BRITISH MACHINE VISION CONFERENCE 2016
19-22 September
YORK

Contents

People

General Chair	Richard Wilson
Programme Chair	Edwin Hancock
Publicity	William Smith
Tutorials and Workshops	Nick Pears
Local Arrangements	Bob French
	Adrian G. Bors
Industry liaison	Supannee Tanathong

Area Chairs

Adrian Clark	University of Essex
Alessio Del Bue	IST Genova
Andrea Cavallaro	Queen Mary, University of London
Andrea Torsello	University of Venice
Andrew Fitzgibbon	Microsoft Research Cambridge
Antonio Robles-Kelly	NICTA - Canberra
Barbara Caputo	University of Roma
Bjorn Stenger	Toshiba Research Europe
Carole Twining	University of Manchester
Daniel Cremers	Technical University Munchen
Dima Damen	University of Bristol
Edmond Boyer	INRIA Grenoble Rhone-Alpes
Frederic Jurie	University of Caen
Guoping Qiu	University of Nottingham
Hongdong Li	Australian National University
Jianguo Zhang	University of Dundee
John Collomosse	University of Surrey
Jon Starck	The Foundry
Julia Schnabel	Kings College London
Lewis Griffin	University College London
Ling Shao	Northumbria University
Majid Mirmehdi	University of Bristol
Manuel Trucco	University of Dundee
Mathieu Bredif	Institut Géographique National, France
Michel Valstar	University of Nottingham
Nassir Navab	Technical Univeristy Munich
Neill Campbell	University of Bath
Ondrej Chum	Czech Technical University
Paul Siebert	University of Glasgow
Peter Hall	University of Bath

Reyer Zwiggelaar	Aberystwyth University
Shaogang Gong	Queen Mary, University of London
Stefanos Zafeiriou	Imperial College London
Stephen Pollard	Hewlett Packard Labs UK
Stephen Maybank	Birkbeck, University of London
Steven Beauchemin	University of Western Ontario
Sudeep Sarkar	University of South Florida
Tao Xiang	Queen Mary, University of London
Tilo Burghardt	University of Bristol
Tim Cootes	University of Manchester
Toby Breckon	Durham University
Umberto Castellani	University Verona
Xianghua Xie	Swansea University
Xuelong Li	Birkbeck, University of London
Yiannis Patras	Queen Mary, University of London
Zhao Zhang	Soochow University

Foreword

It is our great pleasure to welcome you to York for the 27th British Machine Vision Conference (BMVC). The University of York recently celebrated its 50th birthday (in 2013) and has grown rapidly since its founding. It now home to more than 16,000 students. York is a campus University, sited in parkland and famous for its lakes and waterfowl. The conference is sited on the Heslington West campus on the outskirts of the historic city of York.

BMVC is one of the top events in the Computer Vision conference calendar and must now be considered as a truly international event with the majority of papers coming from outside the UK. This year, BMVC attracted a total of 365 valid submissions. Although this is lower that the record submissions level of recent years, it still represents a very active and healthy conference. The paper review process was unchanged from previous BMVCs, and we recruited more than 300 reviewers to process the papers. All papers received three reviews and each paper was then handled by two area chairs from our pool of 50 subject experts. Accepted papers required strong support from both reviewers and area chairs. We would like to thank all the reviewers and area chairs for their hard work and prompt responses.

Of the 365 submissions, just 144 were accepted for presentation in BMVC 2016, which is a 39% acceptance rate. Only the very highest quality papers were selected for oral presentation, with 38 papers gaining a podium spot, or 10% of the submissions. The accepted papers represent a truly international research community, with 18% of the papers from the UK, 36% from the EU excluding the UK, 22% from Asia, 20% from North America, and 4% from the rest of the world. As is now standard for BMVC, the proceedings are published entirely online, without the use of USB drives, for environmental reasons.

BMVC has always has strong links with industry, and again we are very grateful to our industrial sponsors for supporting the event. ARM, Disney Research, OSRAM, Ocado technology, HP, the IET, Edmund Optics and DigitalBridge kindly supported the main conference. Our thanks also go to NVIDIA, CRC press and Springer for sponsoring the best paper prizes.

We have put together an interesting programme of invited speakers and are delighted to welcome Dr Abhijeet Ghosh, Professor Katsushi Ikeuchi and Professor Raquel Urtasun to the conference. Dr Ghosh will deliver the tutorial on appearance modelling, and Profs. Ikeuchi and Urtasun will give the two keynote presentations in the main conference.

BMVC 2016 has been organised by the Computer Vision and Pattern Recognition group in the Department of Computer Science at the University of York. The organisation of such a large conference would not be possible without the help of many people, and we are grateful to everyone who has contributed. Particular thanks must go to Bob French for his help in sorting out the logistical details and the student helpers for their support during the conference. We would also like to thank Xianghua Xie for his support and advice as the outgoing BMVC chair, and the BMVA committee for their extremely helpful suggestions and advice.

We hope you find BMVC 2016 in York both and enjoyable and valuable experience.

Richard Wilson, Edwin Hancock, Will Smith, Nick Pears, Adrian Bors
BMVC 2016 organising committee

BMVC 2016 Programme

Programme: Monday 19th September: Tutorial

11:00-17:00 Registration - Exhibition Centre

Tutorial - Conference Hall PX001

Chair: William Smith

13:00-14:30 Measurement-based Appearance Modelling

Abhijeet Ghosh 39

14:30-15:00 Break

Tutorial continued - Conference Hall PX001

15:00-16:30 Measurement-based Appearance Modelling

Abhijeet Ghosh 39

19:00-21:00 Welcome Reception - Exhibition Centre

21:00

1

Programme: Tuesday 20th September

08:15-17:45 Registration - Exhibition Centre

08:45-09:00 Welcome

Keynote - Conference Hall PX001

Chair: Edwin Hancock

09:00-10:00 e-Intangible Heritage

Katsushi Ikeuchi 41

Session 1: Segmentation - Conference Hall PX001

Chair: Will Smith

10:00-10:20 Local Shape Transfer for Image Co-segmentation

Wei Teng, Yu Zhang, Xiaowu Chen, Jia Li and Zhiqiang He 47

10:20-10:40 SMURFS: Superpixels from Multi-scale Refinement of Super-regions

Imanol Luengo, Mark Basham and Andrew P. French . . . 48

10:40-11:00 Break

Session 2: Low-level vision and computational photography - Conference Hall PX001

Chair: Eraldo Ribeiro

11:00-11:20 Boundary Detection Through Surround Modulation

Arash Akbarinia, C. Alejandro Parraga 49

11:20-11:40 Bio-inspired Collision Detector with Enhanced Selectivity for Ground Robotic Vision System

Qinbing Fu, Shigang Yue, Cheng Hu 50

11:40-12:00 A Deep Primal-Dual Network for Guided Depth Super-Resolution

Gernot Riegler, David Ferstl, Matthias Rüther and Horst Bischof 51

12:00-12:20 Towards Deep Style Transfer: A Content-Aware Perspective

Yi-Lei Chen, Chiou-Ting Hsu. 52

12:20-12:40 Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularization

Christian Reinbacher, Gottfried Graber and Thomas Pock 53

12:40-13:40 Lunch

Posters 1 - Exhibition Centre

Posters 1 continued - Exhibition Centre

Posters 1 continued - Exhibition Centre

Posters 1 continued - Exhibition Centre

Posters 1 continued - Exhibition Centre

Posters 1 continued - Exhibition Centre

13:40-14:40 Attribute Embedding with Visual-Semantic Ambiguity Removal for Zero-shot Learning

Yang Long, Li Liu and Ling Shao 115

13:40-14:40 NRSfM-Flow: Recovering Non-Rigid Scene Flow from Monocular Image Sequences

Vladislav Golyanik, Aman Shankar Mathur and Didier Stricker 116

13:40-14:40 Better Together: Joint Reasoning for Non-rigid 3D Reconstruction with Specularities and Shading

Qi Liu-Yin, Rui Yu and Andrew Fitzgibbon, Lourdes Agapito, and Chris Russell 117

13:40-14:40 STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition

Wei Liu, Chaofeng Chen, Kwan-Yee K. Wong, Zhizhong Su and Junyu Han 118

13:40-14:40 Context Matters: Refining Object Detection in Video with Recurrent Neural Networks

Subarna Tripathi, Zachary C. Lipton, Serge Belongie, Truong Nguyen 119

Posters 1 continued - Exhibition Centre

Posters 1 continued - Exhibition Centre

Posters 1 continued - Exhibition Centre

13:40-14:40 Fine-grained Recognition in the Noisy Wild: Sensitivity Analysis of Convolutional Neural Networks Approaches

Erik Rodner, Marcel Simon, Robert B. Fisher and Joachim Denzler 135

13:40-14:40 Near-Field Photometric Stereo in Ambient Light

Fotios Logothetis, Roberto Mecca, Yvain Quéau and Roberto Cipolla 136

Session 3: 3D Computer Vision - Conference Hall PX001

Chair: Richard Bowden

14:40-15:00 Adding Synchronization and Rolling Shutter in Multi-Camera Bundle Adjustment

Thanh-Tin Nguyen and Maxime Lhuillier 54

15:00-15:20 EMVS: Event-based Multi-View Stereo

Henri Rebecq, Guillermo Gallego and Davide Scaramuzza 55

15:20-15:40 Occlusion-aware 3D Morphable Face Models

Bernhard Egger, Andreas Schneider, Clemens Blumer, Andreas Morel-Forster, Sandro Schönborn and Thomas Vetter 56

15:40-16:00 Next-Best Stereo: Extending Next-Best View Optimisation For Collaborative Sensors

Oscar Mendez, Simon Hadfield, Nicolas Pugeault and Richard Bowden 57

16:00-16:20 Break

Session 4: Motion and Tracking - Conference Hall PX001

Chair: David Suter

16:20-16:40 Shape-based Image Correspondence

Berk Sevilmis and Benjamin B. Kimia 58

16:40-17:00 Reprojection Flow for Image Registration Across Seasons

Shane Griffith and Cédric Pradalier 59

17:00-17:20 Unsupervised Learning of Shape-Motion Patterns for Objects in Urban Street Scenes

Dirk Klostermann, Aljoša Ošep, Jörg Stückler and Bastian Leibe 60

17.20-18.10 Poster discussion sessions - Exhibition Centre

17.20-18.10 BMVA Members AGM - PX001

Programme: Wednesday 21st September

8:15-17:45 Registration - Exhibition Centre

Keynote - Conference Hall PX001

9:00-10:00 Towards Affordable Self-driving Cars

Raquel Urtasun 43

Chair: Richard Wilson

Session 5: Statistical Methods and Learning - Conference Hall PX001

Chair: Andrea Cavallaro

10:00-10:20 Efficient Learning for Discriminative Segmentation with Supermodular Losses

Jiaqian Yu and Matthew B. Blaschko 61

10:20-10:40 Variational Weakly Supervised Gaussian Processes

Melih Kandemir, Manuel Haußmann, Ferran Diego, Kumar Rajamani, Jeroen van der Laak and Fred A. Hamprecht 62

10:40-11:00 Break

Session 6: Recognition and Physics-based vision - Conference Hall PX001

Chair: Andres Bruhn

11:00-11:20 Regional Gating Neural Networks for Multi-label Image Classification

Rui-Wei Zhao, Jianguo Li, Yurong Chen, Jia-Ming Liu, Yu-Gang Jiang and Xiangyang Xue 63

11:20-11:40 Multispectral Deep Neural Network for Pedestrian Detection

Jingjing Liu, Shaoting Zhang, Shu Wang and Dimitris N. Metaxas 64

11:40-12:00 L1 Graph Based Sparse Model for Label De-noising

Xiaobin Chang, Tao Xiang and Timothy M. Hospedales 65

12:00-12:20 PatchIt: Self-Supervised Network Weight Initialization for Fine-grained Recognition

Patrick Sudowe and Bastian Leibe 66

12:20-12:40 Combining Shape from Shading and Stereo: A Variational Approach for the Joint Estimation of Depth, Illumination and Albedo

Daniel Maurer, Yong Chul Ju, Michael Breuß, Andrés Bruhn 67

12:40-13:40 Lunch

Posters 2 - Exhibition Centre

Posters 2 continued - Exhibition Centre

13:40-14:40 Optimized Regressor Forest for Image Super-Resolution

Chia-Yang Chang, Wei-Chih Tu and Shao-Yi Chien 145

13:40-14:40 Convolutional aggregation of local evidence for large pose face alignment

Adrian Bulat and Yorgos Tzimiropoulos . . . 146

13:40-14:40 Wide Residual Networks

Sergey Zagoruyko and Nikos Komodakis . . . 147

13:40-14:40 Deep Part-Based Generative Shape Model with Latent Variables

Alexander Kirillov, Mikhail Gavrikov, Ekaterina Lobacheva, Anton Osokin and Dmitry Vetrov . . . 148

13:40-14:40 General Human Traits Oriented Generic Elastic Model for 3D Face Reconstruction

Joi San Tan, Ibrahim Venkat, Iman Yi Liao and Philippe De Wilde 149

13:40-14:40 Attend Refine Repeat: Active Box Proposal Generation via In-Out Localization

Spyridon Gidaris and Nikos Komodakis . . . 150

13:40-14:40 Crafting a multi-task CNN for viewpoint estimation

Francisco Massa, Renaud Marlet and Mathieu Aubry 151

13:40-14:40 Improving Weakly-Supervised Object Localization By Micro-Annotation

Alexander Kolesnikov and Christoph H. Lampert . . . 152

Posters 2 continued - Exhibition Centre

13:40-14:40 MBest Struct: M-Best diverse sampling for structured tracker

Ivan Bogun and Eraldo Ribeiro 153

13:40-14:40 Event-Based Hough Transform in a Spiking Neural Network for Multiple Line Detection and Tracking Using a Dynamic Vision Sensor

Sajjad Seifozzakerini, Wei-Yun Yau, Bo Zhao and Kezhi Mao 154

13:40-14:40 Holistically Constrained Local Model: Going Beyond Frontal Poses for Facial Landmark Detection

KangGeon Kim, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency and Gérard Medioni 155

13:40-14:40 Bag of Surrogate Parts: one inherent feature of deep CNNs

Yanming Guo and Michael S. Lew 156

13:40-14:40 Multi-scale Colorectal Tumour Segmentation Using a Novel Coarse to Fine Strategy

Kun Zhang, Danny Crookes, Jim Diamond, Minrui Fei, Jianguo Wu, Peijian Zhang and Huiyu Zhou . . . 157

13:40-14:40 Learning Additive Kernel For Feature Transformation and Its Application to CNN Features

Takumi Kobayashi 158

13:40-14:40 Robust 3D Car Shape Estimation from Landmarks in Monocular Image

Yanan Miao, Xiaoming Tao and Jianhua Lu . . . 159

Posters 2 continued - Exhibition Centre

13:40-14:40 Projective Unsupervised Flexible Embedding with Optimal Graph

Wei Wang, Yan Yan, Feiping Nie, Xavier Alameda Pineda, Shuicheng Yan and Nicu Sebe 160

13:40-14:40 Towards Automatic Image Editing: Learning to See another You

Amir Ghodrati, Xu Jia, Marco Pedersoli and Tinne Tuytelaars 161

13:40-14:40 Dense Labeling with User Interaction: an Example for Depth-Of-Field Simulation

Ana B. Cambra, Adolfo Muñoz, José J. Guerrero and Ana Murillo 162

13:40-14:40 MLBoost Revisited: a Faster Metric Learning Algorithm for Identity-Based Face Retrieval

Romain Negrel, Alexis Lechervy and Frederic Jurie 163

13:40-14:40 Learning Neural Network Architectures using Backpropagation

Suraj Srinivas and R. Venkatesh Babu 164

13:40-14:40 Enhancing pose estimation through efficient patch synthesis

Pierre Rolin, Marie-Odile Berger and Frédéric Sur 165

Posters 2 continued - Exhibition Centre

13:40-14:40 Discovering motion hierarchies via tree-structured coding of trajectories

Juan-Manuel Péréz-Rúa, Tomas Crivelli, Patrick Peréz and Patrick Bouthemy 166

13:40-14:40 Coplanar Repeats by Energy Minimization

James Pritts, Denys Rozumnyi, M. Pawan Kumar and Ondrej Chum 167

13:40-14:40 Two-Stream SR-CNNs for Action Recognition in Videos

Yifan Wang, Jie Song, Limin Wang, Luc Van Gool and Otmar Hilliges 168

13:40-14:40 An Efficient Convolutional Network for Human Pose Estimation

Umer Rafi, Ilya Kostrikov, Ilya Kostrikov and Bastian Leibe 169

13:40-14:40 A data augmentation methodology for training machine/deep learning gait recognition algorithms

Christoforos C. Charalambous and Anil A. Bharath 170

13:40-14:40 Mean Box Pooling: A Rich Image Representation and Output Embedding for the Visual Madlibs Task

Ashkan Mokarian, Mateusz Malinowski and Mario Fritz 171

Posters 2 continued - Exhibition Centre

Posters 2 continued - Exhibition Centre

13:40-14:40 Online Feature Selection for Visual Tracking

Giorgio Roffo and Simone Melzi 180

13:40-14:40 Deep Aggregation of Local 3D Geometric Features for 3D Model Retrieval

Takahiko Furuya and Ryutarou Ohbuchi 181

13:40-14:40 Learning Grimaces by Watching TV

Samuel Albanie and Andrea Vedaldi 182

13:40-14:40 Accurate and robust face recognition from RGB-D images with a deep learning approach

Yuancheng Lee, Jiancong Chen, Ching Wei Tseng and Shang-Hong Lai 183

13:40-14:40 Global Deconvolutional Networks for Semantic Segmentation

Vladimir Nekrasov, Janghoon Ju and Jaesik Choi 184

13:40-14:40 Convolutional Sparse Coding-based Image Decomposition

He Zhang and Vishal M. Patel 185

13:40-14:40 Domain Adaptive Subspace Clustering

Mahdi Abavisani and Vishal M. Patel 186

13:40-14:40 Filtering 3D Keypoints Using GIST For Accurate Image-Based Localization

Charbel Azzi, Daniel Asmar, Adel Fakih and John Zelek 187

Posters 2 continued - Exhibition Centre

13:40-14:40 Probabilistic Obstacle Partitioning of Monocular Video for Autonomous Vehicles

Ryan W. Wolcott and Ryan M. Eustice 188

13:40-14:40 Track Facial Points in Unconstrained Videos

Xi Peng, Qiong Hu, Junzhou Huang, Dimitris N. Metaxas 189

Session 7: Recognition - Conference Hall PX001

Chair: Krystian Mikolajczyk

14:40-15:00 Structured Prediction of 3D Human Pose with Deep Neural Networks

Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua 68

15:00-15:20 Multi-task Relative Attribute Prediction by Incorporating Local Context and Global Style Information Features

Yuhang He, Long Chen and Jianda Chen. 69

15:20-15:40 Deep Multi-task Attribute-driven Ranking for Fine-grained Sketch-based Image Retrieval

Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy Hospedales and Xiang Ruan 70

15:40-16:00 The Role of Context Selection in Object Detection

Ruichi Yu, Xi Chen, Vlad I. Morariu and Larry S. Davis 71

16:00-16:20 Break

Session 8: Face and Gesture - Conference Hall PX001

Chair: Nick Pears

16:20-16:40 Highly Efficient Regression for Scalable Person Re-Identification

Hanxiao Wang, Shaogang Gong and Tao Xiang . . . 72

16:40-17:00 Reflective Regression of 2D-3D Face Shape Across Large Pose

Xuhui Jia, Heng Yang, Xiaolong Zhu, Zhanghui Kuang, Yifeng Niu and Kwok-Ping Chan 73

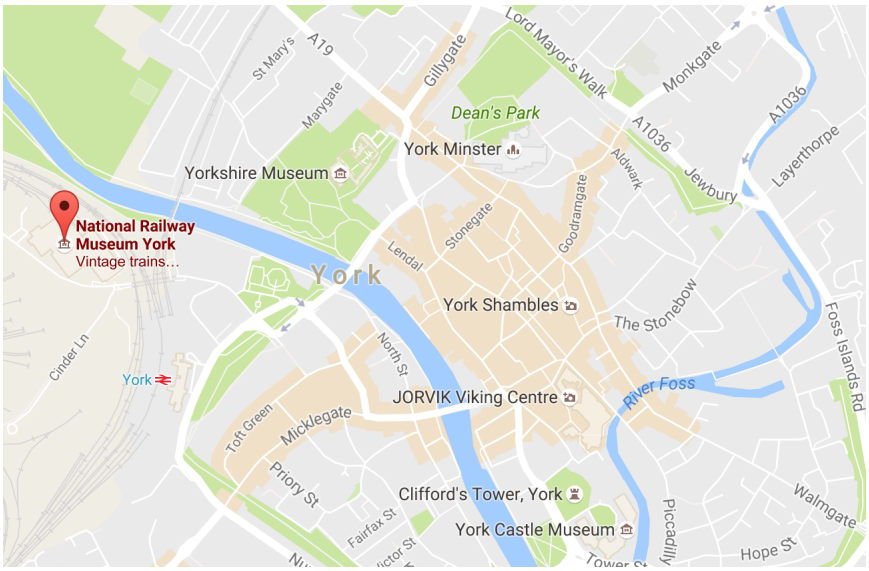
17:00-17:20 Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition

Oscar Koller, Sepehr Zargaran, Hermann Ney and Richard Bowden 74

17.20-18.10 Poster discussion sessions

19:00-22:00 Banquet

The BMVC 2016 Banquet takes place at the National Railway Museum, York



Programme: Thursday 22nd September

Session 9: Recognition, Optimisation and Performance Evaluation - Conference Hall PX001

Chair: Paolo Remagnino

09:00-09:20 Measuring the effect of nuisance variables on classifiers

Alhussein Fawzi and Pascal Frossard 75

09:20-09:40 Learning Robust Graph Regularisation for Subspace Clustering

Elyor Kodirov, Tao Xiang, Zhenyong Fu and Shaogang Gong 76

09:40-10:00 Solving Jigsaw Puzzles with Linear Programming

Rui Yu, Chris Russell and Lourdes Agapito 77

10:00-10:20 Detecting Tracking Errors via Forecasting

Obaid Ullah Khalid, Andrea Cavallaro and Bernhard Rinner 78

10:20-10:40 Oracle Performance for Visual Captioning

Li Yao, Nicolas Ballas, Kyunghyun Cho, John R. Smith, Yoshua Bengio, Frederico A. Limberger, Richard C. Wilson . . . 79

10:40-11:00 Break

Session 10: Video events, robot vision and deep learning - Conference Hall PX001

Chair: Adrian Bors

11:00-11:20 Beyond Action Recognition: Action Completion in RGB-D Data

Farnoosh Heidarivincheh, Majid Mirmehdi and Dima Damen 80

11:20-11:40 Video Stream Retrieval of Unseen Queries using Semantic Memory

Spencer Cappallo, Thomas Mensink, and Cees G. M. Snoek 81

11:40-12:00 Mapping Auto-context to a Deep, Sparse ConvNet for Semantic Segmentation

David L. Richmond, Dagmar Kainmueller, Michael Y. Yang, Eugene W. Myers and Carsten Rother 82

12:00-12:20 Fully-trainable deep matching

James Thewlis, Shuai Zheng, Philip H. S. Torr and Andrea Vedaldi 83

12:20-12:40 Detection of fast incoming objects with a moving camera

Fabio Poiesi and Andrea Cavallaro 84

12:40-12:50 Conference closing remarks

The End of the 27th British Machine Vision Conference 2016

BMVC 2016 Tutorial
Monday 19th Sep. 13:00-14:30, 15:00-16:30

Abhijeet Ghosh
Imperial College, London

Measurement Based Appearance Modelling

Abstract

This tutorial will cover measurement based appearance modelling for graphics and vision, with a focus on acquisition of facial and material appearance including shape and reflectance properties. The tutorial will first introduce fundamentals such as the BRDF and the BSSRDF and physically based surface and subsurface scattering models. Building upon these, the tutorial will present various techniques for acquisition of facial geometry and reflectance including state of the art techniques employed for high quality facial acquisition for visual effects and games. Central to these acquisition techniques will be measurements under various types of controlled illumination. Specifically, we will cover practical ways of measuring layered skin reflectance including surface and subsurface scattering using a small set of measurements as well as state of the art techniques for multi-view facial geometry and reflectance acquisition with polarized spherical gradient illumination. The tutorial will also cover a recent technique for measuring skin micro-geometry at the resolution of a few microns for very high resolution (16K) rendering of skin for increased realism.

The second half of the tutorial will focus on measurement and modelling of material reflectance properties. Here, the discussion will be restricted to BRDFs and spatially varying BRDFs for representing material appearance. Once again, controlled illumination techniques using various lighting setups will be presented for estimation of spatially varying diffuse and specular reflectance properties including albedo, surface normals, specular roughness and in some cases anisotropy. The tutorial will conclude with discussion of some recent advances in material appearance acquisition using commodity hardware such as LCD screens and mobile devices.

Biography

Dr Abhijeet Ghosh is currently a Lecturer in the Department of Computing at Imperial College London. His main research interests are in appearance modelling, realistic rendering, and computational photography. Previously, he was a senior researcher and research assistant professor at the University of Southern California Institute for Creative Technologies where he worked on Light Stage based acquisition. Abhijeet received his PhD in computer science from the University of British Columbia. His doctoral dissertation, "Realistic Materials and Illumination Environments", received an Alain Fournier Award and his doctoral work on BRDF acquisition received a Marr Prize Honorable Mention (ICCV 2007). He currently holds a Royal Society Wolfson Research Merit Award and an EPSRC Early Career fellowship at Imperial College London.

BMVC 2016 Invited Speaker
Tuesday 20th Sep. 9:00-10:00

Katsushi Ikeuchi
Microsoft Research Asia

e-Intangible Heritage

Abstract

Tangible heritage, such as temples and statues, is disappearing day-by-day due to human and natural disaster. In-tangible heritage, such as folk dances, local songs, and dialects, has the same story due to lack of inheritors and mixing cultures. We have been developing methods to preserve such tangible and in-tangible heritage in the digital form. This project, which we refer to as e-Heritage, aims not only record heritage, but also analyze those recorded data for better understanding as well as display those data in new forms for promotion and education.

This talk mainly covers how to preserve in-tangible heritage, in particular, preservation of Japanese and Taiwanese folk dances. The first half of my talk covers how to display such a Japanese folk dance on a humanoid robot. Here, we follow the paradigm, learning-from-observation, in which a robot learns how to dance from observing human dance. Due to the physical difference between a human and a robot, the robot cannot mimic the entire human actions. Instead, the robot first extracts important actions of a dance, referred to key poses, only exactly mimics those key poses and then interpolates interval trajectories as much as possible but within the limit of the robot capabilities. The second half of my talk covers our effort to apply similar technics to Taiwanese folk dances. Here, I concentrate on the analysis of the key poses and how such key poses relate to their social institutions.

Biography

Dr. Katsushi Ikeuchi is a Principal Researcher of Microsoft Research. He received a Ph.D. degree in Information Engineering from the University of Tokyo in 1978. After working at AI Lab of MIT as a pos-doc fellows for three years, Electrotechnical Lab, Japan as a researcher for five years, Robotics Institute of Carnegie Mellon University as a faculty member for ten years, the University of Tokyo as a faculty member for nineteen years, he joined Microsoft Research Asia in 2015. His research interest spans computer vision, robotics, and computer graphics. He has received several awards, including IEEE-PAMI Distinguished Researcher Award, the Okawa Prize from the Okawa foundation, and the Medal of Honor with Purple ribbon from the Emperor of Japan. He is a fellow of IEEE, IEICE, IPSJ, and RSJ.

BMVC 2016 Invited Speaker
Wednesday 21st Sep. 9:00-10:00

Raquel Urtasun
University of Toronto

Towards Affordable Self-driving Cars

Abstract

The revolution of self-driving cars will happen in the near future. Most solutions rely on expensive 3D sensors such as LIDAR as well as hand-annotated maps. Unfortunately, this is neither cost effective nor scalable, as one needs to have a very detailed up-to-date map of the world.In this talk, I?ll review our current efforts in the domain of autonomous driving. In particular, I'll present our work on stereo, optical flow, appearance-less localization, 3D object detection as well as creating HD maps from visual information alone. This results in a much more scalable and cost-effective solution to self-driving cars.

Biography

Raquel Urtasun is an Associate Professor in the Department of Computer Science at the University of Toronto and a Canada Research Chair in Machine Learning and Computer Vision. Prior to this, she was an Assistant Professor at the Toyota Technological Institute at Chicago (TTIC), an academic computer science institute affiliated with the University of Chicago. She received her Ph.D. degree from the Computer Science department at Ecole Polytechnique Federal de Lausanne (EPFL) in 2006 and did her postdoc at MIT and UC Berkeley. Her research interests include machine learning, computer vision and robotics. Her recent work involves perception algorithms for self-driving cars, deep structured models and exploring problems at the intersection of vision and language. She is a recipient of an NVIDIA Pioneers of AI Award, a Ministry of Education and Innovation Early Researcher Award, two Google Faculty Research Awards, a Connaught New Researcher Award and a Best Paper Runner up Prize awarded at the Conference on Computer Vision and Pattern Recognition (CVPR). She is also Program Chair of CVPR 2018, an Editor of the International Journal in Computer Vision (IJCV) and has served as Area Chair of multiple machine learning and vision conferences (i.e., NIPS, UAI, ICML, ICLR, CVPR, ECCV, ICCV).

Abstracts of papers presented at BMVC 2016, York

Local Shape Transfer for Image Co-segmentation

Wei Teng[1]
tengw@buaa.edu.cn
Yu Zhang[1]
zhangyulb@gmail.com
Xiaowu Chen[†1]
chen@buaa.edu.cn
Jia Li[12]
jiali@buaa.edu.cn
Zhiqiang He[3]
lirong2@lenovo.com

[1] State Key Laboratory of Virtual Reality Technology and Systems
Beihang University
Beijing, China

[2] International Research Institute for Multidisciplinary Science
Beihang University
Beijing, China

[3] Lenovo Research

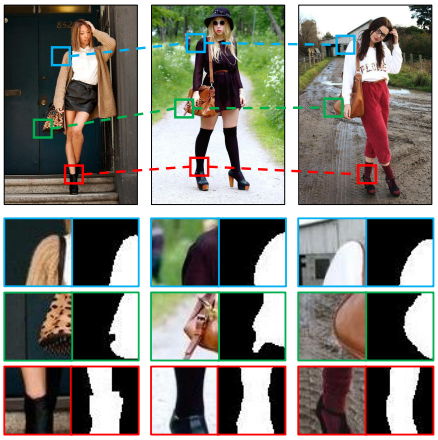


Figure 1: The motivation of this paper. The common objects in these images have different poses, rendering their global shapes inconsistent. However, the local object shapes in different images are highly consistent and provide important cues for co-segmentation.

Image co-segmentation is a challenging computer vision task that aims to segment all pixels of the common objects in an image set. In real-world cases, the common objects often vary greatly in poses, locations and scales, making their global shapes highly inconsistent across images and difficult to be segmented. However their local shapes are often highly consistent (see Fig. 1) and thus transferable. Based on the observation, we propose a novel co-segmentation approach, which transfers patch-level local object shapes and appears more consistently across different images. Given a group of M images, our framework first estimates coarse initial foreground segmentations by thresholding saliency maps [3]. Meanwhile, we build inter-image connections by constructing a weighted graph on patches sampled from different images using [1], where weights are learned by Locally Linear Embedding [2]. With the patch graph, we refine the initial segmentation in each image by transferring the local shapes among different images. Formally, we minimize the objective

$$\min_{\mathbf{y}} \sum_{i=1}^{M} E_{\text{seg}}(\mathbf{y}^{[i]}) + \alpha \sum_{i=1}^{P} \left\| \vec{y}_i - \sum_{j \in N_i} w_{ij} \vec{y}_j \right\|^2,$$
$$\text{s.t. } \mathbf{y} \in \{0,1\}^{|\mathbf{y}|},$$

where $\mathbf{y}$ concatenates the binary labels of all pixels in the image set, $\mathbf{y}^{[i]}$ is the part from the ith image. The energy E_{seg} implements intra-image foreground/background segmentation, for which we use the popular Markov Random Field energy. The problem is NP-hard and usually large scale as it operates on pixels, which is approximately solved by half quadratic splitting.

We evaluate the proposed approach on two public benchmarks: iCoseg and Fashionista. Experiments show that our approach performs comparably with or better than the state-of-the-arts on iCoseg dataset, while achieving more than 31% relative improvements on Fashionista dataset.

[1] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *CVPR*, pages 2307–2314, 2013.

[2] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[3] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, and B. Price. Minimum barrier salient object detection at 80 fps. In *ICCV*, pages 1404–1412, 2015.

† correspondence should be addressed to Xiaowu Chen.

Tuesday
10:20-10:40

SMURFS: Superpixels from Multi-scale Refinement of Super-regions

Imanol Luengo[1]
imanol.luengo@nottingham.ac.uk
Mark Basham[2]
mark.basham@diamond.ac.uk
Andrew P. French[1]
andrew.p.french@nottingham.ac.uk

[1] School of Computer Science,
University of Nottingham,
Nottingham, UK, NG8 1BB

[2] Diamond Light Source Ltd,
Harwell Science & Innovation Campus,
Didcot, UK, OX11 0DE

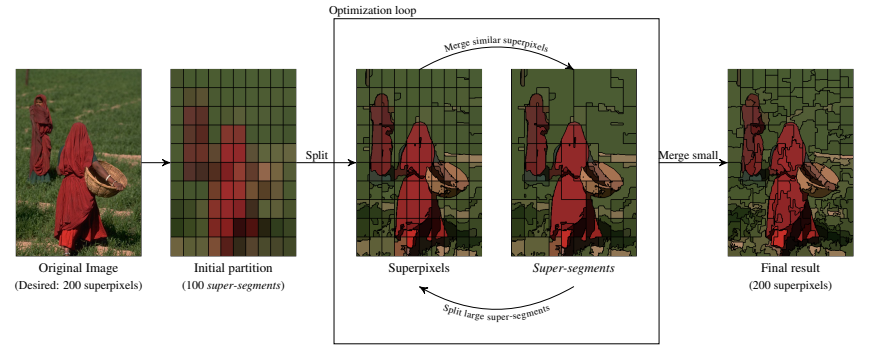


Figure 1: Overview of our algorithm. Iterative refinement over two scales of regions yields increasingly more robust superpixels that better capture global image features.

Here we present a new superpixel algorithm: Superpixels from MUlti-scale ReFinement of Super-regions (SMURFS), which not only obtains state of the art superpixels, but can also be applied hierarchically to form what we call n-th order super-regions. In essence, starting from a uniformly distributed set of super-regions, the algorithm iteratively alternates graph-based split and merge optimization schemes which yield superpixels (1st order super-regions) that better represent the image. We define a super-region hierarchy forming the level i by grouping elements of the level $i-1$. Denoting the pixel grid as level $i=0$, superpixels (level $i=1$) are formed by grouping similar adjacent pixels while *supersegments* (level $i=2$) are formed of multiple superpixels. To be able to better represent the image, we alternate optimization schemes at both level $i=1$ and $i=2$ with the aim of refining both superpixels and supersegments simultaneously. The split step is performed over the pixel grid to separate large supersegments into different smaller superpixels. This step is fully parallelizable as every region is split independently, and produces superpixels that better capture local information of the supersegments. The merging process, conversely, is performed over the superpixel graph to create *supersegments* with the aim of better capturing global image features. This iterative two-scale procedure refines the super-region boundaries of the image without shape or boundary initialization constraints, present in most of state of the art superpixels. Results show state of the art Achievable Segmentation Accuracy (ASA) in the Berkley Segmentation dataset (BSD500) [1].

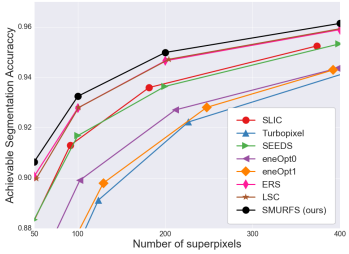


Figure 2: ASA comparison on the BSD500 dataset.

[1] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

Boundary Detection Through Surround Modulation

Arash Akbarinia
www.cvc.uab.es/people/sakbarinia/
C. Alejandro Parraga
www.cvc.uab.es/people/aparraga/

Centre de Visió per Computador
Universitat Autònoma de Barcelona
Barcelona, Spain

Edges are key components of any visual scene to the extent that we can recognise objects merely by their silhouettes. Human visual system captures edge information using neurons that are sensitive to both intensity discontinuities and particular orientations. The "classical approach" assumes that these cells are only responsive to the stimulus present within their receptive fields (RF), however, recent studies demonstrate that surrounding regions and interareal feedback connections influence their responses significantly. In this work we propose a biologically-inspired edge detection model based on these physiological findings.

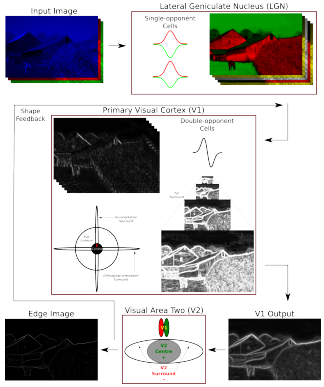


Figure 1: The flowchart of proposed model.

Figure 1 shows the schematics of our model. The original image is processed by balanced and imbalanced single opponent cells in the retina and sent though the lateral geniculate nucleolus in the form of colour opponent channels [4]. Orientation is obtained in the primary visual cortex (V1) by convolving these channels with double-opponent cells (known to be responsive to colour edges [4]), whose RF we modelled through the first derivative of a Gaussian function. To consider the RF surround: we define a short range circular (isotropic) region corresponding to full surround [2], long range iso- and orthogonal-orientation surrounds along the primary and secondary axes of the RF [1], and we model far surround via feedback connections. These interactions are inversely dependant on the contrast of the RF [5]. V1 output signal is pooled at V2 by a contrast-variant centre-surround mechanism applied orthogonally to the preferred direction of the V1 RF [3]. To account for the impact of global shapes on local contours [2], we feed the output of V2 back into V1.

Our experiments suggest that V1 surround modulation strengthens edges while V2 suppresses undesired textural elements (Figure 2).

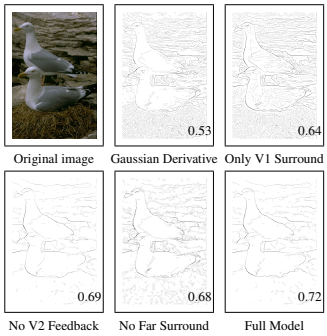


Figure 2: Components evaluation.

[1] David J Field et al. Contour integration and the association field. *The new visual neurosciences*, pages 627–638, 2013.

[2] Gunter Loffler. Perception of contours and shapes: Low and intermediate stage mechanisms. *Vision Research*, 48(20):2106–2127, 2008.

[3] F Poirier and H R Wilson. A biologically plausible model of human radial frequency perception. *Vision research*, 46(15):2443–2455, 2006.

[4] Robert Shapley and Michael J Hawken. Color in the cortex: single-and double-opponent cells. *Vision research*, 51(7):701–717, 2011.

[5] S Shushruth et al. Comparison of spatial summation properties of neurons in macaque v1 and v2. *J. of neurophysiology*, 102(4):2069–2083, 2009.

Bio-inspired Collision Detector with Enhanced Selectivity for Ground Robotic Vision System

Tuesday
11:20-11:40

Qinbing Fu
10230460@students.lincoln.ac.uk
Shigang Yue
syue@lincoln.ac.uk
Cheng Hu
chu@lincoln.ac.uk

Computational Intelligence Laboratory
Department of Computer Science
University of Lincoln
Lincoln, UK

There are many ways of building collision-detecting systems. In this paper, we propose a bio-inspired collision detector based on the juvenile locust vision pathway (Fig.1). Two types of motion detectors, LGMD1 and LGMD2 have been identified in locusts' visual system [1]. Compared to LGMD1, LGMD2 matures early in juvenile locusts which mainly live on the ground whereas already represent evasive responses to swooping predators. An important feature is its looming sense is only for light-to-dark luminance change. It is able to detect dark looming objects embedded in the bright background selectively whilst not responding to light objects against the dark background. There are two defects in LGMD1 modeling works (*e.g.* [2]): first, the approaching and receding stimulus are not properly distinguished in depth; second, the translating stimulus regularly leads to collision mis-detection. The revealed neural characteristics of LGMD2 make it ideal to handle those defects for ground vision-based platforms. Compared to some state-of-the-art collision detectors, the proposed bio-inspired computational model can cope with unpredictable environments without using specific object recognition algorithms. It detects potential collision via reacting to expansion of the object edges, rather than the strategy of recognizing the target or analyzing the scene.

The core of this framework is a biophysical architecture of ON and OFF visual pathways, which underlies motion detection circuit, and reveals the fundamental principle of splitting visual signals into parallel channels encoding brightness increments (ON) and decrements (OFF) as illustrated in Eq.1. Moreover, in LGMD2 modeling work, we put forth a bias in ON pathway to achieve its specific collision selectivity.

The proposed framework was set up in a vision-based ground miniature robot and tested against systematic and comparative real-time experiments. Compared to other computer vision techniques, this neural system performs quickly and robustly in the very limited hardware. The experimental results also demonstrate two main contributions: first, the collision selectivity to dark objects against bright background is enhanced which makes it ideal for ground mobile robots; second, the selectivity to approaching objects versus translation has been shaped which is expected for a practical collision-detecting system.

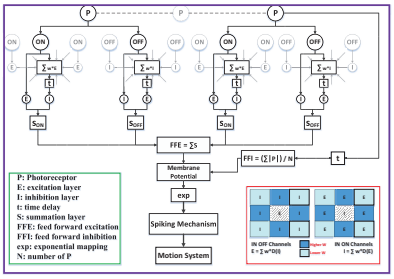


Figure 1: The schematic overview of LGMD2 vision system: Model notations are illustrated in green-box. The convolution illustrations are in red-box.

$$
\begin{aligned}
P_{x,y}^{ON}(t) &= (P_{x,y}(t) + |P_{x,y}(t)|)/2, \\
P_{x,y}^{OFF}(t) &= |(P_{x,y}(t) - |P_{x,y}(t)|)|/2
\end{aligned}
\tag{1}
$$

[1] P. J. Simmons and F. C. Rind. Responses to object approach by a wide field visual neurone, the lgmd2 of the locust: Characterization and image cues. *J Comp Physiol A*, 180:203–214, 1997.

[2] S. Yue and F. C. Rind. Collision detection in complex dynamic scenes using a lgmd based visual neural network with feature enhancement. *IEEE Trans. Neural Netw.*, 17(3):705–716, 2006.

A Deep Primal-Dual Network for Guided Depth Super-Resolution

Gernot Riegler, David Ferstl,
Matthias Rüther, Horst Bischof
{riegler,ferstl,ruether,bischof}@icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

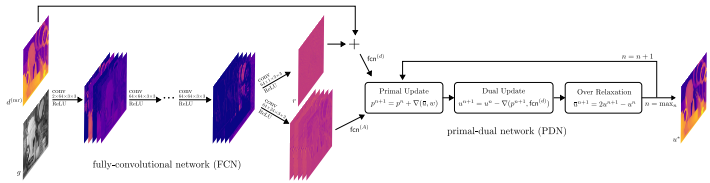


Figure 1: Overview of our proposed *deep primal-dual network*.

Sensors that measure pixel-wise depth have become increasingly popular and enabled a broad range of novel computer vision applications. However, these sensors suffer from a low spatial resolution and depth noise due to physical limitations of the measurement principles.

In this work we present a novel method to increase the spatial and lateral resolution of noisy depth images. We combine a deep fully convolutional network (*FCN*) with a non-local variational method in a *deep primal-dual network* (see Fig. 1) extending our work presented in [4]. The input to our method is a low resolution, noisy depth map $d^{(\mathrm{lr})}$ and a high-resolution intensity image g that is used as guidance in the upsampling process. This guidance image is essential for higher upsampling factors, as shown in our experiments. First, the fully convolutional network computes the residual to the bilinear upscaled low resolution depth input. Further, the *FCN* outputs non-local weighting terms, which are utilized in the subsequent primal-dual network (*PDN*) as weighting coefficients and correspond to discontinuities in the high resolution depth data.

In the primal-dual network we compute the optimizer u_k^* of a variational energy functional

$$u_k^* = \arg\min_u \lambda D(u, \mathsf{fcn}_{s_k}^{(d)}) + R(u, \mathsf{fcn}_{s_k}^{(A)}),$$

by unrolling the computation steps of the first-order primal-dual scheme by [1]. D is the data term penalizing deviations from the initial solution, R is the regularization term encoding smoothness assumptions, and λ is a trade-off parameter. We evaluate in this work several popular choices of regularization terms, which are especially suited for depth data and found that a non-local Huber regularization, as given by

$$R(u) = \int_\Omega \int_{\mathcal{N}(x)} w(x,y)|u(x) - u(y)|_\varepsilon \,\mathrm{d}x\,\mathrm{d}y,$$

in combination with a ℓ_2 data term yields the best trade-off between accuracy and computational requirements. The benefit of unrolling the optimization algorithm in the network are that we can learn all parameters of the variational method, as well as, all hyper-parameter of the optimization scheme itself. Further, the fully-convolutional network adapts in the joint training to the subsequent primal-dual network.

To train our network we generate high-quality depth maps and corresponding color images with a physically based renderer in large quantities. Using this data, we pre-train the fully-convolutional network and subsequently train the complete model end-to-end.

In our experimental evaluation we show the influence of the energy functional and the non-local neighborhood size on the performance of our method. Further, we compare our method on two standard benchmarks for depth super-resolution to other recent approaches: On the noisy Middlebury images [3] and the realistic ToFMark dataset [2]. With this novel combination we are able create visually appealing results and outperform state-of-the-art on both datasets.

Acknowledgment: This work was supported by the Austrian Research Promotion Agency project TOFUSION (FIT-IT Bridge program).

[1] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[2] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image Guided Depth Upsampling using Anisotropic Total Generalized Variation. In *ICCV*, 2013.

[3] J. Park, H. Kim, Y. Tai, M. Brown, and I. Kweon. High Quality Depth Map Upsampling for 3D-TOF Cameras. In *ICCV*, 2011.

[4] G. Riegler, M. Rüther, and H. Bischof. ATGV-Net: Accurate Depth Super-Resolution. In *ECCV*, 2016.

Tuesday
12:00-12:20

Towards Deep Style Transfer: A Content-Aware Perspective

Yi-Lei Chen
https://sites.google.com.tw/site/fallcolor
Pixart Imaging Inc.
Hsinchu, Taiwan

Chiou-Ting Hsu
http://www.cs.nthu.edu.tw/~cthsu/candy.html
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

Motivation

Recently, it has been shown that one can invert a deep convolutional neural network originally trained for classification tasks to transfer image style. There is, however, a dearth of research on content-aware style transfer. In this paper, we generalize the original neural algorithm [1] for style transfer from two perspectives: *where to transfer* and *what to transfer*. To specify where to transfer, we propose a simple yet effective *masking out* strategy to constrain the transfer layout. To illustrate what to transfer, we define a new style feature by high-order statistics to better characterize content coherency.

Methodology

Given a source image (or content image) $\mathbf{c}$ and a target image (or style image) $\mathbf{s}$, [1] aims to synthesize an image $\mathbf{x}$ which simultaneously shares the visual content of $\mathbf{c}$ and the style representation of $\mathbf{s}$. Specifically, the image rendering was modelled as an optimization problem by minimizing the difference between $\mathbf{c}$ and $\mathbf{x}$ and the difference between $\mathbf{s}$ and $\mathbf{x}$ in terms of content and style features, respectively. The authors characterize both features by the deep convolutional neural network (CNN). The desired image was obtained by

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \lambda \sum_{l\in l_c} \left\| \mathbf{F}_{l(\mathbf{x})} - \mathbf{F}_{l(\mathbf{c})} \right\|^2 + \sum_{l\in l_s} \left\| \mathbf{G}_{l(\mathbf{x})} - \mathbf{G}_{l(\mathbf{s})} \right\|^2 + \gamma\Gamma(\mathbf{x}) \quad (1)$$

where the content feature $\mathbf{F}_{l(\mathbf{x})}$ is the layer-wise response and the style feature $\mathbf{G}_{l(\mathbf{x})} = \mathbf{F}_{l(\mathbf{x})}{}^{\mathrm{T}}\mathbf{F}_{l(\mathbf{x})}$ encodes cross-feature dependencies globally.

We formulate the generalized style transfer based on Equation (1) under two additional constraints: *where to transfer* and *what to transfer*. To constrain *where to transfer*, we introduce a diagonal matrix $\mathbf{M}_{l(\mathbf{x})}$, whose $(i,i)^{\text{th}}$ entry m_i $(0 \le m_i \le 1)$ is a soft indictor of feature aggregation, to specify the spatial correspondence. To constrain *what to transfer*, we propose a new feature statistics $\hat{\mathbf{G}}_{l(\mathbf{x})} = (\mathbf{P}_{l(\mathbf{x})}\mathbf{F}_{l(\mathbf{x})})^{\mathrm{T}}(\mathbf{P}_{l(\mathbf{x})}\mathbf{F}_{l(\mathbf{x})})$, by introducing a high-order convolutional matrix $\mathbf{P}_l$, to better match the style representation. Finally, we propose to embed both two constraints into the style loss of Equation (1) and derive the layer-wise gradient in a general form:

$$\nabla_{\mathbf{F}_{l(\mathbf{x})}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{P}_l^{(j)\mathrm{T}}\mathbf{M}_{l(\mathbf{x})}^{(k)\mathrm{T}}\left(\mathbf{M}_{l(\mathbf{x})}^{(k)}\mathbf{P}_l^{(j)}\mathbf{F}_{l(\mathbf{x})}\right)\left(\hat{\mathbf{G}}_{l(\mathbf{x})}^{(k)} - \hat{\mathbf{G}}_{l(\mathbf{s})}^{(k)}/N_l^{(k)}\right),$$

$$\text{where } \hat{\mathbf{G}}_{l(\mathbf{x})}^{(k)} = \left(\mathbf{M}_{l(\mathbf{x})}^{(k)}\mathbf{P}_l^{(j)}\mathbf{F}_{l(\mathbf{x})}\right)^{\mathrm{T}}\left(\mathbf{M}_{l(\mathbf{x})}^{(k)}\mathbf{P}_l^{(j)}\mathbf{F}_{l(\mathbf{x})}\right). \quad (2)$$

Results

We show an example for real-life photo transfer in Fig. 1. Using the semantic masks estimated by image matting, we successfully transfer the dogs' appearance without either changing background or producing noticeable artifacts. Please refer to our paper for more style transfer results

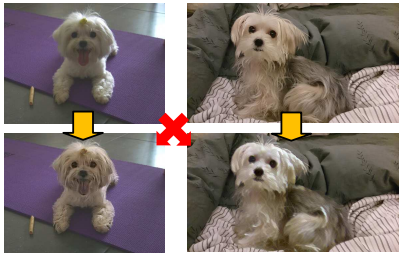


Figure 1: An example of content-aware style transfer. The face appearance of two different breeds of dogs, *Maltese* and *Yorkshire terrier*, are exchanged by our method.

[1] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, 2016.

Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularisation

Christian Reinbacher[1]
reinbacher@icg.tugraz.at
Gottfried Graber[1]
graber@icg.tugraz.at
Thomas Pock[1,2]
pock@icg.tugraz.at

[1] Graz University of Technology
Institute for Computer Graphics and Vision

[2] Austrian Institute Of Technology
Vienna

(a) Raw Events (b) Reconstructed Image (c) Event Manifold

Figure 1: Sample results from our method. The image (a) shows the raw events and (b) is the result of our reconstruction. The time since the last event has happened for each pixel is depicted as a surface in (c) with the positive and negative events shown in green and red respectively.

Event cameras or neuromorphic cameras mimic the human perception system as they measure the per-pixel *intensity change* rather than the actual *intensity level*. In contrast to traditional cameras, such cameras capture new information about the scene at MHz frequency in the form of sparse events. The high temporal resolution comes at the cost of losing the familiar per-pixel intensity information.

In this work we aim to bridge the gap between the time-continuous domain of events and frame-based computer vision algorithms. We propose a simple method for simultaneous denoising and intensity reconstruction for neuromorphic cameras in real-time (see Fig. 1 for a sample output of our method). In contrast to very recent work on the same topic by Bardow *et al*. [1], we formulate our algorithm on an event-basis, avoiding the need to simultaneously estimate the optical flow. We cast the intensity reconstruction problem as an energy minimisation, where we model the camera noise in a data term given by the *generalised Kullback-Leibler divergence*. The optimisation problem is defined on a manifold induced by the timestamps of new events (see Fig. 1(c)). Benosman *et al*. [2] fittingly call this manifold the *surface of active events*. We show how to optimise this energy using the Primal-Dual algorithm of Chambolle and Pock [3] and achieve real-time performance by implementing the energy minimisation on a graphics processing unit (GPU). We release software to provide live intensity image reconstruction to all users of DVS cameras[1]. We believe this will be a vital step towards a wider adoption of this kind of cameras.

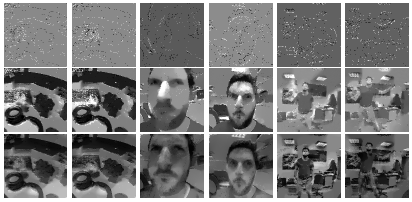


Figure 2: Comparison to the method of [1]. The first row shows the raw input events that have been used for both methods. The second row depicts the results of Bardow *et al*., and the last row shows our result. We can see that out method produces more details (e.g. face, beard) as well as more graceful gray value variations in untextured areas, where [1] tends to produce a single gray value.

[1] Patrick Bardow, Andrew Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *CVPR*, 2016.

[2] R. Benosman, C. Clercq, X. Lagorce, S. H. Ieng, and C. Bartolozzi. Event-based visual flow. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):407–417, 2014.

[3] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1), 2011.

[1] `https://github.com/VLOGroup/dvs-reconstruction`

Adding Synchronization and Rolling Shutter in Multi-Camera Bundle Adjustment

Tuesday
14:40-15:00

Thanh-Tin Nguyen

Maxime Lhuillier
http://maxime.lhuillier.free.fr

Institut Pascal
CNRS UMR 6602, Université Blaise Pascal, IFMA
Aubière, FR

This paper introduces a new bundle adjustment for multi-cameras, that simultaneously estimates not only the usual parameters (poses and points) but also the synchronization and the rolling shutter. We start from an initial calibration with a global shutter camera model (GS) and a frame-accurate synchronization provided by previous self-calibration methods [1]. Our BA provides subframe-accurate synchronization (SFA), *i.e.* it estimates the residual time offsets Δ_j between a reference camera and the others. It also estimate the rolling shutter (RS) coefficient, *i.e.* the time delay τ between two adjacent lines of a frame.

We present a continuous camera trajectory model that provides the multi-camera pose at every time. Let $\mathcal{R}$ be a $\mathcal{C}^1$ continuous function that maps $\Omega \subseteq \mathbb{R}^k$ to rotation set $SO(3)$. The camera trajectory is modeled by a $\mathcal{C}^3$ continuous function $M : [0,1] \to \mathbb{R}^3 \times \Omega$ such that $M(t)^T = \begin{pmatrix} T_M(t)^T & E_M(t)^T \end{pmatrix}$ and $T_M(t) \in \mathbb{R}^3$ is the translation and $\mathcal{R}(E_M(t)) \in SO(3)$ is the rotation. We approximate M at every time t by using M at few times $t_i \in [0,1]$ corresponding to the key-frames provided by standard GS-multi-camera structure-from-motion: $M(t)$ is a linear combination of three $M(t_i)$ thanks to Taylor approximations and by neglecting their remainders. The y-th line of the j-th camera in the i-th keyframe is taken at time $t = t_i + \Delta_j + y\tau$.

Following [2], a minimal parametrization $\mathcal{R}$ is preferred to avoid any constraints on the $\mathcal{R}$ entry and limit the number of estimated parameters: $k = 3$. However, $\mathcal{R}$ is a global parametrization of all rotations of the continuous camera motion and $\mathcal{R}$ has singularities like every 3D parametrization of $SO(3)$. We propose to define $\mathcal{R}$ using a careful use of the Euler parametrization and keep away from the Euler singularities thanks to an assumption on the camera (all yaw angles are possible, but the pitch and roll are small). This assumption is reasonable for an user exploring the environment without a special objective like grasping at objects on the ground.

We experiment in cases that we believe useful: several and identical consumer cameras mounted on a helmet under varying conditions: bike riding, walking, and flying. Ground truth is available for τ (using a strobe) and Δ_j (synthetic videos). At first glance, our approximations seem hazardous if the user does a motion that is not consistent with the neighboring keyframes. Anyway, the majority of keyframes provides accurate enough approximation to obtain the following results in our non trivial datasets. The relative error of the estimated line delay is less than 7.9% except in the most difficult case (flying) with faster head motions; the simultaneous estimation of line delay and time offsets can provide bias but it also provides the best result (5.1%) for the most difficult case. The best time offsets are given by the simultaneous estimation.

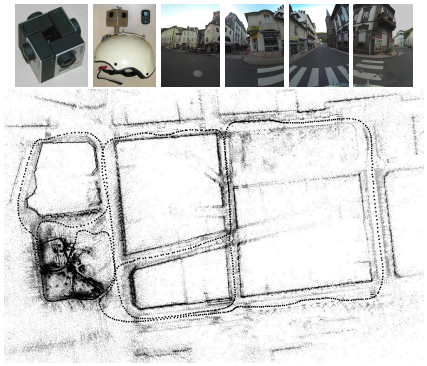


Figure 1: Helmet-held multi-camera formed by four GoPro Hero3 cameras, images taken at a viewpoint, reconstruction of a 900m long video sequence (walking in a town) by RS-SFA bundle adjustment without loop closure.

[1] M. Lhuillier and T.T. Nguyen. Synchronization and self-calibration for helmet-held consumer cameras, applications to immersive 3d modeling and 360 videos. In *3DV'15*.

[2] L. Oth, P. Furgale, L. Kneip, and R. Siegwart. Rolling shutter camera calibration. In *CVPR'13*.

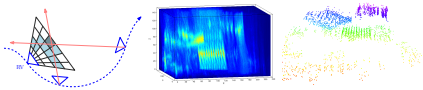
EMVS: Event-based Multi-View Stereo

Henri Rebecq
rebecq@ifi.uzh.ch
Guillermo Gallego
guillermo.gallego@ifi.uzh.ch
Davide Scaramuzza
sdavide@ifi.uzh.ch

Robotics and Perception Group
University of Zurich
http://rpg.ifi.uzh.ch
Zurich, Switzerland

Event cameras are bio-inspired vision sensors that output pixel-level brightness changes instead of standard intensity frames. They offer significant advantages over standard cameras, namely a very high dynamic range, no motion blur, and a latency in the order of microseconds. However, because the output is composed of a sequence of asynchronous events rather than actual intensity images, traditional vision algorithms cannot be applied, so that a paradigm shift is needed.

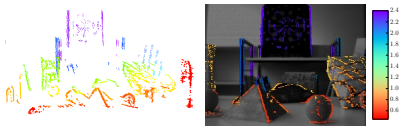
We introduce the problem of Event-based Multi-View Stereo (EMVS) for event cameras and propose a solution to it. Unlike traditional MVS methods, which address the problem of estimating *dense* 3D structure from a set of known viewpoints, EMVS estimates *semi-dense* 3D structure from an event camera with known trajectory. Our EMVS solution elegantly exploits two inherent properties of an event camera: (*i*) its ability to respond to scene edges—which naturally provide semi-dense geometric information without any pre-processing operation—and (*ii*) the fact that it provides continuous measurements as the sensor moves. Despite its simplicity (it can be implemented in a few lines of code), our algorithm is able to produce accurate, semi-dense depth maps. We successfully validate our method on both synthetic and real data. Our method is computationally very efficient and runs in real-time on a CPU.



(a) (b) (c)

Figure 1: Events are back-projected into rays (a), which are counted in a ray density DSI (b), from which a semi-dense 3D reconstruction of the scene edges is extracted by detecting local maxima (c).

We solve the EMVS problem in a similar way to the Space-Sweep approach for MVS [2], showing how sparsity can be leveraged to estimate 3D structures without the need for explicit data association or photometric information.



(a) Front view. (b) Projection on a frame.

Figure 2: Reconstruction of a desk scene with objects of different shapes that cause occlusions. Depth is colored from red (close) to blue (far).

The Event-based Space-Sweep Method consists of three steps:

- Events are back-projected according to the viewpoint of the camera. Each event produces a viewing ray (Fig. 1(a)).
- Rays are counted on a discretized volume containing the 3D scene, yielding a DSI (Fig. 1(b)) that measures the spatial density of rays.
- Scene points are detected in the regions of high ray-density. The location and value of local maxima of the DSI provide the depth and confidence of reconstructed 3D points, respectively. The most confident values yield a semi-dense depth map or point cloud (Fig. 1(c))

Results. Our algorithm produces accurate 3D reconstructions (relative depth errors $< 5\%$) in the presence of high dynamic range (HDR) illumination and/or high-speed motions that cannot be handled by standard cameras, in spite of the low spatial resolution of the sensor [1] and the high amount of noise typical of event cameras.

Multimedia Material. A supplemental video for this work is available on the authors' webpage: http://rpg.ifi.uzh.ch

[1] C. Brandli et al. A 240x180 130dB 3us latency global shutter spatiotemporal vision sensor. *IEEE J. of Solid-State Circuits*, 2014.

[2] R. T. Collins. A space-sweep approach to true multi-image matching. In *IEEE CVPR*, 1996.

Occlusion-aware 3D Morphable Face Models

Tuesday
15:20-15:40

Bernhard Egger
bernhard.egger@unibas.ch
Andreas Schneider
andreas.schneider@unibas.ch
Clemens Blumer
clemens.blumer@unibas.ch
Andreas Morel-Forster
andreas.forster@unibas.ch
Sandro Schönborn
sandro.schoenborn@unibas.ch
Thomas Vetter
thomas.vetter@unibas.ch

Department of Mathematics and Computer Science
University of Basel
Basel Switzerland
http://gravis.cs.unibas.ch

We propose a probabilistic occlusion-aware extension to 3D Morphable Face Models [1, 2] for face image analysis based on the Analysis-by-Synthesis setup. In natural images, parts of the face are often occluded by a variety of objects. Such occlusions are a challenge for face model adaptation. We propose to segment the image into face and non-face regions and model them separately. The segmentation and the face model parameters are not known in advance and have to be adapted to the target image. A good segmentation is necessary to obtain a good face model fit and vice-versa. Therefore, face model adaptation and segmentation are solved together using an EM-like procedure. We use a stochastic sampling strategy based on the Metropolis-Hastings algorithm for face model parameter adaptation [3] and a modified Chan-Vese segmentation for face region segmentation. Previous robust methods are limited to homogeneous, controlled illumination settings and tend to fail for important regions such as the eyes or mouth. We propose a RANSAC-based robust illumination estimation technique to handle complex illumination conditions. We do not use any manual annotation and the algorithm is not optimised to any specific kind of occlusion or database. We evaluate our method on a controlled and an "in the wild" database.

[1] Blanz and Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.

[2] Paysan et al. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009.

[3] Schönborn et al. A monte carlo strategy to integrate detection and model-based face analysis. In *Pattern Recognition*, 2013.

[4] Martinez and Benavente. The ar face database. In *CVC Technical Report*, 1998.

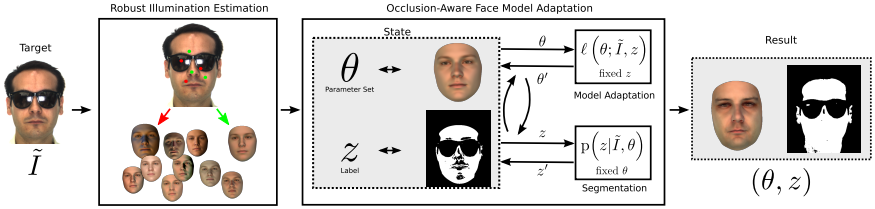


Figure 1: Algorithm overview: First we perform a RANSAC-like robust illumination estimation for initialisation of the segmentation label z and the illumination setting. Then our face model and the segmentation are simultaneously adapted to the target image $\tilde{I}$. The result is a set of face model parameters θ and a segmentation into face and non-face regions. The presented target image is from the AR face database [4].

Next-Best Stereo: Extending Next-Best View Optimisation For Collaborative Sensors

Oscar Mendez, Simon Hadfield[1]
(O.Mendez,S.Hadfield)@surrey.ac.uk
Nicolas Pugeault[2]
N.Pugeault@exeter.ac.uk
Richard Bowden[1]
R.Bowden@surrey.ac.uk

[1] CVSSP
University of Surrey
Guildford, UK

[2] Department of Computer Science
University of Exeter
Exeter, UK

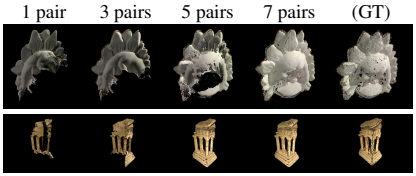


Figure 1: Reconstruction using autonomous stereo pair selection on the Middlebury Dataset.

Reconstruction algorithms that are capable of selecting data to maximise performance, while reducing computational time are necessary to perform reconstruction in the real world. This work proposes an approach to intelligently filter large amounts of data for 3D reconstructions of unknown scenes using monocular cameras. Fig. 1 shows how the reconstruction progresses with a limited number of views. We can achieve state-of-the-art results using as little as 3.8% of the views on the Middlebury dataset. Furthermore, view selection is efficient, taking only 1.1ms per pose pair.

We first present a novel criterion for Next-Best View (NBV) optimisation based on a compromise between the competing objectives of coverage and accuracy. The coverage objective will drive the system to collect views of previously unobserved parts of the scene (e.g., due to restrictions on the field of view or occlusion), whereas the accuracy objective will drive the system to choose the next pose to reduce point cloud uncertainty. These two criteria are optimised jointly using parameter γ. Fig. 2 illustrates how a γ of 1 will give the highest cost to unobserved voxels, preferring to reduce the uncertainty of observed voxels, while 0 will give them the lowest, preferring exploratory behaviour.

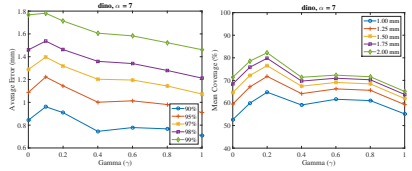


Figure 2: Average Error (Left) and Average Coverage (Right) with different values of γ.

When there are multiple collaborating sensors available, we can extend NBV to also optimise the stereo arrangement of the sensors. This can be achieved by selecting another view, with respect to the NBV, to create the stereo pair with the best possible vergence and baseline.

	Thresh.	[1]	[1]	[2]	Prop.
# Frames	-	41	41	?	**26**
Err(mm)	80%	0.64	0.59	0.64	**0.53**
	90%	1.00	0.88	0.91	**0.74**
	99%	2.86	2.08	1.89	**1.68**
Cvg(%)	0.75mm	79.5	82.9	72.9	**87.3**
	1.25mm	90.2	93.0	73.8	**96.4**
	1.75mm	94.3	96.9	73.9	**98.4**

Table 1: Middlebury Evaluation for different NBV and MVS approaches.

Tab. 1 shows experimental evaluation against the Middlebury benchmark. The proposed method allows efficient selection of stereo pairs for reconstruction, such that a dense model can be obtained with only a small number of images. Once a complete model has been obtained, the remaining computational budget is used to intelligently refine areas of uncertainty, achieving results comparable to state-of-the-art batch approaches on the Middlebury dataset, using as little as 3.8% of the views.

Both contributions are extremely efficient, taking 0.8ms and 0.3ms per pose, respectively. More importantly, neither uses any image-based information instead relying on cues from the partially reconstructed geometry. This allows the proposed approach to sample areas of space that have not been imaged, and is therefore inherently applicable to robotic problems such as path-planning and goal estimation.

[1] Alexander Hornung, Boyi Zeng, and Leif Kobbelt. Image selection for improved multi-view stereo. In *CVPR*, 2008.

[2] Michal Jancosek, Alexander Shekhovtsov, and Tomas Pajdla. Scalable multi-view stereo. In *ICCV*, 2009.

Tuesday
16:20-16:40

Shape-based Image Correspondence

Berk Sevilmis
berk_sevilmis@brown.edu
Benjamin B. Kimia
benjamin_kimia@brown.edu

LEMS
Brown University
Providence, RI 02912 USA

Many of the computer vision tasks such as stereo correspondence, optical flow, biometric user verification, and object recognition require the establishment of dense pixel correspondences between pair of images which can differ in image acquisition setting, *i.e.,* scene content and scene configuration. On the one end of the spectrum is the narrow-baseline stereo correspondence, where these variations are at a minimum since the same 3D scene is captured from slightly different viewpoints. On the other extreme is the *semantic image alignment*, involving images captured from different 3D scenes sharing similar characteristics such as containing same but different instances of objects.

Recent state-of-the-art approaches [2, 5, 7, 8, 10] attempt to compute correspondences between pair of images by matching image signatures, *e.g.,* color histograms, SIFT descriptor [9], CNN features [6], extracted locally from pixels and enforce smoothness on the correspondence field by enforcing spatial regularity. This type of a variational approach is challenged by semantically related images featuring large visual variations given that the variation measure does not capture any semantic aspect of the scene beyond a local histogram over a neighborhood.

Our approach is to introduce certain *semantic concepts* into the correspondence process. Specifically, in this paper, we explore the effect of shape as an additional guideline to the variational correspondence process. We ask whether specifying a pair of corresponding shapes can influence the correspondence process significantly and under what scenarios. We also ask whether shape should be specified in the form of a contour fragment or in the form of a closed curve bounding a region. Finally, when such corresponding shape constraints are not available, we ask whether object proposals can serve this purpose and under what conditions.

In our experiments, we consider three types of tasks: *(i)* optical flow, *(ii)* wide-baseline stereo correspondence, *(iii)* semantic image alignment and use four publicly available datasets, *i.e.,* the MPI Sintel Flow dataset [3], the DTU Robot Image datasets [1], the CUB-200-2011 dataset [11], and the PASCAL-Part dataset [4].

The qualitative and quantitative experiments reveal that *(i)* for datasets depicting slight visual variations, traditional methods are effective and do not benefit from the introduction of a shape correspondence constraint; *(ii)* for datasets depicting large visual variation with the same scene context, the shape correspondence constraints improve the correspondence in the range of 7%; and *(iii)* for datasets depicting instance and configuration variation, there are significant improvements up to 170%. Shape seems to help bring pixels into proper registration. The experiments on the form of the shape constraining the correspondence show that closed curves generally perform better than contour fragments which in turn perform better than shape presented as an unorganized cloud of points. Moreover, the use of object proposals to automatically obtain a shape constraint is also very promising. Compared to the performance obtained when ground truth segmentations are used, a ~26% drop in segmentation accuracy in terms of Jaccard index leads to a ~10% drop in performance.

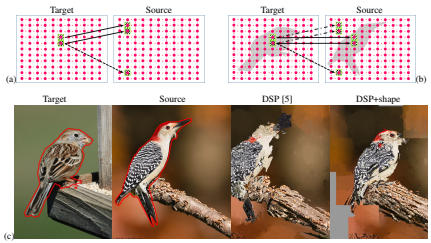


Figure 1: **Shape aligned dense correspondence.** (a) Spatial regularity in current state-of-the-art methods only disambiguate matches which are not locally consistent, *i.e.,* preferring the solid line correspondence to dashed one. (b) Shape alignment can reduce the ambiguity further by ruling out correspondences which violate inside-outside consistency. (c) A visual result. The warped source using shape alignment constraint is clearly superior.

[1] H. Aanæs, A.L. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *IJCV*, 97:18–35, 2012.

[2] Connelly Barnes, Eli Shechtman, Dan B. Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*, pages 29–43, 2010.

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pages 611–625, 2012.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[5] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *CVPR*, pages 2307–2314, 2013.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[7] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5): 978–994, 2011.

[8] Jonathan Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *NIPS*, pages 1601–1609, 2014.

[9] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[10] Weichao Qiu, Xinggang Wang, Xiang Bai, Alan L. Yuille, and Zhuowen Tu. Scale-space SIFT flow. In *WACV*, pages 1112–1119, 2014.

[11] 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, 2011.

Reprojection Flow for Image Registration Across Seasons

Shane Griffith[1,2,3]
sgriffith7@gatech.edu
Cédric Pradalier[2,3]
cedric.pradalier@georgiatech-metz.fr

[1] College of Computing,
Georgia Institute of Technology,
Atlanta, USA

[2] GeorgiaTech Lorraine,
Metz, France

[3] CNRS UMI 2958 GT-CNRS,
Metz, France

We address the problem of robust visual data association across seasons and viewpoints. The predominant methods in this area are typically appearance–based, which lose representational power in outdoor and natural environments that have significant variation in appearance. After a natural environment is surveyed multiple times, we recover its 3D structure in a map, which provides the basis for robust data association. Our approach is called Reprojection Flow (see Fig. 1).

A map can make robust data association possible, but acquiring one that is composed of landmarks from different seasons is a feat in and of itself. First, images are registered (low-res) between near–time surveys to identify images of the same scenes (aided by GPS). Full resolution image registration is performed on the set that aligns well in order to acquire inter-survey observations of KLT–tracked landmarks. A map is recovered from the set of intra- and inter-survey landmark observations using visual SLAM.

Given the optimized map and camera poses, reprojected map points are used for 1) appearance–invariant viewpoint selection and 2) the robust registration of images. First, images of the same scenes from multiple surveys are found by maximizing the co-visibility of reprojected map points. Second, the pixel locations of reprojected map points are used to indicate correspondences between them. This *reprojection flow* directly provides sparse data association among images of the same scenes, which is applied with matching constraints to maximize the use of appearance–invariant information.

We evaluated this approach using a dataset of 24 surveys of a natural environment that span over a year. This approach significantly improves dense correspondence across seasons compared to SIFT Flow [1]. It also provides robustness to changes in viewpoint.

[1] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5):978–994, 2011.

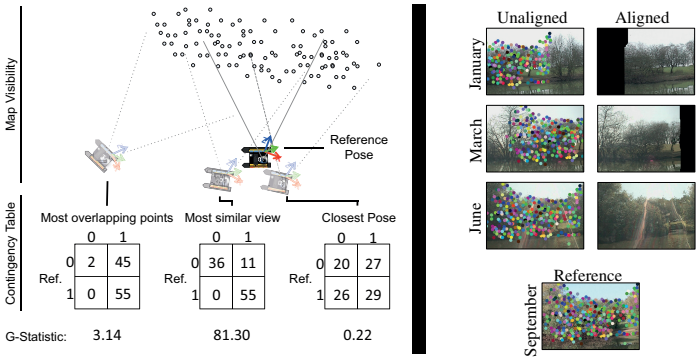


Figure 1: Reprojection Flow: **left)** Finding the most similar view by maximizing co-visibility using the G-statistic, compared to a closest pose heuristic, and a heuristic to maximize the number of overlapping points. The contingency tables are shown for each case. **right)** Using reprojected map points to guide image registration. The KLT points of the reference image are shown projected onto the unaligned images.

Unsupervised Learning of Shape-Motion Patterns for Objects in Urban Street Scenes

Tuesday
17:00-17:20

Dirk Klostermann
dirk.klostermann@rwth-aachen.de
Aljoša Ošep
osep@vision.rwth-aachen.de
Jörg Stückler
stueckler@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Computer Vision Group,
Visual Computing Institute,
RWTH Aachen University
Aachen, Germany

Analysing and predicting the movement of objects is a vital ability for self-driving cars and autonomous mobile robots. We propose an unsupervised approach to learn typical motion patterns of object categories from example data. In our approach, object categories are not limited to predefined classes. Instead, we categorize objects by similarity in shape and trajectory.

Maneuver-based methods (e.g. [2]) such as ours find patterns in previously observed trajectories to predict the future evolution of the trajectory. Compared to previous maneuver-based methods, we additionally distinguish the observed objects by their shape, which allows us to assign object category-specific motion patterns. While most related work in the research area of object categorization focuses on supervised methods (e.g. [1]), only a small number of approaches tackle the semi-/unsupervised categorization of objects.

Based on noisy stereo data we cluster objects based on shape and trajectory information in an unsupervised way. We propose to describe objects in a hierarchical approach, as visualized in Fig. 1: (1.) First, we run a tracker on the training set and gather training data by sampling the trajectories of the tracked objects. A training example contains a shape model of the object together with its trajectory. (2.) We cluster the instances from the training set based on their shape. The shape clusters $c_S \in \mathcal{C}_S$ represent categories of objects, differentiating view points within an object class. (3.) Various instances in each shape cluster can have different motion models, e.g. a car can be parked or drive with a large velocity. Hence, we cluster the trajectories $m \in \mathcal{M}(c_S)$ of each shape cluster c_S to obtain shape-specific trajectory clusters $c_{M|c_S} \in \mathcal{C}_{M|c_S}$. (4.) Each trajectory cluster in a shape forms one shape-motion pattern (SMP)

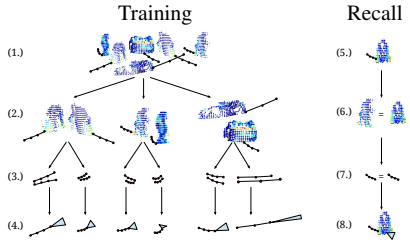


Figure 1: Motion prediction pipeline based on hierarchical clustering.

$p = (c_S, c_{M|c_S})$. The trajectories in the SMP can be described by the Gaussian distribution on the past and future positions relative to their current position. (5.)-(7.) Using our learned model, we classify novel object shapes and trajectories into one of the learned SMPs. (8.) Based on the selected motion model we predict the future motion and infer a probability distribution describing where the object could move based on its shape.

We demonstrate in our experiments that our approach outperforms Kalman filter based predictions and that reasoning on shape can increase prediction performance especially for object categories *e.g.* pedestrians whose trajectory models differ significantly from the typical trajectories of other observed objects. In addition, we demonstrate qualitatively that our method can predict possible future motions for static objects and thus foresee potentially dangerous situations.

[1] D. Mitzel, J. Diesel, A. Ošep, U. Rafi, and B. Leibe. A fixed-dimensional 3d shape representation for matching partially observed objects in street scenes. In *ICRA*, 2015.

[2] V. Romero-Cano, J. Nieto, G. Agamennoni, et al. Unsupervised motion learning from a moving platform. In *Intel. Vehicles Symp.*, 2013.

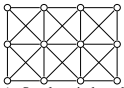
Efficient Learning for Discriminative Segmentation with Supermodular Losses

Jiaqian Yu
jiaqian.yu@centralesupelec.fr

CentraleSupélec
Université Paris-Saclay & Inria
Châtenay-Malabry, France

Matthew B. Blaschko
matthew.blaschko@esat.kuleuven.be

Center for Proc. Speech and Images
Dept. Elektrotechniek - ESAT
KU Leuven, Belgium



(a) Our loss is based on an 8-connected neighborhood.

$$E = -\overbrace{\begin{pmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \end{pmatrix}}^{\text{inference pairwise potential}} - \underbrace{\begin{pmatrix} 0 & \gamma \\ 0 & 0 \end{pmatrix}}_{\text{loss pairwise potential}}$$

(b) Pairwise potential construction for an edge with $y^{*k} = +1$ and $y^{*l} = -1$ following Equation (1).

Figure 1: Non-submodularity of the joint loss augmented inference using the same mapping to a set function for inference and loss functions.

Several non-modular loss functions have been considered in the context of image segmentation. These loss functions do not necessarily have the same structure as the segmentation inference algorithm, and in general, we may have to resort to generic submodular minimization algorithms for loss augmented inference. Although these come with polynomial time guarantees, they are not practical to apply to image scale data.

In this work, we first propose a supermodular loss function that is itself optimizable with graph cuts. It counts the number of incorrect pixels plus the number of pairs of neighboring pixels that both have incorrect labels

$$\Delta(y^*, \bar{y}) = \sum_{j=1}^{n} [y^{*j} \neq \bar{y}^j] + \sum_{(k,l)\in\mathcal{E}_\ell} \gamma [y^{*k} \neq \bar{y}^k \wedge y^{*l} \neq \bar{y}^l] \quad (1)$$

where $[\cdot]$ is Iverson bracket, $\mathcal{E}_\ell$ is a loss specific edge set and γ is a positive weight. We may identify this function with a set function to which the argument is the set of mispredicted pixels.

While being incorporated in a joint loss-augmented inference leads to non-submodular potentials, we therefore use the alternating direction method of multipliers (ADMM) based decomposition strategy (Algorithm 1). It consists of alternatingly optimizing the loss function and performing MAP inference, with each process augmented by a quadratic term enforcing the labeling determined by each to converge to the optimum of the sum. In this way, we gain computational efficiency, making new choices of loss functions practical, while simultaneously making the inference algorithm employed during training closer to the test time procedure.

Algorithm 1 ADMM in scaled form for finding a saddle point of the Lagrangian $\mathcal{L}(y_a, y_b, \lambda) = -\langle w, \phi(x, y_a)\rangle - \Delta(y^*, y_b) + \lambda^T(y_a - y_b) + \frac{\rho}{2}\|y_a - y_b\|_2^2$

1: Initialization $u^0 = 0$
2: **repeat**
3: $y_a^{t+1} = \arg\min_{y_a} -\langle w, \phi(x, y_a)\rangle + \frac{\rho}{2}(\|y_a - y_b^t + u^t\|_2^2)$
4: $y_b^{t+1} = \arg\min_{y_b} -\Delta(y^*, y_b) + \frac{\rho}{2}(\|y_a^{t+1} - y_b + u^t\|_2^2)$
5: $u^{t+1} = u^t + (y_a^{t+1} - y_b^{t+1})$
6: $t = t + 1$
7: **until** stopping criterion satisfied

(a) Hamming (b) 8-connected (c) Hamming (d) 8-connected

(e) Hamming (f) 8-connected (g) Hamming (h) 8-connected

Figure 2: The segmentation results of prediction trained with Hamming loss and our supermodular loss.

We show improvement both in accuracy and computational performance on the MR Grabcut database (Fig. 2) and a brain structure segmentation task, empirically validating the use of a supermodular loss during training and the improved computational properties of the proposed ADMM approach over the Fujishige-Wolfe minimum norm point algorithm. We envision that this can be of use in a wide range of application settings, and an open source general purpose toolbox for this efficient segmentation framework with supermodular losses is available for download from `https://github.com/yjq8812/efficientSegmentation`.

Wednesday
10:00-10:20

Variational Weakly Supervised Gaussian Processes

Melih Kandemir[1]
melih.kandemir@iwr.uni-heidelberg.de
Manuel Haußmann[1]
manuel.haussmann@iwr.uni-heidelberg.de
Ferran Diego[1]
ferran.diego@iwr.uni-heidelberg.de
Kumar Rajamani[2]
KumarThirunellai.Rajamani@in.bosch.com
Jeroen van der Laak[3]
Jeroen.vanderLaak@radboudumc.nl
Fred A. Hamprecht[1]
fred.hamprecht@iwr.uni-heidelberg.de

[1] Heidelberg University, HCI Heidelberg, Germany
[2] Robert Bosch Engineering Bangalore, India
[3] Radboud University Medical Center Nijmegen, Netherlands

Wednesday 10:20-10:40

We introduce the first model to perform weakly supervised learning with Gaussian processes (GPs) on up to millions of instances. The key ingredient to achieve this scalability is to replace the standard assumption of MIL that the bag-level prediction is the maximum of instance-level estimates with the accumulated evidence of instances within a bag. Given data set of N instances $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] = \{\mathbf{X}_1 \cup \mathbf{X}_1 \cup \cdots \cup \mathbf{X}_B\}$ composed of B disjoint partitions, called bags, and supervised by bag labels $\mathbf{T} = [t_1, \cdots, t_B]$, we propose the following model to infer a Bayesian weakly supervised predictor

$$p(\mathbf{u}|\mathbf{Z}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{ZZ}}), \quad (1)$$

$$p(\mathbf{f}|\mathbf{u}, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{f}|\mathbf{K_{XZ}}\mathbf{K_{ZZ}^{-1}}\mathbf{u}, \quad (2)$$
$$diag(\mathbf{K_{XX}} - \mathbf{K_{XZ}}\mathbf{K_{ZZ}^{-1}}\mathbf{K_{ZX}})),$$

$$p(\mathbf{T}|\mathbf{f}) = \prod_{b=1}^{B} \text{Bernoulli}(T_b|\sigma(\mathbf{f}_b^T\mathbf{1})), \quad (3)$$

where $\mathbf{Z} = [\mathbf{z}_1; \cdots; \mathbf{z}_P]$ with $P \ll N$ is a tiny pseudo data set called the *inducing point set* and $\mathbf{u}$ the corresponding outputs. We refer to this model as the *Variational Weakly Supervised Gaussian Process (VWSGP)* [1]. Here, Equations 1 and 2 constitute the sparse GP prior and Equation 3 is a Bernoulli likelihood ensuring the model to predict binary outputs. Thanks to the sum term $\mathbf{f}_b^T\mathbf{1}$ in the likelihood, this model can be trained by closed-form variational inference updates. Hence, keeping all parameters but one fixed, the remaining parameter can be analytically fit to the global optimum. This virtue leads to charmingly fast convergence, fitting perfectly to large-scale learning setups.

We evaluate our VWSGP on the Pascal VOC '07 benchmark and two medical image analysis applications: i) Diabetic Retinopathy screening (DR), and ii) metastatic tumor detection from histopathology images of lymph node tissues (Lymph). While VOC '07 consists of 19M instances (2000 region proposals per image), DR and Lymph have 361K and 1M instances, respectively. The results are summarized in Table 1. Our model proves to outperform various scalable MIL algorithms, as well as state-of-the-art adaptations of deep learning to weakly supervised learning.

Table 1: Bag-level average precision scores on two medical data sets.

	VOC'07	DR	Lymph
VWSGP (Ours)	**83.7**	**0.98**	**0.68**
VGG-S [1]	82.4	-	-
DMIL [4]	75.5	-	-
mi-FV [3]	-	0.92	0.48
e-MIL [2]	-	0.93	0.61

[1] The source code of our model is publicly available under `https://github.com/melihkandemir/vwsgp`

[1] A. Vedaldi et al. Return of the devil in the details: Delving deep into CNNs. In *BMVC*, 2014.
[2] G. Krummenacher et al. Ellipsoidal multiple instance learning. In *ICML*, 2013.
[3] X.-S. Wei et al. Scalable multi-instance learning. In *ICDM*, 2014.
[4] J. Wu et al. Deep multiple instance learning for image classification and auto-annotation. In *CVPR*, 2015.

Regional Gating Neural Networks for Multi-label Image Classification

Rui-Wei Zhao[1]
rwzhao14@fudan.edu.cn
Jianguo Li[2]
jianguo.li@intel.com
Yurong Chen[2]
yurong.chen@intel.com
Jia-Ming Liu[3]
james.liu.n1@gmail.com
Yu-Gang Jiang[1]
ygj@fudan.edu.cn
Xiangyang Xue[1]
xyxue@fudan.edu.cn

[1] Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science,
Fudan University,
Shanghai, China

[2] Intel Labs China
Beijing, China

[3] Department of Control Science and Engineering
Tongji University
Shanghai, China

In this paper we propose a novel deep learning framework named as regional gating neural networks (RGNN) for multi-label image classification. It mainly focuses on integrated contextual object region selection. The motivation arises from the fact that successful global CNN features ignore the underlying context information among different image objects. However, when people attempt to use information from objectness regions, current objectness region proposal algorithms usually produce too many irrelevant or even noisy regions as well. Thus it is meaningful to study how to effectively select useful contextual regions for image classification in the deep architecture.

The proposed RGNN is an end-to-end deep learning framework that can automatically select contextual region features with specially designed gate units, which are then fused for better classification. The feed-forward path of RGNN consists of 5 steps: (1) For each image, object proposals are used to generate multiple candidate regions. (2) Shared Conv + ROI pool + FC layers are then applied to obtain feature representations of regions. (3) Region/feature level gate units are imposed on each regional representation to control whether to be turned on/off so as to select useful contextual region features. (4) Multi-scale cross region pooling are further applied to get contextual image level feature representation. (5) Fused contextual representation are fed into FC layers to predict image labels.

The whole network is optimized with multi-label loss. When object level bounding box annotations are available, we further define a localization loss to aid effective region selection and optimize the network with multi-task learning. Because the gate units and the classifier are integrated in the same deep neural network pipeline, we can learn parameters of the network simultaneously.

Method	VOC'07	VOC'12
VGG-16+19 [1]	89.7	89.3
HCP-VGG [2]	90.9	90.5
HCP++ [2]	-	93.2
RGNN-RL	93.7	93.4
RGNN-FL	93.7	93.3

Table 1: Classification results (AP in %) comparison on VOC'07 and VOC'12 benchmarks.

We evaluate on PASCAL VOC 2007/2012 and MS-COCO benchmarks, and results show that RGNN is superior to existing state-of-the-art methods. Partial comparison results with state of the arts are displayed in Table 1. We can see from these results that our proposed networks with region level gate (RGNN-RL) and feature level gate (RGNN-FL) outperform the global VGG-16+19 networks. Compared to existing algorithms based on objects information like HCP-VGG and HCP++ (with multiple models fusion), RGNNs also work better thanks to the effective integrated contextual region selection in the deep networks. We also find that the introduced localization loss can effectively improve RGNN performances from our ablation studies on VOC 2007 data set.

[1] Simonyan et al. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv.org*, September 2014.

[2] Wei et al. HCP: A Flexible CNN Framework for Multi-label Image Classification. *IEEE TPAMI*, pages 1–8, 2015.

Multispectral Deep Neural Networks for Pedestrian Detection

Jingjing Liu[1]
jl1322@cs.rutgers.edu
Shaoting Zhang[2]
szhang16@uncc.edu
Shu Wang[1]
sw498@cs.rutgers.edu
Dimitris N. Metaxas[1]
dnm@cs.rutgers.edu

[1] Department of Computer Science
Rutgers University
Piscataway, NJ, USA

[2] Department of Computer Science
UNC Charlotte
Charlotte, NC, USA

Multispectral pedestrian detection is essential for around-the-clock applications, *e.g.*, surveillance and autonomous driving. In some sense, color and thermal images provide complementary visual information. As shown in Figure 1, thermal images usually present clear silhouettes of human objects [1], but losing fine visual details of human objects (*e.g.* clothing) which can be captured by RGB cameras (depending on external illumination), Nevertheless, except very recent efforts (*e.g.*,[2]), most of previous studies concentrated on detecting pedestrians with color or thermal images only. It is still unknown how color and thermal image channels can be properly fused in DNNs to achieve the best pedestrian detection synergy.

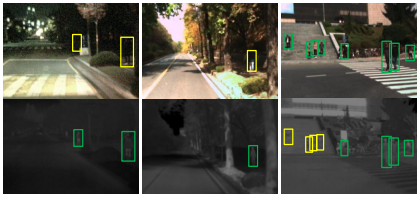


Figure 1: Yellow bounding boxes indicate detection failures with one image channel.

In this paper, we focus on how to make the most of multispectral images (color and thermal) for pedestrian detection. With the recent success of DNNs on generic object detection, it becomes very natural and interesting to exploit the effectiveness of DNNs for multispectral pedestrian detection. We deeply analyze Faster R-CNN [3] for this task and then model it into a convolutional network (ConvNet) fusion problem. We carefully design four distinct ConvNet fusion architectures that integrate two-branch ConvNets on different DNNs stages, *i.e.*, convolutional stages, fully-connected stages, and decision stage, corresponding to information fusion on low level, middle level, high level, and confidence level. All these models outperform the strong baseline detector Faster-RCNN on KAIST multispectral pedestrian dataset (KAIST) [4].

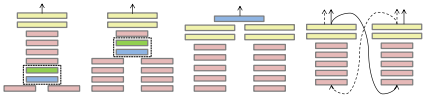


Figure 2: ConvNet fusion models for color and thermal images. From left to right are fusions at low level (Early Fusion), middle level (Halfway Fusion), and high level (Late Fusion), confidence level (Score Fusion), respectively.

We reveal that our Halfway Fusion model – fusion of middle-level convolutional features, provides the best performance on multispectral pedestrian detection. Our Halfway Fusion model significantly reduces the missing rate of baseline method Faster R-CNN by 11%, yielding a 37% overall missing rate on KAIST, which is also 3.5% lower than the other proposed fusion models. We speculate that middle-level convolutional features from color and thermal branches are more compatible in fusion: they contain some semantic meanings and meanwhile do not completely throw all fine visual details.

[1] Y. Socarrás, S Ramos, D. Vázquez, A.M. López, and T. Gevers. Adapting pedestrian detection from synthetic to far infrared images. In *ICCVW*, 2011.

[2] J. Wagner, V Fischer, M. Herman, and S. Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. 2016.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPs*, 2015.

[4] S. Hwang, J. Park, N. Kim, Y. Choi, and I.S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015.

L_1 Graph Based Sparse Model for Label De-noising

Xiaobin Chang
x.chang@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

We aim to learn recognition models from widespread user-provided social media tags, rather than costly purpose created annotations. To address this challenge, we propose a label de-noising algorithm to rectify noisy (incorrect and missing) labels. Subsequent supervised learning tasks then benefit from using the de-noised labels rather than the original noisy ones.

Our model is based on two intuitions: learning the typical noise pattern between observed noisy labels and latent true labels, and exploiting the expected smoothness true labels with regards to the image manifold. Notably, we handle both visual and label outliers with robust L_1-norm based regularisers. Our L_1 Graph based Sparse model with explicit noise pattern model (L_1GSP) is shown in Eq. (1), with two key components: the robust L_1 visual similarity graph regulariser ($||S\hat{Y}||_1$) and the robust L_1 label regulariser with explicit label noise pattern modelling ($||\hat{Y} - YQ||_1$):

$$\min_{\hat{Y},Q} ||S\hat{Y}||_1 + \gamma||\hat{Y} - YQ||_1 + \frac{\beta}{2}||Q||_F^2, \quad (1)$$

where S encodes the visual similarity graph, Y and $\hat{Y}$ represent observed noisy labels and latent de-noised labels respectively, and Q the learned noise pattern transition matrix. The optimisation of Eq. (1) is non-trivial because the two L_1 norm terms make it significantly harder than the more common case of a single L1 norm. Therefore, multiple stages of alternating optimisation procedures are formulated in order to break it into more tractable sub-problems.

Our experiments apply label de-noising algorithms to train sets and evaluate de-noising performance. The cleaned labels are then used for classifier learning, and performance is evaluated on test sets. L_1GSP achieves better performance than its competitors on both label de-noising and follow-up classification tasks across datasets, as shown in Table 1 and 2. Qualitative label de-noising results are shown in Fig. 1. The first example shows that incorrect labels can be eliminated from the top ranking predictions of our de-noising model. The effectiveness of the proposed model to recover missing labels is illustrated in the second image of Fig. 1. The last image of Fig. 1 shows a failure case using our model, which is mainly due to the unconventional appearance of toys.

		GT	NL	L_2VG	L_2VGLG	RPCA	L_1GSP
Denoising	mAP	-	-	52.21	55.01	56.39	**60.09**
Testing	mAP	71.98	42.34	40.33	41.10	53.54	**58.66**

Table 1: Pascal VOC 2007 de-noising performance and testing performance (mAP, %). GT for Ground-truth; NL for Noisy Labels.

	De-noising		Testing	
	mAPc	mAPi	mAPc	mAPi
GT	-	-	47.76	74.31
NL	-	-	30.07	47.88
L_2VG	52.39	57.45	33.81	48.52
L_2VGLG	53.02	59.68	34.69	49.45
RPCA	48.89	64.10	31.20	54.21
L_1GSP	**58.46**	**66.98**	**35.70**	**57.84**

Table 2: De-noising (left) and testing (right) performance (mAP, %) on NUS-WIDE. GT for Ground-truth; NL for Noisy Labels.

Figure 1: Illustrations of label de-noising results on NUS-WIDE (top 3 scoring of the de-noised labels by L_1GSP are shown). Red indicates incorrect labels, green for missing labels and blue for correct labels. Failure case in red dashed line.

[1] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Ml-mg: Multi-label learning with missing labels using a mixed graph. In *ICCV*, 2015.

[2] Wenxuan Xie, Zhiwu Lu, Yuxin Peng, and Jianguo Xiao. Graph-based multimodal semi-supervised image classification. *Neurocomputing*, 2014.

PatchIt: Self-Supervised Network Weight Initialization for Fine-grained Recognition

Patrick Sudowe
sudowe@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Visual Computing Institute
RWTH Aachen University
Germany

ConvNet training is highly sensitive to initialization of the weights. A widespread approach is to initialize the network with weights trained for a different task, an *auxiliary task*. The ImageNet-based ILSVRC classification task is a very popular choice for this, as it has shown to produce powerful feature representations applicable to a wide variety of tasks. However, this creates a significant entry barrier to exploring non-standard architectures. In this paper, we propose a self-supervised pretraining, the *PatchTask*, to obtain weight initializations for fine-grained recognition problems, such as person attribute recognition, pose estimation, or action recognition. Our pretraining allows us to leverage additional unlabeled data from the same source, which is often readily available, such as detection bounding boxes. We experimentally show that our method outperforms a standard random initialization by a considerable margin and closely matches the ImageNet-based initialization.

The *PatchTask* presented in this paper provides a viable alternative to the popular ImageNet-based pretraining. The core idea is to leverage data from the *same domain* as the target task for pretraining. The pretraining is self-supervised, *i.e.*, it solely relies on automatically generated rather than human annotated labels. We target fine-grained recognition tasks that appear in person analysis applications (*e.g.*, pose estimation, re-identification, action and attribute recognition). Their common aspect is that they make predictions for an object that has been located before (*e.g.*, by a detector). So, we will assume such a specific input domain.

The *PatchTask* idea is inspired by the work of Doersch *et al.* [1], who propose an auxiliary task defined by the spatial layout of pairs of patches. In contrast to their work on general images, we focus on fine-grained recognition, where the input images come from a *restricted data domain* (*i.e.*, bounding boxes showing persons). In this restricted setting, it is feasible to directly predict the original location of single patches (Fig. 1).

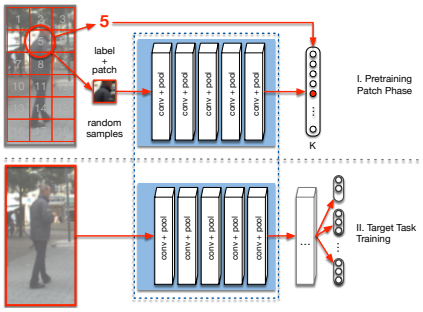


Figure 1: Patch Task: Classify the extraction position given one 32×32 pixel patch. During the pretraining phase, the model needs to encode local patch structure. The parameters are transferred to the target task net. Subsequent fine-tuning benefits from a better initialization.

This paper makes the following contributions: (1) We describe a family of self-supervised *patch tasks* for fine-grained analysis. (2) We demonstrate and evaluate their use for human attribute recognition, where we achieve state-of-the-art performance without using external labels (in particular, without ImageNet). This facilitates further exploration of architectures. (3) We provide data for person analysis pretraining and supporting code that may be used to improve person representations in other ConvNet architectures.

[1] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 2015.

Combining Shape from Shading and Stereo: A Variational Approach for the Joint Estimation of Depth, Illumination and Albedo

Daniel Maurer
maurer@vis.uni-stuttgart.de
Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

Yong Chul Ju
ju@vis.uni-stuttgart.de
Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

Michael Breuß
breuss@tu-cottbus.de
Institute for Applied Mathematics and Scientific Computing, BTU Cottbus-Senftenberg, Germany

Andrés Bruhn
bruhn@vis.uni-stuttgart.de
Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

Shape from shading (SfS) and stereo are two fundamentally different strategies for image-based 3-D reconstruction. While approaches for SfS infer the depth solely from pixel intensities, methods for stereo are based on finding correspondences across images.

In this paper we propose a joint variational method that combines the advantages of both strategies. By integrating recent stereo and SfS models into a single minimisation framework, we obtain an approach that exploits shading information to improve upon the reconstruction quality of robust stereo methods. To this end, we fuse a Lambertian SfS approach with a robust stereo model and supplement the resulting energy functional with a detail-preserving anisotropic second-order smoothness term. Moreover, we extend the novel model in such a way that it jointly estimates depth, albedo and illumination. This in turn makes the approach applicable to objects with non-uniform albedo as well as to scenes with unknown illumination.

Experiments for synthetic and real-world images show the advantages of our combined approach: While the stereo part overcomes the albedo-depth ambiguity inherent to all SfS methods, the SfS part improves the degree of details of the reconstruction compared to pure stereo methods. An example of the reconstruction quality of our combined approach using only two views is given in Figure 1. As one can see, the reconstructed depth is quite detailed. Moreover, the computed illumination direction as well as the estimated albedo look reasonable.

[1] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

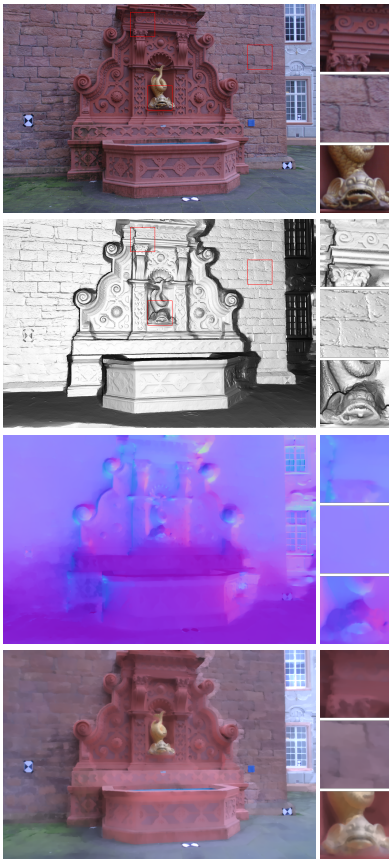


Figure 1: Two-view results for the *Fountain-P11* [1]. **Top to bottom**: Reference image, computed depth (shaded reconstruction), computed illumination direction, computed albedo.

Wednesday
12:20-12:40

Structured Prediction of 3D Human Pose with Deep Neural Networks

Bugra Tekin*[1]
bugra.tekin@epfl.ch
Isinsu Katircioglu*[1]
isinsu.katircioglu@epfl.ch
Mathieu Salzmann[1]
mathieu.salzmann@epfl.ch
Vincent Lepetit[2]
lepetit@icg.tugraz.at
Pascal Fua[1]
pascal.fua@epfl.ch

[1] CVLab
EPFL,
Lausanne, Switzerland

[2] CVARLab
TU Graz,
Graz, Austria

In this paper, we introduce a Deep Learning regression architecture for structured prediction of 3D human pose from monocular images that relies on an overcomplete auto-encoder to learn a high-dimensional latent pose representation and account for joint dependencies.

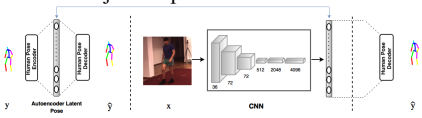


Figure 1: **(a)** An overcomplete denoising autoencoder is trained. **(b)** CNN is mapped into the latent representation learned by the autoencoder. **(c)** The latent representation is mapped back to the original pose space using the decoder.

For this purpose, we first train an overcomplete auto-encoder that projects joint positions to a high dimensional space represented by its middle layer, as depicted by Fig. 1(a). We then learn a CNN-based mapping from the input image to this high-dimensional pose representation as shown in Fig. 1(b). This is inspired by Kernel Dependency Estimation (KDE) [2], which maps both input and output to high-dimensional Hilbert spaces via kernel functions and learns a mapping between these spaces. In fact, it can be understood as replacing kernels by the auto-encoder layers to predict the pose parameters in a high dimensional space that encodes complex dependencies between different body parts. As a result, it enforces implicit constraints on the human pose, preserves the body statistics, and improves prediction accuracy. Finally, as in Fig. 1(c), we connect the decoding layers of the auto-encoder to this network, and fine-tune the whole model for pose estimation. Our contribution is to show that combining traditional CNNs for supervised learning with auto-encoders for structured learning preserves the power of CNNs while also accounting for dependencies, resulting in increased performance.

Results: We evaluate our method on Human3.6m [2] and report our results along with three state-of-the-art approaches [2, 3, 4] in Table 1. Our method outperforms all the baselines.

Following [1], we show in Table 2 the differences between the ground-truth limb ratios and the limb ratios obtained from predictions based on KDE, CNN regression and our approach. These results evidence that our predictions better preserve these limb ratios, and thus better model the dependencies between joints.

Model	Discussion	Eating	Greeting	Taking Photo	Walking	Walking Dog
LinKDE([2]	183.09	132.50	162.27	206.45	97.07	177.84
DconvMP-HML [3]	148.79	104.01	127.17	189.08	77.60	146.59
StructNet-Max [4]	149.09	109.93	136.90	179.92	83.64	147.24
StructNet-Avg [4]	134.13	97.37	122.33	166.15	68.51	132.51
OURS	**129.06**	**91.43**	**121.68**	**162.17**	**65.75**	**130.53**

Table 1: Average Euclidean distance error in mm for [2, 3, 4] and ours.

Model	Lower Body	Upper Body	Full Body
KDE [2]	1.02	7.18	16.43
CNN	0.57	6.86	14.97
OURS no FT	0.62	5.30	11.99
OURS with FT	0.77	5.43	11.90

Table 2: Sum of the log of limb length ratio errors for different parts of the human body.

[1] C. Ionescu, F. Li, and C. Sminchisescu. Latent Structured Models for Human Pose Estimation. In *ICCV*, 2011.

[2] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014.

[3] S. Li and A.B. Chan. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In *ACCV*, 2014.

[4] S. Li, W. Zhang, and A. B. Chan. Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation. In *ICCV*, 2015.

* indicates equal contribution

Multi-task Relative Attributes Prediction by Incorporating Local Context and Global Style Information Features

Yuhang He
www.heyuhang.com
Long Chen
http://www.carlib.net
Jianda Chen
http://www.carlib.net

School of Data and Computer Science
Sun Yat-sen University
Guangzhou, P.R China

Relative attribute represents the correlation degree of one attribute between an image pair. Fine-grained or appearance insensitive relative attribute prediction still remains as a challenging task. To address this challenge, we propose a multi-task trainable deep neural networks by incorporating an object's both local context and global style information to infer the relative attribute. In particular, we leverage convolutional neural networks (CNNs) to extract feature, followed by a ranking network to score the image pair. In CNNs, we treat features arising from intermediate convolution layers and full connection layers in CNNs as local context and global style information, respectively. Our intuition is that local context corresponds to bottom-to-top localised visual difference and global style information records high-level global subtle difference from a top-to-bottom scope between an image pair. We concatenate them together to escalate overall performance of multi-task relative attribute prediction. Finally, experimental results on 5 publicly available datasets demonstrate that our proposed approach outperforms several other state of the art methods and further achieves comparable results when comparing to very deep networks, like 152-ResNet and inception-v3.

0.1 Feature Learning

We propose to learning discriminative feature by incorporating both local context and global style information feature. Local context stores object's local and obvious feature, while global style information stores more abstract and high-level feature. We achieve this by extracting final full connection layer feature and intermediate layer feature, and further concatenate them together to form the final feature vector [1] (see fig.2 for framework pipeline):

$$\psi^i = \psi^i_{fc} + \psi^i_{local} + \psi^i_{global} \quad (1)$$

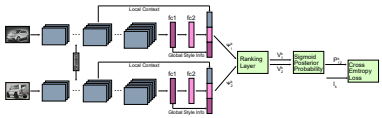


Figure 1: Framework pipeline: we feed the image pair to two CNNs with the same network architecture and shared parameters. Features learned by CNNs intermediate layers and final several full connection layers are concatenated together to form the final feature. The feature pair is further fed to the ranking layer to score each attribute.

0.2 Relative Attribute Prediction via Ranking

The image feature extracted above is mapped to a real value (or a real value vector) through a matrix W and a bias term b: $v = W \cdot \psi + b$. Then We calculate the posterior probability for each relative attribute and squash it between [0-1] via a sigmoid function[2],

$$P^k_{1,2} = \frac{1}{1+e^{-(v^i_1 - v^i_2)}} \quad (2)$$

Finally, we utilise cross entropy loss to rank each relative attribute,

$$\mathcal{L}_i = \sum_{k=1}^{K} l^i_k \log(P^k_{1,2}) - (1-l^i_k)\log(-P^k_{1,2}) \quad (3)$$

0.3 Experiment

see the paper for detailed experiment discussion.

[1] W. Choi F. Yang and Y. Lin. Exploit all layers: fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proc. CVPR*, 2016.

[2] E. Adeli-Mosabbed Y. Souri, E. Noury. Deep relative attributes. *arXiv preprint arXiv:1512.04103*, 2015.

Deep Multi-task Attribute-driven Ranking for Fine-grained Sketch-based Image Retrieval

Jifei Song[1]
j.song@qmul.ac.uk
Yi-Zhe Song[1]
yizhe.song@qmul.ac.uk
Tao Xiang[1]
t.xiang@qmul.ac.uk
Timothy Hospedales[1]
t.hospedales@qmul.ac.uk
Xiang Ruan[2]
ruanxiang@gmail.com

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

[2] TIWAKI Corporation, Ltd.
Japan

Wednesday
10:20-10:40

With touch-screen devices becoming ever more ubiquitous, sketch holds great promise as an intuitive and efficient mode of input compared to classic alternatives. This has motivated a major revival of interest in vision-based analysis of sketches, notably in sketch-based image retrieval (SBIR). Superior to classic SBIR methods, fine-grained SBIR (FG-SBIR) methods [1] are proposed to make fine-grained retrieval in category-level.

In this work, we introduce a multi-task learning (MTL) model for FG-SBIR (as illustrated in Fig. 1), where the main task is a retrieval task with triplet-ranking objective similar to [1], and attributes are detected and exploited in two additional side tasks: The *first* side task is to predict the attributes of the input sketch and photo images. By optimising this task at training, we encourage the learned representation to more meaningfully encode the semantic properties of the photo/sketch; The *second* side-task is to perform retrieval ranking based on the attribute predictions themselves. At test time, this means that the retrieval ordering is explicitly driven by semantic attribute-level similarity as well as the similarity of the internally learned representation. The multi-task loss is formulated as Eq. 1 (full details can be found in our paper).

$$\begin{aligned} L\left(s,p^{+},p^{-}\right) &= L_{\theta}\left(s,p^{+},p^{-}\right)+\lambda_{a}L_{a}\left(s,p^{+},p^{-}\right) \\ &+\lambda_{s}L_{p}\left(s,t^{s}\right)+\lambda_{p^{+}}L_{p}\left(p^{+},t^{p^{+}}\right) \\ &+\lambda_{p^{-}}L_{p}\left(p^{-},t^{p^{-}}\right)+\lambda_{\theta}\|\theta\|_{2}^{2} \end{aligned} \tag{1}$$

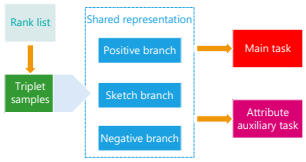


Figure 1: Diagram of the proposed deep multi-task fine-grained SBIR model.

By introducing multiple tasks in the network, the model generalises better and further can rely less on expensive human ranking annotation. Specifically, we show that the highly non-scalable step of triplet annotation required by the model in [1] can now be avoided and an automatic attribute-based strategy is developed instead to focus on the most informative 'hard' training samples for more efficient learning of the model.

Contributions Our contributions are two-fold: (1) A novel deep MTL model is proposed to exploit two attribute-based auxiliary tasks for learning semantically meaningful and domain-invariant representation for FG-SBIR. (2) A new attribute-based triplet generation and sampling strategy is developed to boost the effectiveness of the deep MTL model.

Experiments Extensive experiments are carried out on two benchmarks and the results demonstrate that the proposed model significantly outperforms the state-of-the-art while simultaneously requiring less costly annotation. Partial results are shown in Table 1 (full comparisons can be found in our paper).

Table 1: Comparative results against state-of-the-art retrieval performance.

Shoe Dataset	top 1	top 10	trip-acc	Chair Dataset	top 1	top 10	trip-acc
Triplet model [1]	39.13%	87.83%	69.49%	Triplet model [1]	69.07%	97.94%	72.30%
Ours	**50.43%**	**91.30%**	**70.59%**	Ours	**78.35%**	**98.97%**	**73.13%**

[1] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016.

The Role of Context Selection in Object Detection

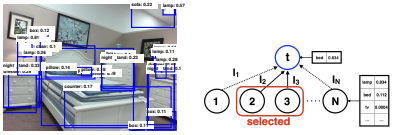
Ruichi (Rich) Yu[1]
richyu@cs.umd.edu
Xi (Stephen) Chen [2]
chnxi@microsoft.com
Vlad I. Morariu[1]
morariu@umiacs.umd.edu
Larry S. Davis[1]
lsd@umiacs.umd.edu

[1] University of Maryland
College Park, MD. USA.

[2] Microsoft Corporation
One Microsoft Way,
Redmond, WA. USA.

We investigate why the utility of context information in object detection is limited through the evaluation of the effect of different pure context cues. We analyze the predictive potential of context in an idealized case where the labels of all contextual objects are known, and only these labels and their relationships to a target object are used to predict the target object label. These experiments reveal that, despite ignoring the appearance of the target object, pure context is effective at predicting the target object class. Not surprisingly, different categories vary in their ability to predict certain target objects. Based on this study, we propose a region-based context re-scoring method with dynamic context selection, illustrated in figure 1(b), which tries to eliminate false positive contextual regions while emphasizing likely true positive and informative ones. Specifically, we introduce a latent variable for each contextual region that determines if that region will be selected to provide context information. In practice, it is intractable to select the optimal set of contextual regions that provide the most trustworthy information when contradictory evidence exists, *for* and *against* the target object being in a certain class. Instead, we decompose the problem by selecting informative regions providing the strongest supporting and refuting evidence independently to compute a *For upper-bound* (FUB) and an *Against upper-bound* (AUB) of the confidence score, and then re-score the confidence for that object being in that class with the difference between the two upper-bounds. The model for computing the two upper-bounds is trained by latent-SVM [1].

The proposed method is evaluated on the SUN RGB-D dataset and achieves 48.25% mean average precision (mAP), an improvement of $\sim$ 2.8% over using object detections without context (45.47%). We also conduct experiments to study the performance of the selection model. Both the simulations on pure context and the real-world experiments using the proposed selection method demonstrate the importance of object-to-object context and the gain attributed to the context selection scheme.



(a) Detection Results (b) Context Selection

Figure 1: (a) Imperfect detections from the Fast R-CNN detector; (b) The proposed context selection method.

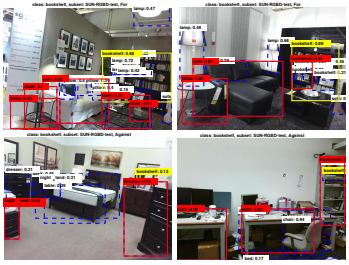


Figure 2: 1st row: FUB model. 2nd row AUB model. The yellow boxes are the target objects, the red boxes are the selected contextual regions, and the blue dashed boxes are the ones that are not selected.

[1] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9): 1627–1645, September 2010.

Wednesday
10:00-10:20

Highly Efficient Regression for Scalable Person Re-Identification

Hanxiao Wang
hanxiao.wang@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk
Tao Xiang
t.xiang@qmul.ac.uk

Vision Group,
School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

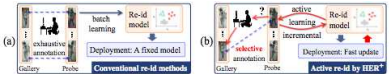


Figure 1: (a) Conventional re-id: A re-id model is trained on a fully labelled training set, then fixed for deployment; (b) Active re-id by HER$^+$: A training set is actively labelled incrementally on-the-fly as a re-id model is incrementally learned, and further updated without re-training during future deployment.

This work is motivated by two very intuitive requirements for a scalable re-id system [2]: (1) Low model complexity with scalable computational cost and memory usage in model training; and (2) High model adaptability supporting fast model update to incorporate any new and increasingly larger data. A Highly Efficient Regression (HER) model is formulated by embedding the Fisher's criterion to a ridge regression model for very fast re-id model learning with scalable memory/storage usage. Importantly, this new HER model supports faster than real-time incremental model updates therefore making real-time active learning feasible in re-id with human-in-the-loop (Fig. 1).

Our Highly Efficient Regression (HER) solution for re-id has a very simple and fast closed-form solution, involved with only a set of linear equations. It is readily scalable to large data with many off-the-shelf efficient implementation available. The base HER model for adopts the form of minimising a least mean squared error:

$$\mathbf{P} = \arg\min_{\mathbf{P}} \tfrac{1}{2}\|\mathbf{X}^\top\mathbf{P} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{P}\|_F^2, \qquad (1)$$

where $\mathbf{X} \in \mathbb{R}^{d\times n}$ refers to the labelled data, and $\mathbf{P} \in \mathbb{R}^{d\times k}$ refers to the discriminative projection to be learned. To make the estimated subspace person identity discriminative, FDA [1] criterion are further embedded. Moreover, to incorporate new and increasingly larger data in a real-world, we further introduce an incremental learning formulation HER$^+$, enabling fast model updates without the need for re-training from scratch.

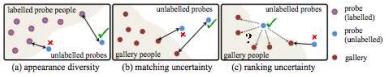


Figure 2: Joint exploration-exploitation criteria for active re-id.

The efficient model updates achieved by HER$^+$ makes makes *active learning re-id* with human-in-the-loop feasible with reduced human labelling costs. A joint exploration-exploitation (*jointE*^{2}) active sampling strategy is further proposed (Fig. 2). Three criterion are considered for selecting most useful samples to maximise the re-id model's discriminative power (1) Appearance diversity exploration, (2) Matching uncertainty exploitation, and (3) Ranking uncertainty exploitation. Finally, these criterion are combined into the final active sampling strategy.

For experimental results, when evaluated under the conventional supervised re-id setting on three popular re-id benchmarks, VIPeR, CUHK01, and CUHK03, HER achieves Rank-1 rates of 45.1%, 68.3% and 60.8% respectively, outperforms all existing competitors. The computational efficiency of HER is also evaluated and it is shown that HER is the fastest in batch training over other state-of-the-art models. When evaluated under the active re-id setting where a model is trained incrementally, it is shown that: (1) HER$^+$ incremental updates is much more efficient than re-training from scratch; and (2) The proposed *jointE*^{2} sampling strategy effectively reduces human labelling effort and achieves better re-id performances.

[1] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

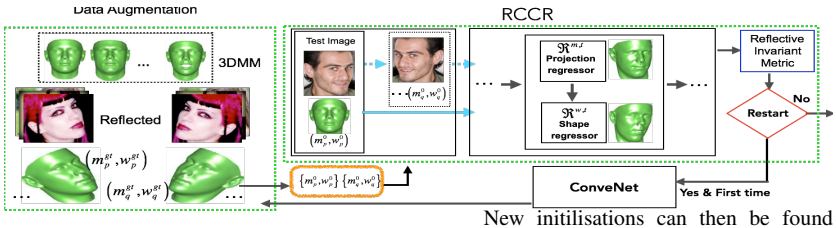
[2] Shaogang Gong, Marco Cristani, Change Loy Chen, and Timothy M. Hospedales. The re-identification challenge. In *Person Re-Identification*. Springer, 2014.

Wednesday
10:20-10:40

Reflective Regression of 2D-3D Face Shape Across Large Pose

Xuhui Jia[1]
xhjia@cs.hku.hk
Heng Yang[2]
yanghengnudt@gmail.com
Xiaolong Zhu[3]
lucienzhu@gmail.com
Zhanghui Kuang[4]
kuangzhanghui@sensetime.com
Yifeng Niu[2]
niuyifeng@nudt.edu.cn
Kwok-Ping Chan[1]
kpchan@cs.hku.hk

[1] The University of Hong Kong
[2] National University of Defense Technology
[3] Tencent Inc.
[4] Sensetime Inc.



In this paper we present a novel reflective method to estimate 2D-3D face shape across large pose. Based on the fact that 2D face image is a projection of 3D face model, we parameterise the configuration of landmarks into 3D Morphable Model and the projection matrix, and regress them in a unified framework. First, two regressors are learned for each cascaded stage, one for predicting the update of camera projection matrix, and the other for 3D shape parameters. They work collaboratively to refine the predicted shape towards true shape; Second, to tackle failures which always occur in large-pose problem, we propose a novel reflective invariant metric to quantitatively estimate the alignments, subsequently the estimation will guide the model whether there is a need to restart the algorithm with different initialization. This is motivated by the fact that CPR are more sensitive to horizontal reflection, and the reflective variance are highly correlated to the misalignment error; Third, instead of using mean shape or random shapes for initialisation [1], we propose a head pose (from a ConveNet estimator) based initilisation scheme, which will relax failure alignments. New initilisations can then be found by searching samples with similar head pose in the training set. The main contributions of this paper are: 1) Large pose face alignment by fitting a dense 3DMM; 2) A novel reflective invariant metric, by investigating the relation between reflective variance and misalignment error; 3) A Reflective Cascaded Collaborative-Regressor algorithm that reduces large pose face alignment failures greatly.

In experiments, we evaluate the effectiveness of our proposed method in component-wise manner on AFLW test set. We compare to 1) RCCR without reflective feedback (CCR). 2) RCCR with reflective feedback and 5 random restart initialisations (RCCR). 3) RCCR with reflective feedback and 5 smart restart initialization (RCCR + SR). The comparison can be found in our paper, which shows, by using the reflective feedback, we achieve big improvement over CCR, which suggests us a failure-alarm mechanism is indeed very useful. Moreover, by using the head pose based initialisations, we achieve even better performance, though the improvement is relatively minor.

[1] Heng Yang and Ioannis Patras. Mirror, mirror on the wall, tell me, is the error small? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition

Oscar Koller[1]
koller@cs.rwth-aachen.de
Sepehr Zargaran[1]
sepehr.zargaran@rwth-aachen.de
Hermann Ney[1]
ney@cs.rwth-aachen.de
Richard Bowden[2]
r.bowden@surrey.ac.uk

[1] Human Language Technology and Pattern Recognition Group
RWTH Aachen University
Aachen, Germany

[2] Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, UK

This paper introduces the end-to-end embedding of a CNN into a HMM, while interpreting the outputs of the CNN in a Bayesian fashion. The hybrid CNN-HMM combines strong discriminative abilities of CNNs with sequence modeling capabilities of HMMs. Most current approaches in the field of gesture and sign language recognition disregard the necessity of dealing with sequence data both for training and evaluation. With our presented end-to-end embedding we are able to improve over the state-of-the-art on three challenging benchmark continuous sign language recognition tasks by between 15% & 38% relative & up to 13.3% absolute.

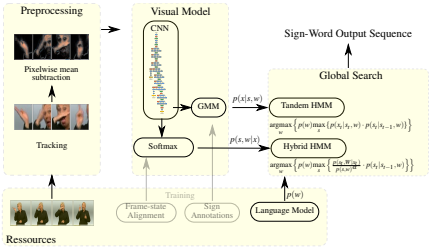


Figure 1: Overview of the proposed CNN-HMM hybrid approach. For clarification the tandem approach is also depicted.

Gesture is a key part in human communication. However, it does not have a well defined structure. Sign language on the other hand provides a clear framework with a defined inventory and grammatical rules that govern joint expression by hand (movement, shape, orientation, place of articulation) and by face (eye gaze, eye brows, mouth, head orientation). This makes sign languages a perfect test bed for computer vision and human language modeling algorithms targeting human computer interaction and gesture recognition.

Following the recent popularity of CNNs in computer vision, several works have made use of it in gesture and sign language recognition. However, in most previous CNN-based approaches the temporal domain is not elegantly taken into consideration. Most approaches use a sliding window or simply evaluate the output on the frame level. We present a hybrid modeling scheme that incorporates a CNN into a HMM. Inspired by the hybrid approach known from speech recognition [1], we use the CNN to model the posterior probability $p(s|x)$ for a hidden state s given the input image x. In this way only the CNN needs to be trained. Opposed to previous works combining CNNs with HMMs, we convert the posteriors into scaled likelihoods using Bayes' rule such that they neatly integrate into the HMM-framework.

We make several contributions:

1. We are the first to embed a deep CNN in a HMM framework in the context of sign language and gesture recognition, while treating the outputs of the CNN as true Bayesian posteriors and training the system as a hybrid CNN-HMM in an end-to-end fashion.

2. We present a large relative improvement of over 15% compared to the state-of-the-art on three challenging standard benchmark continuous sign language data sets.

3. We analyse the impact of the alignment quality on the hybrid performance & experimentally compare the hybrid & tandem approach, which has not been done in the domain of gesture before.

[1] Herve A. Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 1994.

Wednesday
10:20-10:40

Measuring the effect of nuisance variables on classifiers

Alhussein Fawzi
alhussein.fawzi@epfl.ch
Pascal Frossard
pascal.frossard@epfl.ch

Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de
Lausanne (EPFL)
Lausanne, Switzerland

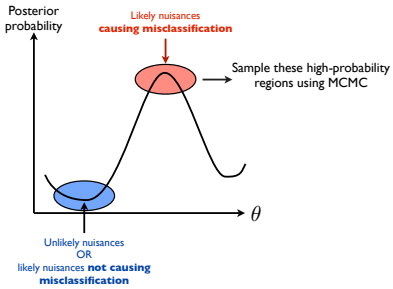
In real-world classification problems, *nuisance* can cause wild variability in the data. Nuisance corresponds for example to geometric distortions of the image, occlusions, illumination changes or any other deformations that do not alter the ground truth label of the image. We propose a probabilistic framework for efficiently *estimating* the robustness of state-of-the-art classifiers and for *sampling* problematic nuisances.

Ingredients.
Classifier: The classifier is provided through its conditional distribution $p_{\text{cl}}(c|x)$, which represents the probability that an image x is classified as c by the classifier.
Nuisance: We denote by $p_{\mathcal{T}}(\theta)$ a prior probability distribution on the nuisance set $\mathcal{T}$. The prior captures the region of interest in the nuisance space.

Measuring the robustness to nuisance.
We define the robustness $\mu_{\mathcal{T}}(x)$ as the average confidence of the classifier on the transformed samples: $\mu_{\mathcal{T}}(x) := \mathbb{E}_{\theta\sim p_{\mathcal{T}}}\left[p_{\text{cl}}(\ell(x)|T_\theta x)\right]$, where $\ell(x)$ is the ground truth label of x, $T_\theta x$ is the image x transformed by θ and where $p_{\text{cl}}(\ell(x)|T_\theta x)$ represents the probability that the transformed image $T_\theta x$ is classified as $\ell(x)$.
A *global* robustness measure $\rho_{\mathcal{T}}$ is then computed by averaging $\mu_{\mathcal{T}}(x)$ over a data distribution $x \sim p_d$. We estimate the average robustness $\rho_{\mathcal{T}}$ using a Monte-Carlo approximation.

Sampling problematic nuisances.
While $\rho_{\mathcal{T}}$ measures the *average* likelihood of the classifier, it is also crucial to visualize the *problematic* regions of the nuisance space where the classifier has low confidence on transformed images.
The problematic regions are mathematically described by a posterior distribution. Sampling from this distribution allows us to visualize the likely "weak spots" of the classifier.

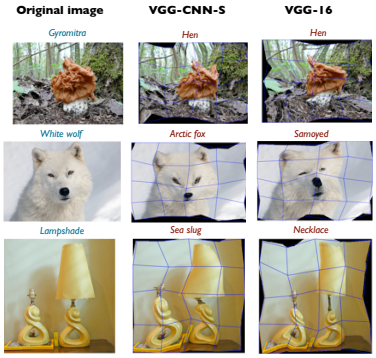
Illustrative experiments.
We evaluate the robustness to affine transformations of different CNN architectures, and show



that:

- Deeper networks are more robust to nuisances,
- While dropout leads to significant improvements in test accuracy, it has no effect on the robustness,
- Data augmentation and spatial transformers can lead to a quantitatively significant boost of the robustness.

We use our sampling method to reveal the problematic nuisance vectors. The following figure shows examples of transformed images that misclassify state-of-the-art networks trained on ImageNet, when the nuisance set is the set of piecewise affine transformations.



Thursday
9:00-9:20

Learning Robust Graph Regularisation for Subspace Clustering

Elyor Kodirov, Tao Xiang
{e.kodirov,t.xiang}@qmul.ac.uk
Zhenyong Fu, Shaogang Gong
{z.fu,s.gong}@qmul.ac.uk

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

Abstract. Various subspace clustering methods have benefited from introducing a graph regularisation term in their objective functions [2]. In this work, we identify two critical limitations of the graph regularisation term employed in existing subspace clustering models and provide solutions for both of them. First, the squared l_2-norm used in the existing term is replaced by a l_1-norm term to make the regularisation term more robust against outlying data samples and noise. Solving l_1 optimisation problems is notoriously expensive and a new formulation and an efficient algorithm are provided to make our model tractable. Second, instead of assuming that the graph topology and weights are known a priori and fixed during learning, we propose to learn the graph [1] and integrate the graph learning into the proposed l_1-norm graph regularised optimisation problem. Extensive experiments were conducted on five benchmark datasets.

Methodology. To address the aforementioned problems, we propose following objective function:

$$\min_{\mathbf{D},\mathbf{Y},\mathbf{W}} \quad \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{Y}\|_F^2+\lambda_1\|\mathbf{Y}\|_1+\lambda_2\|\mathbf{Y}\mathbf{A}_\mathbf{W}\|_1+\lambda_3\|\mathbf{W}\|_F^2$$
$$s.t. \quad \|\mathbf{d}_i\|^2 \leq 1, \mathbf{W}^\mathbf{T}\mathbf{1}=1, \mathbf{W}\geq 0. \tag{1}$$

where $\mathbf{X}\in\mathbb{R}^{r\times N}$ is a data matrix with N r-dimensional data feature vectors as columns, $\mathbf{D}\in\mathbb{R}^{r\times d}$ is a dictionary with d number of atoms, $\mathbf{W}$ is an affinity matrix that captures the topology of the data, $\mathbf{A}_\mathbf{W}$ is a matrix that is obtained by applying eigendecomposition on $\mathbf{W}$, and $\mathbf{Y}\in\mathbb{R}^{d\times N}$ is a sparse code matrix. In the following, we give explanation for each term:

(1) $\|\mathbf{X}-\mathbf{D}\mathbf{Y}\|_F^2$ is the reconstruction error term evaluating how well a linear combination of the atoms (columns) of the dictionary $\mathbf{D}$, can approximate the data matrix $\mathbf{X}$.

(2) $\lambda_1\|\mathbf{Y}\|_1$ is a sparsity regularisation term on $\mathbf{Y}$, with a weighting factor λ_1 to favour a small number of atoms to be used for the reconstruction.

(3) $\lambda_2\|\mathbf{Y}\mathbf{A}_\mathbf{W}\|_1$ is our proposed robust graph regularisation term. Note that we are using l_1-norm instead of l_2-norm weighted by λ_2.

(4) $\lambda_2\|\mathbf{Y}\mathbf{A}_\mathbf{W}\|_1+\lambda_3\|\mathbf{W}\|_F^2$ is the term with proper constraints ($\mathbf{W}^\mathbf{T}\mathbf{1}=1$ and $\mathbf{W}\geq 0$) for graph learning weighted by λ_2 and λ_3.

The constraints, $\mathbf{W}^\mathbf{T}\mathbf{1}=1$ and $\mathbf{W}\geq 0$, are there to ensure the validity of the learned graph, while the constraint $\|\mathbf{d}_i\|^2\leq 1$ ($\mathbf{d}_i$ is a column of $\mathbf{D}$ with $i=1,\ \ldots\ ,r$) enforces the learned dictionary atoms to be compact.

Remark. Terms (3) and (4) are robust graph regularisation and graph learning terms, while first two terms, (1) and (2) constitute the conventional objective function of dictionary learning.

Optimisation. To solve the objective Eq. (1), we develop an algorithm based on ADMM.

Experiments. Table 1 shows experiments on benchmark datasets which are CMU-PIE (C-PIE for short), COIL, ORL, Yale, and YaleB. CA stands for clustering accuracy (%).

Table 1: Comparative results on C-PIE, COIL, ORL, Yale, and YaleB. 'G' stands for graph.

		CA (%)				
Methods	**G**	**C-PIE**	**COIL**	**ORL**	**Yale**	**YaleB**
l_1G	No	70.3	67.1	66.8	40.0	48.4
SSC	No	72.1	58.9	55.5	38.7	52.6
LRR	No	71.5	45.1	66.2	46.2	65.0
LSR	No	77.9	56.0	56.5	48.5	62.4
CASS	No	82.6	59.1	68.3	45.6	81.9
GSC	Yes	**100**	80.9	61.5	43.4	74.2
Rl_1G	Yes	89.5	79.4	62.0	41.3	68.5
SMR	Yes	85.4	65.6	57.6	45.3	73.5
NSLRR	Yes	85.1	61.8	55.3	NA	NA
PCAN	Yes	82.5	76.5	49.1	54.4	59.2
LML	Yes	90.2	80.2	46.7	46.7	60.9
SDRAM	Yes	95.6	86.3	70.6	51.8	92.3
Ours	Yes	**100**	**88.1**	**76.3**	**59.6**	**95.2**

References

[1] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 977–986. ACM, 2014.

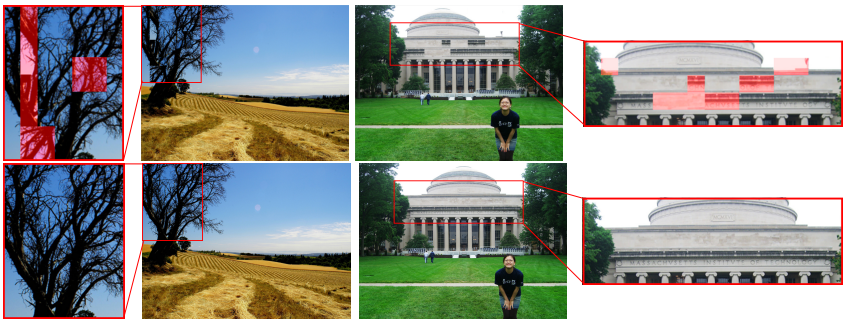
[2] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *Image Processing, IEEE Transactions on*, 20(5):1327–1336, 2011.

Thursdday
9:20-9:40

Solving Jigsaw Puzzles with Linear Programming

Rui Yu
R.Yu@cs.ucl.ac.uk
Chris Russell
CRussell@turing.ac.uk
Lourdes Agapito
L.Agapito@cs.ucl.ac.uk

Dept. of Computer Science
University College London
London, UK



Comparison of Type 2 puzzle reconstruction results between Son *et al.*[2] (top row) and our approach (bottom row). Mistakes are highlighted in red.

We propose a novel Linear Program (LP) based formulation for solving jigsaw puzzles. We formulate jigsaw solving as a set of successive global convex relaxations of the standard NP-hard formulation, that can describe both jigsaws with pieces of unknown position and puzzles of unknown position and orientation. The main contribution and strength of our approach comes from the LP assembly strategy. In contrast to existing greedy methods, our LP solver exploits all the pairwise matches simultaneously, and computes the position of each piece globally. The main advantages of our LP approach include: *(i)* a reduced sensitivity to local minima compared to greedy approaches, since our successive approximations are global and convex and *(ii)* an increased robustness to the presence of mismatches in the pairwise matches due to the use of a weighted L1 penalty. To demonstrate the effectiveness of our approach, we test our algorithm on public jigsaw datasets and show it outperforms state-of-the-art methods.

Starting with an initial set of pairwise matches, the method increasingly builds larger and larger connected components that are consistent with the LP. Our LP formulation naturally addresses the so called Type 1 puzzles[1], where the orientation of each jigsaw piece is given and location of each piece is unknown. However, we show that our approach can be directly extended to the more difficult Type 2 puzzles (see table 1), where the orientation of the pieces is also unknown.

	Direct	Neighbor	Comp.	Perfect
Gallagher [1]	82.2%	90.4%	88.9%	9
Son [2]	94.7%	94.9%	94.6%	12
Ours	**95.6%**	**95.3%**	**95.6%**	**14**

Table 1: Reconstruction performance on Type 2 puzzles from the MIT dataset, Each jigsaw has 432 pieces. Please see paper for the meaning of the scores.

[1] Andrew C Gallagher. Jigsaw puzzles with pieces of unknown orientation. In *CVPR 2012*.

[2] Kilho Son, James Hays, and David B. Cooper. Solving square jigsaw puzzles with loop constraints. In *ECCV 2014*.

Thursday
9:00-9:20

Detecting tracking errors via forecasting

ObaidUllah Khalid[1,2]
o.khalid@qmul.ac.uk
Andrea Cavallaro[1]
a.cavallaro@qmul.ac.uk
Bernhard Rinner[2]
bernhard.rinner@aau.at

[1] Centre for Intelligent Sensing
Queen Mary University of London
London, UK

[2] Inst. of Networked and Embedded Sys.
Alpen-Adria-Universitat Klagenfurt
Klagenfurt, Austria

We propose a framework that detects the failures of a tracker using its output only. The framework is based on a tracker state-background discrimination approach that generates a track quality score, which quantifies the ability of the tracker to remain *on target*.

Let S_t be the region defined by the estimated tracker state x_t in frame I_t at time t. Using motion information $\vec{\mathbf{v}}_{\Delta t_1}$ from a past short temporal window Δt_1 and x_{t-1}, we select the background region $\mathbf{B}_t$ in I_t. We split $\mathbf{B}_t$ into four smaller equally sized regions, b_t^a, each with the same width and height of S_t. We then determine the distribution for S_t, d''_{S_t}, and each of the smaller background regions b_t^a, $d''_{b_t^a}$, using colour distribution fields (DF) [3] (Figure 1). The tracking quality score y_t is determined by averaging the L_1 distances measured between each of the $d''_{b_t^a}$ and d''_{S_t}, where low (high) values of y_t indicate similarity (dissimilarity) between the two regions.

Employing time series analysis, we use the Auto Regressive Moving Average (ARMA) model to forecast future values $\hat{y}_{t+l}$ of $\mathbf{Y} = \{y_t\}_{t=1}^{T}$ over the forecast length $l \geq 1$ at t. We calculate the forecasting error, $|\tilde{e}_{t+l}| = y_{t+l} - \hat{y}_{t+l}$ that highlights the significant changes (tracking errors) and allows the method to detect time instants when a tracking error occurs.

The proposed approach Detecting Tracking Errors via Forecasting (DTEF) is first trained over a dataset D1[1] and then tested using 20 sequences from the Object Tracking Benchmark (OTB)[1] dataset. Using precision (P), recall (R), F-score (F) and false positive rate (FPR), we compare DTEF with two variations of the proposed approach: NAIVE and RAW; one state-of-the-art (SOA) method for tracker error detection [2]: CovF; and two SOA features employed for video tracking [1]: RgbHist and RLHist. Results on the OTB dataset are presented in Table 1. Finally, we demonstrate the flexibility of DTEF via an experimental comparison with the respective SOA methods using baseline tracking results of four trackers and sequences from the VOT2014 challenge.

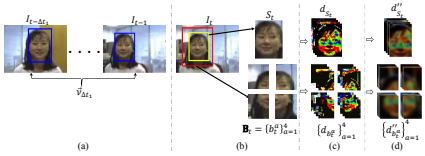


Figure 1: Background and tracker state region selection. (a) $x_{t-\Delta t}$, ..., x_{t-1} (blue bounding boxes) and motion information $\vec{\mathbf{v}}_{\Delta t_1}$ over a past temporal window Δt_1; (b) background region $\mathbf{B}_t$ (red bounding box) and tracker state region S_t (yellow bounding box) selected at frame I_t; (c)-(d) distributions of $\mathbf{B}_t$ and S_t represented with colour DF [3].

	DTEF	NAIVE	RAW	CovF	RgbHist	RLHist
P	.110	.111	**.122**	.087	.083	.078
R	**.714**	.667	.405	**.714**	.595	.667
F	**.191**	.190	.188	.155	.146	.140
FPR	.037	.035	**.019**	.048	.042	.051

Table 1: Performance comparison of tracking error detection over the OTB dataset. Bold font indicates the best result.

[1] J. Ning, L. Zhang, D. Zhang, and C. Wu. Robust object tracking using joint color-texture histogram. *Int. Journal Pattern Recog. Artificial Intell.*, 23(07):1245–1263, Feb 2009.

[2] J.C. SanMiguel and A. Calvo. Covariance-based online validation of video tracking. *IEEE Elec. Lett.*, 51(3):226–228, Feb 2015.

[3] Y. Wang, H. Chen, S. Li, J. Zhang, and C. Gao. Object tracking by color distribution fields with adaptive hierarchical structure. *The Visual Comput.*, pages 1–13, Nov 2015.

[1] http://www.eecs.qmul.ac.uk/~andrea/dtef.html

Thursdday
9:20-9:40

Oracle Performance for Visual Captioning

Li Yao[1]
li.yao@umontreal.ca
Nicolas Ballas[1]
nicolas.ballas@umontreal.ca
Kyunghyun Cho[3]
kyunghyun.cho@nyu.edu
John R. Smith[2]
jsmith@us.ibm.com
Yoshua Bengio[1]
yoshua.bengio@umontreal.ca
Frederico A. Limberger[1]
http://www.cs.york.ac.uk/~fal504
Richard C. Wilson[1]
http://www.cs.york.ac.uk/~wilson

[1] Université de Montréal
[2] IBM T.J. Watson Research
[3] New York University
[4] Dept. of Computer Science
University of York
York, UK

With standard datasets publicly available, such as COCO and Flickr in image captioning, and YouTube2Text, MVAD and MPI-MD in video captioning, the field has been progressing in an astonishing speed. For instance, the state-of-the-art results on COCO image captioning has been improved rapidly from 0.17 to 0.31 in BLEU. Similarly, the benchmark on YouTube2Text has been repeatedly pushed from 0.31 to 0.50 in BLEU score.

While obtaining encouraging results, captioning approaches involve large networks, usually leveraging convolution network for the visual part and recurrent network for the language side. It therefore results model with a certain complexity where the contribution of the different component is not clear.

Instead of proposing better models, the main objective of this work is to develop a method that offers a deeper insight of the strength and the weakness of popular visual captioning models. In particular, we propose a trainable oracle that disentangles the contribution of the visual model from the language model. To obtain such oracle, we follow the assumption that the image and video captioning task may be solved with two steps. Consider the model $P(\mathbf{w}|\mathbf{v})$ where $\mathbf{v}$ refers to usually high dimensional visual inputs, such as representations of an image or a video, and $\mathbf{w}$ refers to a caption, usually a sentence of natural language description. In order to work well, $P(\mathbf{w}|\mathbf{v})$ needs to form higher level visual concept, either explicitly or implicitly, based on $\mathbf{v}$ in the first step, denoted as $P(\mathbf{a}|\mathbf{v})$, followed by a language model that transforms visual concept into a legitimate sentence, denoted by $P(\mathbf{w}|\mathbf{a})$. $\mathbf{a}$ referes to *atoms* that are visually perceivable from $\mathbf{v}$. We define the configuration of $\mathbf{a}$ as an orderless collection of unique atoms. That is, $\mathbf{a}^{(k)} = \{a_1, \ldots, a_k\}$ where k is the size of the bag and all items in the bag are different from each other.

The above assumption suggests an alternative way to build an oracle. In particular, we assume the first step is *close to perfect* in the sense that visual concept (or hints) is observed with almost 100% accuracy. And then we train the best language model conditioned on hints to produce captions.

We consider a simple parametrization of $P(\mathbf{w}|\mathbf{a})$ with Long-short term memory networks (LSTMs) in Hochreiter and Schmidhuber [1]

$$\begin{bmatrix} p(\mathbf{w}_t \mid \mathbf{w}_{<t}, \mathbf{a}^{(k)})) \\ \mathbf{h}_t \\ \mathbf{c}_t \end{bmatrix} = \psi(\mathbf{h}_{t-1}, \mathbf{c}_{t-1}, \mathbf{w}_{t-1}, \mathbf{a}^{(k)}), \tag{1}$$

where $\mathbf{h}_t$ and $\mathbf{c}_t$ represent the RNN state and memory of LSTMs at timestep t respectively.

Despite its simplicity, the proposed model serves as a "performance upper bound" in visual captioning tasks. For the comparison of such oracle models with SOTA, please refer to the paper for details.

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Thursday
9:00-9:20

Beyond Action Recognition: Action Completion in RGB-D Data

Farnoosh Heidarivincheh
farnoosh.heidarivincheh@bristol.ac.uk
Majid Mirmehdi
majid@cs.bris.ac.uk
Dima Damen
dima.damen@bristol.ac.uk

Department of Computer Science
University of Bristol
Bristol, UK

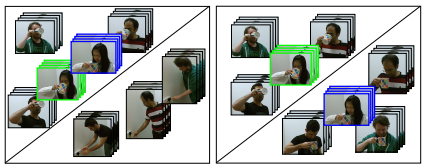
Robust motion representations for action recognition have achieved remarkable performance in both controlled and 'in-the-wild' scenarios. Such representations are primarily assessed for their ability to label a sequence according to some predefined action classes (e.g. *walk, wave, open*). Although increasingly accurate, these classifiers are likely to label a sequence, even if the action has not been fully completed, because the motion observed is similar enough to the training set. Consider the case where one attempts to drink but realises the beverage is too hot. A *drink*-vs-all classifier is likely to recognise this action as positive regardless.

We introduce **action completion** as a step beyond the task of action recognition. It aims to recognise whether the action's goal has been successfully achieved. The notion of completion differs per action and could be infeasible to verify using a visual sensor, however, for many actions, an observer would be able to make the distinction by noticing subtle differences in motion.

We address incompletion in a supervised approach (Fig. 1), on a new dataset that contains 414 complete as well as incomplete sequences, captured using a depth sensor, and spanning 6 actions (*switch, plug, open, pull, pick* and *drink*). For each action, we varied the conditions so the action cannot be completed.

Since the notion of completion differs per action, a general action completion method should investigate the performance of different types of features to accommodate the various action classes. We propose a method that chooses the feature(s) suitable for recognising completion from a pool of depth features using 'leave-one-person-out' cross validation on the training set and automatically selecting the most discriminative feature(s).

We present results on a pool of five features: {Local Occupancy Pattern, Joint Positions, Joint Relative Positions, Joint Relative Angles and Joint Velocities} encoded by the Fourier temporal pyramid.



(a) *drink* vs. *plug* (b) comp. vs. incomp. drink

Figure 1: For a complete *drink* (green) and an incomplete *drink* (blue) sequences from our dataset, both are classified as *drink* when using *drink* vs. *plug* classifier (a). The proposed supervised action completion model (b) identifies the incomplete sequence.

On a sequence of experiments, we show that the various features (i) produce high and comparable % accuracy for action recognition on our dataset, yet (ii) behave differently on *incomplete* action sequences with only some able to distinguish the subtle changes between *complete* and *incomplete* sequences of an action.

By automatic feature selection to build the completion model as a binary classifier, we achieve 95.7% accuracy for recognising action completion across the whole dataset (Fig. 2).

Figure 2: Sample frames of correctly (a), (b) and incorrectly (c) classified test sequences. Dataset is publicly available.

Thursdday
9:20-9:40

Video Stream Retrieval of Unseen Queries using Semantic Memory

Spencer Cappallo
cappallo@uva.nl
Thomas Mensink
tmensink@uva.nl
Cees G. M. Snoek
cgmsnoek@uva.nl

Institute of Informatics
University of Amsterdam
Science Park 904
Amsterdam
The Netherlands

Search among live, user-broadcast videos is an under addressed and increasingly relevant challenge. Every day, more content is shared via services like Meerkat, Periscope, and Twitch. As streaming video becomes more prevalent, it is necessary to develop retrieval systems that can address the unique consequences of live video. In contrast to pre-recorded videos, live streams frequently are transmitted without any accompanying textual description. The nature of streaming video means that even if text is available, there is no guarantee that it will adequately describe the content of a live broadcast. For this reason, the video content itself must be used. The full range of possible future search queries is unknowable, which motivates the framing of stream retrieval as a no-example retrieval problem, where visual examples of a query are assumed to be unavailable beforehand.

We adapt existing approaches from the zero shot classification community, and rely on a word2vec semantic embedding to relate textual queries to pre-trained visual classifier confidence scores [2]. For a given query q, we score a stream with

$$\text{score}(q, x_t) = s(q)^\intercal \phi(x_t)$$

where x_t is the softmax scores of a deep neural network across some set of pre-trained classifiers C, $s(q)$ denotes the semantic similarity between q and C in the semantic embedding, and $\phi(x_t)$ encodes the classifier scores in a sparse manner.

Traditional video tasks assume the *whole* video is available for, a luxury that is not possible in a streaming setting. Also, especially in longer streams, content can change significantly and abruptly throughout the stream. A stream retrieval approach must provide up-to-date representations of the stream content. We explore three ways to emphasize only recent stream content. Two of these methods, Mean Memory Pooling and Max Memory Pooling, perform pooling over a fixed window from the past into the present. We introduce Memory Welling,

$$w(x_t) = \max\left(\frac{m-1}{m} w(x_{t-1}) + \frac{1}{m} x_t - \beta, 0\right)$$

where the current value of the well, w, is built on its previous state, diminished by a memory parameter m, and a constant leaking term β. Memory Wells emphasize recent, reliable content.

We test our approach and competitive baselines on the ActivityNet data set and a motivated subset of the FCVID data set. We synthesize two additional data sets of longer videos through concatanation of random videos. Two tasks are identified and targeted: Instantaneous Retrieval of relevant video streams at one moment, and Continuous Retrieval of streams relevant to a query over a long viewing session. Scoring metrics for both tasks are developed, and videos are evaluated in a simulated streaming setting. To test responsiveness to unseen queries, the test set queries are disjoint from the validation set.

In both target tasks, and on all data sets, either Memory Welling or an adaptation, Max Memory Welling, performs the strongest. We further validate our approach through comparison to state of the art on a traditional, non-streaming video task. Max Memory Welling demonstrates improvement over [1] on zero-shot event retrieval within the TRECVID MED 13 data set, using the setting described in [1].

[1] M. Jain, J. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015.

[2] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *ICLR*, 2014.

Mapping Auto-context Decision Forests to Deep ConvNets for Semantic Segmentation

David L. Richmond*[1]
Dagmar Kainmueller*[1]
Michael Y. Yang[2]
Eugene W. Myers[1]
Carsten Rother[2]

[1] Max Planck Institute of Molecular Cell Biology and Genetics
Dresden, DE

[2] Technical University of Dresden
Dresden, DE

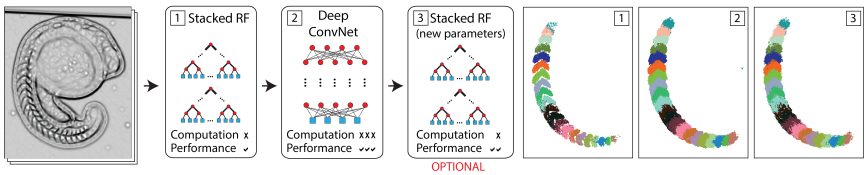


Figure 1: Our method and exemplary results for semantic segmentation of developing zebrafish. (1) A stacked Random Forest is trained, and then (2) mapped to a deep ConvNet and further trained by back-propagation. (3) Optionally, the ConvNet is mapped back to a stacked Random Forest with updated parameters.

In this paper, we propose a mapping from the Auto-context model to a deep Convolutional Neural Network (ConvNet), bridging the gap between these two models, and helping address the challenge of training ConvNets with limited training data.

Auto-context (AC) is a simple model for semantic segmentation that has proven powerful for, e.g., body-pose, facade, and brain segmentation. However, AC is limited by the fact that each classifier is trained greedily. ConvNets on the other hand, benefit from end-to-end training, and have demonstrated remarkable performance when large training data sets are available.

Here we demonstrate that AC can be mapped directly to a ConvNet, and thereby trained end-to-end. This mapping can be seen as an intelligent initialization of the ConvNet, enabling training of large models on limited data. We also describe an approximate mapping of our sparse, deep ConvNet back to a stacked Random Forest with updated parameters, for more computationally efficient evaluation. See Figure 1 for our proposed workflow.

Other works have explored the space between Random Forests and ConvNets [1, 2, 3]. In [2, 3], the authors map a single Random Forest to a shallow Neural Network model. We extend this work to Auto-context and deep ConvNets, and apply it to semantic segmentation.

The mapping that we propose leads to an interesting new ConvNet architecture that, to the best of our knowledge, hasn't previously been explored. One feature of the architecture is that it has large, sparse convolutional kernels. It also avoids max pooling and strided convolutions, which tend to lead to coarse outputs in semantic segmentation. Another result of the initialization from a stack of classifiers, is that specific layers in the ConvNet are directly interpretable as intermediate predictions of the net.

We experimentally verify that the mapping outperforms stacked Random Forests for two different applications: Kinect-based body part labeling from depth images, and somite segmentation in microscopy images of developing zebrafish. By directly visualizing the intermediate prediction layers, we observe that the ConvNet learns to smooth the intermediate predictions, a strategy that was previously developed to improve the performance of stacked classifiers.

[1] Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulo. Deep neural decision forests. In *ICCV*, 2015.

[2] Ishwar Sethi. Entropy nets: from decision trees to neural networks. *Proceedings of the IEEE*, 78 (10):1605–1613, 1990.

[3] Johannes Welbl. Casting random forests as artificial neural networks (and profiting from it). In *GCPR*, 2014.

Thursdday
9:20-9:40

Fully-Trainable Deep Matching

James Thewlis
http://www.robots.ox.ac.uk/~jdt/
Shuai Zheng
http://www.robots.ox.ac.uk/~szheng/
Philip H. S. Torr
http://www.robots.ox.ac.uk/~tvg/
Andrea Vedaldi
http://www.robots.ox.ac.uk/~vedaldi/

Robotics Research Group
Department of Engineering Science
University of Oxford
Oxford, United Kingdom

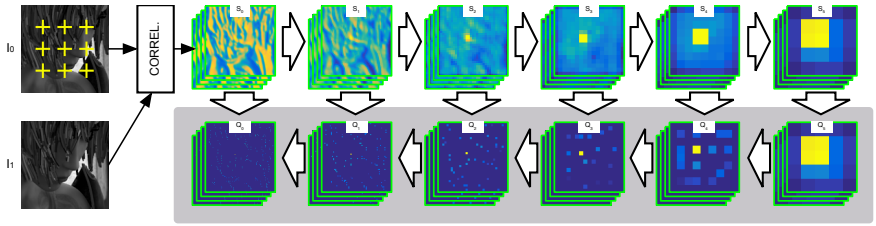


Figure 1: **Fully-Trainable Deep Matching.**

Deep Matching (DM) [3] is a popular method for establishing quasi-dense correspondences between images. An important application of DM is optical flow, where it is used for finding an initial set of image correspondences, which are then interpolated and refined by local optimisation. However, DM, as originally proposed, is *not* a deep neural network and cannot be trained end-to-end via backpropagation.

Our key contribution is to show that the full DM pipeline can be formulated as a Convolutional Neural Network (CNN) with a U-topology (Fig. 1). The fine-to-coarse stage of DM was already given as a CNN in [3]. Here, we complete the construction and show that the DM recursive decoding stage can be implemented by convolutional operators which reverse the ones used in the fine-to-coarse stage. The architecture can be trained using backpropagation, for which we propose a structured-output loss.

We demonstrate the utility of the approach by improving the performance of DM, replacing the HOG image features with the first layers of a pretrained classification CNN. We then train it end-to-end on an image matching task using synthetic data, showing that the accuracy of the matches improves (Table. 1). We further use [2] to interpolate the matches and give dense optical flow, which also sees an improvement.

We compare to FlowNet [1], an alternative end-to-end trainable CNN architecture for image matching, noting the significant advantage of the DM architecture that is further improved by the ability to train from data.

Method	Training	Acc@10 (matches)	EPE (flow)
FlowNet [1]	Flying Chairs	—	4.76
DM-HOG	—	89.39%	3.72
DM-CNN	Pretrained	89.48%	3.63
DM-CNN	Flying Chairs	**90.03%**	**3.50**

Table 1: **Sintel Final training set performance.** DM variants evaluated on Sintel Final training set. EPE (endpoint-error, in pixels) gives mean distance between predicted and correct flow vectors. Acc@10 [3] assigns each pixel a nearby match, measuring the proportion correct within 10px, hence illustrating the quality of the quasi-dense matches.

[1] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE ICCV*, 2015.

[2] Jérôme Revaud, Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *IEEE CVPR*, 2015.

[3] Jérôme Revaud, Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Deepmatching: Hierarchical deformable dense matching. *IJCV*, 2015.

Detection of fast incoming objects with a moving camera

Fabio Poiesi
fabio.poiesi@qmul.ac.uk
Andrea Cavallaro
a.cavallaro@qmul.ac.uk

Centre for Intelligent Sensing
Queen Mary University of London
London, UK

We tackle the problem of detecting fast incoming objects from a moving camera (e.g. on a flying robot) before an impact. We detect these objects using the optical flow computed from an uncalibrated camera without extracting any feature points [1].

We calculate (and compensate) the motion induced by the camera and the time-to-contact (TTC) [3] to infer the position and the closeness of the incoming objects, respectively. We divide the optical flow into a grid of $g \times g$ pixel cells and we detect motion that is dissimilar from the one induced by the moving camera. For each cell j we compute

$$\hat{\alpha}_j = \left(1 + e^{-(m_j - M)}\right)^{-1}, \qquad (1)$$

where m_j is the compensated motion within the cell and M is the 98% percentile of the overall compensated motion in the frame. We use this motion to adaptively learn the background motion model while reducing background motion noise. Unlike [2], our strategy to adaptively learn the background motion adapts to different camera velocities and scene depths.

We merge the optical flow information and use a Bayesian collision avoidance method to locate object-free regions (whose centre is represented as *safe point*) on the image plane. The flying robot can then use this safe point to infer where to go and avoid the object. The likelihood function used by the Bayesian collision avoidance uses measurements from the compensated motion to fit a Gaussian with an active variance that is inversely proportional to the object closeness, which is measured based on the TTC.

Experiments show that our method to detect incoming objects with a moving camera outperforms baselines and alternative state-of-the-art methods. Moreover, our approach to learn the background motion reduces false positive detections.

Our method can be synergistically used with other features and combined with other collision detection methods.

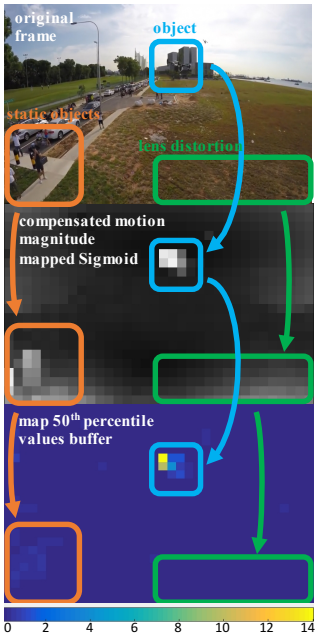


Figure 1: Example of spurious motion removal. (a) Original frame. (b) Magnitude of the compensated motion after mapping via a sigmoid function. (c) Difference from the learnt background indicating the presence of an incoming object.

[1] G. Alenya, A. Negre, and J.L. Crowley. Time to contact for obstacle avoidance. In *European Conference on Mobile Robots*, pages 19–24, Mlini, Croatia, Sep. 2009.

[2] O. Barnich and M. Van Droogenbroeck. ViBE: a universal background subtraction algorithm for video sequences. *Trans. on Image Processing*, 20 (6):1709–1724, Jun. 2011.

[3] Y. Watanabe, F. Sakaue, and J. Sato. Time-to-contact from image intensity. In *Proc. of Computer Vision and Pattern Recognition*, pages 4176–4183, Boston, MA, USA, Jun. 2015.

Thursday
9:20-9:40

Multi-view Multi-illuminant Intrinsic Dataset

Shida Beigpour
shida@mpi-inf.mpg.de
Mai Lan Ha
hamailan@informatik.uni-siegen.de
Sven Kunz
sven.kunz83@gmail.com
Andreas Kolb
andreas.kolb@uni-siegen.de
Volker Blanz
blanz@informatik.uni-siegen.de

Institute for Vision and Graphics
University of Siegen
Siegen, Germany

Decomposing an image into its intrinsic components (e.g. reflectance and shading) is a fundamental concept in computer vision. This paper proposes a novel high-resolution multi-view dataset of complex multi-illuminant scenes with precise reflectance and shading ground-truth as well as raw depth and 3D point cloud.

Our dataset challenges the intrinsic image methods by providing complex coloured cast shadows, highly textured and colourful surfaces, and specularity. This is the first publicly available multi-view real-photo dataset at such complexity with pixel-wise intrinsic ground-truth. Our work improves over the state-of-the-art intrinsic datasets [1, 2]. In the effort to help evaluating different intrinsic image methods, we propose a new perception-inspired metric that is based on the reflectance consistency. We provide the evaluation of three intrinsic image methods using our dataset and metric.

Fig. 1 demonstrates an example of a scene captured with six different cameras under different illumination conditions along with its raw depth, point cloud, and a rough surface reconstruction. Here the advantage of using reflectance instead of the captured pixel colour for the 3D surface colour is evident. In total the dataset consists of 20 illumination conditions, 5 scenes (Fig. 2), and 6 cameras. Our complete dataset consists of 600 high-resolution images along with their ground-truth and is publicly available online at:

http://www.cg.informatik.uni-siegen.de/data/iccv2015/intrinsic/

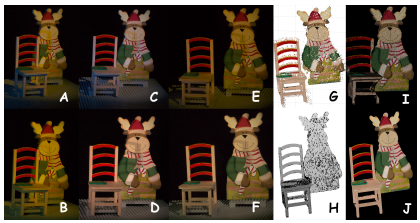


Figure 1: Examples of different views and illuminations (A-F), rough point cloud (G), raw depth (H), 3D surface (I), and ground-truth reflectance (J).

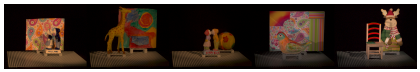


Figure 2: The proposed scenes present complex shapes and coloured textures.

We believe that our dataset and metric can help in improving the quality of intrinsic image methods in complex scenes and lighting conditions. Please refer to our paper for more information and formulas.

[1] R. Grosse et. al. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *IEEE ICCV*, 2009.

[2] Sh. Beigpour et. al. A comprehensive multi-illuminant dataset for benchmarking of the intrinsic image algorithms. In *IEEE ICCV*, 2015.

Tuesday
13:40-14:40

Accurate Closed-form Estimation of Local Affine Transformations Consistent with the Epipolar Geometry

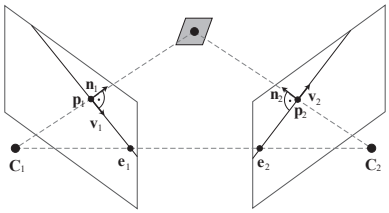
Daniel Barath, Levente Hajder
{barath.daniel,hajder.levente}@sztaki.mta.hu
MTA SZTAKI
Budapest, Hungary

Jiri Matas
matas@cmp.felk.cvut.cz
CMP, Czech Technical University
Prague, Czech Republic

A novel method is proposed for accurate estimation of local affine transformations for a pair of images satisfying the epipolar constraint. The method returns the closest, in least squares sense, affine transformation to an initial estimate consistent with the fundamental matrix.

The contributions of the paper: (i) the introduction of two novel constraints for a local affine transformation making it consistent with the fundamental matrix, and (ii) a method estimating an EG-L_2-Optimal affinity – transformation which is consistent with the epipolar geometry (EG) –, by enforcing the proposed constraints.

An affine correspondence consists of a point pair $\mathbf{p}_1$, $\mathbf{p}_2$ and a local affine transformation $\mathbf{A}$ mapping the neighborhood of the points.



The constraints state that the 2×2 matrix $\mathbf{A}$ transforms the normal $\mathbf{n}_1$ of the epipolar line related to point $\mathbf{p}_1$ into $\beta\mathbf{n}_2$, where $\mathbf{n}_2$ is the normal of the epipolar line related to point $\mathbf{p}_2$ and $\beta \in \mathbb{R}$ is a scalar. This statement is equivalent to $\mathbf{n}_1\mathbf{A}^{-T} = \beta\mathbf{n}_2$. It is proven as well that β is determined by the epipolar geometry.

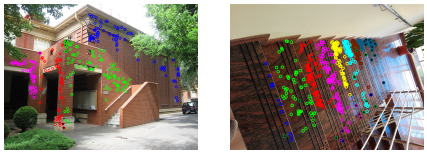
The method requires an affine correspondence $\mathbf{p}_1, \mathbf{p}_2, \mathbf{A}'$, i.e. estimated by an affine-covariant detector. The points $\mathbf{p}_1$ and $\mathbf{p}_2$ are optimally be corrected w.r.t. the epipolar geometry, in least squares sense, by the method of [4]. The proposed technique corrects $\mathbf{A}'$ by simultaneously minimizing $||\mathbf{A} - \mathbf{A}'||_F^2$ and enforcing the introduced constraints using a closed-form approach. It is proven that $||\mathbf{A} - \mathbf{A}'||_F^2$ has both geometric and algebraic interpretations.

The processing time of the method is ≈0.04 ms in C++.

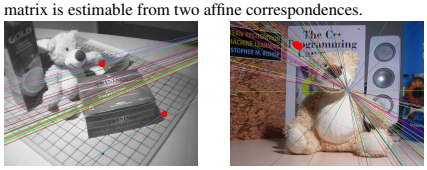
Evaluation. The method is validated on synthetic data and publicly available benchmarks. The corrected affinities are always more accurate than the output of the affine-covariant detector. As a side-effect, the detectors are compared – *the most accurate is the Hessian-Affine* augmented by view-synthesis a la ASIFT.

Conclusions. The algorithm has negligible time demand and always makes the input affinities more accurate. In problems involving local affine transformations in rigid scenes, the proposed method should always be used.

Application 1. Using the proposed results the detection and segmentation of multiple planes becomes more accurate [1].



Application 2. Using equation $\mathbf{n}_1\mathbf{A}^{-T} = \beta\mathbf{n}_2$ the fundamental matrix is estimable from two affine correspondences.



Application 3. Surface normal estimation benefiting from precise affine correspondences [2].

Application 4. Precise affine correspondences significantly improve camera calibration as well as 3D reconstruction [3].

Application 5. In the paper, we use the method to compare the geometric precision of affine-covariant feature detectors.

[1] D. Barath, J. Matas, and L. Hajder. Multi-H: Efficient recovery of tangent planes in stereo images. In *BMVC*, 2016.

[2] D. Barath, J. Molnar, and L. Hajder. Novel methods for estimating surface normals from affine transformations. In *VISIGRAPP Selected Papers*, 2016.

[3] I. Eichhardt and L. Hajder. Improvement of camera calibration using surface normals. In *ICPR*, 2016.

[4] R. I. Hartley and P. Sturm. Triangulation. *CVIU*, 1997.

Recognition of Transitional Action for Short-Term Action Prediction using Discriminative Temporal CNN Feature

Hirokatsu Kataoka[1]
hirokatsu.kataoka@aist.go.jp
Yudai Miyashita[2]
undermmusic@gmail.com
Masaki Hayashi[34]
m.hayashi@liquidinc.asia
Kenji Iwata[1]
kenji.iwata@aist.go.jp
Yutaka Satoh[1]
yu.satou@aist.go.jp

[1] National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan

[2] Tokyo Denki University
Adachi-ku, Tokyo, Japan

[3] Liquid Inc.
Tokyo, Japan

[4] Keio University
Yokohama, Kanagawa, Japan

Transitional actions belong to a class between actions for short-term action prediction (see Figure 1). Early action recognition is necessary for producing action predictions in the early frames of an objective action. Earlier prediction in the initial frames of an objective action is desirable for early action recognition problems, but the solutions depend on the action itself. On one hand, within the setting of a short-term action prediction, understanding a pending human action change is more natural if we have a firm grasp on transitional actions. In a traffic scene, short-term action predictions are particularly crucial for avoiding accidents between humans and vehicles. Figure 1 shows sequential actions that include *Walk straight*, *Walk straight - cross*, and *cross*. Where *Walk straight* and *cross* are conventional action definitions, our proposal adds a transitional action between actions (here *Walk straight - cross*) in order to provide a better action approach to predictions. Our proposed short-term predictions achieve earlier prediction than so-called early activity recognition, since they can recognize a dangerous *cross* action while it is transitional. Intuitively, the recognition difficulty arising from action and transitional action is that they tend to partially overlap each other. We believe that the use of a subtle motion descriptor (SMD) will allow us to identify sensitive differences between actions and transitional actions.

In this paper, we address the recognition of transitional action for short-term action prediction. We also propose a discriminative temporal convolutional neural network (CNN) feature that can be used for recognizing transitional actions in order to overcome the difficulty of indistinguishable feature classification in transitional

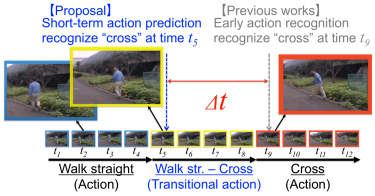


Figure 1: Recognition of transitional actions for short-term action prediction: Identification of transitional actions allow us to understand the next activity at time t_5 before an early action recognition approach at time t_9.

actions. To accomplish this, we employ an SMD that captures subtle differences between consecutive frames. Our paper contains two main contributions: (i) the definition of transitional action for short-term action prediction that achieves earlier prediction than early action recognition, and (ii) identifying CNN-based SMD to create a clear distinctions between action and transitional action. The feature is simply updated from a spatio-temporal CNN feature Pooled Time Series (PoT) proposed in [1].

Our CNN-based SMD demonstrated the best rate of success on three different trial datasets. Even when using the shortest (3-frame) feature accumulation for recognition tuning, we confirmed outstanding results with 85.78% (NTSEL), 69.77% (UTKinect), and 49.93% (Watch-n-Patch) on the three different datasets.

[1] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. CVPR, 2015.

Tuesday
13:40-14:40

Multi-H: Efficient Recovery of Tangent Planes in Stereo Images

Daniel Barath, Levente Hajder
{barath.daniel,hajder.levente}@sztaki.mta.hu
MTA SZTAKI
Budapest, Hungary

Jiri Matas
matas@cmp.felk.cvut.cz
CMP, Czech Technical University
Prague, Czech Republic

Multi-H – an efficient method for the recovery of the tangent planes of a set of point correspondences satisfying the epipolar constraint is proposed. The problem is formulated as a search for a labeling minimizing an energy that includes a data and spatial regularization terms. The number of planes is controlled by a combination of Mean-Shift and α-expansion.

The input of Multi-H are point correspondences with local affine transformations and the epipolar geometry. We use Matching On Demand with view Synthesis (MODS) method [3] since it provides accurate local affinities, the fundamental matrix **F** and point correspondences consistent with **F**. Using Homography from Affine transformation and Fundamental matrix (HAF) method [1] a homography is estimated for every single correspondence.

The alternating minimization stage of the algorithm repeats:

1. Mean-Shift is applied to the density function built on the homographies to reduce the complexity of the problem and to find the modes in the homography space. This procedure assumes that many points have the same tangent plane and these planes form a mode in the space of homographies.
2. α-expansion is applied to the correspondences assigning a label to each. A label is associated with a homography.
3. Least-squares homography re-fitting step uses the HAF method to re-estimate the homographies exploiting the labeling provided by Step 2.

Convergence is reached when both the number of the clusters and the energy remain unchanged. It is guaranteed since the first step does not increase the number of clusters, the others decrease the energy, and the set of labeling is finite.

The speed of Multi-H was measured on two sets consisting of 100 and 500 correspondences. The processing time for the 100 and 500 correspondences were 0.04 and 0.80 sec. on a desktop PC with Intel Core i5-4690 CPU, 3.50 GHz using 4 cores.

Test 1. – Tangent plane estimation is accurately solved by Multi-H as it refines the initial estimates by partitioning the correspondences based on the similarity of their tangents. The figure below shows the recovered surface normals coloured by their labels in two images of fountain-P11 dataset.

The table below shows the improvement in surface normal estimation between selected frames – angular errors in degrees.

Frames	Affine Detector	EG-L_2-Opt	**Multi-H**
1 - 2	35.7°	35.5°	**14.4°**
1 - 5	19.0°	16.7°	**7.0°**
3 - 5	24.9°	23.1°	**9.0°**
5 - 9	20.0°	17.8°	**7.1°**
6 - 8	22.5°	19.9°	**8.8°**

Test 2. – Multiple plane recovery is a long-standing problem [2]. The combination of Multi-H with a compatibility criterion $||H^T F + F^T H||_F > \theta$ leads to results superior to the state-of-the-art multi-plane fitting techniques, where H is a homography, F the fundamental matrix, and θ a threshold.

		mean	median
J-Linkage	(ECCV 2008)	25.50	24.48
SA-RCM	(CVPR 2012)	28.30	29.40
T-Linkage	(CVPR 2014)	24.66	24.53
RPA	(BMVC 2015)	17.20	17.78
Grdy-RansaCov	(CVPR 2016)	26.85	28.77
ILP-RansaCov	(CVPR 2016)	12.91	12.34
Multi-H	(BMVC 2016)	**4.40**	**2.41**

The table above compares the mean and median misclassification errors on the AdelaideRMF dataset. Every algorithm, including Multi-H, has been tuned separately on each image pair to allow comparison with the literature. Results, using a fixed set-up, are shown in the table below.

	T-Linkage	SA-RCM	RPA	**Multi-H**
johnsa	34.28	36.73	10.76	**9.33**
johnsb	24.04	16.46	26.76	**10.14**
ladysymon	24.67	39.50	24.67	**4.49**
neem	25.65	41.45	19.86	**2.00**
old	20.66	21.30	25.25	**1.79**
sene	7.63	20.20	0.42	**0.00**
mean	22.82	29.27	17.95	**4.79**
median	24.36	29.02	22.27	**3.74**

Conclusions. Multi-H is accurate, outperforms state-of-the-art multi-homography fitting techniques for both fixed and per-image parameter setting. In most applications, Multi-H will run significantly faster than the affine-covariant detectors providing the input.

Recovered dominant planes

[1] D. Barath and L. Hajder. Novel ways to estimate homography from local affine transformations. In *VISAPP*, 2016.

[2] H. Isack and Y. Boykov. Energy-based geometric multi-model fitting. *IJCV*, 2012.

[3] D. Mishkin, J. Matas, and M. Perdoch. MODS: Fast and robust method for two-view matching. *CVIU*, 2015.

Jointly Learning Non-negative Projection and Dictionary with Discriminative Graph Constraints for Classification

Weiyang Liu[12]
wyliu@pku.edu.cn
Zhiding Yu[3]
yzhiding@andrew.cmu.edu
Yandong Wen[3]
yandongw@andrew.cmu.edu
Rongmei Lin[4]
rongmei.lin@emory.edu
Meng Yang*[1]
yang.meng@szu.edu.cn

[1] College of Computer Science & Software Engineering,
Shenzhen University, China

[2] School of ECE,
Peking University, China

[3] Dept. of ECE,
Carnegie Mellon University, USA

[4] Dept. of Math & Computer Science,
Emory University, USA

Different from the conventional wide variety of discriminative Dictionary learning (DL) literatures, our work casts an alternative view on this problem. One major purpose of this paper is to jointly learn a feature projection that improves DL. Instead of keep exploiting additional discrimination from the dictionary representation, we consider optimizing the input feature to further improve the learned dictionary. We believe such process can considerably influence the quality of learned dictionary, while a better learned dictionary may directly improve subsequent classification performance.

Given that mid-level object parts are often discriminative for classification, we aim to learn a feature projection that mines these discriminative patterns. It is well-known that non-negative matrix factorization (NMF) [1] can learn similar part-like components. In the light of NMF and projective NMF (PNMF) [2], we consider the projective self-representation (PSR) model where the set of training samples $\boldsymbol{Y}$ is approximately factorized as: $\boldsymbol{Y} \approx \boldsymbol{MPY}$. The model jointly learns both the intermediate basis matrix $\boldsymbol{M}$ and the projection matrix $\boldsymbol{P}$ with non-negativity such that the additive (non-subtractive) combinations leads to learned projected features $\boldsymbol{PY}$ accentuating spatial object parts. In the paper, we propose a novel NMF-like feature projection learning framework on top of the PSR model to simultaneously incorporate label information with discriminative graph constraints. One shall see, our proposed framework can be viewed as a tradeoff between NMF and feature learning [4].

The dictionary representation is further discriminatively learned given the projected input features. An overview of the joint non-negative projection and dictionary learning (JNPDL) framework is illustrated in Fig. 1. The construction of discriminative graph constraints in both non-negative projection and dictionary learning follows the graph embedding framework [3]. While the inputs of graph constraints are essentially the same, they form different regularization terms for the convenience of optimization. Finally, a discriminative reconstruction constraint is also adopted so that coding coefficients will only well represent samples from their own classes but poorly represent samples from other classes. We test JNPDL in both image classification and image set classification with comprehensive evaluations, showing the excellent performance of JNPDL.

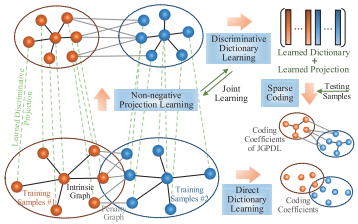


Figure 1. An illustration of JNPDL.

[1] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[2] Xiaobai Liu, Shuicheng Yan, and Hai Jin. Projective non-negative graph embedding. *IEEE T. IP*, 19(5):1126–1137, 2010.

[3] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE T. PAMI*, 29(1):40–51, 2007.

[4] Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng. Deep learning of invariant features via simulated fixations in video. In *NIPS*, 2012.

Tuesday
13:40-14:40

A MultiPath Network for Object Detection

Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O. Pinheiro,
Sam Gross, Soumith Chintala, Piotr Dollár

Facebook AI Research (FAIR)

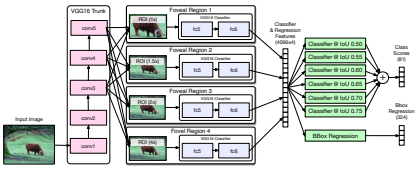


Figure 1: Proposed MultiPath architecture. COCO contains objects at multiple scales, in context and among clutter, and under frequent occlusion. Moreover, the COCO evaluation metric rewards high quality localization. To addresses this, we propose the MultiPath network pictured above, which contains three key modifications: skip connections, foveal regions, and and an integral loss. Together these modifications allow information to flow along multiple paths through the network, enabling the classifier to operate at multiple scales, utilize context effectively, and perform more precise object localization. Our MultiPath network, coupled with DeepMask object proposals [4, 5], achieves major gains on COCO detection.

The recent COCO dataset presents several new challenges for object detection. In particular, it contains objects at a broad range of scales, less prototypical images, and requires more precise localization. To address these challenges, we test three modifications to the standard Fast R-CNN object detector: (1) skip connections that give the detector access to features at multiple network layers, (2) a foveal structure to exploit object context at multiple object resolutions, and (3) an integral loss function and corresponding network adjustment that improve localization.

The result of these modifications is that information can flow along multiple paths in our network, including through features from multiple network layers and from multiple object views. We refer to our modified classifier as a 'MultiPath' network. We couple our MultiPath network with DeepMask object proposals, which are well suited for localization and small objects, and adapt our pipeline to predict segmentation masks in addition to bounding boxes. The combined system improves results over the baseline Fast R-CNN detector with Sel-Search by 66% overall and by 4× on small objects.

Figure 2: Selected detection results on COCO. Only high-scoring detections are shown. While there are missed objects and false positives, many of the detections are quite good.

	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
ResNet [3]	27.9	51.2	27.6	8.6	30.2	45.3
MultiPath	25.0	45.4	24.5	7.2	28.8	39.0
ResNet [3]	37.1	58.8	39.8	17.3	41.5	52.5
MultiPath	33.2	51.9	36.3	13.6	37.2	47.8
ION [1]	30.7	52.9	31.7	11.8	32.8	44.8
Fast R-CNN* [2]	19.3	39.3	19.9	3.5	18.8	34.6
Faster R-CNN* [6]	21.9	42.7	—	—	—	–

Table 1: **Top**: COCO test-standard segmentation results. **Bottom**: COCO test-standard bounding box results (top methods only). Leaderboard snapshot from 01/01/2016. *Note: Fast R-CNN and Faster R-CNN results are on test-dev as reported in [6], but results between splits tend to be quite similar.

Our system placed second in both the COCO 2015 detection and segmentation challenges, without using ResNets. Source code is available.

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural nets. In *CVPR*, 2016.

[2] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *NIPS*, 2015.

[5] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *ECCV*, 2016.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

Fast Eigen Matching

Accelerating Matching and Learning of Eigenspace method

Yusuke Sekikawa, Koichiro Suzuki,
Kosuke Hara, Yuichi Yoshida, Ikuro Sato
DENSO IT Laboratory, Inc. Japan.
{ysekikawa,ksuzuki,khara,yyoshida,isato}@d-itlab.co.jp

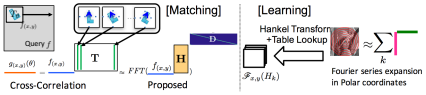


Fig. 1: Overview of the proposed method

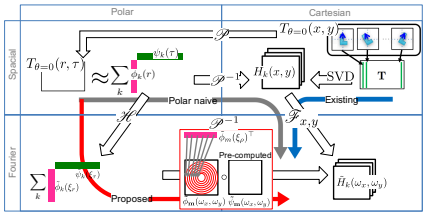


Fig. 2: The learning algorithm.

We propose *Fast Eigen Matching*, a method for accelerating the matching and learning processes of the eigenspace method for rotation invariant template matching (RITM).

Correlation-based template matching is one of the basic techniques used in computer vision. Among them, rotation invariant template matching (RITM), which locates a known template in a query irrespective of the template's translation and orientation, has been widely put to use in many industrial applications. A naive implementation of RITM requires intensive computation since one needs to correlate query f with N rotated templates $\boldsymbol{T}$ (Fig.1 left). Eigenspace methods takes advantage of the fact that a set of correlated images $\boldsymbol{T}$ can be approximately represented by a small set of *eigenimages*. Once *eigenimages* and it's 2D-Fourier transform are computed in learning process, matching process of RITM can be performed very efficiently using these 2D-Fourier transform*ed eigenimages*[1].

It is also important to speedup the learning process, especially for applications such as global robot localization, where a template changes frame by frame and efficient online learning is required. The existing eigenspace methods are not feasible for problem settings of this kind, because it requires a lot of time for generation of rotated templates, SVD and 2D-FFT.

To speed up the matching and the learning process of existing Eigenspace methods, we propose *Fast Eigen Matching* by exploiting FFT and Hankel Transform. Our contributions are as follows:

Speedup the Matching process By focusing on the circularity of in-plane rotation and concentration of power spectrums to low frequency, we compute *fast-eigenimages* $\boldsymbol{H}$ by expanding a templates using Fourier basis, which leads to the use of FFT in a matching process (Fig.1 left).

Speedup the Learning process By utilizing the fact that Fourier expansion in polar coordinates is efficiently transformed to frequency domain using Hankel transform[2], our method computes 2D-Fourier transform of each *fast-eigenimages* $\hat{\boldsymbol{H}}$ in polar coordinate (Fig.1 right Fig.2). This computation is equivalent to existing learning method, i.e., time-consuming rotated template generation, numerical SVD and 2D-FFTs in Cartesian Coordinates, but substantially boosts the learning process by avoiding these time-consuming computation.

Our experiments revealed that the learning, matching, and total processes respectively becomes 120, 3, and 36 times faster, while keeping comparable matching performance compared to previous method. As a representative example, we show an application to global localization with a Particle Filter.

[1] Gou Koutaki and Keiichi Uchimura. Occlusion Robust Pattern Matching Using Shape Based Eigen Templates. *IEEJ Transactions on Electronics, Information and Systems*, 133(1):134–141, 2013.

[2] Robert Piessens. The hankel transform. 2000.

Tuesday
13:40-14:40

Recoding Color Transfer as a Color Homography

Han Gong[1]
http://www2.cmp.uea.ac.uk/~ybb15eau/
Graham D. Finlayson[1]
g.finlayson@uea.ac.uk
Robert B. Fisher[2]
http://homepages.inf.ed.ac.uk/rbf/

[1] School of Computing Sciences
University of East Anglia
Norwich, UK

[2] School of Informatics
University of Edinburgh
Edinburgh, UK

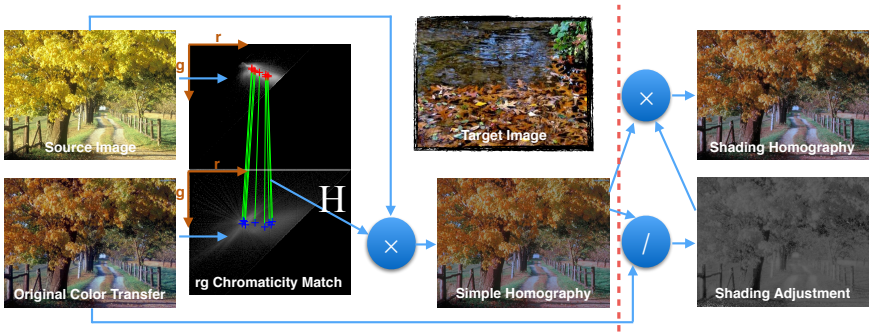


Figure 1: Pipeline of color-homography-based color transfer recoding.

The color homography theorem shows that colors across a change in photometric viewing condition are related by a homography [1]. In this paper, we propose a color-homography-based color transfer decomposition which encodes color transfer as a combination of chromaticity shift and shading adjustment. Our experiments show that the proposed color transfer decomposition provides a close approximation to many popular color transfer methods. We believe that our color transfer model is useful and fundamental for developing simple and efficient color transfer algorithms. Our model also enables users to amend the imperfections of a color transfer result or extract a concise form of the original desired effect for an efficient reapplication (see our paper for examples).

In Figure 1, we start with the outputs of the prior-art algorithms. Assuming we relate source image I_s to target image I_t with a pixel-wise correspondence, we represent the RGBs of I_s and I_t as two $n \times 3$ matrices A and B respectively where n is the number of pixels. These $n \times 3$ matrices can be reconstituted into the original image grids. The chromaticity mapping is modeled as a 3×3 linear transform but because of the relative positions of light and surfaces there might also be per-pixel shading perturbations. Assume the Lambertian image formation is an accurate physical model,

$$DAH \approx B \tag{1}$$

where D is an $n \times n$ diagonal matrix of shading factors and H is a 3×3 chromaticity mapping matrix. A color transfer can be decomposed into a diagonal shading matrix D and a homography matrix H. The homography matrix H is a global chromaticity mapping. The matrix D can be seen as a change of surface reflectance or position of illuminant. Equation 1 can be solved by Alternating Least Squares [1]. To apply the extracted color transfer effect to a different scene, the shading adjustment D can be further modeled as a smooth brightness-to-shading function f as follows:

$$\mathbf{diag}(D) \approx f(\mathbf{brightness}(AH)) \tag{2}$$

[1] Graham D. Finlayson, Han Gong, and Robert B. Fisher. Color homography color correction. In *Color Imaging Conference*. Society for Imaging Science and Technology, 2016.

Pose-Robust 3D Facial Landmark Estimation from a Single 2D Image

Brandon M. Smith
http://www.cs.wisc.edu/~bmsmith
Charles R. Dyer
http://www.cs.wisc.edu/~dyer

Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI USA

Despite much research interest in facial landmark estimation in recent years, relatively little work has been done to handle the full range of head poses encountered in the real world (*e.g.*, beyond $\pm 45^\circ$ rotation). As a result, the large majority of face alignment algorithms are limited to near fronto-parallel faces, and break down on profile faces. We propose an approach to face alignment that can handle 180° of head rotation.

The foundation of our approach is cascaded shape regression (CSR), which has emerged as the leading strategy (see, *e.g.*, [2]). To better handle a wide range of head poses, we extend the 2D CSR approach to 3D. That is, instead of fitting a 2D face model to single 2D images, we fit a 3D face model to single 2D images (3D-to-2D). Intuitively, as the range of head poses increases, the 3D geometry of the face becomes increasingly important in explaining its 2D image projection.

Recent facial landmark estimation methods, including 3D-to-2D approaches [3], employ *local* optimization algorithms at each cascade level, which can fail on face collections with large head pose variation. It is unlikely that a single cascade of generic domain maps (from input features to output landmark updates) will consistently find the true solution. We therefore partition the shape regression problem into a set of simpler *viewpoint domains*, and learn a separate cascade of regressors for each. Each viewpoint domain corresponds to an automatically learned range of camera viewpoints/head poses, as shown in Figure 1. At test time our algorithm adaptively chooses which CSR to apply.

Despite a recent trend toward modeling face shape nonparametrically (*e.g.*, directly updating landmark coordinates), we adopt a parametric model and show empirically that there are no significant differences in accuracy between parametric and nonparametric shape models.

CSR methods commonly use off-the-shelf feature mapping functions (*e.g.*, SIFT) to produce features from the image. Instead, we employ regression random forests [1] to *learn* local binary features that predict ideal shape parameter updates.

Figure 1: The first four modal viewpoints found for $V = 8$ viewpoint domains. The modal occlusion state is stored for each viewpoint domain (green is visible, red is occluded).

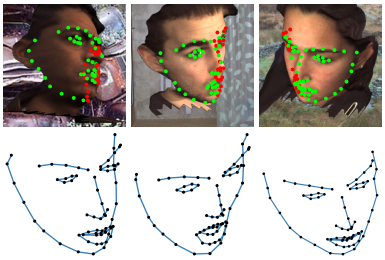


Figure 2: Qualitative results on faces from BU-4DFE [4]. Top row: estimated visibility of each landmark (green is visible, red is occluded). Bottom row: estimated 3D shape.

Results demonstrate quantitatively that the proposed approach is significantly more accurate than recent work. Figure 2 shows a sample of qualitative results.

[1] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] Sergey Tulyakov and Nicu Sebe. Regressing a 3D face shape from a single image. In *IEEE International Conference on Computer Vision*, 2015.

[4] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

Tuesday
13:40-14:40

Bottom-up Instance Segmentation using Deep Higher-Order CRFs

Anurag Arnab
anurag.arnab@eng.ox.ac.uk
Philip H.S. Torr
philip.torr@eng.ox.ac.uk

Department of Engineering Science
University of Oxford
United Kingdom

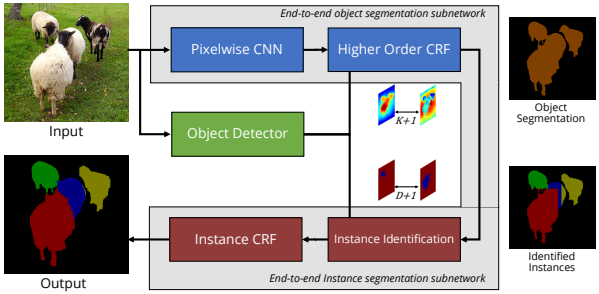


Figure 1. Overview of our end-to-end method. Our system consists of an initial network for semantic segmentation, and then additional modules for instance segmentation. Please refer to our paper for details.

Object detection and semantic segmentation have been two of the most popular Scene Understanding problems within the Computer Vision community. In this paper, we focus on the problem of *Instance Segmentation*. Instance Segmentation lies at the intersection of Object Detection – which localises different objects at a bounding box level, but does not segment them – and Semantic Segmentation – which determines the object-class label of each pixel in the image, but has no notion of different instances of the same class. As shown in Fig. 1, the task of instance segmentation localises objects to a pixel level.

Many recent instance segmentation works have built on the "Simultaneous Detection and Segmentation" (SDS) approach of Hariharan *et al*. [2]. These methods all involve first detecting the various objects in an image before refining these detections into instance-level segmentations.

We present a different approach to instance segmentation, where we initially perform a category-level, semantic segmentation of the input image, classifying each pixel into one of K fixed categories. The resulting semantic segmentation is then refined into an instance-level segmentation, where the object class of each instance segment is obtained from the previous semantic segmentation. Both of these stages, while conceptually different, are fully differentiable and the entire system can be implemented as a neural network. We are able to reason about instances because our semantic segmentation network incorporates a differentiable Higher Order Conditional Random Field (CRF) which uses the cues from the output of an object detector. This CRF is inserted as another layer of a neural network [1, 4]. The object detection cues not only improve category-level segmentations, but the original detection scores are also calibrated during inference. This makes our system robust to false-positive detections, and helps us to reason about instances in the second part of the network. Our paper has full details on this formulation.

Our simple, bottom-up method is able to effectively leverage the progress made by state-of-the-art semantic segmentation and object detection networks to perform the related task of instance segmentation. This is emphasised by our state-of-the-art performance on the VOC 2012 dataset.

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip H. S. Torr. Higher order potentials in end-to-end trainable conditional random fields. In *ECCV*, 2016.

[2] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. Springer, 2014.

[3] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.

[4] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

Horizon Lines in the Wild

Scott Workman
scott@cs.uky.edu
Menghua Zhai
ted@cs.uky.edu
Nathan Jacobs
jacobs@cs.uky.edu

Department of Computer Science
University of Kentucky
Lexington, KY, USA

Single image horizon line estimation is one of the most fundamental geometric problems in computer vision. Knowledge of the horizon line enables a wide variety of applications, including: image metrology, geometry-aware object detection, and automatic perspective correction. Despite this demonstrated importance, progress on this task has stagnated. We believe the lack of a suitably large and diverse evaluation dataset is the primary cause. Existing datasets [2, 3] are often small and were created to focus on evaluating methods that use a particular geometric cue (*e.g.*, orthogonal vanishing points). Methods that perform well on such datasets often perform poorly in real-world conditions.

We introduce *Horizon Lines in the Wild* (HLW), a new dataset for single image horizon line estimation. HLW is several orders of magnitude larger than any existing dataset for horizon line detection (containing 100 553 images), has a much wider variety of scenes and camera perspectives, and wasn't constructed to highlight the value of any particular geometric cue. The dataset (including models and sample code) is available for download at our project website [1].

Using HLW, we investigate methods for directly estimating the horizon line using convolutional neural networks (CNNs), including both classification and regression formulations. We focus on the GoogleNet architecture and explore the impact of design and implementation choices on the accuracy of the resulting model. Additionally, we propose two post-processing strategies for aggregating horizon line estimates across subwindows.

Our approach is fast, works in natural and man-made environments, does not fail catastrophically when vanishing point detection is difficult, and outperforms all existing methods on the challenging real-world imagery contained in HLW. Further, when combined with the recent method by Zhai et al. [4], which uses a CNN to provide global context for vanishing point estimation, we obtain state-of-the-art results on two existing benchmark datasets [2, 3].

Figure 1: Example results showing the estimated distribution over horizon lines (ground truth dash green and predicted horizon magenta).

Table 1: Quantitative evaluation.

	HLW	ECD	YUD
Ours	69.97%	83.96%	85.33%
Ours (*average*)	**71.16%**	83.60%	86.41%
Ours (*optimize*)	70.66%	86.05%	86.11%
[4] (CNN = Orig.)	58.24%	90.80%	94.78%
[4] (CNN = Ours)	65.50%	**91.29%**	**95.46%**

Our main contributions are: 1) a novel approach for using structure from motion to automatically label images with a horizon line, 2) a large evaluation dataset of images with labeled horizon lines, 3) a CNN-based approach for directly estimating the horizon line in a single image, and 4) an extensive evaluation of a variety of CNN design choices.

[1] *Horizon Lines in The Wild* project website. `http://hlw.csr.uky.edu/`.

[2] Olga Barinova, Victor Lempitsky, Elena Tretiak, and Pushmeet Kohli. Geometric image parsing in man-made environments. In *ECCV*, 2010.

[3] Patrick Denis, James Elder, and Francisco Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *ECCV*, 2008.

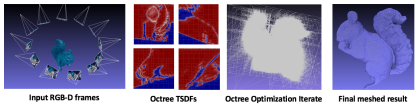
[4] Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting vanishing points using global image context in a non-manhattan world. In *CVPR*, 2016.

Tuesday
13:40-14:40

An Octree-Based Approach towards Efficient Variational Range Data Fusion

Wadim Kehl [1]
kehl@in.tum.de
Tobias Holl [1]
holl@in.tum.de
Federico Tombari [12]
tombari@in.tum.de
Slobodan Ilic [13]
slobodan.ilic@siemens.com
Nassir Navab [1]
navab@cs.tum.edu

[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



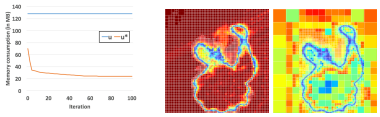


Figure 1: Left: Memory usage between dense and Octree variant during optimization. Center/Right: Octree-slice at iterations 1 and 100.

Figure 2: Meshed result and reconstruction difference between dense and Octree version.

Volume-based reconstruction is usually expensive both in terms of memory consumption and runtime. Especially for sparse geometric structures, volumetric representations produce a huge computational overhead. We present an efficient way to fuse range data via a variational Octree-based minimization approach by taking the actual range data geometry into account. We transform the data into Octree-based truncated signed distance fields and show how the optimization can be conducted on the newly created structures. The challenge is to uphold speed and a low memory footprint without sacrificing the solutions' accuracy during optimization.

We construct Octrees f_i^* from TSDFs f_i in a top-to-bottom manner. Starting from root node n, we define the spread s of values subsumed by node n in f,

$$s_f(n) = \left| \max_{\mathbf{x}\in\Omega_3(n)} f(\mathbf{x}) - \min_{\mathbf{x}\in\Omega_3(n)} f(\mathbf{x}) \right| \quad (1)$$

with $\Omega_3(n)$ being the subvolume that node n represents. Initially, $\Omega_3(n) = \Omega_3$ and the spread will be maximal. From here we recursively apply splitting rules until the partitioning is finished.

We define a functional over the Octrees as

$$\mathcal{E}(u^*) := \int_{\Omega_3} \frac{D(\mathbf{f}^*, \mathbf{w}^*, u^*)}{\sum_i w_i^* + \gamma} + \lambda S(\nabla u^*)\, d\mathbf{x}, \quad (2)$$

and to solve for u^* we determine the steady state of our PDE with a constantly evolving Octree u^*

$$\frac{\partial \mathcal{E}}{u^*} = \lambda \operatorname{div}(S_{\nabla u^*}(\nabla u^*)) - \frac{D_{u^*}(\mathbf{f}^*, \mathbf{w}^*, u^*)}{\sum_i w_i^* + \gamma}. \quad (3)$$

We conduct the optimization by having at all times only one version of u^* in memory and adjusting the structure while we recursively traverse into each node of u^*. Finally, it is shown on three different kinds of data (synthetic, Kinect, high-precision depth scanner) that the presented method performs very well and can decrease both the runtime and memory requirements while retaining a reconstruction quality that is on par with their dense pendants.

Finsler Geodesics Evolution Model for Region-based Active Contours

Da Chen[1]
chenda@ceremade.dauphine.fr
Jean-Marie Mirebeau[2]
jean-marie.mirebeau@math.u-psud.fr
Laurent D. Cohen[1]
cohen@ceremade.dauphine.fr

[1] CEREMADE, CNRS, University Paris Dauphine, PSL Research University, UMR 7534, 75016 PARIS, FRANCE

[2] Laboratoire de mathématiques d'Orsay, CNRS, Université Paris-Sud, Université Paris-Saclay, 91405 ORSAY, FRANCE

We introduce a new deformable model for image segmentation by reformulating a region-based active contours energy into a geodesic contour energy through a Finsler metric.

Let $\Omega \subset \mathbb{R}^2$ be the image domain and $\gamma : [0,1] \to \Omega$ be a regular curve with outward normal vector $\mathcal{N}$. Given a function $f : \Omega \to \mathbb{R}$ of interest, we consider the curve evolution scheme $\partial\gamma/\partial\tau = f(\gamma)\mathcal{N}$, where τ denotes time. This curve evolution equation can be regarded as a gradient descent, thus a minimization procedure for the functional $F(\gamma) = \int_K f(\mathbf{x})\,d\mathbf{x}$, where K is the region inside the closed curve $\gamma := \partial K$.

A complete active contours energy with a curve length regularization is defined by

$$E(\gamma) = \alpha F(\gamma) + \int_0^1 P(\gamma(t))\|\dot{\gamma}(t)\|\,dt, \quad (1)$$

where P is an edge-based potential function, and $\alpha > 0$ is a constant.

Reformulation as Finsler Geodesic Energy: Suppose $\mathcal{V}_\perp : \Omega \to \mathbb{R}^2$ to be a continuously differentiable vector field defined over the domain Ω such that $\mathcal{V}_\perp$ satisfies the divergence equation $\nabla \cdot \mathcal{V}_\perp(\mathbf{x}) = \alpha f(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega$, where f is the first order derivative function used in (1) and $\nabla \cdot \mathcal{V}_\perp(\mathbf{x})$ denotes the divergence value of a vector $\mathcal{V}_\perp(\mathbf{x})$. Letting M be the counter-clockwise rotation matrix with rotation angle $\theta = \pi/2$, by divergence theorem, the regional energy F in (1) can be expressed as $\alpha F(\gamma) = \int_K \nabla \cdot \mathcal{V}_\perp(\mathbf{x})d\mathbf{x} = \int_0^1 \langle \mathcal{V}(\gamma(t)), \dot{\gamma}(t)\rangle dt$, where $\mathcal{V} = M^{\mathrm{T}}\mathcal{V}_\perp$. We introduce a Finsler metric $\mathcal{F} : \Omega \times \mathbb{R}^2 \to \mathbb{R}^+$:

$$\mathcal{F}(\mathbf{x},\mathbf{u}) = P(\mathbf{x})\|\mathbf{u}\| + \langle \mathcal{V}(\mathbf{x}), \mathbf{u}\rangle, \forall \mathbf{x} \in \Omega, \forall \mathbf{u} \in \mathbb{R}^2.$$

This metric should obey the smallness condition $\|\mathcal{V}(\mathbf{x})\| < P(\mathbf{x})$, which is difficult to be satisfied. Therefore, assuming that $\forall \mathbf{x} \in \Omega$, $P(\mathbf{x}) \geq 1$, we make use of the following condition:

$$\|\mathcal{V}(\mathbf{x})\| < \min_{\mathbf{y}\in\Omega}\{P(\mathbf{y})\} = 1, \quad \forall \mathbf{x} \in \Omega. \quad (2)$$

The energy E (1) is converted to the Finsler geodesic energy $\mathcal{L}(\gamma) = \int_0^1 \mathcal{F}(\gamma(t), \dot{\gamma}(t))dt$. The minimization procedure of $\mathcal{L}$ is solved inside a neighbourhood U instead of the whole domain Ω. This means that we only require the vector field $\mathcal{V}^\perp$ defined over U. In order to satisfy (2), it is natural to select the vector field $\mathcal{V}_\perp$ by minimizing an energy for all $\mathbf{x} \in U$

$$\min\left\{\int_U \|\mathcal{V}_\perp\|^2\right\}, s.t. \nabla \cdot \mathcal{V}_\perp(\mathbf{x}) = \alpha f(\mathbf{x}). \quad (3)$$

Note that $\|\mathcal{V}_\perp\|_\infty$ is bounded by the area of U. To obtain a vector field obeying $\|\mathcal{V}_\perp\|_\infty < 1$, one can choose a tubular neighbourhood U with small width hence a small area. On the other hand, U is regarded as the search space for the next evolutional curve. A small U may therefore make the algorithm fall into undesirable local minimas of $\mathcal{L}$. Thus we make use of a nonlinear mapping increasing function $T(x) = 1 - \exp(-x), \forall x > 0$. The vector field $\bar{\mathcal{V}}$ is defined as $\bar{\mathcal{V}}(\mathbf{x}) = T(\|\mathcal{V}_\perp(\mathbf{x})\|)M^{-1}\mathcal{V}_\perp(\mathbf{x})/\|\mathcal{V}_\perp(\mathbf{x})\|$, where the condition (2) will be immediately satisfied. Based on $\bar{\mathcal{V}}$, the new Finsler metric is defined by $\bar{\mathcal{F}}(\cdot,\mathbf{u}) = P(\mathbf{x})\|\mathbf{u}\| + \langle \bar{\mathcal{V}}(\cdot), \mathbf{u}\rangle$ and the respective geodesic energy is defined by $\bar{\mathcal{L}} = \int_0^1 \bar{\mathcal{F}}(\gamma(t), \dot{\gamma}(t))dt$.

The minimization of E (1) is transferred to the minimization of $\bar{\mathcal{L}}$. Note that since in general we induce $\bar{\mathcal{L}}$ with a nonlinear mapping T, there is in fact slight difference in the minimization problems. The use of the non-linear mapping T is reasonable: **1)** Minimizing E is to find a path γ, for which the direction $\dot{\gamma}(t)$ for each $t \in [0,1]$ should be as opposite to $\mathcal{V}(\gamma(t))$ as possible and the norm $\|\mathcal{V}(\gamma(t))\|$ should be as large as possible, giving the relevance between the minimization problems of E and $\bar{\mathcal{L}}$. **2)** When the Finsler geodesics evolution scheme tends to stabilize, one can reduce the width of tubular neighbourhood U. Thus $T(\|\mathcal{V}\|) \approx \|\mathcal{V}\|$ as $\|\mathcal{V}\|$ is small.

Tuesday
13:40-14:40

Patch Based Confidence Prediction for Dense Disparity Map

Akihito Seki[1,2]
akihito.seki@toshiba.co.jp
Marc Pollefeys[2]
marc.pollefeys@inf.ethz.ch

[1]Corporate R&D Center
Toshiba Corporation, Japan
[2]Department of Computer Science
ETH Zürich, Switzerland

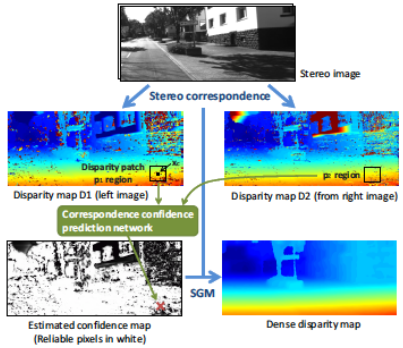


Figure 1: Overview of our method.

Confidence of stereo correspondences is useful information to improve quality of the disparity maps. Learning based confidence measure[4] combines hand-crafted confidence prediction features and is able to outperform their individual usage. These features and classifiers are carefully designed, however beneficial information might be undescribed or their representation might be too redundant. Figure 1 shows our confidence prediction method in order to overcome the problem. Moreover, the confidence is incorporated into Semi Global Matching (SGM) [3] to acquire dense disparity map. In the following, we will briefly explain both methods and their results.

Confidence estimation with a CNN: We leverage the disparity patch and introduce the knowledge of the conventional features. The patch consists in a two channels. 1st channel is coming from an idea that neighboring pixels on a disparity map D_1 which have consistent disparities are more likely to be correct matching. In 2nd channel, a disparity D_2 from another image is considered such that the matches from left to right image should be consistent with those from right to left. We employ a shallow CNN for the sake of reducing potential computation cost of the network, however, the network is still slow computation because the output of the network for each pixel has to be computed from scratch. We also propose speed-up networks by modifying pre-processing of the patches and network structure. Table 1 shows evaluation results based on sparsification curve and its area under curve (AUC) value. Better confidence prediction methods have AUC values that are closer to the optimal curve: It means the method removes incorrect correspondence pixels while keeping the correct ones. Our methods outperform state of the art method[4] on both accuracy and computation time.

Method		AUC[$\times 10^{-2}$]	Time[sec.]
Optimal		3.95	-
Ours		**4.20**	28.5
	fast	4.50	**0.3**
	hybrid	4.48	0.5
Park&Yoon[4]	50 trees	4.70	2.2

Table 1: Overall AUC value and computation time.

Rank	Method	Out-Noc error
1	Ours	**2.36%**
2	Displets v2[2]	2.37%
3	VDS(anonymous)	2.42%
4	MC-CNN-acrt[5]	2.43%

Table 2: Out-Noc error on KITTI 2012 testing dataset by May 1st 2016.

Confidence fusion with SGM: SGM has two parameters in order to control discontinuities of disparity map. We consider the pixels with high confidence should be trusted and are able to be discontinuities easily. Table 2 shows the accuracy of dense disparity map on KITTI 2012[1] testing dataset. We got the best accuracy.

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

[2] F. Guney and A. Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *Proc. CVPR*, 2015.

[3] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *Trans. on PAMI*, 30(2):328–341, 2008.

[4] M.G. Park and K.J. Yoon. Leveraging stereo matching with learning-based confidence measures. In *Proc. CVPR*, 2015.

[5] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proc. CVPR*, 2015.

Boosted Convolutional Neural Networks

Mohammad Moghimi[1]
mmoghimi@cs.cornell.edu
Mohammad Saberian[2]
esaberian@netflix.com
Jian Yang[3]
jianyang@yahoo-inc.com
Li-Jia Li[5]
lijiali@cs.stanford.edu
Nuno Vasconcelos[4]
nvasconcelos@ucsd.edu
Serge Belongie[1]
sjb344@cornell.edu

[1] Cornell Tech, New York, NY
Cornell University, Ithaca, NY
[2] Netflix, Los Gatos, CA
[3] Yahoo Research, Sunnyvale, CA
[4] UC San Diego, La Jolla, CA
[5] Snapchat, Los Angeles, CA

In this work, we propose a new algorithm for boosting Deep Convolutional Neural Networks (BoostCNN) to combine the merits of boosting and these networks. To learn this new model, we propose a novel algorithm to incorporate boosting weights into the deep learning architecture. More specifically, in each iteration of boosting, we train a deep network to approximate the boosting weights, i.e. minimizing

$$\mathcal{L}_{se}(w,g) = \sum_{x_i \in \mathcal{D}} \sum_{j=1}^{M} (g_j(x_i) - w_j(x_i))^2,$$

where $g(x)$ is a deep network, $w(x)$ are boosting weights, M is the number of classes and $\mathcal{D}$ is the training set. Experiments show that the proposed method is able to achieve state-of-the-art performance on several fine-grained classification tasks such as bird, car, and aircraft classification, see Table (1) as an example.

In addition we also show that it is possible to use networks of different structures within the proposed boosting framework. In this case, at each boosting iterations, we train these networks independently to approximate the boosting weights and the network that leads to the best performance will be added to the ensemble. Experiment show that, this not only results in superior performance but also reduces the required manual effort for finding the right network structure, see Table (1).

Our open source implementation is based on Caffe framework and it is available on Github[1].

[1] http://github.com/mmoghimi/BoostCNN

Method	Accuracy
BoostCNN	85.6%
BoostCNN (heterogeneous)	**86.2**%
Bilinear CNN (B-Net)	84.1%
Krause *et al.*	82.0%
Pose Normalized CNN	75.7%
Part-based RCNN	73.9%

Table 1: Performance comparison for bird classification on CUB200 dataset.

Material-Specific Chromaticity Priors

Tuesday
13:40-14:40

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

1 Introduction

Recent advances in machine learning have enabled the recognition of high-level categories of materials with an accuracy of up to 80% [1]. With these techniques, we can construct a per-pixel material labeling from a single image. We observe that groups of materials have distinct chromaticity footprints (see Figure 1). We propose a novel combination of techniques, where we use a material classifier to predict the dominant material category of objects, which in turn is useful to constrain reflectances in the context of the intrinsic images problem [2]. Preliminary evaluation indicates our method has promise. In Figure 1, the chromaticities in YUV color space plotted in heatmap colors, from random samples of different material categories. Blue represents UV values that are the least frequent, red those that are the most frequent. The plots show that various subgroups of materials have different characteristics. Plastic has a much wider range of chromaticity values than sky. Wood spans a limited range of unsaturated colors, while metal has quite a few outliers due to strong specular reflections. This suggests we can use knowledge of the material in a scene to predict the chromaticities and thereby improve estimation of the underlying reflectance.

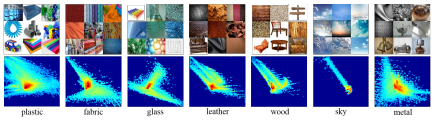


Figure 1: Chromaticity plots in heatmap colors.

2 Our approach

Using recent deep learning techniques, we segment objects in areas of homogeneous materials. We assign a material label to each pixel in these regions. Areas with different materials form islands within the space of chromaticity distributions. If the category is known, it makes sense to use this knowledge to properly constrain the chromaticity values and place more stringent priors on the reflectance of segmented objects parts.

This prior is useful in the context of intrinsic image decomposition, where it can be used to contrain reflectance estimation. Preliminary evaluation of our method on the MIT dataset [2] is shown in Table 1. It shows the MSE error metric of the estimation and the ground truth data. The normal prior trains chromaticity on 10 other images from the MIT dataset. Our method trains the chromaticity prior on images from specific material categories in OpenSurfaces and consistently performs better.

Table 1: Preliminary evaluation of our method on the MIT dataset, showing the MSE error metric of the estimation and the ground truth data.

	normal prior MSE	our prior MSE	Difference
cup2	25.8316	**22.4774**	14.92%
frog2	35.4718	**29.8646**	18.77%
paper2	33.3270	**30.0913**	10.75%
pear	31.8777	**27.6916**	15.11%
potato	29.2384	**26.2526**	11.37%
raccoon	27.5501	**24.2536**	13.59%
sun	36.1633	**32.7922**	10.28%
teabag1	43.5110	**37.6781**	15.48%
squirrel	40.1366	**35.4141**	13.34%

[1] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pages 2335–2342, 2009.

Play and Learn: Using Video Games to Train Computer Vision Models

Alireza Shafaei
http://cs.ubc.ca/~shafaei
James J. Little
http://cs.ubc.ca/~little
Mark Schmidt
http://cs.ubc.ca/~schmidtm

Department of Computer Science
University of British Columbia
Vancouver, Canada

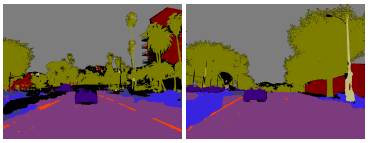


Figure 1: Densely-labeled samples from the synthetic dataset.

Are video games realistic enough to be used in the training of computer vision models to address practical, real-world problems? We explore this idea and deliver a proof of concept by experimenting with the synthetic RGB images that we sample from a video game. We collect over 60,000 synthetic samples with similar conditions to the real-world CamVid [1] and Cityscapes [2] datasets. We provide several experiments to demonstrate that the synthetically generated RGB images can be used to improve the performance of deep neural networks on both image segmentation and depth estimation.

Dataset. We capture the synthetic dataset by sampling the game every second while an autonomous driver is wandering in the city. Each sample contains the RGB image, semantic segmentation, depth image, and the surface normals. See Fig. 1.

Experiments. We use the FCN8 [3] architecture for the dense image classification task, and for the depth estimation experiments we use the approach of Zoran *et al*. [4].

Results. We show that in a cross-dataset setting, the CNNs that we obtain from synthetic data have a similar test error as the networks that we train on real-world data (Fig. 2). Furthermore, the synthetically generated RGB images can provide similar or better results compared to the real-world datasets if a simple domain adaptation technique is applied (Tab. 1). We also show that pre-training on synthetic data results in a better initialization and final local minima in the optimization. For the depth estimation task, we present similar improvements.

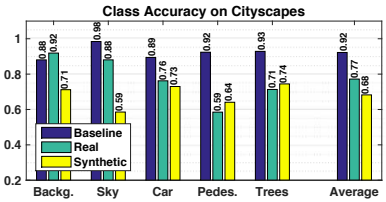


Figure 2: The cross-dataset per-class accuracy. The baseline is trained on the target dataset, the real is trained on the CamVid dataset, and the synthetic is trained on synthetic data only.

	Cityscapes		
Model	Pixel Acc.	Class Acc.	IoU
Baseline	83%	77%	50%
Real	83%	77%	50%
Synthetic	**84%**	**79%**	51%
Mixed	**84%**	**79%**	**52%**

Table 1: Evaluation of different pre-training strategies.

Conclusion. Our results suggest that video games with photorealistic environments are potentially useful for a variety of computer vision tasks as they can offer an alternative way to compile large realistic datasets for training and evaluation.

[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] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[4] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015.

SDF-TAR: Parallel Tracking and Refinement in RGB-D Data using Volumetric Registration

Tuesday
13:40-14:40

Miroslava Slavcheva
mira.slavcheva@tum.de
Technische Universität München
Munich, Germany

Slobodan Ilic
slobodan.ilic@siemens.com
Siemens AG
Munich, Germany

SDF-TAR is a real-time SLAM system which employs volumetric registration in RGB-D data. It is based on the SDF-2-SDF registration energy [1] that minimizes the per-voxel difference of a pair of signed distance fields (SDFs). The energy is used both in the GPU frame-to-frame tracking module, and in the concurrent CPU batch pose refinement module. To minimize runtime and memory consumption, registration is done only over several limited-extent volumes (LEVs), anchored at locations of high curvature.

LEVs The original SDF-2-SDF registration uses regular voxel grids, which become too memory-intensive if scanning large spaces with a fine resolution is desired. To tackle this issue, we carry out registration in a number of partial volumes (dubbed LEVs), which guarantee an upper bound on the runtime and memory requirements. We set them at the most geometrically discriminative regions of a scene, namely at the locations of highest curvature. These anchor points are very fast to compute as the second order derivative directly from the depth image, followed by a non-maximum suppression step. This allows us to select peaks sufficiently far apart, so that volumes do not overlap. Experiments show that this strategy leads to higher tracking accuracy than taking the same numbers of uniformly spaced or randomly placed LEVs.

Registration To estimate the camera pose ξ, represented via 6-element twist coordinates, of an SDF ϕ_c with weight field ω_c relative to a reference SDF ϕ_r with weight field ω_r generated from the identity pose, we minimize the sum of direct per-voxel differences in all p LEVs Ω_i:

$$E_{SDF}(\xi) = \frac{1}{2} \sum_{LEV\ \Omega_i}^{i=1..p} \Big(\sum_{voxels \in \Omega_i} \big(\phi_r \omega_r - \phi_c(\xi)\omega_c(\xi)\big)^2 \Big) . \quad (1)$$

Parallel tracking and refinement *Tracking* is carried out in a frame-to-frame fashion on the GPU, using a first-order Taylor approximation of

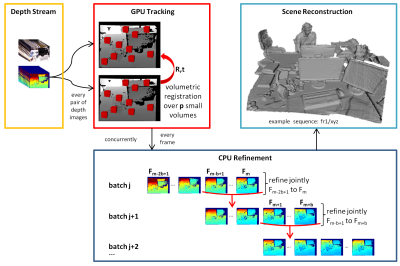


Figure 1: Overview: GPU frame-to-frame tracking, and concurrent CPU batch pose refinement.

Eq. 1 which yields a 6×6 linear system in every iteration. *Pose refinement* is done jointly over batches of $2b$ frames concurrently on the CPU, until the next batch becomes available. The second half of a batch are the most recently tracked b frames, while the first half are the previous b frames that have already been optimized once. The first $b/2$ poses are kept fixed for stability. As tracking ensures a good initialization, this optimization follows a simple gradient descent over the 6-element pose vector of every camera.

Results We evaluate tracking precision on the TUM RGB-D benchmark [2]. Via the use of LEVs, SDF-TAR disregards regions that could impede registration, leading to considerably better rotational and on-par translational motion estimation with related volumetric techniques. In addition, we achieve higher reconstruction fidelity on the 4 objects of the CoRBS dataset [3].

[1] M. Slavcheva, W. Kehl, N. Navab, and S. Ilic. SDF-2-SDF: Highly Accurate 3D Object Reconstruction. In *Proc. ECCV*, 2016.

[2] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proc. IROS*, 2012.

[3] O. Wasenmüller, M. Meyer, and D. Stricker. CoRBS: Comprehensive RGB-D Benchmark for SLAM using Kinect v2. In *Proc. WACV*, 2016.

Probabilistic Semi-Supervised Multi-Modal Hashing

Behnam Gholami
bb510@cs.rutgers.edu
Computer Science Department
Rutgers university

Abolfazl Hajisami
hajisamik@cac.rutgers.edu
Department of Electrical and Computer Engineering
Rutgers university

In this paper, we propose a non-parametric Bayesian framework for multi-modal hash learning that takes into account the distance supervision (similarity/dissimilarity constraints). Our model embeds data of arbitrary modalities into a single latent binary feature with the ability to learn the dimensionality of the binary feature using the data itself. Given supervisory information (labeled similar and dissimilar pairs), we propose a novel discriminative term and develop a new Variational Bayes (VB) algorithm which incorporates that term into the proposed Bayesian framework.

Let $\boldsymbol{T} = [\boldsymbol{X}, \boldsymbol{Y}]$ be the observed bi-modal data matrix where $\boldsymbol{X} = [x_1, x_2, ..., x_d]_{M\times d}$ and $\boldsymbol{Y} = [y_1, y_2, ..., y_d]_{N\times d}$ denote the first modal and the second modal data matrix respectively, and $\boldsymbol{Z} = [z_1, z_2, ..., z_d]_{K\times d}$ denotes the latent binary code matrix.

In our VB framework, we truncate the length of the binary codes (K) and we set it to a finite but large number. If K is large enough, the analyzed multi-modal data using this number of bits, will reveal less than K bits. In order to incorporate the information of the similarity/dissimilarity constraints into the VB algorithm, we first define a regularizer for the binary code z_i as

$$\alpha(z_i) = \frac{1}{|\mathcal{D}_i|} \sum_{j:(i,j)\in\mathcal{D}} KL(q_{z_i}(z_i)||q_{z_j}(z_j)) - \frac{1}{|\mathcal{S}_i|} \sum_{j:(i,j)\in\mathcal{S}} KL(q_{z_i}(z_i)||q_{z_j}(z_j)) \quad (1)$$

where $KL(p||q)$ denotes the KL divergence between two distributions p and q, and $\mathcal{S}(\mathcal{D})$ denotes the set of similar (dissimilar) pairwise constraints. Intuitively, for each binary code z, $\alpha(z)$ should be large such that it best agrees with those constraints.

By defining the regularizer $\Omega(\boldsymbol{Z}) = \sum_{i=1}^{d} \alpha(z_i)$ for the binary code matrix $\boldsymbol{Z}$ using the set of similar/dissimilar pairs, we add this regularizer to the objective function of VB and solve the new optimization problem using the Coordinate Descent method.

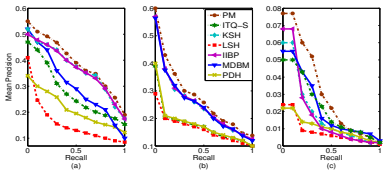


Figure 1: The result of category retrieval for image-to-image queries. (a) PASCAL-Sentence Dataset; (b) SUN Dataset (Euclidean ground truth computed from visual data); (c) SUN Dataset (Class label ground truth)

We evaluate the proposed method on two benchmark bi-modal datasets: (1) The PASCAL-Sentence 2008 dataset [1] consists of 1000 images categorized into 20 classes. (2) The SUN-Attribute dataset [2] contains 102 attribute labels for each of the 14340 images from 717 categories. We compare the performance of the proposed method against five state-of-the-art hashing methods (Fig. 1) using precision-recall curve as an accuracy measure. As can be seen, the proposed method outperforms the other state of the art (multi-modal) hashing methods.

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[2] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012.

Learning to Invert Local Binary Patterns

Tuesday
13:40-14:40

Felix Juefei-Xu
felixu@cmu.edu
Marios Savvides
msavvid@ri.cmu.edu

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA

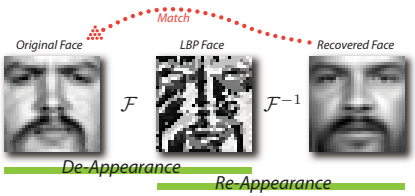


Figure 1: Flowchart of the method. The de-appearance step uses LBP for the forward mapping $\mathcal{F}$ and obtains the LBP face (glyph) in the middle. The re-appearance step tries to learn the inverse mapping $\mathcal{F}^{-1}$ from the LBP domain to the pixel domain. The recovered face is of high fidelity as compared to the original face.

We have proposed to invert the local binary patterns (LBP) descriptor. The success of the inversion gives rise to two applications: face de-appearance and re-appearance. The flowchart of the algorithm is shown in Figure 1.

The de-appearance, based on image-LBP forward mapping, is thorough in the sense that not only the identity information but also the soft-biometric information of the subject is removed. The intuition behind using LBP for de-appearance is straightforward because LBP is a local difference operator and most of the COTS FRS cannot deal with LBP faces / glyphs. They either can not locate the LBP faces from the scene, or the match scores are horribly low.

The re-appearance yields face reconstruction with high fidelity and also enables secure application with a unique encryption key. The re-appearance is achieved by learning the inverse mapping of the LBP descriptors through an ℓ_0-constrained coupled dictionary learning scheme that jointly learns two overcomplete dictionaries in both the pixel and the LBP domains such that inverse mapping from the LBP domain to the pixel domain is made possible without knowing the mapping function explicitly. The enforcement of the sparsity level as well as the sharing of sparse coefficient between the two domains are the added constraints that can uniquely determine the inverse mapping $\mathcal{F}^{-1}$.

Obtaining a consistent sparse encoding between the two domains allows for a more meaningful reconstruction. Given a novel de-appearanced image $\mathbf{y}_{\text{LBP}}$ in the LBP domain, we first obtain the sparse representation $\mathbf{x}$ in $\mathbf{D}_{\text{LBP}}$. We then obtain the reconstruction using $\mathbf{Dx}$. By forcing consistent sparse representations $\mathbf{x}$ during training, we optimize for a low reconstruction error for both domains jointly and simultaneously. A simple rearrangement can lead to solving the formulation using the standard K-SVD dictionary learning approach [1] as previously observed [2]:

$$\underset{\mathbf{D},\mathbf{D}_{\text{LBP}},\mathbf{X}}{\arg\min} \left\| \begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}_{\text{LBP}} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \mathbf{D}_{\text{LBP}} \end{pmatrix} \mathbf{X} \right\|_F^2 \quad (1)$$
$$\text{subject to } \forall i, \|\mathbf{x}_i\|_0 \leq K$$

The procedure also comes naturally with high selectivity when reconstructing the faces with various LBP encryption keys. We have shown the effectiveness of our proposed approach on the FRGC ver 2.0 database which involves large-scale fidelity test and face verification experiments using the state-of-the-art commercial and academic face matchers.

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on signal processing*, 54(11): 4311–4322, 2006.

[2] Zhuolin Jiang, Zhe Lin, and L.S. Davis. Label consistent K-SVD: Learning a discriminative dictionary for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11): 2651–2664, Nov 2013.

Probabilistic Compositional Active Basis Models for Robust Pattern Recognition

Adam Kortylewski
adam.kortylewski@unibas.ch
Thomas Vetter
thomas.vetter@unibas.ch

Department of Mathematics and Computer Science
University of Basel
Basel, Switzerland

In this paper we propose an approach for learning hierarchical compositional active basis models. Our contribution is three-fold: First, we introduce a greedy EM-type algorithm to automatically infer the complete structure of a compositional active basis model (CABM). Second, we formulate the pattern model and the learning process in a fully probabilistic manner. Finally, based on the statistical framework, we augment the pattern model with an implicit geometric background model that reduces the models sensitivity to pattern occlusions and structured clutter. We demonstrate that probabilistic CABMs are capable of recognising patterns under complex non-linear distortions that can hardly be represented by a finite set of training data.
Probabilistic hierarchical compositional models have been proposed as object representation *e.g.* in [1, 2, 3]. However, in contrast to [1], we automatically learn the structure of the hierarchy. The work in [2, 3] is most related to our method. In difference to [3] we embed our model in a statistical inference framework. Compared to [2], we use fully generative compositional units instead of invariant features as part representations. Furthermore, we do not make hard decisions on the detection of parts during learning. Instead the full part likelihoods are used in the structure induction process.
In order to demonstrate the robustness of the proposed object representation, we evaluate it on a complex forensic image analysis task (Fig.1). We learn CABMs for 1175 reference impressions. Subsequently, the posterior probability of 300 probe images under each reference model is inferred within a Bayesian estimation setup. Experimental results show that the forensic image analysis task is processed with unprecedented quality.

[1] Jifeng Dai et al. Unsupervised learning of dictionaries of hierarchical compositional models. In *CVPR*, 2014.

[2] Long Zhu et al. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *ECCV*, 2008.

[3] Sanja Fidler and Aleš Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*, 2007.

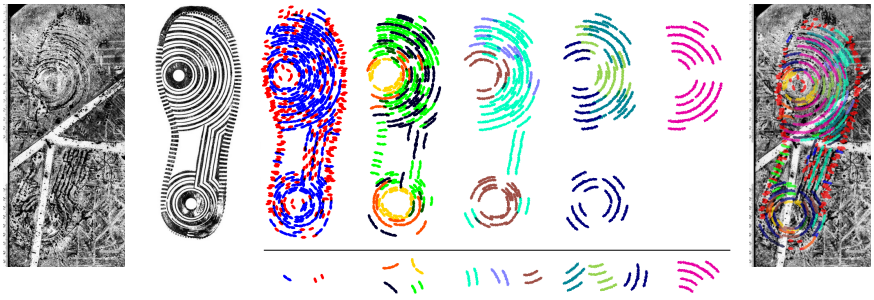


Figure 1: Overview over the process of footwear impression recognition. (a) A typical probe image. (b) The corresponding reference impression; (c-g) The learning result for each layer of the compositional hierarchy. The learned CABMs are illustrated with different colours in their mean position (bottom), together with an encoding of reference impression (top). (h) An overlay of the final CABM over the probe image with the spatial transformation that maximises the posterior probability.

Tuesday
13:40-14:40

Loglet SIFT for Part Description in Deformable Part Models: Application to Face Alignment

Qiang Zhang
q.zhang.13@warwick.ac.uk
Abhir Bhalerao
abhir.bhalerao@warwick.ac.uk

Department of Computer Science
University of Warwick
Coventry, UK

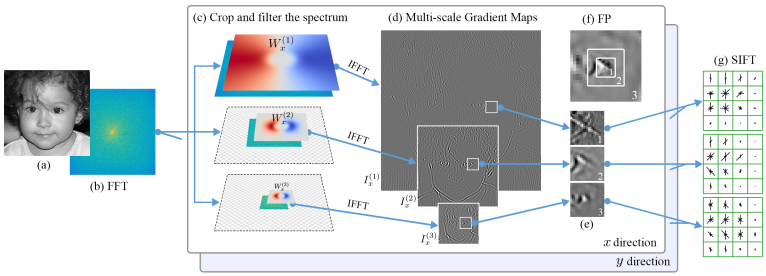


Figure 1: Overview of extracting a loglet-SIFT part descriptor. (a) An image example. (b) The Fourier spectrum of the image. (c) Filter and crop the spectrum with the filter banks. The process with x direction is shown as an example. (d) Pyramids of the gradient maps representing multi-scale structures are obtained directly from the filtered spectrum. (e) At a landmark, patches with the same size in pixels are extracted as a multi-scale local descriptor. (f) The patches represent coherent scales and domain sizes forming a feature pyramid. (g) The Loglet-SIFT descriptor is obtained like SIFT, by accumulating directional gradients from each of the gradient patches.

Deformable Part Model (DPM) have emerged as the leading approach for accurate landmark detection in applications such as face alignment. A DPM describes an object by local parts and the spacial relationships of the parts. Part descriptors seek a representation of local structures which preserves intrinsic properties and discriminative information, while exhibiting invariance to changes such as illumination, scale and variations in appearance across instances. We propose a new local feature descriptor called loglet-SIFT, which enhances a number of invariances, i.e., the invariance to illumination by the local pooling of SIFT and the suppression of slow varying mean level by the wavelets, as well as the invariances to noise by SIFT, and to sample shift by loglets. An overview of the proposed method can be found in Fig 1.

We integrate our descriptor into a DPM driven by a supervised descent method and validate its performance in the face alignment scenario. We compare the performance of our Fourier domain designed filters with spatially designed gradient filters, and compare our descriptor with conventional SIFT. We further present the comparison against several state-of-the-art methods on two popular datasets: HELEN and 300-W. Experimental results show that the new descriptor improves the performance of the DPM by a significant margin. We achieve state-of-the-art performance on HELEN and 300-W common dataset, and comparable performance on the 300-W challenging dataset. The proposed descriptor can be readily integrated in other gradient and SIFT based deformable part models.

Dictionary Replacement for Single Image Restoration of 3D Scenes

T M Nimisha
ee13d037@ee.iitm.ac.in
M Arun
ee14s002@ee.iitm.ac.in
A N Rajagopalan
raju@ee.iitm.ac.in

Image Processing and Computer Vision Lab
Indian Institute of Technology, Madras
Chennai, India

In this paper, we address the problem of jointly estimating the latent image and the depth/blur map from a single space-variantly blurred image using dictionary replacement. While most of the dictionary-based deblurring methods consider planar scenes with space-invariant blur, we handle 3D scenes with space-variant blur caused by either camera motion or optical defocus. For a given blurred image, the dictionary blurred with the corresponding blur kernel provides the best representation with the least error. We formulate our problem of blur map and latent image estimation as a multi-label MRF and solve it using graph-cut.

An image X degraded by space-invariant blur h can be modeled by convolution as

$$Y = h \otimes X = h \otimes D \circ \Lambda = D_b \circ \Lambda \qquad (1)$$

$h \otimes D$ is denoted as D_b, a blurred version of dictionary D. This implies that when kernel h is known (as in non-blind deblurring), the signal X can be recovered from Y using the blur-invariant representation Λ. Dictionary replacement-based deblurring techniques, in fact, work on this principle.

Let Y be the observed blurred image of a 3D scene and h_0 be the blur kernel corresponding to the most blurred region in the image. From the blur-depth relation, we know that the blur at any other position is a scaled down version of h_0. Hence, the problem of depth estimation boils down to estimating the scale of the blur kernel at each location. With the underlying idea that for a given sparsity, the dictionary blurred with the correct scale will represent the blurred patch with minimum error, we formulate this as a Markov Random Field (MRF) problem.

$$\arg\min_i \quad DC_i(k) + \sum_{k' \in \mathcal{N}} SC(\bar{i}_{k'}, i_k) \qquad (2)$$

where $DC_i(k)$ is the data cost and SC is the edge aware smoothness cost.

Input Result by [2] Our result Depthmap

We compared our method with NCSR [1], Hu et al. [2] (blind) and natural prior-based [3] deblurring techniques and showed that our method outperforms others both quantitatively and qualitatively. In our experiments, we included space-variant blur caused by defocus effect as well as motion blur. We also gave two applications of our method in blur magnification and image reblurring. We also considered

Blur Magnification

Reblurring

the case of blur due to object motion. As our method works on local patches and does not assume any global camera motion constraint, it performs well in these scenario too.

[1] Weisheng Dong, Lei Zhang, and Guangming Shi. Centralized sparse representation for image restoration. In *ICCV*, pages 1259–1266. IEEE, 2011.

[2] Zhe Hu, Jia-Bin Huang, and Ming-Hsuan Yang. Single image deblurring with adaptive dictionary learning. In *ICIP*, pages 1169–1172. IEEE, 2010.

[3] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, pages 1107–1114, 2013.

Poisson Noise Removal for Image Demosaicing

Tuesday
13:40-14:40

Sukanya Patil
sukanyapatil1993@gmail.com
Indian Institute of Technology Bombay, Mumbai, India

Ajit Rajwade
ajitvr@cse.iitb.ac.in
Indian Institute of Technology Bombay, Mumbai, India

Most color image cameras today acquire only one out of the R, G, B values per pixel by means of a color filter array (CFA) in the hardware producing the so called 'CFA image'. In-built software routines undertake the task of obtaining the rest of the color information at each pixel through a process termed demosaicing. Studies in [1] have shown that raw CFA images captured by a camera are corrupted predominantly by Poisson noise which affects demosaicing results. While there exist several approaches in the literature to perform demosaicing, most of them do not fully account for the Poisson nature of the noise in the raw CFA images. In this paper, we present two simple but principled methods that infer dictionaries *in situ* from the noisy CFA images, both taking into account the Poisson nature of the noise. These dictionaries are used to denoise the noisy CFA images prior to demosaicing by exploiting the patch-level nonlocal similarity present in CFA images formed under periodic patterns such as the Bayer pattern and the sparsity of the coefficients of a linear combination of dictionary elements to express these patches. The denoised CFA image can be given as input to any off-the-shelf demosaicing routine to generate the full RGB image from the denoised CFA data. Our first approach, which we term the 'Poisson Penalty Approach' (PPA), is based on the direct minimization of an energy function which is the sum of the negative log likelihood of the Poisson noise model and a weighted sparsity-promoting term. Patches from the noisy CFA image are expressed as a non-negative sparse linear combination of dictionary columns, also constrained to be non-negative. Here, the dictionary as well as the sparse coefficients are learned *in situ* from the noisy patches in the CFA image. Our second approach is termed the 'Variance Stabiliser Approach' (VSA). To denoise a Poisson corrupted CFA image Y using this approach, we first compute its Anscombe transform given by $Z = 2\sqrt{Y + 3/8}$, denoise Z using a dictionary-based image denoising algorithm suited for the Gaussian noise model with a fixed, known variance (which equals 1 in this case), and obtain the final image as $W = Z^2/4 - 3/8$. The specific denoising algorithm we use is the spatially varying PCA approach with a Wiener filter. We have performed extensive experiments on both synthetic and real data. Some results on real data captured using a Canon camera are shown in Figure 1 and comparisons are drawn between the noisy image displayed by the camera (after in-built demosaicing without denoising), the Poisson Penalty Approach, the Variance Stabilizer Approach, and a commerically available tool called NeatImage which denoises the RGB image after demoisaicing the noisy CFA image. Both the approaches clearly outperform the results obtained using NeatImage. Our methods have been tested on Bayer pattern CFA images but would work equally well on any other periodic CFA pattern.

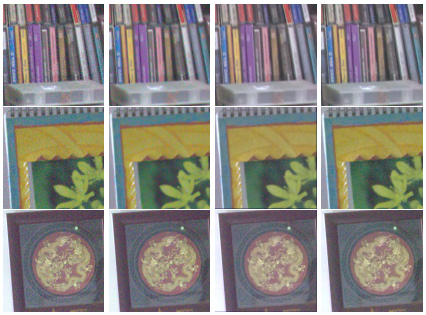


Figure 1: In each row, left to right: noisy image acquired by camera (post in-built demosaicing), output of NeatImage, output of PPA (with appropriate parameters), output of local VSA (with appropriate parameters). See supplemental material for a clearer view.

[1] H. J. Trussell and R. Zhang. The dominance of poisson noise in color digital cameras. In *ICIP*, pages 329–332, Sept 2012.

Factorized Binary Codes for Large-Scale Nearest Neighbor Search

Frederick Tung
ftung@cs.ubc.ca
James J. Little
little@cs.ubc.ca

Department of Computer Science
University of British Columbia
Vancouver, Canada

Nearest neighbor search is a ubiquitous problem in computer vision. Given a previously unseen query point $\mathbf{q} \in \mathbf{R}^d$, we seek its closest matches in a database $\mathbf{X} \in \mathbf{R}^{n \times d}$. One class of techniques for nearest neighbor search is *hashing algorithms* for constructing compact binary codes. Hashing algorithms transform the original data points into compact bit string signatures that require significantly less storage space and can be compared quickly using bit operations.

We can think of the bits in a binary code as the decisions of a set of hash functions or hyperplanes, possibly in some kernelized space. These hyperplanes are learned or generated by the hashing algorithm. In matrix form, we have

$$\mathbf{Y} = \text{sgn}(\mathbf{XW}) \tag{1}$$

where $\mathbf{X} \in \mathbf{R}^{n \times d}$, $\mathbf{W} \in \mathbf{R}^{d \times c}$, $\mathbf{Y} \in \{0,1\}^{n \times c}$, and c is the number of hash functions, or the number of bits in the generated binary code.

Typically, nearest neighbor search performance improves as the number of hash functions increases, i.e. as c increases. However, as the number of hash functions increases, the matrix $\mathbf{Y}$ of binary codes also increases in size, leading to higher storage requirements. For example, if we wish to improve retrieval performance by doubling the number of hash functions, we have to store binary codes that are twice the length.

In this paper, we present a novel factorized binary codes approach that uses an approximate matrix factorization of the binary codes to increase the number of hash functions while maintaining the original storage requirements. Fig. 1 illustrates the factorized binary codes approach. Given $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Y}$ as defined in Eq. (1), define a 'long' code length $c^l > c$, and form the matrix $\mathbf{W^l} \in \mathbf{R}^{d \times c^l}$, which appends $(c^l - c)$ new hash functions to the c existing hash functions in $\mathbf{W}$. The new hash functions are generated using the same procedure as the existing hash functions, according to the underlying hashing algorithm. The augmented matrix $\mathbf{W^l}$ produces 'long' binary codes $\mathbf{Y}^l \in \{0,1\}^{n \times c^l}$. We approximate $\mathbf{Y}^l$

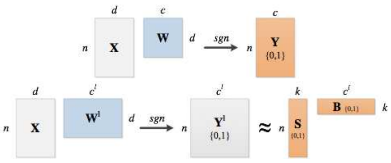


Figure 1: Graphical overview of the factorized binary codes approach

as the Boolean product of two factor matrices $\mathbf{S}$ and $\mathbf{B}$, both of which are also binary:

$$\mathbf{Y}^l \approx \mathbf{S} \circ \mathbf{B} \tag{2}$$

where $\mathbf{S} \in \{0,1\}^{n \times k}$, $\mathbf{B} \in \{0,1\}^{k \times c^l}$, $\circ$ denotes the Boolean product, and k is set such that the factor matrices require no more storage than the original binary codes $\mathbf{Y} \in \{0,1\}^{n \times c}$ (in Fig. 1, the areas highlighted in orange are the same). Given a query $\mathbf{q} \in \mathbf{R}^d$, the 'long' binary code $\mathbf{y_q} \in \{0,1\}^{c^l}$ is computed using the augmented set of hash functions $\mathbf{W}^l$ and matched with the approximate binary codes $\tilde{\mathbf{Y}}^l$ as reconstructed using $\mathbf{S}$ and $\mathbf{B}$. Fig. 2 shows experimental results on the LM+SUN dataset with 384-dimensional Gist descriptors.

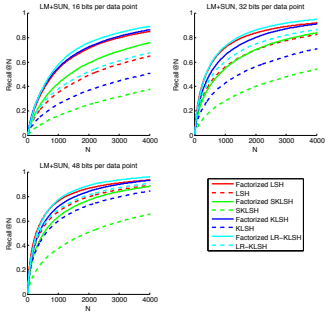


Figure 2: Experimental results on LM+SUN

Edge Enhanced Direct Visual Odometry

Tuesday
13:40-14:40

Xin Wang[1]
xinwang_cis@pku.edu.cn
Wei Dong[1]
dongwei92@pku.edu.cn
Mingcai Zhou[2]
mingcai.zhou@samsung.com
Renju Li[1]
lirenju@3dscan.com.cn
Hongbin Zha[1]
zha@cis.pku.edu.cn

[1] Key Laboratory of Machine Perception
School of EECS
Peking University
Beijing, China

[2] Advanced Research Lab
Samsung Research Center-Beijing
Beijing, China

In this paper, we propose an RGB-D visual odometry method that both minimizes the photometric error and aligns the edges between frames. The combination of the direct photometric information [1] and the edge features leads to higher tracking accuracy and enables the approach to deal with challenging texture-less scenes. In contrast to traditional line feature based methods [2], we involve all edges rather than only line segments, avoiding the aperture problem and the uncertainty of endpoints. Instead of explicitly matching edge features, we design a dense representation of edges to align them, bridging the direct methods and the feature-based methods in tracking. Image alignment and feature matching are performed in a general framework including both pixels and salient visual landmarks.

To track the camera pose, every new frame $\mathcal{F}_c$ is aligned to a reference frame $\mathcal{F}_r$ which is a carefully selected keyframe. First, the visual edges are extracted in $\mathcal{F}_c$ and $\mathcal{F}_r$. Then, error caused by camera pose at $\mathcal{F}_c$ is estimated: non-edge points p in $\mathcal{F}_r$ are reprojected to $\mathcal{F}_c$ using $\omega(p,d,\xi)$ followed by the computation of photometric error; meanwhile, edge points e_r in $\mathcal{F}_r$ are reprojected to a distance field derived from edges in $\mathcal{F}_c$ holding the minimal distance to the nearest edge point per pixel. The bottom picture of middle column in Fig.1 illustrates a distance field, whose intensity reflects the value of distance field: whiter regions are further to edges. By multiplying a weight α, we combine these two types of error and formulate an energy function. We apply Levenberg-Marquardt algorithm to minimize the proposed non-convex objective function.

Evaluations on real-world benchmark datasets show that our method achieves competitive results in indoor scenes. Especially, it outperforms the state-of-the-art algorithms in texture-less scenes.

[1] J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of ECCV*, pages 834–849, 2014.

[2] K. Hirose and H. Saito. Fast line description for line-based SLAM. In *Proceedings of BMVC*, 2012.

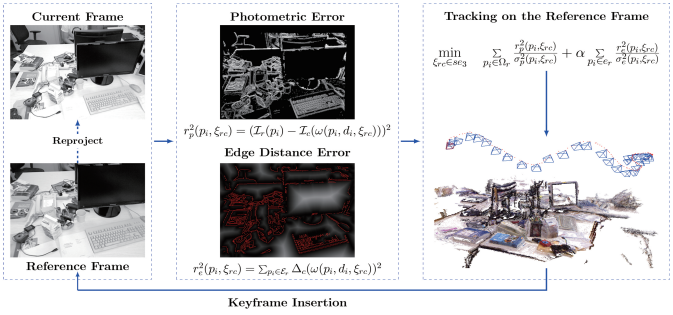


Figure 1: Overview of our approach

Optimised photometric stereo via non-convex variational minimisation

Laurent Hoeltgen[1]
hoeltgen@b-tu.de
Yvain Quéau[2]
queau@enseeiht.fr
Michael Breuß[1]
breuss@b-tu.de
Georg Radow[1]
radow@b-tu.de

[1] Chair for Applied Mathematics
BTU Cottbus-Senftenberg,
Cottbus, Germany

[2] IRIT, UMR CNRS 5505
Université de Toulouse
Toulouse, France

Estimating the shape and appearance of a three dimensional object from flat images is a challenging research topic that is still actively pursued. Among the various techniques available, Photometric Stereo (PS) is known to provide very accurate local shape recovery in terms of surface normals. In this work we propose to minimise non-convex variational models for PS that recover the depth information directly.

Photometric Stereo consists in finding a depth map z that best explains all image irradiance equations (IIEs) $I^i = \mathcal{R}(z; s^i, \rho)$, for several images I^i, considered under different lightings s^i, with $i \in \{1, \dots, m\}$. The function $\mathcal{R}$ describes our reflectance model in terms of the depth z, the lighting s^i, and the albedo ρ. We assume Lambertian reflectance, neglect shadows, and require $m \geqslant 3$.

Our approach uses a variational framework with a least-squares penalisation on the IIEs augmented with a zero-th order Tikhonov regularisation. The obtained energy (1) is non-convex and we make use of matrix differential theory and recent developments in non-convex and non-smooth optimisation to determine good minimisers.

$$\min_{z,\rho} \left\{ \mathcal{E}_{\mathcal{R}}(z, \rho; I) + \frac{\lambda}{2} \|z - z_0\|^2 \right\} \qquad (1)$$

Here, $\mathcal{E}_{\mathcal{R}}$ represents the reprojection criterion based on the IIEs.

Our numerical strategy uses recent findings of Ochs *et al*. [2]. They proposed a novel method to handle such tasks, called *iPiano*. Inspired by the heavy-ball method, it separates non-smooth and non-convex parts in an efficient splitting strategy. Further benefits include a thorough convergence theory. The algorithm makes explicit use of the derivative of the smooth terms in the cost function, which in our case involves derivatives of matrix-valued functions, and we will employ as a technical novelty, matrix differential theory [1] to derive the resulting scheme.

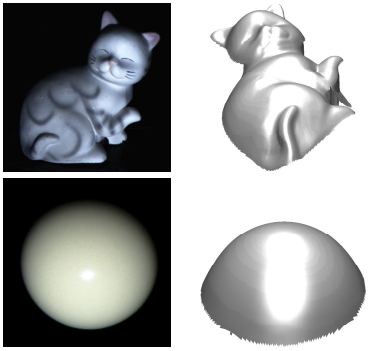


Figure 1: Test data, 3D-reconstructions obtained, 3D-reconstruction results using the *full pipeline*, consisting of a preprocessing, followed by classic PS and finally our proposed method.

Figure 1 presents some visualised results. It consists of two scenes captured under 20 different known illuminants. The experimental setups also demonstrate that our framework performs consistently better than alternative approaches such as [3] in terms of the mean angular error.

[1] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 3rd edition, 2007.

[2] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial proximal algorithm for non-convex optimization. *SIAM Journal on Imaging Sciences*, 7 (2):1388–1419, 2014.

[3] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):134–144, 1980.

U-shaped Networks for Shape from Light Field

Tuesday
13:40-14:40

Stefan Heber[1]
stefan.heber@icg.tugraz.at
Wei Yu[1]
wei.yu@icg.tugraz.at
Thomas Pock[1,2]
pock@icg.tugraz.at

[1] Graz University of Technology
Graz, Austria

[2] Austrian Institute of Technology
Vienna, Austria

Overview This paper presents a novel technique for Shape from Light Field (SfLF), that utilizes deep learning strategies. Our model is based on a fully convolutional network, that involves two symmetric parts, an encoding and a decoding part, leading to a u-shaped network architecture. By leveraging a recently proposed Light Field (LF) dataset, we are able to effectively train our model using supervised training. To process an entire LF we split the LF data into the corresponding Epipolar Plane Image (EPI) representation and predict each EPI separately. This strategy provides good reconstruction results combined with a fast prediction time. In the experimental section we compare our method to the state of the art. The method performs well in terms of depth accuracy, and is able to outperform competing methods in terms of prediction time by a large margin.

Contribution The proposed method is inspired by the method of Heber and Pock [2], that uses a conventional Convolutional Neural Network (CNN) in a sliding window fashion to predict depth information. They showed that CNNs have a large capacity to learn from data to predict the orientation of the lines in the EPIs. However, due to the sliding window approach, their method suffers from considerable high computational costs. Compared to [2] we were able to significantly reduce the computation time by predicting complete EPIs at once using u-shaped networks, cf. Figure 1. Besides drastically reducing the prediction time we were also able to reduce the errors in homogeneous regions, because the proposed model can overcome the patch-nature of the network proposed in [2]. Our experiments demonstrate that the proposed method is able to predict an entire 4D disparity field within a few seconds. Moreover, due to the fact that our network architecture does not include any fully connected layer, our method also allows to process LFs with varying resolutions.

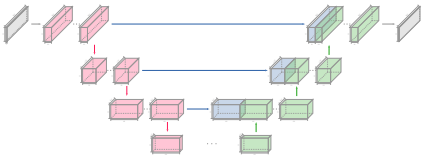


Figure 1: Illustration of the proposed u-shaped network architecture. The encoding and decoding parts of the network are highlighted in purple and green, respectively. The pinhole connections are marked in blue.

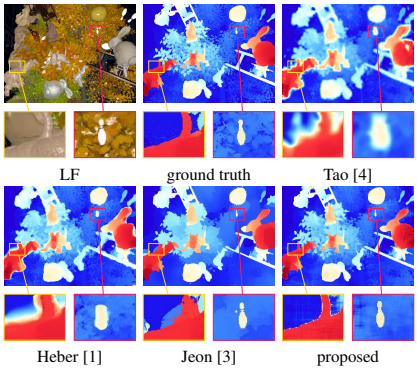


Figure 2: Comparison to state-of-the-art methods on the synthetic POV-Ray dataset.

[1] Stefan Heber and Thomas Pock. Shape from light field meets robust PCA. In *Proceedings of the 13th European Conference on Computer Vision*, 2014.

[2] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1547–1555, June 2015.

[4] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *International Conference on Computer Vision (ICCV)*, December 2013.

Using Shading and a 3D Template to Reconstruct Complex Surface Deformations

Mathias Gallardo
Mathias.Gallardo@gmail.com
Toby Collins
Toby.Collins@gmail.com
Adrien Bartoli
Adrien.Bartoli@gmail.com

ISIT, UMR 6284 CNRS
Université d'Auvergne
Clermont-Ferrand, France

Motivations The goal of Shape-from-Template (SfT) is to register and reconstruct the 3D shape of a deforming surface from a single image and a known deformable 3D template. Most SfT methods use only motion information and require well-textured surfaces which deform smoothly. Consequently they are unsuccessful for poorly-textured surfaces with complex deformations such as creases. However, Shape-from-Shading methods permit to reconstruct textureless surfaces and complex deformations since it uses all image pixels and the photometric relationship. We overcome the shortcomings of previous attempts by proposing a novel, *(i)* fully-integrated approach to combine shading constraints with SfT in order to *(ii)* reconstruct complex deformations on all visible regions, both textured and textureless, *(iii)* without any *a priori* photometric calibration.

Template, illumination and camera modeling We define the template as a texture-mapped thin shell 3D mesh in a known reference pose with M vertices. At each time t, each vertex is deformed into the unknowns 3D camera coordinates $\mathbf{x}_t \in \mathbb{R}^{3\times M}$. We upgrade the template with an photometric texture map which defines how each point of the template's surface reflects light. We assume Lambertian model and compute this map using an intensity-based segmentation of the texture-map. It gives constant albedo regions with $\alpha = \{\alpha_1, ..., \alpha_K\}$, the K unknown albedo values. The scene is illuminated by an unknown illumination $\mathbf{l}$ which is constant over time, fixed in the camera coordinates and modeled by spherical harmonics (4 and 9 coefficients). The camera has a linear response, $\beta_t \in \mathbb{R}^+$, which is unknown and time-varying.

Integrated cost function The deformation $\mathbf{x}_t$ is constrained by image data and deformation priors (*isometry* and *smoothing* constraints), and

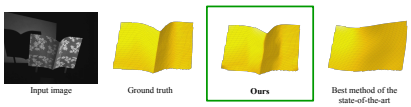


Figure 1: 3D renderings for the input image $n^\circ 6$ of the *floral plane* dataset.

$\mathbf{l}$, β_t and α are constrained by the shading term and the batch of images. We use the *shading* relationship to enforce similarity between the modeled and the measured pixel intensities. As it uses all image pixels, mis-alignement may induce errors. Thus, we use *motion* and *boundary* constraints to align the projected 3D surface with its input image. We also use a robust *smoothing* based on an *M-estimator*, which permits piecewise constant 3D reconstructions, such as creases.

Strategy solution The integrated cost function is large scale and highly non-linear, but all constraints are sparse with respect to $\mathbf{x}_t$. We use a cascaded initialization for the four types of unknowns: first $\mathbf{x}_t$, then using a batch of input images $\mathbf{l}$, β_t and finally α. Using Gauss-Newton iterations with line-search, a refinement process minimizes the whole integrated cost function for the batch of images. We found that a dense mesh with vertices of order $\mathcal{O}(10^4)$ is sufficient to capture the creases.

Experimental results We compare our approach on three datasets with four SfT methods and we see that our method is capable of capturing non-smooth deformations, better than others, as figure 1 shows, using shading without any *a priori* photometric calibration, which was not possible with previous methods in SfT or SfS.

Physics 101: Learning Physical Object Properties from Unlabeled Videos

Tuesday
13:40-14:40

Jiajun Wu[1], Joseph J. Lim[2], Hongyi Zhang[1],
Joshua B. Tenenbaum[1],
William T. Freeman[13]
[1] {jiajunwu,hongyiz,jbt,billf}@mit.edu
[2] lim@csail.mit.edu

[1] Massachusetts Institute of Technology
[2] Stanford University
[3] Google Research

Introduction We study the problem of learning physical properties of objects from unlabeled videos. Humans can learn basic physical laws when they are very young [1], which suggests that such tasks may be important goals for computational vision systems.

There have been early efforts to build computer vision systems with the physical knowledge of an early child. Recently, researchers started to tackle concrete scenarios for understanding physics from vision [2], some involving deep learning. Different from these, we aim to develop a system that can infer physical properties, *e.g.* mass and density, directly from visual input. Our method is general and easily adaptive to new scenarios, and is more efficient compared to analysis-by-synthesis approaches [3].

Physical World Model There exist highly involved physical processes in daily events in our physical world. We can divide all involved physical properties into two groups: the first is the intrinsic physical properties of objects like mass, many of which we cannot directly measure from the visual input; the second is the descriptive physical properties, *e.g.* velocity of objects, which characterize the scenario in the video. The second group of parameters are observable, and are determined by the first group, while both of them determine the content in videos.

Physics 101 Dataset We collected a dataset of 101 objects made of different materials and with various masses and volumes. We started by collecting videos of them from multiple viewpoints in four scenarios: objects slide down an inclined surface and possibly collide with another object; objects fall onto surfaces made of different materials; objects splash in water; and objects hang on a spring. These seemingly straightforward setups require understanding multiple physical properties, *e.g.*, material, mass, volume, density, coefficient of friction, and coefficient of restitution. We called this dataset Physics 101.

Method Our method is a CNN consisting of three components. The bottom component is a *visual property discoverer*, which aims to discover physical properties like material or volume which could at least partially be observed from visual input; the middle component is a *physics interpreter*, which explicitly encodes physical laws into the network structure and models latent physical properties like density and mass; the top component is a *physical world simulator*, which characterizes descriptive physical properties like distances that objects traveled, all of which we may directly observe from videos. Our network corresponds to our physical world model, and learns object properties from unlabeled data.

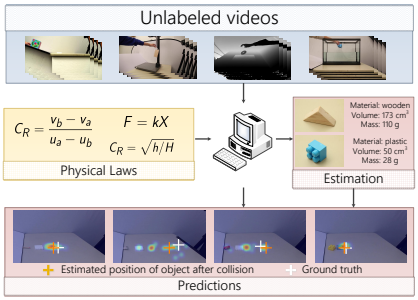


Figure 1: Overview of our model, which learns directly from unlabeled videos, produces estimates of physical properties of objects based on the encoded physical laws, and generalizes to tasks like outcome prediction

Evaluation We demonstrate that our framework develops some physics competency by observing videos. We also show that our generative model can transfer learned physical knowledge from one scenario to the other, and generalize to other tasks like predicting the outcome of a collision.

[1] Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004.

[2] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *PNAS*, 110(45):18327–18332, 2013.

[3] Jiajun Wu, Ilker Yildirim, Joseph J Lim, William T Freeman, and Joshua B Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NIPS*, 2015.

Attribute Embedding with Visual-Semantic Ambiguity Removal for Zero-shot Learning

Yang Long[1]
ylong2@sheffield.ac.uk
Li Liu[2]
li2.liu@northumbria.ac.uk
Ling Shao[2]
ling.shao@ieee.org

[1] Department of Electronic and Electrical Engineering
The University of Sheffield
Sheffield , UK

[2] Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, UK

Conventional *zero-shot learning* (ZSL) methods recognise an unseen instance by projecting its visual features to a semantic space that is shared by both seen and unseen categories [1, 2]. However, we observe that such a one-way paradigm suffers from the *visual-semantic ambiguity* problem. As shown in Fig. 1, the semantic concepts (e.g. attributes or classes) cannot explicitly correspond to visual patterns, and similar visual features may come from different classes. Such a problem can lead to a huge variance in the visual features for each attribute.

In this paper, we propose the ***V**isual-**S**emantic **A**mbiguity **R**emoval* (VSAR) algorithm to address such a problem. In particular, we propose a novel latent attribute space $\mathcal{V}$ to mitigate the gap between visual and semantic spaces $\mathcal{X}$ and $\mathcal{A}$:

$$J = \|\mathcal{X} - U_1\mathcal{V}\|_F^2 + \alpha\|\mathcal{A} - U_2\mathcal{V}\|_F^2 + \lambda\mathcal{R}, \quad (1)$$

where U_1 and U_2 are two projection matrices. $\mathcal{R}$ is a *Dual-graph* regularisation that combines two supervised graphs $W_\mathcal{X}$ and $W_\mathcal{A}$ that model the intrinsic data structures in $\mathcal{X}$ and $\mathcal{A}$. In the embedding space $\mathcal{V}$, we expect that if the vertices in both graphs are connected, each pair of embedded points v_i and v_j are also closed to each other. However, for the *visual-semantic ambiguity* problem, $W_\mathcal{X}$ and $W_\mathcal{A}$ usually give contradictory results. To compromise such conflict, we linearly combine the two graphs, i.e. $W_{ij} = W_{\mathcal{X}_{ij}} + \alpha W_{\mathcal{A}_{ij}}$. The resulted regularisation is:

$$\mathcal{R} = \frac{1}{2}\sum_{i,j=1}^{N}\|v_i - v_j\|^2 W_{ij} = Tr(\mathcal{V}L\mathcal{V}^T), \quad (2)$$

where D is the degree matrix of W, $D_{ii} = \sum_i w_{ij}$. L is known as graph Laplacian matrix $L = D - W$ and $Tr(.)$ computes the trace of a matrix.

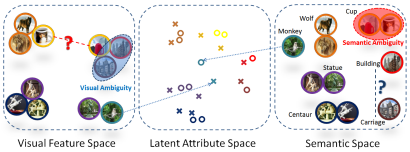


Figure 1: *Visual Ambiguity* (in blue oval): the image of a carriage is taken with a building background. It cannot recover the semantic distance (blue question mark) to the building category. *Semantic Ambiguity* (in red oval): the cup printed with a wolf and the cup-like building share the same name which can lead to a large visual variance (the red question mark). After embedding to the latent attribute space using VSAR, such ambiguity is mitigated.

Once we obtain the latent attribute embedding $\mathcal{V}$ of the seen data, performing zero-shot recognition is straightforward via *least-square approximation* between $\mathcal{V}$ and $\{\mathcal{A}, \mathcal{X}\}$. During the test, given unseen category names and their attributes in pairs: $\{\mathcal{Y}_u, \mathcal{A}_u\}$. We firstly embed all unseen attributes $\mathcal{A}_u$ into the latent embedding space as references: $\mathcal{V}_u = \mathcal{V}\mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}_u$. Given a test unseen instance $\hat{x}$, its embedded latent attribute representation is: $\hat{v} = \mathcal{V}\mathcal{X}^T(\mathcal{X}\mathcal{X}^T)^{-1}\hat{x}$. Finally, we adopt a simple NN classifier to predict the category label $\hat{c}$:

$$\hat{c} = \arg\min_c \|\hat{v} - v_c\|^2, \text{ where } v_c \in \mathcal{V}_u. \quad (3)$$

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

NRSfM-Flow: Recovering Non-Rigid Scene Flow from Monocular Image Sequences

Tuesday
13:40-14:40

Vladislav Golyanik[1,2]
http://av.dfki.de/members/golyanik/
Aman Shankar Mathur[1]
asmathur@rhrk.uni-kl.de
Didier Stricker[1,2]
http://av.dfki.de/members/stricker/

[1] Department of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany

[2] Department Augmented Vision
German Research Center for Artificial Intelligence (DFKI GmbH)
Kaiserslautern, Germany

Recovery of scene flow (a dense 3D velocity vector field) of a dynamic scene from monocular image sequences is an emerging field in computer vision. Being sensitive to occlusions, existing Monocular Scene Flow (MSF) methods are either limited in handling non-rigid deformations [5], or make strong assumptions on scene [2] and camera motion [1]. To overcome these limitations, we propose a framework for MSF estimation based on Non-Rigid Structure from Motion (NRSfM) [4] techniques — NRSfM-Flow. In the continuous domain, relation between a shape $\mathbf{S}(\mathbf{p},t)$, camera motion $\mathbf{R}(t)$ and scene flow $\Theta(\mathbf{p},t)$ can be expressed as

$$\Theta(\mathbf{p},t) = \frac{\partial \mathbf{R}(t)}{\partial t}\mathbf{S}(\mathbf{p},t) + \mathbf{R}(t)\frac{\partial \mathbf{S}(\mathbf{p},t)}{\partial t}. \quad (1)$$

To enhance reconstruction accuracy and speedup computations, two preprocessing steps are proposed — Translation Resolution (TR) and Redundancy Removal (RR). With TR, translation of the scene is resolved using a sparse point tracker. Using RR, frames with insufficient diversity are dropped according to the criterion

$$\left\| \int_{\hat{\Psi}} \int_{t_a}^{t_b} \Xi(\mathbf{v},t)\, dt\, d\hat{\mathbf{v}} \right\|_2 \geq \varepsilon, \quad (2)$$

where $\Xi(\mathbf{v},t)$ is a continuous optical flow function, $\hat{\mathbf{v}} \in \hat{\Psi} \subset \mathbb{R}^2$ are 2D points observed at a reference time τ, and ε is a scalar threshold.
Our approach can handle long image sequences with non-rigid deformations and self-occlusions, with no strong assumptions such as a known camera motion. Performance is demonstrated on several synthetic and real image sequences (see Fig. 1 for an example). With this paper we hope, on the one hand, to draw attention to model-based approaches for MSF estimation and, on the other, to highlight importance of the differential interpretation of the NRSfM problem.

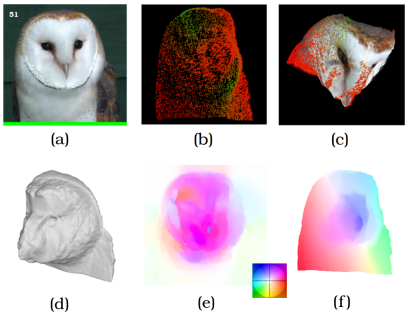


Figure 1: Experimental results on the *barn owl* sequence [3]: (a) frame 51; (b) scene flow between frames 51 and 52; (c) geometry + scene flow; (d) shaded geometry (Poisson) from a novel viewpoint; (e) optical flow between frames 51 and 52; (f) projection of the 3D motion field into the image plane.

[1] N. Birkbeck, D. Cobzaş, and M. Jägersand. Depth and scene flow from a single moving camera. In *3D Data Processing Visualization and Transmission (3DPVT)*, 2010.

[2] N. Birkbeck, D. Cobzaş, and M. Jägersand. Basis constrained 3d scene flow on a dynamic proxy. In *International Conference on Computer Vision (ICCV)*, pages 1967–1974, 2011.

[3] P. Dinning. *Barn Owl at Screech Owl Sanctuary*. `https://www.youtube.com/watch?v=xmou8t-DHh0`, 2014. [online; accessed on 12.05.2016; usage rights obtained from the author].

[4] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279, 2013.

[5] D. Xiao, Q. Yang, B. Yang, and W. Wei. Monocular scene flow estimation via variational method. *Multimedia Tools and Applications (An International Journal)*, pages 1–23, 2015.

Better Together: Joint Reasoning for Non-rigid 3D Reconstruction with Specularities and Shading

Qi Liu-Yin[1]
Qi.Liu@cs.ucl.ac.uk
Rui Yu[1]
R.Yu@cs.ucl.ac.uk
Andrew Fitzgibbon[2]
awf@microsoft.com
Lourdes Agapito[1]
L.Agapito@cs.ucl.ac.uk
Chris Russell[1]
crussell@turing.ac.uk

[1] University College London
London, UK
[2] Microsoft Research Cambridge
Cambridge, UK

In this paper, we demonstrate the use of shape-from-shading (SfS) to improve both the quality and the robustness of 3D reconstruction of dynamic objects captured by a single camera. Unlike previous approaches that made use of SfS as a post-processing step, we offer a principled integrated approach that solves dynamic object tracking and reconstruction and SfS as a single unified cost function. Moving beyond Lambertian SfS, we propose a general approach that models both specularities and shading while simultaneously tracking and reconstructing general dynamic objects. Solving these problems jointly prevents the kinds of tracking failures which can not be recovered from by pipeline approaches.

Figure 1: The reflected intensity is the product of albedo and diffuse shading plus specularities.

Our proposed approach is an online template-based method that captures both the 3D geometry and the reflectance properties (Figure 1) of the non-rigid object. Our main novelty is the photometric error data term of the energy cost that is minimized for each new frame. It models the photometric error as follows

$$E_D = \sum_{i \in \mathcal{V}} \left\| \mathbf{I}(\pi(\mathbf{R}(\mathbf{s}_i)+\mathbf{t})) - \widehat{\boldsymbol{\rho}}_i \mathbf{l} \cdot Y(\mathbf{R}(\mathbf{n}_i(\boldsymbol{S}))) - \boldsymbol{\beta}_i \right\|_\varepsilon$$

For each vertex, it penalizes the difference between its projected and its estimated intensities as a function of albedo $\widehat{\boldsymbol{\rho}}$, diffuse shading $\mathbf{l} \cdot Y(\cdot)$ and specular highlights $\boldsymbol{\beta}$.

We tested our method on synthetically rendered sequences, using the results from [1], and on real sequences. We compare against [2] and show state-of-the-art results both qualitatively (Figure 2) and quantitatively (Table 1).

Figure 2: From top to bottom: synthetic input sequence, results from Yu *et al*., and our results.

	LF	SF	LC	SC
Yu *et al*. [2]	7.29	7.93	9.18	9.28
Ours	2.73	2.89	3.42	3.84

Table 1: Comparison of RMS error (in mm.) with Yu *et al*. on 4 different synthetic sequences.

[1] Levi Valgaerts et al. Lightweight binocular facial performance capture under uncontrolled lighting. *SIGGRAPH Asia*, 2012.

[2] Rui Yu et al. Direct, dense, and deformable: Non-rigid 3d reconstruction from rgb video. *ICCV*, 2015.

Tuesday
13:40-14:40

STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition

Wei Liu[1]
wliu@cs.hku.hk
Chaofeng Chen[1]
cfchen@cs.hku.hk
Kwan-Yee K. Wong[1]
kykwong@cs.hku.hk
Zhizhong Su[2]
suzhizhong@baidu.com
Junyu Han[2]
hanjunyu@baidu.com

[1] Department of Computer Science
The University of Hong Kong, HK

[2] Institution of Deep Learning
Baidu Inc, Beijing

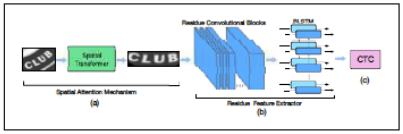
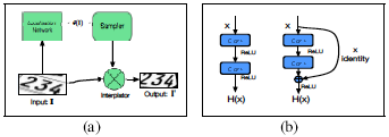


Figure 1: Overview of our STAR-Net for scene text recognition. (a) Spatial attention mechanism. (b) Residue feature extractor. (c) Connectionist Temporal Classification.



(a) (b)

Figure 2: Structures of the spatial transformer, plain and residue convolutional blocks.

In this paper, we present a novel SpaTial Attention Residue Network (**STAR-Net**) for recognising scene texts. The overall architecture of our STAR-Net is illustrated in fig. 1. Our STAR-Net emphasises the importance of representative image-based feature extraction from text regions by the spatial attention mechanism and the residue learning strategy. It is by far the deepest neural network proposed for scene text recognition.

Spatial Attention Mechanism The spatial transformer [2] is responsible for introducing the spatial attention mechanism (see fig. 2(a)). A localisation network is used to determine the transformation parameters $\theta(I)$ of the original text image. Based on these parameters, a sampler locates sampling points on the input image which explicitly define the text region to be unwarped. Finally, an interpolator generates the output image by interpolating the intensity values of the four pixels nearest to each sampling point.

Residue Learning Strategy To fully exploit the potential of convolutional layers and build up a powerful deep feature encoder, we employ residue convolutional blocks [1] (see fig. 2(b)) along with Long Short-Term Memory to extract informative features from the rectified text regions.

Experimental Results Experiments conducted on the SVT-Perspective dataset show that our STAR-Net outperforms other state-of-the-art methods on both lexicon-based and lexicon-free recognition. Besides STAR-Net, we also evaluate three other network architectures (CRNN, STA-CRNN and R-Net) to demonstrate the effectiveness of each component in our STAR-Net. More details are in the experiment part of the paper.

Wang *et al.*	40.5	26.1	-
Mishra *et al.*	45.7	24.7	-
Wang *et al.*	40.2	32.4	-
Phan *et al.*	75.6	67.0	-
Shi *et al.*	91.2	77.4	71.8
CRNN	92.6	72.6	66.8
STA-CRNN	93.0	80.5	69.3
R-Net	93.0	83.6	70.9
STAR-Net	**94.3**	**83.6**	**73.5**

Table 1: Scene text recognition accuracies (%) on SVT-Perspective dataset.

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[2] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

Context Matters: Refining Object Detection in Video with Recurrent Neural Networks

Subarna Tripathi[1]
http://acsweb.ucsd.edu/~stripath/
Zachary C. Lipton[1]
zlipton@cs.ucsd.edu
Serge Belongie[2,3]
sjb344@cornell.edu
Truong Nguyen[1]
tqn001@eng.ucsd.edu

[1] University of California San Diego
La Jolla, CA, USA
[2] Cornell University
Ithaca, NY, USA
[3] Cornell Tech
New York, NY, USA

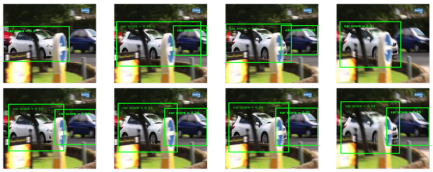


Figure 1: RNN (bottom) recognizes multiple objects more accurately than a state of the art frame-level model (top).

In this paper, we introduce a framework for improving object detection in videos by capturing temporal context and encouraging temporally consistent predictions. First, we train a *pseudo-labeler*, that is, a domain-adapted convolutional neural network for object detection. The pseudo-labeler is first trained individually on the subset of labeled frames, and then subsequently applied to all frames. Then we train a recurrent neural network (RNN) that takes as input sequences of pseudo-labeled frames and optimizes an objective that encourages both accuracy on the target frame and consistency across consecutive frames.

The approach incorporates strong supervision of target frames, weak-supervision on context frames, and regularization via a smoothness penalty. Building on YOLO, a domain-adapted frame-level object detection model [3], we demonstrate that for the sparsely annotated *YouTube Objects* dataset [2], our method achieves mean Average Precision (mAP) of 68.73 on test data, as compared to a best published result of 37.41 [4] and 61.66 for YOLO alone.

As with YOLO [3], our fine-tuned *pseudo – labeler* takes 448×448 frames as input and regresses on category types and locations of possible objects at each one of 7×7 non-overlapping grid cells. For each grid cell, the model outputs class conditional probabilities as well as 2 bounding boxes and their associated confidence scores.

Then, to incorporate temporal context, we train an RNN with gated recurrent units (GRUs) [1] to refine the provisional predictions. This net takes as input sequences of *pseudo-labels*. For this recurrent model, we demonstrate an efficient and effective training strategy. The objective encourages predictions to be close to true labels (for labeled frames), not to deviate too far from the pseudo-labels, and to be similar across adjacent frames. As demonstrated experimentally, our framework proves effective, achieving state-of-the art mAP and producing compelling visual examples.

[1] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[2] Alessandro Prest, Vicky Kalogeiton, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Youtube-objects dataset v2.0, 2014. URL `calvin.inf.ed.ac.uk/datasets/youtube-objects-dataset`. University of Edinburgh (CALVIN), INRIA Grenoble (LEAR), ETH Zurich (CALVIN).

[3] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[4] Subarna Tripathi, Serge J. Belongie, Youngbae Hwang, and Truong Q. Nguyen. Detecting temporally consistent objects in videos through object class label propagation. *WACV*, 2016.

Outlier Rejection for Absolute Pose Estimation with Known Orientation

Tuesday
13:40-14:40

Viktor Larsson[1]
viktorl@maths.lth.se
Johan Fredriksson[1]
johanf@maths.lth.se
Carl Toft[2]
carl.toft@chalmers.se
Fredrik Kahl[1,2]
fredrik@maths.lth.se

[1] Centre for Mathematical Sciences,
Lund University,
Sweden

[2] Department of Signals and Systems,
Chalmers University of Technology,
Sweden

In this paper we present an outlier rejection method for absolute pose estimation. We focus on the special case when the orientation of the camera is known. The problem is solved by projecting to a lower dimension where we are able to efficiently compute upper bounds on the maximum number of inliers. The method guarantees that only correspondences which cannot belong to an optimal pose are removed. Once the majority of the outliers have been removed the problem is greatly simplified and can be solved using standard methods (e.g. RANSAC [1]).

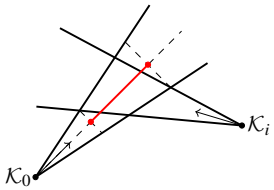
If the orientation is known we can w.l.o.g. assume that $R = I$, by rotating the image points. Each 2D-3D correspondence then constrains the translation $\mathbf{t}$ to a cone,

$$\mathcal{K} = \left\{ \mathbf{t} \in \mathbb{R}^3 \;\middle|\; \|\mathbf{X} - \mathbf{t}\| \leq \frac{1}{\cos(\varepsilon)} \langle \mathbf{x}, \mathbf{X} - \mathbf{t} \rangle \right\}, \quad (1)$$

and we are interested in finding the translation which satisfies as many cones as possible.

To remove outliers we want to determine if a given cone $\mathcal{K}_0$ can be part of an optimal solution. If the answer is negative we can discard the correspondence safely.

In our outlier rejection scheme we first orthogonally project each intersection $\mathcal{K}_0 \cap \mathcal{K}_i$ to the center line in $\mathcal{K}_0$.



Each intersection gives an interval on the line and by finding the maximum number of overlapping intervals we get an upper bound for any solution which includes $\mathcal{K}_0$.

Finding the projection of the intersection between the cones is a convex problem and can be solved using standard solvers. For our application these are however too slow for practical use. Instead we form a polyhedral approximation of the cone $\mathcal{K}_0$. This allows us to find a closed form solution to the projection problem. In experiments we show that the errors introduced by the planar approximation are negligible and that the closed form solution gives significant speed-ups compared to using the standard solvers. For some instances the runtime went from 20-30 minutes to a couple of milliseconds.

We evaluate our method on a new dataset for metric localization from a single image for a car driving through a tunnel. The poor lightning conditions and repetitive textures makes matching difficult and there are a large number of outliers. See Figure 1 for an example image and localization result.

Figure 1: *Left:* Input image with SIFT features (blue points). *Right:* Camera pose in world coordinate frame.

We compare running our outlier rejection followed by a few iterations of RANSAC with performing RANSAC on original correspondences. Our approach gives improved performance, both in terms of localization accuracy and computation time.

[1] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, US, 2011.

Learning from scratch a confidence measure

Matteo Poggi
http://vision.disi.unibo.it/~mpoggi
Stefano Mattoccia
http://vision.disi.unibo.it/~smatt

University of Bologna
Department of Computer Science and Engineering (DISI)
Viale del Risorgimento 2
Bologna, Italy

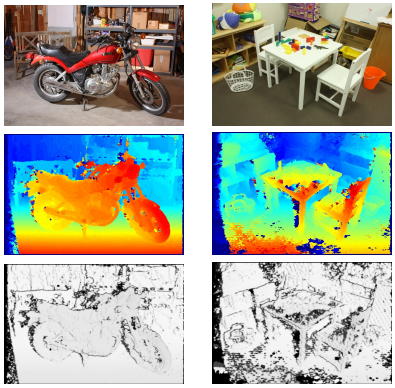


Figure 1: Reference image, disparity map and confidence computed by CCNN for *Motorcycle* and *Playtable* frames of the Middlebury 2014 training dataset.

In this paper, we propose a novel approach, referred to as Confidence Convolutional Neural Network (CCNN)[1], to predict the correctness of stereo matching by deploying a Convolutional Neural Network (CNN). In literature this is usually carried out by means of confidence measures [1] which encode the degree of reliability of the disparity assigned to each pixel by considering different cues: cost volume, reference image, disparity map and so on. Although some standalone measures are quite effective [1], recent works proved that combining a pool of them, within a machine learning framework, enables to significantly improve the overall effectiveness. In particular, Park & Yoon [2] represents state-of-the-art in this field, obtaining the best results according to the Area Under the Curve (AUC) evaluation protocol defined by Hu and Mordohai [1].

This paper proposes the first method that allows to obtain a confidence measure inferred from scratch by a CNN deploying as input cue only the disparity map computed by a stereo algorithm. This strategy makes our proposal suited even for out-of-the-box 3D sensors that typically do not provide the cues required by other methods.

[1]The source code of CCNN and the trained network is available here: `http://vision.disi.unibo.it/~mpoggi`

Dataset/Alg.	Opt.	Park&Yoon	CCNN
KITTI/BM	0.137	0.179	**0.175**
KITTI/SGM	0.038	0.124	**0.099**
Middl./BM	0.093	0.114	**0.107**
Middl./SGM	0.042	0.093	**0.074**

Table 1: Average AUC on KITTI 2015 and Middlebury 2014 training datasets with BM and SGM algorithms. The lower, the better. Values closer to optimum are in bold.

For a fair comparison, we trained the proposed CCNN and Park & Yoon [2] on KITTI 2012 (more than 6 million samples), using the Block Matching stereo algorithm (BM). This provides more than 6 million samples for training. Then, we evaluated CCNN and Park & Yoon on KITTI 2015 training dataset processing the output of BM and Semi-Global Matching (SGM). We also cross-evaluated the two approaches, with BM and SGM stereo algorithms, on Middlebury 2014 training dataset. Table 1 reports average AUCs for CCNN and Park & Yoon, computed on KITTI 2015 and Middlebury 2014, for BM and SGM, in order to assess their effectiveness. Observing the table, we can notice that our proposal always outperforms state-of-the-art.

[1] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 2121–2133, 2012.

[2] Min-Gyu Park and Kuk-Jin Yoon. Leveraging stereo matching with learning-based confidence measures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Localizing Periodicity in Time Series and Videos

Tuesday
13:40-14:40

Giorgos Karvounas[1, 2]
gkarv@ics.forth.gr
Iason Oikonomidis[1]
oikonom@ics.forth.gr
Antonis A. Argyros[1, 3]
argyros@ics.forth.gr

[1] Institute of Computer Science, FORTH, Heraklion, Crete, Greece
[2] TEI of Crete, Department of Informatics Engineering, Greece
[3] Computer Science Department, University of Crete, Greece

Periodic patterns and motions are ubiquitous in both natural and man-made environments. Several well established tools and techniques such as the Fourier Transform [2] can be used to analyse purely periodic signals. However, in many real life scenarios, periodic signals appear as segments of larger signals containing non-periodic parts. The detection and characterization of such periodic parts is an interesting problem that is not yet fully addressed.

In this work we propose a method that, given a time series representing a periodic signal that has a non-periodic prefix and tail, estimates the start, end and period of the periodic part of the signal. The resulting method has a small number of free parameters, is unsupervised and can detect short periodic events occurring in the context of extended non-periodic activities.

Consider as input a univariate time series $\mathbf{x} = < x_1, x_2, \ldots, x_N >$. Assuming that this time series is periodic between times b and e and that the period of that part of the signal is l, our goal is to estimate b, e and l. We formulate this task as an optimization problem in a search space defined by b, e and l. A candidate triplet (b,e,l) defines $n = \lfloor (e-b)/l \rfloor$ segments: $s_i = < x_{b+l\cdot(i-1)}, \ldots, x_{b+l\cdot i-1} >, i \in \{1 \ldots n\}$. We quantify the total dissimilarity of these segments as the mean squared error among all pairs of segments: $\varepsilon_s(l) = \frac{1}{n\cdot l} \sum_{i=1}^{n} \sum_{j=i+1}^{n} ||\mathrm{s_i} - \mathrm{s_j}||_2^2$, where $||\cdot||_2^2$ denotes the squared L_2 norm. Based on this quantity, we formulate an appropriate objective function that is optimized using Particle Swarm Optimization (PSO) [1]. PSO is a stochastic method that iteratively improves a candidate solution with regard to a given measure of quality.

The core of the proposed framework is a method that, given a univariate time series containing a periodic part, detects the start, the end and the period length of that part. In practice, several phenomena can be more effectively represented as multivariate time series. We consider multivariate time series as a set of synchronized, univariate time series. We apply the core periodicity detection method to each of them. Then, we employ a simple yet effective voting method to aggregate partial results towards characterizing the periodicity of the event that is represented with the multivariate time series.

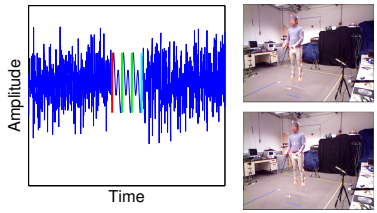


Figure 1: Left: the proposed method aims to find a periodic part within a larger, non-periodic signal. Right: indicative result on a video. Our method detects the start, the end and the period of the periodic motion (jumping). Top frame: The highest point of the jump, begin of an estimated period. Bottom frame: The highest point of the next jump, begin of next period.

We present the results we obtained using the proposed method. We first evaluate the performance of the method on synthetically generated sequences, determining appropriate parameters for PSO. Given these parameters, we evaluate the performance of the method under the presence of varying amounts of noise. Finally, we present results in real-world data. Specifically, we present results on detecting periodic activities using motion capture or video data as input.

[1] Maurice Clerc and James Kennedy. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on*, 6(1):58–73, 2002.

[2] Jean Baptiste Joseph Fourier. *The analytical theory of heat*. The University Press, 1878.

Person Re-id in Appearance Impaired Scenarios

Mengran Gou
http://www1.coe.neu.edu/~mengran/
Xikang Zhang
zhangxk@ece.neu.edu
Angels Rates-Borras
ratesborras.a@husky.neu.edu
Sadjad Asghari-Esfeden
www1.coe.neu.edu/~sadjad/
Octavia Camps
camps@coe.neu.edu
Mario Sznaier
msznaier@coe.neu.edu

Robust Systems Lab
Electrical and Computer Engineering
Northeastern University
Boston

(a) (b)

Figure 1: (a) Examples of persons wearing black suits; (b) Examples of images of the same person but wearing different clothing.

Person re-identification is critical in surveillance applications. Current approaches rely on appearance-based features extracted from a single or multiple shots of the target and candidate matches. These approaches are at a disadvantage when trying to distinguish between candidates dressed in similar colors (Figures 1(a)) or when targets change their clothing (Figures 1(b)). In this paper we propose a dynamics-based feature to overcome this limitation. **The main contributions of this paper are:** (i) A novel dynamics-based and Fisher vector encoded feature DynFV for re-id; (ii) Three new challenging "appearance impaired" datasets for re-id performance evaluation; and (iii) A comprehensive evaluation of the effect of choosing different spatio, spatio-temporal, and dynamics-based features on the performance of (unsupervised) re-id methods.

We propose to use soft-biometric characteristics provided by *sets of dense, short trajectories (tracklets)*, which have been shown to carry useful invariants [1]. All tracklets are encoded by using *pyramids of dense trajectories* with *Fisher vector encoding* [2], as illustrated in Figure 2 and described in detail in the paper.

To illustrate the need for dynamic-based features we collected three challenging "appearance-impaired" datasets. Two of them consist of video sequences of people wearing black/dark clothing. They are subsets

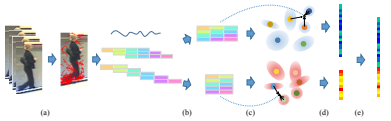
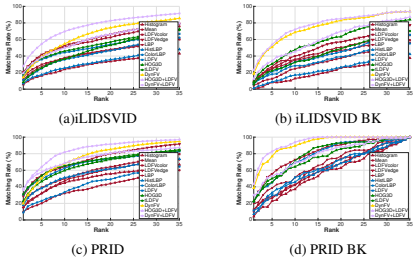


Figure 2: Pipeline of the proposed dynamics-based feature extraction.

of the iLIDS-VID and PRID 2011 datasets and we named them **iLIDSVID BK** and **PRID 2011 BK**, respectively. The third dataset, named the **Train Station dataset (TSD)**, has sequences of persons with different clothing and accessories (Figure 1(b)). We compare unsupervised re-id performance when using different combinations of five different types of features. After combining LDFV and DynFV, the rank-1 accuracies have relative improvements of 142.1% on average for all three new datasets (Figure 3).



(a)iLIDSVID (b) iLIDSVID BK

(c) PRID (d) PRID BK

Figure 3: CMC curves for iLIDSVID, PRID and the BK extension datasets

[1] B. Li, O. Camps, and M. Sznaier. Cross-view activity recognition using hankelets. In *CVPR*, pages 1362–1369, 2012.

[2] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156. 2010.

Learning to Detect and Match Keypoints with Deep Architectures

Tuesday
13:40-14:40

Hani Altwaijry[1,2]
haltwaijry@cs.cornell.edu
Andreas Veit[1,2]
aveit@cs.cornell.edu
Serge Belongie[1,2]
sjb344@cornell.edu

[1] Cornell University
Ithaca, NY, USA
[2] Cornell Tech
New York, NY, USA

In computer vision, the extraction of effective features for the detection and description of important image regions is a key step for many applications. Traditionally, these features are extracted using hand engineered detectors and descriptors. Approaches adopting this paradigm are generally referred to as *keypoint-based* or *feature-based* approaches. Recently, the reintroduction of neural networks into many computer vision tasks broadly replaced hand-engineered feature-based approaches. Neural network based approaches generally learn the feature extraction as part an end-to-end pipeline. While these approaches have shown great success in tasks such as object detection and classification, other tasks such as structure-from-motion (SfM) still depend on purely engineered features, *e.g.* SIFT, to detect and describe keypoints.

In this paper, we propose a model that learns what constitutes a good keypoint, is capable of capturing keypoints at multiple scales and learns to decide whether two keypoints match. We achieve multiscale keypoint detection with a fully-convolutional network that recursively applies convolutions to regresses keypoint scores. With each successive convolution, the network evaluates image patches, i.e., keypoints, at a larger scale. By extracting the keypoint feature map after each convolution we obtain a feature map that resembles a keypoint scale-space. To learn descriptors for keypoint matching, we leverage a triplet network to learn an embedding where patches of matching keypoints are closer to each other than non-matching patches. Figure 1 provides an overview of our proposed model.

There is currently no large-scale dataset for learning both keypoint detectors and descriptors from image patches. Furthermore, finding training examples to train deep neural networks for this task poses a serious challenge, as collecting human annotated examples would be prohibitively expensive. Therefore, we create our own dataset by following a self-supervised approach. We utilize SfM to construct a large-scale model of 1.3 million 3D points, which are used to extract matching patches with varying photometric properties such as scale, illumination, perspective. Although those feature detections and matches were determined originally with engineered features, SfM factors in the underlying geometry. This allows to learn features that extend upon their engineered counterparts.

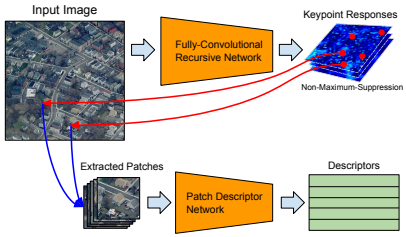


Figure 1: Proposed architecture for learning to detect and describe keypoints at multiple-scales.

We evaluate the proposed model both quantitatively and qualitatively and show its capablity of identifying multiscale keypoints as well as matching them. We show that the descriptors outperform previous approaches and demonstrate the transferability to unseen datasets with different statistics; Figure 2 shows an example.

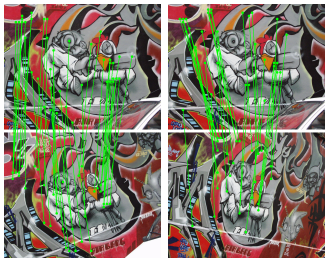


Figure 2: Qualitative evaluation on "Wall' image from the Oxford dataset.

Supervised Incremental Hashing

Bahadir Ozdemir[1]
ozdemir@cs.umd.edu
Mahyar Najibi[1]
najibi@cs.umd.edu
Larry S. Davis[2]
lsd@umiacs.umd.edu

[1] Department of Computer Science
University of Maryland
College Park, MD 20742

[2] Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

This paper presents an incremental strategy for learning hash functions with kernels for large-scale image search. Despite the fact that new images are added to online photo databases every day, no supervised hashing method learns hash functions incrementally for newly added images. We identified three main objectives for our supervised hashing method – as being incremental and parallelizable, avoiding overfitting by better generalization, and balancing $+1/-1$ in learned binary codes.

To address this problem, we introduce *Supervised Incremental Hashing* (SIH); a method based on binary and multi-class SVMs. SIH treats binary codes as intermediate variables between the feature space and the semantic space. In the first stage of classification, binary codes are considered as class labels by a set of binary SVMs; each corresponds to one bit. In the second stage, binary codes become the input space of a multi-class SVM. We formulate our hashing objectives in a joint optimization task that provides better generalization with regularizations and maximizes the entropy by balancing binary codes.

We describe an algorithm that solves the optimization problem efficiently by an incremental strategy. In this approach, the NP-hard problem of finding optimal binary codes is solved via cyclic coordinate descent and the SVMs are trained in a parallel fashion. Considering a dataset with class information, SIH can be adapted to modifications like adding new classes, deleting existing classes, and adding images to existing classes efficiently. Furthermore, we present an upper bound for the convergence of our method when these changes happen.

Figure 1 shows a simulation of our incremental SVM approach on a sample dataset with 6 classes represented by colors in part (a). Assignments of $+1$ and -1 are indicated by filled and empty shapes, respectively. Part (b) shows the changes in the hyperplanes during training.

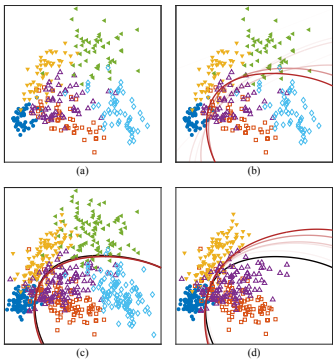


Figure 1: A simulation of SIH.

Increasing transparency of the lines indicates earlier iterations of the execution. In part (c), the number of data points is increased to 600 by adding new points from the same distributions in (b). Red lines represent the hyperplanes when our method is initialized with the solution at (b) shown by a black line. Finally, in part (d) two classes are deleted from the dataset.

We evaluate our method on three large-scale datasets: CIFAR10, MNIST, and NUS-WIDE. Our method outperforms the state-of-the-art hashing methods in retrieval performance while it has competitive execution time. The significance of our incremental strategy is observed when it is applied on dynamic datasets where new images are added and existing images are delete. Our incremental hashing strategy reaches the same retrieval performance as the from-scratch hashing strategy while requiring shorter training time.

Experiments validate that the incremental hashing strategy for dynamic datasets is capable of updating hash functions efficiently. Besides, the proposed approach provides higher quality codes with well-balanced bits and better generalization.

Multi-Scale Fully Convolutional Network for Fast Face Detection

Tuesday
13:40-14:40

Yancheng Bai[1]
yancheng@iscas.ac.cn
Wenjing Ma[1]
wenjing@iscas.ac.cn
Yucheng Li[1]
yucheng@iscas.ac.cn
Liangliang Cao[2]
liangliang.cao@gmail.com
Wen Guo[3]
grewen@126.com
Luwei Yang[4]
luweiy@sfu.ca

[1] Institute of Software, Chinese Academy of Science, Beijing, China
[2] Columbia University and Yahoo Labs New York, USA
[3] Shandong Technology and Business University, Shandong, China
[4] Simon Fraser University Vancouver, Canada

Motivation. Image pyramid is a common strategy in detecting objects with different scales in an image. The computation of features at every scale of a finely-sampled image pyramid is the computational bottleneck of many modern face detectors.

Contributions. In this paper, we propose a new architecture of fully convolutional network framework for fast face detection. In our detector, face models at different scales are trained end-to-end and simultaneously. And more importantly, different scale models share the same convolutional feature maps. During testing, only images at octave-spaced scale intervals need to be processed by our detector. And faces of different scales between two consecutive octaves can be detected by multi-scale models in our system. This makes our detector very efficient and can run about 100 FPS on a GPU for VGA images. Meanwhile, our detector shows superior performance over most of state-of-the-art ones [1] [2] [3] on three challenging benchmarks, including FDDB, AFW, and PASCAL faces, shown in Fig. 1-3.

[1] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv*, 2015.

[2] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015.

[3] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, pages 3676–3684, 2015.

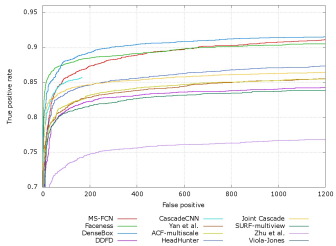


Figure 1: Results on the FDDB dataset.

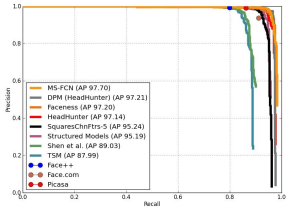


Figure 2: Results on the AFW dataset.

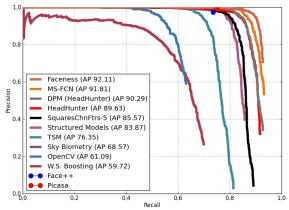


Figure 3: Results on Pascal faces dataset.

Attention Networks for Weakly Supervised Object Localization

Eu Wern Teh
umteht@cs.umanitoba.ca
Mrigank Rochan
mrochan@cs.umanitoba.ca
Yang Wang
ywang@cs.umanitoba.ca

Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada

In this paper, we propose a new approach for localizing objects in weakly labeled data. The novelty of our method is to introduce the concept of "attention" in weakly supervised learning. Our approach starts with generating a set of candidate object regions in each image using standard object proposal techniques. For each object proposal, instead of directly predicting its class label, we first compute an "attention score". This attention score indicates the importance of each object proposal. We then combine the object proposals in the image using their respective attention scores to form a whole image feature vector. This feature vector is then used to classify this image. Since the feature vector for whole image classification is obtained from candidate regions using their attention scores, this will focus the model to learn to assign high attention scores to regions that contain the object of interest. The overview of our approach is illustrated in Fig. 1.

Object proposals: Given a collection of weakly labeled images, the first step of our approach is to generate a shortlist of object proposals in each image. We use the edge boxes method, which is a commonly used technique for generating object proposals. Each proposal is a bounding box that may contain any object.

Let $\mathbf{x}$ be the input image and K be the number of object proposals generated on the image $\mathbf{x}$. We represent each proposal as a fixed length feature vector $\mathbf{x}_i$ ($i = 1, 2, ..., K$).

Proposal attention: For each object proposal $\mathbf{x}_i$, we then compute an *attention score* s_i indicating how likely this object proposal contains the object of interest. This is achieved by applying a linear mapping on $\mathbf{x}_i$ followed by a softmax operation. Let $\mathbf{w}_a$ denote a vector of parameters for the linear mapping, the attention score s_i is

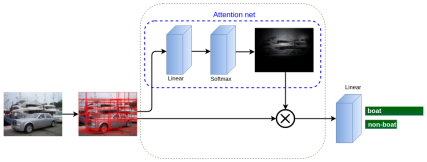


Figure 1: An overview of our architecture.

calculated as:

$$g_i = \mathbf{w}_a^\top \mathbf{x}_i \tag{1a}$$

$$s_i = \frac{\exp(g_i)}{\sum_{j=1}^{K} \exp(g_j)}, \quad i = 1, 2, ..., K \tag{1b}$$

Without loss of generality and to simplify the notation, we use a linear mapping without the bias term in Eq. 1 by assuming that the feature vector $\mathbf{x}$ already has 1 appended to the end.

Image-level classification: Since our data are labeled only at the image-level, we need to use a learning method where the loss function is based on image-level labels. In our work, we use the attention scores to combine the object proposals to get an image-level feature vector $\mathbf{z}$ as $\mathbf{z} = \sum_{i=1}^{K} s_i \mathbf{x}_i$. This image-level feature $\mathbf{z}$ is then used to classify the whole image by a linear classifier with parameters $\mathbf{w}_c$:

$$f(\mathbf{x}; \{\mathbf{w}_a, \mathbf{w}_c\}) = \mathbf{w}_c^\top \mathbf{z} \tag{2}$$

where $f(\mathbf{x}; \{\mathbf{w}_a, \mathbf{w}_c\})$ is the score of classifying $\mathbf{z}$ to be a positive class.

Once the learning is done, we localize the object in weakly labeled data directly using the attention scores. For example, suppose the object of interest is "dog". For each positive "dog" image, we simply choose the object proposal that has the highest attention score s_i as the localized dog instance in this image.

Tuesday
13:40-14:40

Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset

Andrew Shin
andrew@mi.t.u-tokyo.ac.jp
Yoshitaka Ushiku
ushiku@mi.t.u-tokyo.ac.jp
Tatsuya Harada
harada@mi.t.u-tokyo.ac.jp

Graduate School of
Information Science and Technology,
The University of Tokyo
Tokyo, Japan

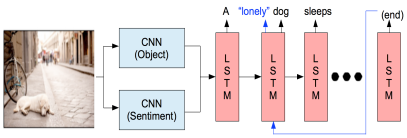


Figure 1: Overall workflow of our model

Image captioning task has become a highly competitive research area with application of convolutional and recurrent neural networks, especially with the advent of long short-term memory (LSTM) architecture. However, its primary focus has been a factual description of the images, mostly objects and their actions. While such focus has demonstrated competence, describing the images with non-factual elements, namely sentiments of the images expressed via adjectives, has mostly been neglected. We attempt to address this issue by fine-tuning an additional convolutional neural network solely devoted to sentiments, where dataset on sentiment is built from a data-driven, multi-label approach.

Building a dataset on sentiments accompanies a number of challenges. First, because sentiments are subjective by nature, it is difficult to label the images in a reliable way. We handle this problem by treating the images as having multiple labels. We utilize *Binary Relevance* [1] in which m training examples x_i whose associated labels form a set Y are viewed as following:

$$D_j = \{(x_i, \phi(Y_i, y_j)) | 1 \leq i \leq m\}$$

$$\text{where } \phi(Y_i, y_j) = \begin{cases} 1, & \text{if } y_j \in Y_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Then, the set of labels for unseen example is determined by the obtained binary classifiers g_j for q classes:

$$Y = \{y_j | g_j(x) > 0, 1 \leq j \leq q\} \quad (2)$$

Second issue is the financial cost of building such dataset. We remark that the viewers' comments associated with the images frequently reflect the dominant sentiments of the images, and exploit them with natural language processing techniques and SentiWordNet [2] to automatically label the images, building a large, weakly-supervised sentiment dataset at zero cost.

Model	Sentiment Term
ImNet+	Young man
Bigram	Poor man
Style	Surgical shirt
Ours	Confused man

Figure 2: Example of generated modifying terms from each model.

We fine-tune a separate convolutional neural network on our sentiment dataset. Roughly inspired by the mechanism in which two hemispheres of human brain perform separate functions of logical and emotional perception, we juxtapose two separate convolutional neural networks, for object and sentiment classification, respectively. We train the obtained representation using two networks with captions, and compare the results with various baseline models. Since automatic evaluation metrics are not designed to handle sentiment terms, we mainly resort to human evaluation as our primary metric. Although ground truth captions contain only a limited amount of sentiment terms, the results demonstrate that our features were able to learn better mapping between the images and sentiment terms than baseline models.

[1] M. Boutell, J. Luo, X. Shen, and C. Brown. *Learning Multi-label Scene Classification*. *Pattern Recognition*, Vol. 37, pp.1757-1771, 2004.

[2] S. Baccianella, A. Esuli, F. Sebastiani. *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. In *Language Resources and Evaluation*, 2010.

Learning of Separable Filters by Stacked Fisher Convolutional Autoencoders

Arash Shahriari
arash.shahriari@anu.edu.au;csiro.au

Australian National University (ANU)
Commonwealth Scientific & Industrial Research Organisation (CSIRO)
Canberra, Australia

Inspired by the overcomplete autoencoders, we introduce Fisher convolutional autoencoders to learn separable filters in a distributed network. These stacked autoencoders employ the linear discriminant analysis to impose the maximum distinction among texture classes whilst holds the minimum separation within each individual one. A network of stacked Fisher autoencoders learns banks of separable filters in parallel and makes an ensemble of deep features with higher separability for better classification. This adjusts the depth of stacks automatically with respect to the capability of each separable filter to extract high order convolutional features for the textures of dataset under study.

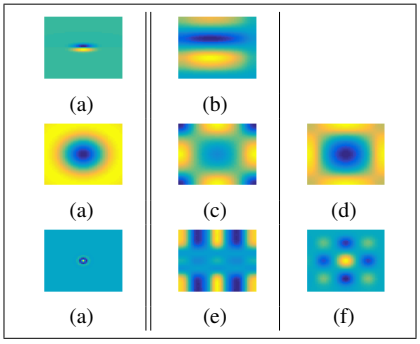


Figure 1: (a) initial filters; (b) UIUC; (c) KTH-TIPS2-a; (d) KTH-TIPS2-b; (e) FMD; (f) DTD.

The figure shows examples of initials and their corresponding learned separable filters from different texture banks. It can be seen that for some filters the changes across various datasets is smaller than the others because they are responsible to extract common features in the texture patterns. In contrast, the filters with high deformations usually connect to deeper stacks of autoencoders capturing high-order convolutional representations.

Dataset	DSIFT [1]	**Proposed**
UIUC	$\mathbf{97.2 \pm 0.8}$	90.1 ± 0.8
KTH-a	82.5 ± 5.3	$\mathbf{85.5 \pm 4.9}$
KTH-b	69.3 ± 0.9	$\mathbf{70.1 \pm 0.7}$
FMD	58.1 ± 1.7	$\mathbf{71.8 \pm 2.2}$
DTD	58.6 ± 2.0	$\mathbf{59.1 \pm 1.3}$

Table 1: Mean accuracy of texture recognition for dense SIFT and our descriptors.

We conduct our experiments on publicly available datasets varying in number of classes and quality of textures on a standard platform. The results prove supremacy of our descriptors over popular dense SIFT features for the purpose of texture understanding.

Dataset	SOA [1][2]	**Ours**
UIUC	$\mathbf{99.3 \pm 0.4}$	96.3 ± 0.1
KTH-a	84.7 ± 1.5	$\mathbf{86.0 \pm 2.3}$
KTH-b	81.1 ± 2.4	$\mathbf{82.3 \pm 0.9}$
FMD	82.4 ± 1.5	$\mathbf{85.7 \pm 3.0}$
DTD	74.7 ± 1.0	$\mathbf{85.9 \pm 1.4}$

Table 2: Mean accuracy of texture recognition for others (SOA) and our framework (Ours).

Our experiments also confirm that the Fisher convolutional autoencoders are successful on imposing distinction among highly-correlated texture patterns when joined the pre-trained deep local descriptors like DeCAF and VGG.

[1] 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.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision*, pages 1–30, 2015.

Deskewing by space-variant deblurring

Tuesday
13:40-14:40

Karthik Seemakurthy
karthikvjit@gmail.com
Subeesh Vasu
ee13d050@ee.iitm.ac.in
A.N. Rajagopalan
raju@ee.iitm.ac.in

Image Processing and Computer Vision Lab
Department of Electrical Engineering
IIT Madras
Chennai, India

Skew and motion blur are significant challenges when camera and scene of interest are in two different media. Skew occurs due to spatially varying refraction on a dynamic water surface, whereas motion blur results from multiple intensities impinging on the imaging sensor during camera exposure time due to time varying refraction. In this paper, we propose a technique to restore underwater images degraded by attenuated water waves. Following others [4] [2], we also assume that the static camera is looking vertically downwards and imaging a planar scene but through spatially decaying periodic water waves (such as waves due to breeze in shallow water bodies). Also, we allow for the attenuation factor as well as the direction of the water waves to undergo changes during the exposure time of the camera (a situation not handled by the state-of-the-art work in [3]). We propose a shot detection method that automatically segments a captured video into groups of frames wherein each group is governed by a dominant water wave direction (i.e. unidirectional) and a single exponential factor of attenuation. Within each segment, we average the frames and show that the blur induced at different pixel locations are scaled versions of each other and model these scale factors through a virtual depth map. Note that despite the scene being planar, the blur induced due to attenuated water waves is space-variant in nature. We pose deskewing of the planar scene as equivalent to a space-variant deblurring problem corresponding to a 3D scene with depth profile being the same as that of the virtual depth map. We propose an alternating framework to solve for the latent image from a single blurred observation. The procedure can likewise be repeated for other segments too, if needed.

We compare our latent image estimation results with the state-of-the-art space-variant deblurring approach [5] along with state-of-the-art deskewing methods [4], [2]. The input to [4] [2] should be in the form of video, while a single blurred image is used for [5] and our proposed algorithm. For synthetic experiments, we use the model described in [1] to simulate a decaying envelope of the water waves which is given as $h(\mathbf{x},t) = A(\mathbf{x})\sin(\omega_x \mathbf{x} + \omega_y \mathbf{y} - t)$, where A($\mathbf{x}$) denotes the decaying amplitude of the sinusoid, ω_x and ω_y are the spatial frequencies. In all our experiments, the decaying function is given by $A(\mathbf{x}) = A_0 \exp(d_f.(\mathbf{d_d}^T \mathbf{x}))$, where d_f is the decay factor and $\mathbf{d_d}$ is the decay direction. We used laminated textured sheets kept at the bottom of an aquarium for indoor experiments. The field-of-view was kept large enough to witness the effect of wave attenuation in the captured video. The source of waves is a fan that acts as a wind blower. We additionally performed outdoor experiments in swimming pool. Here the source of waves was the breeze as well the flow due to water circulation. For all examples, our method outperformed competitive methods, qualitatively and quantitatively.

[1] An-Kuo Liu and Stephen H Davis. Viscous attenuation of mean drift in water waves. *Journal of Fluid Mechanics*, 81(01):63–84, 1977.

[2] Omar Oreifej, Guang Shu, Teresa Pace, and Mubarak Shah. A two-stage reconstruction approach for seeing through water. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1153–1160. IEEE, 2011.

[3] Karthik Seemakurthy and Ambasamudram Narayanan Rajagopalan. Deskewing of underwater images. *Image Processing, IEEE Transactions on*, 24(3):1046–1059, 2015.

[4] Yuandong Tian and Srinivasa G Narasimhan. Seeing through water: Image restoration using model-based tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2303–2310. IEEE, 2009.

[5] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1107–1114, 2013.

Faces in Places: Compound Query Retrieval

Yujie Zhong[1]
yujie@robots.ox.ac.uk
Relja Arandjelović[2]
relja.arandjelovi@inria.fr
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Visual Geometry Group
Dept. of Engineering Science
University of Oxford, UK

[2] WILLOW project
INRIA, France

The goal of this work is to retrieve images containing both a target person and a target scene type (*e.g* Barack Obama on the beach) from a large dataset of images. At run time this compound query is handled using a face classifier trained for the person, and an image classifier trained for the scene type.

We make three contributions: **first** we propose a hybrid convolutional neural network architecture that produces place-descriptors that are aware of faces and their corresponding descriptors. We show that our jointly trained Place-CNN is able to ignore large faces in the images; and it produces descriptors that are amenable to the combination rule we choose - the relevance of a database image to the compound query is computed as the minimum of the query face and query place classification scores.

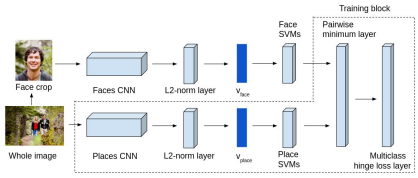


Figure 1: **Hybrid network architecture.** The network is used to generate face-descriptors and face-aware place-descriptors (v_{face} and v_{place}) for the face crop and the whole image respectively.

Second, we propose an image synthesis system to render high quality fully-labelled face-and-place images (as shown in Figure 2), and train the network only from these **synthetic** images. This synthesis system alleviates the difficulty of collecting training data and the problem of severe class-imbalance.

Some examples are shown in Figure 3. The automatic face search and replacement system successfully deals with different lighting condition and head poses, as well as accessories like hats and glasses.

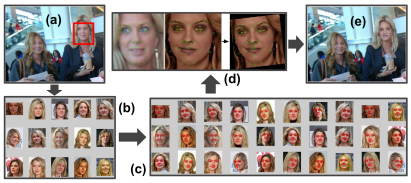


Figure 2: **Automatic synthetic rendering pipeline.** Replacing the face of the unknown person in the airport with the celebrity face (Gage Golightly).

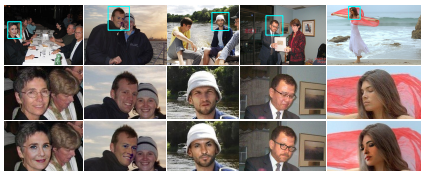


Figure 3: **Example images from the synthetic dataset.** The first row shows the synthetic images, and the second and third row show the close-up of the original and replaced face inside the blue box respectively.

Third, To test our method, and to facilitate research in compound query retrieval, we collect and annotate a 'Celebrity In Places' (CIP) Dataset of **real** images containing celebrities in different places. We test the compound query retrieval on the CIP dataset plus distractor images. As shown in the full paper, retrieval performance for compound queries is significantly improved using the face-aware place-descriptors. Figure 4 shows the top 2 retrieved images of various queries in a dataset of 73k images using our method.

Figure 4: **Examples of top ranking images returned by our retrieval system.**

Tuesday
13:40-14:40

Semi-Supervised Video Object Segmentation Using Multiple Random Walkers

Won-Dong Jang
http://mcl.korea.ac.kr/~dotol1216
Chang-Su Kim
http://mcl.korea.ac.kr

School of Electrical Engineering
Korea University
Seoul, Korea

In this work, we propose a semi-supervised video object segmentation algorithm, which require user annotations about a desired object at the first frame. Figure 1 is an overview of the proposed algorithm. For each frame, we estimate initial distributions and simulate multiple random walkers (MRW) [1]. We perform these two-step processes from the second to the last frames sequentially to yield a segment track.

First, we estimate initial distributions of the foreground and the background based on the segmentation results of previous frames. To this end, we minimize an energy function, which consists of three terms: color Markov energy, motion Markov energy, and guidance energy. Second, we simulate MRW using the two initial distributions. The movements of the foreground agent $\mathbf{p}_\mathrm{f}$ and the background agent $\mathbf{p}_\mathrm{b}$ are modeled by

$$\mathbf{p}_\mathrm{f}^{(\theta+1)} = (1-\varepsilon)\mathbf{A}_\mathrm{c}\mathbf{p}_\mathrm{f}^{(\theta)} + \varepsilon\mathbf{r}_\mathrm{f}^{(\theta)}, \quad (1)$$

$$\mathbf{p}_\mathrm{b}^{(\theta+1)} = (1-\varepsilon)\mathbf{A}_\mathrm{c}\mathbf{p}_\mathrm{b}^{(\theta)} + \varepsilon\mathbf{r}_\mathrm{b}^{(\theta)}, \quad (2)$$

where $\mathbf{r}_\mathrm{f}$ and $\mathbf{r}_\mathrm{b}$ are the restart distributions. $\mathbf{A}_\mathrm{c}$ is a transition matrix. With probability $1-\varepsilon$, the agents move on the graph according to the transition matrix $\mathbf{A}_\mathrm{c}$. On the other hand, with probability ε, the foreground and background agents are forced to restart with the distributions $\mathbf{r}_\mathrm{f}$ and $\mathbf{r}_\mathrm{b}$, respectively. We use a restart rule, which is a hybrid of inference and interactive restart rules. The inference restart rule is time-invariant and inferred from the previous segmentation labels. The time-variant interactive restart rule encourages repulsive interactions between the agents.

Experimental results demonstrate that the proposed algorithm outperforms the state-of-the-art conventional algorithms on the SegTrack v2 dataset [2]. To summarize, this paper has three main contributions:

- Development of an effective restart rule for MRW that yields spatially precise and temporally consistent segment tracks.
- Fixation of parameters, which ensures segmentation qualities on general videos.
- Remarkable performance achievement on the SegTrack v2 dataset.

[1] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim. Multiple random walkers and their application to image cosegmentation. In *CVPR*, pages 3837–3845, 2015.

[2] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013.

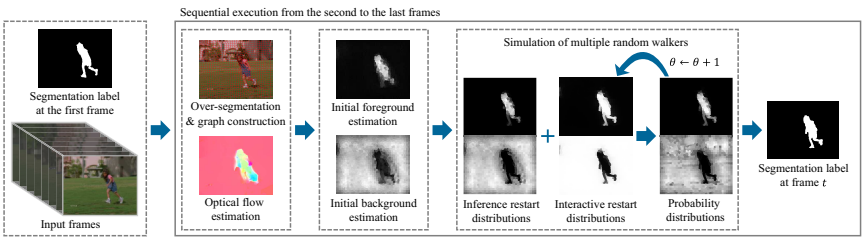


Figure 1: An overview of the proposed algorithm. Using segmentation labels at previous frames, we initialize foreground and background distributions for each frame. Then, we simulate MRW using the inference restart rule and the interactive restart rule. For the segmentation, we compare the foreground and background probabilities at each superpixel.

Deep Learning for Detecting Multiple Space-Time Action Tubes in Videos

Suman Saha[1]
suman.saha-2014@brookes.ac.uk
Gurkirt Singh[1]
gurkirt.singh-2015@brookes.ac.uk
Michael Sapienza[2]
michael.sapienza@eng.ox.ac.uk
Philip H. S. Torr[2]
philip.torr@eng.ox.ac.uk
Fabio Cuzzolin[1]
fabio.cuzzolin@brookes.ac.uk

[1] Dept. of Computing and Communication Technologies
Oxford Brookes University
Oxford, UK

[2] Department of Engineering Science
University of Oxford
Oxford, UK

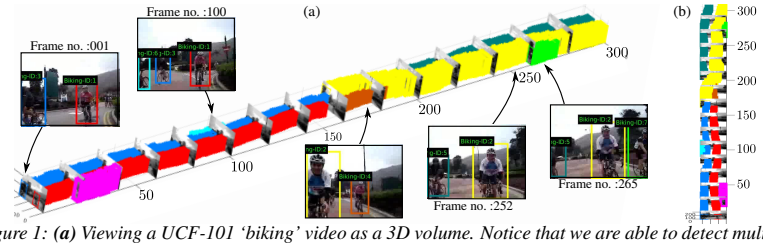


Figure 1: ***(a)*** *Viewing a UCF-101 'biking' video as a 3D volume. Notice that we are able to detect multiple action instances in both space and time.* ***(b)*** *Top-down view.*

In this work, we propose an approach to the spatiotemporal localisation (detection) and classification of multiple concurrent actions within temporally untrimmed videos. Our framework is composed of three stages. In stage 1, appearance and motion detection networks are employed to localise actions from colour images and optical flow. In stage 2, the appearance network detections are boosted by combining them with the motion detection scores, in proportion to their respective spatial overlap. In stage 3, sequences of detection boxes most likely to be associated with a single action instance, called action tubes, are constructed by solving two energy maximisation problems via dynamic programming. While in the first pass, action paths spanning the whole video are built by linking detection boxes over time using their class-specific scores and their spatial overlap, in the second pass, temporal trimming is performed by ensuring label consistency for all constituting detection boxes.

We demonstrate the performance of our algorithm on the challenging UCF101, J-HMDB-21 and LIRIS-HARL datasets, achieving new state-of-the-art results across the board and significantly increasing detection speed at test time. We achieve a huge leap forward in action detection performance when compared to the top competitor [2], and report a **20%** and **11%** gain in mAP on UCF-101 and J-HMDB-21 datasets respectively. The proposed $appearance + motion$ fusion strategy improves the mAPs by 9.4%, 3.6% and 2.5% on the UCF-101, J-HMDB-21 and LIRIS HARL datasets respectively. Further, our 2-pass energy maximisation algorithm contributes to a great extent to significantly boost the performance. Finally, we demonstrate that our action detection pipeline is relatively faster in training and test time detection speeds than the state-of-the-art [1, 2]. Sample qualitative results are provided in the supplementary video [1], and on the project web page [2], where the code and the pretrained models have also been made available.

[1] G Gkioxari and J Malik. Finding action tubes. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015.

[2] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, June 2015.

[1] https://www.youtube.com/embed/vBZsTgjhWaQ
[2] http://sahasuman.bitbucket.org/bmvc2016

Tuesday
13:40-14:40

Exploiting Random RGB and Sparse Features for Camera Pose Estimation

Lili Meng, Jianhui Chen, Frederick Tung, James J. Little, Clarence W. de Silva
University of British Columbia
Vancouver, Canada
lilimeng@mech.ubc.ca

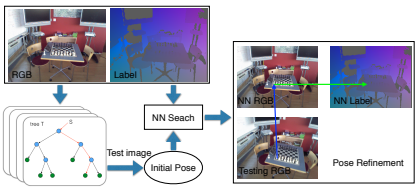


Figure 1: Pipeline. Our method first trains a regression forest using random RGB features. At test time, the camera pose is initially estimated by the random forest predictions, and then is refined by sparse feature matching.

1 Overview

We extend recent advances in scene coordinate regression forests [2] for camera relocalization in RGB-D images to use RGB features, enabling camera relocalization with only a single RGB image at test time. Furthermore, we integrate the random RGB features and sparse feature matching in an efficient and accurate way, broadening the method for fast sports camera calibration in highly dynamic scenes.

2 Methodology

Fig.1 shows the pipeline of our method. During training, the scene information is encoded in a random forest. At test time, an initial camera pose is calculated using the random forest predictions with real-time response. A nearest neighbor (NN) image is queried using the initial camera pose. The camera pose is refined by sparse feature matching between the test image and the NN image. In our method, the labels can be any information associated with pixel locations.

The random RGB features We use features based on pairwise pixel comparison:

$$f_\phi(\mathbf{p}) = \mathtt{I}(\mathbf{p}, c_1) - \mathtt{I}(\mathbf{p}+\delta, c_2) \qquad (1)$$

where δ is a 2D offset and $\mathtt{I}(\mathbf{p}, c)$ indicates an RGB pixel lookup in channel c. Our feature does not require depth information, and so is suitable for large scale sports camera calibration.

Methods	SCRF[2]	PoseNet[1]	Ours
Train	RGB-D	RGB	RGB-D
Test	RGB-D	RGB	RGB
Avg. Err	0.08m,1.60°	0.44m,10.4°	0.17m,5.26°

Table 1: 7 *Scenes* results.

Figure 2: Sports camera calibration examples, best viewed in color. The court lines are overlaid on the images to indicate the accuracy of calibration.

Pose Refinement The 2D-3D correspondences are found by SIFT feature matching between the test image and the NN image. Then, the camera pose $\mathtt{P}$ is optimized by minimizing the reprojection error:

$$\mathtt{P}^* = \arg\min_{\mathtt{P}} \sum_k d(\mathbf{x}_k, \mathtt{P}\mathbf{X}_k)^2 \qquad (2)$$

where $\mathbf{x}$ are the feature locations in the image, and $\mathbf{X}$ are the correspondent 3D world coordinates associated with the NN image.

3 Evaluation

Our method is evaluated on the 7 *Scenes* dataset and a new basketball dataset using standard metrics. Table 1 shows quantitative results in 7 *Scenes* dataset. Fig.2 illustrates qualitative results in the basketball dataset. Experiment results demonstrate the efficacy of our approach, showing superior or on-par performance with the state of the art.

[1] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *ICCV*, 2015.

[2] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013.

Fine-grained Recognition in the Noisy Wild: Sensitivity Analysis of Convolutional Neural Networks Approaches

Erik Rodner[1]
erik.rodner@uni-jena.de
Marcel Simon[1]
marcel.simon@uni-jena.de
Robert B. Fisher[2]
rbf@inf.ed.ac.uk
Joachim Denzler[1]
joachim.denzler@uni-jena.de

[1] Computer Vision Group
Friedrich Schiller University Jena
Germany
www.inf-cv.uni-jena.de

[2] University of Edinburgh
United Kingdom

Overview In this paper, we study the sensitivity of CNN outputs with respect to image transformations and noise in the area of fine-grained recognition. We answer the following questions: (1) how sensitive are CNNs with respect to image transformations encountered during wild image capture?; (2) can we increase the robustness of CNNs with respect to image degradations? and (3) how can we predict CNN sensitivity?

To answer the first question, we provide an extensive empirical sensitivity analysis of common CNN architectures (AlexNet, VGG19, and GoogleNet) across various types of image degradations. We perturb test images of different datasets with noise types including Gaussian and pepper noise, random color shifts, and different geometric image transformations. This allows for predicting CNN performance for new domains comprised by images of lower quality or captured from a different viewpoint. Our experiments show that even small random noise can lead to a dramatic performance decrease.

The question naturally arises if it is possible to increase the robustness either during testing or by adapting the learning. we analyze two intuitive ideas for increasing robustness: data augmentation by applying input dropout to the training data and image pre-processing.

After the empirical analysis, the question remains whether we can quickly detect images with unstable CNN outputs. This question goes beyond a pure sensitivity study but asks for uncertainty estimates often available for Bayesian methods but not for CNNs. We present a novel approach (Figure 1) for estimating the sensitivity given an input using a first-order approximation of the output change.

Take-home message The experiments show that the influence especially of common intensity noise is severe even at low noise levels. The reason is a domain shift between noise-free training and perturbed test data. From our study, we can draw several conclusions:

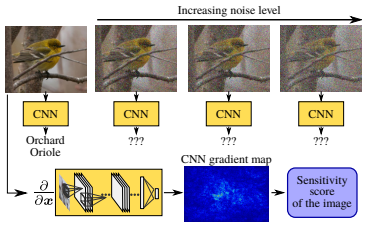


Figure 1: How sensitive are CNNs with respect to image noise and transformations? We study this question and show how to predict CNN sensitivity for a given image.

1. The training images should have the same noise level as the test images and care has to be taken even for small noise applied to intensities.
2. Data augmentation during training or image pre-processing are no solutions as they decrease the accuracy on noise-free images dramatically and are only beneficial for high noise levels at best.
3. Noise sensitivity depends on the CNN architecture and VGG19 has shown to be the most robust one.
4. Sensitivity of CNN outputs can be predicted for small noise levels with our technique allowing for uncertainty estimates of CNN outputs.

These conclusions can be seen as guidelines especially for developers of real-world applications, where, for example, cheap camera sensors deliver low quality images but the training was done on relatively noise-free datasets like ImageNet.

Near-Field Photometric Stereo in Ambient Light

Tuesday
13:40-14:40

Fotios Logothetis[1]
fl302@cam.ac.uk
Roberto Mecca[1,2]
roberto.mecca@eng.cam.ac.uk
Yvain Quéau[3]
yvain.queau@enseeiht.fr
Roberto Cipolla[1]
cipolla@eng.cam.ac.uk

[1] Department of Engineering
University of Cambridge
United Kingdom

[2] Department of Mathematics
University of Bologna
Italy

[3] Université de Toulouse
France

Shape recovery from shading information has recently regained importance as latest Photometric Stereo-based techniques have been improved in terms of appearance of reflective objects. However, 3D scanners based on this technology do not provide reliable reconstructions as long as the considered irradiance equation neglects additive bias such as ambient light. We present a new approach based on ratios of differences of images where perspective viewing geometry, non-linear light propagation, both specular and diffuse reflectance plus the additive bias of the ambient light are tackled simultaneously.

Contribution. Our approach for PS extends the model presented in [1] by considering non-negligible ambient light as an additional pixel-wise component to the usual irradiance model.

Theory. We assume the following irradiance equation for the i^{th} light source:

$$I_i(x,y) = \rho(x,y) a_i(x,y) (\overline{\mathbf{n}}(x,y) \cdot \overline{\mathbf{h}}_i(x,y,z))^{\frac{1}{c(x,y)}} + A(x,y) \tag{1}$$

where ρ is the albedo, c is a material property, a_i is the light attenuation, all in $(0,1]$. The ambient light $A(x,y)$ is a pixel-wise unknown of the problem, independent from $\mathbf{l}_i$. First, we preliminary manipulate the irradiance equation as follows:

$$(\rho a_i)^c \overline{\mathbf{n}} \cdot \overline{\mathbf{h}}_i = (I_i - A)^c \approx {I_i}^c - c I_i^{c-1} A \tag{2}$$

where the approximation comes from truncating the Binomial expansion to the first two terms. This simplifies to (using $\gamma_i = \frac{a_i^c}{I_i^{c-1}}$):

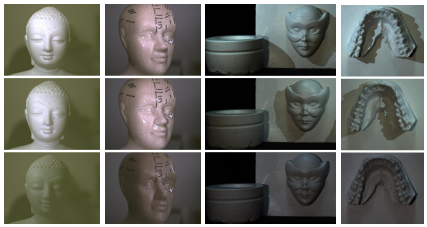
$$I_i - cA \approx \rho^c \gamma_i \frac{\mathbf{n}}{|\mathbf{n}|} \cdot \overline{\mathbf{h}}_i. \tag{3}$$

Then, we consider two pairs of irradiance equations, namely the i^{th}, j^{th} and the q^{th}, r^{th}; in order to let the ambient light cancel out together with the albedo, we concider the following ratio:

$$\frac{I_i - cA - I_j + cA}{I_q - cA - I_r + cA} \approx \frac{\frac{\rho^c}{|\mathbf{n}|}\left[\gamma_i \mathbf{n} \cdot \overline{\mathbf{h}}_i - \gamma_j \mathbf{n} \cdot \overline{\mathbf{h}}_j\right]}{\frac{\rho^c}{|\mathbf{n}|}\left[\gamma_q \mathbf{n} \cdot \overline{\mathbf{h}}_q - \gamma_r \mathbf{n} \cdot \overline{\mathbf{h}}_r\right]}. \tag{4}$$

By using the parameterization of the normal in terms of the depth from [2], a variational problem for the unknown depth is solved.

Experiments. We evaluated the algorithm on various real data sets: a marble Buddha statue (1(a)), a shiny plastic head(1(b)), a plaster mask(1(c)) and a plaster print of teeth(1(d)).



(a) Buddha (b) Head (c) Mask (d) Teeth

Figure 1: Two samples from each object (rows 1-2), the respective ambient light (row 3) and the corresponding reconstructions.

[1] R. Mecca and Y. Quéau. Unifying diffuse and specular reflections for the photometric stereo problem. In *WACV*, 2016.

[2] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *CVPR*, 2013.

Acknowledgments R. Mecca is a Marie Curie fellow of the "INdAM", Italy.

Solving Visual Madlibs with Multiple Cues

Tatiana Tommasi[1]
ttommasi@cs.unc.edu
Arun Mallya[2]
amallya2@illinois.edu
Bryan Plummer[2]
bplumme2@illinois.edu
Svetlana Lazebnik[2]
slazebni@illinois.edu
Alexander C. Berg[1]
aberg@cs.unc.edu
Tamara L. Berg[1]
tlberg@cs.unc.edu

[1] University of North Carolina at Chapel Hill, (NC) USA
[2] University of Illinois at Urbana-Champaign, (IL) USA

This paper focuses on answering multiple choice questions from the Visual Madlibs dataset [2] which was created by asking people to write fill-in-the-blank descriptions about persons (action, attribute, location), objects (affordance, attribute, location), and high-level concepts as future and past events.

We posit that in order to truly understand an image and answer questions about it, it is necessary to leverage rich and detailed global and local information. To explore this assertion, we represent the images by using CNN architectures trained on task-specific sources to recognize more than 200 scenes, 900 actions and 300 attributes (see Fig. 1). We extract the features both from the whole image and from regions selected to best match people and objects mentioned in the answers. We project both the visual and textual information in a joint CCA-embedding space [1] and at test time, we select the putative answer which obtains the highest cosine similarity with the image features. Finally we integrate multiple cues, through low-level visual feature stacking and high-level CCA score combinations. Our results show a significant improvement over the previous state of the art (see Tab. 1), and indicate that answering different question types benefits from examining a variety of image cues and carefully choosing informative image sub-regions.

[1] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.

[2] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. In *ICCV*, 2015.

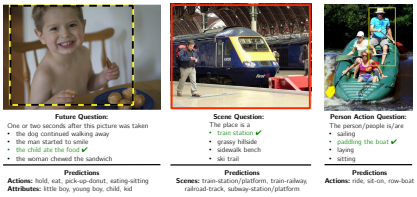


Figure 1: Our method uses multiple deep networks trained on external knowledge sources to predict action, attribute, scene, and other diverse features from specific regions in the image. A CCA model trained on these features allows to score the putative answers and select the correct one for different different types of questions.

	Question Type		Baseline VGG	CCA Ensemble
a)	Interesting	Easy	79.53	83.20
		Hard	55.05	57.70
	Past	Easy	80.24	86.36
		Hard	54.35	60.00
	Future	Easy	80.22	86.88
		Hard	55.49	62.39
b)	Person Attribute	Easy	53.56	68.50
		Hard	42.58	55.90
	Person Action	Easy	84.71	88.34
		Hard	68.04	71.65
	Person Location	Easy	84.95	85.70
		Hard	64.67	63.92
	Person Object Relationship	Easy	73.63	78.93
		Hard	56.19	58.63
c)	Object Attribute	Easy	50.35	58.94
		Hard	45.41	54.50
	Object Affordance	Easy	82.49	87.29
		Hard	64.46	68.37
	Object Location	Easy	67.91	70.03
		Hard	56.71	58.01

Table 1: Improvement in accuracy by combining CCA scores from multiple cues.

LSTM for Image Annotation with Relative Visual Importance

Geng Yan[1]

Yang Wang[2]

Zicheng Liao[1]

[1] College of Computer Science
Zhejiang University

[2] Department of Computer Science
University of Manitoba

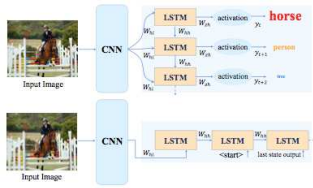


Figure 1: Illustration of the LSTM model. (Top) In our model, the image feature is used as an input to the LSTM at each time step. (Bottom) In the LSTM model used for image captioning, the image feature is only used to start the initial state in the LSTM model.

We consider the problem of image annotations that takes into account of the relative visual importance of tags. Humans have the remarkable ability to selectively process very narrow regions of the scene that are important to us. So when asked to annotate an image, we only mention a subset of the objects appearing in the image, and we mention the important objects first. In this paper, we propose a method for producing such ranked tag list for a given image. Such a ranked tag list can be useful for various applications including image retrieval, image parsing and image caption generation.

Our proposed approach combines the convolutional neural network (CNN) for images and the LSTM for sequential data. Fig. 1 illustrates our model and compare it with the RNN model for image captioning.

Image representation: Following prior work (e.g. [1]), we represent an image as a 4096-dimensional CNN feature vector using pre-trained VGGNet. We then use a fully connected layer to reduce the dimension to d. In other words, given an input image I_m, we represent it as a d-dimensional feature vector as:

$$I = W_I \cdot CNN(I_m) + b_I \tag{1}$$

where $W_I \in \mathbb{R}^{d\times 4096}$ and $b_I \in \mathbb{R}^d$ are the parameters to be learned. $CNN(I_m)$ is the 4096-dimensional CNN feature extracted on the image I.

LSTM for tag list prediction: We modify the standard LSTM, so that the hidden state at each time step considers the image feature $v(I)$ as one of the inputs. In other words, our LSTM model is defined as follows:

$$i_t = \sigma(W^{(i)}I + U^{(i)}\mathbf{h}_{t-1}) \tag{2}$$
$$f_t = \sigma(W^{(f)}I + U^{(f)}\mathbf{h}_{t-1}) \tag{3}$$
$$o_t = \sigma(W^{(o)}I + U^{(o)}\mathbf{h}_{t-1}) \tag{4}$$
$$\tilde{\mathbf{c}}_t = \tanh(W^{(c)}I + U^{(c)}\mathbf{h}_{t-1}) \tag{5}$$
$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \circ \tilde{\mathbf{c}}_t \tag{6}$$
$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{c}_t) \tag{7}$$

At each time step t, we need to predict a tag from a vocabulary of size V. We use another linear layer to project the hidden state h_t into a vector of dimension V, followed by a softmax operator. This will give us the probability of choosing each of the V possible tags as the predicted tag at time t:

$$\mathbf{z}_t = W^{(z)}\mathbf{h}_{(t)} + \mathbf{b}^{(z)} \tag{8}$$
$$p_{t,v} = \frac{\exp(z_{t,v})}{\sum_{k=1}^{V}\exp(z_{t,k})} \tag{9}$$

where $\mathbf{z}_t \in \mathbb{R}^V$, and $p_{t,v}$ denotes the probability of picking the v-th tag in the vocabulary as the predicted tag at time t.

We demonstrate the effectiveness on the PASCAL2007 dataset and the LabelMe dataset.

[1] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

OnionNet: Sharing Features in Cascaded Deep Classifiers

Martin Simonovsky
martin.simonovsky@enpc.fr
Nikos Komodakis
nikos.komodakis@enpc.fr

Imagine Lab
Université Paris Est / École des Ponts
Paris, France

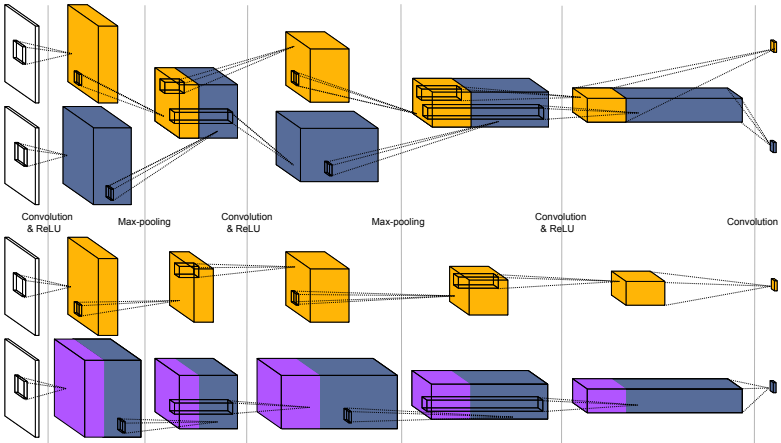


Figure 1: *Top* - OnionNet: the first stage (S1, orange) shares its intermediate feature maps (visualized as cubes) with the second stage (S2, blue). *Bottom* - A traditional cascade: stages are independent and S2 has to be evaluated fully, recomputing certain features (purple).

The focus of our work is speeding up evaluation of deep neural networks in retrieval scenarios. A popular approach to reduce time spent on negative examples is to set up a cascade of classifiers of increasing strength, called stages. As these are trained for the same or a similar objective, the question is how much their features (should) have in common. Without any sharing, a representation presumably at least as powerful as in the previous stage has to be rebuilt in the following one.

We address this by proposing OnionNet, a novel feature-sharing cascaded architecture where the next stage extends the feature map set of the previous stage, preventing repeated computation. Crucially, the architecture is flexible: the next stage may add both new layers as well as new feature channels. We construct our cascades by gradually increasing the width, possibly in addition to depth. This is beneficial as the lowest layers tend to be the most expensive ones to compute while producing weak classifiers on their own.

Figure 1 illustrates a two-stage OnionNet cascade consisting of two branches with the same layer organization. Each takes the same input and is terminated by its own output layer. The core idea is that the branches are linked before every convolutional layer. The feature maps of the first stage (S1) are used as additional input to the following convolutional layer in the second stage (S2) but not the other way round, creating a one-way dependence. The model is trained end-to-end under a joint loss, which makes the cascade learn the proper allocation of features between the stages.

OnionNet is applied to three important tasks: patch matching, proposal-based object detection, and image retrieval. We demonstrate good speed-ups due to cascades and show that OnionNet sharing can bring further gain atop of it, with only a marginal decrease in precision. Specifically, we achieve 2.8x, 2.9x, and 1.7x running time reduction in each respective application. Furthermore, we provide a systematic study in theory and on a synthetic benchmark that sheds further light into the time cost behavior of cascaded architectures.

Wednesday
13:40-14:40

Line reconstruction using prior knowledge in single non-central view

Jesus Bermudez-Cameo[1]
bermudez@unizar.es
Cédric Demonceaux[2]
cedric.demonceaux@u-bourgogne.fr
Gonzalo Lopez-Nicolas[1]
gonlopez@unizar.es
José J. Guerrero[1]
josechu.guerrero@unizar.es

[1] Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza
Zaragoza, Spain

[2] Le2i - Institut Universitaire de Technologie
Le Creusot
Université de Bourgogne
Le Creusot, France

Wednesday
13:40-14:40

Line projections in non-central systems contain more geometric information than central systems. The four degrees of freedom of the 3D line are mapped to the line-image and the 3D line can be theoretically recovered from 4 projecting rays (i.e. line-image points) from a single non-central view [3]. If the non-central system is properly calibrated we obtain a metric reconstruction of the 3D line. In practice, extraction of line-images is considerably more difficult and the resulting reconstruction is imprecise and sensitive to noise.

In this paper we explore the reconstruction accuracy improvements when we impose geometrical constraints [1] exploiting prior knowledge. In particular, when the lines of the scene are arranged in two orthogonal directions and we know prior information about the direction of one of this directions (typically the vertical direction), the complexity of line fitting reduces, the accuracy of the metric reconstruction improves, and the extraction procedure is simplified.

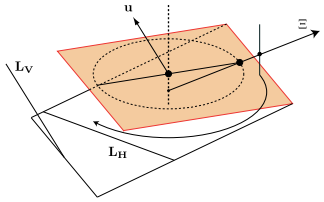
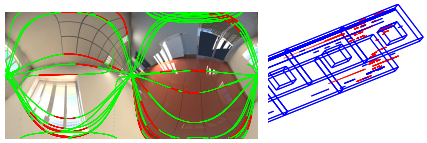


Figure 1: Consider a Manhattan setting with horizontal lines $\mathbf{L}_H$ and vertical lines $\mathbf{L}_V$. A non-central circular panoramic system is considered in unknown position and orientation but the vertical direction $\mathbf{u}$ of lines $\mathbf{L}_V$ in its own reference system is known.

The first restriction considers a 3D line parallel to a plane, and takes advantage of the prior knowledge of the vertical direction to fit horizontal lines. In this case only three rays are needed to fit the 3D line. The second restriction considers a 3D line with known direction, and again takes advantage of the prior knowledge of the vertical direction for fitting vertical lines. Both formulations are integrated in a line-extraction pipeline, which is tested with synthetic and real non-central circular panoramas [2].



(a) (b)

Figure 2: (a) Extraction and (b) reconstruction example from synthetic single non-central panorama.

In addition, we evaluate the performance of the robust extractor and the accuracy of the proposal in comparison with the unconstrained method. We conclude that the proposal outperforms the unconstrained algorithm and provides good results taking into account typical accuracy of standard commercial IMUs (error around 0.5 deg).

[1] Jesus Bermudez-Cameo, Joao P. Barreto, Gonzalo Lopez-Nicolas, and Jose. J Guerrero. Minimal solution for computing pairs of lines in non-central cameras. In *ACCV*, 2014.

[2] Marc Menem and Tomás Pajdla. Constraints on perspective images and circular panoramas. In *BMVC*, pages 1–10, 2004.

[3] Seth Teller and Michael Hohmeyer. Determining the lines through four lines. *Journal of graphics tools*, 4(3):11–22, 1999.

Attribute Recognition from Adaptive Parts

Luwei Yang[1]
luweiy@sfu.ca
Ligeng Zhu[2]
zhuligeng@zju.edu.cn
Yichen Wei[3]
yichenw@microsoft.com
Shuang Liang[4]
shuangliang@tongji.edu.cn
Ping Tan[1]
pingtan@sfu.ca

[1] Simon Fraser University
Vancouver, Canada
[2] Zhejiang University
Hangzhou, China
[3] Microsoft Research Asia
Beijing, China
[4] Tongji University
Shanghai, China

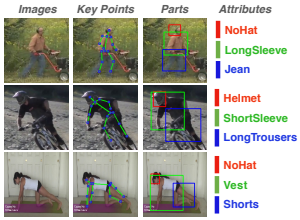


Figure 1: Given an image, we estimate the key points, generate object parts accordingly, and predict attributes of each part. The learning is end-to-end in a single deep neural network.

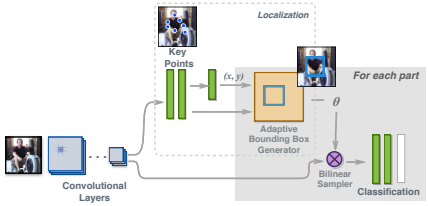


Figure 2: Overview of network architecture. It consists of initial convolutional feature exaction layers, a key point localization, an adaptive bounding box generator for each part, and the final attribute classification network for each part.

We focus on the problem of object attribute recognition. Previous part-based attribute recognition approaches perform part detection and attribute recognition in separate steps. The parts are not optimized for attribute recognition and therefore could be sub-optimal. In this paper, we present an end-to-end deep learning approach to overcome the limitation.

In our network architecture, instead of training part detector and attribute classifier separately, both key point estimation and attribute recognition are learned jointly in a multi-task setting. Figure 1 shows the example of our pipeline. We firstly estimate object key points as an auxiliary task. Because the definition of key point is clear, their annotation is less ambiguous than part bounding boxes. From the key points, the object parts are generated adaptively, with free parameters to be learned for adjusting its spatial extent. This adaptive part generation is inspired by the recent spatial transformer network [1], which can learn image spatial transformation from the image classification goal. Instead of applying the transform to the whole image [1], we apply a spatial transform for each part and use the bilinear sampler [1] to warp the image features for subsequent attribute recognition. The whole network is learned end-to-end in a multi-task setting, with attribute classification as the main task and key point prediction as the auxiliary one.

The network framework is illustrated in Figure 2. It consists of a convolutional network for feature extraction, a localization network for key point estimation, an adaptive bounding box generator for each part, and part based feature sampler and attribute classifier.

Key Point Estimation As shown in Figure 2, three fully-connected layers are appended on the top of the convolutional features as the regression network, with the output dimensions $2N$ (N is the number of key points). We use L2 distance loss for key point estimation, $\sum_i^N \|\hat{p}_i - p_i\|_2^2$, where $\hat{p}_i, p_i \in \mathbb{R}^2$ are the normalized ground truth and estimation for key point i.

Adaptive Part Generation Some attributes are clearly associated with certain object parts. We specify a subset of key points $\mathcal{P}_t$ for each part t. The initial part bounding box $b_t = [w_t, h_t, x_t, y_t]$ encodes the origin and size of a rectangle area that covers all key points in $\mathcal{P}_t$, and can be obtained by finding the maximum and minimum of the subset. The final bounding box is defined as $[w_t(1+\Delta_w), h_t(1+\Delta_h), x_t+\Delta_x, y_t+\Delta_y]$, with additional adjustment parameters $\Delta = [\Delta_w, \Delta_h, \Delta_x, \Delta_y]$ to be learned adaptively.

Bilinear Sampling and Attribute Recognition For each part bounding box, the convolutional feature maps are warped accordingly. We use the Bilinear Sampler in [1] that warps the local feature via bilinear interpolation, and it serves as a bridge that allows the gradient of attribute classification flow into the localization network. The warping employs a 2×3 affine transformation, parameterized as

$$\theta_t = \begin{bmatrix} w_t(1+\Delta_w) & 0 & x_t+\Delta_x \\ 0 & h_t(1+\Delta_h) & y_t+\Delta_y \end{bmatrix}. \quad (1)$$

The transformation θ_t warps the local coordinates, the corresponding content can be sampled by bilinear interpolation subsequently. For attribute parsing, the Softmax multi-class classifier is adopted.

Our approach is validated on human attribute recognition on two datasets, via extensive experiment comparison. The comparable results show the effectiveness of jointly training of localization and classification task.

[1] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Memory-based Gait Recognition

Dan Liu
liudan11060126@gmail.com
Mao Ye
cvlab.uestc@gmail.com

School of Computer Science and Engineering,
Center for Robotics,
University of Electronic Science and Technology of China,
Chengdu 611731, P.R. China

In this paper, inspired by the mechanism of memory and prediction in our brains [2], we propose a straightforward and effective memory-based gait recognition method (MGR) to realize the memory and recognition process of the gait sequences. Because of various covariates including carrying, clothing, surface and view angle, we extract the robust 2D joint location information via the joint extraction model as the gait features. Compared to the traditional neural network, the memory neuron network (MNN), for example, the Long Short-term Memory (LSTM) architecture, simulates the human brain and stores the objects in the weights of neural connections. Besides, by the large-scale parallel computing, MNN can repair the incomplete and tainted data (the extracted 2D gait feature is dirty). It is the first time that we utilize the MNN to address the gait recognition issue. This maybe empower a fresh orientation for solving gait recognition problem. Fig.1 shows the overall framework of the method.

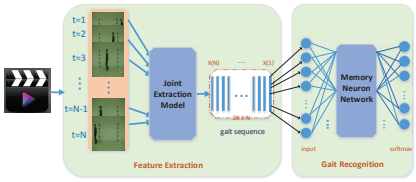


Figure 1: The Memory-based gait recognition framework. Obviously, the process is divided into two stages: feature extraction and gait recognition. *N* denotes the length of a gait sequence.

We compare our method against others on the CASIA A and CASIA B gait datasets. Tab.1, Tab.2 show some experimental results. Getting enlightenment from [1], the longer sequences do not improve the algorithm performance in some cases. Therefore, we reduce the length of sequences to about 45 in average. In Tab.2, Exp1, Exp2 and Exp3 indicate different conditions, respectively. The details can refer to our full paper. Though the presented network configuration is simple, the proposed method still obtains the relatively satisfactory and comparable results.

Methods	0°-view	45°-view	90°-view	avg
Wang1 [4]	65.00	63.75	77.50	68.75
Wang2 [4]	65.00	66.25	85.00	72.08
Wang3 [4]	75.00	81.25	93.75	83.33
Orig-results	82.50	83.75	92.50	86.25
Length-red	**85.00**	**87.50**	**95.00**	**89.17**

Table 1: The comparisons of some algorithms on the CASIA A (0°,45°,90°) dataset. Wang1, Wang2 and Wang3 indicate that the different classifiers and similarity measures are used in the same method.

Methods	Exp1	Exp2	Exp3	Avg
Martin [3]	70.16	74.19	58.60	67.65
Orig-results	83.06	85.48	80.11	82.88
Length-red	**83.87**	**85.48**	**81.72**	**83.69**

Table 2: Algorithms comparisons on the CASIA B dataset on Exp1, Exp2 and Exp3.

[1] Armand Joulin and Tomas Mikolov. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems 28*.

[2] Christof Koch and Joel L. Davis. *Large-scale Neuronal Theories of The Brain*. MIT press, 1994.

[3] Raul Martin-Felez and Tao Xiang. *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I*, chapter Gait Recognition by Ranking. 2012.

[4] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on PAMI*, 25(12):1505–1518, Dec 2003.

Three-Point Direct Stereo Visual Odometry

Jeong-Kyun Lee
leejk@gist.ac.kr
Kuk-Jin Yoon
kjyoon@gist.ac.kr

Computer Vision Laboratory
Gwangju Institute of Science and
Technology (GIST), South Korea

Existing visual odometry methods are generally classified into two groups; feature-based and direct methods. In the feature-based methods, the measurement errors are defined by the re-projection errors of feature points, and the ego-motion is estimated by minimizing the re-projection errors that are assumed to usually conform to the Laplace distribution [1]. However, since the local appearance of the feature point consecutively varies with time, the location of the tracked feature point tends to drift from its initial location during tracking and this results in error accumulation. In contrast, direct methods minimize the measurement errors measured by intensity differences between consecutive images. Since a large number of pixels are utilized for motion estimation, these methods are superior to feature-based methods in the error accumulation aspect. However, in return, the direct methods have a difficulty in handling outliers and are vulnerable to illumination change.

These methods generally use a maximum consensus set of inlier feature points on the Laplace distribution assumption of measurement errors. However, as shown in Fig. 1, the assumption is often violated in the real-world scene because of lots of outliers and/or biases of measurement errors. In this situation, although a motion candidate may be accurately obtained from a small number of point samples in the RANSAC framework, its final estimate optimized from the set of inliers can be rather erroneous. The work of Chum *et al*. [2] tried to solve the similar problem using an iterative local RANSAC scheme but it is still hard to hold thoroughly uncontaminated inliers only.

In this paper, to solve the problem shown in Fig. 1, we propose a method that estimates ego-motion using only uncontaminated feature points. Since it is difficult to distinguish the uncontaminated features among lots of features, we randomly sample a minimum number of features (*i.e*., 3 points) and exploit them for the estimation. Thereby, the proposed method maximally excludes the errors caused by inaccurate inliers. However, using fewer features in feature-based methods can rather degrade the performance because measurement errors caused in feature matching and tracking can be propagated to the ego-motion estimation directly. Therefore, we propose a hybrid method of feature-based and direct methods. It minimizes measurement errors by the ego-motion estimation directly using image intensity values without any intermediate step to search measurements and is more robust than conventional direct methods. Experimental results show the proposed method produces better performance than the existing methods using all inliers.

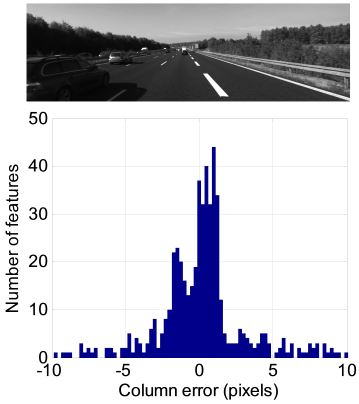


Figure 1: The distributions of re-projection errors for estimated motion in selected frames of the KITTI dataset. The distribution violates the Laplace distribution assumption of the measurement errors because of outliers from moving objects.

[1] Hernan Badino, Akihiro Yamamoto, and Takeo Kanade. Visual odometry by multi-frame feature integration. In *ICCV*, 2013.

[2] Ondrej Chum, Jiri Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition*, pages 236–243. Springer Berlin Heidelberg, 2003.

Semantic Segmentation for Real-World Data by Jointly Exploiting Supervised and Transferrable Knowledge

Li-Hsien Lu
s103062588@m103.nthu.edu.tw

Chiou-Ting Hsu
cthsu@cs.nthu.edu.tw

Dept. of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

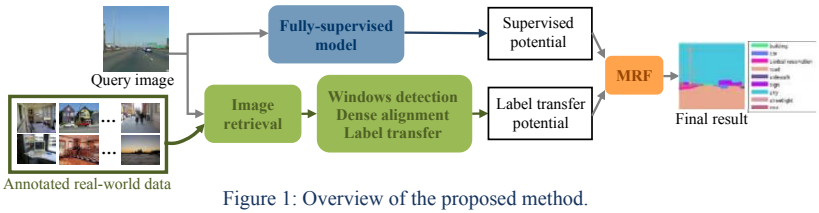


Figure 1: Overview of the proposed method.

This paper addresses two major challenges in semantic segmentation for real-world data. First, with ever-increasing semantic labels, we need a more pragmatic approach other than existing fully-supervised methods. Second, semantic segmentation for rarely-appeared objects are still challenging for existing methods.

We assume that there exist one supervised model and one annotated real-world dataset. Our goal is to leverage the well-learned knowledge from the existing model and infer new labels via label transfer from the real-world dataset. We propose a "content-adaptive" and "label-aware" MRF framework to jointly exploiting both the supervised and transferrable knowledge. The proposed method needs no off-line training and can easily adapt to real-world data.

Assume that the existing supervised model is trained with the label set C_{fs}, and the real-world dataset is annotated with the label set C_r, where $C_{fs} \neq C_r$. Note that the supervised model is unaware of unknown labels $c \in C_r \backslash C_{fs}$. We formulate the MRF energy function as:

$$E(\boldsymbol{c}) = -\sum_{\boldsymbol{p}\in I}[(1-\alpha(I))\cdot\psi_{fs}(c,\boldsymbol{p}) + \alpha(I)\beta(c,I)\cdot\psi_{trans}(c,\boldsymbol{p})] + \lambda\sum_{(\boldsymbol{p},\boldsymbol{q})\in\varepsilon}\theta(c(\boldsymbol{p}),c(\boldsymbol{q})), \quad (1)$$

where $\psi_{fs}(c,\boldsymbol{p})$ is the supervised potential derived by FCN [1]; $\psi_{trans}(c,\boldsymbol{p})$ is the label transfer potential obtained by the modified nonparametric method [2]; and $\theta(\cdot,\cdot)$ is the pairwise potential term and λ is a smoothing constant. $\alpha(I)$ is adaptive to different query image I to dynamically combine the two potentials. $\beta(c,I)$ is a label-aware parameter for balancing the priority of rare labels in the query.

In Table 1, we compare our method with existing methods on SIFT Flow and LMSun dataset, and the results demonstrate the effectiveness of the proposed method.

Table 1: Comparison with existing methods.

Method	Per-pixel (%)	Per-class (%)
SIFT Flow dataset $\lvert C_r\rvert = 33$		
C. H. Ma et al. [2]	78.3	46.1
M. George [3]	81.7	50.1
Ours ($\lvert C_{fs} \cap C_r\rvert = 7$)	81.7	50.0
J. Long et al. [1]	**85.6**	50.1
Ours ($C_{fs} = C_r$)	85.2	**52.0**
LMSun dataset $\lvert C_r\rvert = 232$		
M. George [3]	61.2	16.0
J. Yang et al. [4]	60.6	**18.0**
Ours ($\lvert C_{fs} \cap C_r\rvert = 55$)	**65.4**	16.5

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[2] C. H. Ma, C. T. Hsu and B. Huet. Nonparametric scene parsing with deep convolutional features and dense alignment. In *ICIP*, 2015.

[3] M. George. Image parsing with a wide range of classes and scene-level context. In *CVPR*, 2015.

[4] J. Yang, B. Price, S. Cohen and M. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.

Wednesday
13:40-14:40

Optimized Regressor Forest for Image Super-Resolution

Chia-Yang Chang
cychang@media.ee.ntu.edu.tw
Wei-Chih Tu
wctu@media.ee.ntu.edu.tw
Shao-Yi Chien
sychien@ntu.edu.tw

Media IC and System Lab
Graduate Institute of Electronics Engineering
National Taiwan University
Taipei, Taiwan

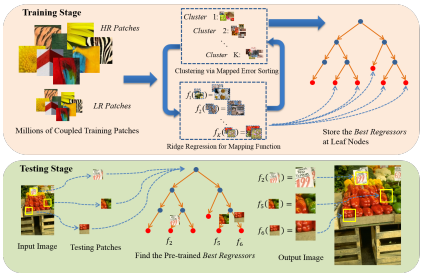
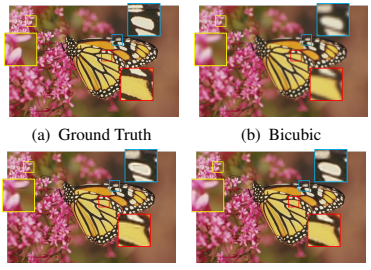


Figure 1: Overview of our framework.



(a) Ground Truth (b) Bicubic

(c) A+ (d) Our ORF-PSNR

Figure 2: Butterfly image from Set14 dataset with upscaling 3x.

The goal of image super-resolution is to recover missing high frequency details of an image given single or multiple low-resolution images. It is a well-known ill-posed problem and requires mature prior knowledges or enough examples to restore high-quality high-resolution images. Recently, many methods [1, 2] formulate image super-resolution as a regression problem. Input image patches are classified into pre-trained clusters, and cluster-dependent mapping functions are employed to super-resolve input patches. The classification following the regression training scheme has a potential problem that the regression step minimizes the error within a cluster. While there might be outliers in the classification process, the learned regression function may not be the optimal for all patches. The key towards high quality and efficient image super-resolution turns to answering the following questions:
(i) *How to classify image patches such that patches in the same cluster can be accurately super-resolved using the same regressor?*
(ii) *How to learn the best regression function such that the super-resolved HR patches have highest numerical accuracy?*

In this paper, we tackle the problem in a reverse manner. We put the regressor at the first place. We propose to learn a set of regression functions from training samples using the EM-algorithm. These regressors are learned to minimize overall reconstruction distortion (*e.g.* peak signal to noise ratio(PSNR) or structure similarity(SSIM)) for all training data. After that, we will obtain a set of patch-regressor pairs. We then train a random forest from these patch-regressor pairs to predict the best regressors for input patches. We call it the *Optimized Regressor Forest* for super-resolution. An overview of the proposed framework is illustrated in Fig. 1. Our experimental results show that the proposed method is able to achieve comparable or better results in numerical evaluation or visual comparison(see in Fig. 2).

[1] J. Salvador and E. Pérez-Pellitero. Naive Bayes Super-Resolution Forest. In *IEEE International Conference on Computer Vision*, pages 325–333, 2015.

[2] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. 2014.

Convolutional aggregation of local evidence for large pose face alignment

Adrian Bulat
adrian.bulat@nottingham.ac.uk
Georgios Tzimiropoulos
yorgos.tzimiropoulos@nottingham.ac.uk

Computer Vision Laboratory
University of Nottingham
Nottingham, UK

Wednesday
13:40-14:40

Methods for unconstrained face alignment must satisfy two requirements: (a) they must not rely on accurate initialization/face detection and (b) they should perform equally well for the whole spectrum of facial pose. To the best of our knowledge, there are no methods meeting these requirements to satisfactory extent, and in this paper, we propose Convolutional Aggregation of Local Evidence (CALE), a Convolutional Neural Network (CNN) architecture particularly designed for addressing both of them.

In particular, CALE by-passes the requirement for accurate face detection by firstly using a CNN detector to perform facial landmark detection, providing at the same time confidence scores for the location of each of the facial landmarks (local evidence). Next, our system aggregates the local evidence for each facial landmark through joint CNN regression of the confidence scores, in order to refine the landmarks' location. Besides playing the role of a graphical model, CNN regression is a key feature of our system, guiding the network to rely on context for predicting the location of occluded landmarks, typically encountered in very large poses. The proposed architecture (Fig. 1) is simple and can be trained end-to-end with intermediate supervision. We show that our system achieves large performance improvement on AFLW-PIFA[1], which is, to the best of our knowledge, by far the most difficult test set for face alignment to date.

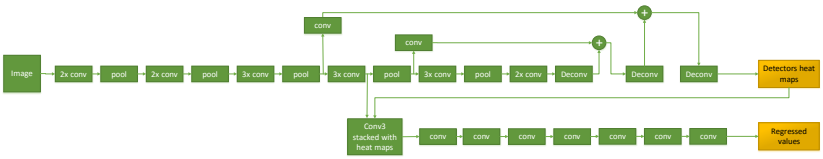


Figure 1: Proposed architecture for Convolutional Aggregation of Local Evidence (CALE).

Our second contribution in this paper is an investigation of CALE's alignment performance beyond human faces and, in particular, on animal faces. As animal faces exhibit a much larger degree of variability in shape and appearance as well as in pose and expression, animal face alignment is a much more difficult problem which, to the best of our knowledge, has never been systematically explored in the past by the Computer Vision community. Although drawing a direct comparison is not possible, our results, show that CALE's performance on animal faces is not far from that on human faces.

When applied to AFLW-PIFA[1], our method provides more than 50% absolute gain in localization accuracy when compared to other recently published methods [2, 3] for large pose face alignment. Note that prior work reports on visible points, only. To the best of our knowledge we are the first to report results on non-visible landmarks too. Remarkably, the performance of CALE when evaluated on all points - both visible and occluded surpasses the performance of all existing methods when these are evaluated on visible points only.

[1] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *ICCV*, 2015.

[2] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, 2016.

[3] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016.

Wide Residual Networks

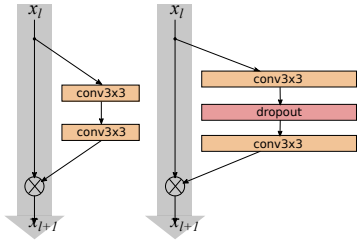
Sergey Zagoruyko
sergey.zagoruyko@enpc.fr
Nikos Komodakis
nikos.komodakis@enpc.fr

Université Paris-Est, École des Ponts
ParisTech
Paris, France

Deep residual networks were shown to be able to scale up to thousands of layers and still have improving performance. However, each fraction of a percent of improved accuracy costs nearly doubling the number of layers, and so training very deep residual networks has a problem of diminishing feature reuse, which makes these networks very slow to train. To tackle these problems we conduct a detailed experimental study on the architecture of ResNet blocks, based on which we propose a novel architecture where we decrease depth and increase width of residual networks. In addition, we propose a new way of utilizing dropout within deep residual networks so as to properly regularize them and prevent overfitting during training. We call the resulting network structures wide residual networks (WRNs) and show that these are far superior over their commonly used thin and very deep counterparts. Our experiments show that:

- our widened architecture consistently improves performance across residual networks of different depth;
- increasing both depth and width helps until the number of parameters becomes too high and stronger regularization is needed;
- there doesn't seem to be a regularization effect from very high depth in residual networks as wide networks with the same number of parameters as thin ones can learn same or better representations. Furthermore, wide networks can successfully learn with a lot more parameters than thin ones, which would require doubling the depth of thin networks, making them infeasibly expensive to train.

Overall, we demonstrate that even a simple 16-layer-deep wide residual network outperforms in accuracy and efficiency all previous deep residual networks, including thousand-layer-deep networks, achieving new state-of-the-art results on CIFAR-10, CIFAR-100 and SVHN (table 1). Our code is available at https://github.com/szagoruyko/wide-residual-networks.



(a) basic (b) wide-dropout

Figure 1: Basic and wide-dropout residual blocks. Batch normalization and ReLU precede each convolution

	depth-k	CIFAR-10	CIFAR-100
NIN		8.81	35.67
FitNet		8.39	35.04
Highway [4]		7.72	32.39
ResNet[1]	110	6.43	25.16
	1202	7.93	27.82
stoc-depth[3]	110	5.23	24.58
	1202	4.91	-
pre-ResNet[2]	110	6.37	-
	164	5.46	24.33
	1001	4.64	22.71
WRN (ours)	40-4	4.97	22.89
	16-8	4.81	22.07
	28-10	**4.17**	**20.50**

Table 1: Test error on CIFAR-10 and CIFAR-100 with moderate data augmentation (flip/translation). k is a widening factor. We don't use dropout for these results.

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016.

[3] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. *CoRR*, abs/1603.09382, 2016.

[4] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.

Deep Part-Based Generative Shape Model with Latent Variables

Alexander Kirillov[1]
alexander.kirillov@tu-dresden
Mikhail Gavrikov[2]
gavrmike@gmail.com
Ekaterina Lobacheva[4]
elobacheva@hse.ru
Anton Osokin[3]
anton.osokin@inria.fr
Dmitry Vetrov[4]
vetrovd@yandex.ru

[1] TU Dresden, Dresden, Germany
[2] Rubbles, Moscow, Russia
[3] INRIA – École Normale Supérieure, Paris, France
[4] National Research University Higher School of Economics (HSE), Moscow, Russia

Wednesday
13:40-14:40

Models of shape play a substantial role in a number of computer vision tasks The Shape Boltzmann Machine (SBM) [2] and its multilabel version MSBM [1] have been recently introduced as deep generative models that capture the variations of an object shape. While being more flexible MSBM requires datasets with labeled parts of the objects for training (Fig. 1c). In the paper we present an algorithm for training MSBMs using binary masks of objects (Fig. 1b) and the seeds which approximately correspond to the locations of objects parts (Fig. 1d). The latter can be obtained from part-based detectors in an unsupervised manner. We derive a latent variable model and an EM-like training procedure for adjusting the weights of MSBM using a deep learning framework. We show that the model trained by our method outperforms SBM in the tasks related to binary shapes and is very close to the original MSBM in terms of quality of multilabel shapes.

SBM and MSBM SBM is a Deep Boltzmann Machine (DBM) with special constraints on its parameters [2] that allow to avoid overfitting for small datasets. It defines a joint distribution $p(\boldsymbol{b}, \boldsymbol{h}^1, \boldsymbol{h}^2|\boldsymbol{\theta})$, where $\boldsymbol{b}$ is a layer that corresponds to the binary shape of an object (fig. 1b), $\boldsymbol{h}^1, \boldsymbol{h}^2$ are two hidden layers and $\boldsymbol{\theta}$ is a vector of all the SBM parameters. MSBM is a generalization of SBM model to multilabel case and it defines the similar distribution $p(\boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\boldsymbol{\theta})$, where $\boldsymbol{m}$ is a layer which corresponds to the multilabel shape of an object (fig. 1c). MSBM is more expressive since the variations of an object's parts are usually smaller than the variation of the whole object.

Our model In our paper a step towards the unsupervised training of a shape model is made. We propose a way to train a multilabel model without using the full multilabel annotation (as in [1]). Instead we use easier-to-obtain binary masks and seeds $\boldsymbol{s}$ of the parts (fig. 1d). For each part there is one seed pixel that approximatelly corresponds to the center of this part.

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

Figure 1: (a) – image of an object, (b) – binary segmentation $\boldsymbol{b}$, (c) – the multilabel segmentation $\boldsymbol{m}$, (d) seeds $\boldsymbol{s}$ for the selected 4 parts: head, front legs, rear legs and croup.

We model the joint distribution $p(\boldsymbol{b}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\boldsymbol{\theta})$ of binary mask $\boldsymbol{b}$, seeds $\boldsymbol{s}$, multilabel masks $\boldsymbol{m}$ and hidden variables $\boldsymbol{h}^1$, $\boldsymbol{h}^2$ using the assumption that binary segmentation $\boldsymbol{b}$ and seeds $\boldsymbol{s}$ are conditionally independent given multilabel segmentation, i.e, $p(\boldsymbol{b}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\boldsymbol{\theta}) = p(\boldsymbol{b}|\boldsymbol{m})p(\boldsymbol{s}|\boldsymbol{m})p(\boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\boldsymbol{\theta})$. To train the unknown parameters $\boldsymbol{\theta}$ of the model we use the variational EM algorithm, i.e. maximize $\sum_{d=1}^{D} \log p(\boldsymbol{b}^d, \boldsymbol{s}^d|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$.

Our experiments show that as the model of binary shape the MSBM trained by our technique performs similar to the MSBM that is trained using full annotation and significantly outperforms the SBM and MSBM that is trained using automatically obtained multilabel annotation from a binary mask and seeds.

[1] S. M. Ali Eslami and Chris Williams. A generative model for parts-based object segmentation. In *NIPS*. 2012.

[2] S. M. Ali Eslami, Nicolas Heess, Chris Williams, and John Winn. The shape boltzmann machine: a strong model of object shape. In *IJCV*, 2013.

General Human Traits Oriented Generic Elastic Model for 3D Face Reconstruction

Joi San Tan[1]
tjs11_com120@student.usm.my
Ibrahim Venkat[1]
ibra@usm.my
Iman Yi Liao[2]
iman.liao@nottingham.edu.my
Philippe De Wilde[3]
p.dewilde@kent.ac.uk

[1] Universiti Sains Malaysia, Pulau Pinang, Malaysia.
[2] The University of Nottingham, Malaysia Campus, Semenyih, Selangor, Malaysia.
[3] The University of Kent, Canterbury Kent CT2 7NZ, United Kingdom.

We propose a *Simplified Generic Elastic Model (S-GEM)* which intends to construct a 3D face from a given 2D face image by making use of a set of general human traits viz., *Gender, Ethnicity and Age (GEA)*. We hypothesise that the variations inherent on the depth information for individuals are significantly mitigated by narrowing down the target information via a selection of specific GEA traits. In this paper, we propose a 3D reconstruction method to retain the robustness of the PCA-based models and in the meantime to provide control over the depth values of 2D facial feature points. We formulate the reconstruction of the 3D face model of a given 2D face image as a posterior estimation of the PC coefficients Φ given the observations of the 2D facial feature points x_f. The depth value Z of the 2D feature points is expressed as the hidden information. The posterior probability is represented as the marginal distribution of $P(\Phi|\boldsymbol{x}_f)$ integrated over Z as shown below:

$$P(\Phi|\boldsymbol{x}_f,\Delta) \propto \sum_{Z_f} P(\Delta|\boldsymbol{x}_f,Z_f)\cdot P(\boldsymbol{x}_f,Z_f|\Phi)\cdot P(\Phi) \tag{1}$$

where $\boldsymbol{x}_f$ represents the x and y coordinates of the 2D input feature points, Δ represents the corresponding GEA group and Z_f represents the hidden Z values of the 2D input feature points. Also, $\boldsymbol{x}_f$ and Δ are the observed variables in this representation. Based on the Bayesian theory, $P(\Delta|\boldsymbol{x}_f,Z_f)$ can be written as,

$$P(\Delta|\boldsymbol{x}_f,Z_f) = \frac{P(\boldsymbol{x}_f,Z_f|\Delta)P(\Delta)}{\sum P(\boldsymbol{x}_f,Z_f|\Delta)P(\Delta)} \tag{2}$$

$P(\boldsymbol{x}_f,Z_f|\Delta)$ is defined using simplified mixture model. The proposed S-GEM method was compared with the PCA-TR method [1] and a method of utilising the Z-coordinates of the model mean-face feature points (we name it M-GEM). M-GEM is chosen for comparisons as it is based on the same principle of the popular GEM model[2], i.e., using the model mean face depth value to represent any individual face depth. The results in Figure 1 shows that the proposed S-GEM method has produced the least MSE out of the three methods in terms of the full 3D face shapes. Figure 2 shows the outputs of the newly reconstructed 3D face models using proposed S-GEM.

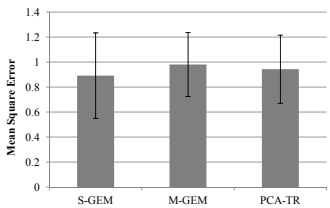


Figure 1: Empirical test of the proposed method (S-GEM), M-GEM and PCA-TR.

Figure 2: Outputs of the newly reconstructed 3D face models using proposed S-GEM.

[1] AY Maghari, Ibrahim Venkat, Iman Yi Liao, and Bahari Belaton. PCA-Based Reconstruction of 3D Face Shapes using Tikhonov Regularization. International Journal of Advances in Soft Computing and its Applications , 5(2):1-15, 2013.
[10] Jingu Heo. 3D Generic Elastic Models for 2D Pose Synthesis and Face Recognition, 2009.

Attend Refine Repeat: Active Box Proposal Generation via In-Out Localization

Spyros Gidaris
spyros.gidaris@enpc.fr
Nikos Komodakis
nikos.komodakis@enpc.fr

Université Paris-Est, École des Ponts ParisTech
Paris, France

Wednesday 13:40-14:40

Introduction: In our work we deal with the problem of *category agnostic box proposal generation*. Its definition is that for a given image a small set of boxes must be generated that will cover with high recall all the objects in the image regardless of their category. Recently, this problem has received an immense amount of attention due to the fact that box proposal generators have became a core component in many vision tasks, ranging from object detection till visual question answering, leading in all of them to state-of-the-art results.

Approach: The dominant paradigm in box proposal generation is that of having a CNN model that given a set of input boxes (uniformly distributed in the image), it predicts their objectness score and refines their coordinates such that they better align with object's borders (i.e. bounding box prediction). In that context, our work improves the previous state-of-the-art in box proposal generation in two ways. **(1)** We improve the object's bounding box prediction step by adapting the succesful LocNet [1] approach for category-specific object localization to that of category-agnostic localization (see Figure 1). **(2)** We employ an active box proposal generation strategy, which we call *Attend Refine Repeat* algorithm (see algorithm on the right), that starts from a set of seed boxes, which only depend on the image size, and it then sequentially produces newer boxes that will better cover the objects of the image while avoiding the "objectless" image areas (see Figure 2).

Results: We exhaustively evaluate our system both on PASCAL and on the more challenging COCO datasets and we demonstrate significant improvement with respect to the state-of-the-art on box proposal generation. Furthermore, we provide strong evidence that our object location refinement module is capable of generalizing to unseen categories.

[1] Spyros Gidaris and Nikos Komodakis. Locnet: Improving localization accuracy for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Algorithm: ***Attend Refine Repeat***

Input : Image $\mathbf{I}$
Output: Bounding box proposals $\mathbf{P}$
$\mathbf{C} \leftarrow \emptyset$, $\mathbf{B}^0 \leftarrow$ seed boxes
for $t \leftarrow 1$ *to* T **do**
  `/* Attend & Refine procedure */`
  $\mathbf{O}^t \leftarrow \mathit{ObjectnessScoring}(\mathbf{B}^{t-1}|\mathbf{I})$
  $\mathbf{B}^t \leftarrow \mathit{ObjectLocationRefinement}(\mathbf{B}^{t-1}|\mathbf{I})$
  $\mathbf{C} \leftarrow \mathbf{C} \cup \{\mathbf{B}^t, \mathbf{O}^t\}$
end
$\mathbf{P} \leftarrow \mathit{NonMaxSuppression}(\mathbf{C})$

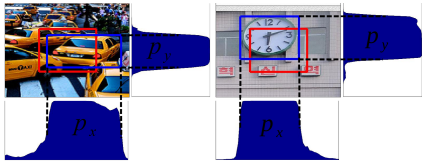


Figure 1: **Object location refinement formulation.** We formulate the problem of box prediction to that of dense classification. Specifically, given an input box (red rectangle), our model defines a search region (depicted image crop) and assings a membership probability to each row and each column of that region (p_y and p_x probability vectors) that represent the likelihood of those elements (rows or columns) to be inside the target bounding box (blue rectangle).

Figure 2: Image areas that are attened by our active proposal generation algorithm as it progress from the first (left column) to the last iteration (right column).

Method	AR@10	AR@100	AR@1000
EdgeBoxes	0.074	0.178	0.338
Geodesic	0.040	0.180	0.359
Selective Search	0.052	0.163	0.357
MCG	0.101	0.246	0.398
DeepMask	0.153	0.326	0.482
Co-Obj	0.189	0.366	0.492
SharpMask	0.192	0.391	0.555
AttractioNet (Ours)	**0.328**	**0.535**	**0.661**

Table 1: **Sampled average recall results on COCO validation set.**

Crafting a multi-task CNN for viewpoint estimation

Francisco Massa
http://imagine.enpc.fr/~suzano-f/
Renaud Marlet
http://imagine.enpc.fr/~marletr/
Mathieu Aubry
http://imagine.enpc.fr/~aubrym/

LIGM, UMR 814, Imagine,
Ecole des Ponts ParisTech, UPEM,ESIEE Paris, CNRS, UPE
Champs-sur-Marne, France

Convolutional Neural Networks (CNNs) were recently shown to provide state-of-the-art results for object category viewpoint estimation. However different ways of formulating this problem have been proposed and the competing approaches have been explored with very different design choices. This paper presents a comparison of these approaches in a unified setting as well as a detailed analysis of the key factors that impact performance. Followingly, we present a new joint training method with the detection task and demonstrate its benefit. We also highlight the superiority of classification approaches over regression approaches, quantify the benefits of deeper architectures and extended training data, and demonstrate that synthetic data is beneficial even when using ImageNet training data. By combining all these elements, we demonstrate a consistent improvement of approximately 5% mAVP over previous state-of-the-art results on the Pascal3D+ dataset [4].

Contributions: In this paper, we study several factors that affect performance for the task of joint object detection and pose estimation with CNNs and introduce a new approach for the joint training. Using the best design options, we rationally define an effective method to integrate detection and viewpoint estimation, quantify its benefits, as well as the boost given by deeper networks and more training data, including data from ImageNet and synthetic data. The relative benefits of each of these elements as well as a comparison with baseline is summarized in table 1. We demonstrate that the combination of all these elements leads to an important improvement over state-of-the-art results on Pascal3D+, going for example from 31.1% to 36.1% AVP in the case of the most challenging 24 viewpoints classification. While several of the elements that we employ have been used in previous work [2, 3], we know of no systematic study of their respective and combined effect, resulting in an absence of clear good practices for viewpoint estimation and sub-optimal performances.

[1] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3D geometry to deformable part models. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proceedings of the IEEE International Conference on Computer Vision*.

[3] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond Pascal: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

Table 1: Summary of results and comparison with baselines using AVP24

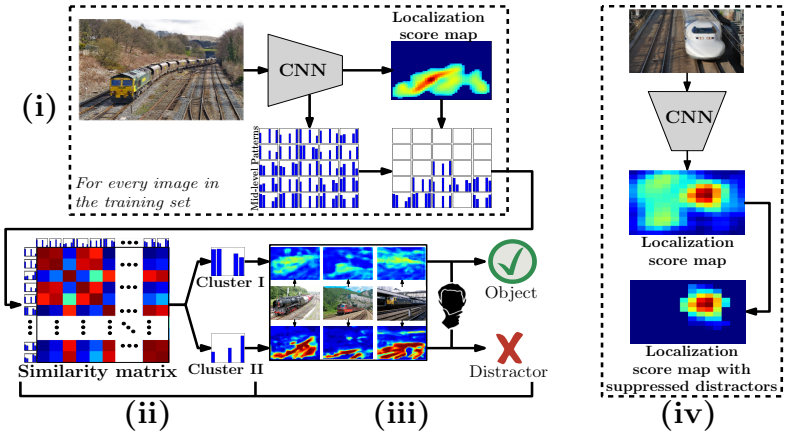
Method	aero	bike	boat	bus	car	chair	table	mbike	sofa	train	tv	mAVP24
DPM-VOC+VP [1]	9.7	16.7	2.2	42.1	24.6	4.2	2.1	10.5	4.1	20.7	12.9	13.6
Render For CNN [2]	21.5	22.0	4.1	38.6	25.5	7.4	11.0	24.4	15.0	28.0	19.8	19.8
Viewpoints & Keypoints [3]	37.0	33.4	10.0	54.1	40.0	**17.5**	19.9	34.3	28.9	43.9	22.7	31.1
Classif. approach & AlexNet	21.6	15.4	5.6	41.2	26.4	7.3	9.3	15.3	13.5	32.9	24.3	19.3
+ our joint training	24.4	16.2	4.7	49.2	25.1	7.7	10.3	17.7	14.8	36.6	25.6	21.1
+ VGG16 instead of AlexNet	26.3	29.0	8.2	56.4	36.3	13.9	14.9	27.7	20.2	41.5	26.2	27.3
+ ImageNet data	42.4	37.0	**18.0**	59.6	43.3	7.6	25.1	**39.3**	29.4	**48.1**	28.4	34.4
+ synthetic data	**43.2**	**39.4**	16.8	**61.0**	**44.2**	13.5	**29.4**	37.5	**33.5**	46.6	**32.5**	**36.1**

Wednesday
13:40-14:40

Improving Weakly-Supervised Object Localization By Micro-Annotation

Alexander Kolesnikov
akolesnikov@ist.ac.at
Christoph H. Lampert
chl@ist.ac.at

IST Austria
Am Campus 1
3400 Klosterneuburg
Austria



Wednesday
13:40-14:40

Object localization is a crucial step needed for building automatic systems for visual scene understanding. This task can be successfully tackled using fully-supervised learning methods, but these require annotations in a form of bounding boxes or per-pixel segmentation masks that are time-consuming and expensive to acquire. Therefore, it is important to develop weakly-supervised object localization learning techniques, which require much cheaper forms of annotation, *e.g.* image-level class labels.

Analyzing the current methods for weakly-supervised object localization we arrive at the conclusion that they tend to fail for object classes that consistently co-occur with the same background elements (distractors), *e.g. trains* on *tracks*. We overcome these failures by developing a new procedure that determines semantic parts that constitute the object detection and then discards distractor parts. The main steps of our approach are (see Figure above) (i) represent all predicted foreground regions of all images by mid-level features learned by a deep neural network, (ii) cluster these features using spectral clustering (the number of clusters is determined automatically), (iii) visualize the clusters and let a human annotator select which ones actually corresponds to the object class of interest. The information about clusters and their annotation can then be used to better localize objects: (iv) for any (new) image, predict a foreground map using only the image regions that match clusters labeled as 'object'.

Note, that the proposed method requires virtually negligible amount of additional supervision: an annotator has to answer a few binary questions (typically 2 or 3) per semantic class. Huge datasets, such as *ILSVRC*, can be annotated by one annotator in just a few hours.

The proposed approach can be readily used in combination with many existing localization methods. In this work we combine it with the current state-of-the art methods for weakly-supervised bounding box prediction [2] and for weakly-supervised semantic segmentation [1], showing improved results on the challenging *ILSVRC* 2014 and *PASCAL VOC* 2012 datasets.

[1] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. *ECCV*, 2016.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

MBestStruck: M-Best Diverse Sampling for Structured Tracker

Ivan Bogun
http://my.fit.edu/~ibogun2010
Eraldo Ribeiro
http://cs.fit.edu/~eribeiro

Department of Computer Sciences and Cybersecurity
School of Computing
Florida Institute of Technology
Melbourne, U.S.A.

We approach the problem of *model-free* visual tracking of objects in videos. Model-free tracking has its state-of-the-art in a class of methods called *tracking-by-detection*, as shown in recent benchmarks. Some top-performing methods use deep neural networks (i.e., convnets) to solve the learning-based steps of the tracking algorithm (e.g., bounding box prediction and evaluation). Despite improving accuracy, *convnets* impose a high computational cost on trackers, limiting their real-time applications. In this paper, we propose to use deep features from a pre-learned deep-convolutional network in a computationally efficient way. Here, we use the *M-Best diverse-sampling* approach for sampling a small yet diverse set of bounding boxes that are likely to contain the object being tracked. These bounding boxes are then used by our method to perform detection using deep features. The resulting tracker, which we call *MBestStruck*, uses high-quality feature representation while being computationally efficient. Our tracking approach compares very well with the state-of-the-art, as we demonstrate by experiments done on popular benchmark datasets.

M-Best-Diverse Labeling. Let $E : \mathcal{Y} \to \mathbb{R}$ be an energy function that we define as a negative ObjStruck [2] discriminative function:

$$E(\mathbf{y}) = -\sum_{i,\bar{\mathbf{y}}} \beta_i^{\bar{\mathbf{y}}} \langle \phi(\mathbf{x_i}, \bar{\mathbf{y}}), \phi(\mathbf{x}, \mathbf{y}) \rangle - \\ - \lambda_s s(\mathbf{y}) - \lambda_e e(\mathbf{y}), \qquad (1)$$

where $\lambda_e, \lambda_s > 0$ are objectness parameters, and $s(\cdot), e(\cdot)$ are the straddling and the edge-density measures of objectness, respectively. Batra et al. [1] uses a greedy sequential procedure for finding M diverse labelings, $\mathbf{y_1}, \ldots, \mathbf{y_M}$, according to the following criterion:

$$\mathbf{y}^m = \arg\min_{y \in \mathcal{Y}} \left[E(\mathbf{y}) - \lambda \sum_{i=1}^{m-1} \Delta(\mathbf{y}, \mathbf{y}^i) \right], \qquad (2)$$

Figure 1: Left row: sampling using MBest procedure. Right row: sampling deterministically using linearly spaced bounding boxes.

for $i = 1, \ldots, M$, where parameter $\lambda > 0$ controls a trade-off between the diversity of the labelings and their quality. The function $\Delta(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}$ is called a *dissimilarity kernel*.

We compared our method with the other trackers on benchmarks OTB50, OTB100, and VOT2015. Results show that our sampling strategy compares favorably to the state-of-the-art while using fewer bounding boxes for detection.

[1] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse M-best solutions in Markov random fields. In *ECCV*, pages 1–16. Springer, 2012.

[2] Ivan Bogun and Eraldo Ribeiro. Object-aware tracking. *ICPR 2016 (to appear)*, 2016.

Wednesday
13:40-14:40

Event-Based Hough Transform in a Spiking Neural Network for Multiple Line Detection and Tracking Using a Dynamic Vision Sensor

Sajjad Seifozzakerini[1]
stusajjad@i2r.a-star.edu.sg
Wei-Yun Yau[1]
wyyau@i2r.a-star.edu.sg
Bo Zhao[1]
zhaob@i2r.a-star.edu.sg
Kezhi Mao[2]
ekzmao@ntu.edu.sg

[1] Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore

[2] School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore

Wednesday
13:40-14:40

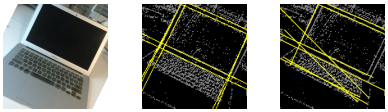
In this paper, we develop an event-based Hough transform and apply it to a new type of camera, namely Dynamic Vision Sensor (DVS). DVSs are a new generation of cameras that are sensitive to logarithmic intensity change [4]. Once the change is larger than a predefined threshold, a positive or negative event will be generated depending on the direction of the change.

The main idea of Hough transform is first transforming every point from the conventional Cartesian coordinates to the parameter space, in which every point defines a specific shape, and then finding local maximums in the parameter space to obtain the shape parameters through a voting procedure [3].

In this paper, we use LIF spiking neurons [1] to build an SNN that represents the parameter space of Hough transform for line detection. every Spiking Neuron (SN) has some inputs and an output. The input is a spike train that influences the neuron's Membrane Potential (MP) which is always decaying by a fixed rate. Whenever the MP exceeds the + or - threshold, a spike with corresponding polarity is generated in the output and MP is reset to zero subsequently.

The parameter space is built up by a two dimensional SNN with one dimension for angle θ and the other for normal distance r. A local lateral inhibition strategy is adopted in our SNN which allows the SNN to suppress noise lines (or redundant lines) from being detected.

In cases that there are more than one moving line in the frame, we need a segmentation procedure to distinguish between them. Since every line is moving smoothly in Cartesian space, the corresponding spikes in parameter space are "moving" smoothly as well and they produce a cluster. We use an event-based clustering method [2] to do the segmentation and tracking of different lines.



(a) (b) (c)

Figure 1: (a) Image captured by a conventional camera; (b) The proposed event-based algorithm's line detection results (yellow) superimposed onto DVS events (grey); (c) Conventional frame-based hough transform's results using MATLAB standard functions for line detection with the same number of the lines.

The efficacy of the proposed algorithm is shown by extensive experiments on both artificially generated events and real DVS output. SNN with local lateral inhibition is efficient in detecting correct lines and tracking them as well as suppressing incorrect ones as seen in figure 1.

[1] A N Burkitt. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biological cybernetics*, 95(1):1–19, jul 2006. ISSN 0340-1200. doi: 10.1007/s00422-006-0068-6.

[2] T Delbruck and P Lichtsteiner. Fast sensory motor control based on event-based hybrid neuromorphic-procedural system. *Circuits and Systems, 2007. ISCAS ...*, (80 cm):845–848, 2007.

[3] Richard O. Duda and Peter E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1): 11–15, jan 1972. ISSN 00010782. doi: 10.1145/361237.361242.

[4] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 X 128 120 dB 15 us latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, feb 2008. ISSN 00189200. doi: 10.1109/JSSC.2007.914337.

Holistically Constrained Local Model: Going Beyond Frontal Poses for Facial Landmark Detection

KangGeon Kim[1]
kanggeon.kim@usc.edu
Tadas Baltrušaitis[2]
tbaltrus@cs.cmu.edu
Amir Zadeh[2]
abagherz@cs.cmu.edu
Louis-Philippe Morency[2]
morency@cs.cmu.edu
Gérard Medioni[1]
medioni@usc.edu

[1] Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA, USA

[2] Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Facial landmark detection is an essential initial step for a number of facial analysis research areas such as expression analysis, 3D face modeling, facial attribute analysis, and person recognition. It is a well researched problem that has seen a surge of interest in the past couple of years.

However, most state-of-the-art methods still struggle in the presence of extreme head pose, especially in challenging in-the-wild images. Furthermore, as most methods operate in a local manner [1, 2], they rely on good and consistent initialization, which is often very difficult to achieve. While some images attempt to combat this by evaluating a number of proposals and initializations, this comes at a computational cost.

In our work, we present a new model – Holistically Constrained Local Model (HCLM), which unifies local and holistic facial landmark detection by integrating head pose estimation, sparse-holistic landmark detection and dense-local landmark detection. Our method's main advantage is the ability to handle very large pose variations, including profile faces. Furthermore, our model integrates local and holistic facial landmark detectors in a joint framework, with a holistic approach narrowing down the search space for the local one.

For a given set of k facial landmark positions $\mathbf{x} = \{x_1, x_2, ..., x_k\}$, our HCLM model defines the likelihood of the facial landmark positions conditioned on a set of sparse landmark positions $X_s = \{x_s,\ s \in S\}$ ($|S| \ll k$) and image $\mathcal{I}$ as follows:

$$p(\mathbf{x}|I, X_s, \mathcal{I}) \propto p(\mathbf{x}) \prod_{i=1}^{k} p(x_i|X_s, \mathcal{I}). \quad (1)$$

In Equation 1, $p(\mathbf{x})$ is prior distribution over set of landmarks $\mathbf{x}$ following a 3D point distribution model (PDM) with orthographic camera projection.

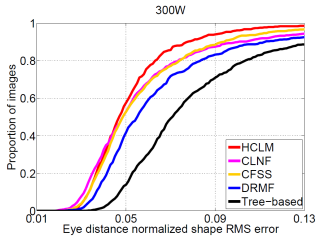


Figure 1: *Cumulative error curves on 300-W dataset.* Measured as the mean Euclidean distance from ground truth normalized by the inter-ocular distance. Note that we use 68 points for this comparison.

Some of the results comparing our HCLM model to state-of-the-art baselines can be seen in Figure 1. Our model demonstrates competitive or better performance to most of the baselines. Furthermore, HCLM demonstrates superior performance in especially difficult images, such as profile ones. This is due to the both better initializations and combination of *holistic* and *local* approaches of our model.

[1] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[2] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

Bag of Surrogate Parts: one inherent feature of deep CNNs

Yanming Guo
y.guo@liacs.leidenuniv.nl
Michael S. Lew
mlew@liacs.nl

LIACS Media Lab
Leiden University
Leiden, the Netherlands

In this paper, we first develop a new feature from the last pooling layer (i.e. pool_5) of VGG, called Bag of Surrogate Parts (BoSP), and its spatial variant, Spatial BoSP (S-BoSP). Next, we propose a scale pooling scheme for better handling the objects that may appear in different shape, positions and scales. Furthermore, we raise a global constrained augmentation method to make more comprehensive predictions. The details of our contributions are described below:

Bag of Surrogate Parts (BoSP)

We take the feature maps as surrogate parts, and for each spatial unit on the feature maps, we calculate its assignment strengths for the surrogate parts by observing the activation values. The one-by-one processing of these spatial units can be viewed as densely sampling and assigning regions of the input image. Finally, we sum the assignment strengths for the surrogate parts and form a vector accordingly, i.e. BoSP, whose length is the same with the number of the feature maps.

On top of BoSP, we further propose its spatial variant, called S-BoSP, by dividing the image equally into multiple sub-regions (9 regions in 3 rows and 3 columns), and concentrating the BoSP inside each sub-region.

Scale Pooling

The BoSP/S-BoSP described above only concern the spatial units at the finest level, and handle them in a disjoint way, which means to sample and assign regions in input images with fixed size and position. To handle the objects with different scales and deformations, we propose a scale pooling scheme.

Specifically, for BoSP, we divide the activations from the pool_5 layer of VGG into 7 scales. Under different scales, we derive coarse spatial units of different numbers and different sizes, in which the coarse spatial units in scale 1 correspond to the fine spatial units on the feature map. For scale i ($i \in [1,7]$), we can derive $(8-i)^2$ coarse spatial units, each coarse unit contains i^2 fine spatial units. Next, we max pool these coarse spatial units and calculate their assignment strengths for the surrogate parts. Finally, we sum the assignment strengths of the coarse spatial units under different scales together to form the refined assignment strengths for the image. For S-BoSP, we utilize the scale pooling schemes inside each sub-region, and then concentrate the features together.

Global Constrained Augmentation

Given an input image, we first resize it to 224×224, and extract the global feature. This feature focuses more on the entire image, and we can achieve the global prediction based on it, denoted as Pre_{global}. Next, we resize the image to make the smallest side equal S while keeping its ratio, and crop regions of 224×224 with stride of 32 pixels. Thereby, we formulate several sub-images from the input image, each sub-image may only contain part of the original object. The image feature is the average of the sub-image features, and this feature concerns more about parts of the image, and based on it, we make the part prediction, denoted as Pre_{part}. The final prediction is the product of the global prediction and part prediction.

Experiment

There are some valuable findings through our experiment:

1) In terms of efficiency and accuracy, it is more beneficial to derive the BoSP feature from higher layer, i.e. pool_5, than lower layers.

2) Scale pooling method could improve the performance of BoSP/S-BoSP without enlarging the feature dimension, and the improvement can be very large. For example, scale pooling increases the BoSP features of Caltech101 and Oxford102 by more than 3%.

3) Regardless of the differences of the global-based prediction and part-based prediction, it is always beneficial to incorporate them by utilizing the global constrained augmentation scheme.

4) Our approach improves the state-of-the-art of Caltech101, Oxford102, SUN397 considerably, and achieves comparable result with previous best performance on Indoor67 dataset, while comes in lower dimension.

Multi-scale Colorectal Tumour Segmentation Using a Novel Coarse to Fine Strategy

Kun Zhang[1,2]
Danny Crookes[3]
Jim Diamond[4]
Minrui Fei[1]
Jianguo Wu[2]
Peijian Zhang[2]
Huiyu Zhou[3]
[1] Shanghai University, China
[2] Nantong University, China
[3] Queen's University Belfast, UK
[4] PathXL Ltd., Belfast, UK

Colorectal cancer is the third most common form of cancer worldwide [1]. In general, there are two challenges in automatic detection of colorectal cancer in histopathological images. One is the enormous volume of data which the algorithms have to cope with. The second is that in histopathology slides, cancerous tissue can look similar to noncancerous tissue.

In this paper, we propose an efficient and novel coarse to fine framework to address the problem of colorectal tumour segmentation for the purpose of tumour detection, as shown in Fig 1. In this way, we propose the use of colour modelling and morphological operations to extract the initial Region of interest (ROI). In order to reduce the noisy margin, we use Euclidean distance based histograms as a criterion. Further, to perform tumour segmentation at the best resolution, we deploy a Convergence Index (CI) approach to detect nuclei by fitting circles. At the tumour classification stage, when the resolution is at a high level, we use a rotation invariant feature and random projection based l_2-norm sparse representation technique for more accurate segmentation. The main contributions of our work include: 1) a multi-scale strategy for tumour segmentation, which simulates the decision making of a pathologist from the coarse-to-fine processing, and 2) a novel Rotation Invariant Raw Statistics (RIRS) feature and random projection based l_2-norm sparse representation method. These make the tumour classification process more effective than other published methods.

The RIRS-pixel feature can be calculated by:

$$x^{pixel} = [x_{0,0}, (x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4}), ..., (x_{8,1}, x_{8,2}, x_{8,3}, x_{8,4})]^T \quad (1)$$

where $x_{i,j}$ is a pixel with different scales and orientations. i is scale and j is orientation.

The new texton learning model is formed by:

$$min_{D,\alpha}(\| x - DK \|_2^2 + \lambda \| K \|_2^2 + \eta \sum_{i=1}^{n} \| \alpha_i - \delta \|_2^2) \ s.t \ d_j^T d_j = 1 \quad (2)$$

where $D = [d_1, d_2, ..., d_j], d_k \in R^m$ is the texton learning model. $K = [\alpha_1, \alpha_2, ..., \alpha_l]$ are coding coefficients. δ is the mean of all α_i which may vary by 50% more or less than the mean value. $\delta = 1/n \sum_{i=1}^{n} \alpha_i$ when distortion does not appear. When distortion appears, $\delta = 1/(n-u)(\sum_{i=1}^{n} \alpha_i - \sum_{f=1}^{u} \alpha_f)$, where u is the distortion.

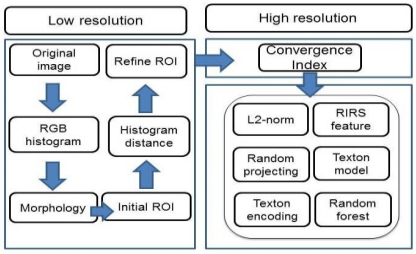


Figure 1: Overview of the proposed segmentation algorithm.

A total of 20 H&E stained colorectal cancer slides were supplied by a medical imaging company and used as the basis for training and testing. We compare our method against recently published texture classification methods (ELBP, MR8, TEISF, Patch, and SRP). The results favor the proposed approach in terms of classification accuracy.

[1] J. Ferlay, et al., Cancer incidence and mortality worldwide. International Agency for Research on Cancer, 2010.

Wednesday 13:40-14:40

Learning Additive Kernel For Feature Transformation and Its Application to CNN Features

Takumi Kobayashi
takumi.kobayashi@aist.go.jp

National Institute of Advanced Industrial Science and Technology
Tsukuba, Japan

Wednesday
13:40-14:40

Feature transformation, such as L_2-Hellinger [1] and an explicit map of χ^2 kernel [2], favorably improves classification performance for BoW and Fisher kernel features. In this paper, we propose a method to learn the feature transformation function of high generality and discriminative power in a bottom-up manner based on actual (annotated) data. The learned function corresponds to an explicit map of an additive kernel and the bottom-up learning approach endows the kernel function with discriminative power, adapting it even to the features of unknown characteristics; this is the case of CNN features which are of our main interest in this paper.

Given a pair of feature vectors, $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^D$, an additive kernel is defined by

$$\bar{\mathtt{k}}(\boldsymbol{x},\boldsymbol{y}) = \textstyle\sum_{i=1}^{D} \mathtt{k}(x_i,y_i) \approx \sum_{i=1}^{D} \boldsymbol{\phi}(x_i)^\top \boldsymbol{\phi}(y_i),$$

where x_i and y_i are the i-th elements of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\boldsymbol{\phi}$ is the explicit map function of the kernel $\mathtt{k}$. While the explicit map $\boldsymbol{\phi}$ can be determined according to the predefined kernel $\mathtt{k}$ [2], we learn the function from data in a bottom-up manner, resorting to the approximated representation of $\boldsymbol{\phi}$ by basis functions like the Fourier fashion:

$$\boldsymbol{\phi}(x) = \boldsymbol{W}^\top [\mathtt{f}_1(x), \cdots, \mathtt{f}_M(x)]^\top = \boldsymbol{W}^\top \mathbf{f}(x),$$

where $\boldsymbol{W} \in \mathbb{R}^{M \times K}$ is the coefficient matrix for the basis functions, and we define the basis functions $\{\mathtt{f}_m\}_{m=1}^{M}$ for CNN features as $\mathtt{f}_{2m-1}(x) = x\cos(2\pi\eta_m x)$, $\mathtt{f}_{2m}(x) = x\sin(2\pi\eta_m x)$ with the frequency parameters η_m. Thereby, our objective is to learn the coefficients $\boldsymbol{W}$ from data.

The feature vector $\boldsymbol{x}$ is transformed into $\boldsymbol{W}^\top \mathbf{F}(\boldsymbol{x}) \triangleq \boldsymbol{W}^\top [\mathbf{f}(x_1), \cdots, \mathbf{f}(x_D)]$ and thus the linear classifier is written by $\mathtt{tr}\{\boldsymbol{W}^\top \mathbf{F}(\boldsymbol{x})\boldsymbol{A}\} + b$ with a classifier weight $\boldsymbol{A}$ and a bias b. We apply an efficient approach to learn $\boldsymbol{W}$, providing a good trade-off between the generality and discriminativity, in two steps. First, we optimize the joint weight $\boldsymbol{V}_{\mathcal{D}} = \boldsymbol{W}\boldsymbol{A}_{\mathcal{D}}$ on various datasets, and then we extract $\boldsymbol{W}$ via SVD, which is shared across the various optimizers $\{\boldsymbol{V}_{\mathcal{D}_c}\}_{c=1}^{C}$.

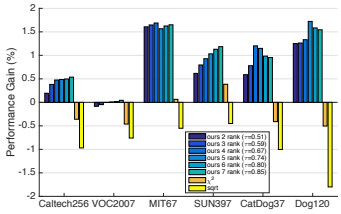


Figure 1: Performance results for CNN features.

The proposed method is applied to transform the pre-trained CNN features using the models of `Alex`, `VGG` and `C3D` on various datasets. We first assessed the generality by measuring how the learned (kernel) function fluctuates according to the training datasets, and found that the learned kernels exhibit quite high similarity, being almost identical, even though they are trained on different datasets of enough size. Thus, we can say that the proposed method produces the *generic* additive kernel (or feature transformation).

Next, the proposed method is compared with the other feature transformation methods, explicit map of χ^2 kernel [2] and Hellinger (square root) transformation [1], both of which are successfully applied to the hand-crafted features, such as BoW and Fisher kernel features. The performance results of `Alex` feature are shown in Fig. 1. In contrast to the case of hand-crafted features, the χ^2 and Hellinger transformations do not work well; particularly, the Hellinger transformation [1] degrades performance in all cases. On the other hand, the learned transformation favorably boosts the performance, outperforming the other methods. Through the bottom-up learning approach, the proposed method adapts the transformation function, *i.e.*, the additive kernel, to the CNN features whose characteristics are not fully revealed.

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

[2] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

Robust 3D Car Shape Estimation from Landmarks in Monocular Image

Yanan Miao
miaoyn12@mails.tsinghua.edu.cn
Xiaoming Tao
taoxm@mail.tsinghua.edu.cn
Jianhua Lu
lhh-dee@mail.tsinghua.edu.cn

Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Electronic Engineering
Tsinghua University
Beijing, P. R. China

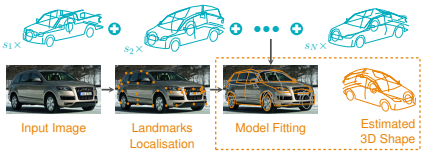


Figure 1: The framework for 3D shape estimation. **Top**: A series of prior 3D shape basis [2]. **Bottom**: The shape estimation procedure for a given input image.

Estimation of the 3D shape of a object from monocular image is an under-determined problem, which becomes harder when the observations are severely contaminated. In this paper, we propose a robust model to estimate 3D shape $\mathbf{X}$ from 2D landmarks $\mathbf{x} \in \mathbb{R}^{2\times p}$ with unknown camera pose M. The 3D shape of the object is assumed as a linear combination of predefined shape basis $\{X_i\}_{i=1}^N \in \mathbb{R}^{3\times p}$ weighted by $\mathbf{s} = [s_1,\ldots,s_N]^\mathsf{T} \in \mathbb{R}^N$. To estimate $\mathbf{s}$ and M, we fit the model by minimizing the error between the observations $\mathbf{x}$ and the projected model points M$\mathbf{X}$ (as shown in Figure 1).

Model. To address the outliers in the observed 2D points, which result from the complex background and illumination conditions, we propose a robust 3D shape estimation model. We explicitly model the outliers with an additional sparse error term $E \in \mathbb{R}^{2\times p}$. Thus, the robust model is then formulated as

$$\min_{\mathbf{s},\mathsf{M}} \frac{1}{2}\|\mathbf{x}-\mathbf{t}-\mathsf{M}\mathbf{X}-E\|_F^2 + \underbrace{\lambda\|\mathbf{s}\|_1 + \eta\|E\|_1}_{\text{non-smooth}}$$
$$s.t.\ \overbrace{\mathsf{M}\mathsf{M}^\mathsf{T} = \mathbf{I}_2}^{\text{non-convex}}, \mathbf{X} = \sum_{i=1}^N s_i X_i + \mu \tag{1}$$

where $\mathbf{t} = [t_x, t_y]^\mathsf{T} \cdot \mathbf{1}_{1\times p}$ is the translation, and λ, η are the regularization parameters, and μ is the mean shape. The objective function in (1) is non-convex and non-smooth constrained on Stiefel manifold, where the coupling of the unknown shape representation coefficients $\mathbf{s}$ and camera pose M makes it more difficult to be solved.

Method. We propose an efficient numerical algorithm based on Alternative Direction Method of Multipliers (ADMM) [1] to solve this problem. With an auxiliary variable $V \in \mathbb{R}^{2\times 3}$ introduced, the augmented Lagrangian is,

$$\mathcal{L}_{\mathsf{M},V,\mathbf{s},E,\mathbf{t},\Lambda} = \tfrac{1}{2}\|\mathbf{x}-\mathbf{t}-\mathsf{M}\mathbf{X}-E\|_F^2 + \lambda\|\mathbf{s}\|_1$$
$$+\eta\|E\|_1 + <\Lambda, \mathsf{M}-V> + \tfrac{\tau}{2}\|\mathsf{M}-V\|_F^2$$
$$s.t.\ \mathsf{M} = V, VV^\mathsf{T} = \mathbf{I}_2, \mathbf{X} = \textstyle\sum_{i=1}^N s_i X_i + \mu$$

where Λ is the multiplier and τ is penalty parameter. We update each block with all the others fixed. Based on some analysis on non-convex optimization of ADMM [3], we set the orthogonality constraints into the smooth sub-problem (V-minimization),

$$\min_V \{\|V - (\mathsf{M}^k + \Lambda^k/\tau^k)\|_F^2 : VV^\mathsf{T} = \mathbf{I}_2\}.$$

The closed-form solution is given by $V^{k+1} = U\mathbf{I}_{2\times 3}W^T$, where U and W satisfy $[U,S,W] = \mathcal{SVD}\left[\mathsf{M}^k + \Lambda^k/\tau^k\right]$. The other sub-problems can be easily solved. Both the optimization of M and $\mathbf{t}$ admit closed-form solutions. The updating of $\mathbf{s}$ is a *Lasso*-problem, and the sparse error pattern E can be efficiently solved by element-wise soft-thresholding. The convergences of ADMM with more than two blocks cannot be always guaranteed [1], and may be influenced by the update ordering. We set a fixed update ordering that can always lead convergence in our experiments.

[1] S. Boyd and etc. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[2] Y. Lin and etc. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, pages 466–480. Springer, 2014.

[3] Y. Zhang. Recent advances in alternating direction methods: Practice and theory. In *IPAM workshop*, 2010.

Projective Unsupervised Flexible Embedding with Optimal Graph

Wei Wang[1]
wei.wang@unitn.it
Yan Yan[1]
yan.yan@unitn.it
Feiping Nie[2]
feipingnie@gmail.com
Xavier Alameda Pineda[1]
xavier.alamedapineda@unitn.it
Shuicheng Yan[3]
eleyans@nus.edu.sg
Nicu Sebe[1]
niculae.sebe@unitn.it

[1] DISI,
University of Trento
Trento, Italy

[2] OPTIMAL,
Northwestern Polytechnical University,
Xi'an, China

[3] ECE,
National University of Singapore,
Singapore

Wednesday
13:40-14:40

Graph based dimensionality reduction techniques have been successfully applied to clustering and classification tasks. The fundamental basis of these algorithms is the constructed graph which dominates their performance. Usually, the graph is defined by the input affinity matrix. However, the affinity matrix is sub-optimal for dimension reduction as there is much noise in the data. To address this issue, we propose the projective unsupervised flexible embedding with optimal graph (PUFE-OG) model. We build an optimal graph by adjusting the affinity matrix. To tackle the out-of-sample problem, we employ a linear regression term to learn a projection matrix. The optimal graph and projection matrix are jointly learned by integrating the manifold regularizer and regression residual into a unified model. An efficient algorithm is derived to solve the challenging model. The experimental results on several public benchmark datasets demonstrate that the presented PUFE-OG outperforms other state-of-the-art methods.

When the labels are available, the most popular dimensionality reduction algorithm is linear discriminant analysis (LDA). It has excellent performance as LDA utilizes discriminant information to learn the subspace. In addition, simultaneously performing clustering and subspace learning can yield even better clustering result. [3] proposed an effective discriminating K-Means (DisKmeans) algorithm by integrating LDA and K-Means. However, labeled data are often very costly to obtain.

When the labels are unavailable, unsupervised dimensionality reduction methods become the only choice. For example, PCA is widely used because of its simplicity and efficiency. The unsupervised graph based dimensionality reduction methods usually outperform PCA. This is because these methods take advantage of manifold information. Many graph based dimensionality reduction methods have been explored, such as locally linear embedding (LLE), Laplacian eigenmap (LE), and ISOMAP. However, these methods suffer from out-of-sample problem. They can not map the new data points that are not included in the training set. To tackle this problem, many works [1] integrated the manifold regularizer with the ridge regression loss into the subspace learning framework. Similar to other manifold learning algorithms, their performance is also controlled by the graph constructed by the fixed affinity matrix, which might lead to a sub-optimal result [2]. To address this issue, we propose a projective unsupervised flexible embedding with optimal graph (PUFE-OG) framework. Instead of utilizing the fixed affinity matrix to preserve the manifold structure, we construct an optimal graph by adapting the affinity matrix for subspace learning.

[1] Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang, and Changshui Zhang. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *TIP*, 2010.

[2] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *SIGKDD*. ACM, 2014.

[3] Jieping Ye, Zheng Zhao, and Mingrui Wu. Discriminative k-means for clustering. In *NIPS*, 2008.

Towards Automatic Image Editing: Learning to See another You

Amir Ghodrati[3,1]
http://homes.esat.kuleuven.be/~aghodrat/
Xu Jia[3,1]
http://homes.esat.kuleuven.be/~xjia/
Marco Pedersoli[2]
marco.pedersoli@inria.fr
Tinne Tuytelaars[1]
http://homes.esat.kuleuven.be/~tuytelaa/

[1] ESAT-PSI
KU Leuven, iMinds
Leuven, Belgium

[2] THOTH
INRIA Grenoble
Grenoble, France

[3] equal contribution to this work and listed in alphabetical order

Problem Definition. We propose a method that aims at automatically editing an image by altering its attributes. More specifically, given an image of a certain class (*e.g.* a human face), the method should generate a new image as similar as possible to the given one, but with an altered visual attribute (*e.g.* the same face with a new pose or a different illumination).

Contributions. The main contributions of this paper are: i) the definition of a new problem, where the goal is to generate images as similar as possible to a source image yet with one attribute changed; ii) a solution that follows an encoder-decoder pipeline, where the desired attribute modification is first encoded then integrated at feature map level; iii) the insight that the result can be refined by adding another convolutional encoder-decoder model; and iv) good qualitative and quantitative results on different tasks on the MultiPIE dataset [1].

How. We propose a model following the encoder-decoder architecture. It takes a face image and the desired attribute as inputs. The image is encoded into a feature map by means of several convolutional layers. The desired attribute is also encoded into a feature map, so that the two flows of information can be deeply fused in the feature map level. A new image is then generated with a convolutional decoder module. The output image of this network is already reasonable, however it still has some missing details and artifacts. Thus, in order to refine the previously generated image, we introduce a second stage network. It takes as input the source image and the image generated from the first stage and generates a refined output image using again a convolutional encoder-decoder network. In summary, the first stage is in charge of rendering a global representation of the desired object, while the second stage focuses on local refinements to remove some artifacts.

Evaluation. We evaluate our method on three different tasks on the MultiPIE dataset. The main task is to rotate a face (Figure 1). We extensively evaluate our method for this task, showing both qualitative results and quantitative results which are measured in terms of per-pixel mean squared error (MSE) between the generated image and the ground-truth image. Our method shows better performance when compared with [2]. The other two tasks are generating faces with different illumination and filling in the missing part of a face image on synthetic data generated from MultiPIE (Figure 2).

Figure 1: Qualitative results of our image generation from test data of MultiPIE. From left to right are the input image, the ground-truth target image, the output of the first stage and the output of the second stage.

Figure 2: Qualitative results for the task of image inpainting. From left to right are the input image, images generated with our method and the original images without the occluding pattern.

[1] Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, 2010.

[2] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Du-Sik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015.

Wednesday
13:40-14:40

Dense Labeling with User Interaction: an Example for Depth-Of-Field Simulation

Ana B. Cambra
acambra@unizar.es
Adolfo Muñoz
adolfo@unizar.es
José J. Guerrero
jguerrer@unizar.es
Ana C. Murillo
acm@unizar.es

Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza
Zaragoza, Spain

Wednesday
13:40-14:40

The main contribution of this work is a novel pipeline for interactive dense labeling, which provides a framework that can be applied in any application that involves dense labeling and user interaction. Our approach is focused on efficient dense labeling estimation and is particularly well suited for the use of continuous magnitudes.

We define the image dense labeling problem as a linear system of equations, where the unknowns are the labels. We model the problem as a graph of superpixels [1]. We consider two equations types: *Unary equations* assign to the unknown label a numerical value based on the individual superpixel properties and on the user input; *Binary equations* establish relations between labels of two connected superpixels. The over-determined system obtained is solved with a common least squares method to find an approximate solution minimizing the error.

Our experiments are focused on the part of the pipeline that estimates the labeling (step 4), since it is the only common part in all the approaches compared. We compare our algorithm to: the state-of-the-art MRF-based dense labeling approaches [4], block coordinate descent algorithm [2] and a Random Walk based approach [3]. Our experiments show that our approach is the fastest to obtain a solution compared to related approaches while keeping comparable quality in the results.

Besides, we demonstrate how our pipeline is suitable for interactive applications developing an interactive application for depth-of-field simulated effects from a single image, which requires a fast dense depth estimation. Users only need to mark a few depth values in the image, from which our application estimates a dense depth map. Each pixel affected by a user stroke generates a unary equation that is added to the system (step 3). This, combined with the pre-computed binary equations (step 2), leads to a system that is interactively solved and yields a dense estimated depth map and the corresponding effect: we generate the simulated image by applying a variable Gaussian blur filter.

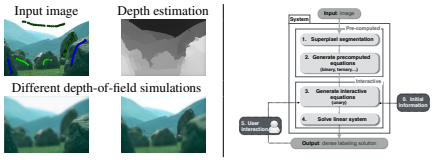


Figure 1: Left: Steps of the Interactive depth-of-field application. Right: Steps of our dense-labeling pipeline.

The proposed dense labeling technique has great flexibility to model this problem and has the advantage of providing an interactive solver. Besides, since we target an interactive technique the user can always refine and improve the input iteratively. We believe that our approach will inspire future research for interactive editing applications based on dense labeling.

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

[2] Qifeng Chen and Vladlen Koltun. Fast mrf optimization with application to depth reconstruction. In *CVPR*, pages 3914–3921, 2014.

[3] Leo Grady. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28 (11):1768–1783, 2006.

[4] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):1068–1080, 2008.

MLBoost Revisited: A Faster Metric Learning Algorithm for Identity-Based Face Retrieval

Romain Negrel
romain.negrel@unicaen.fr
Alexis Lechervy
alexis.lechervy@unicaen.fr
Frederic Jurie
frederic.jurie@unicaen.fr

Normandie Univ, UNICAEN,
ENSICAEN, CNRS
France

This paper focuses on the problem of identity-based face retrieval [2], a problem heavily depending on the quality of the similarity function used to compare the images. Instead of using standard or handcrafted similarity functions, one of the most popular ways to address this problem is to learn adapted metrics, from sets of similar and dissimilar example pairs. This is generally equivalent to projecting the face signatures into an adapted (possibly low-dimensional) space in which the similarity can be measured by the Euclidean distance. For large scale applications, the dimension of this subspace should be as small as possible to limit the storage requirements, and the projections should be fast to compute. Since the Euclidean distance fulfill the second requirement, producing face representations adapted to the Euclidean metric is interesting. However, such representations usually have very large sizes. Several methods have been proposed to learn projections that are capable of reducing the size of the signatures while preserving their performance. Most of these approaches are based on metric leaning algorithms [1] used to learn Mahalanobis-like distances:

$$D_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i-\mathbf{x}_j)^\top \mathbf{W}(\mathbf{x}_i-\mathbf{x}_j), \quad (1)$$

with $\mathbf{W}$ a positive semi-definite matrix. To guarantee this property and to reduce the size of the signatures, these methods use the factorization $\mathbf{W}=\mathbf{L}\mathbf{L}^\top$ with $\mathbf{L}\in\mathcal{M}_{D\times d}$ as projection matrix: $\mathbf{y}_i=\mathbf{L}^\top\mathbf{x}_i$. It is important to control the rank of $\mathbf{W}$ so that the dimension of the reduced signature is as small as possible.

In this paper, we focus on a particular metric learning algorithm so-called MLBoost [3], a supervised method based on boosting. MLBoost learns the metric incrementally by aggregating several weak metrics:

$$D_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j) = \sum_t \alpha^{(t)} \mathrm{D}_{\mathbf{z}^{(t)}}\left(\mathbf{x}_i,\mathbf{x}_j\right) \quad (2)$$

Sign.	Final Dim.	n= 1	n= 10	n= 20	n= 50	n= 100
LBP $D=9860$	FULL	31.9	53.7	60.5	68.8	74.7
	16	18.7	43.5	52.7	64.3	74.0
	32	31.4	57.0	63.1	72.1	77.3
	128	36.4	54.8	62.9	71.6	77.5
	512	**38.5**	**58.6**	**63.6**	**74.0**	**79.2**
VGG-Face $D=4096$	FULL	89.6	96.9	97.4	98.1	98.3
	16	82.0	94.1	96.7	97.6	98.6
	32	89.4	96.2	97.4	98.1	98.6
	128	90.8	95.7	97.2	98.1	**98.8**
	512	**92.4**	**96.7**	**97.6**	**98.3**	98.6

Table 1: Performance of MLBoost with low-cost weak metrics ($\tau = 5\%$) and rank constraints ($R \in \{16,32,128,512\}$) for face retrieval on the LFW Dataset [2] ("FULL" rows correspond to the performance of the non-reduced original signatures).

with $\alpha^{(t)}$ the weights of the weak metrics and $\mathbf{z}^{(t)}$ the projector vectors of these weak metrics. Here, we propose two improvements over MLBoost [3]: (i) A new method for **computing weak metrics at a lower computational cost**; (ii) An explicit way **to control the rank of the produced metric** during the learning phase, allowing to fix the size of the low-dimensional space in which the images are represented.

Our experiments demonstrated a more than $10\times$ speedup in addition to a significant improvement in the performance of the signatures even when their dimensions are strongly reduced (Table 1).

[1] A. Bellet, A. Habrard, and M. Sebban. *Metric Learning*. Morgan & Claypool Publishers, 2015.

[2] B. Bhattarai, G. Sharma, F. Jurie, and P. Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *ECCV Workshops*, pages 1–13, 2014.

[3] R. Negrel, A. Lechervy, and F. Jurie. Boosted metric learning for efficient identity-based face retrieval. In *BMVC*, volume 13, pages 1007–1036, 2015.

Learning Neural Network Architectures using Backpropagation

Suraj Srinivas
surajsrinivas@grads.cds.iisc.ac.in
R. Venkatesh Babu
venky@cds.iisc.ac.in

Department of Computational and Data Sciences
Indian Institute of Science
Bangalore, India

Recent work on model compression of deep neural networks have shown that some smaller models can perform just as well as large ones. In this work, we introduce the problem of architecture-learning, i.e; learning the architecture of a neural network along with weights. We start with a large neural network, and then learn which neurons to prune. We also propose a smooth regularizer which encourages the total number of neurons after elimination to be small. The resulting objective is differentiable and simple to optimize.

Architecture-Learning We wish to minimize the following objective.

$$\hat{\theta}, \hat{\Phi} = \arg\min_{\theta,\Phi} \ell(\hat{y}(\theta,\Phi), y) + \lambda \|\Phi\| \qquad (1)$$

Here, θ denotes the weights of the neural network, Φ the architecture, and ℓ denotes the loss function. The λ parameter trades-off between a good fit and a low complexity model. Here, $\|\Phi\|$ denotes the model complexity, which is simply the total number of neurons in our case.

Selecting width The strategy we follow to solve the Architecture Learning problem is outlined in Figure 1. To prune neurons in a layer, we multiply auxiliary gate variables to each neuron. These are either '0' or '1'. As a result, neurons with corresponding gate variables with '0' values can be pruned away. If we learn these gates along with weights, we effectively learn the width of neural network layers.

Selecting depth To reduce the depth of the network, we attempt to replace non-linearities with linear functions wherever possible. As a result, whenever linearities are present, two layers can be merged, as shown in Figure 1.

Tri-State ReLU The two disparate observations mentioned above are combined into a single non-linearity called the Tri-State ReLU (TSReLU), which is defined as follows.

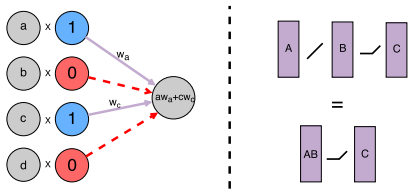


Figure 1: Our strategy for selecting width and depth. Left: Grey blobs denote neurons, coloured blobs denote the proposed auxiliary gate parameters. Right: Purple bars denote weight-matrices.

$$tsReLU(x) = \begin{cases} wx, & x \geq 0 \\ wdx, & otherwise \end{cases} \qquad (2)$$

The various modes of behaviour of this function for different sets of values of w and d is given in Table 1. As indicated, the proposed function can behave either can ReLU, a zero function or an identity function.

w	d	Behaviour
1	0	ReLU
1	trainable	Parameteric-ReLU [1]
0	any value	Returns 0 always
1	1	Identity function
0 or 1	0 or 1	Tri-State ReLU

Table 1: Various modes of behviour for different values of w, d.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

Wednesday
13:40-14:40

Enhancing pose estimation through efficient patch synthesis

Pierre Rolin — Université de Lorraine
https://members.loria.fr/PRolin/
Marie-Odile Berger — INRIA
http://www.loria.fr/~berger/
Frédéric Sur — Université de Lorraine
https://members.loria.fr/FSur/

Estimating the pose of a camera from a scene model is a challenging problem when the camera is in a position not covered by the views used to build the model, because feature matching is difficult. Several viewpoint simulation techniques have been recently proposed in this context. They generally come with a high computational cost, are limited to specific scenes such as urban environments or object-centred scenes, or need an initial pose guess. This paper presents a viewpoint simulation method well suited to most scenes and query views. Two major problems are addressed: the positioning of the virtual viewpoints with respect to the scene, and the synthesis of geometrically consistent patches.

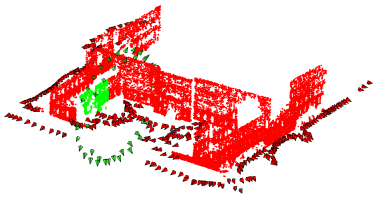


Figure 1: Virtual viewpoints (in green) positioning relative to one of the segmented patches (green points). The red viewpoints are the ones used to construct the model.

We propose a method to position the virtual viewpoints with respect to a segmentation of the scene in planar parts, the only assumption being that the scene is piecewise planar, which is not restrictive in most human-made environments. A set of virtual viewpoints is associated with each planar part. An adapted measure for viewpoint changes ensures that the existing viewpoints are completed with relatively few virtual viewpoints. Figure 1 illustrates the virtual viewpoints positioning relative to a segmented patch.

Viewpoint simulation techniques based on a scene model often generate images of the scene using a dense model [2] or generate local patches around every interest point [1]. The first approach fails in cases where some parts of the scene are not correctly densified, and the latter is computationally demanding without any pose guess. We propose an intermediate approach consisting in synthesizing semi-local planar patches of the scene and enriching the scene model with descriptors from these synthesized patches, using a visibility constraint.

Figure 2: In this scene the pose cannot be accurately estimated with a standard RANSAC-PnP approach (left). Patch synthesis improves the registration of the query view to the scene, as proved by the superposed scene edges (right).

The experiments show that a model enriched with the proposed synthesis method leads to more accurate poses, and even gives accurate poses when pose estimation simply fails without synthesis. In addition the time needed for pose computation is reduced, since the RANSAC step needs less iterations. Figure 2 shows an example where patch synthesis gives a dramatic improvement of pose estimation.

[1] P. Rolin, M.-O. Berger, and F. Sur. Viewpoint simulation for camera pose estimation from an unstructured scene model. In *Proc. International Conference on Robotics and Automation (ICRA)*, pages 6320–6327, 2015.

[2] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. pages 1808–1817, June 2015.

Discovering motion hierarchies via tree-structured coding of trajectories

Juan-Manuel Pérez-Rúa[1]
juanmanuel.perezrua@technicolor.com
Tomas Crivelli[1]
http://www.technicolor.com/en/tomas-crivelli
Patrick Pérez[1]
http://www.technicolor.com/en/patrick-perez
Patrick Bouthemy[2]
patrick.bouthemy@inria.fr

[1] Technicolor R&I
Cesson Sévigné, France

[2] Inria
Centre Rennes
Bretagne Atlantique, France

Wednesday
13:40-14:40

Figure 1: Hierarchical organization of visual motions in a natural scene.

The dynamic content of physical scenes is largely compositional, that is, the movements of the objects and of their parts are hierarchically organized and relate through composition along this hierarchy. This structure also prevails in the apparent 2D motion that a video captures (see Fig.1). Accessing this visual motion hierarchy is important to get a better understanding of dynamic scenes and is useful for video manipulation. In this paper, we aim at capturing the hierarchical video representation through learned, tree-structured sparse coding of point trajectories.

Given an input video sequence of $M+1$ frames and N input point trajectories extracted from it $(\mathbf{x}^n_{0:M}) \in \mathbb{R}^{2\times(M+1)}$, $n = 1\cdots N$, we define the data matrix $X \in \mathbb{R}^{2M\times N}$ as:

$$X = \begin{bmatrix} \Delta\mathbf{x}_1^1 & \Delta\mathbf{x}_1^2 & \cdots & \Delta\mathbf{x}_1^N \\ \Delta\mathbf{x}_2^1 & \Delta\mathbf{x}_2^2 & \cdots & \Delta\mathbf{x}_2^N \\ . & . & \cdots & . \\ \Delta\mathbf{x}_M^1 & \Delta\mathbf{x}_M^2 & \cdots & \Delta\mathbf{x}_M^N \end{bmatrix}, \qquad (1)$$

where $\Delta\mathbf{x}^n_m = \mathbf{x}^n_m - \mathbf{x}^n_{m-1}$. In this matrix, each column stems for the sequence of displacements along one trajectory.

A powerful way to discover multiple structures in such data is through sparse coding with a learned dictionary. However, this does not enforce any structure among the atoms of the dictionary and on the associated codes. We formulate the problem so that the dictionary and the encoding are constrained in certain way by a tree structure. We want the movement of a given scene element to be represented *only with dictionary atoms stemming from a same branch of the tree*. For a given rooted tree $\mathcal{T}$ of K nodes we learn a dictionary $D = [\mathbf{d}_{1:K}] \in \mathbb{R}^{2M\times K}$ of K trajectory atoms organized according to this tree structure, together with the corresponding matrix $A = [\alpha_{1:N}] \in \mathbb{R}^{K\times N}$ of sparse codes. To this end, we consider the following constrained minimization problem:

$$\arg\min_{D,A} \|X - DA\|_2^2, \quad \text{sb.t. } \alpha_n \in \mathcal{A}(\mathcal{T}),\ \forall n \quad \text{and} \quad \|\mathbf{d}_k\|_2 = 1,\ \forall k, \qquad (2)$$

where $\mathcal{A}(\mathcal{T}) \subset \mathbb{R}^K$ is the set of tree-structured codes defined as:

$$\mathcal{A}(\mathcal{T}) = \{\alpha \in \mathbb{R}^K : \text{supp}(\alpha) = \text{anc}(\text{k}(\alpha))\}, \qquad (3)$$

where $\text{anc}(k)$ denotes the ancestor set of node k in $\mathcal{T}$ (the nodes, including itself, that form the unique path from k to root node 1), $\text{supp}(\alpha)$ is the support of α, that is the index set of its non-zero entries, and $\text{k}(\alpha) = \max(\text{supp}(\alpha))$ stands for the last atom in the code.

We show in the paper how to use our algorithm for hierarchical and flat clustering of trajectories taken from video sequences. We also analyze hierarchical motion patterns that are common in human activities like walking and jumping. We show that our algorithm is capable, up to some level, to separate nested and independent motions.

Coplanar Repeats by Energy Minimization

James Pritts[1]
http://cmp.felk.cvut.cz/~prittjam
Denys Rozumnyi[1]
rozumden@cmp.felk.cvut.cz
M. Pawan Kumar[2]
http://mpawankumar.info
Ondrej Chum[1]
http://cmp.felk.cvut.cz/~chum

[1] The Center for Machine Perception
Faculty of Electrical Engineering
Czech Technical University
Prague, CZ

[2] Department of Engineering Science
University of Oxford
Oxford, UK

This paper proposes an automated method to detect, group and rectify arbitrarily-arranged coplanar repeats via energy minimization. We propose a global energy model for grouping coplanar repeats that combines features that encourage (i) the geometric consistency of repeated coplanar elements, (ii) the appearance similarity of planar repeated elements, (iii) the spatial and color coherence of scene planes, (iv) the spatial and color coherence of the background, (v) and the parsimony of detected coplanar repeat groups and scene planes.

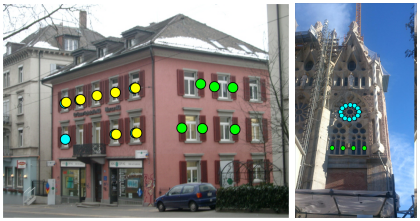


Figure 1: Selected detected coplanar repeats.

A block-coordinate descent framework is proposed for energy minimization that alternately assigns keypoints to coplanar repeats by labeling via a recent variant of α-expansion, and regresses the continuous parameters that model the geometries and appearances of coplanar repeat groups and their underlying scene planes.

We introduce a dataset of 113 images containing coplanar repeated patterns with translated, rotated and reflected symmetries that repeat arbitrarily or periodically. The dataset will be made publicly available[1].

To evaluate the performance of the proposed method, we compare against two state-of-the-art geometric multi-model fitting methods: J-Linkage and MultiRANSAC [3, 4]. The accuracy of rectifications constructed from vanishing lines computed from estimated coplanar repeat groups are used to compare the methods.

The cumulative distribution of distortions on the dataset (truncated at 10 pixels) is shown in Fig. 2. At 1 pixel of distortion, the proposed method solves 163% more scene planes than the next best; at 2 pixels, 94% more; and at 5 pixels, which can be considered a threshold for meaningful rectification, 51% more scene planes. The proposed energy minimization formulation demonstrates a distinct increase in the quality of rectifications estimated from detected coplanar repeat groups on the evaluated dataset with respect to two state-of-the-art geometric multi-model fitting methods. The advantage can be attributed to the global scene context that is incorporated into the energy functional of the proposed method.

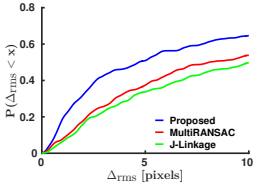


Figure 2: CDF of rectification distortions (Δ_{rms}).

[1] O. Chum and J. Matas. Planar affine rectification from change of scale. In *ACCV*, 2010.

[2] J. Pritts, O. Chum, and J. Matas. Detection, rectification and segmentation of coplanar repeated patterns. In *CVPR*, 2014.

[3] R. Toldo and A. Fusiello. Robust multiple structures estimation with j-linkage. In *ECCV*, 2008.

[4] M. Zuliani, C. Kenney, and Manjunath B. The multiransac algorithm and its application to detect planar homographies. In *ICIP*, 2005.

[1] http://ptak.felk.cvut.cz/personal/prittjam/bmvc16/coplanar_repeats.tar.gz

Wednesday
13:40-14:40

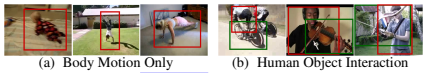
Two-Stream SR-CNNs for Action Recognition in Videos

Yifan Wang[1]
yifan.wang@student.ethz.ch
Jie Song[1]
jsong@inf.ethz.ch
Limin Wang[2]
07wanglimin@gmail.com
Luc Van Gool[2]
vangool@vision.ee.ethz.ch
Otmar Hilliges[1]
otmar.hilliges@inf.ethz.ch

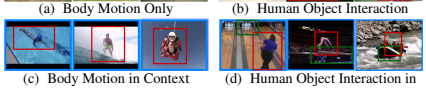
[1] Advanced Interactive Technologies Lab, ETH Zurich, Switzerland
[2] Computer Vision Lab, ETH Zurich, Switzerland

Wednesday
13:40-14:40

Action recognition belongs to the most challenging tasks in computer vision. An action is usually defined by multiple elements, called "cues", e.g. person, object and scene. Accordingly, common actions can be divided into four types as shown in Figure 1.



(a) Body Motion Only (b) Human Object Interaction



(c) Body Motion in Context (d) Human Object Interaction in Context

Figure 1: Action types with different composition of semantic cues, e.g. human body (red box), interacting objects (green box), and global context (blue box)

Despite the high overall classification accuracy [4], the conventional two-stream CNN approach [2] performs poorly on human-centric categories (shadowed plot in Figure 3). This discovery indicates overfitting from uninformative variances possibly from "scene". Hence we propose a semantically aware CNN-based framework for action recognition in video, which uses the locational information of various semantic cues as an explicit attention guidance during training and testing.

1 Approach

First of all, we propose a generic and efficient method to extract action relevant persons and objects from video sequences using the output of an object detector, e.g. Faster R-CNN [1]. This method recovers detection errors and removes irrelevant "bystanders" devoid of ground truth. The obtained bounding boxes are incorporated into the conventional two-stream CNNs network via a RoiPooling layer as shown in Figure 2. Each semantic cue constructs an individual channel, which is combined by a fusion layer to produce the final prediction.

2 Experiment and Result

We conduct a series of experiments on UCF101 dataset [3] and determine the best performing model, namely SR-CNNs with sum-fused person and scene

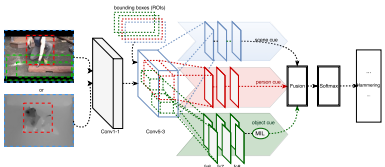


Figure 2: Architecture of two-stream SR-CNNs

channels, denoted as *S+P*. Our empirical study demonstrates that (1) our approach outperforms the original two-stream CNNs in terms of global accuracy (Table 1) (2) the robustness against ambiguous variances in scene (Figure 3) is improved (3) semantic channels exhibit complementary properties and improves spatial and temporal streams differently.

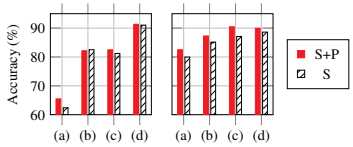


Figure 3: Performance comparison on UCF101 split1 in different action types defined in Figure 1. (left: spatial stream; right: temporal stream)

Models	Spatial	Temporal	Two Stream
S	77.93	86.79	91.15
S+P	**78.32**	**88.29**	**92.60**

Table 1: Comparison to conventional two-stream CNNs on UCF101 averaged over 3 splits

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[2] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[3] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[4] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

An Efficient Convolutional Network for Human Pose Estimation

Umer Rafi[1]
rafi@vision.rwth-aachen.de
Ilya Kostrikov[1]
ilya.kostrikov@rwth-aachen.de
Juergen Gall[2]
gall@iai.uni-bonn.de
Bastian Leibe[1]
leibe@vision.rwth-aachen.de

[1] Computer Vision Group, RWTH Aachen University, Germany

[2] Computer Vision Group, University of Bonn, Germany

In recent years, human pose estimation has greatly benefited from deep learning and huge gains in performance have been achieved on popular benchmarks [1, 3, 4]. The trend to maximise the accuracy on benchmarks, however, resulted in computationally expensive deep network architectures that require expensive hardware and pre-training on large datasets. In this work, we propose an efficient deep network architecture that can be efficiently trained on mid-range GPUs without the need of any pre-training and that is on par with much more complex models on the benchmarks [1, 3, 4].

Our proposed Fully Convolutional GoogLeNet (FCGN) network (see Figure 1) is based on the network architecture from [2]. We take the first 17 layers of [2] and add a deconvolution layer to make it fully convolutional. In addition, we introduce a skip layer and combine two FCGNs with shared weights to obtain a multi-resolution network. Belief maps for each joint are then obtained by a deconvolution layer with large kernel size in combination with a sigmoid function for normalisation and spatial drop out for regularisation.

We compare the performance of the proposed architecture against convolutional pose machines [5] on the well-known FLIC, LSP, and MPII benchmarks [1, 3, 4]. Our proposed network outperforms most previous approaches and achieves competitive performance to the more complex model of [5], while requiring only 3GB of memory and far less training time.

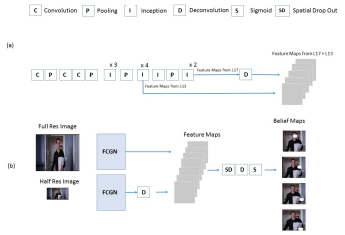


Figure 1: (a) Proposed fully convolutional GoogLeNet (FCGN) (b) The proposed multi-resolution network combines two FCGNs.

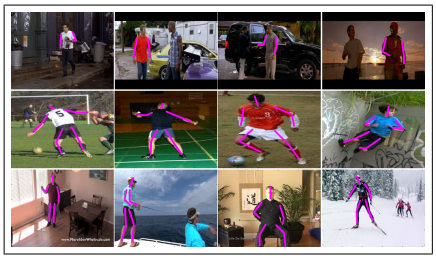


Figure 2: Our Qualitative results on FLIC [4], LSP [3] and MPII [1].

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New BenchMark. In *CVPR*, 2014.

[2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015.

[3] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*, 2010.

[4] B. Sapp and B. Taskar. MODEC : Multimodel Decomposable Models for Human Pose Estimation. In *CVPR*, 2013.

[5] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *CVPR*, 2016.

A data augmentation methodology for training machine/deep learning gait recognition algorithms

Christoforos C. Charalambous
http://www.bicv.org/team/christoforos-charalambous/
Anil A. Bharath
http://www.bicv.org/team/anil-anthony-bharath/

BICV Research Group
Department of Bioengineering
Imperial College London
London, UK

In this paper we propose a data augmentation methodology for training machine/deep learning gait recognition algorithms. While previously published methods generated synthetic data to augment training and/or testing sets [1] or to learn features invariant to certain conditions [2], they have incorporated a very limited number of covariate factors. To our knowledge, this is the first attempt to provide the ability to simultaneously generate synthetic data with so many controllable conditions for gait recognition.

The combination of real motion capture, data preparation, avatar construction and scripted rendering allows synthetic frames to be generated, with almost arbitrary degrees of variation. Figure 1 shows a small sample of the controllable confounding factors that can be generated with the proposed methodology.

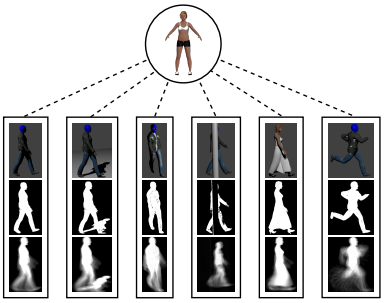


Figure 1: A small sample of the controllable confounding factors. It is readily apparent how the extracted gait features are directly affected.

Using the Gait Energy Image (GEI) [3] as gait features, extracted directly from the synthetically generated frames, we performed experiments to assess the level to which identity is preserved within the synthetic data sets. Results from our experiments – presented in Figure 2 – suggest that information about the identity of subjects is retained within the synthetically generated data. The experiments using GEIs as features within our dataset showed that augmenting a limited amount of real data with the synthetically generated data can yield identification of subjects with an accuracy of more than 95%. The results of this study suggest that synthetic data can be used to augment the training of gait recognition algorithms provided that the feature set (e.g. principal components) is expanded to include both real and synthetic examples of data.

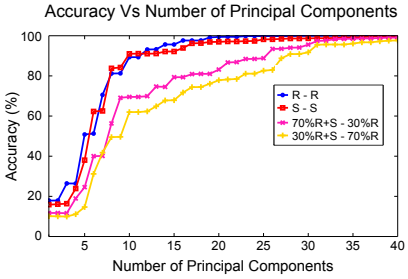


Figure 2: Results from our experiments, with different training and testing sets, showing accuracy achieved with GEI features being projected onto different number of principal components.

The proposed methodology offers the possibility to generate sequences with multiple confounding factors, allowing exploratory work into training machine/deep learning algorithms for fully-invariant gait recognition, with a far greater amount of synthetic training and test data that would otherwise be impossible. The dataset and simulation files will be made publicly available.

[1] J. Shotton et al. Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[2] J. Zhang et al. Arbitrary view action recognition via transfer dictionary learning on synthetic training data. *IEEE International Conference on Robotics and Automation*, page 1678–1684, 2016.

[3] J. Han, B. Bhanu. Statistical feature fusion for gait-based human recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:0–5, 2004.

Mean Box Pooling: A Rich Image Representation and Output Embedding for the Visual Madlibs Task

Ashkan Mokarian
ashkan@mpi-inf.mpg.de
Mateusz Malinowski
mmalinow@mpi-inf.mpg.de
Mario Fritz
mfritz@mpi-inf.mpg.de

Scalable Learning and Perception
Max Planck Institute for Informatics
Saarbrücken, Germany

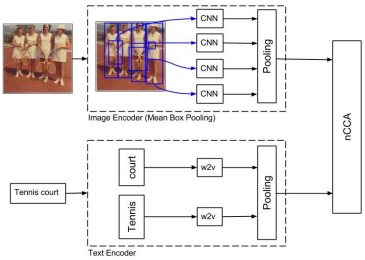


Figure 1: Our proposed image representation with Mean Box Pooling.

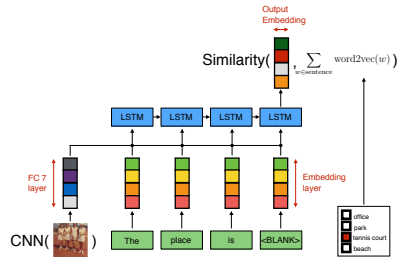


Figure 2: Embedded CNN+LSTM directly optimizes loss of answers in the embedding space.

Question answering about real-world images is a relatively new research direction that requires a chain of machine visual perception, natural language understanding, and deductive capabilities to successfully come up with an answer on a question about the visual content. In our experiments we consider the multi-choice Visual Madlibs dataset [1] as the ambiguities in the output space are rather minimal for this task.

In our paper, we present two novel architectures for this task. First, we argue for a rich image representation in the form of pooled CNN representations of highly overlapping object proposals, which we call Mean Box Pooling. Such a representation allows for a more fine grained, multi-scale and multi-part object analysis compared to global CNN representations. The overview is shown in Figure 1. Second, motivated by the popularity of deep architectures for visual question answering, which combine a global CNN image representation with an LSTM question representation, as well as the leading performance of nCCA on the multi-choice Visual Madlibs task, we propose a novel extension of the CNN+LSTM architecture, which we call Embedded CNN+LSTM, which chooses a prompt completion out of four candidates by comparing them directly in the embedding space at test time. This contrasts with the prior work of [1] that uses a post-hoc comparison between the discrete output of the CNN+LSTM method and all four candidates. Our method integrates more tightly with the multi-choice filling the blanks task, and significantly outperforms the prior CNN+LSTM methods [1]. This architecture is depicted in Figure 2.

We evaluate our methods on the multiple choice task of the aforementioned Visual Madlibs dataset. Interestingly, a large number of highly overlapping proposals significantly improves performance and even outperforms MSCOCO ground truth bounding boxes. nCCA with the proposed image representation outperforms the prior work [1] by 5.9 and 1.4 percent points in average on Easy and Hard tasks respectively. Although nCCA tops the leaderboard on the Visual Madlibs task, the largest body of work on the question answering about images combines a CNN with an LSTM. We hypothesize that the comparison for the multiple choice task should be directly done in the output embedding space. Our results, improving over [1] by 7 and 7.6 percent points in average on Easy and Hard tasks respectively, confirm our hypothesis.

[1] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual madlibs: Fill in the blank image generation and question answering. In *ICCV*, 2015.

Wednesday
13:40-14:40

Practical View on Face Presentation Attack Detection

Naser Damer
naser.damer@igd.fraunhofer.de
Kristiyan Dimitrov
kristiyan.dimitrov@igd.fraunhofer.de

Smart Living & Biometric Technologies
Fraunhofer Institute for Computer Graphics Research (IGD)
Darmstadt, Germany

Wednesday
13:40-14:40

Face recognition is one of the most socially accepted forms of biometric recognition. The recent availability of very accurate and efficient face recognition algorithms leaves the vulnerability to presentation attacks as the major challenge to face recognition solutions. Previous works have shown high preforming presentation attack detection PAD solutions under controlled evaluation scenarios. This work tried to analyze the practical use of PAD by investigating the more realistic scenario of cross-database evaluation and presenting a state-of-the-art performance comparison. The work also investigated the relation between the video duration and the PAD performance. This is done along with presenting an optical flow based approach that proves to outperform state-of-the-art solutions in most experiment settings

The presented solution is tested on multiple databases: the *REPLAY-ATTACK*, the *MSU-MFSD* and the *CASIA-FASD*. Each of these datasets includes subsets for training and testing to evaluate the algorithm performance.

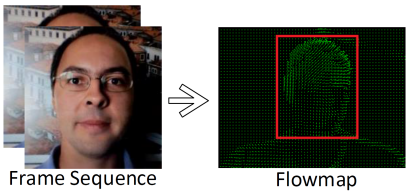


Figure 1: Optical flows from a face video.

The Presented PAD solution was based on optical flow in a similar manner to the Histogram of Oriented Optical Flow feature extractor (Fig.1) along with an AdaBoost classifier. Feature and score-level fusion approaches were utilized to enhance the performance.

A table of performance comparison between the proposed solution and state-of-the-art works were presented in the paper. This comparison included cross-database performance evaluation measures.

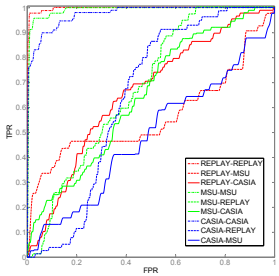


Figure 2: ROC curves achieved by the "F-triple: S-mean" approach on intra- and cross-database evaluation.

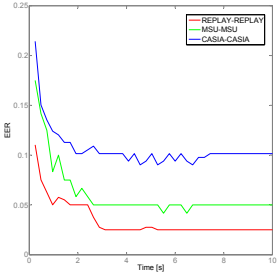


Figure 3: The performance (represented by the EER) development at different video lengths (time).

The performance degradation is clear when performing cross-database evaluation (Fig.2). However, the cross-database performance improves when considering whole videos with score-level fusion. The development of the PAD performance over the duration of the video is studied and shown in Fig.3.

To conclude, this work addressed the issue of realistic face PAD performance through evaluating it over a number of databases. This was presented along with a state-of-the-art performance comparison and a proposed solution based on optical flows.

Fast Feature-Less Quaternion-based Particle Swarm Optimization for Object Pose Estimation From RGB-D Images

Giorgio Toscana
giorgio.toscana@polito.it
Politecnico di Torino, Turin, Italy
Stefano Rosa
stefano.rosa@polito.it

We present a novel quaternion-based formulation of *Particle Swarm Optimization* (PSO) for pose estimation which, differently from other approaches, does not rely on 2D or 3D features or machine learning. The quaternion formulation avoids the gimbal lock problem and, unlike other rotation formalisms, doesn't require conversions to and from rotation matrix form at each step. The objective function is based on raw 2D depth information only, under the assumption that the object region is segmented from the background. This makes the algorithm suitable for pose estimation of objects with large variety in appearance, from lack of texture to strong textures, for the task of robotic grasping. We find candidate object regions using a graph-based image segmentation approach, but the PSO is agnostic to the segmentation algorithm used. The algorithm is implemented on GPU, and the nature of the objective function allows high parallelization.

The orientation of the i-th particle is updated using the discrete form of the quaternion kinematics. Each particle renders its pose hypothesis against the depth map of the cluster. The fitness value of the j-th particle is thus computed as:

$$\Phi_j = \frac{\alpha}{N_{R_j}} \sum_{i=1}^{N_{R_j}} \left(z_{K_i} - z_{R_{ij}} \right)^2 + \beta \, \frac{\mu_j + \kappa_j}{2} \qquad (1)$$

where: N_{R_j} is the number of pixels of the depth map rendered by the j-th particle, $z_{R_{ij}}$ is the depth value of the pixel i rendered by the j-th particle, while z_{K_i} is the corresponding depth value of the cluster at pixel i. α and β are two constant parameters used to weight the two terms of the fitness function. μ_j gives the percentage of cluster pixels that are not covered by the rendering of the object 3D model of the particle j. κ_j is the complement of μ_j i.e., it gives the percentage of rendered pixels of the particle j that are not covered by valid depth values in the cluster depth map.

The segmentation phase provides a rough approximation of the 3D centroid of a cluster.

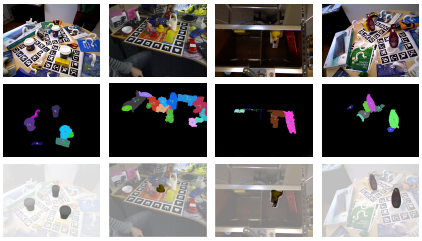


Figure 1: Examples of the approach on different datasets. First row: RGB images; second row: segmented objects; third row: detected objects superimposed.

The segmentation step cannot generate an estimate of the object orientation, so the object attitude is initialized on the surface of the northern hemisphere of $\mathbb{S}^3$. The rendering of each particle is done directly in GPU. Our rendering pipeline is based on the optimized version of the **edge function**.There are two levels of parallelism: each particle renders its object model independently and the rendering of the triangles for each model is also parallelized.

We tested our approach on two public datasets for 3D pose estimation [1] and [2]. The software has been developed using the CUDA library in C++ under Linux and runs on GPU. In our experiments we used 1024 particles and global topology for the PSO and run 10 PSO iterations for each segmented cluster and the total time is 85ms for each cluster. Results are comparable to [1] and [2], without requiring a training phase. Results are shown in Figure 1.

[1] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. 2011.

[2] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In *Computer Vision–ECCV 2014*, pages 462–477. Springer, 2014.

Graph Based Convolutional Neural Network

Michael Edwards

Xianghua Xie
http://www.csvision.swan.ac.uk

Department of Computer Science
Swansea University
Swansea, UK

In this paper we present a method for the application of Convolutional Neural Network (CNN) operators for use in domains which exhibit irregular spatial geometry by use of the spectral domain of a graph Laplacian, Figure 1. This allows learning of localized features in irregular domains by defining neighborhood relationships as edge weights between vertices in graph G. By formulating the domain as a fixed graph representation and projecting the observed data onto G as a graph signal f we are able to utilize the convolution theorem via a graph Fourier transform, matrix multiplication with the column-wise eigenvector matrix U, and elementwise multiplication with spectral filters k to learn feature maps (1).

$$y = U \sum_{i=1}^{I} U^T f_{s,i} \odot k_{i,o} \tag{1}$$

We introduce novel gradient calculations for the convolution operator backpropagation step in regards to both f (2) and k (3). These new calculations are shown to provide higher accuracy and stability compared to calculations presented by [2] and [1].

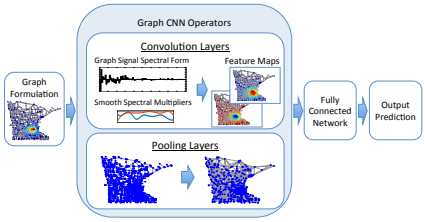


Figure 1: Graph based Convolutional Neural Network components.

The gradient calculation in regards to signal $f_{s,i}$ is given by (2), a spectral convolution of output loss $\nabla y_{s,o}$ with current weights of the spectral filters $k_{i,o}$.

$$\nabla f_{s,i} = U \sum_{o=1}^{O} U^T \nabla y_{s,o} \odot k_{i,o} \tag{2}$$

Gradients for the spectral filters are provided by (3), which are shown to improve over those of [2] in Figure 2.

$$\nabla k_{i,o} = \sum_{s=1}^{N} U^T \nabla y_{s,o} \odot U^T f_{s,i}. \tag{3}$$

We also present the use of Algebraic Multigrid as a method of graph coarsening, an analogy to the pooling operator of conventional CNNs, agglomerating nodes from the previous layer into a singular node in the subsequent layer. As with standard CNNs this provides both a reduction in graph complexity and generalization of learnt features.

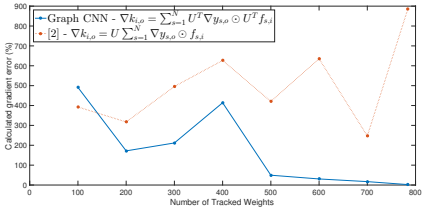


Figure 2: Gradient calculation errors for interpolation of various numbers of tracked weights.

Although this method is adaptable to numerous domains, we evaluate performance on a regular 2D pixel grid and an irregular grid with subsampled spatial geometry with the MNIST digit classification problem projected onto the graph. By utilizing (2) and (3) we obtain accuracy rates of 94.23% and 94.96% for the regular and irregular spatial domains respectively.

[1] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *CoRR*, abs/1312.6203, 2013.

[2] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *CoRR*, abs/1506.05163, 2015.

I have seen enough: Transferring parts across categories

David Novotny[1]
david@robots.ox.ac.uk
Diane Larlus[2]
diane.larlus@xrce.xerox.com
Andrea Vedaldi[1]
http://www.robots.ox.ac.uk/~vedaldi

[1] Visual Geometry Group
University of Oxford

[2] Computer Vision Group
Xerox Research Centre Europe

Figure 1: We study the transferability of object parts and ask a question: What is the minimal amount of supervision so that part detectors trained on source classes perform well on unseen target classes? To provide an answer, we experiment with a novel Animal Parts dataset.

Recent progress in image understanding, while dramatic, has been primarily fueled by the availability of increasingly large quantities of labeled data. However, it is unclear whether manual supervision will be able to keep up with the demands of increasingly sophisticated and data-hungry algorithms. This raises the obvious question: *When is supervision enough?*

In this paper we examine this question from the viewpoint of learning shareable semantic parts, focusing on two research problems.
(i) We conduct the first careful investigation of part transferability across a large set of visually dissimilar classes. To this end, the problem is viewed from the *Domain Adaptation* (DA) perspective where domains correspond to animals.
(ii) We investigate how many annotations are required to train a part detector. We consider an *Active Learning* scenario, studying which images should be chosen and how many are needed to saturate performance of the part detector.

ImageNet Animal Parts dataset. A thorough evaluation of the transferability of parts requires a suitable dataset with a large number of classes. Because existing datasets are insufficient for our task, we annotated "foot" and "eye" part keypoints in 14711 existing ImageNet images from 100 animal classes. Example annotations can be seen in Figure 1.

Proposed methods. Our part detectors rely on convolutional neural networks. We adapt the widely used *uncertainty* sampling principle to our scenario. We further propose an efficient ensemble of detectors which is optimized for generalization to new classes and can be used to guide active learning.

Main results

Part transferability. We challenge the common assumption that parts are visually "shareable" across different classes and therefore it suffices to learn them from a limited number of classes to understand them equally well in all cases. Given an unseen target class, by comparing performances of detectors trained on its semantically nearest/farthest classes we have verified the relevance of semantic distance for cross-domain transfer. Furthermore, we have shown that learning parts from a sufficiently diverse set of classes allows satisfactory transfer to novel classes.

Have I seen enough? We also ask how many source images need to be annotated in the source domains such that the performance of a part detector saturates as quickly as possible. Results indicate that our proposed method outperforms other baselines on foot detection while performing on par for eyes. Besides the relative merits of compared methods, the performance reaches 98% of the accuracy of the fully annotated scenario by providing only a few thousand annotations, showing that excellent performance can be achieved by annotating a small representative subset of classes and images.

Wednesday
13:40-14:40

Impatient DNNs – Deep Neural Networks with Dynamic Time Budgets

Manuel Amthor
manuel.amthor@uni-jena.de
Erik Rodner
erik.rodner@uni-jena.de
Joachim Denzler
joachim.denzler@uni-jena.de

Computer Vision Group
Friedrich Schiller University Jena
Germany
www.inf-cv.uni-jena.de

Wednesday
13:40-14:40

We introduce Impatient Deep Neural Networks (DNNs) for dealing with dynamic time budgets during inference time. They allow for either individual budgets given before for each test example or for anytime prediction, *i.e.* a possible interruption at multiple stages during inference while still providing output estimates. Figure 1 visualizes these advantages. Our approach can therefore tackle the computational costs and energy demands of DNNs in an adaptive manner, a property essential for real-time applications.

Our framework for learning dynamic budget predictors is based on risk minimization. We consider inference algorithms f providing predictions $y \in \mathcal{Y}$ for input examples $\boldsymbol{x} \in \mathcal{X}$ at different times $t \in \mathbb{R}$, *i.e.* we have $f : \mathcal{X} \times \mathbb{R} \to \mathcal{Y}$. Learning the parameters $\boldsymbol{\theta}$ of f is done by minimizing the regularized risk

$$\min_{\boldsymbol{\theta}} \iiint_{(t,y,\boldsymbol{x})} \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x},t),y) \cdot p(\boldsymbol{x},y,t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}y\,\mathrm{d}t + \mathcal{R}(\boldsymbol{\theta})$$

where $\mathcal{L}$ is a suitable loss function, $\mathcal{R}(\boldsymbol{\theta})$ a regularization term, and $p(\boldsymbol{x},y,t)$ the joint distribution of an input-output pair $(\boldsymbol{x},y)$ and the available time t.

This framework leads to a very flexible and simple learning scheme for deep neural networks. In particular, the resulting architecture of our networks contains "early prediction layers" directly connected to loss layers. The parameters of all of the layers are learned jointly by minimizing a weighted combination of the loss layers where the loss weights are directly computed using the distribution of time budgets for the application.

Experiments and Evaluation We present experimental results for different architectures, *e.g.* AlexNet and VGG19, on various object classification datasets to answer the most interesting question: Does our joint training scheme provide superior results compared to learning predictors independently? We compare our approach with different baselines that learn several SVM classifiers based on extracted CNN activations at each early prediction layer using an original CNN pretrained on ImageNet (ORIG) and a pre-trained CNN fine-tuned on the current dataset (FT). Our joint learning of early prediction layers shows superior results for most time budget distributions (*cf*. Table 1) with up to 10% absolute improvement on MIT-67 compared to the baseline.

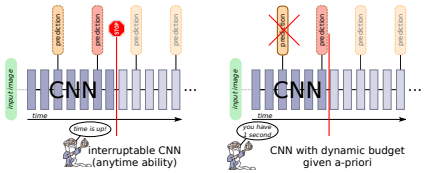


Figure 1: Convolutional neural network prediction in dynamic budget scenarios: (left) prediction can be interrupted at **any time** or (right) the budget is given **before** each prediction.

Table 1: Comparison of an Impatient VGG19 with several baselines on MIT-67. Performance is measured by expected accuracy in % based on the particular budget distribution $p(t)$.

Budget Scheme	Orig	FT	Ours
– uniformly distributed time budgets	46.7	48.1	**53.9**
◞ large time budgets are likely	62.8	67.1	**69.7**
◟ small time budgets are likely	25.6	25.7	**35.1**
∧ normal distributed time budgets	47.5	47.9	**55.4**

Summary In our paper, we present a novel approach for anytime prediction with deep neural networks, which can be easily adapted to any convolutional neural network architecture. Joint training with weighted losses give superior results for different time budget distributions compared to independently trained early predictors. Furthermore, we show that early prediction layers allow for reducing computational costs in the case of being already certain about intermediate classification results.

Maximum Margin Linear Classifiers in Unions of Subspaces

Xinrui Lyu[1,2]
xinrui.lyu@epfl.ch
Joaquin Zepeda[1]
joaquin.zepeda@technicolor.com
Patrick Pérez[1]
patrick.perez@technicolor.com

[1] Technicolor
35576, Cesson-Sevigne, France

[2] École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland

Formulation. We propose learning linear classifiers $\mathbf{w}_k = \mathbf{D}\mathbf{z}_k$ for classes $k = 1, \ldots, K$ so that $\mathbf{z}_k \in \mathbb{R}^A$ is sparse and learned from a training set $\{\mathbf{x}_i\}_{i=1}^M$ along with the dictionary $\mathbf{D} \in \mathbb{R}^{d \times A}$ using an adaptation of the ℓ_2 Support Vector Machine (SVM) objective:

$$\underset{\mathbf{D},\{\mathbf{z}_k\}_k,b}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{i=1}^{M} \max\left(0, 1 - y(\mathbf{x}_i^\top \mathbf{D}\mathbf{z}_k + b)\right) + \frac{\alpha}{2}\|\mathbf{D}\mathbf{z}_k\|_2^2 + \beta\|\mathbf{z}_k\|_1. \quad (1)$$

The resulting $\mathbf{D}$ can be fixed in a latter stage for a never-before-seen class where only the classifier's sparse $\mathbf{z}$ is learned. Regardless, the resulting classifiers will exist in a union of subspaces, with each subspace being the span of a small subset of atoms from $\mathbf{D}$. Hence we refer to our proposed classifier as a *Union-of-Subspaces SVM* (US-SVM). We further introduce Non-Negative (NN), Elastic Net (EN), and mixed NN+EN US-SVM variants.

Advantages. One first benefit of classifiers of the form $\mathbf{w} = \mathbf{D}\mathbf{z}$ is that only the sparse vector $\mathbf{z}$ needs to be stored for each class, resulting in a smaller storage footprint. Furthermore, for fixed $\mathbf{D}$, learning $\mathbf{z}$ results in lower computational cost both at training and testing time. Another benefit is that the atoms (columns) of the learned dictionary will inherit semantic properties shared by different classes and hence can often be interpreted as semantic *attributes* (Fig. 1), thus opening a possible path to weakly supervised attribute discovery. In a similar manner, atoms of the learned dictionary will often correspond to modalities of the underlying feature distribution that can likewise have interesting semantic interpretations. Forcing the classifier to be sparse using a learned dictionary can also be interpreted as a novel SVM regularization scheme. Unlike other schemes that constrain the norm of the classifier, our regularization requires that all classifiers be represented in terms of a common dictionary, in effect enabling the system to leverage the annotations for all classes when learning any given class.

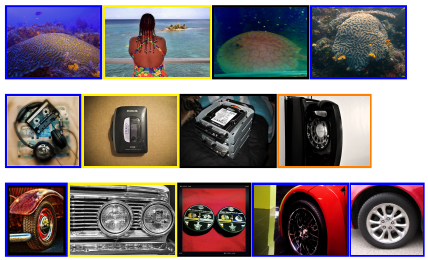


Figure 1: Examples of visual attributes captured by three different learned atoms across multiple classes. *Top*: water; *middle*: rectangular shape; *bottom*: round shape. For each row, images with the same border color belong to the same class.

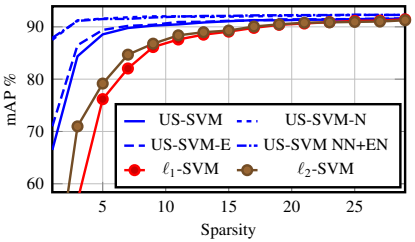


Figure 2: Classification performance versus sparsity (average $|\mathbf{z}_k|_0$) on a subset of ImageNet for US-SVMs with and without Non-negativity (N) constraints and Elastic net (E) penalization, and ℓ_1/ℓ_2-SVMs.

Experiments. In Fig. 2 we present example results on ImageNet that illustrate how US-SVM enjoys nearly constant performance for drastically low sparsity levels of < 5 (for feature vectors $\mathbf{x}_i \in \mathbb{R}^{128}$), where the performance is close to 20 mAP points better (a +20% difference) than that of ℓ_1-SVM or ℓ_2-SVM.

Wednesday
13:40-14:40

Aerial image geolocalization from recognition and matching of roads and intersections

Dragos Costea[1,2]
onorabil@gmail.com

Marius Leordeanu[1,2]
leordeanu@gmail.com

[1] University "Politehnica" of Bucharest
Bucharest, Romania

[2] Autonomous systems
Bucharest, Romania

Wednesday
13:40-14:40

We propose a complete pipeline for aerial image geolocalization based on roads and intersections. The main steps are road detection, followed by intersection detection and intersection matching with publicly available road vectors from the OpenStreetMap project (OSM) [1]. An initial localization is proposed, further improved by geometric alignment of roads.

The main goal of this project is to provide drones a reliable localization system in the areas we expect most autonomous deployments will be made - that is, urban and suburban areas.

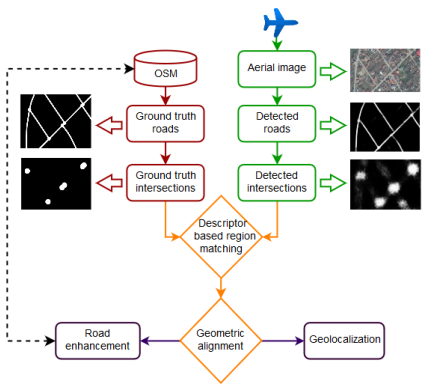


Figure 1: Framework overview

A key insight of our approach is the observation that intersections tend to have a unique road pattern surrounding them and thus can play a key role in localization, by reducing this difficult task to a sparse feature matching problem followed by a local refined road map alignment.

For pixel-wise road detection we used a dual-stream local-global CNN model proposed in [3]. Alexnet was used for intersection detection based on the detected road pattern.

[1] https://www.openstreetmap.org/

For intersection matching, we introduce a novel dataset consisting of images centered on intersections from two cities (one for training and the other one for testing), totalling 7204 600x600px images. A 4096-element descriptor is generated for each intersection using the surrounding detected roads and a neural network trained for intersection detection, in a way that is similar to [2]. We further fine-tune the network in a Siamese-like fashion in order to improve matching performance.

After an initial set of corresponding intersections is returned, we pick the best one by geometrical alignment of road maps for intersection candidates using shape context[1] and RANSAC. Further road enhancements for OSM are possible once the location has been determined.

We notice that most errors (around or above 90% of them) are below 2.5 meters, that is below 3 pixels for the image resolution available in our experiments. We believe that our results demonstrate high level of localization accuracy for our system, which could be very effective in most cases when the GPS signal is lost, for both day and nighttime.

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, volume 2, page 3, 2000.

[2] Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. Deep learning of binary hash codes for fast image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 27–35, 2015.

[3] Alina Marcu and Marius Leordeanu. Dual local-global contextual pathways for recognition in aerial imagery. *arXiv preprint arXiv:1605.05462*, 2016.

Learning local feature descriptors with triplets and shallow convolutional neural networks

Vassileios Balntas[1]
http://www.iis.ee.ic.ac.uk/~vbalnt
Edgar Riba[2]
eriba@cvc.uab.es
Daniel Ponsa[2]
daniel@cvc.uab.es
Krystian Mikolajczyk[1]
k.mikolajczyk@imperial.ac.uk

[1] Imperial College London
London, UK

[2] Computer Vision Center, Computer Science Department
Universitat Autònoma de Barcelona
Bellaterra (Barcelona), Spain

Finding correspondences between images via local descriptors is one of the most extensively studied problems in computer vision due to the wide range of applications. Recently, end-to-end learnt descriptors based on Convolutional Neural Networks (CNNs) have significantly outperformed state of the art features.

In this paper we investigate the use of triplets in learning local feature descriptors with CNNs and we propose a novel in-triplet hard negative mining step to achieve a more effective training and better descriptors. Our method reaches state of the art results without the computational overhead typically associated with mining of negatives and with lower complexity of the network architecture. This is a significant advantage over previous CNN descriptors since it makes our proposal suitable for practical problems involving large datasets.

Learning with triplets involves training from samples of the form $\{\boldsymbol{a},\boldsymbol{p},\boldsymbol{n}\}$, where $\boldsymbol{a}$ is the *anchor*, $\boldsymbol{p}$ is a *positive* example, which is a different sample of the same class as $\boldsymbol{a}$, and $\boldsymbol{n}$ is a *negative* example, belonging to a different class than $\boldsymbol{a}$. In our case, $\boldsymbol{a}$ and $\boldsymbol{p}$ are different viewpoints of the same physical point, and $\boldsymbol{n}$ comes from a different keypoint. The goal is to learn the embedding $f(\boldsymbol{x})$ s.t. $\delta_+ = ||f(\boldsymbol{a}) - f(\boldsymbol{p})||_2$ is low (i.e., the network brings $\boldsymbol{a}$ and $\boldsymbol{p}$ close in the feature space) and $\delta_- = ||f(\boldsymbol{a}) - f(\boldsymbol{n})||_2$ is high (i.e., the network pushes the descriptors of $\boldsymbol{a}$ and $\boldsymbol{n}$ far apart).

Anchor swap for in-triplet negative mining Previous proposals based on triplet based learning use only two of the possible three distances within each triplet, ignoring the distance $\delta'_- = ||f(\boldsymbol{p}) - f(\boldsymbol{n})||_2$. We take it into account to define the *in-triplet hard negative* as $\delta_* = min(\delta_-, \delta'_-)$. If $\delta_* = \delta'_-$, we swap $\{\boldsymbol{a},\boldsymbol{p}\}$, and thus $\boldsymbol{p}$ becomes the *anchor*, and $\boldsymbol{a}$ becomes the *positive* sample. This ensures that the hardest negative inside the triplet is used for backpropagation.

We build our descriptor by training a shallow network architecture from $5M$ triplets sampled on-the-fly using patches extracted around interest points. We evaluate its performance in patch pair classification, where we measure the ability of the descriptor to discriminate positive patch pairs from negative ones, and in nearest neighbour patch matching, where we measure the descriptor precision in matching feature points between different views of a same scene. Our networks outperform previously introduced convolutional feature descriptors. Moreover, they are 10 to 50 times faster than previous approaches. In fact, when running on GPU we reach speeds of $10\mu s$ per patch, which is comparable with the CPU speeds of fast binary descriptors. Details of our proposal are described more fully in the paper, along with extensive experimental work. We provide all the learned models and the training code for all descriptor variants at `https://github.com/vbalnt/tfeat`.

[1] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, 2015.

[2] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.

[3] G. H. M. Brown and S. Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1), pp.43-57, 2011.

Wednesday
13:40-14:40

Online Feature Selection for Visual Tracking

Giorgio Roffo
giorgio.roffo@univr.it
Simone Melzi
simone.melzi@univr.it

Department of Computer Science
University of Verona
Verona, Italy

Wednesday
13:40-14:40

Object tracking is one of the most important tasks in many applications of computer vision. Many tracking methods use a fixed set of features ignoring that appearance of a target object may change drastically due to intrinsic and extrinsic factors. The ability to dynamically identify discriminative features would help in handling the appearance variability by improving tracking performance. The contribution of this work is threefold. Firstly, this paper presents a collection of several modern feature selection approaches selected among filter, embedded, and wrapper methods (e.g., Inf-FS [2], mRMR, SVM-RFE, among others). Secondly, we provide extensive tests, regarding the classification task, on the PASCAL VOC-2007 dataset intended to explore the strengths and weaknesses of the selected methods with the goal to identify the right candidates for online tracking. Taking advantage from the results obtained from the offline scenario, we decided to use the following four candidate methods: *MutInf*, *Fisher*, *Inf-FS*, and *mRMR*. In particular, we take care that execution times of these methods meet the requirements for a real-time application. Finally, we show how the selected algorithms can be successfully employed for ranking features used by the Adaptive Color Tracking (ACT) system proposed in [1], maintaining high frame rates. ACT system is one of the most recent solutions for tracking, it exploits color naming [3] (i.e., the action of assigning linguistic color labels to image pixels), to target objects and learn an adaptive correlation filter by mapping multi-channel features into a Gaussian kernel space. We evaluated four variants of ACT on the OTB-50 [4] benchmark, throughout three different robustness evaluation metrics: one-pass evaluation (OPE), temporal robustness evaluation (TRE), and spatial robustness evaluation (SRE) reporting the average precision and success rate for quantitative analysis. In Figure 1 only the top 10 (out of 34) trackers are displayed for clarity. In particular, the unsupervised method ACTinffs turns out to be the best trade-off between accuracy (62.0%) and speed (111.4 fps). The ACTinffs has the same order of magnitude of the baseline ACT (196 fps) in terms of fps, while ACTMI operates at 19 fps. This work demonstrates the importance of feature selection for real-time applications, resulted in what is clearly a very impressive performance, our solutions improve up to 7% the baseline ACT while providing superior results compared to 29 state-of-the-art tracking methods.

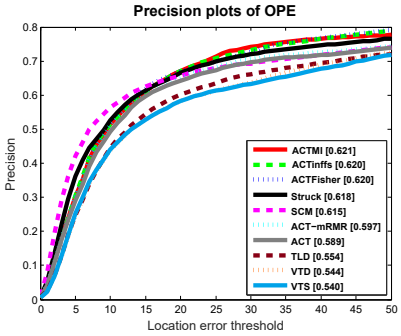


Figure 1: Precision plots over all 50 sequences provided by the OTB-50 benchmark. The mean precision scores for each tracker are reported in the legend.

[1] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer. Adaptive color attributes for real-time visual tracking. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] Giorgio Roffo, Simone Melzi, and Marco Cristani. Infinite feature selection. In *IEEE Conf. International Conference on Computer Vision (ICCV)*, 2015.

[3] Joost van de Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 2009.

[4] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Deep Aggregation of Local 3D Geometric Features for 3D Model Retrieval

Takahiko Furuya
takahikof AT yamanashi.ac.jp

Ryutarou Ohbuchi
ohbuchi AT yamanashi.ac.jp

Integrated Graduate School of Medicine,
Engineering, and Agricultural Sciences,
University of Yamanashi,
Kofu, Japan

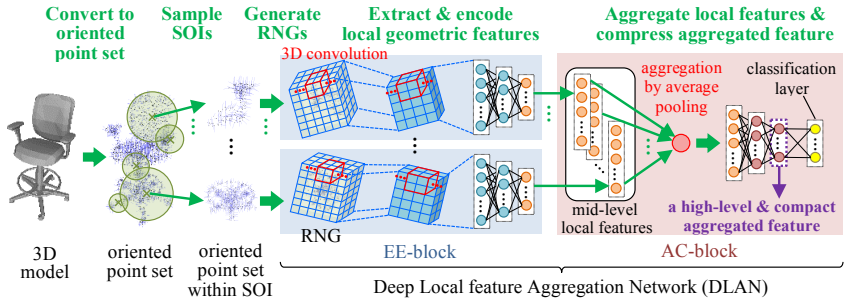


Figure 1: DLAN architecture for extracting a rotation-invariant and salient feature per 3D model.

Aggregation of Local Features (*ALF*) is a well-studied approach for image as well as 3D model retrieval (3DMR). A carefully designed local 3D geometric feature is able to describe detailed local geometry of 3D model, often with invariance to geometric transformations including 3D rotation of local 3D regions. For efficient 3DMR, these local features are aggregated into a feature per 3D model by using Bag-of-Features, Fisher Vector coding, etc.

Recent trend is to use end-to-end 3D Deep Convolutional Neural Network (3D-DCNN) (e.g., [1]) for 3DMR. 3D-DCNNs have often shown accuracies better than methods based on ALF. However, current 3D-DCNN based methods have weaknesses; they lack invariance against 3D rotation and/or they often miss geometrical details as they coarsely quantize shapes into voxels in applying 3D-DCNN.

Our goal is to extract a 3D model feature that is invariant against 3D rotation and more accurate than the existing ALF and 3D-DCNN based approaches. To this end, we combine ALF with 3D-DCNN.

We propose a novel deep neural network for 3DMR called *Deep Local feature Aggregation Network* (*DLAN*) that performs extraction of rotation-invariant 3D local features and their aggregation by using a single deep architecture. A DLAN (Figure 1) first describes local 3D regions of a 3D model by using "mid-level" local features invariant to 3D rotation. The set of local features is aggregated into a rotation-invariant and compact feature vector per 3D model.

Experimental evaluation using three benchmark datasets shows effectiveness of the DLAN. Here, we present results on the ModelNet40 dataset [1]. The proposed DLAN significantly outperforms the state-of-the-arts including 3D-DCNN based [1] and 2D-DCNN based [2][3] 3DMR algorithms.

algorithms	MAP [%]
3D ShapeNets [1]	49.2
MVCNN [2]	79.5
GIFT [3]	81.9
DLAN (proposed)	**85.0**

Table 1: Comparison of retrieval accuracy.

[1] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling. *Proc. CVPR 2015*, 1912–1920, 2015.

[2] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-View Convolutional Neural Networks for 3D Shape Recognition. *Proc. ICCV 2015*, 945–953, 2015.

[3] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki. GIFT: A Real-time and Scalable 3D Shape Search Engine, *Proc. CVPR 2016*, 5023–5032, 2016.

Learning Grimaces by Watching TV

Sam Albanie
http://www.robots.ox.ac.uk/~albanie
Andrea Vedaldi
http://www.robots.ox.ac.uk/~vedaldi

Engineering Science Department
Univeristy of Oxford
Oxford, UK

Differently from computer vision systems which require explicit supervision, humans can learn facial expressions by simply observing other humans in their environment. In this paper, we consider the problem of developing similar capabilities in machine vision. As a starting point, we look at the problem of relating facial expressions to objectively measurable events occurring in videos and make four contributions towards this goal. Firstly we construct and make available *FaceValue*, a dataset of facial expressions labelled with events to facilitate the study of this problem. Second, we evaluate existing emotion recognition CNN architectures on standard benchmarks and demonstrate the value of pre-training on face related tasks to compensate for a scarcity of labelled training data for emotion recognition. Third, we provide human baselines for the difficulty of emotion recognition in general, and specifically the difficulty of predicting events from expressions on our new dataset. Finally, we extend the standard emotion recognition architectures to predict events in videos and learn nameable expressions from them.

FaceValue Dataset

The *FaceValue* dataset comprises 192,030 faces collected from 102 episodes of the TV gameshow "Deal or No Deal" which provides a diverse source of facial expressions and game events. We associate event labels with face tracks, where an event label consists of a sum of money that has been removed from the contestant's potential prizes (see Figure 1 for examples), together with its position in the sequence of events that have taken place in the game.

Emotion Recognition Models and Human Baselines

We train and evaluate a number of CNN architectures on the FER and SFEW 2.0 emotion recognition benchmarks and show that pre-training for the task of face verification produces a substantial jump in performance (+6% average relative improvement on FER vs Imagenet pre-training), and single model test accuracies of 72.89% (FER) and 59.41% (SFEW 2.0).

Figure 1: ***FaceValue*** dataset. *Top row:* detection of an event in the game, and the corresponding reaction of the contestant's face. *Bottom:* four example tracks, the top two for "good" events and the bottom two for "bad" events (see paper for details)

Learning Expressions

We adapt the CNN architectures for emotion recognition to our primary task of learning expressions from events in the *FaceValue* dataset. Despite its challenging nature, we show that CNNs can perform well at the task of predicting event labels directly from expressions. Similarly to the FER and SFEW 2.0 benchmarks, the best model marginally outperforms the accuracy of a committee of human annotators.

Conclusions

Experimental results show that learning facial expressions from contextual events rather than directly labelled data is challenging, but feasible. The dataset and emotion recognition models are available at `http://www.robots.ox.ac.uk/~vgg/data/facevalue`.

Accurate and robust face recognition from RGB-D images with a deep learning approach

Yuancheng Lee
http://cv.cs.nthu.edu.tw/php/people/profile.php?uid=150
Jiancong Chen
http://cv.cs.nthu.edu.tw/php/people/profile.php?uid=153
Ching-Wei Tseng
http://cv.cs.nthu.edu.tw/php/people/profile.php?uid=156
Shang-Hong Lai
http://www.cs.nthu.edu.tw/~lai/

Computer Vision Lab,
Department of Computer Science,
National Tsing Hua University,
Hsinchu, Taiwan

In this paper, we propose a face recognition system based on deep learning, which can be used to verify and identify a subject from the colour and depth face images captured with a consumer-level RGB-D camera. To recognize faces with colour and depth information, our system contains 3 parts: depth image recovery, deep learning for feature extraction, and joint classification.

In the depth image recovery and enhancement stage, a pipeline is applied to improve quality of depth face images from a consumer-level depth camera. First, re-meshing and coarse-to-fine depth fusion are used to alleviate the random depth loss and noise of a depth map and project the facial surface point clouds into 3-D space. Second, a template facial landmark set is computed and frontalized, so that we can align the other faces (point clouds) onto the template using landmark-based transformation for frontalization and compute their head poses (vertical and horizontal rotation). Third, by re-projecting the fused and frontalized 3-D point cloud onto a canvas which is 2 times of the original resolution, we can obtain a super-resolved depth image by resampling. If there are still any holes on the depth image. we can fill the holes by Poisson Blending [1], with super-resolution depth image as background, pre-computed mean depth face image as foreground, and detected hole pixel map as mask. Last, to get high quality 3-D face mesh model, we just mesh and project the super-resolved depth map into 3-D space again. To synthesize depth maps from different view angles for a single 3-D face model, we rotate the model horizontally and then vertically before rendering.

To learn discriminative feature transformation by deep network, we first train our network on CASIA-WebFace [2] dataset for colour (RGB and greyscale) face images. The model for greyscale images is further fine-tuned on the merged depth dataset for transfer learning.

Database similarity standard deviation is proved to be highly correlated to reliability of similarity, and can be viewed as an estimation of image quality. A support vector classifier [3] with probability output and pairwise information as input is used to estimate the confidence score that a pair of images are come from the same subject. The SVM is trained with the following feature: group-wise colour/depth similarity, Average database colour/depth similarity standard deviation of 2 images, and estimated capture-time head pose difference. A binary label indicates whether the 2 images of each pair are from the same subject.

Our experiments show that higher accuracy can be achieved by using the proposed bi-model confidence estimation, especially under harsh illumination environment or large head pose variation.

[1] P. Perez, M. Gangnet, and A. Blake, "Poisson Image Editing" *ACM Siggraph*, 2003.

[2] D. Yi, Z. Lei, S. Liao and S. Z. Li, "Learning Face Representation from Scratch" *Computer Vision and Pattern Recognition*, 2015.

[3] Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classifiers." Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992.

Global Deconvolutional Networks for Semantic Segmentation

Vladimir Nekrasov
nekrasowladimir@unist.ac.kr
Janghoon Ju
janghoon.ju@unist.ac.kr
Jaesik Choi
jaesik@unist.ac.kr

Ulsan National Institute of Science and Technology
50 UNIST, Ulju, Ulsan, 44919 Korea

Wednesday
13:40-14:40

Motivation. Semantic segmentation is a crucial computer vision task, solving which would enable a thorough scene understanding of the environment. The areas that already benefit from the automatic semantic segmentation include biomedical imaging [2], autonomous driving [4]. Further enhancement of current models will necessarily increase the number of possible applications, as well as quality of performance. The transfer learning of deep convolutional networks pre-trained for image classification on ImageNet has proven to be successful in semantic segmentation. In these models, last fully-connected layers are replaced by convolutional ones followed by a learnable deconvolution or fixed interpolation to acquire the output of the same spatial size as the input. Usually, the segmented mask is coarse. Several ways to deal with this have been proposed, including the 'skip'-layer architecture [3] and post-processing with probabilistic graphical models [1].

Algorithm. The combination of graphical models with deep networks requires carefully designed differentiable operations to mimic approximate inference, while traditional upsampling approaches tend to operate only locally. To overcome these issues, we propose an alternative novel architecture aimed to perform an upsampling globally, as well as enforce the correct label recognition. For the first task, we propose the equivalent of deconvolution, which we call *'global interpolation'*. We denote the decoded information of the RGB-image $\mathbf{I} : \mathbf{I} \in \mathbb{R}^{3\times H\times W}$, as $\mathbf{x} : \mathbf{x} \in \mathbb{R}^{C\times h\times w}$, where C represents the number of channels, h and w define the reduced height H and width W, respectively. To acquire $\mathbf{y} : \mathbf{y} \in \mathbb{R}^{C\times H\times W}$, an upsampled signal, we apply the following formula:

$$\mathbf{y_c} = \mathbf{K_h}\mathbf{x_c}\mathbf{K_w}^{\mathrm{T}}, \forall \mathbf{c} \in \mathbf{C} \qquad (1)$$

where the matrices $\mathbf{K_h} \in \mathbb{R}^{H\times h}$ and $\mathbf{K_w} \in \mathbb{R}^{W\times w}$ are interpolating each feature map of $\mathbf{x}$ through the corresponding spatial dimensions. Contrary to a simple bilinear interpolation, which operates only on the closest four points, Eq. (1) allows to include much more information on the rectangular grid. Besides that, we append an additional *multi-label classification loss* to correct the wrong predictions of the network. The complete architecture can be seen in Figure 1.

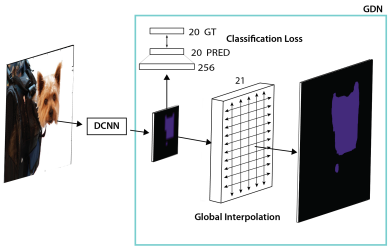


Figure 1: **Global Deconvolutional Network**. Our adaptation of FCN-32s [3]. We upsample the reduced signal with the help of global interpolation and append the multi-label classification loss to increase the recognition accuracy.

Results. We evaluate the proposed approach extending two publicly available models: FCN [3] and DeepLab [1]. We show the superior performance over them and achieve 74.02% mean IoU on the test set of the PASCAL VOC benchmark, which is close to the state-of-the-art performance without exploiting larger datasets.

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[2] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *NIPS*, 2012.

[3] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[4] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.

Convolutional Sparse Coding-based Image Decomposition

He Zhang
he.zhang92@rutgers.edu
Vishal M.Patel
http://www.rci.rutgers.edu/~vmp93

Electrical and Computer Engineering
Rutgers, the state University of New Jersey
NJ, USA

Most of the existing dictionary-based image decomposition methods are patch-based and features learned with these methods often contain shifted versions of the same features [1]. To overcome this issue, we propose a novel sparsity-based method for cartoon and texture decomposition based on Convolutional Sparse Coding (CSC). Our method first learns a set of generic filters that can sparsely represent cartoon and texture type images. Then using these learned filters, we propose a sparsity-based optimization framework to decompose a given image into cartoon and texture components. By working directly on the whole image, the proposed image decomposition algorithm does not need to divide the image into overlapping patches for leaning local dictionaries.

Our goal is to separate the given input image $\mathbf{y}$ into a cartoon part $\mathbf{y}_c$ and a texture part $\mathbf{y}_t$. Assume that we have already learned the convolutional filters corresponding to $\mathbf{y}_c$ and $\mathbf{y}_t$ by solving the CSC problem for the cartoon and the texture components separately. That is, we have learned $\{\mathbf{d}_{c,k}\}_{k=1}^{K_c}$ and $\{\mathbf{d}_{t,k}\}_{k=1}^{K_t}$ such that $\mathbf{y}_c = \sum_{k=1}^{K_c} \mathbf{d}_{c,k} * \mathbf{x}_{c,k}$ and $\mathbf{y}_t = \sum_{k=1}^{K_t} \mathbf{d}_{t,k} * \mathbf{x}_{t,k}$, where $\mathbf{x}_{c,k}$ and $\mathbf{x}_{t,k}$ are the sparse coefficients that approximate $\mathbf{y}_c$ and $\mathbf{y}_t$ when convolved with the filters $\mathbf{d}_{c,k}$ and $\mathbf{d}_{t,k}$, respectively. We propose to estimate $\mathbf{y}_c$ and $\mathbf{y}_t$ via $\mathbf{x}_{c,k}$ and $\mathbf{x}_{t,k}$ by solving the following CSC-based optimization problem

$$\begin{aligned}\arg\min_{\mathbf{x}_{c,k},\mathbf{x}_{t,k}} \frac{1}{2}\left\|\mathbf{y} - \sum_{k=1}^{K_c} \mathbf{d}_{c,k} * \mathbf{x}_{c,k} - \sum_{k=1}^{K_t} \mathbf{d}_{t,k} * \mathbf{x}_{t,k}\right\|_2^2 \\ + \lambda_c \sum_{k=1}^{K_c} \left\|\mathbf{x}_{c,k}\right\|_1 + \lambda_t \sum_{k=1}^{K_t} \left\|\mathbf{x}_{t,k}\right\|_1 \\ + \beta TV\left(\sum_{k=1}^{K_c} \mathbf{d}_{c,k} * \mathbf{x}_{c,k}\right).\end{aligned} \tag{1}$$

We present the results of our proposed CSCD algorithm for image decomposition and compare them with the MCA method [4], adaptive MCA (A-MCA) method [3], and a recent stat-of-the-art Block Nuclear Norm (BNN) method [2]. In these experiments, we use the Peak Signal to Noise Ratio (PSNR) to measure the performance of the routines tested (See Figure 1). Various experiments show the significance of our CSC-based image separation method over the other methods.

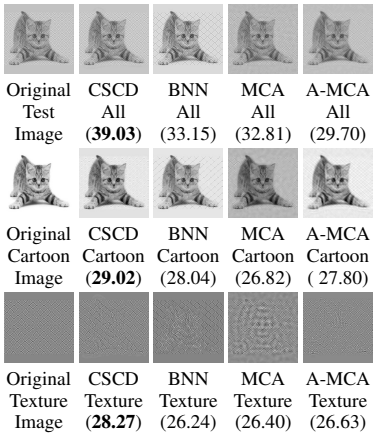


Figure 1: Image decomposition results on the *Cat+Cage* image.

[1] Michael Elad. Prologue. In *Sparse and Redundant Representations*, pages 3–15. Springer, 2010.

[2] Shintaro Ono, Takamichi Miyata, and Isao Yamada. Cartoon-texture image decomposition using blockwise low-rank texture characterization. *Image Processing, IEEE Transactions on*, 23(3): 1128–1142, 2014.

[3] Gabriel Peyré, Jalal Fadili, and Jean-Luc Starck. Learning the morphological diversity. *SIAM Journal on Imaging Sciences*, 3(3):646–669, 2010.

[4] Jean-Luc Starck, Michael Elad, and David L Donoho. Image decomposition via the combination of sparse representations and a variational approach. *Image Processing, IEEE Transactions on*, 14(10):1570–1582, 2005.

Domain Adaptive Subspace Clustering

Mahdi Abavisani
mahdi.abavisani@rutgers.edu
Vishal M. Patel
http://www.rci.rutgers.edu/~vmp93

Electrical and Computer Engineering Department
Rutegrs, The State University of New Jersey
New Brunswick, NJ, USA

Many practical applications in image processing and computer vision require one to analyze and process high-dimensional data. It has been observed that these high-dimensional data can be represented by a low-dimensional subspace. As a result, the collection of data from different classes can be viewed as samples from a union of low-dimensional subspaces. In subspace clustering, given the data from a union of subspaces, the objective is to find the number of subspaces, their dimensions, and the segmentation of the data and a basis for each subspace. In many applications, one has to deal with heterogeneous data. For example, clustering face images collected in the wild, one may have to cluster images of the same individual collected using different cameras and possibly under different resolution and lighting conditions. Clustering of heterogeneous data is difficult because it is not meaningful to directly compare the heterogeneous samples with different distributions which may span different feature spaces. In recent years, various domain adaptation methods have been developed to deal with the distributional changes that occur after learning a classifier for supervised and semi-supervised learning [3]. However, to the best of our knowledge, these methods have not been developed for clustering heterogeneous data that lie in a union of low-dimensional subspaces.

In this paper, we present domain adaptive versions of the sparse and low-rank subspace clustering methods (i.e. SSC [1] and LRR [2]). Figure 1 gives an overview of the proposed method. Given data from K different domains, we simultaneously learn the projections and find the sparse or low-rank representation in the projected common subspace. Once the projection matrices and the sparse or low-rank coefficient matrix is found, it can be used for subspace clustering.

We evaluated the performance of our domain adaptive subspace clustering methods on three publicly available datasets - UMD-AA01 face dataset, Amazon-DLSR-Webcam office data- sets, and USPS-MNIST-Alphadigits handwritten digits datasets. Table 1 shows the results of our proposed domain adaptive LRR (DA-LRR) method on the UMD-AA01 face dataset. Our method performs better than some recent stat-of-the-art domain adaptive subspace clustering methods.

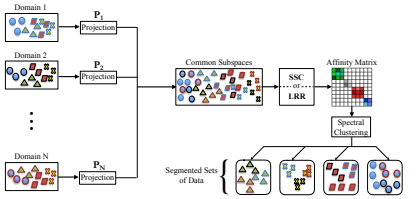


Figure 1: An overview of the proposed domain adaptive subspace clustering framework.

Method	$\{1\} \rightarrow \{2\}$	$\{2\} \rightarrow \{1\}$
LRR	52.04	53.26
CO-LRR	46.73	47.96
DA-LRR	**36.76**	**35.51**
ED-LRR	42.45	42.04
GM-LRR	47.76	47.14

Table 1: Average clustering errors on the UMD-AA01 face dataset. Note that $\{1\}$ and $\{2\}$ correspond to session 1 and session 2, respectively. DA-LRR denotes our proposed method.

[1] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11): 2765–2781, 2013.

[2] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (1):171–184, 2013.

[3] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *Signal Processing Magazine, IEEE*, 32(3): 53–69, 2015.

Filtering 3D Keypoints Using GIST For Accurate Image-Based Localization

Charbel Azzi[1]
cazzi@uwaterloo.ca
Daniel Asmar[2]
da20@aub.edu.lb
Adel Fakih[1]
afakih@uwaterloo.ca
John Zelek[1]
jzelek@uwaterloo

[1] Systems Design Engineering
University of Waterloo
Waterloo, Canada

[2] Vision and Robotics Lab
Mechanical Engineering Department
American University of Beirut
Beirut, Lebanon

Image-Based localization (IBL) addresses the problem of estimating the 6 DoF camera pose in an environment, given a query image and a representation of the scene. The tree-based approach [2] is the standard solution for IBL. When dealing with large-scale environments, the need to reduce the search space of the tree-based becomes the main focus. Sattler et al. [3] reduced the search space by clustering the 3D points into bag-of-words. This approach is well known to trade accuracy for speed due to the quantization effect. Recently, Kendall et.al [1] used deep convolutional neural networks to solve the problem. The accuracy of this approach is enough for location recognition applications, but is not enough to compete with the accuracy of the main IBL systems. In this paper we propose the Gist-based Search Space Reduction (GSSR) system to reduce the search space by finding candidates keyframes in the database, then match against the 3D points that are only seen from these candidates.

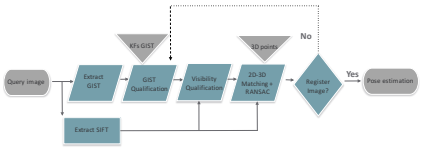


Figure 1: GSSR system overview.

Figure 1 shows an overview of the proposed system, where the GIST distance between the query and all the keyframes is computed. If the distance is below a certain threshold than the keyframe is considered a candidate match. In order to remove outlier keyframes, each candidate keyframe is checked using Eq. 1:

$$F_k = \frac{\sum_{i=1}^{N} P_i(KF_i, KF_k)}{N}, \qquad (1)$$

where N is the total number of candidate keyframes and P_i is the number of 3D points in common between the tested candidate KF_k and the keyframe KF_i at i. If the ratio F_k is high enough the candidate keyframe qualifies for localization. The 3D points of those candidates will be matched to the SIFT features of the image before removing the outliers via RANSAC. Only images with enough inliers will qualify to the pose estimation step.

Table 1: GSSR benchmarked against tree-based [2], PoseNet [1] and ACG Localizer [3].

Dataset	# from Images	# Query Images	Median Error				Average Time (s)			
			Tree-Based	PoseNet	ACG Localizer	GSSR	Tree-Based	PoseNet	ACG Localizer	GSSR
Kings College	1220	343	0.229m, 0.094deg	1.66m, 5.24deg	0.418m, 0.763deg	0.213m, 0.188deg	1.8	0.01	0.74	0.102
Old Hospital	895	182	0.241m, 0.272deg	2.45m, 7.96deg	0.604m, 0.846deg	0.299m, 0.220deg	1.179	0.01	0.43	0.063
Shop Facade	231	103	0.138m,0.220deg	1.89m, 4.29deg	0.548m, 0.575deg	0.160m, 0.217deg	0.802	0.01	0.39	0.044
Average			**0.260m, 0.208deg**	**2.49m, 6.667deg**	**0.672m, 0.773deg**	**0.239m, 0.296deg**	**0.91**	**0.01**	**0.58**	**0.085**

Experimental results on major standard datasets validates the advantages of GSSR. Table 1 shows that GSSR scores the best localization accuracy among all the approaches on the Cambridge 5 Scenes dataset. Note that GSSR has 10 times better accuracy than PoseNet and significantly better than ACG Localizer which is one of the best feature-based IBL systems. GSSR was able to accurately localize a query image in less than 0.1 sec which makes it the fastest feature-based IBL system for large-scale scenes.

[1] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2938–2946, 2015.

[2] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2:331–340, 2009.

[3] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *Computer Vision–ECCV 2012*, pages 752–765. Springer, 2012.

Probabilistic Obstacle Partitioning of Monocular Video for Autonomous Vehicles

Ryan W. Wolcott
rwolcott@umich.edu
Ryan M. Eustice
eustice@umich.edu

Ford Motor Company
Dearborn, MI
Perceptual Robotics Laboratory
University of Michigan
Ann Arbor, MI, USA

Wednesday
13:40-14:40

Localization is a key task for autonomous cars; systems such as the Google driverless car rely on precise and detailed maps for safe operation. Light detection and ranging (LIDAR) sensors are capable of providing rich information—including metric range and point appearance. Robust methods can use this data for vehicle localization by extracting the ground-plane for alignment to a prior map, as in [2].

Vision sensors as part of the localization pipeline can be a great enabler for autonomous platforms. Contrary to LIDAR methods, identifying the ground-plane from a camera image is a more challenging task. In our previous work [3], we considered localizing with a monocular camera by aligning the image to a prior map. As we demonstrated, this can be difficult as the ground-plane can be obscured by obstacles within view of the camera. In this work, we are interested in partitioning an image stream into obstacles and prior map, as shown in Fig. 1, so we can mask out obstacles during registration.

Similar to previous work [1, 4], we use a 1D-Markov random field (MRF) to model a horizontal image partition between obstacles and ground-plane, as in Fig. 1. However, rather than formulating our MRF potentials using image appearance alone (using learned [1] or hand-tuned features [4]), we instead consider the temporal stream of images and inferred parallax. We probabilistically evaluate optical flow against expected optical flow derived from known scene structure and camera egomotion, as in Fig. 2.

Our approach is evaluated on a challenging urban dataset with grayscale imagery, where lighting is non-uniform. We demonstrate our proposed algorithms by looking at errors with respect to hand-labeled groundtruth and present results showing improved image registration when obstacle masks are used.

Acknowledgements: This work was supported by a grant from Ford Motor Company via the Ford-UM Alliance under award N015392.

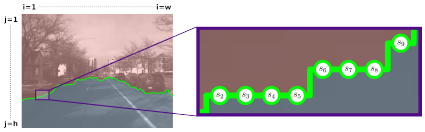


Figure 1: 1D-MRF to partition images into ground-plane and obstacles; each variable node in our MRF partitions an image column. Various unary potentials can be applied to each node; our work emphasizes a potential derived from optical flow.

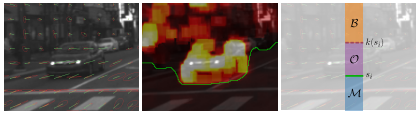


Figure 2: (Left) Optical flow vectors and expected flow vectors with uncertainties. (Middle) Optical flow likelihood and resulting partition. (Right) Optical flow potential implicitly considers segmenting image column into background ($\mathcal{B}$), obstacle ($\mathcal{O}$), and ground-map ($\mathcal{M}$).

[1] Dan Levi, Noa Garnett, and Ethan Fetaya. Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In *Proc. British Mach. Vis. Conf.*, pages 109.1–109.12, Swansea, United Kingdom, September 2015.

[2] Jesse Levinson and Sebastian Thrun. Robust vehicle localization in urban environments using probabilistic maps. In *Proc. IEEE Int. Conf. Robot. and Automation*, pages 4372–4378, Anchorage, AK, May 2010.

[3] Ryan W. Wolcott and Ryan M. Eustice. Visual localization within LIDAR maps for automated urban driving. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, pages 176–183, Chicago, IL, USA, September 2014.

[4] Jian Yao, Srikumar Ramalingam, Yuichi Taguchi, Yohei Miki, and Raquel Urtasun. Estimating drivable collision-free space from monocular video. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 420–427, Waikoloa Beach, HI, USA, January 2015.

Track Facial Points in Unconstrained Videos

Xi Peng[1]
xipeng.cs@rutgers.edu
Qiong Hu[1]
qionghu.cs@rutgers.edu
Junzhou Huang[2]
jzhuang@uta.edu
Dimitris N. Metaxas[1]
dnm@cs.rutgers.edu

[1] Department of Computer Science
Rutgers University
New Jersey, USA

[2] Department of Computer Science
The University of Texas at Arlington
Texas, USA

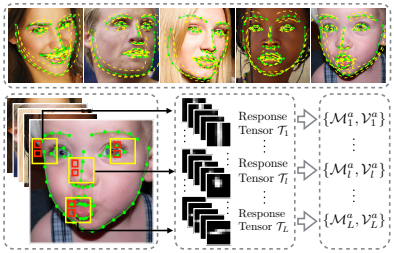


Figure 1: Part-based representation.

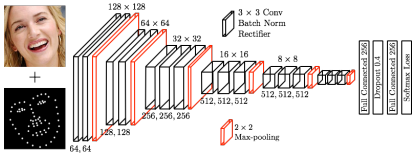


Figure 2: Deep fitting evaluation.

Tracking Facial Points in unconstrained videos is challenging due to the non-rigid deformation that changes over time. In this paper, we propose to exploit incremental learning for person-specific alignment in wild conditions.

Our approach takes advantage of part-based representation, as illustrated in Figure 1 and cascade regression for robust and efficient alignment on each frame. Unlike existing methods that usually rely on models trained offline, we incrementally update the representation subspace and the cascade of regressors in a unified framework to achieve personalized modeling.

Blind model adaptation without correction would inevitably result in model drifting. How to effectively detect misalignment is still a challenging question that is seldom investigated. To address this issue, we propose a deep neural network for robust fitting evaluation to pick out well-aligned faces from misalignment. The architecture of the network is shown is Figure 2. These well-aligned faces are then used to incrementally update the representation subspace and fitting strategy for person-specific modeling on the fly. In summary, our work makes the following contributions:

(1) We propose a novel approach for sequential face alignment. To the best of our knowledge, this is the first time that person-specific modeling is investigated to jointly learn the representation subspace and the fitting parameters in a unified framework.

(2) The proposed part-based representation together with the cascade regression guarantees robust alignment in unconstrained conditions. More importantly, they are critical to efficiently construct personalized models for real-time or large-scale applications.

(3) We propose to leverage deep neural networks for efficient and robust fitting evaluation. It significantly alleviates the drifting issue which would severely deteriorate learned models and inevitably lead to failure.

To validate our approach, we provide a detailed experimental analysis of each component of our approach, as well as performance comparisons with existing approaches. Four image datasets (MultiPIE, LFPW, Helen, and AFLW), four video datasets (FGNET, ASLV, 300-VW, and YtbVW), and four state of the arts (RLMS, DRMF, IFA, and ESR) are employed to conduct the experiments. The results demonstrate that the proposed incremental learning can significantly improve the fitting accuracy with an affordable computational cost, especially in unconstrained videos with extensive variations.

Wednesday
13:40-14:40

Author Index

Notes