Table 1: Comparison of mean average precision with 20% noisy labels on the Voc 2007 dataset

Normal SVM without 0 labels	Our method with 0 labels	Our method without 0 labels
72.7	72.0	**73.5**

Table 2: Mean average precision of Voc 2007 dataset using normal SVM excluding examples labelled as 0 , our method in the presence of the 0 labelled examples and our method after removing the 0 labelled examples.

Thus we show how a simple adaptation of RANSAC SVM can be used to achieve robustness to noise. We further show how it can be used to detect hard examples in the training data.

[1] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 6:381–395, 1981.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*.

[3] K. Nishida and T. Kurita. Ransac-svm for large-scale datasets. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*.

Beyond MSER: Maximally Stable Regions using Tree of Shapes

Petra Bosilj[1]
petra.bosilj@irisa.fr
Ewa Kijak[2]
ewa.kijak@irisa.fr
Sébastien Lefèvre[1]
sebastien.lefevre@univ-ubs.fr

[1] Université de Bretagne-Sud
IRISA
Vannes, France

[2] Université de Rennes 1
IRISA
Rennes, France

Detection of local features which are distinctive, invariant and discriminative is used to construct compact image representations in many computer vision applications. Achieving robustness against viewpoint change motivated the development of affine invariant detectors responding to image gradient or contrast changes, edges or corners. We focus on the Maximally Stable Extremal Regions (MSER) detector [3] which responds to blobs of high contrast to produce affine invariant, distinctive arbitrary shaped regions. Exploiting the tree-based MSER computation algorithm [6], we replace the Min and Max-trees [7] in the algorithm with the Tree of Shapes [5], thus changing the pixel ordering used for region extraction.

Min and Max-trees [7] represent the composition of complex images by encoding hierarchical relations of regions on various scales. The leaves of the Min-tree correspond to local image minima, while the inner nodes are (maximal) connected regions $\mathcal{R}_k$ at gray level k, such that all region pixel intensities $f(p)$ are lower than k. The root region $\mathcal{R}_{max}$ at the highest gray level covers the whole image. Distance between two nodes is their gray level difference: $d(\mathcal{R}_k,\mathcal{R}_l) = |l-k|$. The Max-tree is the *dual hierarchy*, corresponding to the Min-tree of an inverted image $-I$.

Extremal regions used in MSER computation [3] correspond to the Min and Max-tree nodes. *Minimal* extremal regions $\mathcal{R}_k$ are connected regions in which all the elements on the outer boundary have strictly greater intensity than all the adjacent region elements, and are contained in the Min-tree. Similarly, the Max-tree comprises the *maximal* extremal regions. MSER computation is based on finding the local minima of the stability function $q(\cdot)$ for the extremal regions along the nested sequences:

$$q(\mathcal{R}_k) = \frac{|\mathcal{R}_{k+\Delta} \setminus \mathcal{R}_k|}{|\mathcal{R}_k|}. \quad (1)$$

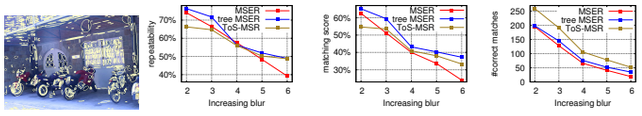
where $|\cdot|$ denotes cardinality. Larger Δ values require region stability through a greater range of gray levels. The region $\mathcal{R}_{k+\Delta}$ is determined from the sequence of nested regions to be the largest region such that $d(\mathcal{R}_k,\mathcal{R}_{k+\Delta}) \leq \Delta$, and found among the ancestral regions of the region $\mathcal{R}_k$ in the corresponding tree. The stability function $q(\cdot)$ can then be calculated concurrently with tree construction [6].

Substituting the Min and Max-tree with the Tree of Shapes (ToS) [5] became viable with the introductions of a near-linear construction algorithm [1]. The ToS models both dark and bright structure by encoding the image composition in terms of shapes and their contrast with their background. It has the *self-dual* property, being unchanged if constructed for the inverted image $-I$. In order to construct a *ToS based Maximally Stable Regions (ToS-MSR)* detector, we have to define the region distance to be used for ToS.

The leaves of ToS correspond to both image maxima and minima, and the regions of the hierarchy are obtained by filling the holes in the extremal regions in Min and Max-trees. A shape $\mathcal{R}$ is the direct parent of the shape $\mathcal{Q}$ if $\mathcal{R}$ is the smallest shape containing $\mathcal{Q}$. Any region $\mathcal{R}$ corresponding to an inner node is composed of the image elements of all of its children and some additional elements, which are always on the same gray level k. The node whose all additional elements are on the level k is referred to as $\mathcal{R}_k$. The distance between any two nodes of ToS in a vertical relation is then defined based on the pair-wise difference between the neighboring node levels. The distance between regions $\mathcal{R}_k \subseteq \mathcal{R}_l$ amounts to the sum of consecutive distances of all the nested regions on a path between those regions, and is equal to:

$$d(\mathcal{R}_k,\mathcal{R}_l) = |k-k_0| + |k_0-k_1| + \cdots + |k_x-l|. \quad (2)$$

The constructed ToS-MSR detector returns slightly more detected regions than MSER, which are still of arbitrary shape but better centralized after affine construction of measurement regions (cf. Fig. 1(a)). We evaluate our detector in the matching framework of Mikolajczyk et al. [4], as well as in a retrieval setup, and compare it to the MSER implementation provided for [4], as well as a tree-based MSER implementation. Typically, we achieve comparable repeatability and matching scores as the MSER detectors with a 20–40% more correct matches (shown in Figs. 1(b)–1(d)). This mitigates one of the main drawbacks of MSER, which occasionally returns too few regions even for applications where a small number of regions is an advantage, such as matching or retrieval. In our retrieval experiments using the VLAD indexing scheme [2], we evaluate the performance in terms of *mean average precision (MAP)* on *INRIA 'holidays'* and *'oxford5k'*. The vocabulary was built using the *'paris6k'* dataset. In addition to the small but consistent increase in the number of features, our ToS-MSER detector outperforms both MSER versions in terms of MAP (shown in Tab. 1). The improvement in the retrieval experiments, output in terms of arbitrary shapes organized in a single hierarchy as well as slightly increasing the number of MSER outputs prompts further investigation of component trees for region detection.



(a) (b) (c) (d)

Figure 1: Performance of ToS-MSR on the *'bikes'* dataset [4] (typical across other framework datasets). Detector responses for image 1 are shown in (a). Repeatability and matching scores are displayed in (b) and (c). The number of correct matches per image pair can be found in (d).

detector	*'holidays'* features	*'holidays'* MAP mean	*'holidays'* MAP high	*'oxford5k'* features	*'oxford5k'* MAP mean	*'oxford5k'* MAP high
MSER	914.78	0.434	0.451	874.02	0.227	**0.252**
tree MSER	1000.57	0.419	0.431	931.08	0.222	0.232
ToS-MSR	1295.85	**0.451**	**0.462**	1160.98	**0.239**	0.250

Table 1: Results of the image retrieval experiments, using *'paris6k'* for vocabulary construction, and *'holidays'* and *'oxford5k'* for validation. Mean and best MAP values for 8 randomly reinitialized vocabularies.

[1] T. Géraud, E. Carlinet, S. Crozet, and L. Najman. A Quasi-linear Algorithm to Compute the Tree of Shapes of nD Images. In *ISMM 2013*, pages 98–110. Springer, 2013.

[2] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR 2010*, pages 3304–3311. IEEE, 2010.

[3] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *BMVC 2002*, pages 384–396, 2002.

[4] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[5] P. Monasse and F. Guichard. Scale-space from a level lines tree. *J. Visual Commun. and Image Represent.*, 11(2):224–236, 2000.

[6] D. Nistér and H. Stewénius. Linear time maximally stable extremal regions. In *ECCV 2008*, pages 183–196. Springer, 2008.

[7] P. Salembier, A. Oliveras, and L. Garrido. Antiextensive connected operators for image and sequence processing. *IEEE T. Image Process.*, 7(4):555–570, 1998.

Exploring Locally Rigid Discriminative Patches for Learning Relative Attributes

Yashaswi Verma
http://researchweb.iiit.ac.in/~yashaswi.verma/
C. V. Jawahar
http://www.iiit.ac.in/~jawahar/

CVIT
IIIT-Hyderabad, India
http://cvit.iiit.ac.in

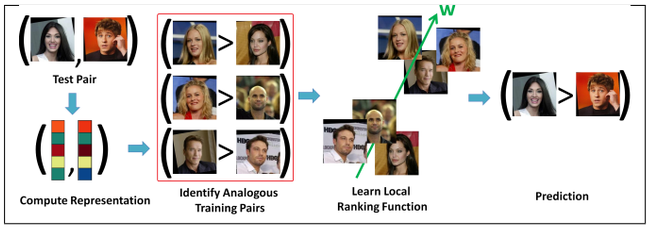


Figure 1: Approach overview: Given a test pair, first its patch-based representation is computed. Then using this representation, its analogous training pairs are identified. These pairs are used to learn a (local) ranking function, which is finally used for relative attribute prediction ("smiling" in this above illustration).

Relative attributes help in comparing two images based on their visual properties [4]. These are of great interest as they have been shown to be useful in several vision related problems such as recognition, retrieval, and understanding image collections in general. In the recent past, quite a few techniques (such as [3, 4, 5, 6]) have been proposed for the relative attribute learning task that give reasonable performance. However, these have focused either on the algorithmic aspect or the representational aspect. In this work, we revisit these approaches and integrate their broader ideas to develop simple baselines. These not only take care of the algorithmic aspects, but also take a step towards analyzing a simple yet domain independent patch-based representation [1] for this task.

Given an image, we compute HOG descriptors from non-overlapping square patches and concatenate them. This basic representation efficiently captures local shape in an image, as well as spatially rigid correspondences across regions in an image pair. The motivation behind using this for the relative attribute learning task is the observation that images in several domain-specific datasets (such as shoes and faces) are largely aligned, and spatial variations in the regions of interest are globally minimal (Figure 2). We integrate this representation with two state-of-the-art approaches: (i) "Global" [4] that learns a single, globally trained ranking model (Ranking SVM [2]) for each attribute, and (ii) "LocalPair" [6] that uses a ranking model trained locally using *analogous* training pairs for each test pair. Its another variant, "LocalPair+ML", uses a learned distance metric while computing the analogous pairs. The motivation behind the LocalPair approach is that as visual differences within an image-pair become more and more subtle, a single prediction model trained using the whole dataset may become inaccurate. This is because it captures only the coarse details, and smoothens the fine-grained properties. This approach proposes to consider only the few training pairs for each test pair that are most analogous to it. These can be thought of as the K training pairs that are most similar to the given test pair. In LocalPair+ML, a learned distance metric is used to give more importance to those feature dimensions that are more representative of a particular attribute while computing the analogous pairs. Using the identified pairs, both LocalPair and LocalPair+ML learn a local (specific to the given test pair) ranking model similar to [4]. Note that the "Global" approach can be thought of as a special case of the LocalPair approach where K is the total number of training pairs, and thus all of them are considered while learning a ranking model. This is illustrated in Figure 1.

We refer the above baselines as Global+Hog, LocalPair+Hog and LocalPair+ML+Hog. These baselines are extensively evaluated on three challenging relative attribute datasets: OSR (natural outdoor scenes), LFW-10 (faces) and UT-Zap50K (shoes). While comparing with previous works, we use the representations used by them (wherever applicable). Table 1 summarizes the quantitative results. We can observe that the baselines achieve promising results on the OSR and LFW-10 datasets, and perform better than the current state-of-the-art on the UT-Zap50K dataset (note that UT-Zap50K-2 dataset with fine-grained within-pair visual differences is the most challenging among these datasets). For detailed comparisons, please refer to the paper.

	LFW-10	OSR	UTZ-1	UTZ-2
RelativeParts [5]	78.5	–	–	–
Global [4]	63.4	88.0	88.1	61.7
LocalPair [6]	62.4	85.7	87.6	63.1
LocalPair+ML [6]	63.6	90.2	88.7	64.7
Global+Hog	72.9	86.8	90.2	65.8
LocalPair+Hog	72.3	87.9	88.8	67.0
LocalPair+ML+Hog	72.8	88.6	90.0	67.5

Table 1: Average accuracy comparison for all the three datasets. The best results (this work and those of the previous approaches) are underlined.

To analyse the performance gains achieved by Global+Hog on the LFW-10 dataset, we try to visualize what a global ranking model learns using the HOG descriptor. Since all the bins in the HOG descriptor have non-negative values, the aggregate weight of the ranking model in the interval corresponding to each cell can be thought of as a measure of confidence for identifying the relative importance of cells, as learned by the model. In Figure 2, we show these weights for two attributes. Surprisingly, the top two cells with maximum aggregate weights fall at almost the right place, thus demonstrating the possibility of attribute semantics being encoded in the ranking model.

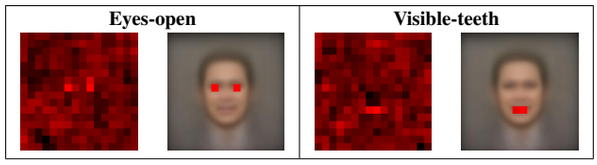


Figure 2: Learned HOG weights using Global+Hog baseline for two attributes from the LFW-10 datasets Left: Normalized distribution of weights. Right: Top few largest weights overlaid on the average image of this dataset (best viewed in color).

In this work, our goal was to develop intuitively simple baselines rather than to create a new method for learning relative attributes. The results suggest that along with the learning algorithm, choosing the right representation also plays a crucial role in the visual comparison task, and it is possible to achieve significant performance gains even by employing a simple but more appropriate representation. Domain knowledge can also prove to be useful in designing/selecting the representation and learning algorithm, as observed in the case of LFW-10 and UT-Zap50K datasets. As evident from the general performance level of the proposed baselines as well as existing methods, there is a lot of scope for improvement, especially on the challenging UT-Zap50K-2 dataset.

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[3] S. Li, S. Shan, and X. Chen. Relative forest for attribute prediction. In *ACCV*, 2012.

[4] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[5] Ramachandruni N. Sandeep, Yashaswi Verma, and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014.

[6] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014.

Semantic Description of Medical Image Findings: Structured Learning Approach

Pavel Kisilev[1]
pavelki@il.ibm.com
Eugene Walach[1]
walach@il.ibm.com
Sharbell Hashoul[1,2]
sharbelh@il.ibm.com
Ella Barkan[1]
ella@il.ibm.com
Boaz Ophir[1]
boazo@il.ibm.com
Sharon Alpert[1]
asharon@il.ibm.com

[1] IBM Haifa Research Lab, Haifa, Israel
[2] Carmel Medical Center, Haifa, Israel

Computer Aided Diagnosis (CADx) systems are designed to assist doctors in medical image interpretation. However, a CADx is often thought of as a "black box" whose diagnostic decision is not intelligible to a radiologist. Therefore, a system that uses semantic image interpretation, and mimics human image analysis, has clear benefits. In this paper, we propose a system which automatically generates textual description of medical image findings, such as lesions.

Having found a lesion, a radiologist examines its visual appearance characteristics to make a final diagnosis. The visual appearance is usually described in terms of semantic descriptors such as: shape, orientation, margin, boundary type, contrast enhancement, localization, mass effect on surrounding tissues, and others.

The estimation of semantic descriptor values requires explicit or implicit representation by a diverse set of image measurements that describe each one of the semantic descriptors quantitatively. We use various image measurements to calculate the informative features, such as histograms of pixel values, shape and texture descriptors and others.

We pose the problem of semantic description of a lesion as learning to map a set of image based informative features to a set of semantic descriptor values. A lesion is described by a set of J semantic descriptors. Semantic description of the i-th lesion is an assignment $y_i = [y_{i,1}, \ldots, y_{i,J}]$ where each j-th semantic descriptor $y_{.,j}$ can have one of the possible discrete values $Y_j \in \{1, \ldots, V_j\}$ corresponding to the radiological lexicon. Following the standard practice in structured learning, the energy function of the above assignment is a sum of unary and pair-wise terms [1]:

$$E(\mathbf{y}_i) = \sum_j \mathbf{u}_1^T \phi_1(y_{ij}, \mathbf{X}_i) + \sum_{j,k \in \mathcal{S}} \mathbf{u}_2^T \phi_2(y_{ij}, y_{ik}, \mathbf{X}_i), \quad (1)$$

where $\mathbf{X}_i$ are image measurements (visual features); ϕ_1 are the unary potentials that capture the relationship between the image measurements and the semantic descriptor values; ϕ_2 are the pairwise potentials that capture joint relationships between the semantic descriptors, and reflect the likelihood of semantic descriptors to jointly have particular values; $\mathcal{S}$ is the set of all possible pairs of semantic descriptors; $\mathbf{u_1}, \mathbf{u_2}$ are the model parameters.

We learn the parameters $\mathbf{u} = [\mathbf{u}_1^T \ \mathbf{u}_2^T]^T$ of the model (1), using Structured SVM (SSVM) framework. In particular, given N training examples, the model parameters are learned by optimizing the regularized large-margin objective [1], [2]:

$$\hat{\mathbf{u}} = \min_{\mathbf{u}, \xi \geq 0} \frac{1}{2} \|\mathbf{u}\|^2 + C\xi$$
$$\text{s.t.} \quad \frac{1}{N} \sum_{i=1}^{N} \max_{\bar{\mathbf{y}}_i \in \mathcal{Y}} \left[\Delta(\bar{\mathbf{y}}_i, \mathbf{y}_i^*) - \langle \mathbf{u}, \psi(I_i, \bar{\mathbf{y}}_i) \rangle + \langle \mathbf{u}, \psi(I_i, \mathbf{y}_i) \rangle \right] \leq \xi \quad (2)$$

where the unary and the pairwise potentials are concatenated into a column vector ψ. Once the model parameters are learned, the inference goal is, given a new lesion, to find the best assignment whose semantic values result in the lowest energy (1). This is achieved by:

$$\hat{\mathbf{y}}_i = \arg\min_{\bar{\mathbf{y}}_i \in \mathcal{Y}} < \hat{\mathbf{u}}, \psi(\bar{\mathbf{y}}_i, \mathbf{X}_i) >, \quad (3)$$

whose solution is found using the Sequence Alignment algorithm [2].

The main mode of operation of the proposed method is illustrated in Figure 1. Two examples of input images, of the detected lesion contours, and of the corresponding automatically generated textual descriptions of lesions using the proposed method are depicted in Figure 1a-1c, respectively. Given a medical image, the first step is to localize a lesion and to find its contour. In this work, we concentrate on the problem of automatic generation of semantic description of lesions. We, therefore, use semi-automatic lesion detection and contour extraction, instead of a fully automatic approach. We assume that the bounding box around a lesion is found or given by a radiologist. Then, an active contour type method is applied to find the lesion contour inside of the bounding box. Given the found lesion contour, we calculate various image measurements and, based on it, construct visual features. Finally, we use the learned in advance model of mapping from visual features to semantic values, and generate the semantic description of a new lesion.

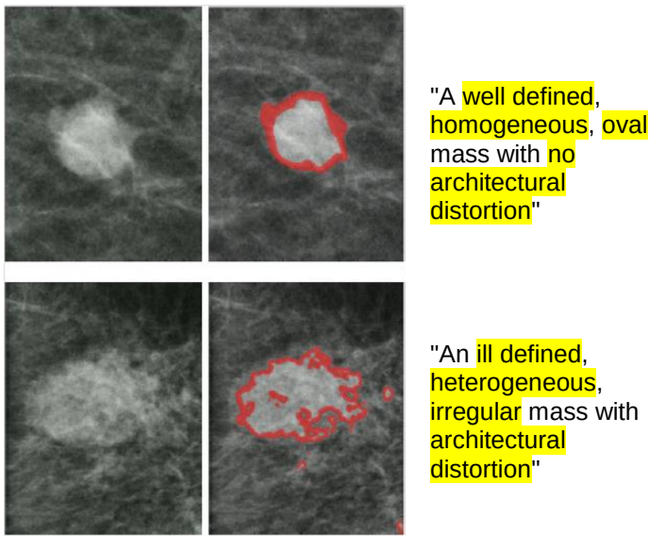


Figure 1: Examples of automatically generated textual description of lesions, using the proposed method: (a) region of interest (ROI) taken from a breast mammography image, (b) found lesion contour (c) textual description of the lesion (estimated semantic descriptors are marked in yellow colour).

The proposed approach generates radiological lexicon descriptors used to make a diagnosis of various diseases. This can help radiologists easily understand a diagnosis recommendation made by an automatic system, such as CADx. We apply the proposed method to publicly available and to proprietary breast and brain imaging datasets, and show that our method generates more accurate descriptions, as compared to other alternative approaches.

[1] Nowozin, Sebastian, and Christoph H. Lampert. "Structured learningand prediction in computer vision." Foundations and Trends in Computer Graphics and Vision 6.3–4 (2011): 185-365.

[2] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. Journal of Machine Learning Research (JMLR), 6(Sep):1453-1484, 2005.

Latent Structure Preserving Hashing

Ziyun Cai[1]
cziyun1@sheffield.ac.uk
Li Liu[2]
li2.liu@northumbria.ac.uk
Mengyang Yu[2]
m.y.yu@ieee.org
Ling Shao[2]
ling.shao@ieee.org

[1] Department of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD, UK

[2] Computer Vision and Artificial Intelligence Group
Department of Computer Science and Digital Technologies
Northumbria University
Newcastle upon Tyne, NE1 8ST, UK

Aiming at efficient similarity search, hash functions are designed to embed high-dimensional feature descriptors to low-dimensional binary codes such that similar descriptors will lead to the binary codes with a short distance in the Hamming space. It is critical to effectively maintain the intrinsic structure and preserve the original information of data in a hashing algorithm. In this paper, we propose a novel hashing algorithm called Latent Structure Preserving Hashing (LSPH), with the target of finding a well-structured low-dimensional data representation from the original high-dimensional data through a novel objective function based on Non-negative Matrix Factorization (NMF) [2]. Via exploiting the probabilistic distribution of data, LSPH can automatically learn the latent information and successfully preserve the structure of high-dimensional data. After finding the low-dimensional representations, the hash functions can be acquired through multi-variable logistic regression. Experimental results on two large-scale datasets, i.e., **SIFT 1M** and **GIST 1M**, show that LSPH can significantly outperform the state-of-the-art hashing techniques. The outline of the proposed approach is depicted in Fig. 1.

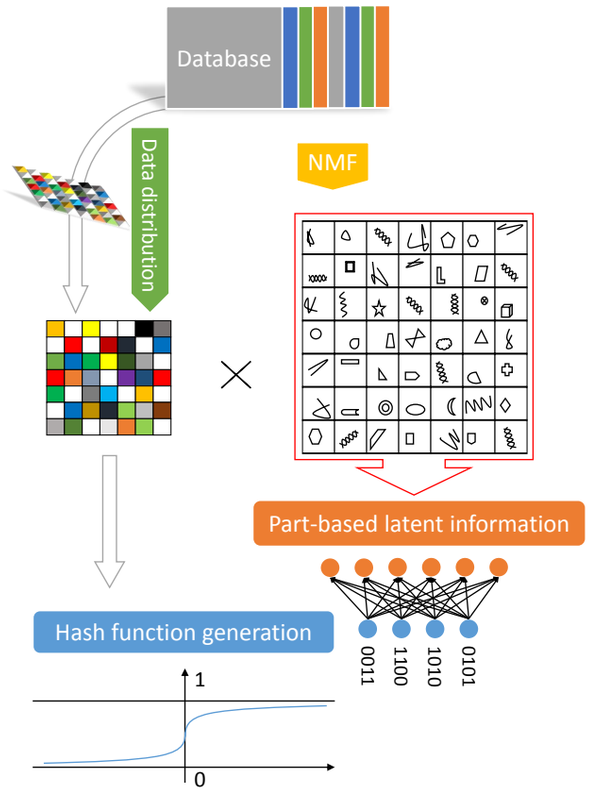


Figure 1: The outline of the proposed method. Part-based latent information is learned from NMF with the regularization of data distribution.

Theoretically, it is expected that the low-dimensional data V given by NMF can obtain locality structure from the high-dimensional data X. We propose to minimize the Kullback-Leibler divergence between the joint probability distribution in the high-dimensional space and the joint probability distribution in the low-dimensional space. Then through combining the data structure preserving part and the NMF technique, we can obtain the following new objective function:

$$O_f = \|X - UV\|^2 + \lambda KL(P\|Q), \tag{1}$$

where P is the joint probability distribution in the high-dimensional space, Q is the joint probability distribution in the low-dimensional space, $V \in \{0,1\}^{D\times N}$, $X, U, V \geqslant 0$, $U \in \mathbb{R}^{M\times D}$, $X \in \mathbb{R}^{M\times N}$, and λ controls the smoothness of the new representation.

Motivated by [3], we first relax the data $V \in \{0,1\}^{D\times N}$ to the range $V \in \mathbb{R}^{D\times N}$ for obtaining real values. The Lagrangian of our problem will be:

$$\mathcal{L} = \|X - UV\|^2 + \lambda KL(P\|Q) + tr(\Phi U^T) + tr(\Psi V^T), \tag{2}$$

where matrices Φ and Ψ are two Lagrangian multiplier matrices. By some algebraic deviations, we have the following update rules for any i, j:

$$V_{ij} \leftarrow \frac{(U^T X)_{ij} + 2\lambda \sum_{k=1}^{N} \frac{p_{jk}V_{ik} + q_{jk}V_{ij}}{1+\|\mathbf{v}_j - \mathbf{v}_k\|^2}}{(U^T UV)_{ij} + 2\lambda \sum_{k=1}^{N} \frac{p_{jk}V_{ij} + q_{jk}V_{ik}}{1+\|\mathbf{v}_j - \mathbf{v}_k\|^2}} V_{ij}, \quad U_{ij} \leftarrow \frac{(XV^T)_{ij}}{(UVV^T)_{ij}} U_{ij}. \tag{3}$$

The proof of convergence about U and V is similar to the ones in [1, 4].

Then we need to convert the low-dimensional real-valued representations from $V = [\mathbf{v}_1, \cdots, \mathbf{v}_N] \in \mathbb{R}^{D\times N}$ into binary codes via thresholding: if the d-th element in $\mathbf{v}_n$ is larger than the specified threshold, this real-value will be represented as 1; otherwise it will be 0, where $d = 1, \cdots, D$ and $n = 1, \cdots, N$. Therefore, with the vector-valued sigmoid function

$$h_\Theta(\mathbf{v}_n) = \left(\frac{1}{1+e^{-(\Theta^T \mathbf{v}_n)_i}}\right)^T_{i=1,\cdots,D}$$

for the matrix $\Theta \in \mathbb{R}^{D\times D}$, our cost function for the corresponding regression matrix Θ can be defined as:

$$J(\Theta) = -\frac{1}{N}\left(\sum_{n=1}^{N}\left(\hat{\mathbf{v}}_n^T \mathbf{log}(h_\Theta(\mathbf{v}_n)) + (\mathbf{1} - \hat{\mathbf{v}}_n)^T \mathbf{log}(\mathbf{1} - h_\Theta(\mathbf{v}_n))\right) + \delta\|\Theta\|^2\right)$$

where $\mathbf{log}(\cdot)$ is the element-wise logarithm function and $\mathbf{1}$ is an $D \times 1$ all ones matrix. We use $\delta\|\Theta\|^2$ as the regularization term in logistic regression to avoid overfitting.

The updating equation is shown as follows:

$$\Theta^{(t+1)} = \Theta^{(t)} - \frac{\alpha}{N}\sum_{n=1}^{N}(h_{\Theta^{(t)}}(\mathbf{v}_n) - \hat{\mathbf{v}}_n)\mathbf{v}_n^T - \frac{\alpha\delta}{N}\Theta^{(t)}. \tag{4}$$

where α is the learning rate. Since h_Θ is a sigmoid function, the hash code for the new coming sample $X_{new} \in \mathbb{R}^{M\times 1}$ can be represented as:

$$\hat{V}_{new} = \lfloor h_\Theta(QX_{new}) \rceil, \tag{5}$$

where $\lfloor\cdot\rceil$ means the nearest integer function for each entry of h_Θ and $Q = (U^T U)^{-1}U^T$ which is the pseudoinverse of U.

[1] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI*, 33(8):1548–1560, 2011.

[2] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[3] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *NIPS*, pages 1753–1760, 2009.

[4] Wenbin Zheng, Yuntao Qian, and Hong Tang. Dimensionality reduction with category information fusion and non-negative matrix factorization for text categorization. In *Artificial Intelligence and Computational Intelligence*, pages 505–512. 2011.

Local Feature Binary Coding for Approximate Nearest Neighbor Search

Li Liu
li2.liu@northumbria.ac.uk
Mengyang Yu
m.y.yu@ieee.org
Ling Shao
ling.shao@ieee.org

Computer Vision and Artificial Intelligence Group,
Department of Computer Science and Digital Technologies,
Northumbria University, UK

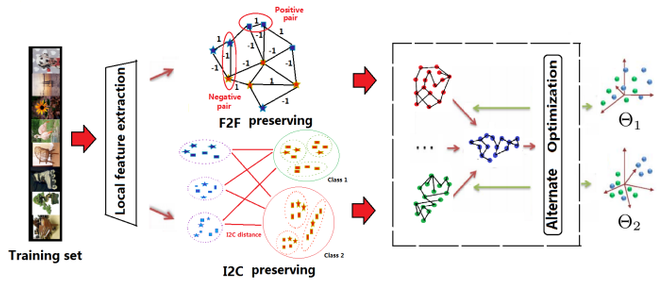


Figure 1: The illustration of the working flow of LFBC learning. The algorithm intends to preserve the pairwise F2F structure and the I2C distances and outputs the optimal bilinear projection matrices Θ_1 and Θ_2.

The potential value of hashing techniques has led to it becoming one of the most active research areas in computer vision and multimedia. However, most existing hashing methods for image search and retrieval are based on global representations, e.g., GIST [3], which lack the analysis of the intrinsic geometric property of local features and heavily limit the effectiveness of the hash code. In this paper, we propose an supervised local feature hashing framework, i.e., Local Feature Binary Coding (LFBC), for visual similarity search, in which the feature-to-feature (F2F) and image-to-class (I2C) structures are successfully preserved and combined together. Specifically, the F2F structure considers the pairwise relationship between local features in the original feature space. While, from a higher-level aspect, I2C structure reflects the connection between images and their corresponding classes, which is derived from [1]. The outline of the proposed method is illustrated in Fig. 1. It is worthwhile to highlight several properties of the proposed method: (1) Different with global representation based hashing, LFBC directly learns hashing function from local features and simultaneously preserves pairwise F2F and I2C structure, which is proved to be more effective for accurate retrieval. (2) Inspired by [2, 4], bilinear projection based hashing function is adopted in our method. Thus, the complexity of the eigen-decomposition, which is the cubic form of the dimensionality, will be significantly reduced. The corresponding integrated LFBC algorithm is depicted in Algorithm 1.

Algorithm 1 Local Feature Binary Coding (LFBC)

Input: Local feature set of each training image $\mathcal{X}_i = \{X_{i1}, \cdots, X_{im_i}\}$ in matrix form, $i = 1, \cdots, n$, the whole local feature set $\mathcal{F} = \bigcup \mathcal{X}_i$, the parameter k for pairwise structure preserving, the number of centroids K in K-means and the label information function $C(\cdot) : \mathcal{F} \to \{1, \cdots, C\}$.

Output: The bilinear projection matrices Θ_1 and Θ_2.

1: Construct local feature pairing set $\mathcal{P} = \{(i, j) | X_i, X_j \in \mathcal{F}\}$ and their corresponding pairwise labels $\ell_{ij} = \{-1, +1\}$, where $\ell_{ij} = +1$ if $X_i \in \mathrm{NN}_k(X_j)$ or $X_j \in \mathrm{NN}_k(X_i)$, and $\ell_{ij} = -1$ otherwise;
2: Employ the K-means clustering algorithm on the set of local features of each class $\bigcup_{C(\mathcal{X}_i)=c} \mathcal{X}_i$, $c = 1, \cdots, C$;
3: Compute pairwise weight W_{ij}^F and I2C similarity W_{ic}^{I2C};
4: Initialize $\Theta_2^{(0)}$ randomly;
5: **repeat**
6: Optimize Θ_1 and Θ_2 alternately;
7: **until** the objective function $\mathcal{L}(\Theta_1, \Theta_2)$ converges.

In the retrieval phase, considering that our method is specifically designed for local features, the original Hamming Ranking and Hamming Table cannot be directly applied to local features for visual indexing. Thus, in this paper, we also introduce an indexing/searching scheme called Local Hashing Voting (LHV) as shown in Fig. 2, which has been demonstrated to be efficient and accurate for image similarity search in our experiments.

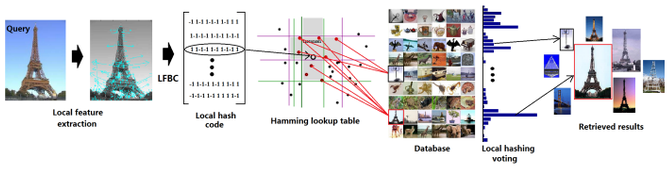
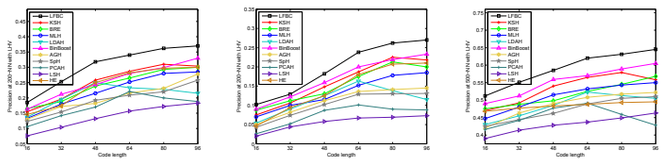
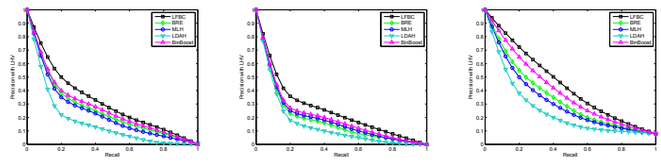


Figure 2: The illustration for the proposed LHV scheme.



(a) Caltech-256 (b) SUN397 (c) NUS-WIDE

Figure 3: Performance comparison with different numbers of bits.



(a) Caltech-256 (b) SUN397 (c) NUS-WIDE

Figure 4: The comparison of precision-recall curves of the supervised algorithms on the three datasets with the code length of 96 bits.

Table 1: Result comparison (32 bits) with/without F2F and I2C term.

Methods	Caltech-256	SUN397	NUS-WIDE
Only F2F preserving	0.189	0.065	0.387
Only I2C preserving	0.227	0.104	0.432
F2F+I2C preserving (LFBC)	**0.253**	**0.129**	**0.551**

For instance, given a bucket with hash code [1,1,-1,-1,1,-1,1,-1,-1,1] of a local feature, we store the indices of the images, which contain the same local feature hash code with this bucket. In this way, we search the hash code $H(\mathbf{q}_i)$ for each local feature $\mathbf{q}_k \in \mathcal{Q}$ in the query image $\mathcal{Q} = \{\mathbf{q}_1, \cdots, \mathbf{q}_m\}$ over the Hamming lookup table within Hamming radius r and return the possible images' indices. Finally, we vote and accumulate the times of each image's indices appearing in relevant buckets and then rank them in decreasing order. The final retrieved samples are returned according to the relevant ranking generated by LHV.

The experimental results on the Caltech-256, SUN397 and NUS-WIDE datasets are demonstrated in Fig. 3, Fig. 4 and Table 1.

[1] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[2] Yunchao Gong, Sanjiv Kumar, Henry A. Rowley, and Svetlana Lazebnik. Learning binary codes for high-dimensional data using bilinear projections. In *CVPR*, 2013.

[3] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[4] Jieping Ye, Ravi Janardan, Qi Li, et al. Two-dimensional linear discriminant analysis. In *NIPS*, 2004.

Revealing Smooth Structure of Visual Data by Permutation on Manifolds

Yi-Lei Chen
https://sites.google.com/site/fallcolor/
Chiou-Ting Hsu
http://www.cs.nthu.edu.tw/~cthsu/candy.html

Department of Computer Science,
National Tsing Hua University, Taiwan

Motivation

The ever-increasing scale of ubiquitous visual data inevitably raises new challenges in data organization. Although the sources of data collection are unpredictable, a collection of visual data usually shares similar topics and most of them are created for some special events. Some pioneer work, called photo sequencing, has been proposed to discover a set of still images for estimating the temporal order of a dynamic scene [1, 2] or the spatial order of an instantaneous event [3]. However, these methods rely on different hypotheses when capturing images, such as similar view points or ring-like camera arrangement, and therefore limit their applicability to general datasets.

In this paper, we address the issue of visual data organization by recovering an intrinsic order from an unorganized dataset. The proposed method exploits the inherent nature of manifold. This new perspective, posing no hypothesis on local topology of observed data, is simply built on the smoothness prior of manifold geometry. Under the observation that strong relation exists among visual content, we assume a visual dataset lies on a manifold and thus changes smoothly from point to point. By exploiting the linearity within nearby data points, our goal becomes to visit all of the data points along a manifold-guided order and to characterize the specific manifold's shape.

Methodology

The main idea of our method is to *construct a template of smooth manifold* and then *discover the optimal permutation between data points to fit this smooth shape*. As shown in Fig. 1, we use a pseudo graph to build the template and define the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$ by

$$w_{ij} = \exp\left(-\frac{(i-j)^2}{\eta_g{}^2}\right) \text{ and } d_{ij} = \begin{cases} \sum_k w_{ik}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}. \quad (1)$$

The edge weight w_{ij} is simply measured according to the adjacency between indices to ensure the well-ordered smoothness, and η_g roughly controls the size of effective neighbourhood. Given N high-dimensional data points $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N]$, we attempt to uncover the intrinsic data order $i \mapsto \phi(i)$ under a guidance of smooth manifold geometry:

$$\begin{aligned} \min\ \mathrm{f}(\mathbf{P}) &= \min \textstyle\sum_{1 \le i,j \le N} \left\| \mathbf{x}_{\phi(i)} - \mathbf{x}_{\phi(j)} \right\|_2^2 w_{\phi(i)\phi(j)} \\ &= \min \mathrm{tr}(\mathbf{XPLP}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}), \end{aligned} \quad (2)$$

where $\mathbf{L}$ is the Laplacian matrix derived from a pseudo graph, and $\mathbf{P} \in \Pi$ is a permutation matrix satisfying

$$\Pi = \{\mathbf{P} \mid p_{ij} = \{0,1\},\ \textstyle\sum_{j=1}^{N} p_{ij} = 1, \forall\, i,\ \sum_{i=1}^{N} p_{ij} = 1, \forall\, j\}. \quad (3)$$

Equation (2) is an NP-hard combinatorial optimization problem and remains very challenging without off-the-shelf solvers. We thus propose a novel algorithm named *Permutation on Manifolds* (PoM) to estimate $\mathbf{P}$ in polynomial time. Specifically, we use variable splitting technique to derive

$$\min \mathrm{tr}\left(\mathbf{X_p L X_p}^{\mathrm{T}}\right) \quad \text{s.t. } \mathbf{X_p} = \mathbf{XP}, \quad (4)$$

Then we follow augmented Lagrange multiplier (ALM) method to solve the constrained optimization problem. The core of our algorithm boils down to one of the sub-problems with the following form:

$$\mathbf{P}^{(t+1)} = \arg\min_{\mathbf{P}} \|\mathbf{XP} - \mathbf{Z}\|_{\mathrm{F}}^2 \quad \text{s.t. } \mathbf{P} \in \Pi, \quad (5)$$

After we rephrase Equation (5) by a sequence of additive costs, our method can approach local optima efficiently with three guidelines:

- Relax binary integer constraints into nonnegative constraints;
- Allow each data point $\mathbf{x}_i$ a limited number of moves to its k nearest data points among $\mathbf{z}_1, \cdots, \mathbf{z}_N$; and
- Determine the best k, as small as possible, by binary search so that every $\mathbf{z}_i$ can be accessed by $\mathbf{x}_1, \cdots, \mathbf{x}_N$ at least once.

Note that we further study the potential of extending PoM algorithm to different applications rather than visual data organization. For example, we can seamlessly embed PoM into linear dimension reduction methods when the data affinity matrix is unavailable. In addition, including PoM into unsupervised manifold priors as a regularization term is also studied in our paper.

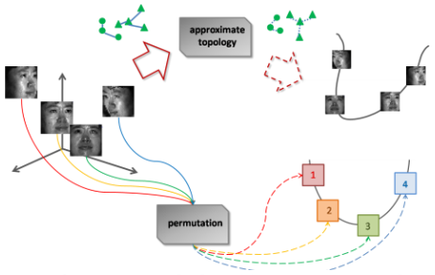


Fig. 1. Illustration of our proposed idea. (a) Data points in the original high-dimensional space; (b) the low-dimensional manifold approximated by classic methods; and (c) the smooth manifold approximated by permutation.

Results

We evaluate the proposed PoM algorithm with three experiments:

- We sample 300 data points from the Swiss roll dataset and justify our hypothesis by linking the data points along our estimated order (see Fig. 2).
- We embed PoM into linear dimension reduction and one of existing tensor completion method [4] to show the ability of PoM for regularization as manifold priors.
- We use two image sets "Basketball" and "Slide" released from [1] to verify that PoM is able to approximate a perceptually reasonable order from a given image set, which somewhat reveals the spatial order of cameras.

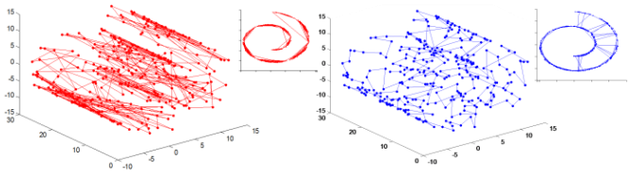


Fig. 2. The data-linking results of the 300 data points sampled from a Swiss roll obtained by ground truth labels (left) and PoM (right).

Conclusion

We summarize our contributions from three aspects: 1) to the best of our knowledge, this is the first work that leverages smooth manifold geometry to organize visual data; 2) the proposed pseudo graph generalizes the use of spectral graph theory and bypasses the measurement of data affinity in traditional methods; and 3) from our experiments, the proposed PoM algorithm shows its potential to serve as a core technique of numerous applications.

[1] T. Basha, Y. Moses, and S. Avidan. Photo Sequencing. In *Proc. ECCV*, 2012.

[2] T. Basha, Y. Moses, and S. Avidan. Space-Time Tradeoffs in Photo Sequencing. In *Proc. ICCV*, 2013.

[3] H. Averbuch-Elor and D. Cohen-Or. RingIt: Ring-ordering Casual Photos of a Temporal Event. In*Proc. SIGGRAPH*, 2015.

[4] Y. L. Chen, C. T. Hsu, and H. Y. Mark Laio. Simultaneous Tensor Decomposition and Completion Using Factor Priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(3): 577-591, 2014.

Normalized Autobinomial Markov Channels For Pedestrian Detection

Cosmin Ţoca
cosmin.toca@gmail.com
Mihai Ciuc
mihai.ciuc@upb.ro
Carmen Pătraşcu
cpatrascu@ceospacetech.pub.ro

University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Applied Electronics and Information Engineering Department, RO

Pedestrian detection represents one of the most important components of engineering devices that use automated vision to help decision systems take quick and accurate actions. Such systems are defined and customized to be useful for different needs, such as monitoring and aided surveillance, or increasing safety features in automotive industry. Given the large spectrum of applications that use pedestrian detection, demand has increased in recent years for the development of feasible solutions which can be integrated in devices such as smartphones or action cameras.

This paper focuses on finding probabilistic features that highlight the human body characteristics regardless of contextual information in images. Adjacent pixels are often spatially correlated, which means that they are likely to have similar values. We view the image as a collection of random variables indexed by certain locations, called sites. The state of a site ξ is conditionally independent of all variables in the random field, except the neighbouring system $\mathcal{N}_\xi = \{\eta \in \Omega \mid \eta \neq \xi, \mathrm{d}^2(\xi,\eta) \leq \Delta\}$, where Δ is a positive integer and $\mathrm{d}^2(\xi,\eta)$ is the squared Euclidean distance between ξ and η. The neighbouring system strictly depends on a collection of cliques $\mathcal{C} = \sum_{k=1}^{\omega(\delta)} \mathcal{C}_k$, where $\omega(\Delta)$ is the number of cliques for each local specification.

Energy function: An unpublished manuscript [2] describes how to interpret the local property of a Markov random field in terms of energy and potential, claiming that the probability at a site ξ is given by:

$$\pi^\xi(\gamma) = \frac{\mathrm{e}^{-\sum_{c \ni \xi} V_C(\gamma)}}{\sum_{\varphi \in \mathcal{F}} \mathrm{e}^{-\sum_{c \ni \xi} V_C(\varphi, \gamma_{\Omega \setminus \xi})}}, \tag{1}$$

where $V_\mathcal{C}$ is the potential function, and $\varphi \in \mathcal{F}$. To get the probability that at a site ξ the state is γ, we need to define a potential function $V_\mathcal{C}(\gamma)$ in the neighbouring system, here denoted by a collection of cliques $\mathcal{C}$. To be able to do this, we refer to the auto-binomial model that was introduced by Besag [1] to describe types of spatial processes, examining some stochastic models that occur in the texture of various physical materials.

Potential function: The potential at a certain state is given by:

$$V_\mathcal{C}(\gamma) = \begin{cases} -\ln \binom{\Gamma}{\gamma_\xi} + \gamma_\xi & \text{if } \mathcal{C} = \{\xi\} \\ \dfrac{\gamma_\xi \cdot \gamma_\eta}{\nu} & \text{if } \mathcal{C} = \{\xi, \eta\} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where ν is a normalization constant. If we replace Eq. 2 in Eq. 1, we get the probability assigned to a local system, and namely Eq. 3.

Feature calculation: The normalized autobinomial Markov channels are given by the probability that at a site ξ the state is γ:

$$\pi^\xi(\gamma) = Z^{-1} \binom{\Gamma}{\gamma_\xi} \sigma^{\gamma_\xi} (1-\sigma)^{\Gamma - \gamma_\xi}, \tag{3}$$

where $Z = \binom{\Gamma}{\Gamma/2}$ is a normalization constant for the binomial distribution, and $\sigma = \sigma(\mathcal{N}_\xi) = (\mathrm{e}^{\langle \alpha, \beta \rangle})/(1+\mathrm{e}^{\langle \alpha, \beta \rangle})$. The scalar product $\langle \alpha, \beta \rangle$ of the two vectors of size $\omega(\delta)$ sums up and weighs (with α) the absolute difference of gray levels $\beta_k = \beta_k(\gamma_{\mathcal{N}_\xi}) = |\gamma_\eta - \gamma_{\eta'}|$. The only sites participating are pairs of cliques found in the same neighbourhood of the analyzed pixel. In other words $\eta \neq \eta'$, and $\eta, \eta' \in \mathcal{N}_\xi$, where $\{\eta, \xi\}$ and $\{\eta', \xi\}$ represent two pairs in $\mathcal{C}_k$ containing ξ.

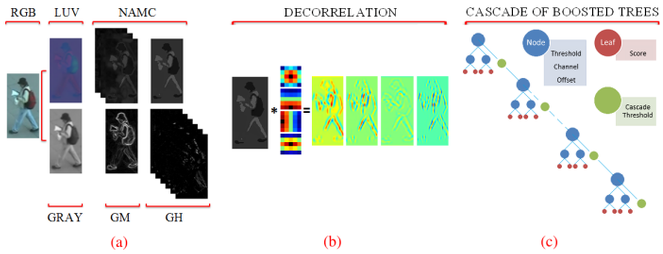


Figure 1: **a) Feature representation.** Several channels (NAMC-normalized autobinomial Markov channels, GM-gradient magnitude and GHs-gradient histograms) have been computed using different transformations of the input image. **b) Local decorrelation.** Multiple decorrelation filters have been learned separately for each channel. **c) Cascade of boosted trees.** The classification is based on a cascade of boosted trees, where each node examines a feature value in a certain channel and each leaf returns a score.

The main advantage of using the normalized autobinomial Markov channels as feature descriptor comes from the property of randomly selecting pixels to be part of the neighbourhood system, having a significant contribution for pedestrian detection in noisy scenarios. Another benefit is given by the fact that it shows several possibilities of optimizations by turning many of the computations into memory accesses.

This paper introduces the normalized autobinomial Markov channels and proves that the general idea of feature scaling is applicable. Moreover, it uses a method of feature decorrelation [3] to substitute the need for oblique splits, and shows that such a cascade of boosted trees [4] outperforms the majority of the existing features and methods for pedestrian detection. The results shown in Fig. 2.a and Fig. 2.b demonstrate the efficiency of our approach against the tested state-of-the art solutions.

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 1974.

[2] J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. *Unpublished*, 1971.

[3] W. Nam, P. Dollar, and J.H. Han. Local decorrelation for improved pedestrian detection. *NIPS*, 2014.

[4] Z. Tu. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. *ICCV*, 2005.

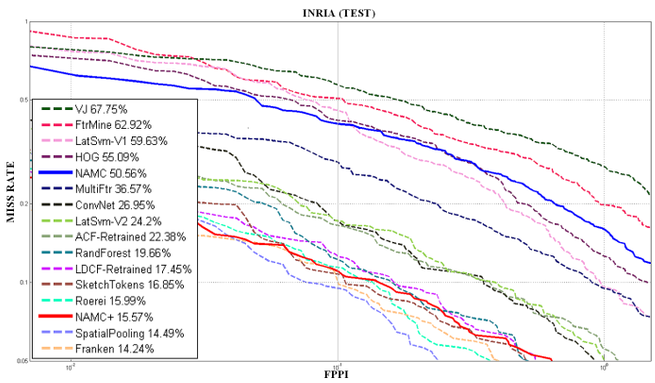


Figure 2.a: **Results on INRIA Pedestrian Dataset.**

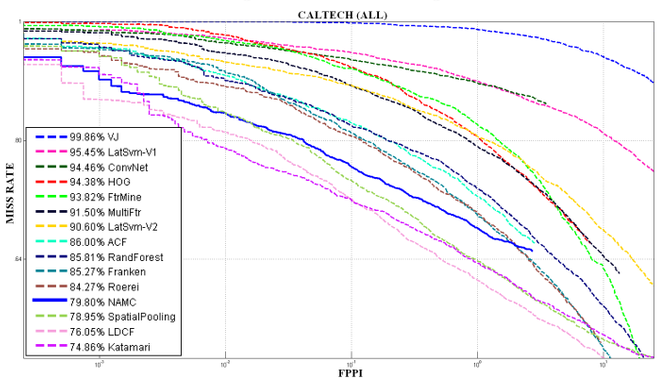


Figure 2.b: **Results on CALTECH Pedestrian Dataset.**

Exploring Prior Knowledge for Pedestrian Detection

Yi Yang[1]
yangyi@nlpr.ia.ac.cn
Zhenhua Wang[2]
wzh@ntu.edu.sg
Fuchao Wu[1]
fcwu@nlpr.ia.ac.cn

[1] National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China

[2] Rapid-Rich Object Search Lab
School of EEE, Nanyang Technological University
Singapore

Pedestrian detection is a classical and hot issue in object detection. Many approaches have been proposed in this area. However, it remains a challenging problem due to the variances in lighting conditions, scene structures, clothes, view angles, postures, scales, occlusions, *etc*.

Previous survey [1] has summarized that using better features plays an important role in improving detection quality. In addition, prior knowledge has shown good success in designing haar-like features for pedestrian detection [4]. Inspired by it, our work aims to integrate more prior knowledge into the design of features to enhance performance of pedestrian detection. By observing the pedestrian samples, we have discovered several important priors that are always ignored by previous methods, *e.g*. the symmetry of human body and the differences among different channels. We therefore utilize these priors to design two kinds of features: 1) symmetric features which capture the difference between two local symmetric regions, and 2) cross-channel features which capture the difference between two different channels of the same region. Figure 1 gives some visual examples of these two features. To the best of our knowledge, we are the first to use symmetric and cross-channel priors in designing features for pedestrian detection. By integrating these prior information into feature design, our detector achieves state-of-the-art performance.



Figure 1: (a) Symmetric features. Rectangles with the same color in the same channel represent one symmetric feature. (b) Cross-channel features. Rectangles with the same color in different channels represent one cross-channel feature.

Symmetric Features: A pedestrian body shape model can be obtained by computing an average edge map based on gradient magnitudes extracted from a large number of training samples [4]. The shape model seems like a fairly standard silhouette of a person standing facing the front, with hands falling naturally on body sides and feet keeping naturally together. It is worth to mention that, this fairly standing body silhouette is symmetrical about the vertical line at the center of the model, as the red dashed lines shown in Figure 1. This symmetry should be a good prior for designing effective features.

Based on this observation, we design a kind of symmetric features to capture the symmetric prior. For convenient implementation and efficient computation, we constrain our feature templates to be rectangles. More specifically, we define the symmetric features to be second-order rectangle features which are composed of two separate local symmetric rectangular regions. The two local symmetric regions share the same size and should be in the same channel. See Figure 1(a) for some examples. The feature values can be effectively computed as the difference between the responses of these two symmetric regions by using integral image.

To enrich the symmetric information, we actually take consideration of 3 more symmetry axes with 0°, 45°, and 135°symmetry angles (see Figure 2) in our implementation. In addition, as the whole human body is not symmetrical about these three additional axes, we actually divide the model region into 4 subregions and generate symmetric rectangles in these subregions to capture the local symmetric information.

Cross-channel Features: In recent pedestrian detectors with multiple channels, *e.g*. HOG+LUV channels, different channels contain different kinds of information. Previous methods mainly use features that

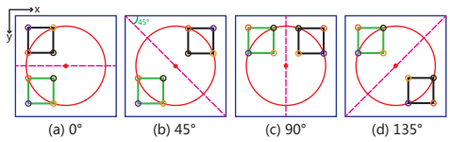


Figure 2: Some simple examples of symmetric rectangles.

capture the information in a single channel, *e.g*., `ChnFtrs` only takes responses of rectangles in one channel at a time. Such features inevitably lose information cross different channels. To deal with this problem, we propose a kind of cross-channel features to capture such valuable information by comparing the responses of the same rectangle in different channels. Note that the cross-channel features are also second-order features. Differently, the two rectangles share the same size and position but locate in different channels. The difference of responses between these two rectangles is recorded as feature value. To make the values of different channels comparable, it is required to normalize all the channels before computing such features. See Figure 1(b) for some examples.

Generating Feature Pool: The details for generating the proposed symmetric and cross-channel features are described as follows. Firstly, we denote a rectangle as a 4 dimensional vector $\mathbf{r} = (x, y, w, h)$ in which x, y are the coordinates of the top-left vertex and w, h are the width and height respectively. Then, the valid rectangle set $\mathcal{R}$ is defined as the set of all possible rectangles that are inside of the model region and larger than a predefined area threshold S. In an usual image coordinate system, *e.g*. with the origin being the top-left vertex of the model region, the positive x-axis pointing right and the positive y-axis pointing down, such a set can be represented as:

$$\mathcal{R} = \{\mathbf{r_i} : x_i \leq W - w_i, y_i \leq H - h_i, w_i \leq W, h_i \leq H, w_i \times h_i \geq S, x_i, y_i, w_i, h_i \in \mathbb{N}\} \quad (1)$$

where W and H are the width and height of the model region respectively. With the valid rectangle set, we can further generate a pool of rectangular templates. For symmetric features, we select two symmetric rectangles $\mathbf{r_i}$ and $\mathbf{r_j}$ from $\mathcal{R}$ and randomly generate their channel index. For cross-channel features, we select one rectangle $\mathbf{r_i} \in \mathcal{R}$ and randomly generate two different channel indexes.

Experiments: We conduct our experiments on two public benchmark data sets: the INRIA [2] and Caltech-USA [3] pedestrian data sets. For Caltech-USA data set, we conduct a dense sampling of the training data (every 4 frames) and report our results on the "reasonable", "scale" and "occlusion" subsets of the test set. Our detector achieves three best performances among all the tested detectors and six best performances among the detectors without using motion features.

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? *In ECCV, CVRSUAD workshop*, 2014.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009.

[4] Shanshan Zhang, Armin B. Cremers, and Christian Bauckhage. Informed haar-like features improve pedestrian detection. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 947 – 954, 2014.

APT: Action localization Proposals from dense Trajectories

Jan C. van Gemert[1,2]
j.c.vangemert@gmail.com
Mihir Jain[1]
m.Jain@uva.nl
Ella Gati[1]
ellagati@gmail.com
Cees G. M. Snoek[1,3]
cgmsnoek@uva.nl

[1] Intelligent Systems Lab Amsterdam
University of Amsterdam
Amsterdam, The Netherlands

[2] Computer Vision Lab
Technical University Delft
Delft, The Netherlands

[3] Qualcomm Research Netherlands
Amsterdam, The Netherlands

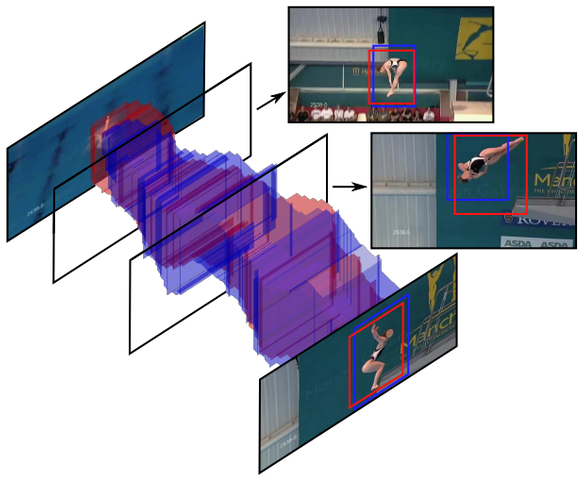


Figure 1: **We propose APT:** an efficient and effective spatio-temporal proposal algorithm for action localization. Different from existing work, which consider different representations for the localization and classification stage, we aptly re-use the dense trajectories as used in the classification representation for localization as well. This allows large-scale deployment with good localization and classification accuracy. For the action *Diving* our blue best proposal results in an overlap of 0.62 with the red ground truth.

	ABO	MABO	Recall	#Proposals
UCF Sports				
Brox & Malik, ECCV 2010 [1]	29.84	30.90	17.02	**4**
Jain et al., CVPR 2014 [3]	63.41	62.71	78.72	1,642
Oneata et al., ECCV 2014 [4]	56.49	55.58	68.09	3,000
Gkioxari & Malik, CVPR 2015 [2]	63.07	62.09	87.23	100
APT (ours)	**65.73**	**64.21**	**89.36**	1,449
UCF 101				
Brox & Malik, ECCV 2010 [1]	15.16	15.06	1.94	**3**
APT (ours)	**47.16**	**46.79**	**46.38**	2,299
MSR-II				
Brox & Malik, ECCV 2010 [1]	2.28	2.34	0	**6**
Jain et al., CVPR 2014 [3]	34.88	34.81	2.96	4,218
Yu & Yuan, CVPR 2015 [6]	n.a.	n.a.	0	37
APT (ours)	**48.02**	**47.87**	**44.33**	6,706

Table 1: **Comparing action proposals** for commonly used metrics against other methods, where *n.a.* denotes not reported values. *ABO* is the Average Best Overlap, *MABO* the Mean ABO over all classes, *Recall* is measured at an IoU overlap $\sigma \geq 0.5$ and the number of proposals is averaged per video. APT outperforms others with a modest number of proposals.

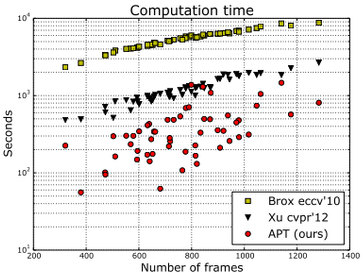


Figure 3: **Evaluating APT computation time** on all videos of MSR-II. APT is faster than the video segmentation of Xu & Corso [5] used in [3]. APT is an order of magnitude faster than the trajectory clustering of Brox & Malik [1].

This paper is on action localization in video with the aid of spatio-temporal proposals. To alleviate the computational expensive segmentation step of existing proposals, we propose bypassing the segmentations completely by generating proposals directly from the dense trajectories used to represent videos during classification. Our Action localization Proposals from dense Trajectories (APT) use an efficient proposal generation algorithm to handle the high number of trajectories in a video. Our spatio-temporal proposals are faster than current methods and outperform the localization and classification accuracy of current proposals on the UCF Sports, UCF 101, and MSR-II video datasets.

[1] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[2] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015.

[3] M. Jain, J. van Gemert, H. Jégou, P. Bouthemy, and C. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014.

[4] D. Oneata, J. Revaud, J. Verbeek, and C Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014.

[5] C. Xu and J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

[6] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015.

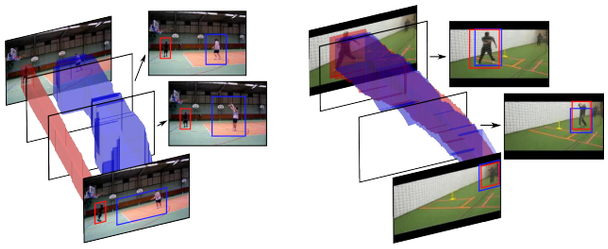


Figure 2: **Success and failure case of APT**. In the left video our method (blue tube) ignores the standing person (red tube) and tracks the moving actor. In the right video, there is ample variation in depth and position yet APT tracks the action well. Our method is intended for actions and thus works best when motion is present.

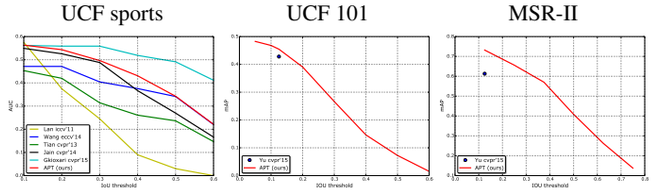


Figure 4: **Classifying action proposals** with a varying IoU threshold σ. Left: AUC on UCF Sports, the fully supervised method of Gkioxari & Malik [2] is best, we are competitive with the unsupervised state of the art. Middle: mAP for UCF 101, we outperform the recent scores on this dataset by Yu & Yuan [6]. Right: mAP for MSR-II, where we outperform the best results on this set by Yu & Yuan [6].

Event Fisher Vectors: Robust Encoding Visual Diversity of Visual Streams

Markus Nagel[1]
mail@markusnagel.com
Thomas Mensink[1]
thomas.mensink@uva.nl
Cees G.M. Snoek[1,2]
cgmsnoek@uva.nl

[1] Intelligent Systems Lab Amsterdam
University of Amsterdam
[2] Qualcom Research Netherlands
Amsterdam

The goal of this paper is to design an effective representation for event recognition in visual streams, such as photo collections and video clips. We are inspired by the success of Fisher Vectors for the encoding of images and videos, where a Gaussian mixture model (GMM) is used as generative probability density function to model 10K-100K local observations per image or video. A visual stream, however, behaves significantly different than local patches or trajectories. Most notably, streams may consist of just tens to hundreds of images, each image in a stream can be described by more discriminative DeepNets, and the temporal structure of the stream can be modelled explicitly.

In order to be more robust against outliers in the small set of observations, we replace the GMM with a Student's-t mixture model (StMM), known for its heavier tails. For observation $x_i \in \mathbb{R}^D$, the StMM is defined as:

$$p(x_i|\theta) = \sum_k \pi_k \, St(x_i|\theta_k), \tag{1}$$

where π_k is the mixing weight and the Student's-t is defined using parameters $\theta_k = \{\text{mean } \mu_k, \text{var } \sigma_k, \text{dof } \nu_k\}$ as:

$$St(x_i|\theta_k) = Z_k \left(1 + \tfrac{1}{\nu_k}\delta_k(x_i)\right)^{-\frac{\nu_k+D}{2}}, \tag{2}$$

with Mahalanobis distance $\delta_k(x_i)$ and normalisation factor Z_k. The *Fisher score*, *i.e.* the derivative w.r.t. the parameters of $\log p(\cdot;\theta)$ is than:

$$G^X_{\mu_k} = \frac{\nu_k + D}{\sigma_k^2 \, \nu_k} \sum_{i=1}^{n} \gamma_i(k) \, \frac{(x_t - \mu_k)}{1 + \frac{1}{\nu_k}\delta_k(x_i)}, \tag{3}$$

where $\gamma_i(k)$ is the *responsibility* value of the k-th mixture component. This score function is, similar to the GMM case, a weighted average of the observations. However, two differences are: (i) the responsibility values $\gamma_i(k)$ are now based on the StMM, and (ii) each observation is weighed by the degrees-of-freedom ν_k and the Mahalanobis distance $\delta_k(x_i)$.

For image classification and retrieval, these Fisher scores are transformed with the *Fisher Information Matrix* (FIM), $F_\theta = \mathbb{E}_{X\sim\theta}[G^X_\theta G^{X\top}_\theta]$ to ensure invariance w.r.t. re-parameterizations of the model. Perronnin *et al.* derived an analytical approximation of the FIM for the GMM, by using the following two assumptions:

Hard-assignment assumption. All patches are sharply peaked around a single component k (*i.e.* $\forall i \, \exists k \; \gamma_k(i) \approx 1$).

Diagonal covariance assumption. Using diagonal covariance matrices in the GMM yields the Fisher scores independent per dimension.

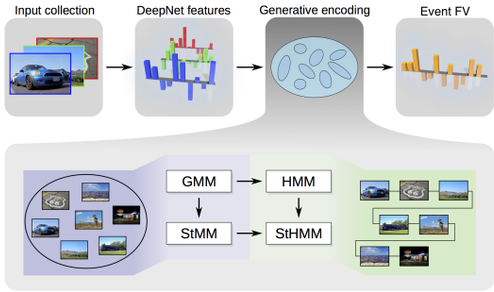
In contrast, the dimensions of the Fisher scores of the StMM are interdependent due to the Mahalanobis distance $\delta_k(x_i)$. In order to derive an analytical approximation, we propose the following assumption:

Constant distance assumption. We assume that in expectation, the Mahalanobis distance $\delta_k(x_i)$ becomes a constant factor. This assumption is based on the *concentration of distances* theorem, which states that for high dimensional data the proportional distance difference between any point and the mean of all data points vanishes. Intuitively, this theorem states that the distance differences $\delta_k(x_i)$, for $k = \{1,\ldots,K\}$ for a specific data point x_i are immaterial.

This results in the following approximation for the mean component:

$$F^{-\frac{1}{2}}_{\mu_k} = \sigma_k \left(\pi_k \, \tfrac{\nu_k}{\nu_k + D}\right)^{-\frac{1}{2}}. \tag{4}$$

To the best of our knowledge, this is the first closed-form approximation of the Fisher information matrix for the Student's-t mixture model. Note that the assumptions above are only used to derive the analytical approximation of the FIM, not for computation of the Fisher scores.



To model the temporal structure of a visual stream, we also propose to use Hidden Markov Models (HMM). The temporal relation in the HMMs is modeled by the latent state z_i, which depends not only on the observation x_i, but also on the latent variable z_{i-1} of the previous observation. The Fisher score of an HMM model w.r.t. the mean μ_k is given by:

$$G^X_{\mu_k} = \sum_{i=1}^{n} \gamma_i(k) \, \nabla_{\mu_k} \log p(x_i|\theta_k), \tag{5}$$

this is identical to the independent models, except that the responsibility values $\gamma_i(k)$ are now computed by:

$$\gamma_i(k) = \frac{p(x_1,\ldots,x_i,z_i{=}k) \; p(x_{i+1},\ldots,x_n|z_i{=}k)}{p(X)}, \tag{6}$$

which reflects the dependence among the images in the collection. For the analytical approximation of the FIM we use once more the crude hard-assignment assumption and obtain the same analytical FIM approximations as for the independent GMM and StMM models.

Different FIM approximations on MED13

FIM	GMM	StMM	HMM	StHMM
Identity	27.6	27.2	28.2	28.5
Empirical	34.7	34.6	34.5	33.8
Closed form	**36.8**	**37.2**	**36.6**	**36.8**

Different models (incl. oracle best per event)

	GMM	StMM	HMM	StHMM	BPE
PEC	80.7	**85.7**	83.6	80.7	88.6
MED13	36.8	**37.2**	36.6	36.8	38.0
CCV	67.4	**69.0**	66.2	66.5	69.1

We evaluated our Event-FV on three recent datasets of photo and video events: Photo Event Collection (PEC, event classification from Flickr collections), TrecVID Media Event Detection (MED13 benchmark with 100 examples per event for training), and Columbia Consumer Video (CCV, Youtube videos of events). For each image in a collection (or sampled frame from a video) we extract the final fully connected layer (4K) from a pre-trained DeepNet (trained on 15K ImageNet classes).

The conducted experiments show that the analytical approximations of the FIM outperform an identity or empirical approximation by a large margin, that the StMM model has a slight edge over the other probabilistic models, and that it results in state-of-the-art performance (when combined with the Mean DeepNet feature and Temporal Pyramids, see paper). This indicates that it is worth to explore more appropriate probabilistic models for the input data within the Fisher vector framework and to derive their analytical FIM approximations. For the task of visual event classification the conclusion is that capturing the heavy tails of the small sample size is more beneficial than modelling the temporal relation of the stream.

Deep Structured Models For Group Activity Recognition

Zhiwei Deng[1]
zhiweid@sfu.ca
Mengyao Zhai[1]
mzhai@sfu.ca
Lei Chen[1]
chenleic@sfu.ca
Yuhao Liu[1]
yla305@sfu.ca
Srikanth Muralidharan[1]
smuralid@sfu.ca
Mehrsan Javan Roshtkhari[2]
mehrsan@sportlogiq.com
Greg Mori[1]
mori@cs.sfu.ca

[1] School of Computing Science
Simon Fraser University
Burnaby, BC, Canada

[2] SportLogiq Inc.
Montreal, QC, Canada

Event understanding in videos is a key element of computer vision systems in the context of visual surveillance, human-computer interaction, sports interpretation, and video search and retrieval. The standard and yet state-of-the-art pipeline for activity recognition and interaction description consists of extracting hand-crafted local feature descriptors either densely or at a sparse set of interest points in the context of a Bag of Words model [4]. In recent years, it has been shown that deep learning techniques can achieve state-of-the-art results for a variety of computer vision tasks including action recognition [3]. However, modeling complex structures and high-level semantic meanings of activities in videos remains a challenge for deep learning methods. Graphical models provide a natural way to hierarchically model group activities and capture the semantic dependencies between group and individual activities [2]. A graphical model defines a joint distribution over states of a set of nodes and the relations between them. Belief propagation is often adopted as a way to infer states or probabilities of variables. However, most of the previous approaches applied a graphical model as a separate phase after using a deep learning model.

In this paper, our main goal is to address the problem of *group activity understanding* and *scene classification* in complex surveillance videos using a deep learning framework. More specifically, we are focused on learning individual activities and describing the scene simultaneously while considering the pair-wise interactions between individuals and their global relationship in the scene. This is achieved by combining a Convolutional Neural Network (CNN) with a probabilistic graphical model as additional layers in a deep neural network architecture into a unified learning framework. The structured deep neural network considers dependencies between individual actions, body poses, and group activities and mimics the message passing process for conducting the inference. Each combination of states is represented as a factor layer. Each neuron represents one type of combination between states and the parameters which are shared across the same type of neurons (weight sharing). We use a sparsely connected factor layer to mimic the message passing from variable nodes to the factor nodes. A layer with reverse connections is used to represent propagation from factor nodes to the variable nodes. The two layers together are considered as one pass iteration. Experimental results show that employing multi-step message passing procedure increases the accuracy of individual and group activity recognition. The multi-step message passing can also be considered as a label refinement procedure through information sharing. In our experimental setup, we use three components to model the interactions between people in the scene: unary terms, scene-action-pose components and global-pose components. Unary terms come from output scores of fine-tuned convolutional neural networks for individual human poses, activities, and the scene label. The scene-action-pose component captures the dependencies between states of human poses, human actions and scene for a group of people. A global-pose component models frequent patterns of different poses for all individuals in the scene. After each step of message passing, the output scores for each activity and scene labels are normalized to probability distribution by a softmax function.

We evaluated both the deep structured model and the features learned

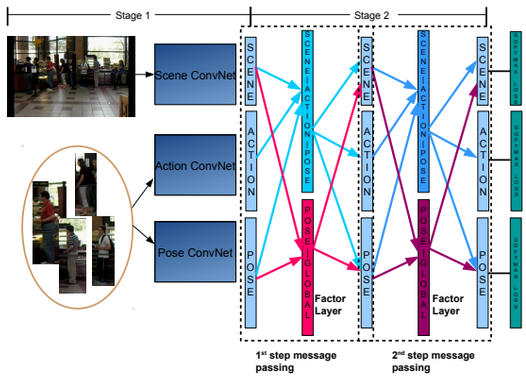


Figure 1: A schematic overview of our message passing CNN framework. Given an image and the detected bounding boxes around each person, our model predicts scores for individual actions and the group activity. The predicted labels are then refined by applying a belief propagation-like neural network. This network considers the dependencies between individual actions, body poses, and the group activity. The model learns the message passing parameters and performs inference and learning in unified framework using back-propagation.

in the factor layers on two challenging group activity datasets: the Collective Activity [1] and the Nursing Home dataset. On both datasets, we observe a significant improvement in the scene classification accuracy by employing the message passing process. Considering the accuracy of a fine-tuned CNN classifier for the scene classification on the collective activity dataset as a baseline which is equal to 48%; the first message passing step reaches 73.6% and the second one improves the accuracy to 78.1%. In addition, by utilizing the learned high-level semantic features and a kernelized SVM as a classifier, we achieved 80.6% accuracy which is comparable to the state-of-the-art. A similar improvement in the accuracy is observed for the Nursing Home dataset, an increase from 69% to 82.5% and 84.0% by applying one and two step message passing.

[1] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *International Conference on Computer Vision Workshops on Visual Surveillance*, pages 1282–1289, 2009.

[2] Tian Lan, Wang Yang, Yang Weilong, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[3] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576. Curran Associates, Inc., 2014.

[4] Heng Wang and C. Schmid. Action recognition with improved trajectories. In *International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013.

Fast Action Retrieval from Videos via Feature Disaggregation

Jie Qin[1]
qinjiebuaa@gmail.com
Li Liu[2]
li2.liu@northumbria.ac.uk
Mengyang Yu[2]
m.y.yu@ieee.org
Yunhong Wang[1]
yhwang@buaa.edu.cn
Ling Shao[2]
ling.shao@ieee.org

[1] Beijing Key Laboratory of Digital Media,
School of Computer Science and Engineering,
Beihang University, China

[2] Computer Vision and Artificial Intelligence Group,
Department of Computer Science and Digital Technologies,
Northumbria University, UK

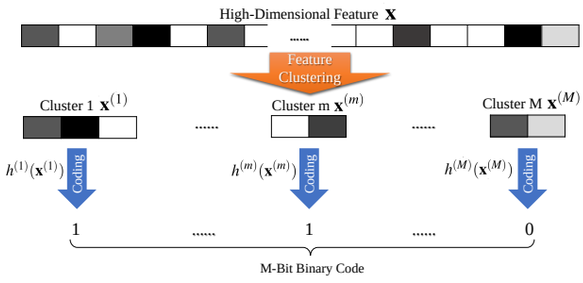


Figure 1: The overall framework of the proposed Disaggregation Hashing.

Motivation. Learning based hashing methods have been actively studied recently. A majority of the approaches [2, 3, 4, 5] have been specifically developed for image retrieval. However, due to the extra temporal information, videos are usually represented by much higher dimensional features compared with images, causing high computational complexity for conventional hashing schemes. Learning hash functions can become quite time-consuming and even intractable when dealing with very high-dimensional video data. Besides, a lot of hashing methods are based on linear projections (*e.g.*, random projection [4]). In terms of mapping the data from the very high-dimensional space to a reduced one, the memory requirements for storing the projection matrix and performing the mapping operation impose heavy burdens.

Contributions. A novel hashing scheme, namely Disaggregation Hashing (DH), is proposed for high-dimensional video data. Our main idea is to disaggregate the original high-dimensional features into several groups of low-dimensional ones, based on which independent hash functions are learned. Figure 1 shows the overall framework of our method.

Given a set of data points $\mathbf{x}$, our goal is to find a binary embedding function $H(\mathbf{x}) = \{-1,+1\}^M$, where M is the length of the code. Disaggregation Hashing learns different hash functions on different subspaces obtained by feature disaggregation. Hence, each bit of the entire code is learned independently, which enhances the scalability on high-dimensional video data and allows fast parallel computation as well. Furthermore, the proposed method only needs to store a D-dimensional projection vector and computing binary codes is of $O(D)$ complexity.

In terms of feature disaggregation or clustering, we show that by incorporating a special structure constraint into the projection matrix $\mathbf{W}$ of PCA, $\mathbf{W}$ can be regarded as the cluster indicator and the solution to feature clustering is identical to the one to PCA. Moreover, as mentioned in [1], k-means clustering has a similar formulation with PCA if we treat $\mathbf{W}$ as the cluster indicator. Therefore, feature clustering can be effectively addressed using k-means clustering along the feature dimensions.

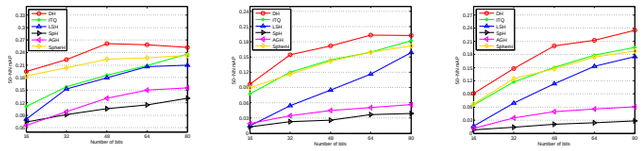
After feature disaggregation, the original feature space $\mathbb{R}^D$ is split into M subspaces $\mathbb{R}^{d_m}$, where $m = 1,...,M$, and $\sum_{m=1}^{M} d_m = D$. The overall hash function $H(\mathbf{x})$ consists of M independent functions $h^{(m)}(\mathbf{x}^{(m)})$, where $\mathbf{x}^{(m)}$ corresponds to the data points from subspace $\mathbb{R}^{d_m}$. Each $h^{(m)}$ maps $\mathbf{x}^{(m)}$ into one binary code $c^{(m)} \in \{-1,+1\}$ and the final code is a concatenation of M-bit codes $(c^{(1)},...,c^{(M)})$. Specifically, we define the above independent hash function as:

$$h^{(m)}(\mathbf{x}^{(m)}) = sgn(\mathbf{w}^{(m)}\mathbf{x}^{(m)} + b^{(m)}) \quad (1)$$

where '$sgn()$' returns '+1' if the argument is positive and '-1' otherwise, and $\mathbf{w}^{(m)} \in \mathbb{R}^{1\times d_m}$ is the projection vector. We aim at learning similarity-preserving codes that hold the minimal average Hamming distance between similar data points. Let $c_p^{(m)}$ denote one bit code for a data point $\mathbf{x}_p$, *i.e.*, $c_p^{(m)} = h^{(m)}(\mathbf{x}_p^{(m)})$, the objective function is defined as:

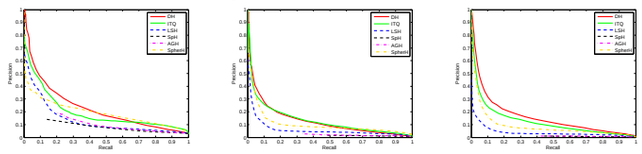
$$\min \sum_{p,q} ||c_p^{(m)} c_q^{(m)} - l_{pq}||^2 \quad (2)$$

where $l_{pq} = +1$ if $\mathbf{x}_p$ and $\mathbf{x}_q$ are similar points, and -1 otherwise. We find the solution to Eq. (2) by addressing a greedy optimization problem and obtain the approximate optimal parameters in Eq. (1).



(a) (b) (c)

Figure 2: Comparison with state-of-the-art methods in terms of mean Average Precision on (a) Hollywood2, (b) HMDB51 and (c) UCF101.



(a) (b) (c)

Figure 3: Comparison with state-of-the-art methods in terms of precision-recall curves@80 bits on (a) Hollywood2, (b) HMDB51 and (c) UCF101.

Results. The experiments are performed for action retrieval on three realistic action datasets, *i.e.*, Hollywood2, HMDB51 and UCF101, and follow the protocols widely used in [2, 3]. We compare our method with five state-of-the-art hashing algorithms. As shown in Figure 2 and 3, our hashing scheme consistently outperforms other methods on different datasets in terms of mean Average Precision and precision-recall curves. Table 1 illustrates the training and coding time of different methods on UCF101. Thanks to feature disaggregation, our method requires much lower time for learning hash functions than most of other methods and computing binary codes for test data points is also very fast.

[1] C. Ding and X. He. K-means clustering via principal component analysis. In *Proc. ICML*, 2004.

[2] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proc. CVPR*, 2011.

[3] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon. Spherical hashing. In *Proc. CVPR*, 2012.

[4] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. STOC*, 1998.

[5] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Proc. NIPS*, 2009.

Bits	Runtime (ms) of Training (Coding)					
	LSH	SpH	AGH	ITQ	SpherH	**DH**
16	(87.0)	4451.8(119.4)	819.0(44.9)	5570.0(82.0)	4774.7(245.0)	1703.5(2.8)
32	(103.9)	4061.8(213.7)	843.9(43.1)	6093.3(114.6)	6756.2(244.3)	1538.7(2.7)
48	(130.6)	4513.7(420.2)	913.3(47.0)	6948.6(140.2)	7809.5(296.7)	1820.2(2.3)
64	(154.3)	4774.9(600.0)	857.7(45.3)	7936.2(159.6)	10139.0(316.9)	2125.3(1.8)
80	(167.9)	5202.6(875.4)	886.4(46.6)	7960.9(190.6)	13720.2(339.9)	2467.7(1.7)

Table 1: Comparison of averaged runtime using different bits on UCF101.

Pose Estimation of Kinematic Chain Instances via Object Coordinate Regression

Frank Michel
Frank.Michel@tu-dresden.de
Alexander Krull
Alexander.Krull@tu-dresden.de
Eric Brachmann
Eric.Brachmann@tu-dresden.de
Michael Ying Yang
Ying.Yang1@tu-dresden.de
Stefan Gumhold
Stefan.Gumhold@tu-dresden.de
Carsten Rother
Carsten.Rother@tu-dresden.de

TU Dresden
Dresden
Germany

Accurate pose estimation of object instances is a key aspect in many applications, including augmented reality or robotics. For example, a task of a domestic robot could be to fetch an item from an open drawer. The poses of both, the drawer and the item have to be known by the robot in order to fulfil the task. 6D pose estimation of rigid objects has been addressed with great success in recent years. In large part, this has been due to the advent of consumer-level RGB-D cameras, which provide rich, robust input data. However, the practical use of state-of-the-art pose estimation approaches is limited by the assumption that objects are rigid. In cluttered, domestic environments this assumption does often not hold. Examples are doors, many types of furniture, certain electronic devices and toys. A robot might encounter these items in any state of articulation.

This work considers the task of one-shot pose estimation of articulated object instances from an RGB-D image. In particular, we address objects with the topology of a kinematic chain of any length, i.e. objects are composed of a chain of parts interconnected by joints. We restrict joints to either revolute joints with 1 DOF (degrees of freedom) rotational movement or prismatic joints with 1 DOF translational movement. This topology covers a wide range of common objects (see our dataset for examples). However, our approach can easily be expanded to any topology, and to joints with higher degrees of freedom.

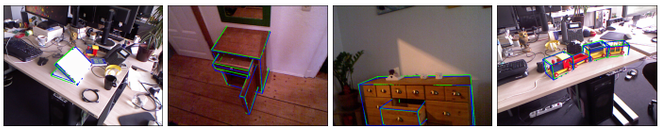


Figure 1: **Our dataset.** These images show results on our dataset. The estimated poses are depicted as the blue bounding volume, the ground truth is shown as the green bounding volume of the object parts.

Articulated Pose Estimation. To estimate the pose of a kinematic chain $\hat{H} = (H_1, \ldots, H_K)$ we need to find the 6D pose H_k for each part k. The problem is however constrained by the joints within the kinematic chain. Therefore, we can find the solution by estimating one of the transformations H_k together with all 1D articulations $\theta_1 \ldots, \theta_{K-1}$, where θ_k is the articulation parameter between part k and $k+1$. We assume the type of each joint and its location within the chain to be known. Given θ_k we can derive the rigid body transformation $A_k(\theta_k)$ between the part k and $k+1$. The transformation $A_k(\theta_k)$ determines the pose of part $k+1$ as follows: $H_{k+1} = H_k A_k(\theta_k)^{-1}$. We can use this to estimate the 6D poses of all parts and thus the entire pose $\hat{H}$ of the chain from a single part pose together with the articulation parameters.

Hypothesis Generation. We use a random forest to produce two kinds of predictions for each pixel i. Given the input depth image, each tree in the forest predicts object probabilities and object coordinates for each separate object part of the kinematic chains (Fig. 2, middle). An articulated pose hypothesis is sampled as follows. We draw a single pixel i_1 from the inner part ($k = 2$) randomly using a weight proportional to the object probabilities $p_k(i)$. We pick an object coordinate prediction $\mathbf{y}_k(i_1)$ from a randomly selected tree t. Together with the camera coordinate $\mathbf{x}(i_1)$ at the pixel this yields a 3D - 3D correspondence $(\mathbf{x}(i_1), \mathbf{y}_k(i_1))$. Two more

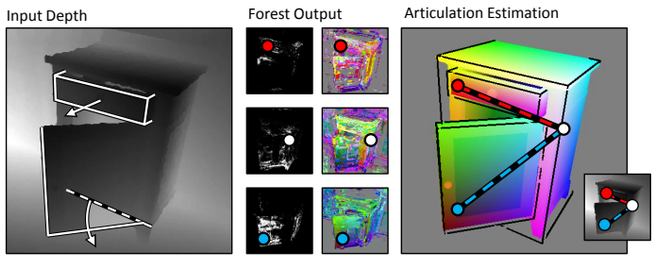


Figure 2: **Articulation estimation.** Left: Input depth image, here shown for the cabinet. The drawer is connected by a prismatic joint and the door is connected by a revolute joint (white lines are for illustration purposes). Middle: Random forest output. Top to bottom: Drawer, base, door, where the left column shows part probabilities and the right the object coordinate predictions, respectively. Right: Articulation estimation between the parts of the kinematic chain using 3D-3D correspondences between the drawer / base and door / base. Note that the three correspondences (red, white, blue) are sufficient to estimate the full 8D pose.

correspondences $(\mathbf{x}(i_2), \mathbf{y}_{k+1}(i_2))$ and $(\mathbf{x}(i_3), \mathbf{y}_{k-1}(i_3))$ are sampled in a square window around i_1 from the neighbouring kinematic chain parts $k+1$ and $k-1$. We can now use these correspondences to estimate the two articulation parameters θ_{k-1} and θ_k between part k and its neighbours. We derive $A_k(\theta_k)$ and $A_{k+1}(\theta_{k+1})$ and map the two sampled points $\mathbf{y}_{k+1}(i_2)$ and $\mathbf{y}_{k-1}(i_3)$ to the local coordinate system of part k. We have now three correspondences between the camera system and the local coordinate system of part k, allowing us to calculate the 6D pose H_k using the Kabsch algorithm. The 6D pose H_k together with the articulation parameters yields the pose $\hat{H}$ of the chain. Fig. 2 illustrates this process

Results. We contribute a new dataset consisting of over 7000 frames annotated with articulated poses of different objects: two cupboards, a laptop and a toy train[1]. The objects show different grades of articulation ranging from 1 joint to 3 joints. When compared to the 6D pose estimation pipeline of Brachmann *et al.* [1] our method shows superior results (89% averaged over all sequences and objects) in comparison to the baseline (29%). Qualitative results are shown in Fig. 1. Employing articulation constraints within the kinematic chain results in better performance on the individual parts as well as for the kinematic chains in its entirety. Our approach of pose sampling for kinematic chains does not only need less correspondences, it is also robust when dealing with heavy self occlusion.

Conclusion. We present a method for pose estimation of kinematic chain instances from RGB-D images. We employ the constraints introduced by the joints of the kinematic chain to generate pose hypotheses using K 3D-3D correspondences for kinematic chains consisting of K parts.

[1] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014.

[1]This dataset will be part of the ICCV 2015 pose challenge: http://cvlab-dresden.de/iccv2015-pose-challenge

Hybrid One-Shot 3D Hand Pose Estimation by Exploiting Uncertainties

Georg Poier[1], Konstantinos Roditakis[2,3], Samuel Schulter[1]
{poier,schulter}@icg.tugraz.at/croditak@ics.forth.gr
Damien Michel[2], Horst Bischof[1], Antonis A. Argyros[2,3]
{michel,argyros}@ics.forth.gr/bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision
Graz University of Technology, Austria

[2] Institute of Computer Science
FORTH, Greece

[3] Computer Science Department
University of Crete, Greece



(a) (b) (c)

Figure 1: (a) A learned joint regressor might fail to recover the pose of a hand due to ambiguities or lack of training data. (b) We make use of the inherent uncertainty of a regressor by enforcing it to generate multiple proposals. The crosses show the top three proposals for the proximal interphalangeal joint of the ring finger for which the corresponding ground truth position is drawn in green. The marker size of the proposals corresponds to degree of confidence. (c) A subsequent model-based optimisation procedure exploits these proposals to estimate the true pose.

Traditionally, the task of hand pose estimation was approached mostly by model-based or data-driven schemes. Model-based approaches have been shown to perform well in a wide range of scenarios. However, they require initialisation and cannot recover easily from tracking failures that occur due to fast hand motions. Data-driven approaches, on the other hand, can quickly deliver a solution, but the results often suffer from lower accuracy or missing anatomical validity compared to those obtained from model-based approaches. We propose to combine the merits of both schemes in a hybrid approach. This way, the method provides anatomically valid and accurate solutions without requiring manual initialisation or suffering from track losses.

Main Idea For the task of 3D hand pose estimation, the substantial similarities between individual fingers and complex finger interactions cause ambiguities and uncertainties which are often disregarded by previous works. In contrast, we have the model-based step exploiting the inherent uncertainties of the data-driven part. First, a learned regressor is employed to deliver multiple initial hypotheses for the 3D position of each hand joint. These proposals approximate the distribution of joint positions under the learned model and thus capture the uncertainty of the model. Subsequently, the parameters of an anatomically valid hand pose are found by model-based optimisation which exploits the uncertainties captured by the proposal distributions. To do this, the optimisation is privy to internal information from the learned regressor. In this way failures of the regressor can be corrected during optimisation (see Fig. 1).

Proposal Generation For the generation of an approximated proposal distribution we build upon a prominent approach for body pose estimation [2], which has also been previously adapted for hand pose estimation [5]. The approach relies on Random Forests [1] to infer a 3D distribution of likely hand joint locations. Using the discriminative Random Forest based method, inference of the individual joint proposals is completely independent from the other joints. While, in this way, the complex dependencies do not need to be modeled, the resulting proposals are not necessarily compatible with anatomical constraints.

Optimisation In order to obtain a valid pose we employ a predefined model of a hand. Such a model can be specified by a number of parameters defining the global position and orientation of the hand as well as the joint angles for each finger joint. During optimisation these parameters are constrained based on anatomical studies [3] which avoids impossible configurations. The goal of the optimisation is to find the model parameters which best describe the modes of the proposal distributions, obtained from the regression forest, while still obeying anatomic constraints.

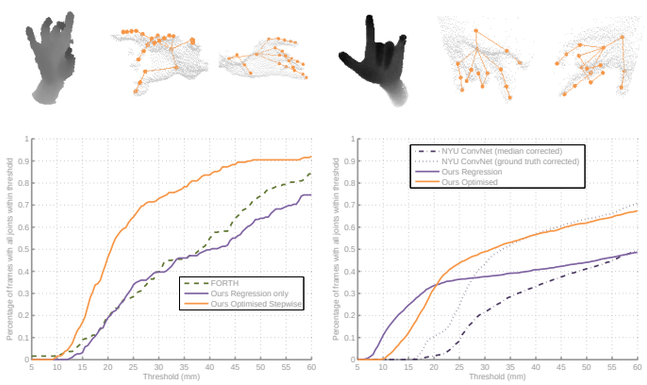


Figure 2: Top row: Qualitative results for two input depth maps rendered under different viewpoints. Bottom row: The ratio of frames with *all* joints within a certain distance to the ground truth as a function of this distance. Comparing the proposed approach (orange) to regression only (purple) and [4] (green, dotted) in the left figure on a synthetic sequence, and to the CNN based hybrid approach [6] (dotted, dark purple) in the right figure on the NYU dataset. See the paper for more results.

Results Experimental results on various datasets (ICVL, NYU, synthetic) show that our specific combination of a data-driven and model-based scheme outperforms state-of-the-art representatives of the model-based, data-driven and hybrid paradigms (see, *e.g.* Fig. 2). In line with this, we found it especially effective to make the optimisation scheme aware of the uncertainty of the regressor. Moreover, using a simple observation we were able to make the optimisation converge much faster: Given the global orientation of the hand, the fingers can move almost independently of each other. This allows us to split the optimisation problem into sub-problems of lower dimensionality, which resulted in a dramatic speed-up without loss of accuracy. See the paper and our website[1] for more details and more results.

Acknowledgment This work has been supported by the EU projects 3D-PITOTI (ICT-2011-600545), FP7-IP-288533 Robohow and FP7-ICT-2011-9 WEARHAP, and by the Austrian Research Promotion Agency under the FIT-IT Bridge programme, project TOFUSION (838513).

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Ross B. Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew W. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. ICCV*, 2011.

[3] John Y. Lin, Ying Wu, and Thomas S. Huang. Modeling the constraints of human hand motion. In *Proc. Workshop on Human Motion*, 2000.

[4] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Proc. BMVC*, 2011.

[5] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proc. ICCV*, 2013.

[6] Jonathan Tompson, Murphy Stein, Yann LeCun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. on Graphics*, 33(5):169:1–169:10, 2014.

[1] Online resources at `http://lrs.icg.tugraz.at/research/hybridhape/`

Particle dynamics and multi-channel feature dictionaries for robust visual tracking

Srikrishna Karanam
karans3@rpi.edu
Yang Li
yangli625@gmail.com
Richard J. Radke
rjradke@ecse.rpi.edu

Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY USA

Introduction: Recent advances in the application of compressive sensing to traditional computer vision problems such as face recognition [5] inspired several visual tracking approaches based on sparse representations. The core idea of these approaches is to build an appearance model of the object using several pre-defined templates. The problem of tracking the object is then cast as finding a sparse approximation in the subspace spanned by the templates. In [3], Mei and Ling introduced the l_1 tracker, demonstrating impressive tracking results. Given an appearance model $\mathbf{A} = [\mathbf{t}_1 \cdots \mathbf{t}_n] \in \mathbb{R}^{m \times n}$ of an object formed using a set of templates $\mathbf{t}_i \in \mathbb{R}^m, i = 1, \ldots, n$, they express a tracking result $\mathbf{y} \in \mathbb{R}^m$ as $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$, where $\mathbf{x} \in \mathbb{R}^n$ is the sparse coefficient vector that is to be recovered, and $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is used to account for partial occlusions. The l_1 tracking algorithm hypothesizes that $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ are sparse for a good tracking candidate, and recovers them by solving an l_1 regularized least squares problem. Subsequently, the candidate with the least projection error in the template subspace is chosen as a tracking target, and a Bayesian state inference model in a particle filter framework is used to track the object over time.

Contributions: In spite of the impressive progress achieved by the l_1 tracker and its recent variants, several issues remain that often lead to tracking failures. First, most related methods employ only a fixed-variance Gaussian distribution to represent the state transition model. Such a fixed-variance state transition model can cause significant drift errors in the approximation of the particle filter, resulting in severe tracking failures. In our work, we propose to mitigate this problem by adaptively learning the variance from past states using a dynamic state transition model. Specifically, we employ an autoregressive model in conjunction with block Hankel matrices to continuously learn the dynamics from past data. Second, most existing approaches use extremely low resolution image intensity features as part of the appearance model. Such features do not capture sufficient visual information required to reliably track the object and avoid drift. To mitigate this problem, we propose a three-channel appearance dictionary comprised of image intensity information, normalized image gradient magnitudes, and histograms of oriented gradients to construct an appearance model of the object. Finally, tracking algorithms typically employ a fixed number of particles to approximate the posterior distribution (e.g., 600 in [3], 400 in [6]). In this work, we demonstrate that many particles are not necessary to reliably track an object, given the initial location. Specifically, we propose to adapt the number of particles required during the state estimation process using the Kullback-Leibler (KL) distance measure [2].

Approach overview: We formulate visual tracking as a sparse representation problem in a particle filtering framework. Given the initial location of the target to be tracked, we warp the image into a 64×64 pixel template, thereby representing the position of the target in each frame using a four-dimensional state vector $\mathbf{s_t} \in \mathbb{R}^4$. By perturbing the initial location by a few pixels (typically, 1–3), we form m such templates. In our experiments, we set $m = 10$. We then construct three appearance dictionaries using these templates: an intensity channel dictionary $\mathbf{A}^1 = [\mathbf{t}_1^1 \cdots \mathbf{t}_m^1]$, a normalized gradient magnitude dictionary $\mathbf{A}^2 = [\mathbf{t}_1^2 \cdots \mathbf{t}_m^2]$, and a Histogram of Oriented Gradients (HOG) [1] dictionary $\mathbf{A}^3 = [\mathbf{t}_1^3 \cdots \mathbf{t}_m^3]$, where each dictionary $\mathbf{A}^j \in \mathbb{R}^{d_j \times m}$. Now, given a potential target particle $\mathbf{y}$, we compute its intensity feature vector $\mathbf{y}^1$, normalized gradient magnitude vector $\mathbf{y}^2$, and HOG vector $\mathbf{y}^3$. In each feature channel, we hypothesize that a good target candidate can be represented as a sparse linear combination of the dictionary templates, and recover the sparse vector by solving the following convex optimization problem:

$$(\mathbf{x}^{j*}, \boldsymbol{\varepsilon}^{j*}) = \underset{\mathbf{x}^j, \boldsymbol{\varepsilon}^j}{\arg\min} \|\mathbf{x}^j\|_1 + \|\boldsymbol{\varepsilon}^j\|_1 \text{ s.t. } \mathbf{y}^j = \mathbf{A}^j\mathbf{x}^j + \boldsymbol{\varepsilon}^j, j = 1,2,3 \quad (1)$$

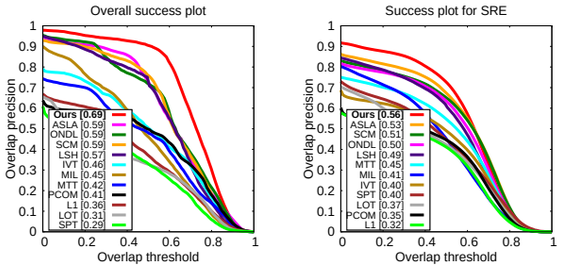
where $\boldsymbol{\varepsilon}^j$ is used to account for error and partial occlusion, and $\mathbf{x}^j$ is the sparse coefficient vector we wish to recover.

Tracking in a particle filtering framework proceeds by generating several hypotheses, and testing each for its likelihood. In our context, each hypothesis is a candidate particle, represented by its state vector. We search for potential candidate particles using an adaptive state transition model incorporating dynamic information. Our key insight is that learning from the dynamics of past state vectors can lead to improved and efficient search for new candidate particles, leading to both accuracy and speed benefits. Formally, if $\mathbf{s}_t \in \mathbb{R}^4$ is the current state vector, we estimate the elements of the next state vector $\mathbf{s}_{t+1}$ using the following transition model:

$$s_{t+1}(i) = s_t(i) + r(i)\sigma_{t+1}(i) \quad (2)$$

where $s_{t+1}(i)$ is the ith element of $\mathbf{s}_{t+1}$, $r(i) \sim \mathcal{N}(0,1)$ is a normally distributed random number, and $\sigma_{t+1}(i)$ is the ith element of the variance vector $\boldsymbol{\sigma}_{t+1} \in \mathbb{R}^4$ that we estimate on-the-fly using the dynamics of the past states.

Results: We tested our algorithm on 25 publicly available video sequences[1] that represent several challenging aspects in visual tracking: illumination variation, scale variation, occlusion, non-rigid object deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutter, and low resolution.



(a) (b)

Figure 1: Success plots averaged over all the 25 test sequences. In both plots, the area under the curve (AUC) is reported in the legend.

The overall success plot and the success plot for the spatial robustness evaluation (SRE) test [4] are shown in Figure 1. We see that our approach results in a significant 10% overall improvement and an improvement of 3% for the SRE test when compared to the state of the art.

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] Dieter Fox. Adapting the sample size in particle filters through KLD-sampling. *IJRR*, 22(12):985–1003, 2003.

[3] Xue Mei and Haibin Ling. Robust visual tracking using l_1 minimization. In *ICCV*, 2009.

[4] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

[5] Allen Y Yang, Zihan Zhou, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Fast l_1-minimization algorithms for robust face recognition. *IEEE T-IP*, 22(8):3234–3246, 2013.

[6] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust visual tracking via structured multi-task sparse learning. *IJCV*, 101(2):367–383, 2013.

[1] Videos demonstrating our tracking performance can be found in the supplementary material.

Collaborative Correlation Tracking

Guibo Zhu[1]
gbzhu@nlpr.ia.ac.cn
Jinqiao Wang[1]
jqwang@nlpr.ia.ac.cn
Yi Wu[2]
ywu.china@yahoo.com
Hanqing Lu[1]
luhq@nlpr.ia.ac.cn

[1] National Laboratory of Pattern Recognition,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.

[2] B-DAT, Nanjing University of Information
Science and Technology,
Nanjing, China

Motivation. Correlation filter based tracking has attracted many researchers' attention in recent years for high efficiency and robustness. Most existing works [1, 2, 4] focus on exploiting different characteristics with correlation filters for visual tracking, *e.g.*, circulant structure, kernel trick, effective feature representation and context information.

Despite its good performance, most of these correlation methods have two main limitations, the first is how to adjust the object scale efficiently. In order to consistently track the object, Danelljan *et.al.* [2] proposed a separate 1-dimensional correlation filter to estimate the target scale, but they only use the original feature space as the object representation.

The second limitation is how to handle the model drift problem caused by the long-term occlusion or out-of-view, which is a very important problem for online tracking [5]. One common mechanism is to introduce a detection module which can select some effective candidates to rectify the base correlation tracker.

Contributions. In this paper, we design a novel online CUR filter for detection. CUR matrix approximation computes the low rank approximation of a given matrix by using the actual rows and columns of the matrix and have been studied in the area of theoretical computer science for large matrix approximation [3].

In the long-term tracking process, all of the historical object representations can form a large data matrix for the current frame which fits for the CUR theory. The large data matrix can be fast approximated by online CUR for representing the intrinsic object structure. In this work we develop an online CUR for learning an online detection filter by random sampling. The online CUR filter can not only exploit the low rank property of object representation [7] in the spatial-temporal domain of tracking, but also project the historical object representation matrix into a subspace with error upper bound so as to achieve a robust object representation. The low rank property of object representation is prevalent in long-term tracking and could be used to alleviate the model drift.

Meanwhile, we propose multi-scale kernelized correlation filter as our tracking filter by embedding the scale variation into the kernelized correlation filter while forming a separate pyramid of object representation. The collaboration between CUR filter and multi-scale kernelized correlation filter constructs the proposed tracker–Collaborative Correlation Tracker (CCT)[1].

Table 1: **Results on the Princeton Tracking Benchmark:** successful rates (%) and rankings (in parentheses) for different categorizations. The best results are in **red** and the second best in blue. hu.:human; an.:animal; ri.:rigid; pa.:passive;ac.:active

Algo.	Avg.	target type			target size		movement		occlusion		motion type	
	Rank	hu.	an.	ri.	large	small	slow	fast	yes	no	pa.	ac.
CCT	**1**	**50(1)**	**51(1)**	**64(1)**	**53(1)**	**57(1)**	**69(1)**	**50(1)**	**44(1)**	**71(1)**	**63(1)**	**53(1)**
Struck	2.82	35(2)	47(3)	53(4)	45(2)	44(4)	58(2)	39(2)	30(4)	64(2)	54(4)	41(2)
VTD	3.27	31(5)	49(2)	54(3)	39(4)	46(2)	57(3)	37(3)	28(5)	63(3)	55(3)	38(3)
RGBdet	4.36	27(7)	41(5)	55(2)	32(7)	46(3)	51(5)	36(4)	35(2)	47(6)	56(2)	34(5)
CT	5.36	31(4)	47(4)	37(7)	39(3)	34(7)	49(6)	31(5)	23(8)	54(4)	42(7)	34(4)
TLD	5.64	29(6)	35(7)	44(5)	32(6)	38(5)	52(4)	30(7)	34(3)	39(7)	50(5)	31(7)
MIL	5.82	32(3)	37(6)	38(6)	37(5)	35(6)	46(7)	31(6)	26(6)	49(5)	40(8)	34(6)
SemiB	7.73	22(8)	33(8)	33(8)	24(8)	32(8)	38(8)	24(8)	25(7)	33(8)	42(6)	23(8)
OF	9.00	18(9)	11(9)	23(9)	20(9)	17(9)	18(9)	19(9)	16(9)	22(9)	23(9)	17(9)

Results. To evaluate the performance of the proposed method, we follow the metric used in [6]. Fig. 1 shows precision and success plots which contains the mean distance and overlap precision over all the 50 sequences. Table 1 has shown the results of CCT on the Princeton Tracking Benchmark. As shown in Fig. 1 and Table 1, our approach CCT outperforms the other trackers with a significant gain in accuracy.

[1] The source code and experimental results are available at `http://www.nlpr.ia.ac.cn/iva/homepage/jqwang/publications.htm`

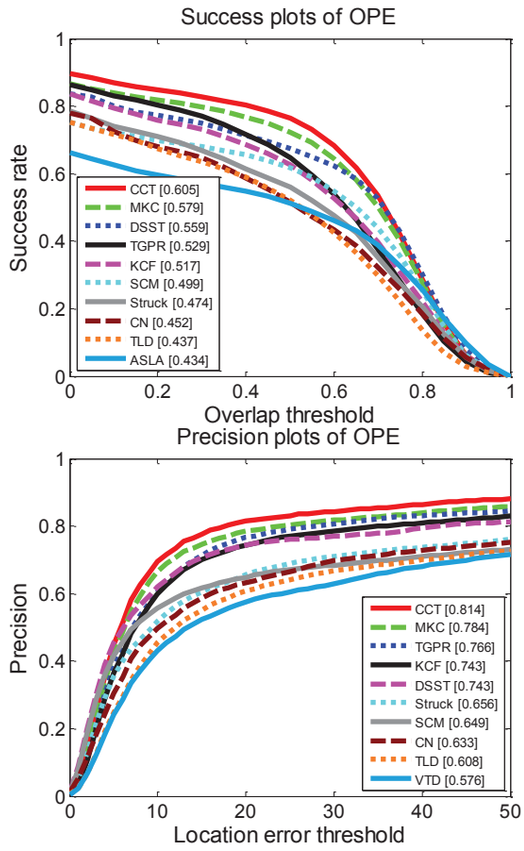


Figure 1: Precision and success plots in the benchmark [6] (best-viewed in high-resolution).

[1] D. Bolme, J. Beveridge, B. Draper, and Y. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, pages 2544–2550. IEEE, 2010.

[2] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014.

[3] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *SIAM symposium on Discrete algorithm*, pages 1117–1126. ACM, 2006.

[4] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015.

[5] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. *TPAMI*, 26(6):810–815, 2004.

[6] Y. Wu, J. Lim, and M. H. Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418. IEEE, 2013.

[7] T. Zhang, K. Jia, C. Xu, Y. Ma, and N. Ahuja. Partial occlusion handling for visual tracking via robust part matching. In *CVPR*, pages 1258–1265. IEEE, 2014.

Enable Scale and Aspect Ratio Adaptability in Visual Tracking with Detection Proposals

Dafei Huang[12]
huangdafei1012@163.com
Lei Luo[†1]
l.luo@nudt.edu.cn
Mei Wen[12]
meiwen@nudt.edu.cn
Zhaoyun Chen[12]
chenzhaoyun09@163.com
Chunyuan Zhang[12]
cyzhang@nudt.edu.cn

[1] College of Computer
National University of Defense Technology
Changsha, China

[2] National Key Laboratory of Parallel and Distributed Processing
National University of Defense Technology
Changsha, China

Among increasingly complicated trackers in visual tracking area, recently proposed correlation filter based trackers have achieved appealing performance despite their great simplicity and superior speed. However, the filter input is a bounding box of fixed size, so they are not born with the adaptability to target's scale and aspect ratio changes. Although scale-adaptive variants have been proposed, they are not flexible enough due to pre-defined scale sampling manners. Moreover, to the best of our knowledge, no correlation filter variant has been proposed to handle aspect ratio variation. To tackle this problem, this paper integrates the class-agnostic detection proposal method, which is widely adopted in object detection area, into a correlation filter tracker, and presents KCFDP tracker.

The correlation filter part of KCFDP is based on KCF[2] with some modifications. We extend the HOG feature in KCF to a combination of HOG, intensity, and color naming by simply concatenating the three features, resulting in 42 feature channels. The model updating scheme in KCF, which is simple linear interpolation, is substituted with a more robust scheme presented in [1]. EdgeBoxes[4] is adopted to generate flexible detection proposals and enable the scale and aspect ratio adaptability of our tracker. It traverses the whole image in a sliding window manner, and scores every sampled bounding box according to the number of contours that are wholly enclosed. To accelerate EdgeBoxes and produce less unnecessary proposals, we set the minimum proposal area and aspect ratio range dynamically in sliding window sampling according to the current target size.

In the tracking pipeline, KCF is firstly performed to estimate the preliminary target location $\mathbf{l}^d$. Within a patch $\mathbf{z}^d$ extracted from current frame, KCF locates the target center according to the location of the maximum element in $\mathbf{f}$:

$$\hat{\mathbf{f}}(\mathbf{z}^d) = \hat{\mathbf{k}}^{\bar{\mathbf{x}}\mathbf{z}^d} \cdot \hat{\alpha}, \tag{1}$$

where $\hat{}$ denotes the discrete Fourier transform (DFT). α and $\bar{\mathbf{x}}$ are the coefficient matrix and current target appearance respectively. $\mathbf{k}$ represents the kernel correlation operation defined as:

$$\mathbf{k}^{\mathbf{x}'\mathbf{x}''} = \exp\left(-\frac{1}{\sigma^2}\left(\|\mathbf{x}'\|^2 + \|\mathbf{x}''\|^2 - 2\mathcal{F}^{-1}\left(\sum_c \hat{\mathbf{x}}'^*_c \cdot \hat{\mathbf{x}}''_c\right)\right)\right), \tag{2}$$

where σ is the bandwidth of Gaussian kernel, $\mathcal{F}^{-1}$ denotes the inverse DFT and $*$ refers to complex conjugation. The subscript c here means the cth channel of feature.

Then EdgeBoxes is performed on a patch $\mathbf{z}^p$ centered at the preliminary location $\mathbf{l}^d$, and of size slightly larger than the current estimated target size. We only take the top 200 proposals and further filter them with proposal rejection: if the intersection over union (IoU) between a proposal and the current detected target is higher than 0.9 or lower than 0.6, the proposal is rejected. We then evaluate each accepted proposal $\mathbf{p}$ using:

$$f(\mathbf{p}) = sum(\mathbf{k}^{\bar{\mathbf{x}}\mathbf{p}} \cdot \alpha), \tag{3}$$

where $sum()$ means the summation of all the elements in a matrix. The proposal with maximum f, whose location and size are denoted as $\mathbf{l}^p$ and (w^p, h^p), is the most promising candidate, and thus adopted to update the target location $\mathbf{l}$ and size (w,h) with a damping factor γ:

$$\mathbf{l} = \mathbf{l}^d + \gamma(\mathbf{l}^p - \mathbf{l}^d); \quad (w,h) = (w,h) + \gamma((w^p,h^p) - (w,h)). \tag{4}$$

The dataset adopted to evaluate our KCFDP is from [3], which consists of 50 sequences (including 51 tracking targets) with many challenging attributes. 29 trackers from [3] and five additional correlation filter trackers are included in comparison. Three experiments are conducted on a 28-sequence subset with significant scale variation, a 14-sequence subset with dramatic aspect ratio variation, and the whole dataset respectively. Evaluation results are provided in two kinds of plots: Precision Plot that indicates the ratio of frames with center location error (CLE) below a certain threshold, where trackers are ranked using a threshold of 20 pixels, and Success Plot that indicates the percentage of frames with IoU larger than a threshold comparing to ground truth, where trackers are ranked using the area under curve (AUC). KCFDP reports the highest accuracy in all the experiments as shown in Fig.1, while running efficiently at an average speed of 20.8 FPS.

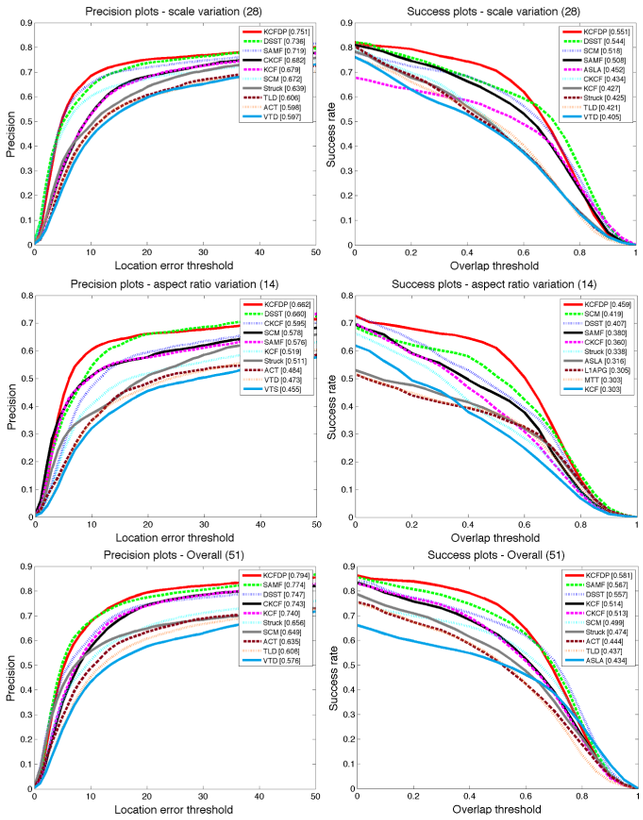


Figure 1: Evaluation results for scale adaptability (upper), aspect ratio adaptability (middle), and on the whole dataset (lower).

[1] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, pages 1090–1097, 2014.

[2] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015. doi: 10.1109/TPAMI.2014.2345390.

[3] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418, 2013.

[4] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405, 2014.

Multiple Frames Matching for Object Discovery in Video

Otilia Stretcu
otiliastr@gmail.com
Computer Laboratory
University of Cambridge, UK

Marius Leordeanu
marius.leordeanu@imar.ro
Institute of Mathematics of the Romanian Academy
Teamnet International, Romania

Automatic discovery of foreground objects in video sequences is important in computer vision, with applications to object tracking, video segmentation and weakly supervised learning. This task is related to co-segmentation [4, 5] and weakly supervised localization [2, 6]. We propose an efficient method for the simultaneous discovery of foreground objects in video and their segmentation masks across multiple frames. We offer a graph matching formulation for bounding box selection and refinement using second and higher order terms. It is based on an Integer Quadratic Programming formulation and related to graph matching and MAP inference [3]. We take into consideration local frame-based information as well as spatiotemporal and appearance consistency over multiple frames. Our approach consists of three stages. First, we find an initial pool of candidate boxes using a novel and fast foreground estimation method in video (VideoPCA) based on Principal Component Analysis of the video content. The output of VideoPCA combined with Edge Boxes [8] is then used to produce high quality bounding box proposals. Second, we efficiently match bounding boxes across multiple frames, using the IPFP algorithm [3] with pairwise geometric and appearance terms. Third, we optimize the higher order terms using the Mean-Shift algorithm [1] to refine the box locations and establish appearance regularity over multiple frames. We make the following contributions:

1. A novel formulation with efficient discrete and continuous optimization for joint selection and refinement of object bounding boxes in video. Our approach encourages appearance, geometric and spatiotemporal consistency over multiple frames with a formulation that considers relations between neighboring as well as farther away frames (Figure 1).

2. A fast method for estimating foreground regions based on Principal Component Analysis of the video content that is able to handle cases of moving or changing backgrounds.

Given a video V as a sequence of temporally ordered frames $V = \{I_1, I_2, ..., I_n\}$, the goal is to find the main foreground object present in the sequence. For each frame I_i we have a pool of n_i candidate bounding boxes B_{ia}'s, obtained automatically. For each box a we store its xy location in image i in θ_{ia}. Let x_{ia} be an indicator variable corresponding to bounding box B_{ia} (- the a-th bounding box of frame i), such that x_{ia} is 1 if the bounding box B_{ia} is selected, and 0 otherwise. The indicator variables are arranged in vector form $\mathbf{x}$ such that its ia-th element corresponds to x_{ia}. We impose the constraint that a single box can be selected per frame: $\sum_a x_{ia} = 1$. Thus, vector $\mathbf{x}$ represents a discrete solution that indicates which box is selected. Similarly, we keep the continuous location parameters in a global vector θ, with θ_{ia} being the parameters of box B_{ia}. The problem becomes one of joint box matching and location refinement over multiple frames, in which we optimize over both $\mathbf{x}$ and θ:

$$\begin{aligned} (\mathbf{x}^*, \theta^*) &= \underset{x,\theta}{\operatorname{argmax}} (\mathbf{x}^T \mathbf{M} \mathbf{x} + \sum_{i=1}^{n} H_i(\mathbf{x}, \theta)) \\ \text{s.t.} \quad & \sum_a x_{ia} = 1 \ \ \forall i, \ \ \mathbf{x} \in \{0,1\}^n. \end{aligned}$$

We adopt a two stage approach. First stage performs discrete optimization in which the quadratic function is optimized by finding the correct frame-to-box matches $\mathbf{x}$, given a fixed θ. The second stage, when $\mathbf{x}^*$ is fixed, the location θ is refined by the non-parametric Mean-Shift to locally optimize the foreground pixels density.

The **pairwise terms** have the following form:

$$M_{ia;jb} = \exp(\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb}). \quad (1)$$

$\mathbf{g}_{ia;jb} = [(f_{ia} + f_{jb}), (v_{ia} + v_{jb}), (c_{ia} + c_{jb}), m_{ia;jb}, o_{ia;jb}, d_{ia;jb}, s_{ia;jb}, r_{ia;jb}]$, such that: 1) f_{ia} (and f_{jb} respectively) measure the absolute difference between the average foreground soft-segmentation values in box a (and b respectively) vs. average foreground values in the surrounding background from its frame i (and j respectively). 2) Similarly, v_{ia} and v_{jb} measure the absolute difference in the relative mean speed between the box and the surrounding background, computed using the DeepFlow method [7]. 3) c_{ia} and c_{jb} measure the distance between the box center and the image center. 4) $m_{ia;jb}$ reflects the quality of the match between the standard HOG descriptor of box a and that of box b. 5) $o_{ia;jb}$ measures the overlap-over-union between the boxes. 6) $d_{ia;jb}$ is the distance between the boxes' centers. 7) $s_{ia;jb}$ is the ratio of the difference between the boxes' areas to the maximum of the two areas. 8) $r_{ia;jb}$ estimates the change in shape, as difference between the boxes' aspect ratios. We learn parameters $\mathbf{w}$ such that $\exp(\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb})$ approximates a target $t = 1$ if the matched pair is correct and equal to a small positive value ($t = 0.1$) otherwise. We want $\exp(\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb}) \approx t$. We take the log on both sides $\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb} \approx \log t$ and obtain a linear system of equations over a set of training samples, which we approximately solve by ridge regression.

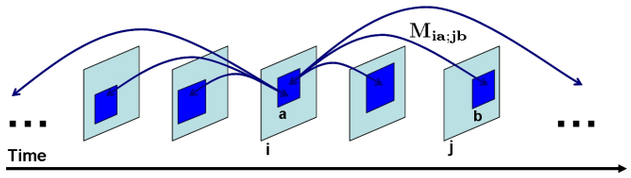


Figure 1: The structure of our box-matching formulation allows links between neighboring frames as well as farther away ones.

For the **higher order terms** we estimate foreground and background color probability distributions from the $2k+1$ bounding boxes and their frames connected to the current one (including itself) and estimate the soft foreground segmentation by classifying individual pixels based on their color. The higher order term $H_i(\mathbf{x}, \theta) = \lambda c_k(i)$ measures the difference between average foreground segmentation values inside the box and outside of it.

We test our method on the large scale YouTube-Objects dataset and obtain state-of-the-art results on several object classes, at a low computational cost. We conclude that our approach, by proposing efficient bounding box generation with VideoPCA, fast matching over multiple frames, and high quality soft-segmentation estimation, covers and extends current approaches in object discovery in video.

[1] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence*, 24(5), 2002.

[2] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3), 2012.

[3] M. Leordeanu, M. Hebert, and Rahul Sukthankar. An integer projected fixed point method for graph matching and map inference. In *NIPS*, 2009.

[4] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[5] J.C. Rubio, J. Serrat, and A. López. Video co-segmentation. In *ACCV*, 2012.

[6] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013.

[7] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *ICCV*, 2013.

[8] C. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.

Surface Based Object Detection in RGBD Images

Siddhartha Chandra*[1]
siddhartha.chandra@inria.fr
Grigorios G. Chrysos*[2]
grigoris.chrysos@gmail.com
Iasonas Kokkinos[1]
iasonas.kokkinos@inria.fr

[1] INRIA GALEN
Centrale Supélec
Paris, France
[2] Imperial College
London, U.K.

*These authors contributed equally to this work. G. Chrysos was at Centrale Supélec while conducting this research.

Introduction Intra-class variation in object detection is due to both intrinsic (texture and shape), and extrinsic (viewpoint and illumination) factors. RGBD data simplifies the problem of detection by introducing depth maps, which are invariant to texture and illumination. However, viewpoint variation still remains the Achilles heel for most object-detection frameworks. In this work we describe strategies to improve a HOG based object detection pipeline by introducing viewpoint based mixture components. We learn accurate mixtures of object detectors for RGBD data using the latent SVM framework. Our contributions are threefold.

- To better exploit depth information, we develop a novel depth-based dense feature extraction method that provides a robust statistical description of scene geometry.
- We use publicly available 3D mesh models to learn strongly supervised viewpoint classifiers. These are used to guide the first stages of model learning, and help avoid inaccurate local minima of latent SVM training.
- We develop a geometric dataset augmentation scheme that uses scene geometry to 'take another look' at the training data, simulating the effect of camera viewpoint changes.

We evaluate our learned detectors on the NYU dataset, and demonstrate that each of our advances results in systematic performance improvements over the traditional HOG-based detection pipeline.

Viewpoint Mixture Model Our object detection framework consists of a mixture model where each mixture component corresponds to a viewpoint. Formally, we denote the parameters of our mixture model by $\mathbf{w} = \{\mathbf{w}_1, \cdots, \mathbf{w}_C\}$, where C is the total number of mixture components. Given a bounding box $\mathbf{x}$ of an RGBD image, as well as a component ID $c \in \{1, \cdots, C\}$, the model provides a score, which is equal to the dot product of its parameters $\mathbf{w}_c$ and the feature vector of the bounding box $\Psi(\mathbf{x})$. The feature vector $\Psi(\mathbf{x})$ consists of two types of features: (i) the HOG (histogram of gradient) features, denoted by $\phi_g(\mathbf{x})$; (ii) the depth features denoted by $\phi_d(\mathbf{x})$, which are a combination of HOD (histograms of depth) features and our novel displacement features. In other words, the score $s_{\mathbf{w}}(\mathbf{x}, c)$ for a bounding box $\mathbf{x}$, and component ID c is given by

$$s_{\mathbf{w}}(\mathbf{x}, c) = \mathbf{w}_c^\top \Psi(\mathbf{x}),$$
$$\Psi(\mathbf{x}) = \left[\phi_g(\mathbf{x}); \phi_d(\mathbf{x})\right] \quad (1)$$

Displacement Features Displacement features are local, depth-based descriptors that are computed over dense grid. Given a region/bounding box $\mathbf{x}$ in a depth image, and a cellsize s, we first subsample $\mathbf{x}$ using average pooling. Each $s \times s$ non-overlapping cell in $\mathbf{x}$ is replaced by the average depth of the pixels in the cell. In the subsampled image, we compute the displacement of each pixel p_i from the center pixel p_0. This displacement is given by, $\delta_i = depth^{p_0} - depth^{p_i}$. We quantize δ_i into a set of N displacement bins, using a hard quantization function: $q(\delta) : \mathbb{R} \rightarrow \mathbb{R}^N$. Thus, each cell in $\mathbf{x}$ is represented by a sparse indicator vector of size N. We concatenate these sparse vectors to get the displacement feature of $\mathbf{x}$. The displacement features capture the depth variations in a region with respect to the center, and give substantial gains in performance when used alongside HOG and HOD features.

Learning A Viewpoint Mixture Model Given a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \cdots, N\}$, we wish to estimate the parameters of our mixture model such that it provides accurate detections for previously unseen test images. Here $\mathbf{x}_i$ is a training sample (specifically, a bounding box of an RGBD image) and $y_i \in \{+1, -1\}$ indicates whether the sample $\mathbf{x}_i$ belongs to the object class or the background class. We denote the indices of all object samples by $\mathcal{O}$ and the indices of all background samples as $\mathcal{B}$. In the absence of explicit ground truth viewpoint annotations, we treat these viewpoints as latent variables. We employ the following latent support vector machine (latent SVM) formulation:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C_O \sum_{i \in \mathcal{O}} \xi_i + C_B \sum_{j \in \mathcal{B}} \xi_j, \\ \text{s.t.} \quad & \max_c \mathbf{w}_c^\top \Psi(\mathbf{x}_i) \geq 1 - \xi_i, \xi_i \geq 0, \forall i \in \mathcal{O}, \\ & \max_c \mathbf{w}_c^\top \Psi(\mathbf{x}_j) \leq -1 + \xi_j, \xi_j \geq 0, \forall j \in \mathcal{B}. \end{aligned} \quad (2)$$

The above optimization problem is a difference-of-convex program whose local minimum can be obtained using the CCCP algorithm. However the CCCP algorithm is sensitive to its initialization and the initial values of the component IDs can greatly affect its performance in practice. To overcome this difficulty, we exploit publicly available 3D models to better initialize the viewpoints.

Component Initialization with 3D Models The aspect ratio of an object is a reasonable- yet not an invincible- cue to what its viewpoint is. It can help distinguish the front and side views of a car, however, it cannot tell the front and rear views apart. In this work, we use a combination of aspect ratio cues, and viewpoint classifiers learnt from publicly available 3D models to estimate viewpoints of objects. We develop a two step strategy to initialize the component IDs for the object samples $\mathcal{O}$. First, we use 3D models of object categories from the Google Warehouse to render synthetic images with known viewpoints. The synthetic samples are used to learn a viewpoint classifier. Next the cues obtained by the viewpoint classifier are combined with the aspect ratio information, in an energy minimization framework, to obtain the component IDs of the object samples.

Dataset Augmentation with Geometric Jittering Camera viewpoint variation is a major challenge in object recognition; camera rotations affect object appearance radically and mixtures of viewpoint-tuned classifiers are imperative for multi-view detection. Here we exploit DIBR (Depth-Image-Based-Rendering) methods to simulate the effects of small camera rotations around the object, and obtain new samples that see the object from novel viewpoints. We use these synthesized images from novel views to augment our training set and enhance its variability.

Experimental Results

Method	Bed	Chair	M.+TV	Sofa	Table	Avg.
rgbd-hog	0.4660	0.2773	0.2480	0.2295	0.1430	0.2628
rgbd-hog + disp	0.5178	0.2771	0.2591	0.3440	0.1683	0.3133
rgbd-hog + disp + aug	0.5406	0.2919	0.2583	0.3470	0.1653	0.3206
rgbd-hog + disp + aug + vp	**0.5675**	**0.3023**	**0.3093**	**0.3719**	**0.1957**	**0.3493**
improvement over rgbd-hog	0.1015	0.0250	0.0613	0.1424	0.0527	0.0766

The table reports Average Precision values for the object detection task on 5 categories in the NYU dataset. The baseline method uses rgbd-hog (HOG+ HOD) features. We get systematic improvements in detection performance as we introduce displacement features (+disp), augmented data (+aug), and viewpoint initialization (+vp) to the baseline detector. Compared to the *rgbd-hog* baseline, we show an average improvement of 7.7%. While the displacement features allow us to better capture the depth variations of the surface, the augmentation and viewpoint initialization allow us to better model viewpoint effects.

Future Work The number of viewpoints dictates the number of components in our mixture model, and thus controls the model complexity. In the future we would like to develop models where we can express the viewpoint/orientation of an object in continuous space, while at the same time keeping a check on the model complexity. We would also like to extend our objective to solve the tasks of object detection and pose estimation simultaneously.

Context Forest for Object Class Detection

Davide Modolo
davide.modolo@gmail.com
Alexander Vezhnevets
vezhnick@gmail.com
Vittorio Ferrari
vittoferrari@gmail.com

University of Edinburgh
Edinburgh, UK

Global image appearance carries information about properties of objects in the image. For instance, a picture of a highway taken from a car is more likely to contain cars from the back viewpoint than from the side (fig. 1). This shows how the global image appearance of images can help understanding what objects are present and what they look like. Moreover, another property that can be inferred from global image appearance is the rough location of object instances [7]. For instance, an urban scene with cars parked in front of a building, shows cars in the bottom half of the image.

In this paper we exploit this observation for object class detection. We propose *Context Forest (ConF)*: a technique for learning the relation between the global appearance of an image and the properties of the objects it contains. Given only the global appearance of a test image, ConF retrieves a subset of training images that contain objects with similar properties. ConF is based on the Random Forest [2] framework, which provides high computational efficiency and the ability to learn complex, non-linear relations between global image appearance and objects properties. It is very flexible and only requires these properties to be defined through a distance function between two object instances, e.g. their appearance similarity or difference in location. We demonstrate ConF by learning to predict three properties: aspects of appearance, location in the image, and class membership.

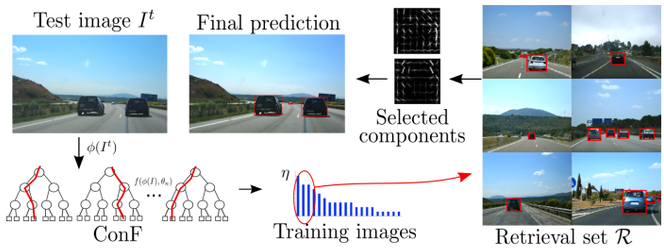


Figure 1: Illustration of ConF selecting components for a test image.

Aspects of appearance. Multi-component detectors [3, 5] model an object class as a mixture of components, each trained to recognize a different aspect of appearance. For example, different viewpoints (e.g. front and back view of a car) or articulation states (e.g. a person sitting vs standing). When trained on a large set, such a detector has many components [9], which all need to be evaluated on a test image, making it slow. Instead, we use ConF to select a subset of model components which is most relevant to a particular test image. We then run only those components, obtaining a speed-up.

We present experiments on two detectors: DPM [3] and EE-SVM [5] on a large 2-class dataset we call *BigCH*. This combines 6 existing datasets and, in total, the dataset has 15766 images containing 28548 car instances and 10107 images containing 13071 horse instances. We compare to building retrieval sets by kNN, and to a baseline which randomly selects components without looking at the test image. ConF outperforms the baseline and kNN for both object classes, for both detectors, and over the whole range of experiments. By employing ConF, we closely match the performance of the full DPM model by running roughly half of the components. We match the performance of a full EE-SVM when running less than 10% of the components. Even in the extreme case of running just *one* EE-SVM component, the AP is about 90% of that of the full model. Interestingly, for EE-SVM on the horse class, ConF *improves AP by 3%* over the full ensemble using all components, when running *10× fewer components*.

Location. At test time, a typical detector scores windows in a test image, based on their appearance only. We propose here to augment the detector's scores by adding knowledge about likely positions and scales



Figure 2: Detection obtained before (top row) and after (bottom row) applying ConF as location model. Green bounding-boxes highlight correct detections, while red ones show false positives.

of the object class, derived purely from the global appearance of the image. We train a second ConF to predict at which positions and scales objects are likely to appear in a given test image, analogue to [4, 7, 8]. By incorporating this information in the detector score at test time, we reduce the false positive rate by removing detections at unlikely locations.

We experiment with DPM and EE-SVM on BigCH, as for the aspect of appearance property. Results show that ConF improves AP for both classes and both detectors (+2% for cars and +1% for horses). Instead, kNN does not bring any improvement.

Class membership. In multi-class problems, a typical system would run detectors for all classes on all test images [6]. Instead, here we use ConF to predict what classes are present in each image, and run only the corresponding detectors. This greatly reduces the number of detectors run, and removes some false-positives.

We experiment with EE-SVM on the ILSVRC2014 dataset [1], which has a large number of classes (200). Without any context, EE-SVM achieves an mAP of 16.3%. Selecting classes based on kNN retrieval sets improves performance by +3.3% (mAP 19.6%), while ConF delivers a larger improvement of +4.8% (mAP 21.1%). The improvements are due to removing false positives produced by detectors of classes unlikely to be present in the image. Interestingly, ConF selects less than 10 classes per image on average, and therefore runs 20× fewer EE-SVM detectors than the context-free baseline.

To conclude, all these experiments demonstrate that ConF is a general technique that can predict various kinds of object properties. An extensive comparison to standard nearest-neighbour techniques for such context-based predictions also shows that ConF predicts object properties from global image appearance more accurately, while being 60× more memory efficient and much faster: kNN requires a number of distances computations linear in the number of training images, whereas ConF requires only a logarithmic number of thresholding operations.

[1] Imagenet large scale visual recognition challenge (ILSVRC). http://www.image-net.org/challenges/LSVRC/2014/index, 2014.

[2] L. Breiman. Random forests. *ML Journal*, 45(1):5–32, 2001.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on PAMI*, 32(9), 2010.

[4] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[5] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. doi: 10.1007/s11263-015-0816-y.

[7] B. Russell, A. Torralba, C. Liu, R. Ferugs, and W. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[8] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):153–167, 2003.

[9] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012.

Spatiotemporal Deformable Prototypes for Motion Anomaly Detection

Robert Bensch
bensch@cs.uni-freiburg.de
Thomas Brox
brox@cs.uni-freiburg.de
Olaf Ronneberger
ronneber@cs.uni-freiburg.de

Department of Computer Science and
BIOSS Centre for Biological Signalling Studies
University of Freiburg
Germany
http://lmb.informatik.uni-freiburg.de

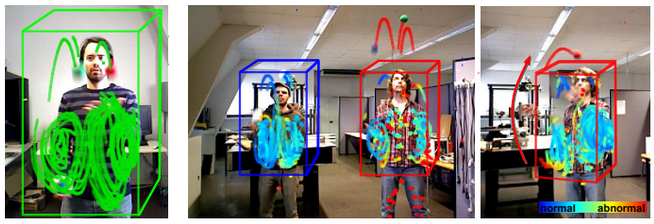
Overview This paper presents an approach for motion-based anomaly detection from video in a new setting. The task is formulated as detecting instances of a well-defined motion pattern prototype, and localizing subtle anomalies in its context, see Fig. 1. To this end, the prototype pattern is detected and elastically registered against a test sample. This is quite different from the traditional anomaly detection setting for surveillance scenarios (*e.g.* [2]) where the absolute position is a valuable and predominant feature. While our setting is most related to reconstruction based approaches (*e.g.* [1]), it is different in that reconstruction is performed in a direct manner by registration. "Supertrajectories" based on hierarchical clustering of dense point trajectories serve as an efficient and robust representation of motion patterns. A statistical prototype model is learned from training sequences to define accepted variations in terms of spatiotemporal deformations and remaining deviations. We start with a robust detection, that allows for temporal shift and 3D spatial translation and rotation. An efficient hashing approach provides transformation hypotheses, that are refined by a spatiotemporal elastic registration of the deformable prototype. We propose a new method for elastic registration of 3D+time trajectory patterns that induces spatial elasticity from trajectory affinities. Finally, an unseen motion pattern is reconstructed by prototype placements, while subpatterns that remain poorly reconstructed are detected as abnormal. Compared to existing methods, that usually deal with 2D+time data, our method copes with 3D+time data. Fig. 1 illustrates our setting of anomaly detection with an example of juggling patterns. The method is evaluated on a new motion anomaly dataset (Figs. 1-2) and performs well in detecting subtle anomalies (Fig. 3). Moreover, we demonstrate the applicability to biological motion patterns (Fig. 4).

Results Our anomaly detection approach is basically different from most existing approaches. Accordingly, we found existing benchmarks to be inappropriate for demonstrating our method. We recorded a new motion anomaly dataset from persons juggling balls using a Kinect camera and generated 3D motion trajectories (Fig. 2). We compare against chaotic invariants (CI) for anomaly detection in crowded scenes [3]. Among the methods with code available, CI is the most related one as it provides invariance to position and magnitude via a bag of features approach. We show that our approach outperforms CI to its accurate modeling of the normal pattern variation (Fig. 3). Especially in the case of viewpoint variations compared to the training data, our approach clearly outperforms CI, see Fig. 3(a). Our method profits from its strong generalization capability due to spatial normalization by detection and elastic registration in 3D+time. Moreover, we demonstrate the relevance and general applicability in experiments on biological motion patterns (Fig. 4). We were able to quantify a significant difference between wild type (WT) and genetically modified embryos (MO) from global endodermal cell dynamics in zebrafish. In addition, we were able to find evidence for our hypothesis, that a time scaling might partially explain these differences. We will make the new motion anomaly dataset and the code publicly available.

[1] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3449–3456. IEEE, 2011.

[2] V. Mahadevan, Weixin Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1975–1981, June 2010.

[3] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2054–2060, 2010.

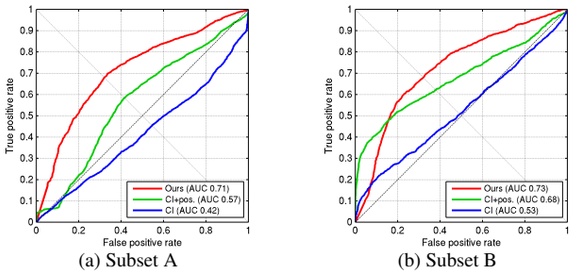


(a) (b) (c)

Figure 1: Motion anomaly detection in juggling patterns. (a) Motion pattern prototype. (b-c) Detection of local anomalies in context of prototype detections (*bounding boxes*). The anomaly score is plotted for supertrajectories from normal (*blue*) to abnormal (*red*).

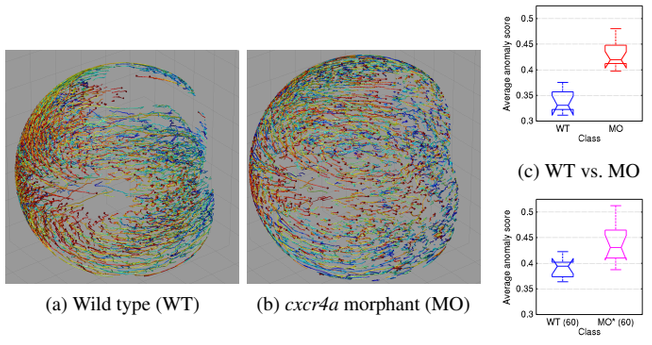


Figure 2: Juggling pattern test set. Examples of the same juggling pattern show possible variations due to different jugglers and viewpoints. Supertrajectories are plotted over a range of 45 frames (1.5sec), the color indicates depth obtained from Kinect camera.



(a) Subset A (different viewpoints) (b) Subset B (similar viewpoints)

Figure 3: Anomaly detection ROC curves. Comparison of our method (*red*) and the chaotic invariants (CI) [3] (*green and blue*) for the two subsets A and B. The average performance of all methods is shown.



(a) Wild type (WT) (b) *cxcr4a* morphant (MO) (c) WT vs. MO (d) WT vs. MO*

Figure 4: Anomaly detection in biological motion patterns of cell dynamics in the early development of zebrafish embryos. (a-b) Example motion patterns for each group. Trajectories are rendered in 3D with time colour-coded from blue to red. (c-d) Boxplots of global anomaly scores. (c) Significant difference between WT and MO patterns. (d) Time scaling of MO patterns (MO*) partially compensates the differences.



BMVC 2015 Workshop on Machine Vision of Animals and their Behaviour (MVAB 2015) September 10, 2015, Swansea, UK

The Machine Vision of Animals and their Behaviour (MVAB) workshop aims to bring together members of the community researching computer vision for animals, from such diverse application areas as wildlife study, animal farming, and post-mortem inspection.

We are pleased to present a programme that, while having an aquatic ring to it, with papers on reef fish, white sharks, crustaceans, and fish captured for consumption, is varied enough to also include dairy cows, chicken, flying birds, butterflies, and even the bottom of the crested black macaque. Applications range from work more or less directly related with food production to the study of animals in their natural habitat. Echoing this diversity of application areas, the workshop's oral sessions will start with two stimulating keynote speeches from very different perspectives: Prof. Ilias Kyriazakis will address aspects of livestock production that can benefit from the use of computer vision, whereas Prof. Robert Fisher will discuss the wide range of applications of the Fish4Knowledge database.

From the onset we thought that all accepted papers should be published, in order to make the call for papers as appealing as possible. We struck a balance between the number of submissions and the duration of the workshop to decide on accepted oral and poster presentations that will encourage interesting exchanges during the workshop. Each paper was reviewed by three programme committee members in a double-blind process and, ultimately, we accepted all papers that had no recommendation for rejection from any reviewer. This resulted in 9 accepted manuscripts out of 15 submissions, or a 60% acceptance rate. In order to leave some generous time for discussion after each oral presentation, while keeping the talks themselves reasonable in length, we chose to have 6 oral presentations and dedicate a poster session to the remaining 3 papers.

We would like to express our sincere thanks to the BMVC organisers for their prompt help with all our administrative queries; to Prof. Kyriazakis and Prof. Fisher (our co-chair) for their availability to contribute with keynote speeches; to all authors who have submitted their work, making the workshop possible; and of course to all members of our programme committee for their hard work reviewing, which enabled us to come up with a balanced, high-quality programme.

7th August 2015
The MVAB chairs

Further information is available at: https://openlab.ncl.ac.uk/mvab2015/

Invited Talks:
Ilias Kyriazakis (Newcastle University, UK): What can machine vision do for livestock?
Robert Fisher (University of Edinburgh, UK): Fish detection, tracking, recognition, and analysis with the Fish4Knowledge dataset

Workshop Organizers:
- Telmo Amaral (Newcastle University, UK)
- Stephen Matthews (Newcastle University, UK)
- Thomas Plötz (Newcastle University, UK)
- Stephen McKenna (University of Dundee, UK)
- Robert Fisher (University of Edinburgh, UK)

Program Committee:
Alexandra Branzan Albu (University of Victoria, Canada)
Douglas Armstrong (University of Edinburgh, UK)
Vinay Bettadapura (Georgia Institute of Technology, US)
Elsbeth van Dam (Noldus Information Technology, Netherlands)
Patrick Dickinson (University of Lincoln, UK)
Alfonso Pérez-Escudero (Massachusetts Institute of Technology, US)
Yasuyo Kita (National Institute of Advanced Industrial Science and Technology, Japan)
Anders Ringgaard Kristensen (University of Copenhagen, Denmark)
Eric Pauwels (Centrum Wiskunde & Informatica, Netherlands)
Sai Ravela (Massachusetts Institute of Technology, US)
Georgios Tzimiropoulos (Nottingham University, UK)

Machine Vision of Animals and their Behaviour (MVAB 2015)

Workshop Program

13:15 - 13:30 Setup
13:30 - 13:35 Welcome
13:35 - 14:20 Keynote by Ilias Kyriazakis
What can machine vision do for livestock? *Ilias Kyriazakis, Newcastle University*

14:20 - 15:00 Livestock
Non-intrusive automated measurement of dairy cow body condition using 3D video*, Mark Hansen (University of the West of England), Melvyn Smith (University of the West of England), Lyndon Smith (University of the West of England) and Duncan Forbes (Kingshay Conservation and Farming Ltd.)*

Detecting Gallbladders in Chicken Livers using Spectral Imaging*, Anders Jørgensen (Aalborg University), Eigil Mølvig Jensen (IHFood) and Thomas B. Moeslund (Aalborg University)*

15:00 - 15:50 Coffee break and Poster Session
A Computer Vision Approach to Classification of Birds in Flight from Video Sequences*, John Atanbori (University of Lincoln), John Murray (University of Lincoln), Wenting*

Duan (University of Lincoln), Kofi Appiah (Nottingham Trent University) and Patrick Dickinson (University of Lincoln)

Microfluidic environment and tracking analysis for the observation of Artemia Franciscana, *Jose Alonso Solis-Lemus (City University London), Yushi Huang (RMIT University), Donald Wlodkowic (RMIT University) and Constantino Carlos Reyes-Aldasoro (City University London)*

Invariant Image-Based Species Classification of Butterflies and Reef Fish, *Hafeez Anwar (Vienna University of Technology), Sebastian Zambanini (Vienna University of Technology) and Martin Kampel (Vienna University of Technology)*

15:50 - 16:35 Keynote by Robert Fisher

Fish detection, tracking, recognition, and analysis with the Fish4Knowledge dataset
Robert Fisher, University of Edinburgh

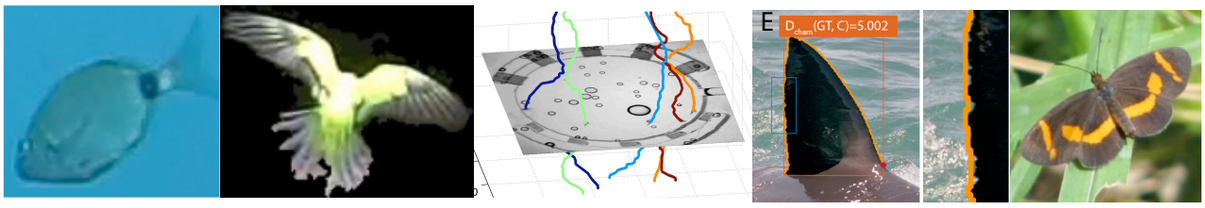
16:35 - 18:00 Fish and Macaque

Croatian Fish Dataset: Fine-grained classification of fish species in their natural habitat, *Jonas Jäger (Hochschule Fulda), Marcel Simon (Friedrich Schiller University Jena), Joachim Denzler (Friedrich Schiller University Jena), Viviane Wolff (Hochschule Fulda), Klaus Fricke-Neuderth (Hochschule Fulda) and Claudia Kruschel (University of Zadar)*

Convolutional Neural Networks for Counting Fish in Fisheries Surveillance Video, *Geoffrey French(University of East Anglia), Mark Fisher (University of East Anglia), Michal Mackiewicz (University of East Anglia) and Coby Needle (Marine Scotland)*

Affinity Matting for Pixel-accurate Fin Shape Recovery from Great White Shark Imagery
Benjamin Hughes (University of Bristol) and Tilo Burghardt (University of Bristol)

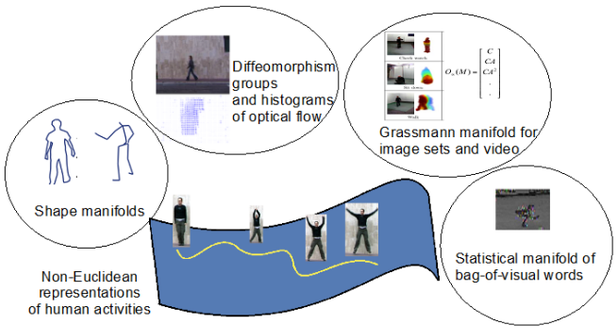
Detecting and Tracking Bottoms and Faces of the Crested Black Macaque in the Wild
John Chiverton (University of Portsmouth), Jerome Micheletta (University of Portsmouth) and Bridget Waller (University of Portsmouth)



1st International Workshop on DIFFerential Geometry in Computer Vision for Analysis of Shapes, Images and Trajectories

In conjunction with BMVC 2015
Swansea, September 10, 2015





Workshop Organizers

Hassen Drira, Institut Mines-Telecom/TELECOM Lille, France; Sebastian Kurtek, The Ohio State University, USA; Pavan Turaga, Arizona State University, USA

Workshop Description

Riemannian geometric computing has received a lot of recent interest in the computer vision community. In particular, Riemannian geometric principles can be applied to a variety of difficult computer vision problems including face recognition, activity recognition, object detection, biomedical image analysis, and structure-from-motion to name a few. Besides their nice mathematical formulation, Riemannian computations based on the geometry of underlying manifolds are often faster and more stable than their classical counterparts. Over the past few years, the popularity of Riemannian algorithms has increased several-fold. Some of the mathematical entities that benefit from a geometric analysis include rotation matrices, medial representations, subspace comparisons, symmetric positive-definite matrices, function-spaces, and many more.

Workshop Program

13:15 - 13:30 Setup

13:30 - 13:35 Opening Remarks by DIFF-CV Organizers

13:35 - 14:35 Keynote by Anuj Srivastava

Some Thoughts on the Role of Geometry in Computer Vision, *Anuj Srivastava, Florida State University*

14:35 - 15:00 Poster Session

Zero-Shot Domain Adaptation via Kernel Regression on the Grassmannian, *Yongxin Yang, Timothy Hospedales (Queen Mary, University of London)*

Karcher Mean in Elastic Shape Analysis, *Wen Huang (Universite Catholique de Louvain), Yaqing You (Florida State University), Pierre-Antoine Absil (Universite Catholique de Louvain), Kyle Gallivan (Florida State University)*

A Novel Riemannian Framework for Shape Analysis of Annotated Surfaces, *Jiaqi Zaetz, Sebastian Kurtek (The Ohio State University)*

A Generalized Lyapunov Feature for Dynamical Systems on Riemannian Manifolds, *Rushil Anirudh, Vinay Venkataraman, Pavan Turaga (Arizona State University)*

Improving 3D Facial Action Unit Detection with Intrinsic Normalization, *Eric Yudin, Aaron Wetzler, Matan Sela, Ron Kimmel (Technion)*

Gauge Invariant Framework for Trajectories Analysis, *Hassen Drira, Barbara Tumpach, Mohamed Daoudi (University Lille1)*

15:00 - 15:30 Coffee break

15:30 – 16:00 Poster Session Continued

16:00 - 18:00 Oral Session

16:00-16:20 Distance Metric Learning by Optimization on the Stiefel Manifold, *Ankita Shukla, Saket Anand (IIIT- Delhi)*

16:20-16:40 Geometric Analysis of Axonal Tree Structures, *Adam Duncan, Anuj Srivastava (Florida State University), Xavier Descombes (INRIA, Sophia Antipolis), Eric Klassen (Florida State University)*

16:40-17:00 Second Order Elastic Metrics on the Shape Space of Curves, *Martin Bauer (University of Vienna), Martins Bruveris, Philipp Harms (ETH Zurich), Jakob Møller-Andersen (Technical University of Denmark)*

17:00-17:20 Fast Alignment of 2D Closed Curves with Application to Elastic Shape Analysis, *Gunay Dogan (Theiss Research), Javier Bernal, Charles Hagwood (National Institute of Standards and Technology)*

17:20-17:40 Temporal Reflection Symmetry of Human Actions: A Riemannian Analysis, *Qiao Wang, Rushil Anirudh, Pavan Turaga (Arizona State University)*

17:40-18:00 Cylindrical Surface Reconstruction by Fitting Paths on Shape Space, *Chafik Samir (University of Clermont), Pierre-Yves Gousenbourger (Universite Catholique de Louvain), Shantanu Joshi (University of California, Los Angeles)*

18:00 Adjourn



BMVC 2015 Workshop on Computer Vision Problems in Plant Phenotyping (CVPPP 2015) September 10, 2015, Swansea, UK

Thursday PM 13:30-18:00

The goal of this second workshop, following on from the first successful CVPPP at ECCV 2014, is to continue to showcase the challenges raised by and extend the state of the art in computer vision for plant phenotyping. Plant phenotyping is the identification of effects on plant structure and function (the phenotype) resulting from genotypic differences (i.e., differences in the genetic code) and the environmental conditions a plant has been exposed to. Knowledge of plant phenotypes is a key ingredient of the knowledge-based bioeconomy, which not only literally helps to feed the world, but is also essential for feed, fibre, and fuel production.

While collection of phenotypic traits was previously manual, non-invasive, image-based methods are now increasingly utilized in plant phenotyping and the resulting images need to be analysed in a high throughput, robust, accurate, and reliable manner. The problems raised differ from the usual tasks addressed by the computer vision community, due to the requirements posed by this challenging application scenario.

Plants are complex, self-changing systems whose complexity increases over time. Typical phenotyping problems include measuring the size, shape, 3D surface structure, architecture, and other structural traits of plants and their organs (leaves, fruit, roots etc.). Many scenarios require quantitative description of plant populations, where core problems include reliable detection and multi-label segmentation of many similar objects, or the reconstruction of specular, almost featureless, and overlapping surfaces. Quantitative description of the growth of these complex, deforming objects is vital, and requires suitable tracking, optical flow and/or scene flow estimation methods. Inherently, the tracked objects change their appearance over time. In some cases images may be acquired under controlled conditions, but they are increasingly likely to be taken in more challenging natural environments like greenhouses, or in the field. Automated image acquisition protocols are highly desirable, generating large numbers of images.

Without automated and accurate computer vision to extract the phenotypes, a bottleneck is formed, hampering our understanding of plant biology. We thus want to identify key but unsolved problems, expose the current state-of-the-art, and broaden the field and the community.

Target audience are

- computer vision experts interested in novel application fields, well accessible to computer vision, but different in requirements
- plant phenotyping scientists with rich expertise in image processing and computer vision interested in standardization, as exact problem formulations in fact allow defining standards.

Further information is available at: http://www.plant-phenotyping.org/CVPPP2015

Leaf Segmentation and Counting Challenges

Associated with the workshop are two computer vision challenges:

- the 2nd "Plant Leaf Segmentation Challenge" and
- the Leaf Counting Challenge.

For the challenges we released training sets (containing raw images and annotations) and testing sets (containing raw images, only). Papers were evaluated and ranked according to their outcome, the validity of the algorithm, and suitability of the approach. Only fully automated approaches were requested and accepted. See http://www.plant-phenotyping.org/CVPPP2015-challenge

Invited Talk:
Tim Cootes: Image Segmentation using Statistical Shape and Appearance Models

Workshop Organizers:

- Sotirios A. Tsaftaris (IMT Lucca, Italy and Northwestern University, USA), S.Tsaftaris@imtlucca.it
- Hanno Scharr (Forschungszentrum Jülich, Germany), H.Scharr@fz-juelich.de
- Tony Pridmore (University of Nottingham, UK), Tony.Pridmore@nottingham.ac.uk

Program Committee:
Andrew French, University of Nottingham, UK
Arijit Biswas, Xerox Research Center India
Christian Klukas, LemnaTec, Germany
Tim Brown, Australian National University, Australia
Babette Dellen, Hochschule Koblenz, Germany
Guillerme Desouza, University of Missouri-Columbia
Jurgen Fripp, CSIRO, Australia
Pablo M. Granitto, CIFASIS, Argentina
Markus Hoeferlin, Robert Bosch Start-up, Germany
Gustavo Pereyra Irujo, CONICET, Argentina
Toni Kazic, University of Missouri, USA
Gert Kootstra, Wageningen, The Netherlands
Xiaoming Liu, University of Missouri, USA
Guillaume Lobet, Université de Liège, Belgium
B. S. Manjunath, UCSB, Santa Barbara, USA
Vasileios Mezaris, CERTH, Thessaloniki, Greece
Massimo Minervini, IMT Lucca, Italy
Marie Neal, Aberystwyth University, UK
Gerrit Polder, Wageningen, The Netherlands
David Rousseau, Univ. of Lyon, France
Kyle Simek, University of Arizona, USA
Concetto Spampinato, University of Catania, Italy
Junsong Yuan, NTU, Singapore
Rick van de Zedde, Wageningen, The Netherlands

Computer Vision Problems in Plant Phenotyping (CVPPP 2015)

Workshop Program

13:15 - 13:30 Setup
13:30 - 13:40 Opening Remarks
13:40 - 14:20 Keynote by Tim Cootes

Image Segmentation using Statistical Shape and Appearance Models, *Tim Cootes, University of Manchester*

14:20 - 15:00 Counting Leaves

Results from the LCC Challenge:

Learning to Count Leaves in Rosette Plants, *Mario Valerio Giuffrida, Massimo Minervini, Sotirios Tsaftaris (IMT Lucca)*

Counting leaves without ``finger-counting'' by supervised multiscale frequency analysis of depth images from top view, *David Rousseau (Université Lyon 1), Henricus J. Van de Zedde (Wageningen)*

15:00 - 15:30 Coffee break
15:30 - 17:10 Segmentation, Annotation, and Reconstruction

Results from the LSC Challenge:

Utilizing machine learning approaches to improve prediction of leaf counts and individual leaf segmentation of rosette plants, *Jean-Michel Pape (IPK-Gatersleben), Christian Klukas (LemnaTec GmbH)*

Gaussian process shape models for Bayesian leaf segmentation, *Kyle Simek, Kobus Barnard (University of Arizona)*

Fuzzy c-means based plant segmentation with distance dependent threshold, *Mads Dyrmann (University of Southern Denmark)*

An interactive tool for semi-automated leaf annotation, *Massimo Minervini, Mario Valerio Giuffrida, Sotirios Tsaftaris (IMT Lucca)*

3D Surface Reconstruction of Plant Seeds by Volume Carving, *Johanna Roussel, Andreas Fischbach, Siegfried Jahnke, Hanno Scharr (Forschungszentrum Jülich)*

17:10 - 17:20 Closing Remarks
17:20 Adjourn

7th UK Computer Vision Student Workshop (BMVW 2015) September 10, 2015, Swansea, UK

The British Machine Vision Student Workshop (BMVW) takes place in the afternoon of the 10th September 2015. The workshop has become a regular feature of British Machine Vision Conference. It provides an excellent opportunity for postgraduate students in computer vision to present their on-going research, to network and to start collaborations with other students. Furthermore, students have the opportunity to attend keynote sessions about hot topics in the field presented by experienced researcher. The workshop is a single track containing oral presentations. All accepted papers will be digitally published on-line at the BMVA website.

Topics

As with the main BMVC conference topics include, but are not limited to:
Statistics and machine learning for vision; Stereo, calibration, geometric modelling and processing; Face and gesture recognition; Early and biologically inspired vision; Motion, flow and tracking; Segmentation and grouping; Model-based vision; Image processing techniques and methods; Texture, shape and colour; Video analysis; Document processing and recognition; Vision for quality assurance, medical diagnosis, etc.; Vision for visualization, interaction, and graphics; Object detection and recognition; Shape-from-X; Video analysis and event recognition; Illumination and reflectance

Traditionally, BMVW workshop only accepted student submissions from British institutions. This year we also encourage student submissions from institutions outside UK. Priority is given to student participants from UK institutions, however.

Further information is available at: http://bmvc2015.swan.ac.uk/?p=2227

Keynote Speaker

Prof. Mark Nixon (University of Southampton)

Title: Frontiers of Identification: Biometrics and Computer Vision

Abstract: Biometrics has a shorter history than computer vision though it has come to affect one of the largest populations in the world through India's Unique Identification project. Much of the progress is consistent with advances in computer vision, since many approaches to biometrics use computer vision techniques. This talk will survey the progress in biometrics and the techniques in computer vision that have been used to support that progress. The talk will naturally focus on areas of my particular expertise, namely feature extraction in computer vision and gait, ear and soft biometrics. Of these topics, computer vision implements the approaches, especially motion-based approaches. Gait concerns recognising people by the way they walk, thus involving moving feature analysis; ears are static shapes which though rather ugly appear to retain their shape throughout a person's life. The newer soft biometrics concern using human descriptions (attributes) for recognition, which when appropriately formulated can provide new means for recognition in surveillance footage. This is a form of automated witness description for recognition, using attributes which are at the forefront of modern computer vision approaches.

Workshop Organizers:
Workshop chair: Dr. Gary KL Tam (On behalf of BMVC 2015 Chairs)
Workshop assistants: Mike Edward, Robert Palmer

Program Committee:

Ashish Gupta	Ohio State University
Dengxin Dai	ETH Zurich
David Fouhey	CMU
David Ferstl	Graz University of Technology
Erik Rodner	Friedrich Schiller University of Jena
Guillaume Bourmaud	University of Bordeaux
Gernot Riegler	Graz University of Technology
Heng Yang	University of Cambridge
Hamid Izadinia	Washington University
Jan Knopp	KU Leuven
Jingjing Zheng	University of Maryland
Junhua Mao	UCLA
Karel Lebeda	University of Surrey
Kevin Shih	Univeristy of Illinois
Karan Sikka	UCSD
Mohamed Elhoseiny	Rutgers University
Ming Jiang	NUS
Mark Hansen	University West of England
Massimo Camplani	Univeristy of Bristol
Michael Sapienza	University of Oxford
Nazli FarajiDavar	University of Surrey
Osian Haines	Univeristy of Bristol
Pablo Alcantarilla	Georgia Institute of Technology
Ruixuan Wang	University of Dundee
Shenping Zhang	Harbin Institute
Simone Buoncompagni	University of Bologna
Varun Nagaraja	University of Maryland
Vinay Venkataraman	Arizona State University
Violet Snell	Southampton Solent University
Wenbin Li	University of Bath
Yin Li	Georgia Tech
Zhi Yang	University of Buffalo

Message from Chair
This year, there are 11 (6 UK and 5 Non-UK) submissions to BMVW2015. Non-UK submissions come from Africa (20%), North America (20%), and Asia (60%), according to the first author's affiliation. 5 papers were accepted for oral presentations. We also have record number of reviewers (32). Each paper was reviewed by exactly 4 reviewers. We gratefully thank all reviewers for providing punctual and insightful reviews for the very smooth running of BMVW2015. We are grateful to our keynote speaker Prof. Mark Nixon (University of Southampton) who is an expert in Biometrics. We are also delighted to have Dr Adeline Paiement (Bristol University) to give an invited postdoc talk, covering not only technical research, but also her experience on PhD and Post-doc life.

7th UK Computer Vision Student Workshop (BMVW) Workshop Programme

13:15 - 13:30 Setup

13:30 - 13:40 Opening Remarks

13:40 - 14:20 Keynote: Prof. Mark Nixon
Frontiers of Identification: Biometrics and Computer Vision
Mark Nixon, University of Southampton

Oral Presentation (15 mins for talk, 5 mins for Q&A)
14:20 - 14:40
Generating Local Temporal Poses from Gestures with Aligned Cluster Analysis for Human Action Recognition
Mike Edwards, Xianghua Xie (Swansea University)
14:40 - 15:00
Advancements in Contact-free Heart Rate Measurements Using Human Face Videos
Muhammad Waqar, Bernard Tiddeman (Aberystwyth University)

15:00 - 15:30 Coffee break

15:30 - 16:00 Invited Post-doc Talk
Computer Vision for Health: Classical Applications and New Perspectives… and My Life as a Post-doc
Dr. Adeline Paiement (Bristol University)

16:00 - 16:20
Analysis of Face and Segment Level Descriptors for Robust 3D Co-Segmentation
David George, Gary Tam, Xianghua Xie (Swansea University)
16:20 - 16:40
Using 3D Representations of the Nasal Region for Improved Landmarking and Expression Robust Recognition
Jiangning Gao, Adrian Evans (University of Bath)
16:40 - 17:00
Evaluating the Resilience Face Recognition Systems against Malicious Attacks
Luma Omar, Ioannis Ivrissimtzis (Durham University)

17:00 - 17:10 Closing Remarks



Author Index

Afonso, Mariana 149
Ahmadi, Seyed-Ahmad 125
Aichert, André 86
Akashi, Takuya 124
Al-Amoudi, Asharf 21
Albarqouni, Shadi 21
Alexiou, Ioannis 150
Al-Hamadi, Ayoub 127
Alpert, Sharon 174
Amaral, Telmo 193
Amor, Heni Ben 84
Andrews, Walter 64
Angelova, Anelia 36
Arefiyan, Mostafa 22
Argyros, Antonis A. 126, 147, 185
Argyros, Antonis 81
Arnab, Anurag 44
Avgerinakis, Konstantinos 42
Avrithis, Yannis 17
Azizpour, Hossein 22
Babu, R. Venkatesh 35
Banerjee, Anjan 171
Barkan, Ella 174
Baró, Xavier 132
Barrios, Juan Manuel 167
Bartlett, Marian 75
Baskurt, Atilla 166
Baust, Maximilian 21
Bekaert, Philippe 20
Ben-Ari, Rami 137
Bensch, Robert 192
Bergevin, Robert 170
Bertalmío, Marcelo 129
Bertalmio, Marcelo 23
Beyerer, Jürgen 91
Bharath, Anil A. 150
Bieth, Marie 88
Bischof, Horst 62, 105, 185
Biswas, Arijit 14
Boia, Raluca 118
Boix, Xavier 19
Bosilj, Petra 172
Botelho, Silvia S. C. Costa 157
Bourmaud, Guillaume 43
Brachmann, Eric 184
Brahmbhatt, Samarth 84
Briasouli, Alexia 42
Brox, Thomas 46, 192
Buf, J.M.H. du 63
Buoncompagni, Simone 136
Burghardt, Tilo 95, 148
Burns, Randal 85
Busto, Pau Panareda 18
Byeon, Moonsub 72
Čadík, Martin 34
Čadik, Martin 49
Caesar, Holger 33
Cai, Jianfei 15
Cai, Ziyun 175
Camplani, Massimo 148
Carlone, Luca 139
Carlsson, Stefan 22
Cavallaro, Andrea 128
Cesar, Vinicius 138, 140
Chai, Xiujuan 39
Chandra, Siddhartha 190
Chang, Hong 168
Chang, Kuang-Yu 59
Chang, Ming-Ching 107
Charpiat, Guillaume 68
Chateau, Thierry 110
Chellappa, Rama 99, 111, 122
Chen, Ching-Hui 122
Chen, Chu-Song 59
Chen, Da 90
Chen, Lei 182
Chen, Xilin 39, 168
Chen, Yani 141
Chen, Yi-Lei 177
Chen, Zhaoyun 188
Cheng, Hong 107
Chevaldonné, Marc 110
Chevallier, Louis 27
Chevillet, Mark A 85
Chia, Alex YS 51
Chinellato, Eris 83
Chittka, Lars 128
Choi, Jin Young 72
Cholakkal, Hisham 162
Christensen, Henrik 84
Chrysos, Grigorios G. 190
Chum, Ondřej 34
Cipolla, Roberto 130
Ciuc, Mihai 178
Codevilla, Felipe 157
Cohen, Laurent D. 90
Cohn, Anthony G. 83
Conjeti, Sailesh 21
Cottrell, Garison 164
Crowley, Elliot J. 69
Cui, Zhaopeng 50
Cyriac, Praveen 23
Cyrille, Beaudry 121
Damen, Dima 148
Das, Sukhendu 134
Davis, Larry S. 57
Debnath, Subhabrata 171
Dellaert, Frank 139
Deng, Zhiwei 182
Dhall, Abhinav 38
Doermann, David 53
Đonlić, Matea 158
Donner, Rene 88
Dörfler, Arnd 86
Drayer, Benjamin 46
Drira, Hassen 196
Drummond, Tom 58
Duffner, Stefan 166
Einabadi, Farshad 47
Elgammal, Ahmed 97
Elhoseiny, Mohamed 97
Emberton, Simon 128
Erofeev, Mikhail 102
Escalante, Hugo Jair 132
Escalera, Sergio 32, 132
Eum, Sungmin 53

Facciolo, Gabriele 94
Fauvet, Eric 119
Fawzi, Alhussein 109
Fedorov, Alexey 102
Ferguson, Dave 36
Ferrari, Vittorio 31, 33, 191
Ferstl, David 105
Fetaya, Ethan 112
Fidler, Sanja 96
Filho, Nelson Duarte 157
Fisher, Robert 193
Fitzgibbon, Andrew 1
Florea, Corneliu 118
Florea, Laura 118
Foroughi, Homa 100
Forster, Deborah 164
Fowlkes, Charless 8, 26
Franchis, Carlo de 94
Frommer, Yuval 137
Frossard, Pascal 109
Fusiello, Andrea 24
Gaidon, Adrien 7, 114
Gall, Juergen 18, 61, 104
Gamage, Dinesh 58
Garnett, Noa 112
Gati, Ella 180
Gaya, Joel De O. 157
Gemert, Jan C. van 180
Gevers, Theo 80
Gherardi, Riccardo 130
Ghiasi, Golnaz 26
Giremus, Audrey 43
Giri, Ritwik 75
Gitman, Yury 102
Goecke, Roland 38
Goldstein, Thomas 152
Golodetz, Stuart 44
Gong, Shaogang 48
Gool, Luc Van 19
Gordo, Albert 114
Graham, Ben 153
Gräser, Lukas 145
Grau, Oliver 47
Grauman, Kristen 5
Gumhold, Stefan 184
Gunes, Hatice 133
Gupta, Raj Kumar 51
Gurevich, Lev 165
Hager, Gregory D 85
Hannuna, Sion 148
Hansard, Miles 93
Hao, Zongbo 54
Hashoul, Sharbell 174
He, Zheqi 159
Hernandez, Monica 89
Highlander, Tyler 163
Hobbs, Kevin 141
Hogg, David C. 83
Hoiem, Derek 131
Hornegger, Joachim 86, 146
Hospedales, Timothy 11
Hradiš, Michal 10, 34
Hsu, Chiou-Ting 177
Hu, Ninghang 80
Huang, Dafei 188
Huang, Kaiqi 70
Huang, Xiaohua 38
Huang, Yongzhen 151
Hughes, Benjamin 95
Hussein, Mohamed E. 98
Ilic, Slobodan 40, 125
Ismail, Mohamed A. 98
Izadi, Shahram 44
Izquierdo, Ebroul 54
Jacobs, David 14
Jain, Mihir 180
Jawahar, CV 117, 120, 173
Jayasumana, Sadeep 123
Jetley, Saumya 123
Ji, Hongkun 144
Ji, Zexuan 144
Jiang, Ming 19
Jiang, Nianjuan 50
Jiang, Xinwei 154
Johnson, Jubin 162
Jordan, Johannes 146
Joshi, Manjunath V. 108
Joshi, Padmaja 134
Jurie, Frederic 27, 142
Kambhamettu, Chandra 78, 115
Kanaeva, Ekaterina 165
Kane, David 23
Karanam, Srikrishna 77, 186
Kashino, Kunio 29
Kaynig-Fittkau, Verena 85
Kehl, Wadim 40, 125
Kelner, Judith 138, 140
Khatri, Nilay 108
Kijak, Ewa 172
Kim, Kikyung 72
Kim, Kyuwon 116
Kim, Seungryong 156
Kimmel, Ron 3, 37, 92
Kiryati, Nahum 137
Kisilev, Pavel 174
Kleissas, Dean M 85
Kneip, Laurent 103
Kodirov, Elyor 48
Köhler, Thomas 146
Kokkinos, Iasonas 190
Kolagunda, Abhishek 78
Kompatsiaris, Ioannis 42
Kong, Chen 96
Kotera, Jan 10
Krizhevsky, Alex 36
Kröger, Peer 52
Krull, Alexander 184
Kulkarni, Praveen 27
Kurtek, Sebastian 196
Kyriazis, Nikolaos 126, 147
Laligant, Olivier 119
Langs, Georg 88
Laurent, Mascarilla 121
Lechervy, Alexis 142
Lee, Guangqing 15
Lee, Hyungtae 53
Lefèvre, Sébastien 172
Lenc, Karel 9, 16
Leordeanu, Marius 189
Lepetit, Vincent 40, 136
Levi, Dan 112
Li, Feng 143
Li, Hongdong 103
Li, Luyuan 159

Li, Wei 101
Li, Xue 152
Li, Yang 77, 186
Liebelt, Joerg 18
Limberger, Frederico A. 60
Lin, Bo-Yao 59
Lin, Dahua 96
Liu, Jundong 141
Liu, Li 175, 176, 183
Liu, Pengcheng 70
Liu, Weifeng 32
Liu, Xiaoming 25
Liu, Xicheng 159
Liu, Yang 73
Liu, Yu 79
Liu, Yuhao 182
López-Sastre, Roberto 67
Lou, Zhongyu 80
Lu, Boyu 111
Lu, Guoyu 78
Lu, Hanqing 187
Lu, Linlin 54
Lu, Qinglin 119
Lu, Shijian 15
Lucassen, Marcel P. 80
Lui, Vincent 58
Luisier, Florian 64
Luo, Lei 188
Luo, Yong 161
Lyu, Siwei 107
Ma, Lei 154
Maddern, Will 74
Maggiori, Emmanuel 68
Magri, Luca 24
Maier, Andreas 86, 146
Maio, Dario 136
Makris, Alexandros 81
Mallya, Arun 131
Malmir, Mohsen 164
Marshall, Dave 145
Matas, Jiri 16
Matthews, Stephen 193
McKenna, Stephen 193
Meinhardt, Enric 94
Mensink, Thomas 181
Menze, Bjoern 88
Meshry, Moustafa 76
Michel, Damien 185
Michel, Frank 184
Michiels, Nick 20
Miksik, Ondrej 44
Milletari, Fausto 125
Mirebeau, Jean-Marie 90
Mirmehdi, Majid 148
Mishkin, Dmytro 16
Modolo, Davide 191
Moeslund, Thomas B. 32
Mohammed, Hussein Adnan 63
Monari, Eduardo 91
Morariu, Vlad I. 57
Morency, Louis-Philippe 170
Mori, Greg 41, 182
Mou, Wenxuan 133
Moujtahid, Salma 166
Movellan, Javier 164
Muralidharan, Srikanth 182
Murasaki, Kazuhiko 55
Nagaraja, Varun K. 57
Nagel, Markus 181
Nakano, Gaku 82
Namboodiri, Anoop M. 117
Namboodiri, Vinay P. 169
Namboodiri, Vinay 171
Nasrabadi, Nasser M. 111
Navab, Nassir 21, 40, 125
Negrel, Romain 142
Newman, Paul 74
Newson, Alasdair 106
Niedermayer, Johannes 52
Ogale, Abhijit 36
Oh, Songhwai 72
Okatani, Takayuki 55, 65
Olsen, Søren Ingvor 32
Ophir, Boaz 174
Padalkar, Milind G. 108
Paiement, Adeline 148
Palmer, Robert Ieuan 87
Pandey, Megha 51
Panteleris, Paschalis 126
Papanelopoulos, Nikos 17
Parizi, Sobhan Naderi 22
Park, Kihong 156
Parkhi, Omkar M. 45, 69
Pascoe, Geoffrey 74
Patras, Ioannis 133
Pătraşcu, Carmen 178
Pekala, Michael 85
Perdoch, Michal 16
Pérez, Patrick 27
Perona, Pietro 56
Perronnin, Florent 114
Pessoa, Saulo 138, 140
Petković, Tomislav 158
Pfister, Hanspeter 85
Pietikäinen, Matti 38
Plötz, Thomas 193
Pock, Thomas 62
Poier, Georg 185
Ponce-López, Víctor 132
Porikli, Fatih 143
Pribanić, Tomislav 158
Přibyl, Bronislav 49
Pridmore, Tony 198
Put, Jeroen 20
Qammaz, Ammar 147
Qi, Honggang 107
Qin, Jie 183
Qu, Chengchao 91
Radenović, Filip 34
Radke, Richard J. 77, 186
Raghuramu, Anil C. 66
Raj, Anant 169
Rajan, Deepu 162
Ramadasan, Datta 110
Ramalingam, Srikumar 117
Ranftl, René 62
Ray, Nilanjan 100
Redondo-Cabrera, Carolina 67
Reinbacher, Christian 105
Reis, Bernardo 138, 140
Ren, Huamin 32
Renaud, Péteri 121
Rhein, Stephen 115
Ricci, Elisa 12

Richard, Alexander 61
Riegler, Gernot 62, 105
Rivera-Rubio, Jose 150
Robinson, Peter 133
Roditakis, Konstantinos 185
Rodriguez, Andres 163
Rodríguez, Raquel Gil 129
Roig, Gemma 19
Romera-Paredes, Bernardino 123
Roncal, William Gray 85
Ronchi, Matteo Ruggero 56
Ronneberger, Olaf 192
Roshtkhari, Mehrsan Javan 182
Rosin, Paul L. 145
Rother, Carsten 184
Rui, Yong 101
Rüther, Matthias 62, 105
Rychlowska, Magdalena 145
Ryu, Seungchul 156
Saavedra, Jose M. 167
Saenko, Kate 28
Safdarnejad, Seyed Morteza 25
Sakurada, Ken 65
Salakhutdinov, Ruslan 64
Santos, Ismael 138, 140
Sapienza, Michael 44
Sapiro, Guillermo 106
Saponaro, Philip 115
Sarfraz, M. Saquib 13
Saxen, Frerk 127
Schaffert, Roman 86
Scharr, Hanno 198
Schuchert, Tobias 91
Schulter, Samuel 185
Schwaiger, Markus 88
Sebe, Nicu 12
Shakeri, Moein 100
Shao, Ling 175, 176, 183
Sharaf, Amr 98
Sharma, Renu 134
She, Yiyuan 135
Shen, Xiaobo 144
Shi, Bibo 141
Shi, Haibo 161
Shi, Xin 135
Shih, Kevin J. 131
Shin, Yongho 155
Sikka, Karan 75, 164
Singh, Saurabh 131
Slossberg, Ron 37, 92
Smith, Charles D. 141
Snoek, Cees G. M. 180, 181
Sohn, Kwanghoon 116, 156
Song, Chunfeng 151
Song, Jingkuan 12
Song, Yi-Zhe 11
Soni, Nishit 117
Sorensen, Scott 115
Sridhar, Muralikrishna 83
Srikantha, Abhilash 104
Srinivas, Suraj 35
Srinivasan, Natesh 139
Srivastava, Nitish 64
Šroubek, Filip 10
Stenger, Björn 130
Stent, Simon 130
Stiefelhagen, Rainer 13
Stretcu, Otilia 189
Su, Hang 30
Sudo, Kyoko 55
Sukhwani, Mohak 120
Sun, Baochen 28
Sun, Quansen 144
Tachos, Stavros 42
Taha, Ahmed 76
Tam, Gary 87, 201
Tan, Ping 50
Tan, Tieniu 70
Tang, Chengzhou 50
Tang, Zhi 159
Taniguchi, Yukinobu 55
Tao, Lili 148
Tarabalka, Yuliya 68
Tavanai, Aryana 83
Teixeira, Luis F. 149
Tepper, Mariano 106
Terzić, Kasim 63
Ţoca, Cosmin 178
Tombari, Federico 40, 125
Torki, Marwan 76
Torr, Philip H. S. 44
Torr, Philip 123
Tsaftaris, Sotirios 198
Turaga, Pavan 71, 196
Tuytelaars, Tinne 169
Udpa, Lalita 25
Uijlings, Jasper 33
Urtasun, Raquel 96
Vakhitov, Alexander 165
Valentin, Julien 44
Vandeventer, Jason 145
Vanhoucke, Vincent 36
Vašíček, Jan 34
Vatolin, Dmitriy 102
Vazquez-Corral, Javier 129
Vedaldi, Andrea 9, 45
Venkataraman, Vinay 71
Verma, Yashaswi 173
Vetter, Thomas 160
Vezhnevets, Alexander 31, 191
Vig, Eleonora 7
Vlachos, Ioannis 71
Vogelstein, Joshua T 85
Vogelstein, R Jacob 85
Walach, Eugene 174
Wang, Changhu 101
Wang, Chong 70
Wang, Jian 86
Wang, Jinqiao 187
Wang, Jue 102
Wang, Liang 151
Wang, Shaofei 8
Wang, Xin 107
Wang, Yongtao 159
Wang, Yu 30
Wang, Yunhong 183
Wang, Zhenhua 179
Wassell, Ian 73
Wei, Fangyun 135
Wen, Mei 188
Wen, Yonggang 161
Werner, Philipp 127
Wetzler, Aaron 37, 92
Wilson, Richard C. 60

Wu, Fuchao 179
Wu, Hong 151
Wu, Jie 54
Wu, Shuang 30
Wu, Xiaomeng 29
Wu, Yi 187
Wu, Ziyan 77
Xiang, Tao 11, 48
Xiao, Huaxin 79
Xie, Xianghua 87
Xu, Chao 161
Xu, Dan 12
Xu, Hongyu 99
Xu, Juan 19
Xu, Wei 79
Xu, Zheng 152
Yamaguchi, Kota 55
Yan, Wang 41
Yan, Yan 12
Yang, Heng 133
Yang, Hua 30
Yang, Huei-Fang 59
Yang, Jie 154
Yang, Juanyu 54
Yang, Kuiyuan 152
Yang, Michael Ying 184
Yang, Peipei 70
Yang, Yi 179
Yang, Yiping 154
Yang, Yongxin 11
Yap, Jordan 41
Ye, Mingquan 168
Yi, Zhou 103
Yin, Fang 39
Yoo, Haan-Ju 72
Yoon, Kuk-Jin 155
Yu, Mengyang 175, 176, 183
Yu, Qian 11
Zakharova, Anastasia 119
Zemčík, Pavel 10, 49
Zepeda, Joaquin 27
Zha, Shengxin 64
Zhai, Mengyao 182
Zhang, Bo 101
Zhang, Chao 124, 135
Zhang, Chunyuan 188
Zhang, Hong 100
Zhang, Hongyang 135
Zhang, Jianguo 79
Zhang, Kaihao 151
Zhang, Lei 101
Zhang, Maojun 79
Zhang, Qianni 54
Zhang, Yichi 133
Zhang, Zheng 79
Zhao, Guoying 38
Zhao, Jun 54
Zhao, Qi 19
Zhao, Qiyang 113
Zheng, Jingjing 99
Zheng, Shibao 30
Zhou, Qin 30
Zhou, Yu 39
Zhu, Guibo 187
Zhu, Hongyuan 15
Ziaeefard, Maryam 170
Zisserman, Andrew 45, 69
Zivanov, Jasenko 160



Notes

Notes

Notes

Notes

IET Journals
The Institution of Engineering and Technology

OPEN ACCESS PUBLISHING OPTIONS AVAILABLE

Publish world-class research with IET *Computer Vision*

IET *Computer Vision* publishes original research papers in a wide range of electronic visual interpretation and recognition, including reconstruction of 3D depth information, estimation of object motion, attribute-based recognition and high-level scene understanding. It publishes the most relevant and topical research in its field, explores new horizons, and aims to set the agenda for future research in computer vision.



Why publish in IET *Computer Vision*?

- Worldwide readership and database coverage, allows your research to be easily accessed
- Prompt and rigorous peer review for quick decisions on publication
- E-First publishing, making your research available at the earliest opportunity
- Discounted pre-submission editing services are available, particularly useful if English is not your first language
- Author URL's upon publication allowing you to share free access to your paper for 6 months

To find out more or to submit a paper, visit

www.**ietdl**.org/IET-CVI

The Institution of Engineering and Technology is registered as a Charity in England and Wales (No. 211014) and Scotland (No. SCO38698).

Venue Maps & Local Information

The confercence is hosted in Singleton Campus of the Swansea University. The campus is located in a park, west to the city centre, and next to a sandy beach.

The oral sessions of the main confernce are hosted in the Taliesin Building, and poster sessions and tea/coffee breaks are in the Taliesin Annexe. Lunches are served in Fulton House.

The onsite residential office is in the Preseli Building, behind Fulton House.

The Conference Reception takes place in Swansea Waterfront Museum in the Marina Quarter. It is 2 miles from the Singleton Campus.

The Conference Banquet is hosted in Brangwyn Hall, which is located between the campus and Swansea Waterfront Museum and is within walking distance from the Singleton Campus.

There are frequent buses from the campus to city centre. The bus stop is in front of Fulton House, where you also find a taxi stand.
Local taxi companies you may use: CityTax 01792 475200 and Station Cabs 01792 477477.



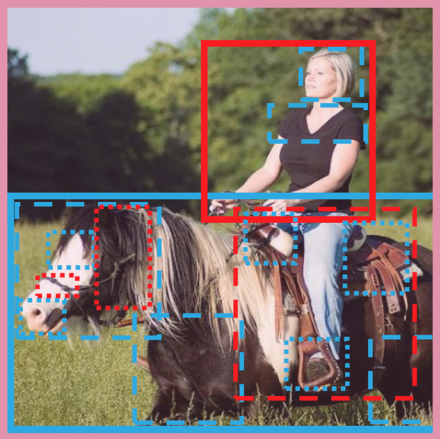




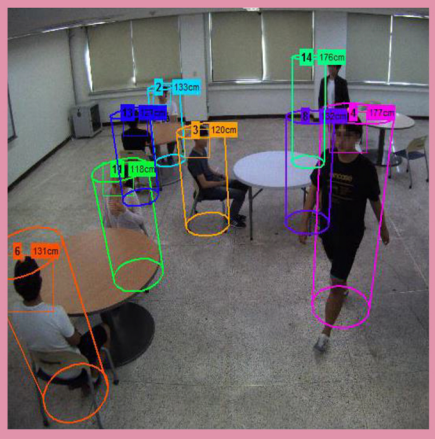





Kitchen
Cellar






14 176cm
2 133cm
3 120cm
11 116cm
6 131cm

