Foreword

We are delighted to welcome you to the 24th British Machine Vision Conference, and in particular once again to the University of Bristol, where BMVC is held for the second time in its 24-year history. The conference is at the peak of its strength and considered as one of the top events in the computer vision community's calendar. This year it has received the most number of papers it has ever received – 439 submissions. Although "British" is in the title, 86% of the accepted submissions are from overseas, reflecting the irrefragable importance and status of the conference internationally. Using the first author's affiliation, 14% of the accepted papers are from a UK-based institute, 44% from Europe (excluding UK), 21% from Asia, 19% from North America and 2% from Australia.

To facilitate the process of peer reviewing such a considerable number of papers, we depended on a team of 225 referees, each of whom reviewed at least 4 papers, and 2 Area Chairs (ACs) per paper, with each AC responsible for at least 15 papers. In total, 49 ACs took part in the reviewing procedure. Following a discussion process, where necessary, involving the reviewers and ACs of each paper, consensus was reached and final decisions were made. The results are the papers you see before you comprising 29 podium presentations and 102 poster presentations. This represents a 7% podium acceptance rate and 30% overall acceptance rate. We would like to thank all the authors, referees, and ACs for their hours of work and contributions to this conference.

This was not the end of the road for the ACs as many of them were also involved in selecting the best papers for our various prizes. Four prizes will be awarded, three of which are traditional prizes for the best Science paper, best Industrial paper, and the best Poster. This year we are also awarding a prize in memory of our nonpareil colleague and computer vision pioneer, Professor Maria Petrou (1953-2012). We are grateful to Qualcomm, Dyson, Microsoft, HP, the IET and all other sponsors for their financial support of the conference.

We are thrilled to have two distinguished scholars as invited speakers, Prof. Andrew Zisserman from the University of Oxford and Prof. Frank Dellaert from Georgia Institute of Technology, and two esteemed tutorial presenters, Dr. Andreas Krause from ETH Zürich and Dr. Adam Coates from Stanford University. In the Friday PhD Student Workshop, we are pleased to have two highly prominent invited speakers, Dr. Gabriel Brostow from University College London and Dr. Edward Rosten from Computer Vision Consulting. We thank all of our presenters in turn for their time and endeavour.

BMVC2013 has been organised by members of the Visual Information Laboratory in the Merchant Venturers School of Engineering at the University of Bristol. We would like to sincerely thank Antonina Timofejeva for her relentless and efficient handling of the conference administration and secretariat duties. We are also indebted to a number of others who have contributed enormously to the organisation of the conference, Jose Martinez-Carranza for organising the workshop and all our valued helpers: Pui Anantrasirichai, Geoffory Daniels, Austin Gregg-Smith, Anthony Glynn, Osian Haines, David Hanwell, Adeline Paiement, Toby Perrett, Roz Sandwell, Laszlo Talas and Richard Vigars. We also received invaluable support from the organisers of the previous BMVC at Surrey University, namely, John Collomosse, Krystian Mikolajczyk, Richard Bowden, and Helen Cooper, who we thank gratefully.

BMVC and Bristol have much to offer in their own imitable ways, and we hope that you find the conference stimulating and informative, and the City of Bristol with its history and sights enjoyable and rewarding.

Tilo Burghardt, Dima Damen, Walterio Mayol-Cuevas, Majid Mirmehdi

Bristol, September 2013

PROGRAMME: Monday 9th September 2013

BMVC2013 Tutorials

12:00 PM	**Registration Desk Open**	
1:30 PM	**TUTORIAL by Adam Coates (Stanford University, USA)** 'Deep Learning for Machine Vision'	**11**
3:00 PM	**Coffee Break**	
3:30 PM	**TUTORIAL by Andreas Krause (ETH Zürich, Switzerland)** 'Submodularity in Machine Learning and Computer Vision'	**12**
5:00 PM	**End of Sessions**	
7:30 PM	***British Machine Vision Association Company Meeting*** *(BMVA internal)*	

PROGRAMME: Tuesday 10th September 2013

Conference Opening

8:30 AM	**Registration Desk Open**	
9:00 AM	**OPENING REMARKS**	**1**

Session 01: Opening Keynote + Statistics and Learning for Computer Vision I

Tuesday, 9:10 AM

Chair: Matti Pietkainien, University Oulu

9:10 AM	**OPENING KEYNOTE by Andrew Zisserman *(University of Oxford, UK)*** 'Towards On-the-fly Large Scale Video Search'	**13**
10:10 AM	**Metric Regression Forests for Human Pose Estimation** Gerard Pons-Moll , Jonathan Taylor , Jamie Shotton , Aaron Hertzmann , Andrew Fitzgibbon	**14**
10:30 AM	**Label Embedding for Text Recognition** Jose Rodriguez , Florent Perronnin	**15**
10:50 AM	**Coffee Break**	

Session 02: Statistics and Learning for Computer Vision II

Tuesday, 11:20 AM

Chair: Tom Drummond, University of Monash

11:20 AM	**Change Detection in Dynamic Scenes using Local Adaptive Transform** Hakan Haberdar , Shishir Shah	**16**
11:40 AM	**Learning Beautiful (and Ugly) Attributes** Luca Marchesotti , Florent Perronnin	**17**
12:00 PM	**Fisher Vector Faces in the Wild** Karen Simonyan , Omkar Parkhi , Andrea Vedaldi , Andrew Zisserman	**18**
12:20 PM	**Perception Preserving Projections** Saining Xie , Jiashi Feng , Shuicheng Yan , Hongtao Lu	**19**
12:40 PM	**A Novel Approach for Efficient SVM Classification with Histogram Intersection Kernel** Gaurav Sharma , Frederic Jurie	**20**
1:00 PM	**Lunch**	

Session 03: Posters — Tuesday, 2:00 PM

Board #01	**Inferring Ongoing Human Activities Based on Recurrent Self-Organizing Map Trajectory** Qianru Sun , Hong Liu	21
Board #02	**Match-time covariance for descriptors** Eric Christiansen , Vincent Rabaud , Andrew Ziegler , David Kriegman , Serge Belongie	22
Board #03	**Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces** Pablo Alcantarilla , Jesus Nuevo , Adrien Bartoli	23
Board #04	**From Bikers to Surfers: Visual Recognition of Urban Tribes** Iljung Kwak , Ana Murillo , Peter Belhumeur , David Kriegman , Serge Belongie	24
Board #05	**Ultra-wide Baseline Aerial Imagery Matching in Urban Environments** Hani Altwaijry , Serge Belongie	25
Board #06	**FRIF: Fast Robust Invariant Feature** Zhenhua Wang , Bin Fan , Fuchao Wu	26
Board #07	**Spatial Co-Training for Semi-Supervised Image Classification** Yi Hong	27
Board #08	**Distribution Fields with Adaptive Kernels for Large Displacement Image Alignment** Benjamin Mears , Laura Sevilla-Lara , Erik Learned-Miller	28
Board #09	**Fractal Approximate Nearest Neighbour Search in Log-Log Time** Martin Stommel , Stefan Edelkamp , Thiemo Wiedemeyer , Michael Beetz	29
Board #10	**Flexible and User-Centric Camera Calibration using Planar Fiducial Markers** Shreyansh Daftry , Michael Maurer , Andreas Wendel , Horst Bischof	30
Board #11	**Large-scale Learning of Sign Language by Watching TV (Using Co-occurrences)** Tomas Pfister , James Charles , Andrew Zisserman	31
Board #12	**Detection of Abnormal Fish Trajectories Using a Clustering Based Hierarchical Classifier** Cigdem Beyan , Robert Fisher	32
Board #13	**Early Recurrence Improves Edge Detection** Xun Shi , Bo Wang , John Tsotsos	33
Board #14	**Modelling Visual Objects Invariant to Depictive Style** Qi Wu , Peter Hall	34
Board #15	**Free your Camera: 3D Indoor Scene Understanding from Arbitrary Camera Motion** Axel Furlan , Stephen D. Miller , Domenico G. Sorrenti , Li Fei-Fei , Silvio Savarese	35
Board #16	**Exploring SVM for Image Annotation in Presence of Confusing Labels** Yashaswi Verma , CV Jawahar	36
Board #17	**Removing Mistracking of Multibody Motion Video Database Hopkins155** Yasuyuki Sugaya , Yuichi Matsushita , Kenichi Kanatani	37
Board #18	**Enforcing Monotonous Shape Growth or Shrinkage in Video Segmentation** Yuliya Tarabalka , Guillaume Charpiat , Ludovic Brucker , Bjoern Menze	38
Board #19	**Deformable Part Models with Individual Part Scaling** Charles Dubout , François Fleuret	39
Board #20	**Focusing Attention on Visual Features that Matter** Grace Tsai , Benjamin Kuipers	40
Board #21	**Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition** Tomas McCandless , Kristen Grauman	41
Board #22	**Learning to approximate global shape priors for figure-ground segmentation** Daniel Kuettel , Vittorio Ferrari	42
Board #23	**What is a good evaluation measure for semantic segmentation?** Gabriela Csurka , Diane Larlus , Florent Perronnin	43
Board #24	**Harmonic Variance: A Novel Measure for In-focus Segmentation** Feng Li , Fatih Porikli	44
Board #25	**Robust Image Matching with Line Context** Wei Guan , Suya You	45
Board #26	**Sketch Recognition by Ensemble Matching of Structured Features** Yi Li , Yi-Zhe Song , Shaogang Gong	46
Board #27	**Learnable Stroke Models for Example-based Portrait Painting** Tinghuai Wang , John Collomosse , Andrew Hunter , Darryl Greig	47
Board #28	**Bayesian region selection for adaptive dictionary-based Super-Resolution** Eduardo Pérez-Pellitero , Jordi Salvador , Javier Ruiz-Hidalgo , Bodo Rosenhahn	48
Board #29	**Replicator Graph Clustering** Michael Donoser	49

Session 03: Posters (continued) — Tuesday, 2:00 PM

Board #30	**Of Gods and Goats: Weakly Supervised Learning of Figurative Art** Elliot Crowley , Andrew Zisserman	50
Board #31	**Multi-scale Joint Encoding of Local Binary Patterns for Texture and Material Classification** Xianbiao Qi , Yu Qiao , Chun-Guang Li , Jun Guo	51
Board #32	**Spectral Imaging Using Basis Lights** Antony Lam , Art Subpa-Asa , Imari Sato , Takahiro Okabe , Yoichi Sato	52
Board #33	**Generalised Perspective Shape from Shading with Oren-Nayar Reflectance** Yong Chul Ju , Silvia Tozza , Michael Breuss , Andres Bruhn , Andreas Kleefeld	53
Board #34	**High Frequency 3D Reconstruction from Unsynchronized Multiple Cameras** Yumi Kakumu , Fumihiko Sakaue , Jun Sato , Kazuhisa Ishimaru , Masayuki Imanishi	54
Board #35	**Properties of Datasets Predict the Performance of Classifiers** Omid Aghazadeh , Stefan Carlsson	55
3:00 PM	**Coffee Break**	

Session 04: Human Pose & Tracking — Tuesday, 3:30 PM

Chair: Graeme Jones, University of Kingston

3:30 PM	**Multi-view Pictorial Structures for 3D Human Pose Estimation** Sikandar Amin , Mykhaylo Andriluka , Marcus Rohrbach , Bernt Schiele	56
3:50 PM	**Transitive Re-identification** Yulia Brand , Tamar Avraham , Michael Lindenbaum	57
4:10 PM	**Domain Adaptation for Upper Body Pose Tracking in Signed TV Broadcasts** James Charles , Tomas Pfister , Derek Magee , David Hogg , Andrew Zisserman	58
4:30 PM	**Multi-view Body Part Recognition with Random Forests** Vahid Kazemi , Magnus Burenius , Hossein Azizpour , Josephine Sullivan	59
4:50 PM	**End of Sessions**	

4:50 PM ***British Machine Vision Association AGM***

BMVC 2013 Reception *(Great Hall of Wills Memorial Building)* — Tuesday, 7:00 PM

Featuring a presentation by Aardman's Tom Barnes (Head of Technical Team) on the latest visual animation projects made in Bristol.



PROGRAMME: Wednesday 11th September 2013

8:30 AM	**Registration Desk Open**	

Session 05: Keynote + Motion & Flow

Wednesday, 9:00 AM

Chair: Roy Davies, Royal Holloway College

9:00 AM	**KEYNOTE by Frank Dellaert *(Georgia Tech, USA)*** 'Factor Graphs for Fast and Scalable and Computer Vision'	**60**
10:00 AM	**The Complete Rank Transform: A Tool for Accurate and Morphologically Invariant Matching of Structures** Oliver Demetz , David Hafner , Joachim Weickert	**61**
10:20 AM	**Efficient Dense 3D Rigid-Body Motion Segmentation in RGB-D Video** Jörg Stückler , Sven Behnke	**62**
10:40 AM	**Coffee Break**	

Session 06: Video Analysis

Wednesday, 11:10 AM

Chair: Majid Mirmehdi, University of Bristol

11:10 AM	**Enhancing Action Recognition by Cross-Domain Dictionary Learning** Fan Zhu , Ling Shao	**63**
11:30 AM	**Unsupervised Object Discovery and Segmentation in Videos** Samuel Schulter , Christian Leistner , Peter Roth , Horst Bischof	**64**
11:50 AM	**Evaluating Superpixels in Video: Metrics Beyond Figure-Ground Segmentation** Peer Neubert , Peter Protzel	**65**
12:10 PM	**Solving Person Re-identification in Non-overlapping Camera using Efficient Gibbs Sampling** Vijay John , Gwenn Englebienne , Ben Krose	**66**
12:30 PM	**Spacetime Forests with Complementary Features for Dynamic Scene Recognition** Christoph Feichtenhofer , Axel Pinz , Richard Wildes	**67**
12:50 PM	**Lunch**	

Session 07: Posters and Demos

Wednesday, 2:00 PM

Board #01	**Hyperspectral Face Recognition using 3D-DCT and Partial Least Squares** Muhammad Uzair , Arif Mahmood , Ajmal Mian	**68**
Board #02	**Merging Pose Estimates Across Space and Time** Xavier Burgos-Artizzu , David Hall , Pietro Perona , Piotr Dollar	**69**
Board #03	**Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition** Xiaojiang Peng , Yu Qiao , Qiang Peng , Xianbiao Qi	**70**
Board #04	**Can Your Eyes Tell Me How You Think? A Gaze Directed Estimation of the Mental Activity** Laura Florea , Corneliu Florea , Ruxandra Vranceanu , Constantin Vertan	**71**
Board #05	**Incremental Solid Modeling from Sparse and Omnidirectional Structure-from-Motion Data** Vadim Litvinov , Maxime Lhuillier	**72**
Board #06	**Egomotion Estimation Using Binocular Spatiotemporal Oriented Energy** Hao Zhong , Richard Wildes	**73**
Board #07	**Online Motion Agreement Tracking** Zheng Wu , Jianming Zhang , Margrit Betke	**74**
Board #08	**A Topic Model Approach to Representing and Classifying Football Plays** Jagannadan Varadarajan , Indriyati Atmosukarto , Shaunak Ahuja , Bernard Ghanem , Narendra Ahuja	**75**

Session 07: Posters and Demos (continued)

Board #09	**Sketch Retrieval via Dense Stroke Features** Chao Ma , Xiaokang Yang , Chongyang Zhang , Xiang Ruan , Ming-Hsuan Yang	**76**
Board #10	**PedCut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues** Fabian Flohr , Dariu Gavrila	**77**
Board #11	**Fast Approximation of Distance Between Elastic Curves using Kernels** Hedi Tabia , David Picard , Hamid Laga , Philippe-Henri Gosselin	**78**
Board #12	**Line extraction in uncalibrated central images with revolution symmetry** Jesus Bermudez-Cameo , Gonzalo Lopez-Nicolas , J.J. Guerrero	**79**
Board #13	**Tree-like Shapes Distance Using the Elastic Shape Analysis Framework** Alejandro Mottini , Xavier Descombes , Florence Besse	**80**
Board #14	**Image Blur Classification and Parameter Identification Using Two-stage Deep Belief Networks** Ruomei Yan , Ling Shao	**81**
Board #15	**Discriminative tree-based feature mapping** Miroslav Kobetski , Josephine Sullivan	**82**
Board #16	**Unsupervised (parameter) learning for MRFs on bipartite graphs** Boris Flach , Tomas Sixta	**83**
Board #17	**Generic Object Crowd Tracking by Multi-Task Learning** Wenhan Luo , Tae-Kyun Kim	**84**
Board #18	**Discriminative Generative Contour Detection** Chao Zhang , Xiong Li , Xiang Ruan , Yuming Zhao , Ming-Hsuan Yang	**85**
Board #19	**Accurate & Computationally-inexpensive Recovery of Ego-Motion using Optical Flow & Range Flow with Extended Temporal Support** Graeme Jones	**86**
Board #20	**Emotion recognition from facial images with arbitrary views** Xiaohua Huang , Guoying Zhao , Matti Pietikainen	**87**
Board #21	**RPM: Random Points Matching for Pair wise Face-Similarity** M. Saquib Sarfraz , Muhammad Adnan Siddique , Rainer Stiefelhagen	**88**
Board #22	**Segmentation Driven Low-rank Matrix Recovery for Saliency Detection** Wenbin Zou , Kidiyo Kpalma , Zhi Liu , joseph Ronsin	**89**
Board #23	**Action Attribute Detection from Sports Videos with Contextual Constraints** Xiaodong Yu , Ching Lik Teo , Yezhou Yang , Cornelia Fermuller , Yiannis Aloimonos	**90**
Board #24	**Decomposed Learning for Joint Object Segmentation and Categorization** Yi-Hsuan Tsai , Jimei Yang , Ming-Hsuan Yang	**91**
Board #25	**Action Chart: A Representation for Efficient Recognition of Complex Activity** Hyung Jin Chang , Jiyun Kim , Jungchan Cho , Songhwai Oh , Kwang Yi , Jin Young Choi	**92**
Board #26	**Simultaneous Pose, Focal Length and 2D-to-3D Correspondences from Noisy Observations** Adrian Penate-Sanchez , Eduard Serradell , Francesc Moreno-Noguer , Juan Andrade-Cetto	**93**
Board #27	**Discriminative Tensor Sparse Coding for Image Classification** Yangmuzi Zhang , Zhuolin Jiang , Larry Davis	**94**
Board #28	**Hierarchical Scene Annotation** Michael Maire , Stella Yu , Pietro Perona	**95**
Board #29	**Blockwise Linear Regression for Face Alignment** Enrique Sánchez-Lozano , Enrique Argones-Rúa , Jose Luis Alba-Castro	**96**
Board #30	**Fast and Robust L1-averaging-based Pose Estimation for Driving Scenarios** German Ros , Angel Sappa , Daniel Ponsa , Antonio López , Julio Guerrero	**97**
Board #31	**Learning Speed Invariant Gait Template via Thin Plate Spline Kernel Manifold Fitting** Sheng Huang , Ahmed Elgammal , Dan Yang	**98**
Board #32	**Parsing Clothes in Unrestricted Images** Nataraj Jammalamadaka , Ayush Minocha , Digvijay Singh , CV Jawahar	**99**
Board #33	**Image Fusion using Gaussian Mixture Models** Stacey Levine , Katie Heaps , Joshua Koslosky , Glenn Sidle	**100**
3:00 PM	**Coffee Break**	

Session 08: Stereo and Geometry I — Wednesday, 3:30 PM

Chair: Andrew Fitzgibbon, Microsoft & BMVA Chairman

3:30 PM	**Real Projective Plane Mapping for Detection of Orthogonal Vanishing Points** Markéta Dubská , Adam Herout	101
3:50 PM	**Efficient Shape Matching using Vector Extrapolation** Emanuele Rodola , Tatsuya Harada , Yasuo Kuniyoshi , Daniel Cremers	102
4:10 PM	**Incremental Line-based 3D Reconstruction using Geometric Constraints** Manuel Hofer , Andreas Wendel , Horst Bischof	103
4:30 PM	**Spline Fusion: A continuous-time representation for visual-inertial fusion with application to rolling shutter cameras** Steven Lovegrove , Alonso Patron-Perez , Gabe Sibley	104
4:50 PM	**End of Sessions**	

BMVC 2013 Banquet *(Bristol Council Hall)* — **Wednesday, 7:00 PM**

PROGRAMME: Thursday 12th September 2013

8:30 AM	**Registration Desk Open**

Session 09: Stereo and Geometry II — Thursday, 9:20 AM

Chair: Stephen Pollard, HP Labs

9:20 AM	**Incremental Surface Extraction from Sparse Structure-from-Motion Point Clouds** Christof Hoppe , Manfred Klopschitz , Michael Donoser , Horst Bischof	105
9:40 AM	**Dense, Auto-Calibrating Visual Odometry from a Downward-Looking Camera** Jacek Zienkiewicz , Robert Lukierski , Andrew Davison	106
10:00 AM	**A Novel Trilateral Filter based Adaptive Support Weight Method for Stereo Matching** Dongming Chen , Mohsen Ardabilian , Liming Chen	107

Session 10: Posters — Thursday, 10:20 PM

Board #01	**Exploring Cross-Channel Texture Correlation for Color Texture Classification** Xianbiao Qi , Yu Qiao , Chun-Guang Li , Jun Guo	108
Board #02	**Depth really Matters: Improving Visual Salient Region Detection with Depth** Karthik Desingh , Madhava Krishna K , Deepu Rajan , CV Jawahar	109
Board #03	**Oriented pooling for dense and non-dense rotation-invariant features** Wan-Lei Zhao , Guillaume Gravier , Herve Jegou	110
Board #04	**Efficient Discriminative Convolution Using Fisher Weight Map** Hideki Nakayama	111
Board #05	**Single-Image Super-Resolution via Adaptive Joint Kernel Regression** Chen Huang , Xiaoqing Ding , Chi Fang	112
Board #06	**Anatomical Structure Sketcher for Cephalograms by Bimodal Deep Learning** Yuru Pei , Bin Liu , Hongbin Zha , Bing Han , Tianmin Xu	113
Board #07	**Reconstructing faces from their signatures using RBF regression** Alexis Mignon , Frederic Jurie	114

Session 10: Posters (continued)

Board #08	**Isowarp and Conwarp: Warps that Exactly Comply with Weak-Perspective Projection of Deforming Objects** Daniel Pizarro , Adrien Bartoli , Toby Collins	115
Board #09	**Point Light Source Estimation based on Scenes Recorded by a RGB-D camera** Bas Boom , Sergio Orts-Escolano , Xi Ning , Steven McDonagh , Peter Sandilands , Robert Fisher	116
Board #10	**Closed form solution for radial distortion estimation from a single vanishing point** Horst Wildenauer , Branislav Micusik	117
Board #11	**Robust Scene Stitching in Large Scale Mobile Mapping** Filip Schouwenaars , Radu Timofte , Luc Van Gool	118
Board #12	**Local Zernike Moment Representation for Facial Affect Recognition** Evangelos Sariyanidi , Hatice Gunes , Muhittin Gökmen , Andrea Cavallaro	119
Board #13	**An Adaptive Subdivision Scheme for Quadratic Programming in Multi-Label Image Segmentation** Marko Rak , Tim König , Klaus-Dietz Tönnies	120
Board #14	**Visual Tracking via Subspace Motion Model** Jun Wang , Fan Zhong , Guofeng Wang , qunsheng Peng , Xueying Qin	121
Board #15	**Place recognition from disparate views** Rob Frampton , Andrew Calway	122
Board #16	**An In Depth View of Saliency** Arridhana Ciptadi , Tucker Hermans	123
Board #17	**Insulin Granule Segmentation in 3-D TEM Beta Cell Tomograms** David Nam , Judith Mantell , David Bull , Paul Verkade , Alin Achim	124
Board #18	**Plane-based Odometry using an RGB-D Camera** Carolina Raposo , Miguel Lourenco , Michel Antunes , Joao Barreto	125
Board #19	**Reflective Symmetry Detection by Rectifying Randomized Correspondences** Zhongwei Tang , Mariano Tepper , Guillermo Sapiro	126
Board #20	**An Iterative 5-pt Algorithm for Fast and Robust Essential Matrix Estimation** Vincent Lui , Tom Drummond	127
Board #21	**Reduced Dimensionality Extended Kalman Filter for SLAM** Dinesh Gamage , Tom Drummond	128
Board #22	**Learning Smooth Pooling Regions for Visual Recognition** Mateusz Malinowski , Mario Fritz	129
Board #24	**Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges** Philipp Werner , Ayoub Al-Hamadi , Robert Niese , Steffen Walter , Sascha Gruss , Harald Traue	130
Board #25	**Overlapping camera clustering through dominant sets for scalable 3D reconstruction** Massimo Mauro , Hayko Riemenschneider , Luc Van Gool , Riccardo Leonardi	131
Board #26	**Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues** Isarun Chamveha , Yusuke Sugano , Yoichi Sato , Akihiro Sugimoto	132
Board #27	**RLBP: Robust Local Binary Pattern** Jie Chen , Vili Kellokumpu , Guoying Zhao , Matti Pietikainen	133
Board #28	**3D Deformable Shape Reconstruction with Diffusion Maps** Lili Tao , Bogdan Matuszewski	134
Board #29	**Towards a minimal solution for the relative pose between axial cameras** Francisco Vasconcelos , João Barreto	135
Board #30	**Combining Features and Intensity for Wide-Baseline Non-Rigid Surface Registration** Jim Braux-Zin , Romain Dupont , Adrien Bartoli	136
Board #31	**Single Image Shadow Detection Using Multiple Cues in a Supermodular MRF** Tomas F Yago Vicente , Chen-Ping Yu , Dimitris Samaras	137
Board #32	**Recognizing Humans in Motion: Trajectory-based Aerial Video Analysis** Yumi Iwashita , Michael Ryoo , Thomas Fuchs , Curtis Padgett	138
Board #33	**Combining Local and Global Cues for Closed Contour Extraction** Vida Movahedi , James Elder	139
Board #34	**Lagrangian Strain Tensor Computation with Higher Order Variational Models** Alexander Hewer , Joachim Weickert , Henning Seibert , Tobias Scheffer , Stefan Diebels	140
Board #35	**Object Recognition in Multi-View Dual Energy X-ray Images** Muhammet Bastan , Wonmin Byeon , Thomas Breuel	141
11:20 AM	**Coffee Break**	

Session 11: Segmentation & Features — **Thursday, 10:50 AM**
Chair: Paul Siebert, University of Glasgow

11:50 AM	**Relatively-Paired Space Analysis** Zhanghui Kuang , Kenneth Wong	**142**
12:10 PM	**Sparse-Coded Features for Image Retrieval** Tiezheng Ge , Qifa Ke , Jian Sun	**143**
12:30 PM	**Combination of paths for interactive segmentation** Julien Mille , Sébastien Bougleux , Laurent Cohen	**144**
12:50 PM	**Collaboratively Regularized Nearest Points for Set Based Recognition** Yang Wu , Michihiko Minoh , Masayuki Mukunoki	**145**
1:10 PM	**CONCLUDING REMARKS**	
1:20 PM	**End of Main Conference**	

PROGRAMME: Friday 13th September 2013

Workshop Session 01 *(Workshop Contributions)* — **Friday, 9:00 AM**

9:00 AM	**WORKSHOP KEYNOTE by Gabriel Browstow *(UCL, UK)*** 'If Only We Had Tracked Something Like This Before'	**146**
	Multi-Scale Azimuthal Projection Distance Image of Normal Maps for 3D Facial Skin Texture Analysis Alassane Seck , Hannah Dee , Bernard Tiddman	
	Synthesising unseen image conditions to enhance classification accuracy for sparse datasets: applied to chimpanzee face recognition Roz Sandwell , Alexander Loos , Tilo Burghardt	
10:40 AM	**Coffee Break**	

Workshop Session 02 *(Talks on BMVC Posters)* — **Friday, 11:20 AM**

11:20 AM	**Sketch Recognition by Ensemble Matching of Structured Features** Yi Li , Yi-Zhe Song , Shaogang Gong	**46**
	Modelling Visual Objects Invariant to Depictive Style Qi Wu , Peter Hall	**34**
	Place recognition from disparate views Rob Frampton, Andrew Calway	**122**
	Generic Object Crowd Tracking by Multi-Task Learning Wenhan Luo , Tae-Kyun Kim	**84**
1:00 PM	**Lunch**	

Workshop Session 03 *(Workshop Contributions)* — **Friday, 2:00 PM**

2:00 PM	**WORKSHOP KEYNOTE by Edward Rosten *(Cambridge, UK)*** 'If Only We Had Tracked Something Like This Before'	**147**
	Improving photometric stereo through per-pixel light vector calculation Jahanzeb Ahmad , Jiuai Sun , Lyndon Smith , Melvyn Smith	
	A Multiple Motion Model Tracker Handling Occlusion and Rapid Motion Variation Muhammad Haris Khan , Michel Valstar , Tony Pridmore	
3:40 PM	**CONCLUDING REMARKS**	
3:50 PM	**End of Student Workshop**	

ABSTRACTS

TUTORIAL by Adam Coates *(Stanford University, USA)*

Deep Learning for Machine Vision

Abstract. Machine learning algorithms have freed practitioners from many error-prone, hand-engineered components for making decisions in common machine vision tasks such as object recognition. A major source of difficulty, however, is that such learning systems still rely on many hand-built components like sophisticated feature extractors that attempt to identify higher-level patterns in images that typical learning algorithms cannot discover on their own. "Deep learning" and "representation learning" algorithms aim to remove this hurdle by learning higher-level representations automatically from data and have led to recent successes in vision, speech, and language tasks. This tutorial will introduce the basic components of deep learning algorithms and practical techniques for debugging and applying these methods to machine vision problems. The first part of the tutorial will cover neural network models and basic training approaches including error back-propagation and numerical optimization methods, with image classification as a motivating application. The second part will cover additional (sometimes domain-specific) techniques to improve the performance of these algorithms and apply them to other vision tasks including detection and image segmentation. With these tools, audience members will understand how deep learning algorithms work and how they are used in practical applications with sufficient knowledge to complete a hands-on tutorial available on the web. We will conclude with a brief high-level overview of other important topics and results in deep learning research.

Biography. Dr Adam Coates received his PhD in Computer Science from Stanford University in 2012. He is currently a post-doctoral researcher at Stanford and a Visiting Scholar at Indiana University, Bloomington. His research focuses on scaling up machine learning and representation learning algorithms to enable machines to acquire knowledge from unsupervised experience. His interests cover related key topics in computer vision, reinforcement learning, and robotics. He has received Best Student Paper awards from ICML and ICDAR; but his favorite endorsement is still the Instrument Rating on his pilot's license.

TUTORIAL by Andreas Krause *(ETH Zürich, Switzerland)*

Submodularity in Machine Learning and Vision

Abstract. Numerous problems in machine learning and vision are inherently discrete. More often than not, these lead to challenging optimization problems. While convexity is an important property when solving continuous optimization problems, submodularity, often viewed as a discrete analog of convexity, is key to solving many discrete problems. Its characterizing property, diminishing marginal returns, appears naturally in a multitude of settings. While submodularity has long been recognized in combinatorial optimization and game theory, it has seen a recent surge of interest in theoretical computer science, machine learning and computer vision. This tutorial will introduce the concept of submodularity and its basic properties, and outline recent research directions -- such as new approaches towards large-scale optimization and sequential decision making tasks. We will discuss recent applications to challenging machine learning and vision problems such as high-order graphical model inference, structured sparse modeling, multiple object detection, active sensing etc. The tutorial will not assume any specific prior knowledge on the subject.

Biography. Dr Andreas Krause is an Assistant Professor of Computer Science at ETH Zurich, where he leads the Learning & Adaptive Systems Group. Before that he was Assistant Professor of Computer Science at Caltech (2009-2012). He received his Ph.D. and M.Sc. in Computer Science from Carnegie Mellon University (2008) and his Diplom in Computer Science and Mathematics from the Technical University of Munich, Germany (2004). He is a Microsoft Research Faculty Fellow and a Kavli Frontiers Fellow of the US National Academy of Sciences. He received an ERC Starting Investigator grant, the Deutscher Mustererkennungspreis, an NSF CAREER award, the Okawa Foundation Research Grant recognizing top young researchers in telecommunications as well as the ETH Golden Owl teaching award. His research in learning and adaptive systems that actively acquire information, reason and make decisions in large, distributed and uncertain domains received awards at premier conferences (AAAI, KDD, IPSN, ICML, UAI) and journals (JAIR, JWRPM).

OPENING KEYNOTE by Andrew Zisserman *(Oxford, UK)*

Towards On-the-fly Large Scale Video Search

Abstract. We would like to be able to find anything in an image or video dataset. The talk will describe our progress on visual search for finding people, specific objects and categories in large scale video datasets. The novelty is that the item of interest can be specified at run time by a text query, and a discriminative classifier for that item is then learnt on-the-fly using images downloaded from Google Image search. We will compare state of the art encoding methods for the problem, and discuss the choices in achieving the best trade-off between three important performance measures for a realtime system of this kind, namely: (i) accuracy, (ii) memory footprint, and (iii) speed. We will also describe steps to achieving `total recall'. There will be demonstrations on a large scale video dataset of BBC broadcasts. This is joint work with Relja Arandjelovic, Ken Chatfield and Omkar Parkhi.

Biography. Professor Andrew Zisserman leads the Visual Geometry Group at the University of Oxford, UK. Andrew's research interests include visual recognition, image retrieval, multi-view geometry, and other aspects of computer vision. Some of Andrew's papers are amongst the most highly cited works in the field. His contributions received multiple awards at the top computer vision conferences including three Marr prizes at the International Conferences on Computer Vision. He has published several books including "Visual Reconstruction" (with Andrew Blake) and "Multiple View Geometry in Computer Vision" (with Richard Hartley). He is a fellow of the Royal Society.

Metric Regression Forests for Human Pose Estimation

Gerard Pons-Moll[12]
http://www.tnt.uni-hannover.de/~pons/
Jonathan Taylor[13]
jtaylor@cs.toronto.edu
Jamie Shotton[1]
jamiesho@microsoft.com
Aaron Hertzmann[14]
hertzman@adobe.com
Andrew Fitzgibbon[1]
awf@microsoft.com

[1] Microsoft Research
Cambridge, UK
[2] TNT
Leibniz University of Hannover, Germany
[3] University of Toronto
[4] Adobe Research

Traditionally, human pose estimation algorithms could be classified into generative [2] and discriminative [4] approaches. Generative approaches model the likelihood of the observations given a pose estimate, however, they are susceptible to local minima and thus require good initial pose estimates. Discriminative approaches learn a direct mapping from image features to pose space from training data, however, they struggle to generalize to unseen poses. Building on previous work [3], Taylor *et al.* [5] bypass some of these limitations using a hybrid-approach that discriminatively predicts, for each pixel in a depth image, a corresponding point on the surface of a humanoid mesh model. This mesh model is then robustly fit to the resulting set of correspondences using local optimization. Surprisingly though, these correspondences are actually inferred using a random forest whose structure was trained using a *classification* objective that arbitrarily equates target model points belonging to the same predefined body part [3].

In this paper, we address Taylor *et al.*'s use of this proxy classification objective by proposing Metric Space Information Gain (MSIG), a replacement objective function for training a random forest to directly minimize the uncertainty over the target model points, naturally encoding the correlation between these points as a function of the geodesic distance. To this end, we view the surface of the model $\mathbb{U}$ as a metric space $(\mathbb{U}, d_{\mathbb{U}})$ defined by the geodesic distance metric $d_{\mathbb{U}}$ (see first panel of Figure 1). The natural objective function to minimize the uncertainty in the resulting true distributions that result from a split function s in such a space, is the information gain $I(s)$ [1]. This is generally approximated using an empirical distribution $Q = \{\mathbf{u}_i\} \subseteq \mathbb{U}$ drawn from the true unsplit distribution p_U as

$$I(s) \approx \hat{I}(s;Q) = \hat{H}(Q) - \sum_{i \in \{L,R\}} \frac{|Q_i|}{|Q|} \hat{H}(Q_i), \tag{1}$$

where Q_L and Q_R are the two resulting empirical distributions from applying s, and $\hat{H}(Q)$ is some approximation to the differential entropy

$$H(U) = \mathbb{E}_{p_U(\mathbf{u})}[-\log p_U(\mathbf{u})] = -\int_{\mathbb{U}} p_U(\mathbf{u}) \log p_U(\mathbf{u}) d\mathbf{u}. \tag{2}$$

of the distribution p_U on $\mathbb{U}$ from which Q arose. We provide this approximation by first estimating the true continuous distribution $p_U(\mathbf{u})$ using Kernel Density Estimation (KDE). Let $N = |Q|$ be the number of datapoints in the sample set. The approximated density $f_U(\mathbf{u})$ is then given by

$$p_U(\mathbf{u}) \simeq f_U(\mathbf{u}) = \frac{1}{N} \sum_{\mathbf{u}_j \in Q} k(\mathbf{u}; \mathbf{u}_j), \tag{3}$$

where $k(\mathbf{u};\mathbf{u}_j)$ is a kernel function centered at $\mathbf{u}_j$. Unfortunately, the obvious way to estimate (2) using Monte Carlo is quadratic in N (see full paper) and thus a key contribution of this work is to demonstrate how to efficiently estimate it in linear time.

To this end, we discretize the space as $\mathbb{U}' = (\mathbf{u}'_1, \mathbf{u}'_2 \ldots, \mathbf{u}'_V) \subseteq \mathbb{U}$. The main advantage is that the discrete metric simplifies to a matrix of distances $D_{\mathbb{U}} = \left(d_{\mathbb{U}}(\mathbf{u}'_i, \mathbf{u}'_j)\right)$ that can be precomputed and cached beforehand. Even better, the kernel functions can be cached for all pairs of points $(\mathbf{u}'_i, \mathbf{u}'_j) \in \mathbb{U}'$. For our experiments, we choose the kernel function on this space to be an exponential $k(\mathbf{u}'_i;\mathbf{u}'_j) = \frac{1}{Z} \exp\left(-\frac{d_{\mathbb{U}}(\mathbf{u}'_i,\mathbf{u}'_j)^2}{2\sigma^2}\right)$ where $d_{\mathbb{U}}\left(\mathbf{u}'_i, \mathbf{u}'_j\right)$ is the geodesic distance on the model and σ is the bandwidth of the kernel.

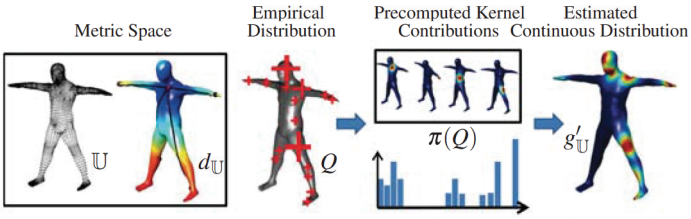


Figure 1: We propose a method to quickly estimate the continuous distributions on the manifold or more generally the metric space induced by the surface model. This allows us to efficiently train a random forest to predict image to model correspondences using a continuous entropy objective.

Using our discretization $\mathbb{U}'$ we then smooth the empirical distribution provided by Q over this discretization using the pre-computed kernel contributions as

$$g_{U'}(\mathbf{u}'_i; Q) \simeq \frac{1}{N} \sum_{j \in \mathcal{N}_i} \pi_j(Q) k(\mathbf{u}'_i; \mathbf{u}'_j) \tag{4}$$

where the weights $\pi_j(Q)$ are the number of data points in the set Q that are mapped to the bin center $\mathbf{u}'_j$. In other words, $\{\pi_j(Q)\}_{j=1}^V$ are the unnormalized histogram counts of the discretization given by $\mathbb{U}'$. We can use this to further approximate the continuous KDE entropy estimate of the underlying density in Eq. 3 as

$$p_U(\mathbf{u}) \simeq f_U(\mathbf{u}; Q) \simeq g_{U'}(\alpha(\mathbf{u}); Q) \tag{5}$$

where $\alpha(\mathbf{u})$ maps $\mathbf{u}$ to a point in our discretization. Using this, we approximate the differential entropy of $p_U(\mathbf{u})$ using the discrete entropy of $g_{U'}$ defined on our discretization. Hence, our MSIG estimate of the entropy on the metric space for an empirical sample Q is

$$\hat{H}_{\text{MSIG}}(Q) = -\sum_{u_i \in \mathbb{U}'} g_{U'}(\mathbf{u}'_i; Q) \log g_{U'}(\mathbf{u}'_i; Q)\ . \tag{6}$$

Only the calculation of the histogram counts scales with the number of training examples and thus, the complexity of calculating (6) is linear.

We find that forests trained using our MSIG objective function can provide substantially better correspondences in comparison to the forests trained using the objective from [5]. These improved correspondences translates into modest improvements in pose estimation that allows us to achieve state of the art pose estimation results with orders of magnitude less training data.

[1] S. Nowozin. Improved information gain estimates for decision tree induction. In *ICML*, 2012.

[2] G. Pons-Moll and B. Rosenhahn. Model-based pose estimation. *Visual Analysis of Humans*, pages 139–170, 2011.

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304. IEEE, 2011.

[4] C. Sminchisescu, L. Bo, C. Ionescu, and A. Kanaujia. Feature-based pose estimation. *Visual Analysis of Humans*, pages 225–251, 2011.

[5] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *CVPR*, 2012.

Label embedding for text recognition

Jose A. Rodriguez-Serrano
jose-antonio.rodriguez@xrce.xerox.com
Florent Perronnin
Florent.Perronnin@xrce.xerox.com
Xerox Research Centre Europe
Meylan, France

Summary The standard approach to recognizing text in images consists in first classifying local image regions into candidate characters and then combining them with high-level word models such as conditional random fields (CRF). *This paper explores a new paradigm that departs from this bottom-up view.*

In our approach, every label from a lexicon is *embedded* to an Euclidean vector space. We refer to this step as *label embedding*. Each vector of image features is then projected to this space. To that end, we formulate the problem in a structured support vector machine (SSVM) framework [3] and learn the linear projection that optimizes a proximity criterion between word images and their corresponding labels: matching label-image pairs should be closer than non-matching pairs. In this space, the "compatibility" between a word image and a label is measured simply as the dot product between their representations. Therefore, given a new word image, recognition amounts to finding the closest label in the common space (Fig. 1).

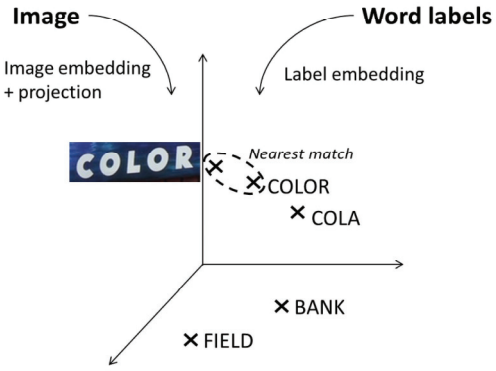


Figure 1: Illustration of recognition with label embedding.

This method presents the following advantages: (i) it does not require costly pre- or post-processing operations, (ii) it allows for the recognition of never-seen-before words, (iii) the recognition process is efficient.

Model Let $\theta : \mathcal{X} \rightarrow \mathbb{R}^D$ be a function that acts on the pixels of x and extracts a D-dimensional feature vector $\theta(x)$ (feature embedding).

Let $\varphi : \mathcal{Y} \rightarrow \mathbb{R}^E$ denote a function that computes a fixed-length feature vector from the label y (label embedding).

We use the following similarity function between the (projected) image embeddings and the label embeddings :

$$F(x,y;W) = \tilde{\theta}^T(x)\varphi(y) = \theta^T(x)W\varphi(y). \quad (1)$$

If the matrix W is known, recognizing the text in image x amounts to scanning the lexicon $\mathcal{Y}$ for a best match:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(x,y;W) \quad (2)$$

The goal of learning is to find the optimal matrix W.

Embeddings For the image embeddings, we use the widely adopted bag-of-patches framework. We choose to compute the patch statistics using the Fisher Vector (FV) principle [4].

For the image embeddings, we propose a Spatial Pyramid of Characters (SPOC). Given a text label, the SPOC counts the frequencies of character appearances at certain subdivisions of the text label, as illustrated in Fig. 2. This embedding is data-free (*i.e.* any label can be easily embedded on-the-fly), respects the lexical similarity between words and is expressed with a fixed-length feature vector.

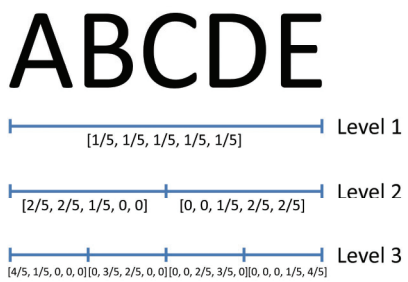


Figure 2: SPOC label embedding.

Learning We note that Eqs. (1) and (2) can be re-written in the form of a ranking SSVM with an objective of the form

$$w^* = \arg\min_w \frac{1}{N}\sum_{n=1}^{N} B_2(x_n, f(y_n)) + \frac{\lambda}{2}||w||^2, \quad (3)$$

where

$$B_2(y_n, f(x_n)) = \sum_{y \in \mathcal{Y}} \Delta(y_n, y) - F(x_n, y_n; w) + F(x_n, y; w), \quad (4)$$

which can be optimized with Stochastic Gradient Descent (SGD) [1].

Experiments Experiments are performed on a private license plate recognition dataset and on the IIIT-5K scene text dataset [2] show that the proposed method is competitive with standard bottom-up approaches to text recognition.

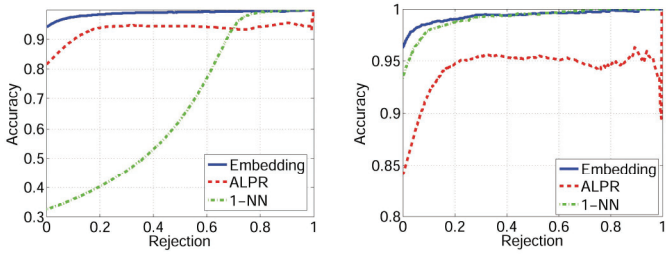


Figure 3: License plate results. Left: using the whole test set. Right: Using only the subset of images which have a true match in the database.

[1] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998.

[2] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.

[3] S. Nowozin and C. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 2011.

[4] F. Perronnin, J. Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.

Change Detection in Dynamic Scenes using Local Adaptive Transform

Hakan Haberdar
hhaberdar@uh.edu,www.haberdar.org
Shishir K. Shah
sshah@central.uh.edu,www.qil.uh.edu

Quantitative Imaging Lab
University of Houston
Houston, Texas, U.S.A

In this paper, we propose a framework that can be used for detecting relevant changes in highly dynamic scenes, where the background has several changing elements (Fig. 1).

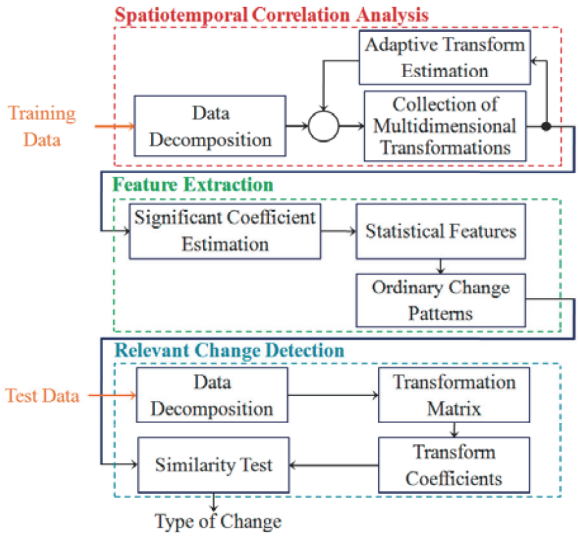


Figure 1: Our algorithm consists of 3 main blocks. Given a video containing only ordinary changes, our goal is to find representations where the spatiotemporal features of the ordinary changes can be captured. Then, we use the training examples to extract spatiotemporal signatures of the ordinary change patterns. Finally, we estimate the existence of relevant change in a test input by interpolating from the training samples.

In dynamic environments, there are almost always changes in the scene. To establish a clear distinction between what is relevant change and what is not, we first categorize the change into two main classes; namely, *ordinary change* and *relevant change*. Changes are considered as irrelevant if they are recurrent elements and changes pertaining on the dynamic background of the scene. Pixels, which belong to regions of the ordinary change, are typically correlated in space and/or time among a set of consecutive frames. This correlation stems from the repetitive nature of ordinary change patterns and induces spatiotemporal signatures [6], which are specific to the ordinary change patterns. We propose that one can make use of the spatiotemporal signatures to discriminate ordinary changes from relevant changes.

The proposed framework makes use of a collection of orthogonal linear transforms to capture spatiotemporal signatures of local ordinary change patterns and subsequently employ them in the detection of relevant changes. We employ three orthogonal linear transforms as the base transforms: i) discrete cosine transform (DCT) [2], ii) Walsh-Hadamard transform (WHT) [1], and iii) Slant transform (ST) [5]. DCT, WHT, and ST are used together because they have complementary basis vectors that enable the framework to capture different types of ordinary change patterns. Our goal is to estimate the most suitable transform for each local ordinary change pattern.

Given a set of frames containing only the ordinary changes, we divide every frame into regions of 8 by 8 pixels in order to improve the localized correlation. Then, 8 consecutive frames are grouped to form a stack. Every stack is composed of 8 8 8 blocks called as *cubes*. To estimate the most suitable transform to extract the spatial signature of a cube, we define the compactness metric ξ_s as follows:

$$\xi_s = \sum_{i=1}^{N} \left(\frac{1}{N} - \mathfrak{p}_i \right)^2, \text{ and } \xi_s \in [0, 1 - \frac{1}{N}] \quad (1)$$

where $\mathfrak{p}_i$ represents a pixel in a cube $\mathbb{C}$ and N is the number of the pixels

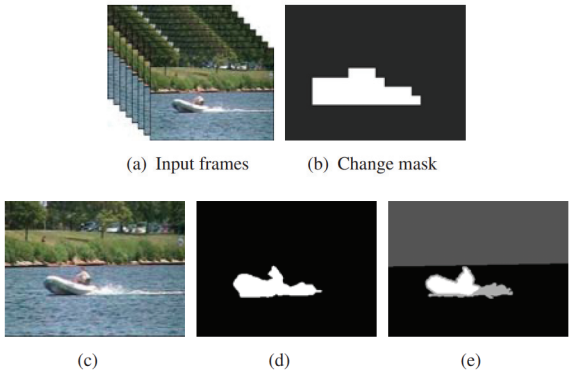


Figure 2: In (b), we present the binary change mask estimated for the 8 consecutive frames given in (a). The change mask is a two-dimensional projection of spatiotemporal changes in these 8 frames. In (c), (d), and (e), we present an input image, relevant changes detected, and the ground truth. The gray levels in (e) are 0:*ordinary change*, 255:*relevant change*, 85:outside region of interest, and 170:unknown motion [3]. Our goal is to detect pixels labeled as *relevant change*.

in the cube $\mathbb{C}$. The transform having the largest compactness coefficient value is considered as the most suitable for the cube. Given a test cube, we construct a maximum likelihood model by using the bivariate inequality of Lal [4] and interpolating from the training instances:

$$P(\lambda_{L_X} < X < \lambda_{U_X}, \lambda_{L_Y} < Y < \lambda_{U_Y}) \geq P_{XY}, \text{ and} \quad (2)$$

$$P_{XY} = 1 - \frac{1}{2k_X^2 k_Y^2}(k_X^2 + k_Y^2 + \sqrt{(k_X^2 + k_Y^2)^2 - 4\rho^2 k_X^2 k_Y^2}), \quad (3)$$

where $\lambda_{L_X} + \lambda_{U_X} = 2\mu_X$, $\lambda_{L_Y} + \lambda_{U_Y} = 2\mu_Y$, $k_X = (\lambda_{U_X} - \lambda_{L_X})/2\sigma_X$, and $k_Y = (\lambda_{U_Y} - \lambda_{L_Y})/2\sigma_Y$. Eq. 2 gives a lower bound for the joint probability of the interval $[\lambda_{L_X}, \lambda_{U_X}]$ around μ_X and the interval $[\lambda_{L_Y}, \lambda_{U_Y}]$ around μ_Y for the random variables X and Y. At the frame level, this corresponds to a two-dimensional projection of spatiotemporal changes within the stack of 8 consecutive frames (Fig. 2 (b)). This summary image is called *binary change mask*, where 1 and 0 indicate the relevant and ordinary change, respectively. We apply two-dimensional version of the framework to obtain the relevant change at the pixel level.

The use of this framework is demonstrated in a variety of videos with highly dynamic backgrounds including lakes, pools, and roads. The major limitation of our method is that estimating base transforms requires a set of frames without relevant changes. Another limitation arises from cube-based computations, which may cause blocking artifacts. Compared to other methods demonstrated on the same test videos, our method shows significant improvement in change detection results.

[1] N. Ahmed and K.R. Rao. Walsh-Hadamard transform. In *Orthogonal Transforms for DSP*, pages 99–152. Springer, 1975.

[2] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *Computers, IEEE Transactions on*, C-23(1):90–93, 1974.

[3] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. Changedetection.net: A new change detection benchmark dataset. In *CVPRW 2012*, pages 1–8, 2012.

[4] D.N. Lal. A note on a form of tchebycheff's inequality for two or more variables. *The Indian J. of Statistics*, 15(3):317 320, 1955.

[5] W. Pratt, W.H. Chen, and L. Welch. Slant transform image coding. *Communications, IEEE Transactions on*, 22(8):1075–1093, 1974.

[6] James V Stone. Object recognition: View-specificity and motion-specificity. *Vision Research*, 39(24):4032–4044, 1999.

Learning beautiful (and ugly) attributes

Luca Marchesotti
luca.marchesotti@xerox.com
Florent Perronnin
florent.perronnin@xerox.com

XRCE
Xerox Research Centre Europe
Meylan, France

The problem. The amount of visual content we handle on a daily basis has grown exponentially. In this ocean of images and videos, there are many questions that artificial systems could help us answer. In the last decade, the focus of the computer vision community had been on semantic recognition. While this is still a very active research field, new questions are arising. For instance, we might want to predict what people like in an image or a video. Although this is a very challenging question, even for humans, it was shown experimentally that aesthetics/preference can be predicted using data-driven approaches [1, 2]. Current approaches to aesthetic image analysis either provide accurate or interpretable results. We thus raise the following question: *can we preserve the advantages of generic features **and** get interpretable results?*.

The solution. In this work, we will address this problem by *discovering and learning attributes automatically*. In particular, our main contribution is a novel approach to aesthetic image analysis which combines the benefits of "attribute-based" and "generic" techniques. It consists of (i) automatically discovering a vocabulary of visual attributes and (ii) learning their visual appearance using generic features. For this purpose, we leverage the AVA dataset [4] which contains more than 250,000 images together with their aesthetic preference ratings and textual comments. Preference ratings allow us to supervise the creation of the attribute vocabulary (step (i)) and to learn automatically the visual appearance of attributes (step (ii)). Finally we apply these learned attributes to three different scenarios: aesthetic quality prediction, image classification and retrieval.

The data. We use AVA, a recently introduced database [4] which contains more than 250,000 images downloaded from www.dpchallenge.com. An interesting characteristic of this dataset is that images are accompanied by natural language text and attractiveness scores. This dataset was assembled for large-scale evaluation of attractiveness classification and regression tasks. But it was also recently used to study the dependence of attractiveness on semantic information [3]. Another peculiarity of this corpus is the organization of photos in *contests*: an equivalent of Flickr groups where images are ranked according to attractiveness scores left by users. Consider the sample images in figure 1, they were taken from the contest *"**Green Macro**: Get up close and personal with the subject of your choice, using green as your primary color"*. Photos in the first row scored highly, the others were ranked at the bottom of the contest. While all six images contain a lot of green, the top ones have brighter, more vivid green elements and the photographic technique *"Macro"* is much better represented.

Discovering aesthetic attributes. Mining attributes by hand-picking photographic rules from a book is problematic: this is a non-exhaustive procedure and it does not give any indication of how much, and when, these techniques should be used. Therefore, we mine beautiful and ugly attributes by discovering which terms can predict the aesthetic score of an image from AVA textual comments. For this purpose, we train an Elastic Net to predict aesthetic scores and, at the same time, select textual features. It is a regularized regression method that combines an ℓ_2-norm and a sparsity-inducing ℓ_1-norm. Let N be the number of textual documents. Let D be the dimensionality of the BOW histograms. Let $\mathbf{X}$ be the $N \times D$ matrix of documents. Let y be the $N \times 1$ vector of scores of aesthetic preference (the score of an image is the average of the scores it received). We learn:

$$\hat{\beta} = \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||^2 \quad (1)$$

where λ_1 and λ_2 are the regularization parameters.

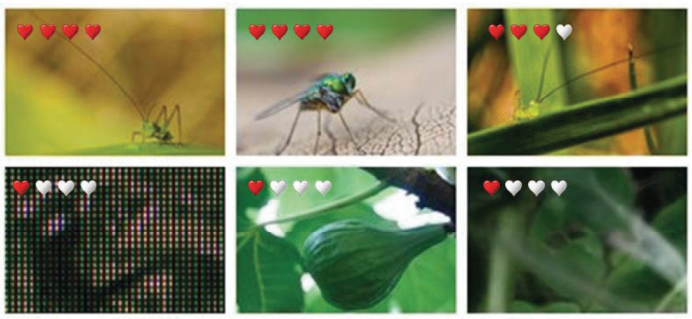


Figure 1: Sample photos from the challenge "Green Macro"

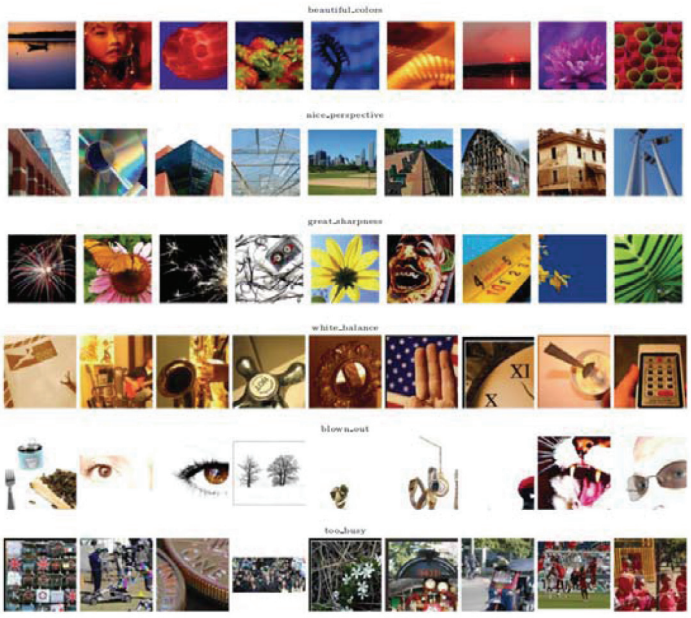


Figure 2: Images with top scores for some representative beautiful and ugly attributes. From top to bottom: beautiful colors, nice perspective, great sharpness, color cast, white balance, blown out, too busy.

Assessing attributes visualness. We enforced interpretability and discriminability of the attribute labels using attractiveness scores as a supervision mechanism. However, this choice does not ensure that all these attributes can be recognized by a computer. This is the reason why we measure "visualness" using Area Under the ROC Curve (AUC) calculated for each individual attribute. In particular, we benchmark the classification performances of each attribute (1-vs-all) and we rank them using AUC. We show some qualitative results in Figure 2 for Ugly and Beautiful attributes.

Conclusions. In this paper, we tackled the problem of visual attractiveness analysis using visual attributes as mid-level features. Despite the great deal of subjectivity of the problem, we showed that we can learn automatically meaningful attributes that can be used in various applications such as score prediction, auto-tagging or retrieval. Future work will focus on testing with users the advantage of our beautiful and ugly attributes and on mitigating biases introduced by semantic information.

[1] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[2] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[3] N. Murray, L. Marchesotti, and F. Perronnin. Learning to rank images using semantic and aesthetic labels. In *BMVC*, 2012.

[4] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

Fisher Vector Faces in the Wild

Karen Simonyan
karen@robots.ox.ac.uk
Omkar M. Parkhi
omkar@robots.ox.ac.uk
Andrea Vedaldi
vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

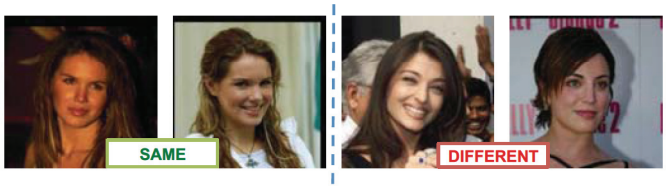


Figure 1: **Face Verification.** Given a pair of face images, the task is to verify if both depict the same person. This is a challenging task due to the variations in lighting, pose, face expression, etc.

Overview. The face verification and recognition domain has mostly been dominated by carefully designed representations based on features computed around numerous facial landmarks [1, 3]. On the other hand, for image classification, image representations are capable of capturing discriminative image information without any domain-specific knowledge by using densely computed features, coupled with a non-linear encoding. For instance, the high-dimensional Fisher vector encoding [6] of SIFT features [5] achieves state-of-the-art performance on several image classification benchmarks [2].

In this paper, we address this discrepancy and make two contributions. First, we show that Fisher vector encoding of dense SIFT, an off-the-shelf image representation, achieves state-of-the-art face verification performance on the challenging "Labeled Faces in the Wild" (LFW) benchmark. Second, we show that the high-dimensional face representation using Fisher vector encoding is amenable to discriminative dimensionality reduction. The resulting compact descriptor has equal or better recognition accuracy and is very well suited to large-scale face recognition tasks.

The overview of our face descriptor computation pipeline is shown in Fig. 2. The learnt dimensionality reduction model is visualised in Fig. 3. It is able to automatically extract the discriminative regions of the face.

Results. In the unrestricted setting of LFW, we achieve $93.03 \pm 1.05\%$ face verification accuracy, closely matching $93.18 \pm 1.07\%$ obtained by the state-of-the-art method [3], based on high-dimensional LBP descriptors sampled around 27 face landmarks. It should be noted that the best result of [3] using SIFT is 91.77%. In the restricted setting of LFW, we achieve the verification accuracy of $87.47 \pm 1.49\%$, setting the new state-of-the-art. This is better than the second best method of [4] by 3.4%.

The ROC curves of our method as well as the other methods are shown in Fig. 4.

[1] T. Berg and P. N. Belhumeur. Tom-vs-Pete classifiers and identity-preserving alignment for face verification. In *Proc. BMVC.*, 2012.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. BMVC.*, 2011.

[3] D. Chen, X. Cao, F. Wen, and Sun J. Blessing of dimisionality: High dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013.

[4] H. Li, G. Hua, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Proc. CVPR*, 2013.

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

[6] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

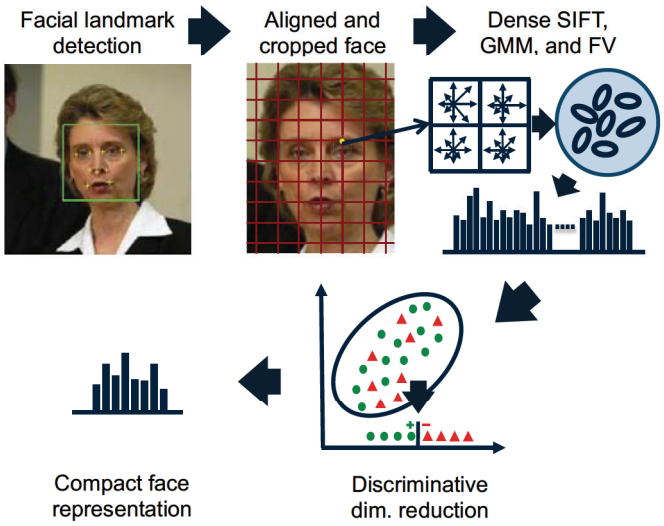


Figure 2: **Method overview.** Given an image, we first run an off-the-shelf face and facial landmark detectors to align the face to a canonical frame. Spatially-augmented SIFT features are then **densely** extracted from the face region and encoded into a high-dimensional **Fisher vector face representation**. Finally, discriminative dimensionality reduction is learnt on these features to compress them into a compact representation, while improving the discriminative ability.

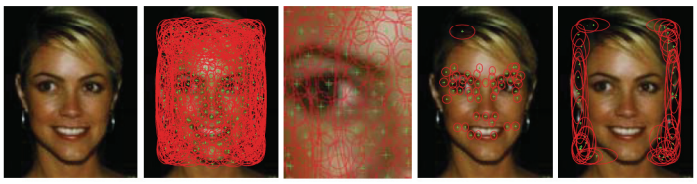


Figure 3: **Learnt model.** A Fisher vector, coupled with discriminative dimensionality reduction, can automatically capture the discriminative parts of the face. (a): an aligned face image; (b): unsupervised GMM clusters densely span the face; (c): a close-up of a face part covered by the Gaussians; (d): 50 Gaussians of the GMM, corresponding to the learnt projection matrix columns with the highest energy; (e): 50 Gaussians corresponding to the learnt projection matrix columns with the lowest energy.

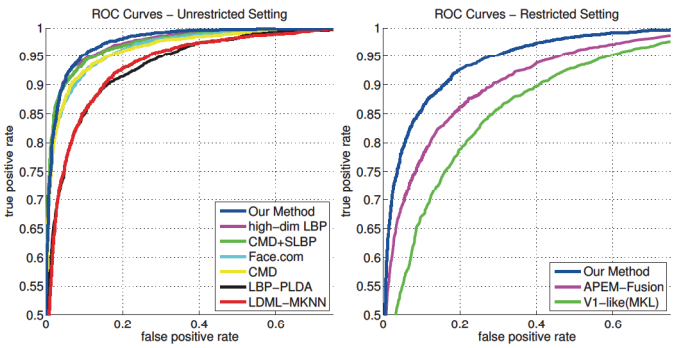


Figure 4: **Comparison with the state-of-the-art.** ROC curves of our method and the state-of-the-art techniques in LFW-unrestricted (left) and LFW-restricted (right) settings.

Perception Preserving Projections

Saining Xie[1]
xiesaining@gmail.com
Jiashi Feng[2]
a0066331@nus.edu.sg
Shuicheng Yan[2]
eleyans@nus.edu.sg
Hongtao Lu[1]
htlu@sjtu.edu.cn

[1] Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China

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

Unsupervised learning of feature projections to low-dimensional linear subspaces from training data has become a standard paradigm in the areas of pattern recognition and computer vision. Principle component analysis (PCA) [2] in particular is one of the most popular projection learning methods for dimensionality reduction and feature extraction. Representations learned with PCA have proven useful for tasks such as face and object recognition, tracking, detection, and background modeling.

Compared with the descriptive features learned from the raw intensity domain, many advanced features work better either for machine perception (pattern recognition accuracy), or human perception (physical interpretability), or both. In those cases, the standard PCA is not tuned to any domain-specific features, thus the discriminative features for certain perception systems may be lost in the projection process.

In this work, we propose a perception preserving projection (PPP) method that is able to preserve the desired feature characteristic for projected images. PPP pursues a set of suitable projection basis which minimize the reconstruction loss for both the original images and the extracted features simultaneously. To quantitatively evaluate the machine perception preserving capability of PPP, we experimentally evaluate the performance of face recognition on the reconstructed face images from both PCA and PPP. It is shown that on the face images reconstructed by PPP, face recognition can achieve significantly improved performance, even in much lower dimensions compared to PCA.

The objective function of PPP can then be formulated as:

$$\min_{U^TU=I_r} \mathcal{L}(U) = \|\mathcal{P}'(X) - \mathcal{P}'(UU^TX)\|_F^2, \tag{1}$$

There are various feature extractors can be seen as linear operators $\mathcal{P}$ over the data vector $\mathbf{x} \in \mathbb{R}^d$. For example, the convolution of data with linear filter, pixel-wise "masking", and the sum of filters. Here $P' = [(1-\alpha)P, \alpha I_d]$, and α is a trade-off parameter between the original data space and feature space. The above objective function states that we aim to find a set of projection basis $U \in \mathbb{R}^{d\times r}$, such that the extracted features from the reconstructed data $\mathcal{P}(UU^TX)$ will not deviate from the features extracted from the original data $\mathcal{P}(X)$ too much.

The most straightforward method to solve the problem (1) is performing gradient descent on the Stiefel manifold defined by $U^TU = I$. Though state-of-the-art algorithms can be directly exploited, the computational cost is quite high.

Inspired by the Robust PCA work [1], which seeks a low rank matrix to approximate the original data matrix, here we also relax the orthogonal constraint in the objective function and just seek a low rank matrix as the transformation matrix.

Therefore, the objective function in (1) can be relaxed as follows:

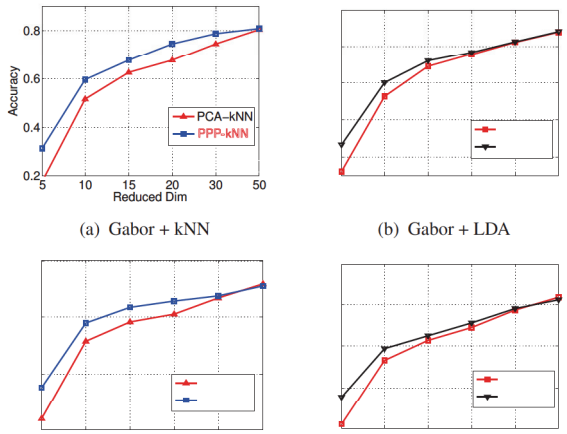
$$\min_W \|\mathcal{P}(X) - \mathcal{P}(WX)\|_F^2, \text{ s.t. } \operatorname{rank}(W) \le r, \tag{2}$$

The above objective function can be further relaxed as:

$$\min_W \|W\|_* + \lambda\|E\|_F^2, \text{ s.t. } \mathcal{P}(X) - \mathcal{P}(WX) = E, \tag{3}$$

where E explicitly accommodates the reconstruction errors.

The optimization problem in (3) can be solved by the Alternating Direction Method (ADM) [3] efficiently. One of the advantages of ADM is that the original optimization problem can be decomposed into several subproblems which are relatively easier to solve. However, it is difficult to solve the sub-problem for optimizing the function w.r.t. W. Directly minimizing the subproblem above will lead to solving a discrete-time Sylvester equation in each iteration, which is infeasible for the ADM algorithm and makes the low-rank relaxation rewardless.



(a) Gabor + kNN (b) Gabor + LDA

(c) LoG + kNN (d) LoG + LDA

Figure 1: The performance comparison for reconstructed face recognition from (a) PCA and (b) PPP on FRGC dataset based on Gabor/LoG feature.

To alleviate such difficulty, we adopt the recently developed Linearized Alternating Direction Method (LADM) [4] and linearize the quadratic term in the above Lagrangian function at the point W^k:

$$\mathcal{L}(W,W_k) = \|W\|_* + \langle Y, -\mathcal{P}(W_kX)\rangle + \mu\left\langle \mathcal{P}^*\left(\mathcal{P}(X) - \mathcal{P}(W^kX) - E\right)X^T, W - W_k\right\rangle + \frac{\mu\eta}{2}\|W - W_k\|_F^2. \tag{4}$$

Here $\eta > (\|\mathcal{P}\|\|X\|)^2$ is the Lipschitz constant of the linear operators imposed on variable W. After several algebra computation, the above objective function can be written as:

$$\mathcal{L}(W,Y,W_k) = \|W\|_* + \frac{\mu\eta}{2}\|W - M_k\|_F^2, \tag{5}$$

where $M_k = W_k - \mathcal{P}^*(\mathcal{P}(X) - \mathcal{P}(W_kX) - E)X^T/\eta + \mathcal{P}^*YX^T/\mu\eta$. $\mathcal{P}^*$ denotes the adjoint of the operator $\mathcal{P}$, which is defined as $\langle \mathcal{P}(X), Y\rangle = \langle X, \mathcal{P}^*(Y)\rangle$. It is well known that the above objective function has following closed form solution:

$$W_{k+1} = U\mathcal{S}_{\frac{1}{\mu\eta}}(\Sigma)V^T, \tag{6}$$

where U, Σ, V are from SVD on the matrix M_k. And $\mathcal{S}(\cdot)$ is a shrinkage operator defined as $\mathcal{S}_\varepsilon[x] = sgn(x)max(|x| - \varepsilon, 0)$. Here the shrinkage operator is performed element-wisely for the involved matrix. To guarantee a good convergence rate, we adopt following adaptive penalty strategy [4]:

$$\mu_{k+1} = \begin{cases} \rho_0\mu_k, & \text{if } \mu_k \max(\sqrt{\eta}\varepsilon_W, \varepsilon_E)/\|\mathcal{P}(X)\| < \varepsilon_2, \\ \mu_k, & \text{otherwise.} \end{cases} \tag{7}$$

Here $\varepsilon_W = \|W_{k+1} - W_k\|$ and $\varepsilon_E = \|E_{k+1} - E_k\|$. And the stopping criterion is:

$$\|\mathcal{P}(X) - \mathcal{P}(W^kX) - E\|/\|\mathcal{P}(X)\| < \varepsilon_1. \tag{8}$$

In our implementation, we also adopt partial SVD and rank prediction techniques using PROPACK and represent W as its skinny SVD to avoid full matrix multiplications and thus yielding a complexity of $\mathcal{O}(nd^2)$.

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 11, 2011.

[2] Ian T Jolliffe. *Principal component analysis*. Springer verlag, 2002.

[3] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[4] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *NIPS*, 2011.

A Novel Approach for Efficient SVM Classification with Histogram Intersection Kernel

Gaurav Sharma
gaurav.sharma@unicaen.fr
Frederic Jurie
frederic.jurie@unicaen.fr

GREYC CNRS UMR6072
University of Caen
Boulevard Marechal Juin
14032 Caen, France

We propose a novel approach for learning non-linear SVM corresponding to the histogram intersection kernel without using the kernel trick. We formulate the exact non-linear problem in the original space and show how to perform classification directly in this space. The learnt classifier incorporates non-linearity while maintaining O(d) testing complexity (for d-dimensional input space), compared to O($d \times N_{sv}$) when using the kernel trick. We show that the SVM problem with histogram intersection kernel is quasi-convex in input space and outline an iterative algorithm to solve it. The proposed approach has been validated in experiments where it is compared with other linear SVM-based methods, showing that the proposed method achieves similar or better performance at lower computational and memory costs.

The standard formulation for learning linear classifiers is the SVM primal formulation,

$$\min_{w} \left\{ \frac{\lambda}{2}||\mathbf{w}||_2^2 + \frac{1}{|\mathcal{I}_t|}\sum_i \xi_i \right\}, \quad \text{s.t. } y_i \mathbf{w}^T \mathbf{x}_i \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \quad (1)$$

where $\mathbf{w}$ is the normal to the linear decision hyperplane and $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathcal{I}_t, y_i \in \{-1, +1\}\}$ are the training vector and label pairs. However, general visual tasks *e.g.* scene or object based classification of unconstrained images, are very challenging due to the presence of high variability due to viewpoint, lighting, pose *etc.* and require non-linear decision boundaries for competitive performance. Such non linear classifier are obtained by using the *kernel trick*, $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$, with the dual formulation,

$$\max_{\alpha} \left\{ \sum_i \alpha_i + \left(\frac{1}{2} - \frac{1}{\lambda}\right) \sum_{i,j} \alpha_i \alpha_j y_i y_j \, \mathbf{x}_i^T \mathbf{x}_j \right\}, \quad \text{s.t.} \quad 0 \leq \alpha_i \leq \frac{1}{|\mathcal{I}_t|}. \quad (2)$$

While the dual formulation allows learning of non linear decision boundaries, the computation of classifier decision for a test vector $\mathbf{x}$, $f(\mathbf{x}) \propto \sum_i c_i K(\mathbf{x}, \mathbf{x}_i)$ (c_i being the model parameters), depends on kernel computation with *all support vectors* $\{\mathbf{x}_i \in \mathbb{R}^d | i = 1 \ldots N_{sv}\}$. Hence, the test time and space complexities becomes $O(d \times N_{sv})$ vs $O(d)$ for the linear case *i.e.* $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$. In practice, N_{sv} is of the order of number of training examples, and this leads to significant cost in terms of time and space. As a result of very successful histrogram based image representations *e.g.* Histogram of Oriented Gradients (HOG) and Bag-of-features, the related histogram intersection kernel $\sum_d \frac{x_d y_d}{|x_d||y_d|} \min(|x_d|, |y_d|) \;\; \forall \;\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ has been a natural choice, in visual tasks. This has led to many recent works proposing faster ways to compute the classifier and/or the decision function corresponding to the histogram intersection kernel *e.g.* [1, 2, 3].

In this paper, we take an (as far we know) unexplored route and show that it is possible to learn a nonlinear classifier implementing the histogram intersection kernel directly in the input space without using the dual formulation and the kernel trick, while achieving similar classification performance. We work with the primal optimization problem and view the kernel as a parametrized scoring function inducing a nonlinear boundary in the input space. The SVM primal optimization in feature mapped space reduces to the following, in input space,

$$L(\mathbf{w}) = \frac{\lambda}{2}||\mathbf{w}||_1 + \frac{1}{|\mathcal{I}_t|}\sum_i \max(0, m - y_i f(\mathbf{w}, \mathbf{x})) \quad (3a)$$

$$f(\mathbf{w}, \mathbf{x}) = \sum_d \frac{w_d}{|w_d|} min(x_d, |w_d|) \;\; (\text{as } \mathbf{x} \in \mathbb{R}_+^d). \quad (3b)$$

We show that the optimization problem becomes non convex, but remains *quasi* convex. To solve the problem, we outline an approximate optimization algorithm which starts by solving a, highly smoothed, convex approximation of the objective and then successively solves less smoothed objectives, initialized with the solution of the previous one, converging to the current objective (Fig. 1). In practice, however, we find that a simple stochastic subgradient descent solver, working with the subgradient of the non convex objective, with very small steps and multiple pass over the data achieves competitive performance.

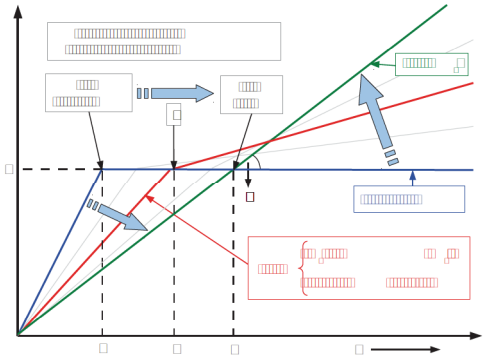


Figure 1: The linear decision function (green) can be seen as a convex relaxation of the histogram intersection like decision function (blue).

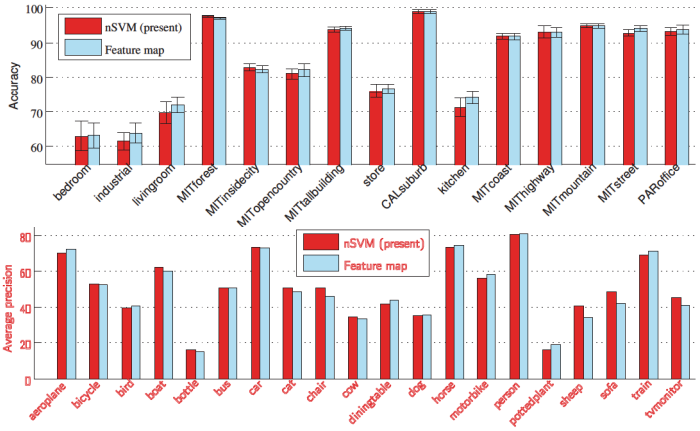


Figure 2: Performance on Scene 15 (top) and Pascal VOC 2007 dataset (bottom) for the proposed method (nSVM) and explicit feature maps [3].

	Time		Memory	
	Secs	Speedup	Kbytes	Reduction
Vedaldi and Zisserman [3]	3.8	1 (ref)	448	1 (ref)
Nonlinear SVM (present)	0.2	19×	64	7×

Table 1: Testing time and memory usage, on test set of aeroplane class of Pascal VOC 2007 dataset.

We validate the method by experiments on two challenging and widely used benchmarks datasets in image classification. While with the current approaches using explicit feature maps [2, 3] two costs are incurred at test time, (i) mapping the test vector and (ii) computing $O(d')$ dot product, in higher d'-dimensional feature space, with the proposed method we only need to perform $O(d)$, with $d < d'$, operations in original space. Our method achieves comparable performance (Fig. 2) to state-of-the-art methods for efficient SVM with histogram kernels while being faster and memory efficient (Tab. 1). Along with being conceptually interesting, our method thus leads to reduced space and time complexities.

[1] Subhransu Maji and Alexander C Berg. Max-margin additive classifiers for detection. In *ICCV*, 2009.

[2] Subhransu Maji, Alexander C Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[3] A. Vedaldi and A. Zisserman. Efficient additive kernels using explicit feature maps. In *CVPR*, 2010.

Inferring Ongoing Human Activities Based on Recurrent Self-Organizing Map Trajectory

Qianru Sun
qianrusun@sz.pku.edu.cn

Hong Liu
hongliu@pku.edu.cn

Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School
Peking University, CHINA

Key Laboratory of Machine Perception(Ministry of Education)
Peking University, CHINA

Automatically inferring ongoing activities is to enable the early recognition of unfinished activities, which is quite meaningful for applications, such as online human-machine interaction and security monitoring. State-of-the-art methods use the spatio-temporal interest point (STIP) based features as the low-level video description to handle complex scenes [1, 2, 3]. While the existing problem is that typical bag-of-visual words (BoVW) focuses on feature distribution but ignores the inherent contexts in sequences, resulting in low discrimination when directly dealing with limited observations. To solve this problem, the Recurrent Self-Organizing Map (RSOM) [4], which was designed to process sequential data, is novelly adopted in this paper for the high-level representation of ongoing activities. The innovation lies that observed features and their spatio-temporal contexts are encoded in a trajectory of the pre-trained RSOM units. Additionally, a combination of Dynamic Time Warping (DTW) distance and Edit distance, named DTW-E, is specially proposed to measure the structural dissimilarity between RSOM trajectories.

RSOM Trajectory: Since the RSOM constitutes a direct extension of SOM, we start from SOM. SOM is to map the data from an input space $\boldsymbol{V}_I$ onto a lower dimensional space $\boldsymbol{V}_L$ (a map) in such way that the topological relationships in $\boldsymbol{V}_I$ are preserved and the SOM units approximate closely the probability density function of $\boldsymbol{V}_I$. Suppose each unit i in SOM is associated with a weight vector $\boldsymbol{w}_i = [w_{i1}, w_{i2}, ..., w_{in}]^T \in \Re^n$ with the same dimension as the input vector $\boldsymbol{x} = [x_1, x_2, ..., x_n]^T \in \Re^n$. Learning process that leads to self-organization on a map can be summarized as,

(i) The feature vector $\boldsymbol{x}(t)$ is input, then its best matching unit (*bmu*) on the map is found by computing the minimum distance as:

$$bmu = \arg\min_{i \in \boldsymbol{V}_L} \{\|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\|\} \tag{1}$$

(ii) The winner *bmu* and its neighbors on the map have their weights $\boldsymbol{w}_i(t)$ updated towards $\boldsymbol{x}(t)$ as:

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \alpha(t) \cdot N_{bmu,i} \cdot \|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\| \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\alpha(t) = \alpha_i \cdot (\alpha_f / \alpha_i)^{T(i)/T_{max}} \in [0,1]$ is the learning rate, where the α_i and α_f denote the initial rate and final rate. $T(i) = \{1, 2, ..., T_{max}\}$ where T_{max} is iteration number. $N_{bmu,i}$ is called neighborhood function and defined over the units on the map. Typically, $N_{bmu,i} = \exp\{-\|r_{bmu} - r_i\|^2 / 2\sigma^2\}$, where $r_{bmu} \in \Re^2$ and $r_i \in \Re^2$ are the location vectors of unit *bmu* and i on the map, and σ defines the Gaussian kernel width.

SOM is not originally designed to accommodate the time series, its temporal extension RSOM is hence adopted here to learn the temporal contexts in activity sequences. It is to utilize both the feature vectors before $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t)$ itself to search the best matching unit of $\boldsymbol{x}(t)$. This is done by associating the following recursive equation to each unit i to compute the difference vector $\boldsymbol{y}_i(t)$:

$$\boldsymbol{y}_i(t) = \lambda \cdot \|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\| + (1-\lambda) \cdot \boldsymbol{y}_i(t-1) \tag{3}$$

where $0 < \lambda < 1$ is a factor determining the influence of earlier difference vectors on the current $\boldsymbol{x}(t)$. When λ is close to 0, the system of Eq. (3) involves a heavy backward memory, whereas when λ is near to 1, Eq. (3) describes a slight memory. Now equations of RSOM for searching bmus and adapting weights are as follows,

$$bmu = \arg\min_{i \in \boldsymbol{V}_L} \{\|\boldsymbol{y}_i(t)\|\} \tag{4}$$

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \alpha(t) \cdot N_{bmu,i} \cdot \|\boldsymbol{y}_i(t)\| \tag{5}$$

Then, we introduce how to map the input feature sequence to a time-varying trajectory of bmus. Supposing that an $M \times M$ map is learned after a fixed number of iterations using Eq. (3)(4)(5). For simplification, the one-dimensional value b of the original coordinate $bmu \in \Re^2$, i.e.,

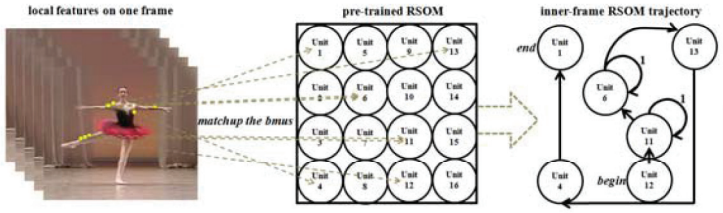


Figure 1: Illuminations of a pre-trained 4×4 RSOM and the assumed trajectory (12, 11, 11, 6, 6, 13, 4, 1) of local features (yellow points). Note that the bmus order in the trajectory is based on their image locations (detected order): from left to right, then from up to down.



Figure 2: Samples of UT-Interaction (a, b) and Rochester Activities dataset (c).

$b = bmu(2) \times M + bmu(1) \in [1, M^2]$, is used as the location index in the final trajectory. During video input at time t, STIP based local features are first extracted on the current frame based on above equations. Then feature vectors search their bmus on the map, and compose an inner-frame trajectory $\boldsymbol{b}$. Finally, $\boldsymbol{b}$ is used to generate the inter-frame trajectory $\boldsymbol{Trj}$.

$$\boldsymbol{b}_f = \{b_k | k = 1, 2, ..., K(f)\}; \boldsymbol{Trj}(t) = \{\boldsymbol{b}_f | f = 1, 2, .., F(t)\} \tag{6}$$

where $K(f)$ is the number of local features on the fth frame and it changes with different frames. $F(t)$ is the frame number until time t. $\boldsymbol{Trj}(t)$ is thus used as a high-level representation of the current observed activity.

DTW-E distance: The structure of RSOM trajectory is clear and special that each subset on one frame (inner-frame) contains the human shape information and the whole sequence (inter-frame) contains the long-range temporal relationships. Therefore, how to reasonably measure the likelihood between RSOM trajectories for pattern classification arises another problem. To solve this, a hierarchical distance based on the combination of DTW distance and Edit distance, named DTW-E, is specially defined (implementation of this algorithm is described in our paper).

Experiments and Conclusions: Two real-world datasets (Figure 2) with different characteristics, complex scenes [5] and inter-class ambiguities [6], serve as sources of data for evaluation. Experimental results based on kNN classifiers confirm that our approach can infer ongoing human activities at any stage with high accuracies.

Our conclusion is that our approach can make efficient inferences even with complex scenes and inter-class ambiguities. It hence confirms the ability of RSOM trajectory to extract sufficient discrimination. Moreover, the RSOM trajectory with the advantage of before-after independence is proved more suitable to the recognition of unfinished patterns.

[1] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie. Behavior recognition via sparse spatio-temporal features. *VS-PETS*, pages 65-72, 2005.
[2] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, pages 124.1-124.11, 2009.
[3] P. Scovanner, S. Ali, M. Shah. A 3-Dimensional SIFT Descriptor and its Application to Action Recognition. *ACM Conf. Multimedia*, pages 357-360, 2007.
[4] M. Varsta, José del R. Millán, J. Heikkonen. A Recurrent Self-Orgnizing Map for Temporal Sequence Processing. *LNCS* 1327, pages 421-426, 1997.
[5] M. S. Ryoo, J. K. Aggarwal. UT-Interaction Dataset. *ICPR Contest on Semantic Description of Human Activities (SDHA)*, 2010.
[6] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, In *Proc. ICCV*, pages 104-111, 2009.

Match-time covariance for descriptors

Eric Christiansen[1]
echristiansen@cs.ucsd.edu
Vincent Rabaud[2]
vrabaud@aldebaran-robotics.com
Andrew Ziegler[3]
amz@gatech.edu
David Kriegman[1]
kriegman@cs.ucsd.edu
Serge Belongie[1]
sjb@cs.ucsd.edu

[1] University of California, San Diego
La Jolla, California, USA
[2] Aldebaran Robotics
Paris, France
[3] Georgia Institute of Technology
Atlanta, Georgia, USA

Local descriptor methods are widely used in computer vision to compare local regions of images. These descriptors are often extracted relative to an estimated scale and rotation to provide invariance up to similarity transformations. We call this extract-time covariance (ETC) following the language of [1]. ETC is an imperfect process, however, and can produce errors downstream. Figure 1 illustrates the deterioration of common methods under changing scale and rotation.

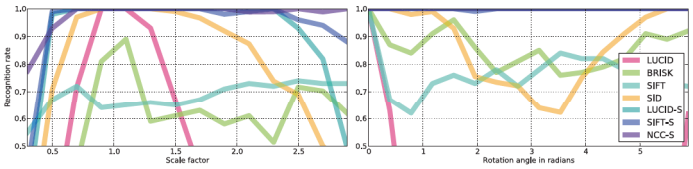


Figure 1: Recognition rates as a function of synthetic scale and rotation of the Oxford `boat` base image for various methods. The rates were obtained following the protocol of [2] and using the 100 strongest keypoints per image. The ETC methods (BRISK and SIFT) pay a penalty for the unreliability of their scale and rotation detectors. The Scale Invariant Descriptor (SID) lacks true scale invariance because it lacks the necessary post-matching normalization. Only the MTC methods (*-S) exhibit true invariance.

In this paper, we propose an alternative to steering we refer to as match-time covariance (MTC). MTC is a general strategy for descriptor design that simultaneously provides invariance in local neighborhood matches together with the associated aligning transformations. We also provide a general framework for endowing existing descriptors with similarity invariance through MTC. The framework, Similarity-MTC, is simple and dramatically improves accuracy. It is illustrated in Figure 2.

Finally, we propose NCC-S, a highly effective descriptor based on classic normalized cross-correlation, designed for fast execution in the Similarity-MTC framework. It is extremely simple, as it is just normalized cross-correlation with a novel normalization scheme, plus some bookkeeping for efficiency. NCC-S is described in Figure 3.

NCC-S is also extremely accurate, dramatically outperforming standard descriptors, which primarily use extract-time covariance. See Figure 4 for comparisons.

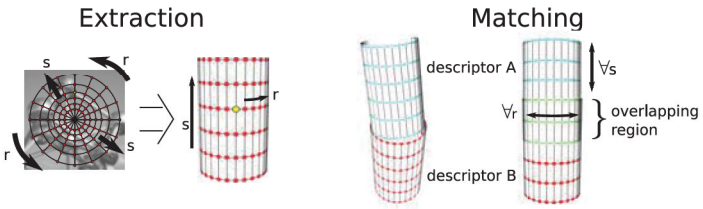


Figure 2: The Similarity-MTC extraction and matching framework, which *wraps* any existing descriptor and provides it with similarity invariance. This description is conceptual; actual implementations may be considerably more efficient. In extraction, descriptors are computed for a range of image scalings and rotations. To do this, a log-polar grid is centered at each keypoint. Each intersection is then associated with the similarity warp which brings the intersection to a fixed canonical location. Taken together, these points describe a cylinder, and descriptor matching is expressed as optimal cylinder alignment. The alignment can be effi-

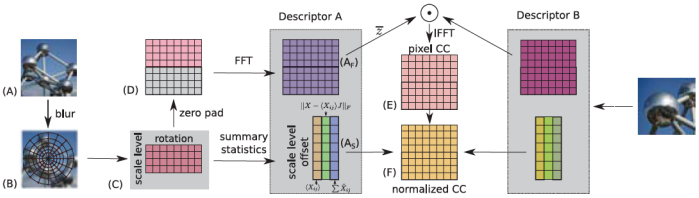


Figure 3: The full pipeline for NCC-S extraction and matching, an efficient descriptor in the Similarity-MTC framework. NCC-S is simply normalized cross-correlation, ported to Similarity-MTC, with some extra tricks for efficiency: 1) It works in Fourier space, 2) It does some bookkeeping to speed normalization.

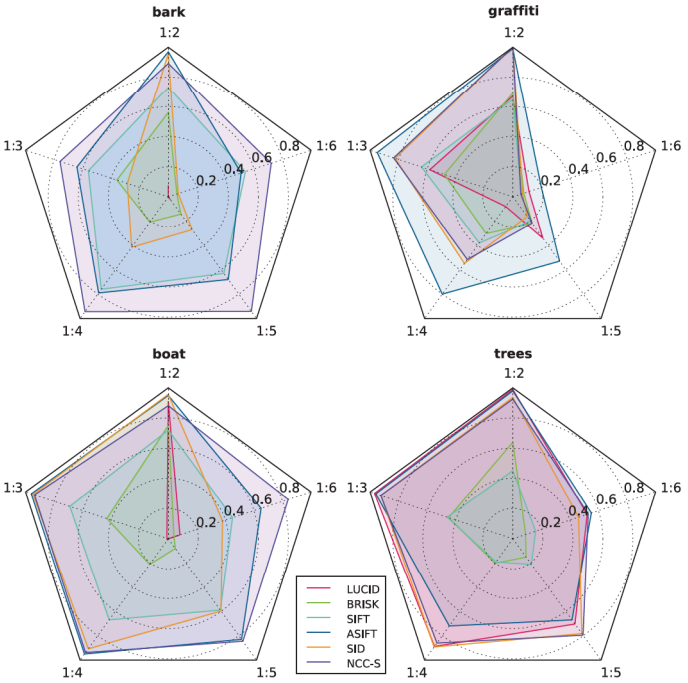


Figure 4: Recognition rates for various methods on a subset of the Oxford image dataset. Each vertex represents an image pair, ranging from `1:2` to `1:6`. The best methods are those that fill their pentagons. NCC-S dramatically outperforms existing methods, with only the much slower ASIFT approaching its accuracy.

The surprising effectiveness of this very simple descriptor suggests that MTC offers fruitful research directions for image matching previously not accessible in the ETC paradigm.

[1] Tinne Tuytelaars and Krystian Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 2007.

[2] Andrew Ziegler, Eric Christiansen, David Kriegman, and Serge Belongie. *NIPS*, 2012.

Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces

Pablo F. Alcantarilla
pablo.alcantarilla@gatech.edu
School of Interactive Computing
Georgia Institute of Technology, Atlanta, GA, USA

Jesús Nuevo
jesus.nuevochiquero@gmail.com
TrueVision Solutions
POBox 191, Toowong, QLD, Australia

Adrien Bartoli
adrien.bartoli@gmail.com
ISIT-UMR 6284 CNRS
Université d'Auvergne, Clermont-Ferrand, France

We propose a novel and fast multiscale feature detection and description approach that exploits the benefits of nonlinear scale spaces. Previous attempts to detect and describe features in nonlinear scale spaces such as KAZE [1] and BFSIFT [6] are highly time consuming due to the computational burden of creating the nonlinear scale space. In this paper we propose to use recent numerical schemes called *Fast Explicit Diffusion* (FED) [3, 4] embedded in a pyramidal framework to dramatically speed-up feature detection in nonlinear scale spaces. In addition, we introduce a *Modified-Local Difference Binary* (M-LDB) descriptor that is highly efficient, exploits gradient information from the nonlinear scale space, is scale and rotation invariant and has low storage requirements. Our features are called *Accelerated-KAZE* (A-KAZE) due to the dramatic speed-up introduced by FED schemes embedded in a pyramidal framework.

1 Accelerated KAZE Features

Building the nonlinear scale space. We use FED schemes for building a nonlinear scale space from which we can detect and describe features. By means of FED schemes, a nonlinear scale space can be built much faster than with any other kind of discretization scheme. Furthermore, FED schemes are extremely easy to implement and are more accurate than revious approaches such as *Additive Operator Splitting* (AOS) [7]. The main idea is to perform M cycles of n explicit diffusion steps with varying step sizes τ_j that originate from the factorization of a box filter:

$$\tau_j = \frac{\tau_{max}}{2\cos^2\left(\pi\frac{2j+1}{4n+2}\right)} \tag{1}$$

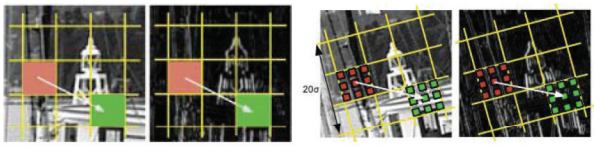
Feature Detection. We compute the determinant of the Hessian for each of the filtered images L^i in the nonlinear scale space. The set of differential multiscale operators are normalized with respect to scale, using a normalized scale factor that takes into account the octave of each particular image in the nonlinear scale space. Then, we search for maxima of the detector response in scale and spatial location.

$$L^i_{Hessian} = \sigma^2_{i,norm}\left(L^i_{xx}L^i_{yy} - L^i_{xy}L^i_{xy}\right) . \tag{2}$$

Feature Description. We propose a *Modified-Local Difference Binary* (M-LDB) that exploits gradient and intensity information from the nonlinear scale space. The LDB descriptor was introduced in [8] and follows the same principle as BRIEF [2], but using binary tests between the average of areas instead of single pixels for additional robustness. In addition to the intensity values, the mean of the horizontal and vertical derivatives in the areas being compared is used, resulting in 3 bits per comparison. Rotation invariance is obtained by estimating the main orientation of the keypoint as in KAZE, and the grid of LDB rotated accordingly. Instead of using the average of all pixels inside each subdivision of the grid, we subsample the grids in steps that are a function of the scale σ of the feature. The scale-dependent sampling in turn makes the descriptor robust to changes in scale.

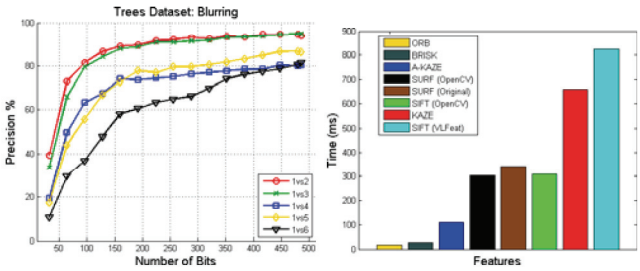
2 Evaluation

We present an extensive evaluation based on the standard Oxford benchmark [5] that shows the excellent compromise between speed and performance of our approach compared to state-of-the-art methods such as BRISK, ORB, SURF, SIFT and KAZE. While A-KAZE is more expensive to compute than BRISK and ORB, it is faster than SURF, SIFT and KAZE. More specifically, A-KAZE is several orders of magnitude faster than KAZE while providing similar or even better performance in some scenarios. In addition, due to the use of binary descriptors the matching step can be computed very efficiently using the Hamming distance.



(a) LDB binary test (b) M-LDB binary test

Figure 1: LDB [8] and proposed M-LDB binary tests between grid divisions around a keypoint, shown for the intensity and the gradients in x. M-LDB includes rotation and subsampling that depends on the scale.



(a) (b)

Figure 2: Descriptor and timing evaluation. (a) Precision vs # Bits (b) Timing evaluation for the joint method (detection and description) considering 1000 features from the first image of the Graffiti dataset.

Open Source Code. The code of the A-KAZE features is implemented in C++ using some functionalities from the OpenCV library. An open source implementation can be downloaded from: `www.robesafe.com/personal/pablo.alcantarilla/kaze.html`.

[1] P. F. Alcantarilla, A. Bartoli, and A. J. Davison. KAZE features. In *Eur. Conf. on Computer Vision (ECCV)*, pages 214–227, 2012.

[2] M. Calonder, V. Lepetit, M. Özuysal, T. Trzinski, C. Strecha, and P. Fua. BRIEF: Computing a local binary descriptor very fast. *IEEE Trans. Pattern Anal. Machine Intell.*, 34(7):1281–1298, 2011.

[3] S. Grewenig, J. Weickert, C. Schroers, and A. Bruhn. Cyclic Schemes for PDE-Based Image Analysis. Technical Report 327, Department of Mathematics, Saarland University, Saarbrücken, Germany, March 2013.

[4] S. Grewening, J. Weickert, and A. Bruhn. From box filtering to fast explicit diffusion. In *Proceedings of the DAGM Symposium on Pattern Recognition*, pages 533–542, 2010.

[5] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Intl. J. of Computer Vision*, 65(1-2):43–72, 2005.

[6] S. Wang, H. You, and K. Fu. BFSIFT: A novel method to find feature matches for SAR image registration. *IEEE Geosci. Remote Sensing Lett.*, 9(4):649–653, 2012.

[7] J. Weickert, B. ter Haar Romeny, and M. A. Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Trans. Image Processing*, 7(3):398–410, 1998.

[8] X. Yang and K. T. Cheng. LDB: An ultra-fast feature for scalable augmented reality. In *IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR)*, 2012.

From Bikers to Surfers: Visual Recognition of Urban Tribes

Iljung S. Kwak[1]
iskwak@cs.ucsd.edu
Ana C. Murillo[2]
acm@unizar.es
Peter N. Belhumeur[3]
belhumeur@cs.columbia.edu
David Kriegman[1]
kriegman@cs.ucsd.edu
Serge Belongie[1]
sjb@cs.ucsd.edu

[1] Dept. of Computer Science and Engineering
University of California, San Diego, USA.
[2] Dpt. Informática e Ing. Sistemas - Inst. Investigación en Ingeniería de Aragón.
University of Zaragoza, Spain.
[3] Department of Computer Science
Columbia University, USA.

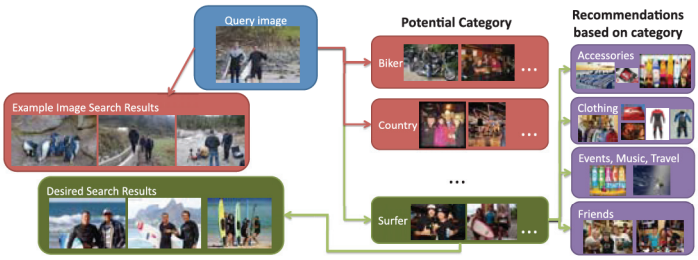


Figure 1: The social groups influence the appearance of their members. This work leverages this intuition to classify images of groups of people into social categories. Automatic urban tribe classification can improve recommendation systems and user experience with social media, and image search engines can take advantage of this classification and provide more meaningful search results.

Motivation. In 1985 Maffesoli described *urban tribes* [3] as a group of people who have similar visual appearances, personal style, and ideals. Visual cues shared by tribe members provide the basis for our work; members from the same urban tribe are expected to look more similar than members of different tribes. As shown in Fig 1, with automatic classification of urban tribes more inspiring image searches can be conducted; more relevant advertisements can enhance the web experience of both businesses and consumers; social networks can provide better recommendations [6]. Unfortunately, this categorization problem is difficult because of the ambiguous nature of social categories and the high intraclass variance. Social categories can evolve and fracture into separate groups; individuals within groups may exhibit features of multiple urban tribes or certain individuals may not present a visually salient style at all.

Contributions. This work attempts to recognize the content of an image from a social perspective, in line with the growing area of social signal processing [4]. A key element common to our work and related works is the need for both global image statistics as well as more semantic individual level attributes. Many recent works recognize the presence or absence of facial attributes, for instance of facial attributes to perform higher level tasks, such as face verification [2]. Rather than approaching this problem of social group image categorization by classifying isolated individuals (as most recent fashion/style analysis works do [5]), we focus on calculating meaningful group features and models. For example, the importance of group structure has been previously studied and shown useful to detect gender, age and even family relationship [1]. Following these ideas, we present a novel recognition pipeline and evaluate different modeling approaches, following common frameworks used for recognition tasks. We present the Urban Tribes dataset, with around 100 labeled images per class from 11 different classes, to facilitate further research on social categorization of group pictures[1]. Fig. 2 and Table 1 show examples from our dataset.

Results. We set up classification experiments where training and testing images were randomly selected from each of the categories for 50 different iterations. For each experiment, a fixed number of the images from each class are used for learning the models and the rest of images are used for testing. An image is correctly classified if the most likely group label matches the ground truth labels. In Fig. 3, we provide confusion matrices for classification results using two different modeling approaches.

[1] http://vision.ucsd.edu/content/urban-tribes



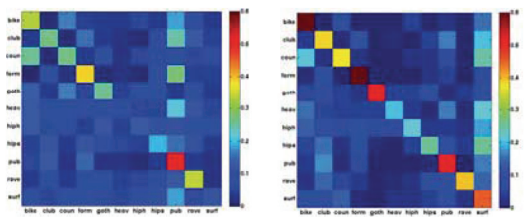
Figure 2: Examples of social groups in our Urban Tribes dataset. The images show a wide range of inter-class and intra-class variability. More details can be seen at the dataset website.

Label	# im	# people	Label	# im	# people	Label	# im	# people
biker	114	443	hip-hop	90	253	club	100	365
country	107	347	hipster	102	288	formal	103	414
goth	99	226	raver	116	305	casual/pub	125	459
heavy-metal	102	266	surfer	100	333			

Table 1: Summary of Urban Tribes dataset.

Both cases show that our techniques have promising recognition results. These experiments show that it is possible to extract semantic meaning from social media group photos, opening opportunities for the previously mentioned applications. Future work can include incorporating semantic attributes to improve the classification performance and analyzing the contribution of different types of semantic attributes.

[1] A. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009.
[2] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009.
[3] M. Maffesoli. *The Time of the Tribes: The Decline of Individualism in Mass Society*. Sage Publications, 1996.
[4] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.
[5] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.
[6] J. Yu, X. Jin, J. Han, and J. Luo. Collection-based sparse label propagation and its application on social group suggestion from photos. *ACM Trans. Intell. Syst. Technol.*, 2011.



(a) (b)

Figure 3: Confusion matrix for classification results obtained with (a) Bag of Parts modeling (BoP_{30k}) and (b) SVM classification framework (SVM_8), using 80% of the data for training. The rows show results in alphabetical order of the labels.

Ultra-wide Baseline Aerial Imagery Matching in Urban Environments

Hani Altwaijry
haltwaij@cs.ucsd.edu
Serge Belongie
sjb@cs.ucsd.edu

Department of Computer Science and Engineering
University of California, San Diego
California, USA

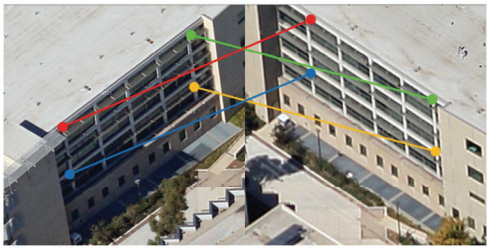


Figure 1: A pair of aerial images with correspondences is shown.

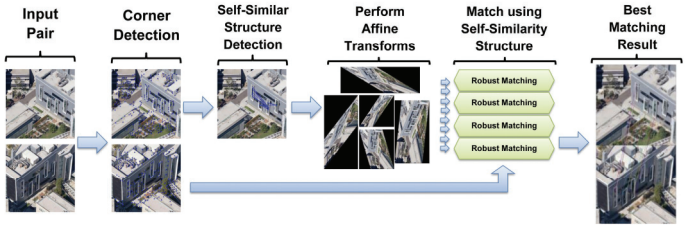


Figure 2: An overview of the matching pipeline is shown here.



Figure 3: An example pair is matched using our method, and the recovered homography is used to stitch the two images together.

Correspondence matching is a fundamental problem in computer vision, having many uses in structure from motion, stereo vision, image registration, pose-estimation, and others. Today, a large amount of aerial imagery is available online via mapping services such as Google Maps. If we were to pick any pair of images from two different aerial views (see Figure 1 for an example), and perform SIFT-based [3] correspondence matching, we would find ourselves with a large number of mismatches due to the large distortions between the images. Even when augmenting these methods with typical robust approaches such as RANSAC [1] and its variants, we would still fail at finding correct correspondences since RANSAC has difficulty calculating the correct model without a large ratio of correct matches to outliers. These difficulties – large distortions, and low ratio of correct matches to outliers – together render traditional methods ineffective.

In this paper, we consider the problem of correspondence matching for aerial imagery in urban environments. Our approach builds on multiple ideas in the literature. Namely, A-SIFT [7], patch-based methods [6], Generalized RANSAC framework [8], self-similarity [5], graph-based image matching [2], and geometric-invariance [4]. The main idea behind this work is to combine view-synthesis with multiple point correspondences under a RANSAC-based scheme. Robust model estimation is supported by self-similarity principles and graph-based modeling that drives the sampling process in a restricted manner that allows the correct model to be extracted. Each of these ideas was chosen to deal with specific problems that cause failures in the earlier approaches as we will now briefly describe.

An overview of our pipeline is shown in Figure 2. We begin by extracting square patches around detected Harris corners, which we then describe using the Histogram of Oriented Gradients. Our model assumes affine distortions, which leads us to synthesizing transformations to account for probable changes between the two images. We apply our transformations to one of the input pairs only, as follows:

$$Scale = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad Shear = \begin{bmatrix} 1 & Sh_x & 0 \\ Sh_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$Rotation = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A = Scale \times Shear \times Rotation, \qquad I_{S_x,S_y,Sh_x,Sh_y,\theta} = A \times I$$

where I is an image.

Buildings, in general, exhibit a lot of repeating patterns by design which in turn leads typical methods to fail. We leverage the repeating patterns by forming a graph over self-similar patches in one of the input images. Similar patches are found by comparing their HOG descriptors. The graph is then simplified by forcing each vertex to have a single neighbor in each radial bin around it. This then yields a connected component essential for our matching step.

During matching, patches are affine transformed and then allowed to match in a many-to-many fashion. Afterwards, our guided RANSAC approach uses the above connected component to enforce a spatial arrangement of the matches. This takes advantage of the fact that a given affine transform that yielded matches reveals more about an underlying transformation that allowed the matches to be made. We attempt to discover this by forcing spatial relationships as opposed to random sampling of probable matches, thereby increasing the probability of finding correct correspondences. A sample result is shown in Figure 3.

The main hurdles to the approach are its high computational complexity and the lack of support to multiple homographies. In our future work, we will pursue the following improvements: (1) reducing the number of affine transformations needed, (2) improving the graph operations, (3) improving the angular binning approach by including distance bins, and (4) including the support of multiple planes.

[1] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*.

[2] David Lavine, Barbara A. Lambird, and Laveen N. Kanai. Recognition of spatial point patterns. *Pattern Recognition*, 16(3):289 – 295, 1983.

[3] David G. Lowe. Object recognition from local scale-invariant features. *ICCV 1999*.

[4] Joseph L. Mundy and Andrew Zisserman, editors. *Geometric invariance in computer vision*. MIT Press, Cambridge, MA, USA, 1992.

[5] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR 2007*.

[6] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV 2012*.

[7] Guoshen Yu and Jean-Michel Morel. ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line 2011*.

[8] Wei Zhang and Jana Kosecka. Generalized RANSAC framework for relaxed correspondence problems. In *3DPVT 2006*.

FRIF: Fast Robust Invariant Feature

Zhenhua Wang
wzh@nlpr.ia.ac.cn
Bin Fan
bfan@nlpr.ia.ac.cn
Fuchao Wu
fcwu@nlpr.ia.ac.cn

National Laboratory of Pattern Recognition,
Institute of Automation,Chinese Academy of Sciences,
100190, Beijing, China

Local image feature is a fundamental component of many computer vision applications such as 3D reconstruction, image retrieval, object recognition and object categorization. The main goal is to find salient image points that can be repeatably detected under various image transformations and then construct distinctive and robust representations for them. Many methods have been proposed in the literature [2, 3, 4, 6]. However, it is still very challenging to obtain a high quality feature whilst maintaining a low computational cost.

This paper aims to tackle this problem by developing a novel Fast Robust Invariant Feature (FRIF). For feature detection, scale invariant and stable keypoints are selected in the scale space according to our Fast Approximated LoG (FALoG) filter responses. For feature description, distinctive binary descriptors are constructed by encoding both local pattern and inter-pattern information. By employing factorization and integral image, FALoG can be computed very fast. Meanwhile, the binary descriptor is constructed directly by intensity comparisons. Thus, both the detection and description of FRIF can be done very efficiently, making it suitable for real-time applications.

Fast Approximated LoG Detector We develop a scale-invariant feature detector based on the LoG function since it is found to be more stable than other derivative based functions in characteristic scale selection [5]. The proposed Fast Approximation of LoG (FALoG) filter assigns the surrounding areas with different weights according to their distances to the central point and keeps the DC response to be zero. As shown in Figure 1, a 9×9 FALoG filter is used to approximate LoG with $\sigma = 1.2$. To efficiently compute the response of FALoG filter, it is factorized into four rectangles with different weights. The response of each rectangle can be computed rapidly using integral image. By summing over all the responses of these four rectangle filters, we obtain the response of FALoG filter.

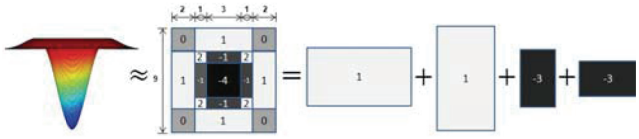


Figure 1: The approximation of LoG kernel.

The scale space is implemented in a way similar to BRISK [3]. To obtain scale invariant keypoints, we first compute the integral image for each octave and intra-octave. This can be done very fast since only one scan over the scale space is required. Next, we compute the 9×9 FALoG filter response for each pixel in the scale space separately, and a threshold T_F is applied to filter out weak responses. Then, for the remaining points, non-maximum suppression is performed so that only the extremas of FALoG responses in a local neighborhood over locations and scales are kept as potential keypoints. Finally, we refine the scale and location of each keypoint via quadratic function fitting.

Similar to LoG, FALoG filter has strong responses along edges, which are poorly localized and sensitive to noise. To filter out these edge responses, we compute the ratio of principal curvatures by the trace and determinant of Harris matrix. Keypoint candidates with ratios larger than T_L are considered as edges and are removed. Note that the Harris matrix is computed on the scale-space image.

Mixed Binary Descriptor The basic idea of our descriptor is to incorporate both the local pattern and inter-pattern information to improve matching performance of current stat-of-the-art binary descriptor, *e.g.* BRISK [3] and FREAK [1]. More specifically, the descriptor is constructed based on a modified BRISK pattern which has more overlapping areas than the original one. The local gradients of pattern locations and the inter-pattern intensity comparisons are combined to create the mixed binary descriptor.

Let $\mathcal{P}$ be the set of all the N pattern locations, for each pattern location $p_i = (x_i, y_i) \in \mathcal{P}$, a set of four points $\mathcal{S}(p_i) = \{s_{i,k}, k = 1,2,3,4\}$ are equally sampled on a circle of radius R centered at p_i. Let $I(x,\sigma)$ be the smoothed intensity of point x with sigma σ and θ be the local dominant orientation estimated by the average local gradients [1, 3]. For each rotated pattern location ${p_i}^\theta$, we compare the pair-wise intensities of its sampling points $s^\theta_{i,k} \in \mathcal{S}(p^\theta_i)$. Then, the descriptor is constructed by assembling all the test results into a binary string, each bit b of which corresponds to:

$$b = \text{sign}(\text{I}(s^\theta_{i,k}, \sigma_\text{i}) - \text{I}(s^\theta_{i,t}, \sigma_\text{i})), \tag{1}$$

$$\forall p^\theta_i \in \mathcal{P} \wedge s^\theta_{i,k}, s^\theta_{i,t} \in \mathcal{S}(p^\theta_i) \wedge k,t = 1,2,3,4 \wedge k \neq t.$$

As four points are sampled for each pattern location, the dimension of the descriptor is $N \times C^2_4 = 6N$ bits. It is worth noting that the intensity comparisons between sampling points $s_{i,k}$ is closely related to the local gradient operator, both of which consider the intensity differences between pairs of local samplings.

The above descriptor encodes the local pattern information into binary strings. We complement it with more global information, which is encoded by the inter-pattern intensity comparisons. Let the set $\mathcal{A}$ be all pairs of pattern locations:

$$\mathcal{A} = \{(p_i, p_j) \mid p_i, p_j \in \mathcal{P} \wedge i \neq j\}. \tag{2}$$

A binary string can be computed by comparing the intensities of all the rotated pairs (p^θ_i, p^θ_j) in a subset $\mathcal{B}$ of $\mathcal{A}$, each bit of which corresponds to $\text{sign}(\text{I}(p^\theta_i) - \text{I}(p^\theta_j))$. To form the subset $\mathcal{B}$, we tested three different selection criteria: 1) randomly select M pairs from $\mathcal{A}$. 2) select the shortest M pairs of $\mathcal{A}$. 3) select the longest M pairs of $\mathcal{A}$. Experiments show that the shortest pairs are more stable than others. This result is consistent with BRISK, which uses the short-distance pairs to build the descriptor. The difference is that, in our method only the shortest M pairs are used as a complementary part of the previous local gradient based binary descriptor.

By concatenating the two kinds of complementary binary strings together, we obtain our mixed binary descriptor. As the BRISK and FREAK descriptor are both 512 bits, we select $M = 512 - 6N$ to make the mixed binary descriptor the same dimension to them.

[1] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition*, pages 510 –517, 2012.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006.

[3] S. Leutenegger, M. Chli, and R.Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *International Conference on Computer Vision*, pages 2548–2555, 2011.

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[5] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, volume 1, pages 525–531, 2001.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *International Conference on Computer Vision*, pages 2564–2571, 2011.

Spatial Co-Training for Semi-Supervised Image Classification

Yi Hong
YHong@walmartlabs.com

@WalmartLabs
San Bruno, CA 94066, USA

Co-training [1] is a famous learning algorithm used when there are small amounts of labeled data and large amounts of unlabeled data, but it has limited applications in image classification due to the unavailability of two independent and sufficient representations of the images. In this paper, we propose a novel co-training algorithm in which these two independent and sufficient representations are automatically learned from the data. We call it as the spatial co-training algorithm (SCT). The main idea of SCT is to divide an image into two subregions and consider each of them as an independent representation. In SCT, the division of the image is firstly learned by an EM style algorithm on small amounts of labeled images, and finally relearned by a co-training style algorithm on many confident unlabeled images; while the classification of the image is performed jointly with the division of the image.

1 Problem definition

Our proposed algorithm is based on the bag-of-words model that represents each image as the histogram of its local image patches. In particular, the bag-of-words model performs to: (1) extract a collection of local descriptors such as SIFT [2] from the images; (2) quantize them as indexes; (3) and represent each image as the histogram of indexes of its local image patches. A lot of visual codebook learning algorithms have been proposed so far, and in this paper we employ the k-means clustering algorithm because of its simplicity and wide applications.

Given a set of labeled images $\{I_1,...,I_l\}$, we extract local sift descriptors densely from each image and express them as a matrix F:

$$F = \{f_{ij} | i = 1,...,w; j = 1,...,v\}, \quad (1)$$

where f_{ij} is the sift descriptor of the local image patch (i,j), w and v are the height and width of the matrix of the SIFT descriptor. To simplify our notations, we assume that all images have the same size and thus have the same values of w and v of the matrix F. Then we apply the k-means clustering algorithm to quantize these local sift descriptors into indexes and rearrange indexes from one image as a matrix C:

$$C = \{c_{ij} | i = 1,...,w; j = 1,...,v\}, \quad (2)$$

where $c_{ij} \in \{1,...,K\}$ is the visual code of the local image patch (i,j) and K is the codebook size. As illustrated in the introduction section, the basic idea of the proposed algorithm is to partition an image into two subregions and consider each of them as an independent representation. Now suppose that I_1^s and I_2^s are a partition of the image I that corresponds to a partition C_1^s and C_2^s of all visual codes in C:

$$C_1^s \bigcap C_2^s = \emptyset; \quad C_1^s \bigcup C_2^s = C, \quad (3)$$

we calculate the histograms of visual codes in both I_1^s and I_2^s as:

$$h_{1,k} = \frac{\sum_{(i,j)\in C_1^s} \delta(c_{ij},k)}{\sum_{(i,j)\in C_1^s} \mathbf{1}}; \quad h_{2,k} = \frac{\sum_{(i,j)\in C_2^s} \delta(c_{ij},k)}{\sum_{(i,j)\in C_2^s} \mathbf{1}}, \quad (4)$$

after normalized for $k = 1,...,K$; and $\delta(a,b) = 1$ if $a = b$, and $\delta(a,b) = 0$ if $a \neq b$. According to (4), each division (C_1^s, C_2^s) of the image will lead to a representation pair (h_1,h_2) of the image:

$$(C_1^s, C_2^s) \Rightarrow (h_1,h_2), \quad (5)$$

where both h_1 and h_2 are the histograms with K bins. Consider the overall number of divisions of each image is around wv that is too large to be processed, in this paper we restrict the division of the image as a vertical line. In this case, the partition of C at the position d is $(C_{1:w,1:d}, C_{1:w,d+1:v})$, and we simplify it as $(C_{1:d}, C_{d+1:v})$ together with their histogram representation pair as $(h_{1:d}, h_{d+1:v})$. Therefore, the candidate pool of all possible representation pairs is:

$$H = \{(h_{1:d}, h_{d+1:v}) | \ d = 1,...,v-1\}. \quad (6)$$

Based on the above notations, the proposed algorithm in this paper aims to learn good divisions $\{d_1,...,d_\ell,d_{\ell+1},...,d_{\ell+u}\}$ of both labeled and unlabeled images such that the co-training algorithm using $\{(h_{i,1:d_i}, h_{i,d_i+1:v}) | i = 1,...,\ell;\ell+1,...,\ell+u\}$ as inputs can succeed. Two points should be mentioned about the above definition: (1) different images may have different good values of d; (2) the above restriction of the division of the image may not be optimal, but our experimental results indicate that it works quite well.

2 Solutions

Fig.1 and Fig.2 are frameworks of our proposed algorithm.

- **Input**
  - A set $\{I_1,...,I_\gamma\}$ of labeled positive training images
  - A set $\{I_{\gamma+1},...,I_\ell\}$ of labeled negative training images
- **Process**
  (1) Set initial divisions $\{d_1,...,d_\gamma\}$ of the images
  (2) Loop:
  (a) $(svm^L, svm^R) \leftarrow$Fix $(d_1,...,d_\gamma)$ and train two one-class SVMs
  (b) $(d_1,...,d_\gamma) \leftarrow$Fix (svm^L, svm^R) and update divisions of the images
  End loop
  (3) Determine divisions $\{d_{\gamma+1},...,d_\ell\}$ of negative training images

Figure 1: The EM style algorithm to learning divisions of labeled images.

- **Input**
  - A set L of labeled training images
  - A set U of unlabeled images
- **Process**
  (1) Learn divisions $D = \{d_1,...,d_\ell\}$ of labeled images in L by EM style algorithm
  (2) Create a pool U' of examples by choosing u' examples at random from U
  (3) Loop for k iterations
  (a) Learn divisions $\{d'_1,...,d'_{u'}\}$ of unlabeled images in U'
  (b) Train a classifier F_1 on $\{h_{1,1:d_1},...,h_{\ell,1:d_\ell}\}$
  (c) Train a classifier F_2 on $\{h_{1,d_1+1:v},...,h_{\ell,d_\ell+1:v}\}$
  (d) Label p positive and n negative on which F_1 is most confident
  (e) Label p positive and n negative on which F_2 is most confident
  (f) Add these self-labeled images to L and their divisions to D
  (g) Update (svm^L, svm^R) by training two one-class SVMs on augmented D
  (h) $\ell = \ell + 2p + 2n$
  (i) Randomly choose $2p+2n$ examples from U to replenish U'
  End loop
  (4) Train a single view classifier F on the augmented training dataset L
  (5) Label all the rest of unlabeled images by F

Figure 2: The spatial co-training algorithm (SCT).

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *COLT*, 1998.

[2] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

Distribution Fields with Adaptive Kernels for Large Displacement Image Alignment

Benjamin Mears
bmears@cs.umass.edu
Laura Sevilla Lara
lsevilla@cs.umass.edu
Erik Learned-Miller
elm@cs.umass.edu

School of Computer Science
UMass Amherst
Amherst, MA

While region-based image alignment algorithms that use gradient descent can achieve sub-pixel accuracy when they converge, their convergence depends on the smoothness of the image intensity values. Image smoothness is often enforced through the use of multi-scale approaches in which images are smoothed and downsampled. Yet, these approaches typically use fixed smoothing parameters which may be appropriate for some images but not for others. Even for a particular image, the optimal smoothing parameters may depend on the magnitude of the transformation. When the transformation is large, the image should be smoothed more than when the transformation is small. Further, with gradient-based approaches, the optimal smoothing parameters may change with each iteration as the algorithm proceeds towards convergence.

We address convergence issues related to the choice of smoothing parameters by deriving a Gauss-Newton gradient descent algorithm based on distribution fields (DFs) and proposing a method to dynamically select smoothing parameters at each iteration. DFs have previously been used in the context of tracking [6]. In this work we incorporate DFs into a full affine model for region-based alignment and simultaneously search over parameterized sets of geometric and photometric transforms. We use a probabilistic interpretation of DFs to select smoothing parameters at each step in the optimization and show that this results in improved convergence rates.

Following the notation of Baker and Matthews [1], let $T(\mathbf{x})$ be an image containing a fixed region for which we want to find the corresponding region in $I(\mathbf{x})$, where $\mathbf{x} = (x,y)^{\mathrm{T}}$ is a column vector. We refer to $T(\mathbf{x})$ as the template image and $I(\mathbf{x})$ as the input image. Let $W(\mathbf{x};\mathbf{p})$ be the parameterized set of warps, where $\mathbf{p}$ are the parameters. The goal of region-based alignment is to find the $\hat{\mathbf{p}}$ that minimizes some distance measure between $T(\mathbf{x})$ and $I(\mathbf{W}(\mathbf{x};\hat{\mathbf{p}}))$.

One of the early image alignment algorithms was the gradient-based Lucas-Kanade (LK) algorithm [3]. In the LK method, the feature space consists of intensity values and the similarity measure is the L2 distance. Our method is inspired by the LK algorithm, but we replace the image and template with their DF representations. The basic idea of a DF is to represent a region in an image as a normalized histogram, i.e., a probability distribution, over feature values at each pixel. In this work we use grayscale intensity values as the feature values although other features could be used instead (e.g. edge intensities, RGB values for color images, etc.). The simplest DF consists of probability distributions over binned intensity values where each probability distribution is degenerate and is given by

$$D(I,\mathbf{x},f) = \begin{cases} 1 & \text{if } I(\mathbf{x}) \in \text{bin } f \\ 0 & \text{otherwise,} \end{cases} \qquad (1)$$

where $D(I,\mathbf{x},f)$ is the value of the DF for image I at position $\mathbf{x}$ and bin f.

Using Eq. (1) alone to represent an image provides few additional benefits over using the image itself. Indeed, when the number of bins equals the number of intensity values, the representation contains the same information as the image. Additional benefits can be gained though if the DF is "smoothed" to "spread" the information in the image. In particular, the DF can be convolved with a three-dimensional Gaussian filter with a standard deviation of σ_{xy} in the spatial directions and σ_f in the feature space dimension. By convolving with a Gaussian filter, some degree of uncertainty is allowed in both the location and value of an image pixel. Like image blurring, convolving a DF with a Gaussian filter spreads information about intensity values to neighboring pixels, but does so with a smaller loss of information.

There are various trade-offs that need to be considered when choosing the σ_{xy} and σ_f values used for smoothing the DFs for alignment. For instance, a larger σ_{xy} value may allow for a larger basin of attraction but may result in a less precise final alignment. And if the σ_{xy} value is chosen to be too large or too small, it can cause the algorithm to diverge.

Rather than choose fixed σ values, the values can be chosen automatically based on the current location in the search space. In this work, the σ values are chosen using the probabilistic view of DFs by maximizing the log likelihood of the current warped input image under the DF of the template image. Treating the DF of one image as an independent pixel model, this is defined to be the sum of the log probabilities of each pixel of the current warped input image under the corresponding probability distribution of the template image's DF. Our method is similar to the approach used by Narayana *et al.* for choosing the pixelwise kernel variances in their background subtraction algorithm [4, 5].

In our approach, the log likelihood of the current warped input image, $I' = W(\mathbf{p})$, under the DF of the template, smoothed using the parameters σ_{xy} and σ_f, is given by

$$l(\sigma = \{\sigma_{xy},\sigma_f\}|T,I',R) = \sum_{\mathbf{x}\in R} \log\left(D_\sigma(T,\mathbf{x},\text{bin}(I'(\mathbf{x}))\right), \qquad (2)$$

where bin is the binning function that takes an intensity value and maps it to the appropriate histogram bin.

Since truncated Gaussian kernels are used for efficiency to smooth the DFs, it is possible that some entries of a DF are zero. To deal with the problem of zero probabilities (in which a single outlier can cause the likelihood to be zero), we replace $\log\left(D_\sigma(T,\mathbf{x},\text{bin}(I'(\mathbf{x}))\right)$ in Eq. (2) with $\log\left(\max\left(.0001, D_\sigma(T,\mathbf{x},\text{bin}(I'(\mathbf{x})))\right)\right)$. At each iteration in our method, an exhaustive search is performed over a finite set of σ_{xy} and σ_f values and the σ_{xy} and σ_f values that maximize Eq. (2) are used to convolve the two DFs.

At each iteration of a forward-inverse compositional Gauss Newton optimization, we use the above method to dynamically select the σ values used to blur the DFs of the current warped input image and template. Further, similar to the simultaneous inverse compositional (SIC) algorithm described by Baker and Matthews [2], we extend the algorithm to also search over bias and gain parameters. We achieve impressive convergence results compared to other existing algorithms and also show that the adaptive kernel parameters produce convergence rates better than or equal to the best convergence rates produced by any of a large set of fixed parameter values.

[1] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004.

[2] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical Report CMU-RI-TR-03-35, CMU Robotics Institute, December 2003.

[3] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[4] M. Narayana, A. Hanson, and E. Learned-Miller. Background modeling using adaptive pixelwise kernel variances in a hybrid feature space. In *CVPR*, 2012.

[5] M. Narayana, A. Hanson, and E. Learned-Miller. Improvements in joint domain-range modeling for background subtraction. In *BMVC*, 2012.

[6] L. Sevilla-Lara and E. Learned-Miller. Distribution fields for tracking. In *CVPR*, 2012.

Fractal Approximate Nearest Neighbour Search in Log-Log Time

Martin Stommel
mstommel@tzi.de
Stefan Edelkamp
edelkamp@tzi.de
Thiemo Wiedemeyer
wiedemeyer@informatik.uni-bremen.de
Michael Beetz
beetz@cs.uni-bremen.de

Institute for Artificial Intelligence
University of Bremen
Am Fallturm 1
28359 Bremen
Germany

Nearest neighbour searches in the image plane are among the most frequent problems in a variety of computer vision and image processing tasks. They can be used to replace missing values in image filtering, or to group close objects in image segmentation, or to access neighbouring points of interest in feature extraction. In particular, we address two nearest neighbour problems:

The *nearest neighbour problem* is usually stated independently of the application as returning the point $p \in S, S = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ that minimises the Euclidean distance $||p-q||_2$ to a query point $q = (x,y)$. The simple solution of a linear scan comprises a comparison of q to all elements of S, which is too time-consuming for most applications, especially those with real-time requirements.

If the nearest neighbour $p \in S$ must be found for every coordinate $q \in I$ of an image $I - \{(0,0),(0,1),(0,2),\ldots,(W,H)\}$ of width W and height H, we obtain the *all nearest neighbours problem*. This problem occurs frequently in modern saliency based approaches, where only robustly detectable image regions are processed (for example in SIFT).

In this paper, we introduce an approximate solution to solve these problems that is based on using a space filling curve. The central idea of the proposed approach is to map the image plane to one dimension using the Hilbert curve (Fig. 1). The nearest-neighbour problem is then solved efficiently in one dimension and mapped back to the 2D-plane.

The proposed method answers the two types of nearest neighbour problems in three steps each: At first, the mapping between 2D and 1D-coordinates is computed. For the all nearest neighbour problem, the set of keypoints S must be written as an array. The nearest neighbour assignment can then be done in two passes throught the array. For the k-nearest neighbour problem, the set of keypoints S must be stored in a priority queue. The neighbours of a query point q can then be found by using the successor function of the queue.

The method has a precomputation time of $O(n)$ for adapting to a fixed image size. Depending on the problem, an additional precomputation time of $O(C+n)$ (all nearest neighbours) or $O(n \lg \lg C)$ (k-nearest neighbours) is needed to adapt to a certain set of key points. The query time is then $O(1)$ or $O(\lg \lg C + k)$, respectively. This is an advantage over a balanced tree (with runtime $\lg n$) if $n > \lg C$ (the latter being a small number for common image formats).

Our experiments show that our method yields a compact and visually meaningful approximation of the Voronoi diagram in the image plane (Fig. 2), which is sufficient for many applications. For 50% of the queries, our method yields the exact result. In a practical example of finding all nearest neighbours of a set of keypoints, our method was 9–18 times faster than a kd-tree.

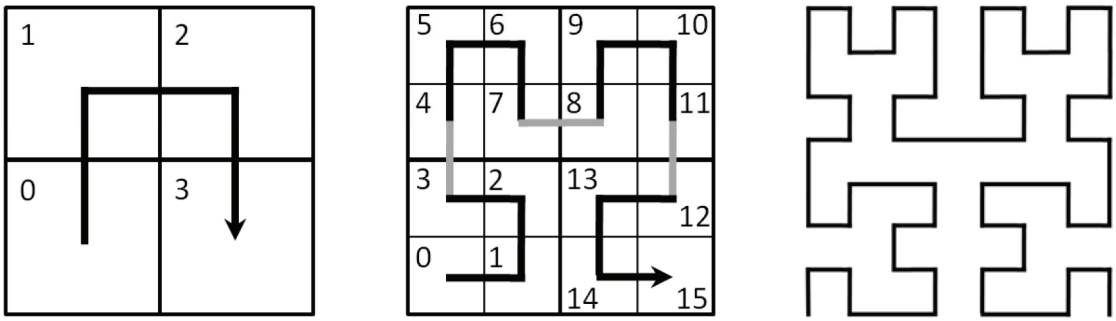


Figure 1: Hilbert curve for a 2×2 (on the left), 4×4 (middle) and 8×8 image (on the right). Pixels are counted along the curve. The second curve (and higher resolutions) can be computed (recursively) from the first one by replacing each corner by a rotated and reflected version of the basic U-shaped pattern.

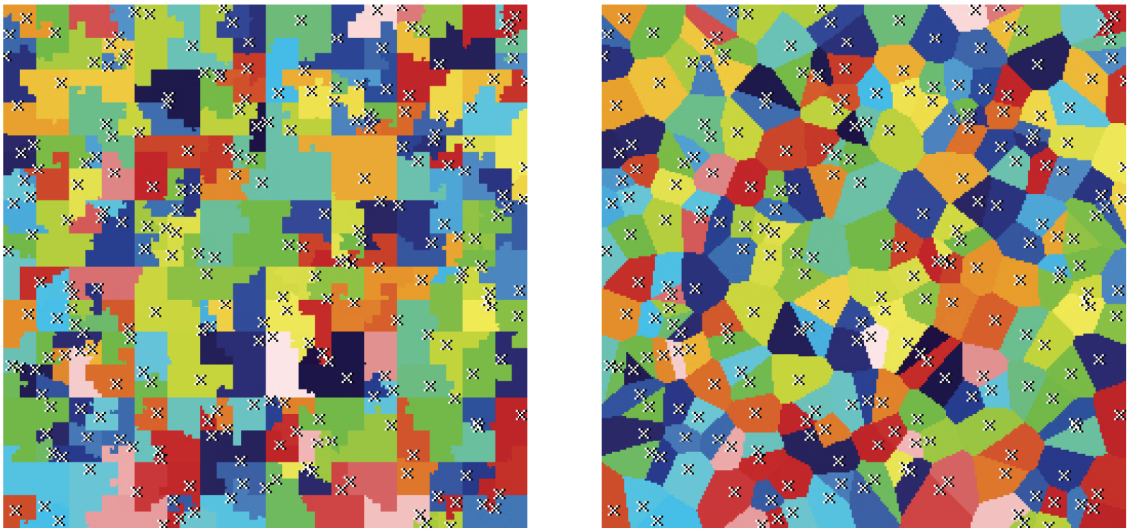


Figure 2: All nearest neighbour assingment using the proposed approximative method (left image) and an exact search (right image) for 240 keypoint (marked by crosses). The resulting cells are coloured randomly but consistent over both diagrams.

Flexible and User-Centric Camera Calibration using Planar Fiducial Markers

Shreyansh Daftry, Michael Maurer
{daftry,maurer}@icg.tugraz.at
Andreas Wendel, Horst Bischof
{wendel,bischof}@icg.tugraz.at
Institute for Computer Graphics and Vision,
Graz University of Technology, Austria

Recently, 3D vision methods have demonstrated increased robustness and resulted in high quality models. Hence multi-view modeling appears more often in industrial projects targeted at large-scale scenes such as Google Earth and Microsoft Bing. Combined with the emergence and availability of affordable high resolution cameras, this development has made 3D vision methods and tools popular not only among researchers, but also lead to the participation of an ever increasing population of end-users, who are not experts in computer vision. Though the advent of sophisticated multi-view stereo systems used in these tools have enabled the creation of visually appealing models, the geometric fidelity mainly depends on the accuracy of the camera parameters as shown in Fig. 1. Consequently, there is a need to increase the flexibility of existing calibration procedures used to estimate the camera intrinsics in image-based modeling, especially when inexperienced end users are involved.

Given its crucial role with respect to overall precision, camera calibration has been a well studied topic over the last two decades. However, in most of the calibration literature, a strict requirement on the target geometry and a constraint to acquire the entire calibration pattern has been enforced. This has recently become a matter of question [3]. Other issues like robust detection of correspondences has also received little attention as most of the standard routines assume that the correspondences have already been found. The poor quality of gathered data, owing to the user's inexperience in calibration procedures (e.g. failure to cover the pattern along the corners of the image) and bad illumination are often sources of inaccuracy as shown in Fig. 2. It is these considerations that motivate our contribution.

The goal of this paper is to present and advocate the use of planar paper based fiducial markers as a target pattern to put together a flexible and reliable calibration pipeline, thereby easing the efforts for the end user. Our contribution is three-fold: First, to improve the accuracy of feature localization and correspondence with the use of simple printed markers having unique extractable ID's and sub-pixel accuracy. Second, to provide sensitivity to partial occlusions of the calibration object, resulting due to bad or uneven illuminations or partially clipped patterns, by releasing the critical requirement to process the entire pattern. Finally, we extend the fidelity of calibration consistency to even wide angle applications and highly distorted lenses by enabling a well distributed set of control points, especially along the fringes of the image and thereby eliminating the need for an elaborate procedure for accurate parametrization.

Our calibration routine follows the basic principles of planar target based calibration [4] and thus requires corresponding points to be imaged in several views. However, we use marker patterns printed on several sheets of paper and laid on the floor in an approximate grid pattern (see Fig. 3) to obtain those correspondences. There is no strict requirement for all markers to be visible in the captured images. To ensure that each marker is classified with a high degree of confidence and to eliminate

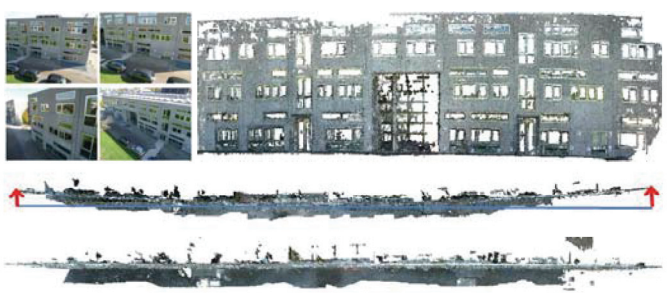
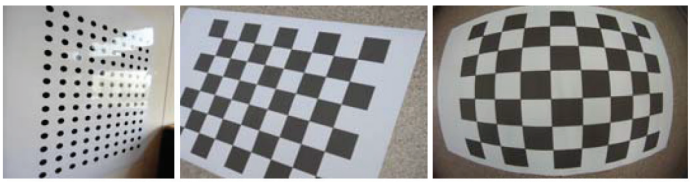
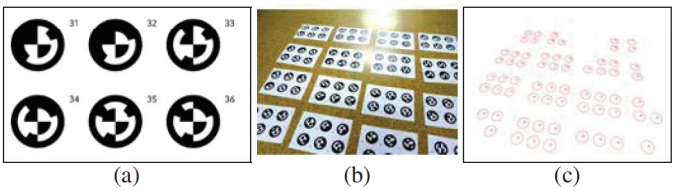


Figure 1: Reconstruction of a facade shows that though visually correct in appearance, geometric inconsistencies (significant bends along the fringes) are prevalent when using calibration and undistortion results from OpenCV (middle). In comparison, using accurate camera parameters delivered by the proposed method results in a straight wall (bottom).



(a) (b) (c)

Figure 2: Images provided by users which *(a)* have uneven illumination or *(b)* are partially clipped serve as common failure points for calibration methods having a critical requirement on pattern geometry. *(c)* On the other hand, if images are taken correctly the critical corners of the image (highest distortion) are never covered.



(a) (b) (c)

Figure 3: (a) Fiducial markers: Each marker encodes a unique ID. (b) A typical calibration image (partially corrupted due to uneven illumination) arbitrarily arranged in a 4x4 grid on the floor. (c) Reliably extracted markers with positions and IDs.

any false positives in case of blurred or otherwise low-quality images, a robust marker detection method is employed. An initial estimate of lens distortion parameters attempts to minimize the reprojection error of extracted feature points based on homographies between the individual views. We follow the approach of Irschara et al. [2], thereby searching for a constant, but unknown focal length f determining the calibration matrix K. Subsequently, bundle adjustment is applied to perform a non-linear optimization of the intrinsics (f, u_0, v_0) and radial distortion (θ).

We comprehensively validate the consistency in estimation of camera calibration parameters, and the robustness to images corrupted by noise or human error with our proposed method using a set of off-the-shelf cameras. The experiments are performed in an uncontrolled environment to simulate conditions similar to that experienced by a common user. The OpenCV Camera Calibration toolbox [1] is used as our primary reference as it represents the most popular method for calibration outside of computer vision and photogrammetry. Pixel reprojection errors and ground truth errors are significantly lowered by our method, even though paper-printable and easy-to-use targets are employed. In addition, we show both qualitative and quantitative benefits of increased flexibility in camera calibration towards multi-view reconstruction of large scale scenes. Empowered with an easy to use GUI with user-feedback facility, the proposed calibration method has been made publicly available at `http://aerial.icg.tugraz.at` so that a wide audience can benefit from our findings.

[1] Gary Bradski, Adrian Kaehler, and Vadim Pisarevsky. Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal*, 9(2):119–130, 2005.

[2] Arnold Irschara, Christopher Zach, and Horst Bischof. Towards wiki-based dense city modeling. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[3] Jean-Marc Lavest, Marc Viala, and Michel Dhome. Do we really need an accurate calibration pattern to achieve a reliable camera calibration? In *European Conference on Computer Vision (ECCV)*. 1998.

[4] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(11):1330–1334, 2000.

Large-scale Learning of Sign Language by Watching TV (Using Co-occurrences)

Tomas Pfister[1]
tp@robots.ox.ac.uk
James Charles[2]
j.charles@leeds.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] School of Computing
University of Leeds
Leeds, UK

We present a framework that automatically and quickly learns a large number of signs from sign language-interpreted TV broadcasts by exploiting supervisory information available in the subtitles.

Our contributions are: (i) we show that, somewhat counter-intuitively, mouth patterns are highly informative for distinguishing words in a language for the Deaf, and their co-occurrence with signing can be used to significantly reduce the correspondence search space; and (ii) we develop a multiple instance learning method using an efficient discriminative search, which determines a candidate list for the sign with both high recall and precision.

The previous approach of Buehler *et al.* [2] for learning signs relies on complex features and a computationally expensive, application-specific learning framework. This has hindered the large scale application of this method. In this paper we describe a method that is much simpler and computationally lighter.

Motivation. TV programmes in many countries across the world are now routinely broadcast with both subtitles and an overlaid signer translating to the Deaf audience (Fig. 2). Our aim is to use this material to learn *signs* corresponding to English *words* in the subtitles [2, 3]. We use this continuous and rich source of training material to build a database of word-sign pairs for a large number of signs and signers. The vision is that this database can later be used to train a large-scale person-independent sign language to text translator.

Summary of method. We cast the problem as one of Multiple Instance Learning (MIL), where the training data are visual descriptors (hand trajectories) with weak supervision from subtitles. We proceed in three steps: (i) the search space for correspondences is significantly reduced by exploiting lip and hand motion co-occurrences to filter away irrelevant intervals of the temporal sequences; (ii) candidates for the signs are obtained using an efficient discriminative search over all remaining sequences by casting each candidate as a classifier for the positive and negative sequences; and finally (iii) these candidates are then selected or rejected using the MIL support vector machine framework (MI-SVM) [1]. Fig. 1 illustrates the processing steps and Fig. 3 shows example results.

We demonstrate the method on videos from BBC TV broadcasts, and achieve higher accuracy and recall than previous methods, despite using much simpler features.

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[2] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. CVPR*, 2009.

[3] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *Proc. CVPR*, 2009.

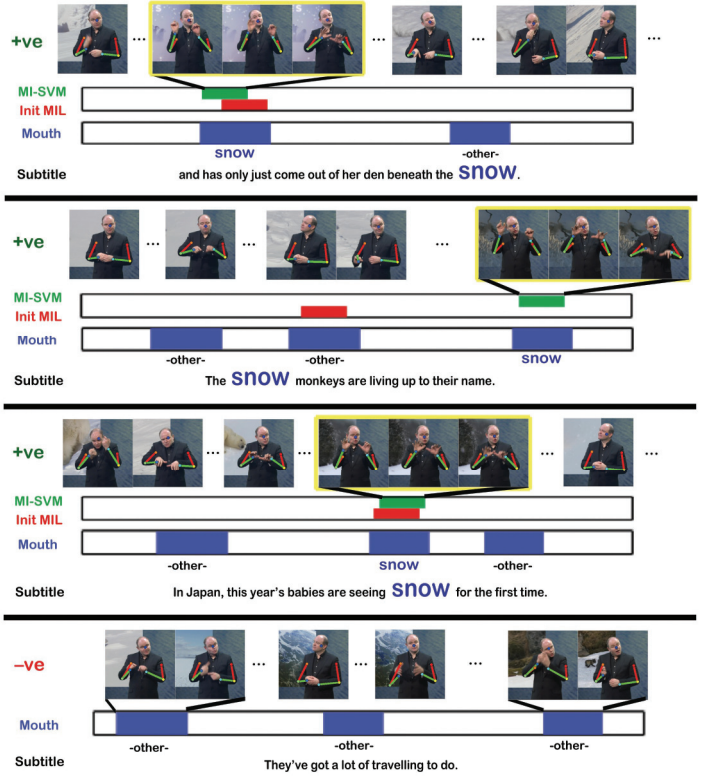


Figure 1: **Learning signs from co-occurrences of subtitle text, mouth and hand motion.** The top three rows are positive subtitle sequences which contain the text word and sign for 'snow'. The final row is an example of a negative subtitle sequence which does not contain 'snow'. Signs are learnt from this weakly aligned and noisy data. A fixed size temporal window is slid across the frames in which mouth motion occurs (blue). The rest of the sequence can be ignored, thus reducing the temporal search space. Candidate signs are proposed by a discriminative MIL search using temporal correlation. A subset of these candidates (red) are used to initialise a MI-SVM, resulting in the final correspondence matches (green). The red and green lines on the signer show the detected limbs and head.

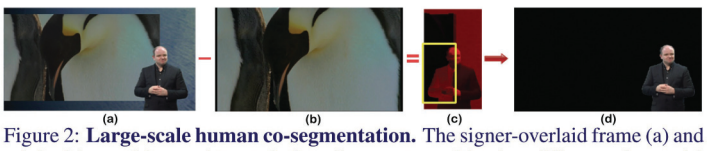


Figure 2: **Large-scale human co-segmentation.** The signer-overlaid frame (a) and original frame (b) are subtracted after alignment, resulting in a difference image (c) that is used as a constraint (yellow) in the segmentation. (d) shows the final result.



Figure 3: **Example output videos for the signs "snow" (top) and "vision" (bottom)** performed by two different signers and learnt automatically.

Detection of Abnormal Fish Trajectories Using a Clustering Based Hierarchical Classifier

Cigdem Beyan
http://homepages.inf.ed.ac.uk/~s1064447/
Robert B. Fisher
http://homepages.inf.ed.ac.uk/rbf/

Inst. Perception Action and Behaviour
School of Informatics
The University of Edinburgh, UK

Automatic analysis of animal and insect behaviour using computer vision techniques is a growing research area with many interesting studies. The traditional way of analyzing the behaviour of animals is by visual inspection of human observers which is very time consuming and also limits the size of data that can be processed. In this study, we present a novel framework for abnormal behaviour detection and especially focus on fish trajectories. Fish behaviour analysis is a fundamental research area in marine biology as it is helpful to detect environmental changes by detecting abnormal fish patterns and to detect new behaviours of fish. Detecting abnormal trajectories is useful when the system does not have any prior knowledge about the data which especially happens in the real-world data due to the uncontrolled naturalness.

When we compare fish trajectory datasets from underwater videos with the other abnormal behaviour detection datasets (traffic surveillance, human abnormal trajectory detection etc.), there are certain differences:

- Fish in the open sea can freely move in 3 dimensions hence there are no defined rules or roads such as exist in a traffic surveillance scenario.
- Fish are not goal-oriented which produces highly complex trajectories in contrast to people or vehicles.
- Fish usually make erratic movements due to currents in the water which increases the complexity of trajectories and also makes encoding the behaviour more difficult than is in human or animal behaviour recognition.

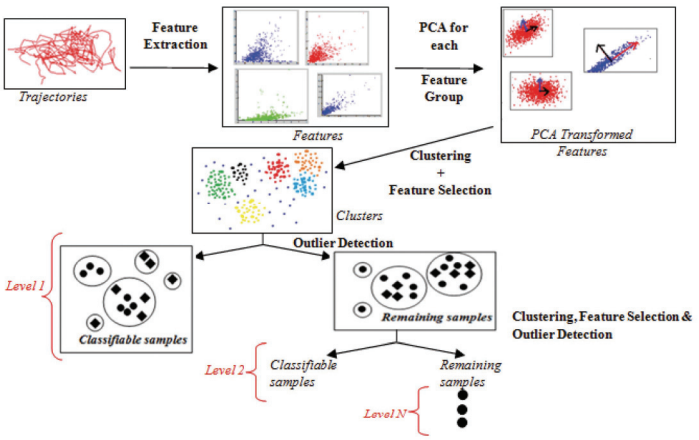


Figure 1: Overview of Proposed Method

In this study, we present an approach to detect abnormal fish trajectories using a hierarchical classifier. Normal fish trajectories are defined as the trajectories which contain frequently observed behaviours while abnormal trajectories are defined as outliers or rare trajectories. Clustered and labelled data are used together and the hierarchy is automatically built using similarity of data instead of using a taxonomy between features or classes as is common in the literature. Different than the studies which use the same feature space for classification, we use different feature sets at different levels of the hierarchy which allows more specific features to be used once the data focuses onto specific subclasses. The construction of hierarchy and new trajectory classification using the hierarchy are described in detail in the paper. Figure-1 also shows the feature extraction, hierarchy construction and the component of the proposed method which is based on clustering, feature selection and outlier detection.

Features: Curvature scale space based features [2], moment descriptors [3], velocity and acceleration based features, turn based features [4], centred distance function [2], vicinity features [5], loop features, fish pass by features (using the image segmentation given in Figure-2), features based on normalized size of bounding box, and features based on displacement on the location (using the image segmentation given in Figure-2).

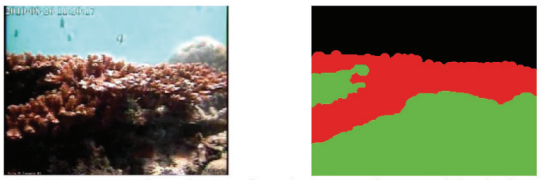


Figure 2: Segmented regions of underwater image; black for open sea, red for above the coral and green for under coral.

Altogether 776 features are obtained in the feature extraction step. As a post processing step of feature extraction, PCA is applied to the data from each feature group individually to get rid of from correlation between the features and handle the over-training. This left 179 features.

Outlier detection is similar to [6]. Present two types of abnormal (outlier) trajectories: those located in small clusters and those that exist in dense clusters but deviate from other trajectories in its cluster.

3102 trajectories (3043 normal, 59 abnormal) were used. The most usual and frequent behaviours in the dataset are freely swimming fish in open sea and hovering fish on the coral (Figure 3a-b) which represent normal behaviours. Abnormal trajectories are: i) fish suddenly (in one frame) diving under the coral, ii) fish suddenly (in one frame) changing direction (predator avoidance, Figure 3c), iii) fish diving quickly between the coral branches when frightened or to hide from predators, and iv) aggressive fish which is moving fast. A trajectory that has normal and abnormal segments is assumed as abnormal.

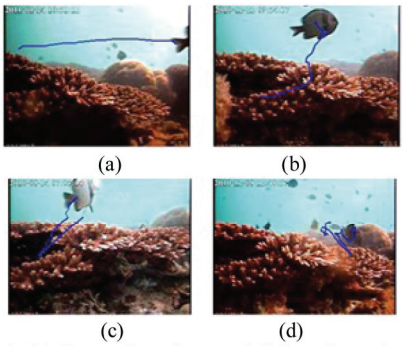


Figure 3: (a-b) Examples of normal fish trajectories, (c-d) Examples of abnormal (rare) fish trajectories.

The proposed method has highest abnormal trajectory detection rate compared to state of art methods of abnormal trajectory detection (fish and in general) while also the best method in overall. *t*-tests show that it is significantly better than all methods. Moreover, the proposed hierarchical classifier is not limited to fish trajectory analysis and is a general framework for classification of imbalanced datasets as shown on popular imbalanced datasets.

[1] B. J. Frey and D. Dueck, Clustering by Passing Messages Between Data Points, *Science*, 972-976, 2007.

[2] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, View-invariant Motion Trajectory based Activity Classification and Recognition, *Multimedia Systems*, 12(1):45-54, 2006.

[3] T. Suk, J. Flusser, Graph method for generating affine moment invariants, In *Proc. ICPR*, 192-195, 2004.

[4] X. Li, W. Hu, and W. Hu, A coarse-to-fine strategy for vehicle motion trajectory clustering, In *Proc. ICPR*, pages 591-94, 2006.

[5] M. Liwicki, and H. Bunke, HMM-Based On-Line Recognition of Handwritten Whiteboard Notes, *10th Int. Workshop on Frontiers in Handwriting Recognition*, pages 595–599, 2006.

[6] N. Anjum, and A. Cavallaro, Multifeature Object Trajectory Clustering for Video Analysis, *IEEE Trans. on Circuit and Systems for Video Technology*, 18(11):1555-1564, 2008.

Early Recurrence Improves Edge Detection

Xun Shi
shixun@cse.yorku.ca
Bo Wang
wangbo.yunze@gmail.com
John K. Tsotsos
tsotsos@cse.yorku.ca

Department of Computer Science and
Centre for Vision Research (CVR)
York University
Toronto, Ontario, Canada

This paper proposes a computational model of early recurrence for edge detection. The work is motivated by studies of primate vision. Features are perceived through the two pathways with different speeds: the dorsal pathway processes information more rapidly than that of the ventral pathway. It is then likely that results formed at higher-dorsal region influence ventral perception via early recurrent connections [1].

The two pathways are modeled to compute different edge features using 2D Gabors: the ventral layer of V1 computes fine-scale edges and the dorsal layer of V1 computes coarse-scale edges. Higher dorsal area MT integrates V1 energy and sends results to modulate ventral V1 (Fig. 1). Modulated ventral V1 responses are then inhibited by the center-surround self-inhibition [3] to form contour. The whole process can be defined as:

$$R^{V1v} = H(E^{V1v} \cdot Inh_{re}^{MT} - \alpha Inh_{cs}^{V1v}) \quad (1)$$

where R^{V1v} denotes ventral V1 representation. $H(s) = max(s,0)$ is a half-wave rectification function. E^{V1v} denotes ventral V1 responses to feed-forward image stimuli, which have been commonly deemed as the edge representation in most existing works. Inh_{re}^{MT} is the early recurrent inhibition generated from area MT. Inh_{cs}^{V1v} denotes the center-surround self-inhibition, and α is a weighting factor.

The inhibition term Inh_{re}^{MT} is defined as weighted multiplication:

$$Inh_{re}^{MT}(x,y,\theta) = \frac{\sum_{\delta\in\Delta} \omega_{(\delta,\theta)} r_{\lambda,\sigma,\psi}^{MT}(x,y,\theta)}{\|\sum_{\delta\in\Delta} \omega_{(\delta,\theta)} r_{\lambda,\sigma,\psi}^{MT}(x,y,\theta)\|_1}, \quad (2)$$

where $\omega_{(\delta,\theta)}$ is the weighting, denoting strength of connection between MT neuron of orientation δ and ventral V1 neuron of orientation θ. $\sum$ sums MT to all orientations $\delta \in \Delta$. $\|.\|_1$ is L1 norm. By setting $\omega_{(\delta,\theta)}$, two types of early recurrence are derived.

Isotropic inhibition causes fine-scale edge of an orientation inhibited by MT with *all orientations in an equal manner*,where $\omega_{(\delta,\theta)} = 1$ for all orientations. The isotropic representation highlights regions corresponding to low-spatial frequency variations to all orientations and is insensitive to variations caused by high-frequency stimuli.

Anisotropic inhibition suppresses fine-scale edge to a preferred orientation by MT responses to *the same orientation*.

Center-surround inhibition, Inh_{cs}^{V1v} is formalized following [3] as a convolution of the maximum energy map $\hat{E}_{\lambda,\sigma,\psi}^{V1}(x,y,\theta)$ with a Difference of Gaussian function weighting function as:

$$Inh_{cs}^{V1v} = \hat{E}_{\lambda,\sigma,\psi}^{V1v}(x,y) * w_{\sigma}^{V1v}(x,y), \quad (3)$$

where the maximum energy map is calculated as:

$$\hat{E}_{\lambda,\sigma,\psi}^{V1v}(x,y) = \max\{E_{\lambda,\sigma,\psi}^{V1v}(x,y,\theta_i) | i = 1..N\}, \quad (4)$$

To search for the best coarse-scale representation, single-scale representation is extended that employs the interval-tree analysis [5].

To investigate whether the proposed model is a general process for machine vision, our work is applied to 2 popular edge detectors: Canny edge detector [2] and multi-scale Brightness/Texture Gradients (BTG) detector [4]. The proposed recurrent representation fits itself into these models by modulating original edge responses before constructing contour. Result (Fig. 2) indicates the proposed computational model may consistently improve edge detections in real scenes.

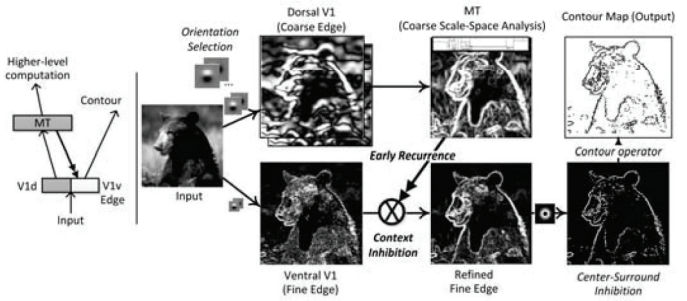


Figure 1: **Model structure. Left:** simplified biological hierarchy. The double arrow denotes the early recurrence. **Right:** an example.

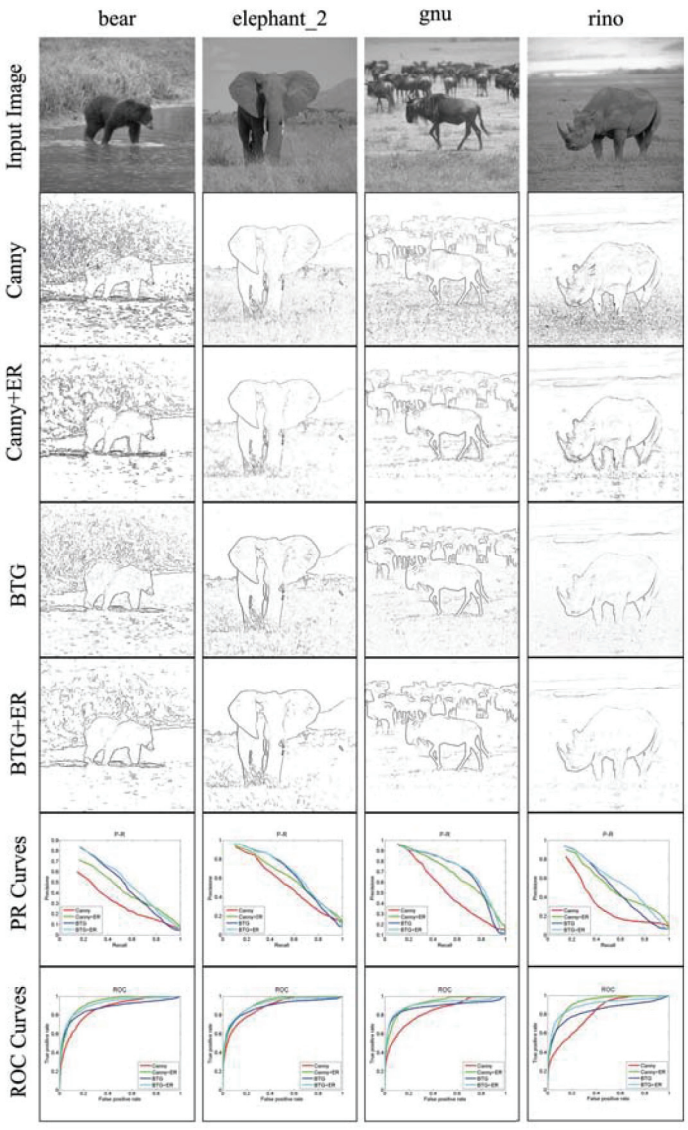


Figure 2: Improved contours.

[1] J. Bullier. *Brain Res. Rev.*, 36(2-3):96–107, 2001.

[2] J Canny. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:679–698, 1986.

[3] Cosmin Grigorescu, Nicolai Petkov, and Michel A. Westenberg. *IEEE Trans. Image Proc.*, 12(7):729–739, 2003.

[4] D.R. Martin, C.C. Fowlkes, and J. Malik. *IEEE Trans. Image Proc.*, 26(5):530–549, 2004.

[5] A. Witkin. In *ICASSP '84.*, 9, 150–153, 1984.

Modelling Visual Objects Invariant to Depictive Style

Qi Wu
http://www.cs.bath.ac.uk/~qw219
Peter Hall
http://www.cs.bath.ac.uk/~pmh

Media Technology Research Centre
Department of Computer Science
University of Bath,
Bath, UK

Representing visual objects is an interesting open question of relevance to many important problems in Computer Vision. State of the art allows thousands of visual objects to be learned and recognised, under a wide range of variations. Only a small fraction of the literature addresses the problem of variation in depictive style (photographs, drawings, paintings *etc*.), yet considering photographs and artwork on equal footing is philosophically appealing and of true practical significance. This paper describes a model for visual object classes that is learnable and which is able to classify over a broad range of depictive styles. When compared to a collection of state-of-art classifiers, our results show a significant increase in robustness to variance in depictive style.

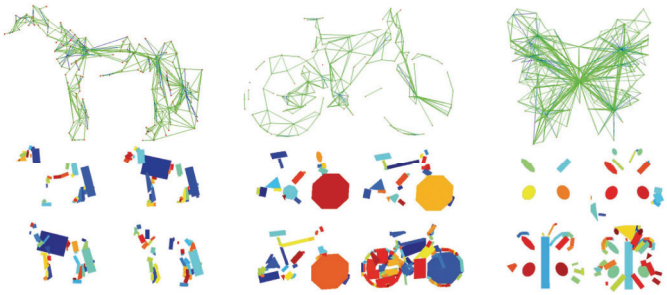


Figure 1: Examples of three graph models generated from 3 categories of objects, which are horses, bicycles, and butterflies. The visualization shows of selected levels below the corresponding model, with the simple shapes fitted. Child-parent arcs are in blue, adjacencies between the nodes in the same level are green.

We argue that models of visual objects should not be premised, even tacitly, on photo-real appearance or indeed on any particular depictive style at all. Rather, visual object models should be based on quasi-invariant properties of the objects in a class. A similar argument is made by those who advocate part-based representations for image. We go further by saying that such models should generalise across depictive styles. This means that if a model is constructed using images in one style, the same object should also be classifiable even when depicted using a different style. In this paper, we investigate a method for modelling visual objects classes in a manner that is invariant to depictive style. The assumption we make is that an object class is characterised by the qualitative shape of object parts and their structural arrangement. Hence we use a graph of nodes and arcs in which qualitative shapes such as triangle, square, and circle to label the nodes. More exactly our model is a hierarchy of levels, yielding a coarse-to-fine representation. Each level contains an undirected graph of nodes and arcs. Nodes between levels are connected via parent-child arcs, which are directed. Child nodes are nested inside their parent.

We learn visual class models from input images, each labelled with the object they contain. There are three major steps. Fig 2 presents a framework of the proposed method.

(i) Build Image Graphs, one for each images: Our modeller uses Berkeley segmentation [1] that automatically yields a hierarchical description of an input image. And then we build a graph from this in which regions label nodes. A novelty here is that we label nodes using qualitative shapes from a collection of shapes. Because we assume that abstracting region shape brings greater robustness to model different depictive styles. These primitive shapes are chosen because they have been shown to explain around 80% of regions in photographs *up to an affine transform*[3].

(ii) Compute an Initial Visual Class Model: The second step is to compute an initial visual class model, by using an approximate median graph generation method [2].

(iii) Refine the Visual Class Model: The initial model contains nodes and arcs that derive from visual clutter in back ground of images in the training set, so we developed a cleaning algorithm to remove such

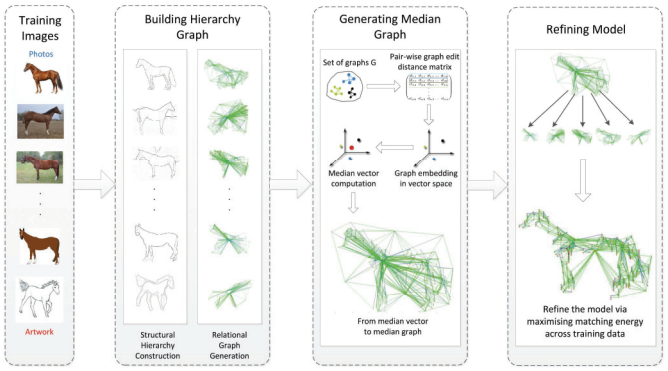


Figure 2: Constructing a class model, from left to right. (a): An input collection (possibly different depictions) used for training. (b): Probability maps for each input image, and graph models for each map. (c): The median graph model for the whole class. (d): The refined median graph as the final class model.

elements.Fig 1 shows some final results.

Our *vcm* has the potential to be used in many applications, here we use cross depiction classification. Using our expanded version of CalTech 256, we compare with three state of art classification methods, one BoW method which uses PHOW features, one shape model-based method, another one uses the first few eigenvalues of Laplacian matrix of the object structure as feature vector. In total we test 800 images and Table 1 shows that our method outperforms the shape and structure only method on all cases. We outperform BoW in all cases except case 1i, when photographs are used in both training and testing.

case 1: Training	5p	5a
case 1: Testing	15p	15a
Dense SIFT	70%	59%
Shape Model	25%	33%
Structure Only	16%	19%
Proposed Method	61%	62%

case 2: Training	8p	10p	8a	10a
case2 : Testing	15a	15a	15p	15p
Dense SIFT	43%	47%	49%	51%
Shape Model	33%	35%	34%	34%
Structure Only	19%	23%	22%	25%
Proposed Method	63%	64%	64%	67%

case 3: Training	3a	5a	3p	5p
case 3: Testing	30m	30m	30m	30m
Dense SIFT	46%	50%	50%	54%
Shape Model	27%	30%	24%	27%
Structure Only	13%	16%	14%	16%
Proposed Method	58%	61%	56%	61%

case 4: Training	6m	10m
case 4: Testing	30m	30m
Dense SIFT	60%	61%
Shape Model	32%	34%
Structure Only	21%	24%
Proposed Method	62%	65%

Table 1: Classification accuracy for different cases. From top to bottom, left to right: (a) single domain task, (b) single cross depiction task, and (c) single to mixture depiction task, (d) mixture cross depiction task. The character 'p' is 'photos', 'a' is 'art' and 'm' is 'mixture'. More detailed results for each single experiment can be found in supplementary material.

The ability to generalise to new depictive styles is important, not least because the number of depictive styles is seemingly unbounded. No training procedure can capture them all and so a class model that is able to generalise to unseen depictive styles is of value. Experiments show that our proposal method performs better than the traditional visual appearance based method in cross-depiction problems, in mixed problems, and in art-only problems.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011. ISSN 0162-8828.

[2] Miquel Ferrer, Ernest Valveny, Francesc Serratosa, Kaspar Riesen, and Horst Bunke. An approximate algorithm for median graph computation using graph embedding. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[3] Qi Wu and Peter Hall. Prime shapes in natural images. In *Proceedings of the British Machine Vision Conference*, pages 45.1–45.12. BMVA Press, 2012. ISBN 1-901725-46-4.

Free your Camera: 3D Indoor Scene Understanding from Arbitrary Camera Motion

Axel Furlan[1]
furlan@disco.unimib.it
Stephen David Miller[2]
sdmiller@stanford.edu
Domenico G. Sorrenti[1]
sorrenti@disco.unimib.it
Li Fei-Fei[2]
feifeili@stanford.edu
Silvio Savarese[3]
silvio@umich.edu

[1] Computer Science Department
University of Milano - Bicocca
Milano, Italy
[2] Computer Science Department
Stanford University
Stanford, CA, USA
[3] EECS Building
University of Michigan
Ann Arbor, MI, USA

Many works have been presented for indoor scene understanding, yet few of them combine structural reasoning with full motion estimation in a real-time oriented approach. In this work we address the problem of estimating the 3D structural layout of complex and cluttered indoor scenes from monocular video sequences, where the observer can freely move in the surrounding space. We propose an effective probabilistic formulation that allows to generate, evaluate and optimize layout hypotheses by integrating new image evidence as the observer moves. Compared to state-of-the-art work, our approach makes significantly less limiting hypotheses about the scene and the observer (e.g., Manhattan world assumption, known camera motion). We introduce a new challenging dataset and present an extensive experimental evaluation, which demonstrates that our formulation reaches near-real-time computation time and outperforms state-of-the-art methods while operating in significantly less constrained conditions.

Figure 1 shows a pictorial representation and a schematic diagram of the whole process.

Sparse 3D reconstruction. As the observer moves in the surrounding environment, we first pre-process sequences with a localization and sparse 3D reconstruction algorithm. In our experiments we compare two such approaches: a real-time implementation of the Monocular V-SLAM approach proposed in [4] and the non-real-time VisualSfM [6]. These 3D reconstructions are in general noisy and sparse.

Candidate layout components. The second step consists of generating a higher level representation of the 3D points estimated in the preprocessing phase. Several types of geometrical primitives are suitable for this purpose. In our case, we believe a piecewise planar representation is the most appropriate for indoor scene representation. We fit a large number of planes to the 3D points so as to generate a large number of (potentially inaccurate) candidates of layout components, i.e. walls, floor, ceiling. In our experiments we implemented an Iterative RanSaC plane fitting procedure, which we optimized for indoor scenes by allowing peripheral fitted points to be re-injected in the iteration process, since these points potentially lay on the intersection of two planes.

Layout Parametrization While much prior work has leveraged the Manhattan world assumption, we believe this is a limiting hypothesis. To overcome this limitation, in this paper we adopt a representation similar to [5] (sometimes referred to as Soft Manhattan), which makes the following assumptions about the environment: i) ground plane and ceiling are parallel; ii) walls are only constrained to be orthogonal to the ground plane (and ceiling); iii) there can be any number of walls and each wall can be displaced at any angle with respect to other walls.

Layout estimation. In the last step, constituting the core of our proposed inference engine, we generate layout hypotheses as random combinations of candidate layout components. Each layout hypothesis is evaluated at each time frame by measuring its compatibility with observations (*e.g.* image points and lines) and geometrical constraints across frames. During this process, each layout is "perturbed" by locally adjusting, optimizing, merging or splitting layout components. There are different approaches to manage sets of hypotheses. In this paper we choose to integrate our probabilistic framework within a particle filter structure. This choice allows to explicitly formulate the problem in a parallel-computing oriented fashion (particles are independent from each other), which can lead to high efficiency gains in computation time. The output of the optimization procedure is an estimation of the 3D scene layout, which is obtained by selecting the layout hypothesis with the best set of layout components.

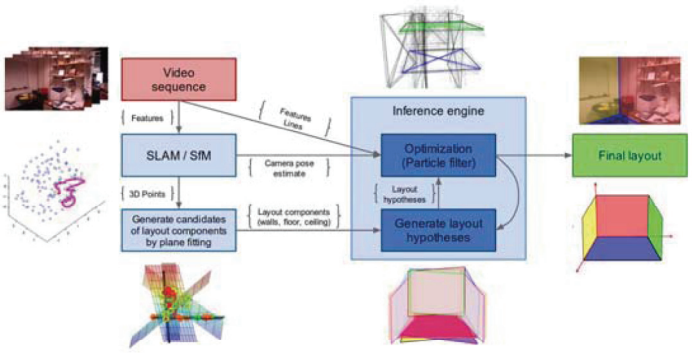


Figure 1: **3D scene layout estimation process.** The video sequence is first processed to obtain camera localization and sparse 3D point cloud reconstruction. Layout components (e.g. floor, ceiling, walls) are generated from the sparse 3D points and combined to generate layout hypotheses. Each layout hypothesis is evaluated and optimized by incorporating new image evidence. The final 3D scene layout is represented by the hypothesis that better describes the scene.

Scoring Hypotheses At each timestep t, we wish to assign a probability to a particular hypothesis, taking into account new observations and geometrical constraints:

$$P_t = \prod_i P_f^i P_o^i(\theta_i) P_r^i(e_r^i) \prod_j P_m^{ij}(\phi_{ij}) P_s^{ij}(d_{ij}^{-1})^{p_{ij}} (P_w^{ij})^{a_{ij}} \quad (1)$$

We have designed this probability to enforce a number of desirable properties (*e.g.* low reprojection error, description simplicity, physical plausibility, etc.). Please refer to the main paper for a complete description of this scoring function and its components.

Experimental validation. We show experimental results of our method when tested on the state-of-the-art dataset [5], as well as on a new challenging dataset [1] that we introduce in this paper. Final reconstruction results are compared to three state-of-the-art approaches (the video-based approach [5] and the two well known single image methods [2, 3]) and a baseline method (to explicitate the importance of the evaluation and optimization process).

Conclusions. In this paper we present a real-time oriented approach for indoor scene understanding in cluttered environments. The proposed probabilistic framework allows to generate, evaluate and optimize layout hypotheses by integrating new image evidence as the observer moves. In the extensive experimental evaluation we demonstrate that our formulation reaches near-real-time computation time and, while operating in significantly less constraining conditions (*e.g.* soft Manhattan assumption, complex scene geometry, freely moving observer), outperforms state-of-the-art methods in both classification accuracy and computation time.

[1] URL `http://www.ira.disco.unimib.it/free_your_camera`.
[2] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009.
[3] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *IJCV*, 75(1), 2007.
[4] Cyril Roussillon, Aurélien Gonzalez, Joan Solà, Jean-Marie Codol, Nicolas Man-sard, Simon Lacroix, and Michel Devy. Rt-slam: A generic and real-time visual slam implementation. *CoRR*, 2012.
[5] Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *ICCV*, 2011.
[6] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M. Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

Exploring SVM for Image Annotation in Presence of Confusing Labels

Yashaswi Verma
http://researchweb.iiit.ac.in/~yashaswi.verma/
C. V. Jawahar
http://www.iiit.ac.in/~jawahar/

CVIT
IIIT-Hyderabad
Hyderabad, India
http://cvit.iiit.ac.in

We address the problem of automatic image annotation in large vocabulary datasets. In such datasets, there exist three practical issues: (a) **Incomplete-labeling**: The training samples are not exhaustively tagged with *all* relevant labels from vocabulary. This is because while building a dataset, human annotators find some labels as "obvious" and miss them in the ground-truth. E.g., an image tagged with *"car"* might not be tagged with *"vehicle"*. (b) **Label-ambiguity**: There are some labels that convey same semantic meaning and thus can be used interchangeably, due to which usually only one of them is assigned by annotator. E.g., an image tagged with *"flowers"* may not be tagged with *"blooms"*, as both convey the same meaning. (c) **Structural-overlap**: There are some labels that, in spite of being different, *share* structural properties. E.g., though *"tiger"* and *"lion"* are two different labels, structurally they are very similar.

Most of the earlier works in this domain have focused on nearest-neighbour based models. A recent work [4] also tries to integrate label information in the nearest-neighbour set-up. In this work, first we demonstrate that even the conventional SVM outperforms several benchmark models. Then we propose a new loss function based on the hinge-loss in order to make SVM tolerant against the three issues discussed above. For this, we introduce a tolerance-parameter "t" that adjusts both the margin as well the gradient update-rule for each sample separately. We call this model as *Suppor Vector Machine with Variable Tolerance* (or *SVM-VT*). Specifically, we formalize the SVM-VT model as that of solving the following optimization problem:

$$\min_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{m}\sum_{j=1}^{m}[1 - y_j t_j(\mathbf{w}\cdot\mathbf{x}_j)]_+, \quad (1)$$

where the additional parameter $t_j \in [0,1]$ controls the tolerance against the errors made in the classification of sample $\mathbf{x}_j$. The hyperplane $\mathbf{w}$ is learnt such that it is more strict towards correctly classifying samples with high value of t_j and any such error leads to a large shift in the hyperplane.

We propose a heuristic approach for determining the t-value for each sample given a label. For a label l, let S^+ and $\bar{S}^+$ be the sets of its positive and negative examples respectively. We consider three factors to determine the semantic relatedness of each sample $\mathbf{x}_j \in \bar{S}^+$ with that label:
(a) *Reverse nearest-neighbours based score*: For a fixed value of K (= 5), let p_k be the number of samples in S^+ that have $\mathbf{x}_j$ as their k^{th} nearest neighbour. Then we define

$$score_1(\mathbf{x}_j|l) = \frac{\sum_{k=1}^{K}(\frac{p_k}{k})}{\sum_{k=1}^{K} p_k + \varepsilon} \quad (2)$$

(b) *Visual similarity based score*: We compute the visual similarity score $sim(\cdot)$ (scaled into range $[0,1]$) of $\mathbf{x}_j$ with its nearest neighbour $\mathbf{x}_i^* \in S^+$ using JEC [3] method and define

$$score_2(\mathbf{x}_j|l) = sim(\mathbf{x}_j, \mathbf{x}_i^*) \quad (3)$$

(c) *Label cooccurrence based score*: For a label l, let $\mathbf{y} \in \{0,1\}^m$ be such that its i^{th} entry is 1 if the i^{th} training image is tagged with l, and 0 otherwise. We compute cooccurrence score $co_occur(l_i, l_j)$ between two labels l_i and l_j using cosine similarity between their corresponding vectors $\mathbf{y}_i$ and $\mathbf{y}_j$. Let $\mathbf{x}_j$ be tagged with labels L_j, then we define

$$score_3(\mathbf{x}_j|l) = \max_{l_j \in L_j} co_occur(l, l_j) \quad (4)$$

From these, we define tolerance parameter for sample $\mathbf{x}_j$ given label l as

$$t_j = 1 - \frac{1}{3}(score_1(\mathbf{x}_j|l) + score_2(\mathbf{x}_j|l) + score_3(\mathbf{x}_j|l)) \quad (5)$$

We compute t_j only for samples in $\bar{S}^+$, and take $t_j = 1$ for all positive samples assuming that they are correctly annotated. From equation 5, it can be seen that for some negative sample, smaller tolerance value corresponds to higher chance of it being related to a given label and vice-versa. Figure 1 shows negative samples (along with their ground-truth labels) with least t-scores for two labels each from three benchmark datasets.



Figure 1: For example labels (in blue) from the three datasets, the top "negative" samples that have least t-scores with corresponding ground-truth labels (smaller t-score implies higher semantic relevance).

Dataset →	Corel-5k	ESP Game	IAPRTC-12
Method ↓	P/R/F1/N+	P/R/F1/N+	P/R/F1/N+
MBRM[1]	0.24/0.25/0.245/122	0.18/0.19/0.185/209	0.24/0.23/0.235/233
JEC[3]	0.27/0.32/0.293/139	0.22/0.25/0.234/224	0.28/0.29/0.285/250
TagProp-ML[2]	0.31/0.37/0.337/146	**0.49**/0.20/0.284/213	**0.48**/0.25/0.329/227
TagProp-σML[2]	**0.33/0.42/0.370/160**	0.39/**0.27/0.319/239**	0.46/**0.35/0.398/266**
KSVM	0.29/0.43/0.346/174	0.30/0.28/0.290/256	0.43/0.27/0.332/266
KSVM-VT (Ours)	**0.32/0.42/0.363/179**	**0.33/0.32/0.325/259**	**0.47/0.29/0.359/268**

Table 1: Performance comparison among different methods. The prefix 'K' corresponds to kernelization using chi-squared kernel.

We evaluate the performance using average precision per label (P), average recall per label (R), average F1 score, and number of labels with positive recall (N+). Table 1 shows the annotation performance of different methods. It can be seen that our method shows promising results on the task of image annotation on three challenging datasets, and establishes a baseline for such models in this domain.

[1] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbour models for image auto-annotation. In *ICCV*, 2009.

[3] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. In *IJCV*, 2010.

[4] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV*, 2012.

Removing Mistracking of Multibody Motion Video Database Hopkins155

Yasuyuki Sugaya[1]
sugaya@iim.cs.tut.ac.jp
Yuichi Matsushita[1]
matusita@iim.cs.tut.ac.jp
Kenichi Kanatani[2]
kanatani2013@yahoo.co.jp

[1] Department of Computer Science and Engineering
Toyohashi University of Technology
Aichi 441-8105 Japan

[2] Okayama University
Okayama 700-8530 Japan

Many mathematical techniques have been presented for classifying feature point trajectories over multibody motion video sequences into different motions, and most are applied to the Hopkins155 database for evaluating their performance. In this paper, we point out that Hopkins155 has problems and that it cannot necessarily evaluate the performance correctly. We create a new database incorrect trajectories from Hopkins155. The basic principle of mistracking removal lies on the fact that correct trajectories all belong to parallel 2-D affine spaces in a high-dimensional space if all motions are translational and that parallel 2-D affine spaces are included in a 3-D affine space. Noting that if the image sequence is divided into short intervals, individual motions can be regarded as approximately translational in each interval, we detect incorrect trajectories by repeated plane fitting in the 3-D using RANSAC.

Suppose we track N feature points $\{p_\alpha\}$ over M frames. Let $(x_{\kappa\alpha}, y_{\kappa\alpha})$, $\kappa = 1,...,M, \alpha = 1,...,N$, be the image coordinates of the αth point p_α in the κth frame. Its motion history is represented by the $2M$-D vector

$$\mathbf{p}_\alpha = (x_{1\alpha}, y_{1\alpha}, x_{2\alpha}, y_{2\alpha}, ..., x_{M\alpha}, y_{M\alpha})^\top, \tag{1}$$

which we simply call the "trajectory" of p_α. We map these $2M$-D points to 3-D points as follows:

1. Compute the centroid $\mathbf{p}_C$ of the trajectories $\mathbf{p}_\alpha$, $\alpha = 1, ..., N$ and the deviations $\tilde{\mathbf{p}}_\alpha$ from $\mathbf{p}_C$:

$$\mathbf{p}_C = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{p}_\alpha, \quad \tilde{\mathbf{p}}_\alpha = \mathbf{p}_\alpha - \mathbf{p}_C. \tag{2}$$

2. Compute the singular value decomposition of the $2M \times N$ matrix

$$(\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_N) = \mathbf{U}\mathrm{diag}(\sigma_1, ..., \sigma_r)\mathbf{V}^\top, \tag{3}$$

where $r = \min(2M, N)$, $\mathbf{U}$ is a $2M \times r$ matrix with r orthonormal columns, $\mathbf{V}$ is an $N \times r$ matrix with r orthonormal columns, and $\sigma_1 \geq \cdots \geq \sigma_r$ (≥ 0) are the singular values.

3. Let $\mathbf{u}_i$ be the ith column of $\mathbf{U}$, and compute the following 3-D vectors $\mathbf{r}_\alpha$, $\alpha = 1, ..., N$:

$$\mathbf{r}_\alpha = ((\tilde{\mathbf{p}}_\alpha, \mathbf{u}_1), (\tilde{\mathbf{p}}_\alpha, \mathbf{u}_2), (\tilde{\mathbf{p}}_\alpha, \mathbf{u}_3))^\top, \tag{4}$$

where and hereafter we denote the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$ by $(\mathbf{a}, \mathbf{b})$.

Geometrically, we are translating the coordinate system of the $2M$-D trajectory space so that the origin is at the centroid $\mathbf{p}_C$, computing the three vectors (the columns of $\mathbf{U}$) that span the affine space, and expressing all the trajectories as their linear combinations.

To these points in 3-D, we fit multiple planes by the following RANSAC procedure:

1. Randomly choose three from among points $\mathbf{r}_\alpha$, $\alpha = 1, ..., N$.
2. Fit a plane to the selected three points and compute θ by LS.
3. Let S be the number of points $\mathbf{r}_\alpha$ that satisfy

$$\frac{(\mathbf{r}_\alpha, \theta)^2}{\theta_1^2 + \theta_2^2 + \theta_3^2} \leq \sigma^2, \tag{5}$$

where θ_i is the ith component of θ. The left hand side is the square distance of point $\mathbf{r}_\alpha$ from the fitted plane, and σ is the standard deviation of feature point detection accuracy, which is empirically set.

4. Repeat the above computation many times and find the value θ that maximize S.



Figure 1: Three video sequences in Hoplins155. The marks □ and × indicate feature point locations regarded as correctly tracked and incorrectly tracked, respectively, by our procedure.



Figure 2: 3-D visualization tracking trajectory in the first five frames in the sequences in Fig. 1.

5. Remove those points $\mathbf{r}_\alpha$ that satisfy

$$\frac{(\mathbf{r}_\alpha, \theta)^2}{\theta_1^2 + \theta_2^2 + \theta_3^2} < \sigma^2 \chi^2_{1;99}, \tag{6}$$

where $\chi^2_{r;a}$ is the ath percentile of χ^2 distribution with r degrees of freedom. This means that we retain those points that cannot be regarded as deviated from the fitted plane by Gaussian noise of mean 0 and standard deviation σ with 1% significance level.

We integrate the result in all intervals by defining the reliability index for the ith interval using the sigmoid function as follows:

$$P(\mathbf{p}_\alpha^{(i)}) = \frac{1}{1 + e^{-(d_\alpha^{(i)} - \sigma^2\chi^2_{1;99})}}. \tag{7}$$

Here, $\mathbf{p}_\alpha^{(i)}$ is the vector that describe the partial trajectory of $\mathbf{p}_\alpha$ over the ith interval, and $d_\alpha^{(i)}$ is the left-hand side of Eq. (6) for the ith interval. However, we regard those points that satisfy Eq. (6) as correct and let $P(\mathbf{p}_\alpha^{(i)}) = 0$. We integrate the results in all the intervals in the following form (the trajectory is more likely to be incorrect if it is larger):

$$L(\mathbf{p}_\alpha) = \Pi_{i|P(\mathbf{p}_\alpha^{(i)})\neq 0}^{K} P(\mathbf{p}_\alpha^{(i)}). \tag{8}$$

Here, K is the number of intervals.

Figure 1 shows a video of Hopkins155. We added new feature tracking and divided the sequence into five-frame intervals with one frame overlaps. Five decimated frames are shown in Fig. 1. The marks □ and × indicate feature point locations regarded as correctly tracked and incorrectly tracked, respectively. We set σ =1.0, which we judged to be reasonable according to our experiences. Figure 2 shows 3-D visualization of partial trajectories in the first five-frame intervals. We can clearly see that correct trajectories lie on two planes and that incorrect trajectories stick out from them.

For correct evaluation, we need a reliable database. For this purpose, we created a new database[1] by removing incorrect trajectories using our technique from 35 scenes of Hopkins155, which were natural scenes consisting of two motions.

[1] http://www.iim.cs.tut.ac.jp/T-Hopkins/

Enforcing Monotonous Shape Growth or Shrinkage in Video Segmentation

Yuliya Tarabalka
yuliya.tarabalka@inria.fr
Inria Sophia-Antipolis, AYIN team
Sophia Antipolis, France

Guillaume Charpiat
guillaume.charpiat@inria.fr
Inria Sophia-Antipolis, STARS team

Ludovic Brucker
ludovic.brucker@nasa.gov
NASA GSFC & Universities Space Research Association
MD, USA

Bjoern H. Menze
bjoern@ethz.ch
Computer Vision Laboratory, ETH Zurich, Switzerland &
Inria Sophia-Antipolis, Asclepios team

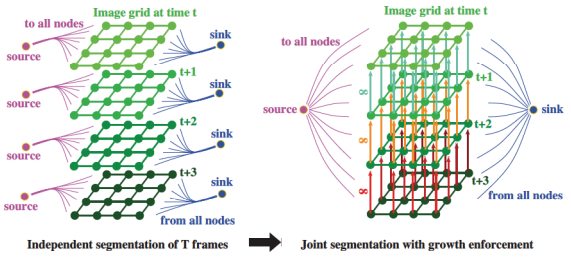


Figure 1: Enforcing shape growth in an image sequence.

One of the great challenges in computer vision is automatic segmentation of objects in videos. This task becomes more difficult when image sequences are subject to low signal-to-noise ratio or low contrast between intensities of neighboring structures. Such challenging data are acquired routinely, for example, in medical imaging or satellite remote sensing.

While individual frames can be analyzed independently, temporal coherence in image sequences provides a lot of information not available for a single image. In this work, we focus on segmenting shapes which only grow or shrink in time, from sequences of extremely noisy images. We consider image sequences, where both foreground and background intensity distributions can vary significantly over time, foreground can be heavily occluded or undistinguishable from a part of the background. Most of previously-proposed spatio-temporal methods rely on coherence of foreground/background intensity distributions in consecutive image frames, and therefore fail when segmenting such noisy data sets. Few approaches have been designed for spatio-temporal segmentation of shapes from magnetic resonance (MR) images with low signal-to-noise ratio [3, 4]. Applied to multi-temporal sequences that show a monotonously growing or shrinking structure, these smoothing methods bias results towards the mean shape obtained from averaging consecutive segmentations and, hence, underestimate rapid growth or shrinkage events.

To address this issue, we propose a new *segmentation framework* based on graph cuts for the joint segmentation of an image sequence. It *introduces growth or shrinkage constraint in graph cuts* by using directed infinite links, connecting pixels at the same spatial locations in successive image frames. By minimizing an energy computed on the resulting spatio-temporal graph of the image set, the proposed method yields a *globally optimal solution*. Differently from the state-of-the-art spatio-temporal techniques, it does not rely on the coherence of the intensity in time, but only on the coherence of the shape.

Graph cut is an optimization tool, which can be used to find the globally optimal binary segmentation of images [1], where the segmentation criterion E is related to a Markov Random Field with submodular interaction terms:

$$E(L) = \sum_{\text{pixels } i} V_i(L_i) + \sum_{i \sim j} W_{i,j}(L_i, L_j), \qquad (1)$$

where L is the binary labelling function to be found (L_i is the label of pixel i), individual potentials V_i are any binary real-valued functions measuring the disagreement between a prior probabilistic model and the observed data, $i \sim j$ denotes a pair of neighboring pixels, and $W_{i,j}$ are any real-valued interaction terms between neighboring pixels expressing spatial coherency of labels. A directed infinite link between two pixels expresses precisely the constraint that this pair of pixels cannot have the pair of labels (0,1). In the case of image binary segmentation, if 0 stands for the background and 1 for the foreground object, then this means that the second pixel may belong to the foreground only if the first one does.

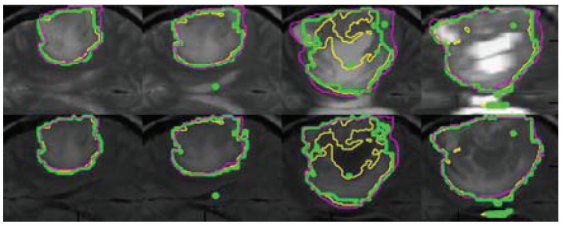


Figure 2: Time series of T2 and FLAIR MR image volumes, each of which acquired about 3-6 months apart. The rapidly growing tumor between the second and fourth scene leads to a suboptimal performance of the initial multimodal segmentation (yellow) [2]. The proposed segmentation with growth constraints (green) delineates areas similar to the manual evaluation (magenta). It does not smooth out outlines of rapidly growing tumors as conventional bi-directional temporal constraints would do.

Shape growth in a sequence of images $I(t)$ can be easily expressed as the property that the foreground object cannot lose any pixel when time advances. Otherwise said, if a pixel belongs to the foreground object at time t_1, then it belongs also to the foreground object for all times $t_2 > t_1$. Equivalently, and simpler: a pair of pixels $((x,y,t),(x,y,t+1))$, sharing the same location and immediately successive in time, cannot have the pair of labels (1,0), with the same binary segmentation notations as above. This can be enforced by setting directed infinite links from all pixels to their immediate predecessor in time.

Given T images $I(t), t \in [1,T]$, and as many associated submodular segmentation criteria E^t, we transform the problem of segmenting independently each image $I(t)$ according to its criterion E^t, into a joint segmentation of all images together, by enforcing the shape growth constraint with directed infinite links (Fig. 1). Thus, instead of applying graph cut T times independently to planar grids of the size of the images $W \times H$, we apply graph cut once to a 3D grid $W \times H \times T$, consisting of the same nodes and edges, but with additional directed infinite links in time. The criterion to be minimized is then $E = \sum_t E^t$ under the constraint of shape growth, and the solution found by graph cut is globally optimal, since the problem is binary and submodular.

One can enforce shape shrinkage instead of shape growth, by reversing the direction of infinite links. Another straightforward extension consists in applying this approach to the case of sequences of 3D images. The directed infinite links are then set for all pairs of voxels of the form $((x,y,z,t),(x,y,z,t-1))$ to enforce 3D shape growth.

We validated the performance of the proposed approach for segmentation of a shrinking ice floe and growing brain tumors from multimodal sequences of 2D satellite and 3D MR images (Fig. 2), respectively. In the latter application, we imposed an additional inter-sequences object inclusion constraint by adding directed infinite links to the joint graph associated to all sequences. The new method proved to be robust to very noisy or low-contrast images, and showed linear complexity.

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 26(9):1124–1137, 2004.

[2] B. H. Menze et al. A generative model for brain tumor segmentation in multi-modal images. In *MICCAI*, pages 151–159, 2010.

[3] T. Riklin-Raviv et al. Segmentation of image ensembles via latent atlases. *Med. image analys.*, 14(5):654–665, Oct. 2010.

[4] R. Wolz et al. Measurement of hippocampal atrophy using 4d graph-cut segmentation: Application to ADNI. *NeuroImage*, 52(1):109–118, 2010.

Deformable Part Models with Individual Part Scaling

Charles Dubout
charles.dubout@idiap.ch
François Fleuret
francois.fleuret@idiap.ch

Computer Vision and Learning Group
Idiap Research Institute
Martigny, Switzerland

École Polytechnique Fédérale de
Lausanne, Switzerland

Current deformable part models such as the ones introduced by Felzenszwalb *et al.* [2] let the parts deform only at a fixed predetermined scale relative to that of the root of the models (typically at twice the resolution). They do so because it enables them to find the optimal placement of each part efficiently, using a fast 2D distance transform algorithm.

We demonstrate in our paper that if one settles for approximately optimal placements, it is possible to efficiently deform the parts across scales as well, by reusing the original convolutions and distance transforms. Allowing parts to move in 3D increases the expressivity of the models, permitting them to compensate for a wider class of deformations, and might approximate an increase in the scanning resolution.

Let H be a feature pyramid and $\mathbf{p} = (x, y, z)$ specify a 2D position (x, y) in the z-th level of the pyramid. Let $\phi(\mathbf{p})$ denote the vector obtained by concatenating the feature vectors in the sub-window of H centered at $\mathbf{p}$, and $\phi_d(\mathbf{p})$ be the deformation features.

A model for an object with n parts is composed of a root filter $\mathbf{w}_0$ and n pairs $(\mathbf{w}_i, \mathbf{d}_i)$, where $\mathbf{w}_i$ is the filter of the i-th part and $\mathbf{d}_i$ is a vector specifying the deformation cost of the part placement.

An object hypothesis $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ specifies the location of the center of each filter in a feature pyramid. Its score is given by the score of each filter at its respective location, minus a deformation cost that depends on the location of each part with respect to the root position,

$$S(\mathbf{p}_0, \ldots, \mathbf{p}_n) = \mathbf{w}_0^\mathsf{T} \phi(\mathbf{p}_0) + \sum_{i=1}^{n} \mathbf{w}_i^\mathsf{T} \phi(\mathbf{p}_i) - \mathbf{d}_i^\mathsf{T} \phi_d(\mathbf{p}_i - \mathbf{p}_0). \quad (1)$$

In our algorithm, we allow the parts to also move across scales, and extend the usual deformation features to include the z component of the disparity between root and parts positions,

$$\phi_d(\mathbf{p}) = (1, x, y, z, x^2, y^2, z^2). \quad (2)$$

Ideally, one would like to now find the optimal location of each part in this way,

$$\mathbf{p}_i^\star = \underset{\mathbf{p} \in \mathbb{Z}^3}{\operatorname{argmax}} \, \mathbf{w}_i^\mathsf{T} \phi(\mathbf{p}) - \mathbf{d}_i^\mathsf{T} \phi_d(\mathbf{p} - \mathbf{p}_0(z_i)), \quad (3)$$

where $\mathbf{p}_0(z_i) = \left(\lambda^{z_i - z_0} x_0, \lambda^{z_i - z_0} y_0, z_0\right)$ are the coordinates of the root position in the i-th part's level z_i, λ being the scaling factor between two successive levels of the feature pyramid.

Unfortunately, $\mathbf{p}_0(z_i)$ is likely to be non-integral, and the generalized distance transform [1] thus cannot be used directly anymore. We thus approximate the root position at the scale of the i-th part by the closest integer one,

$$\tilde{\mathbf{p}}_0(z_i) = \underset{\mathbf{p} \in \mathbb{Z}^2 \times \{z_i\}}{\operatorname{argmin}} \, ||\mathbf{p} - \mathbf{p}_0(z_i)||. \quad (4)$$

Using this approximate root position, we can now again use the generalized distance transform in order to find the optimal part location,

$$\tilde{\mathbf{p}}_i^\star(\mathbf{p}_0) = \underset{\mathbf{p} \in \mathbb{Z}^3}{\operatorname{argmax}} \, \mathbf{w}_i^\mathsf{T} \phi(\mathbf{p}) - \mathbf{d}_i^\mathsf{T} \phi_d(\mathbf{p} - \tilde{\mathbf{p}}_0(z_i)). \quad (5)$$

We expect this location to coincide most of the time with the optimal one for the real root position, the difference between the real and the approximate one being at most 0.5 along the x and y axes. However, the optimal score returned by the transform will generally not match the score of any real root and part configuration so we recompute it in constant time using this time the real root position $\mathbf{p}_0(z_i)$,

$$\tilde{S}(\mathbf{p}_0) = \mathbf{w}_0^\mathsf{T} \phi(\mathbf{p}_0) + \sum_{i=1}^{n} \mathbf{w}_i^\mathsf{T} \phi(\tilde{\mathbf{p}}_i^\star(\mathbf{p}_0)) - \mathbf{d}_i^\mathsf{T} \phi_d(\tilde{\mathbf{p}}_i^\star(\mathbf{p}_0) - \mathbf{p}_0(z_i)), \quad (6)$$

in order to obtain a lower bound on the true optimal score.

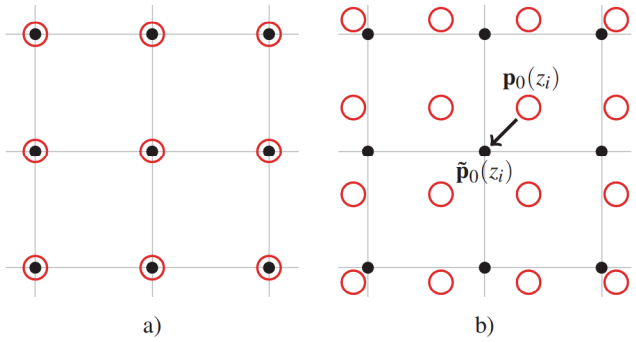


Figure 1: Lattices of part locations (in black) in a particular pyramid level. The red circles indicate root positions. In a) the part and root positions are at the same scale, as is always the case with standard models. In b) there is a mismatch between the scales of the two, and we show how we approximate a root position $\mathbf{p}_0(z_i)$ by rounding it to the closest integral position $\tilde{\mathbf{p}}_0(z_i)$ when looking for its optimal part placement.

	aero	bike	bird	boat	bottle	bus	car
voc-release4 (AP)	28.9	60.2	1.7	8.3	20.6	53.5	51.3
2D DPM (AP)	30.3	57.7	4.4	11.4	25.2	55.0	53.3
3D DPM (AP)	33.5	59.4	6.9	13.1	28.7	59.0	52.9
Rel. gain (%)	10.6	2.8	55.4	15.1	13.9	7.2	-0.8

	cat	chair	cow	table	dog	horse	mbike
voc-release4 (AP)	6.9	3.3	54.5	47.6	18.7	20.1	13.8
2D DPM (AP)	11.1	5.0	59.2	48.5	19.2	22.5	24.4
3D DPM (AP)	19.5	6.5	61.2	48.9	20.6	26.8	25.9
Rel. gain (%)	75.9	29.2	3.5	0.8	7.3	19.4	6.3

	person	plant	sheep	sofa	train	tv	mean
voc-release4 (AP)	38.8	5.8	14.3	28.1	37.3	39.0	**27.6**
2D (AP)	35.7	8.6	18.8	28.4	42.3	42.2	**30.2**
3D (AP)	32.9	11.1	21.5	31.7	44.5	43.8	**32.4**
Rel. gain (%)	-7.7	29.5	14.3	11.6	5.3	3.9	**15.2**

Table 1: Pascal VOC 2007 challenge Average Precision comparison for the models of [2] as well as our 2D and 3D models.

	Brute-force	Approx. DT
mean (AP)	45.0	44.9

Table 2: Comparison between an exact method as well as our approximation to the generalized distance transform.

The idea motivating our work is to make full use of all the convolutions between pyramid levels and part filters evaluated in DPMs, reusing them to deform parts across multiple scales. The extension we presented increases on average the detection accuracy of the models by 15% for a moderate augmentation of its total computational cost, the number of convolutions and distance transforms remaining constant. Despite relying on an approximation to the generalized distance transform, our approach obtains scores virtually equal to its exact but much slower counterpart.

[1] P. F. Felzenszwalb and D. P. Huttenlocher. Distance Transforms of Sampled Functions. Technical report, Cornell Computing and Information Science, 2004.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32 (9):1627–1645, 2010.

Focusing Attention on Visual Features that Matter

Grace Tsai
gstsai@umich.edu
Benjamin Kuipers
kuipers@umich.edu

Electrical Engineering and Computer Science
University of Michigan
Ann Arbor MI 48109 USA

An indoor navigating agent needs to efficiently understand the geometric structure of its local environment in order to act. A common scene understanding approach is to generate a set of hypotheses about the geometric structure of the indoor environment and then test the hypotheses to select the one with the highest rank, from a single image [1, 4, 5, 6] or from a continuous stream of images (e.g. a video) [8, 9]. These methods simply detect features (e.g. lines [1, 4, 6], points [8, 9], and edges [6]) that are easily detectable for evaluating the hypotheses. In fact, some of the most informative features to discriminate the hypotheses may not be extracted if features are detected by fixed thresholds, since the informative regions may not have high image contrasts for features to be detected.

This paper demonstrates that by focusing attention on features in the informative regions, we can evaluate the hypotheses more efficiently. The idea of focusing on informative regions of the image space is inspired by the idea of saliency detection [2, 3, 7]. While these works typically define saliency regions based on image and motion properties of the pixels in the images [2, 3] or based on human fixations [7], our informative regions are defined in terms of the agent's own state of knowledge, the current set of hypotheses about the geometric structure of the indoor environment.

Given a set $\mathbf{M}$ of hypotheses, we divide the image into regions based on the expected information gain that each feature provides, which we call *informativeness* (Figure 1). We define the informativeness $I(p_j, \mathbf{M}) \in [0,1]$ of point p_j, measuring its discriminating power among the set $\mathbf{M}$ as,

$$I(p_j, \mathbf{M}) = \log(|\mathbf{M}|) - H(\mathbf{M}^u | p_j), \quad (1)$$

where $H(\mathbf{M}^u | p_j)$ is the expected entropy of the set $\mathbf{M}$ with uniform prior.



Figure 1: Example of the informative regions. (Best viewed in color.) (**Left**) The current set of hypotheses. (**Middle**) The gray-scale value reflects the informativeness $I(p_j, \mathbf{M}) \in [0,1]$ of each pixel p_j in the current image based on the four hypotheses. Since the hypotheses are qualitatively distinctive, the image divides into several regions based on the informativeness. (**Right**) Since precisely computing the exact boundary of the informative regions can be computationally expensive, we use a set of axis-aligned boxes to approximate these regions. All points within each box are set to the same $I(p_j, \mathbf{M}) > 0$ value (maximum informativeness among all pixels within the box), and any point that is outside the boxes has $I(p_j, \mathbf{M}) = 0$.

To evaluate the set $\mathbf{M}$ of hypotheses, existing approaches extract high-contrast features across the entire image. However, efforts are being wasted when features with high image contrasts lie within uninformative regions, and opportunities may be missed when features in informative regions have relatively low image contrasts. Thus, in this paper, we adjust the threshold for extracting features in the informative regions to allow features to be extracted even if they have lower image contrasts. Moreover, when evaluating the hypotheses, instead of using all extracted features, we only use features that are capable of discriminating among the current set of hypotheses to reduce the computational cost.

We selected a Bayesian filter-based approach to scene understanding [8] to evaluate our attention focusing method. We compare the effectiveness of our method with the baseline [8] by computing the informativeness of the selected features at each frame. Figure 2 shows that with the same set of hypotheses, our method selects more features that are capable of discriminating the hypotheses and wastes no effort on features that are uninformative. In addition, our experimental results demonstrate that this bias of the search toward the most informative point features helps the Bayesian filter to converge to a single hypothesis more efficiently, without loss of accuracy. About 50% of the time, our method converges to a single hypothesis while only about 30% of the time, the baseline method converges to a single hypothesis.

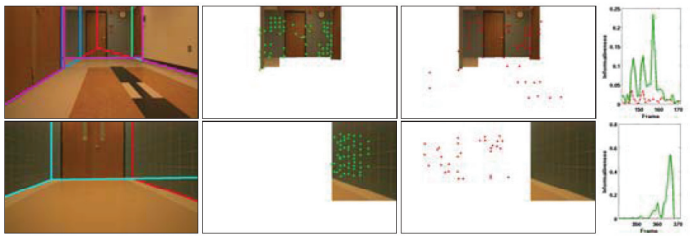


Figure 2: Examples of our attention focusing method. (Best viewed in color.) Each row is a snapshot from our experiment. The first column is the set of hypotheses $\mathbf{M}$ at that frame. The second column visualizes the informative point features (green) $\mathbf{P}_a$ that are selected by our method to evaluate the hypotheses. The third column shows the point features (red) from the baseline set $\mathbf{P}_b$, which are simply point features with high corner responses. For the second and the third column, only the informative regions are shown, and non-informative regions are shown in white. The last column is a comparison of the informativeness of using each feature set. Our proposed attention focusing method $I(\mathbf{P}_a, \mathbf{M})$ is shown in green solid lines, and the baseline method $I(\mathbf{P}_b, \mathbf{M})$ is shown in red dashed lines. Our method achieves higher informativeness because more point features that are capable of discriminating the hypotheses are tracked. In general, there are 1.5 to 6.5 times more point features that are capable of discriminating the hypotheses in the informative set $\mathbf{P}_a$ than in $\mathbf{P}_b$. In some extreme cases (second row), the baseline set $\mathbf{P}_b$ does not contain any features that are capable of discriminating the hypotheses so the informativeness $I(\mathbf{P}_b, \mathbf{M})$ at those frames are zero.

Our main contribution is to show that by using informativeness to control the process of feature acquisition, we can use computational resources more efficiently to discriminate among hypothesized interpretations of a visual scene, with no loss of accuracy. Informativeness allows our method to focus computational resources on regions in the scene where different hypotheses make different predictions. We demonstrate our method using the problem of real-time scene understanding for a mobile agent (e.g. [8, 9]), but it is equally applicable to other scene understanding problems (e.g. [1, 4, 5, 6]). Our experimental results demonstrate that this bias of search towards informative features provides more discriminating power among the hypotheses than simply using features that are easy to detect, with no loss of accuracy.

[1] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. *ICCV*, 2009.

[2] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007.

[3] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998.

[4] David Changsoo Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. *CVPR*, 2009.

[5] David Changsoo Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. *NIPS*, 2010.

[6] Scott Satkin, Jason Lin, and Martial Hebert. Data-driven scene understanding from 3d models. *BMVC*, 2012.

[7] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 2009.

[8] Grace Tsai and Benjamin Kuipers. Dynamic visual understanding of the local environment for an indoor navigating robot. *IROS*, 2012. Dataset: `www.eecs.umich.edu/~gstsai/release/Umich_indoor_corridor_2012_dataset.html`.

[9] Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using Bayesian filtering with motion cues. *ICCV*, 2011.

Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition

Tomas McCandless
tomas@cs.utexas.edu
Kristen Grauman
grauman@cs.utexas.edu

Department of Computer Science,
University of Texas at Austin

Motivation Egocentric computer vision entails analyzing images and video that originate from a wearable camera, which is typically mounted on the head or chest. Seeing the world from this first-person point of view affords a variety of exciting new applications and challenges, particularly as today's devices become increasingly lightweight and power efficient. Nearly all applications of egocentric vision demand robust methods to recognize activities as seen from the camera wearer's perspective.

High-level representations based on detected objects are a promising way to encode video clips when learning egocentric activities [1, 3, 5]. In particular, recent work explores a "bag-of-objects" histogram of all objects detected in a video sequence, as well as an extension that captures the objects' relative temporal ordering [5]. Similarly, for third-person video, activity recognition methods often employ histogram pyramids with regular grid structures to pool local descriptors (e.g., [2, 4]). Coupled with standard discriminative classifiers, this representation shows very good recognition results.

Problem However, existing methods that pool localized visual features into space-time histogram bins do so using hand-crafted binning schemes. The problem with defining the spatio-temporal bins *a priori* is that they may not offer the most discriminative representation for the activity classes of interest. That is, the hand-crafted histogram bins may fail to capture those space-time relationships between the local features that are most informative for activity recognition.

Our idea To overcome this limitation, we propose to *learn* discriminative spatio-temporal histogram partitions for egocentric activities. Rather than manually define the bin structure, we devise a boosting approach (Figure 1) that automatically selects a small set of useful spatio-temporal pyramid histograms among a randomized pool of candidate partitions. To minimize a computationally expensive search, we further propose a way to meaningfully bias the partitions that comprise the candidate pool. We devise an *object-centric* cutting scheme that prefers sampling bin boundaries near objects involved in the egocentric activities. As a result, we focus the randomized pool of space-time partitions to the egocentric setting while also improving training efficiency.

Approach Our approach works as follows. Given a set of egocentric training videos labeled according to their activity class, we first run object detectors on the frames to localize any objects of interest—both those that are "passive" and those that are "active" in an interaction with the camera wearer. Active objects are those being manipulated by the user [5], while passive objects are those which lie inert in the frame background. We then construct a series of candidate space-time pyramids. Given this candidate pool, we compute the corresponding series of object histograms for each training video. Then, we apply multi-class boosting to select a subset of discriminative pyramid structures. At the end, we have a strong classifier that can predict the activity labels of new videos, using only those randomized pyramids selected by the learning algorithm.

While boosting gives an automated way to select informative pyramids, its training time depends linearly on the number of candidates we include in the pool. To avoid an excessive search, we focus the candidate pool in a way that is meaningful for egocentric data. Rather than sample cuts uniformly at random, our idea is to sample the cuts according to the distribution of active objects as they appear in the training videos (Figure 2). We refer to these as *object-centric cuts* (OCC).

Result summary We validate our method on the Activities of Daily Living (ADL) dataset [5], a large and realistic dataset consisting of 18 household actions, like doing laundry, washing dishes, etc. We compare our boosted RSTP+OCC approach to four baselines: 1) a standard bag-of-words (BoW) on space-time interest points with HoG/HoF visual words, 2) a bag-of-objects approach that replaces visual words with detected objects, 3) the Temporal Pyramid (TempPyr) developed in [5], which represents the state of the art performance for this dataset, and 4) a boosting approach just like the proposed method, but without the object-centric cuts (denoted RSTP, for randomized space-time pyramids).

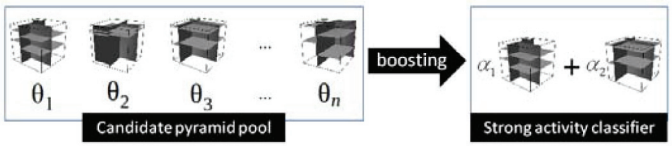


Figure 1: We take a pool of randomized space-time pyramids with object-centric cuts, and use boosting to select those that are most discriminative for egocentric activity recognition. Each pyramid is used to histogram the object occurrences across space and time, revealing the primary interactions of the camera wearer.



(a) Object-centric cuts



(b) Uniformly random shifts

Figure 2: Example partitions using either object-centric (a) or uniformly sampled randomized cuts (b). For display purposes we show cuts on 2D frames, but all cuts are 3D in space-time. The proposed object-centric cuts focus histograms on human-object interactions.

As shown in Table 1, our approach outperforms all four baselines and improves the state of the art. Notably, our proposed object-centric cuts are essential for our strong recognition result. Simply using boosting with purely randomized partitions (RSTP) is noticeably weaker. This shows it is useful to bias bins according to object interactions for egocentric data.

Main contribution Our main contribution is two-fold. We show how to learn the most discriminative partition schemes for spatio-temporal binning in action recognition, and we introduce object-centric cuts for egocentric data. Our proposed approach improves on the current state of the art for recognizing activities of daily living from the first person viewpoint.experiments demonstrate the positive impact of taking active object locations into account via object-centric cuts.

BoW	Bag-of-objects	TempPyr [5]	RSTP	RSTP+OCC
16.5%	34.9%	36.9%	33.7%	**38.7%**

Table 1: Overall classification accuracy on ADL. Our method, RSTP+OCC, improves the state of the art.

[1] A. Behera, D. C. Hogg, and A. G. Cohn. Egocentric activity monitoring and recovery. In *ACCV*, 2012.

[2] J. Choi, W. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *ACM Multimedia*, 2008.

[3] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[5] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

Learning to approximate global shape priors for figure-ground segmentation

Daniel Kuettel
dkuettel@vision.ee.ethz.ch
ETH Zurich

Vittorio Ferrari
vferrari@staffmail.ed.ac.uk
University of Edinburgh



Figure 1: *Training samples drawn from a horse shape prior.*

Many problems in computer vision are formulated as minimizing the energy of a discrete graphical model [2]. This involves designing an energy function to model the problem, and then minimizing it to find the lowest energy labelling. The energy minimization framework allows to cleanly separate modelling and inference. However, it involves a trade-off between expressiveness and optimizability. On the one hand we want sophisticated energy functions which model the problem accurately. On the other hand, we want to find the global optimum of the energy over all possible labellings of its variables. Unfortunately, general efficient minimization algorithms only exist for restricted classes of energy functions. A very popular such class are binary *pairwise* functions [2, 5]. The optimization of higher-order functions is still an open, highly challenging problem [1, 6, 7, 8, 9, 10]. A recent trend is to design energies containing higher-order potentials of a particular form, along with an optimizer specialized for that form [1, 3, 4, 6, 8, 9, 10]. The downside of this approach is that different optimizers needs to be invented for each form of higher-order potential, which requires great skill and knowledge. In this paper, we propose an alternative approach which does not require inventing specialized minimization algorithms. The designer only provides an *arbitrary* energy function E that suits her problem.

Our key idea is to *automatically* learn a pairwise function $\tilde{E}$, which tries to approximate an arbitrary energy function E as well as possible. After learning the parameters of this *proxy* function, it can be efficiently minimized using standard algorithms for pairwise models. We make no assumptions about the form of E. It can be arbitrarily complex and could possibly contain a global potential defined over all variables. The approximation $\tilde{E}$ instead is limited to unary and pairwise potentials. It is learned to hold as well as possible over the entire domain $\tilde{E}(\mathbf{x}) \cong E(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}^N$. The key advantage of learning a pairwise approximation $\tilde{E}$ is that it can be minimized efficiently.

Our goal is to learn the parameters Φ of $\tilde{E}$ so that $\tilde{E}(\mathbf{x};\Phi) \cong E(\mathbf{x})$ for $\forall \mathbf{x} \in \mathcal{X}^N$. We simply treat $E(\mathbf{x})$ as a *black box* mapping a configuration $\mathbf{x}$ to an energy. In principle, with no assumption about the structure of $E(\mathbf{x})$, we would have to evaluate every possible labeling $\mathbf{x} \in \mathcal{X}^N$ to do this. Unfortunately this is not feasible, as there are 2^N different $\mathbf{x}$. We relax the problem by requiring $\tilde{E}$ to approximate E on a very large subset of sample labellings $\tilde{\mathcal{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathcal{X}^N$ (typically millions). The underlying assumption is that E varies smoothly over its domain $\mathcal{X}^N$, and therefore the approximation is good also for most other labelling. This is a weak assumption, as smoothness holds for most practical energy functions. Thus we require $\tilde{E}(\mathbf{x}_k;\Phi) \cong E(\mathbf{x}_k), 1 \leq k \leq K$

We cast learning Φ as a least squares minimization problem

$$\min_\Phi \sum_{k=1}^{K} \left(\tilde{E}(\mathbf{x}_k;\Phi) - E(\mathbf{x}_k)\right)^2 \qquad (1)$$

We show how to efficiently learn the parameters Φ on millions of samples from E.

We apply this idea to figure-ground segmentation, where variables x_i are pixels and the labels correspond to foreground ($x_i = 1$) and background ($x_i = 0$). In this context, $E(\mathbf{x})$ is a global shape prior giving lower energy to segmentations $\mathbf{x}$ that fit a shape class such as horses or mugs. In our experiments we use the smallest Chamfer distance to a set of exemplar shapes as E

$$E_c(\mathbf{x}) = \min_{\mathbf{s} \in \mathcal{S}} \left(\frac{1}{|\partial \mathbf{x}|} \sum_{x \in \partial \mathbf{x}} \min_{s \in \partial \mathbf{s}} d(x,s) + \frac{1}{|\partial \mathbf{s}|} \sum_{s \in \partial \mathbf{s}} \min_{x \in \partial \mathbf{x}} d(x,s) \right) \qquad (2)$$

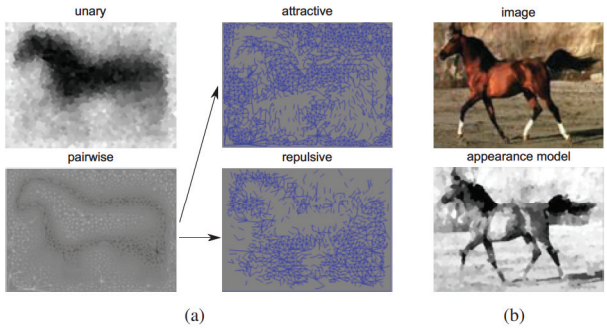


Figure 2: *(a) Learned proxy model for the horse shape prior. (b) Appearance model for a particular test image (darker = more foreground).*

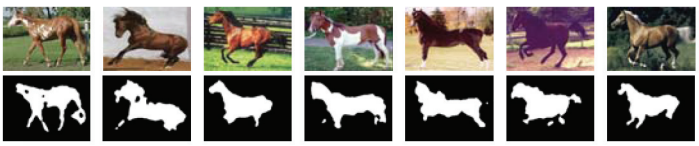


Figure 3: *Example segmentations using a combination of proxy and appearance model.*

where $\mathcal{S}$ is the set of exemplars, d is the Euclidean distance, and ∂ denotes the outline of a segmentation (i.e. the pixels at the boundary between foreground and background).

We conduct our experiments for horse figure-ground segmentation. We build a min-Chamfer prior for horse shapes based on eq. 2. Fig. 1 shows training samples from this shape prior. Fig. 2(a) shows the proxy model learned from E. Our experiments confirm that the proxy closely mimics the original shape prior and separates well between good and bad segmentations.

The shape prior is then combined with an appearance model for horses, see fig. 2(b) and applied to horse test images in fig. 3. When simplifying the min-Chamfer prior to be asymmetric and unnormalized it can be globally optimized with a slow algorithm. For this prior we observe that our learned proxy model achieves virtually the same segmentation accuracy (82.7% versus 83.0%) but at a much lower runtime cost. It also reaches an energy close to the global optimum. Finally, we experiment with the full min-Chamfer prior, that cannot be globally optimized. In this case, the accuracy of the learned proxy improves further to 86.6%. This shows that it is advantageous to use a more advanced shape prior and then solve it approximately.

[1] C. Arora, S. Banerjee, P. Kalra, and S. Maheshwari. Generic cuts: An efficient algorithm for optimal inference in higher order mrf-map. In *ECCV*, 2012.

[2] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, 2001.

[3] T. Kim, K. Lee, and S. Lee. Nonparametric higher-order learning for interactive segmentation. In *CVPR*, 2010.

[4] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[5] Vladimir Kolmogorov and Carsten Rother. Minimizing nonsubmodular functions with graph cuts – a review. *IEEE Trans. on PAMI*, 29(7):1274–1279, 2007.

[6] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *CVPR*, 2009.

[7] L. Ladicky, C. Russel, P. Kohli, and P.H.S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[8] S. Ramalingam, P. Kohli, K. Alahari, and P. Torr. Exact inference in multi-label crfs with higher order cliques. In *CVPR*, 2008.

[9] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, 2009.

[10] V. Vineet, J. Warrell, and P. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In *ECCV*, 2012.

What is a good evaluation measure for semantic segmentation?

Gabriela Csurka
gabriela.csurka@xrce.xerox.com
Diane Larlus
diane.larlus@xrce.xerox.com
Florent Perronnin
florent.perronnin@xrce.xerox.com

Computer Vision Group
Xerox Research Centre Europe
6, chemin de Maupertuis
38920 Meylan, France
http://www.xrce.xerox.com/

Motivation. The goal of semantic segmentation is to assign each pixel of a photograph to one of several semantic class labels, or to none of them. Semantic segmentation has many potential applications including scene understanding, removing undesired objects from photographs, copy-pasting objects from one photograph to another, or local class-based image enhancement. These diverse applications might have different requirements when it comes to judging whether the semantic segmentation algorithm has made "a good job". For instance, for the first application it might be sufficient to segment the scene into rough blobs. On the other hand, in computer graphics applications, having a precise delineation of the contours is important.

Ideally, the success of the segmentation algorithm should be measured by the success of the end application. As this is generally too difficult to evaluate, the computer vision community has resorted to application-independent measures of accuracy. In this paper, we raise the following question: *what is a good semantic segmentation measure?* and show that the answer is not as trivial as it sounds.

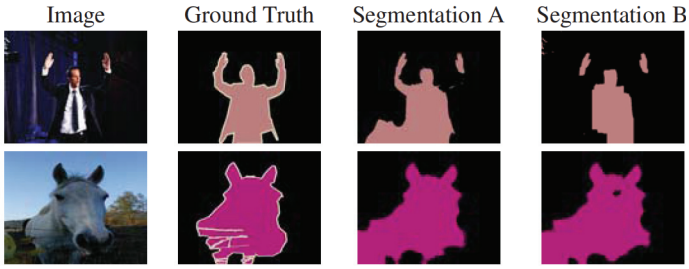


Figure 1: Which one of these two possible segmentations is better? People participating to our study found segmentation A to be more accurate (*i.e.* "closer" to the ground truth segmentation) than segmentation B. Do the segmentation evaluation measures agree ?

Contributions.

First, we draw the attention of the community to the evaluation question that has been well-studied for unsupervised segmentation, but that has been largely overlooked for semantic segmentation. We review the few existing semantic segmentation measures.

Supported by experimental evidences, we show that different segmentation algorithms can be optimal for different segmentation measures.

We propose to measure segmentation accuracy on a per-image basis rather than on the dataset as a whole. This allows visualising the distribution of scores, analysing the statistical differences between algorithms, evaluating methods on a specific image and hence performing user studies.

We propose a new measure based on contours by adapting a segmentation measure initially introduced for unsupervised segmentation [3].

Finally, we perform a user-study – the first we are aware of – to understand how semantic segmentation measures correlate with human preference and we use it to explore a possible combination between region-based and contour-based measures.

Existing segmentation measures. Existing measures are region-based. They consider segmentation as a pixel-level classification problem and to evaluate it using a pixel level confusion matrix accumulated over the entire dataset. Typically, overall (OA) and per-class accuracies (PC) [4] or Jaccard index (JI) [1] are computed. One of the only alternatives is the Trimap (TO) [2] that focuses on the boundary regions.

Proposed contour-based measure. We propose to extend the F1-measure of [3] to semantic segmentation: We make it class-dependent by computing one value per class by comparing the corresponding binarized segmentation maps, and we average the per-class scores over all classes present either in the ground-truth or in the predicted segmentation.

Segmentation algorithms. We consider five different segmentation algorithms in our evaluation. We start from a simple model that only enforces a global but no local consistency. We then add more and more sophisticated consistency terms, which tend to produce more and more precise contours. This enables evaluating the impact of gradually more complex models on the different segmentation measures. The two patch-based classification model are denoted by P and P+MS, and the three CRF models are denoted GCRF, DCRF and DCRFMS.

A first glance at the results. To illustrate the importance of choosing the evaluation measure correctly, we show in Table 1 results obtained for the segmentation methods mentioned above, when evaluated with the standard evaluation measures used to compare semantic segmentations (OA, PC, JI, TO). We can easily notice that the ranking of the methods is highly dependent on the evaluation measure.

	P	P+MS	GCRF	DCRF	DCRFMS
OP	70.7	**74.6**	72.6	**75.0**	**75.2**
PC	43.3	**43.9**	**43.7**	**42.1**	41.9
JI	25.4	**28.1**	26.6	**27.2**	**27.3**
TO	45.4	**54.9**	51.9	**54.2**	**55.3**

Table 1: Pascal VOC 2011. Segmentations are obtained by five different methods. The scores of the different evaluation measures are coloured according to their ranks with green for the highest, blue for the second and red for the third one.

To conclude, we suggest a **list of recommendations:**

- A single segmentation measure does not tell the whole story. Use different measures to assess the different capabilities of your algorithm.
- To allow a fair comparison, the segmentation algorithm parameters should be optimized for each measure.
- Per-image scores should be preferred as they allow a more detailed comparison of methods including statistical significance test, user study, per image reasoning.
- Below a given threshold (*e.g.* < 0.5 for JI) the segmentation scores do not seem to correlate anymore with the perceived quality of a segmentation. Hence, care should be taken when drawing conclusions from such low-scoring images.

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge. http://www.pascal-network.org/challenges/VOC.

[2] P. Kohli, L. Ladický, and P. H S Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(1):302–324, 2009.

[3] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE PAMI*, 26(1), 2004.

[4] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

Harmonic Variance: A Novel Measure for In-focus Segmentation

Feng Li
http://www.eecis.udel.edu/~feli/
Fatih Porikli
http://www.porikli.com/

Mitsubishi Electric Research Labs
Cambridge, MA 02139, USA

Detecting in-focus regions in a low depth-of-field (DoF) image has important applications in scene understanding, object-based coding, image quality assessment and depth estimation because such regions may indicate semantically meaningful objects. Most early approaches consider image sharpness as an image in-focus metric and attempt to measure in-focus directly from high-frequency components, e.g. edges [2] and patch variances [1]. However, in-focus edges with low intensity magnitudes and defocus edges with high intensity magnitudes may give similar responses. Thus, recent work formulates in-focus detection either as a learning task [5] or an optimization problem [7].

Here, we introduce an efficient in-focus measure to estimate the degree of in-focus within an image region and a graph segmentation based regularization method to accurately align the feature responses on the underlying image structure.

A harmonic mean of variances, which we call as *harmonic variance*, is computed from the statistical properties of the given image in its bandpass filtered versions. To find the bandpass responses, we apply the 2D discrete cosine transform (DCT) on image patches, e.g. in 8×8 windows, with a fixed number of bands $M = 64$. These DCT transform responses are bandpass filtered individually and then aggregated into the bandpass filtered image I_m for the DCT channel m.

Note that, even if most of the variances σ_m^2 of a defocus region may have small values, one large variance estimate in a low-frequency channel, which is quite common for textured yet blurry regions, could still make the arithmetic mean of the variances larger than that of in-focus regions, causing the defocus region to be falsely considered as in-focus. A competent in-focus measure should not be sensitive to such outliers. For positive data sets containing at least one pair of unequal values, the harmonic mean is always the smallest of the three means, while the arithmetic mean is the greatest of the three and the geometric mean is in between.

Thus, we use the harmonic mean instead of other possible first order statistics in order to combine the variances σ_m^2 of different bands. For a given patch centered at $\mathbf{x}$, we define the in-focus measure as the harmonic mean of the variances $\sigma_m^2(\mathbf{x})$

$$h(\mathbf{x}) = \left[\frac{1}{M-1} \sum_{m=1}^{M-1} \frac{1}{\sigma_m^2(\mathbf{x})} \right]^{-1} \tag{1}$$

omitting the variance of the zero-frequency channel σ_0^2. This in-focus measure is also closely related to the noise analysis of natural images [6] by a factor α as:

$$\sigma_\eta^2(\mathbf{x}) = \alpha(\mathbf{x}) \cdot h(\mathbf{x}) \,, \tag{2}$$

where σ_η^2 is the image noise variance and

$$\alpha(\mathbf{x}) = \frac{\sqrt{\kappa(\mathbf{x})} - \frac{1}{M-1}\sum_m \sqrt{\kappa_m(\mathbf{x})}}{\sqrt{\kappa(\mathbf{x})}} \,, \tag{3}$$

where κ and κ_m are the kurtosis values of the original image I and bandpass filtered image I_m. In other words, for an in-focus image, the harmonic mean can be used to determine local image noise variances.

We incorporate the harmonic variance measure within a robust segmentation framework based on a graph Laplacian spectrum constraint [4] to segment in-focus regions in low DoF images.

We compute a graph Laplacian matrix L [3] from the image I. The graph Laplacian spectrum constraint $Lf = 0$ enforces the image structure on the prior information, that is the harmonic variance results, in the data fidelity term $\|f-h\|^2$. With this constraint, the optimal f lies in the null-space of L. In other words, f is constant within each connected component of the corresponding graph G. Therefore, we define the in-focus segmentation as a least-squares constrained optimization problem as

$$\min_f \|f-h\|^2 \quad \text{s.t.} \quad Lf = 0. \tag{4}$$

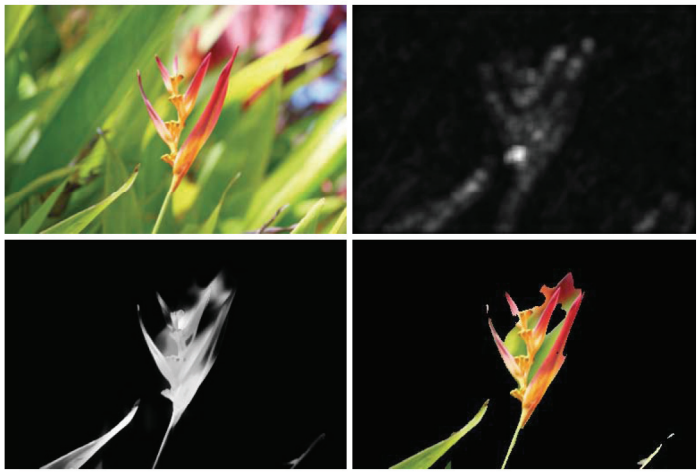


Figure 1: Top-row: original image and its harmonic variance features h. Bottom-row: the optimized in-focus likelihood scores f of h and the final segmentation. Unlike the conventional features, the harmonic variance is an accurate indicator of in-focus regions (please, see the full paper for comparisons). In addition, our method fits the harmonic score responses to the underlying image structure.

The residual $\delta(\mathbf{x}) = |f-h|$ has many spatially continuous outliers. The least-squares fidelity team with equal weights can distort the final estimate in case of outliers. We adapt a robust functional instead of the least-squares

$$\min_f \rho(f-h) + \beta \|Lf\|^2 \,, \tag{5}$$

where ρ is the robust function. We use the Huber function

$$\rho(\mathbf{x}) = \begin{cases} 1 & \text{if } \delta(\mathbf{x}) < \varepsilon \\ \varepsilon/\delta(\mathbf{x}) & \text{if } \delta(\mathbf{x}) \geq \varepsilon \end{cases} \tag{6}$$

for the reason that it is a parabola in the vicinity of 0 and increases linearly when δ is large. Thus, the effects of large outliers can be eliminated significantly. When written in a matrix form, the data fidelity term can be simplified as $\|W(f-h)\|^2$ where W represents the Huber weight function. As a result, the problem Eq.(5) can be solved efficiently in an iterative least-squares approach. At each iteration, the optimal f is updated as

$$f = (\beta L^\top L + W)^{-1} W h \,. \tag{7}$$

In conclusion, the effectiveness of this novel harmonic variance measure coupled with the image structure alignment capability of the graph Laplacian spectrum based robust optimization framework generates highly accurate segmentation results of in-focus regions in low DoF images.

[1] C. Kim. Segmenting a low-depth-of-field image using morphological filters and region merging. *IEEE Trans. on Image Processing*, 14(10): 1503–1511, 2005.

[2] E. Krotkov. Focusing. *International Journal on Computer Vision*, 1: 223–237, 1987.

[3] A. Levin, A. Rav Acha, and D. Lischinski. Spectral matting. *IEEE Trans. PAMI*, 30(10):1699–1712, 2008.

[4] F. Li and F. Porikli. Laplacian spectrum: A unifying approach for enforcing point-wise priors on binary segmentation. *Submitted to ICCV*, 2013.

[5] R. Liu, Z. Li, and J. Jia. Image partial blur detection and classification. *CVPR*, 2008.

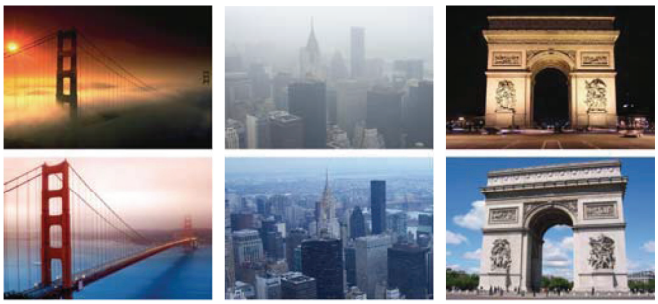
[6] X. Pan, X. Zhang, and S. Lyu. Exposing image splicing with inconsistent local noise variances. *ICCP*, 2012.

[7] Y. Tai and M. Brown. Single image defocus map estimation using local contrast prior. *ICIP*, 2009.

Robust Image Matching with Line Context

Wei Guan
wguan@usc.edu
Suya You
suyay@graphics.usc.edu

Computer Science Department
University of Southern California
Los Angeles, CA, USA



(a) (b) (c)

Figure 1: The challenging cases for image matching in different illuminating environments. The photos on top are taken in (a) the sunlight (b) a rainy day (c) night time. The photos at bottom are taken in a normal daytime.

Some features are specially designed to cope with complex illumination variations [1, 2, 3, 4, 5]. These features are based on intensity orders rather than values so that their descriptors are invariant to non-linear intensity changes. However, one limitation of such features is that they are only able to handle monotonic illumination changes, as the examples shown in Figure 1. In this paper, we try to cope with this problem by proposing a new type of local feature, which describes the context information of neighboring line segments.

Through extensive experiments and based on existing research, we found that the best pixel grouping level to handle large illumination variations is using curves or lines. Since curves require more sophisticated groupings and can be approximated by multiple line segments, we use line segments as the primitives of our feature descriptor. We call the proposed feature *Line Context*, since it is inspired by *Shape Context*.

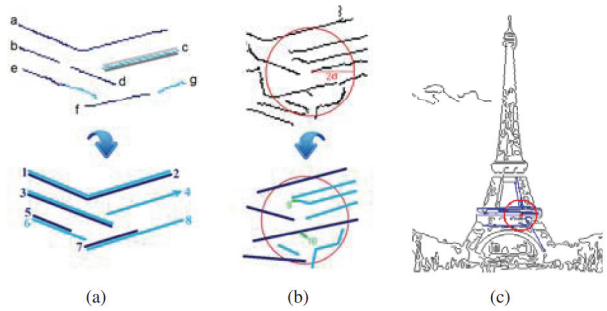
Line Context Detector. It is well known that edges are present at various scales. To detect edges at different scales we use multi-scale Canny edge detector with Gaussian derivatives at several pre-selected scales. Two phases are used to remove unstable edges. The first phase is to apply Laplacian operator. Those edges that do not attain a distinctive extremum over scales will be removed. The second phase is to remove those edges with homogeneous gradients, i.e. the points where the underlying curves have zero curvatures. We apply Harris matrix to achieve this purpose.

Segments in Context. The edge pixels are linked to connected curves at different scales. These curves are then fitted by straight line segments. As shown in Figure 2-(a), several cases need to be considered for representing curves with line segments. One curve may be fitted by multiple segments like *curve a*. Two segments with small gap in between are merged into one larger segment no matter they are on the same scale (*curve b* and *d*) or different scales (*curve f* and *g*). In the latter case, the merged segment only exists in the lower scale (*segment* 8). Besides, all the segments in higher levels are also segments (*segment* 1, 2 and 3) or part of segments (*segment* 5 and 7) in lower levels.

For each keypoint, we need to find line segments in its neighborhood, which is also called *context* of the feature. The line segments lying inside or partially inside the context are called *context segments*. The initial scale σ provides an estimate size of searching area. Let $d_{point-set}(v,S)$ be the shortest distance from point v to a set of points S. With keypoint k and all-segments $\{seg_i\}$, the segments set L in the context is defined as,

$$L(k) = \{seg_i | d_{point-set}(k, seg_i) \leq 2\sigma\} \quad (1)$$

Figure 2-(b) shows an example of context segments. All the segments in scale level σ and *lower* scales are included in the context as long as part of the segment is within distance 2σ. Note that segments with very small lengths are removed because these segments are unstable on different images (*segment* 9 and 10). The feature descriptor will be formed by the remaining segments. Figure 2-(c) shows a keypoint and its context segments on an image after edge detection.



(a) (b) (c)

Figure 2: (a) Curves at various scales are fitted by line segments in different situations. Dark blue denotes higher scale level, and light blue denotes lower scale level. (b) Line segments within distance 2σ are considered context segments. (c) A keypoint and its context segments on an edge image.

Scale and Orientation. The *Line Context* feature is designed to be scale and rotation invariant. Let m_i be the midpoint of segment seg_i in the context of keypoint k, $d_{point-point}(a,b)$ be distance between two points a and b. The scale s for the feature is defined as,

$$s(k) = \frac{\sum_{seg_i \in L(k)} d_{point-point}(k, m_i)}{|L(k)|} \quad (2)$$

During descriptor calculation process, all the context segments of keypoint k will be normalized by a factor of $s(k)$.

Each keypoint descriptor is assigned a canonical orientation so that the descriptor is invariant to rotations. This orientation is determined by the dominant orientation of *context segments* in the *lowest* edge scale.

Line Context Descriptor. We use multiple sampled points as the representation of segments in the context. Four parameters are used to describe each sample point. They are 1) the distance r to the keypoint, 2) the angle $\alpha \in [0, 360)$ between the direction from keypoint to sample point and reference direction (keypoint dominant orientation), 3) the angle $\beta \in [-180, 180)$ between reference direction and the orientation of underlying segment and 4) the underlying segment scale σ. After sampling, all the sample points will be used to form the keypoint descriptor.

Our goal is to produce a compact descriptor for each keypoint by computing a coarse histogram of the relative coordinates of segment sample points. We use a log-polar-like coordinate system to vote for the relative distances. The accumulated weights from all sample points form a 3D descriptor. Many bins have 0 vote which are similar to SIFT descriptors. Through experiments, we found that it is usually not necessary to cover segments in all scales. The scale σ_0 and one level lower are good estimations for most cases.

[1] B. Fan, F. Wu, and Z. Hu. Aggregating gradient distributions into intensity orders: A novel local image descriptor. In *CVPR*, 2011.

[2] R. Gupta, H. Patak, and A. Mittal. Robust order-based methods for feature description. In *CVPR*, 2010.

[3] F. Tang, S. Hwan Lim, N.L. Chang, and H. Tao. A novel feature descriptor invariant to complex brightness changes. In *CVPR*, pages 2631–2638, 2009.

[4] Z. Wang, B. Fan, and F. Wu. Local intensity order pattern for feature description. In *ICCV*, 2011.

[5] Y. Yu, K. Huang, W. Chen, and T. Tan. A novel algorithm for view and illumination invariant image matching. *IEEE Transactions on Image Processing*, 21(1):229–240, 2012.

Sketch Recognition by Ensemble Matching of Structured Features

Yi Li
http://www.eecs.qmul.ac.uk/~yl303
Yi-Zhe Song
http://www.eecs.qmul.ac.uk/~yzs
Shaogang Gong
http://www.eecs.qmul.ac.uk/~sgg

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

Sketch recognition aims to automatically classify human hand sketches of objects into known categories. This has become increasingly a desirable capability due to recent advances in human computer interaction on portable devices. The problem is nontrivial because of the sparse and abstract nature of hand drawings as compared to photographic images of objects, compounded by a highly variable degree of details in human sketches.

Current methods for sketch recognition and sketch-based image retrieval all employ a bag-of-features (BOF) representation of object sketches without considering their spatial structures. A characteristic of sketch is that its basic strokes do not necessarily exhibit strong discriminative cues in isolation. Instead, the structures of a sketch both locally and holistically contain informative visual cues for discriminating different sketches.

In this work, we exploit a star graph as a structured feature representation to encode both local features and the holistic structure of a sketch. We also exploit the ensemble matching for computing the distance metric when comparing star graph representations of different sketches [6]. While star graphs can represent holistic structure of sketches well, current BOF approaches [2, 4, 5] have the benefit of being able to better capture subtle structural details. To that end, we further propose a unified framework to address both holistic structural variations and local detail differences. More specifically, we introduce a separate category filtering process as a first step prior to ensemble matching to keep only a few categories most similar in local details, utilizing Support Vector Machines (SVM) classification on BOF.

For ensemble matching, we consider it as a graph matching problem, where each ensemble is encoded as a star graph. More precisely, we denote a star graph as $G = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V}$, $\mathcal{E}$, $\mathcal{A}$ represent respectively a set of nodes, edges and attributes of the graph. In particular, $\mathcal{V} = \{v_i\}_{i=1}^{N_s} \cup c$ is the set of all N_s sample points $\{v_i\}_{i=1}^{N_s}$ and the center c, and $e_i \in \mathcal{E}$ is the implicit link between v_i and c. Moreover, $\mathbf{a}_{ic} \in \mathcal{A}$ represents the geometrical relationship between v_i and c, and $\mathbf{a}_i \in \mathcal{A}$ denotes the corresponding feature descriptor of v_i.

The computation of the similarity between ensemble q (query) and t (target) is formulated as follows:

$$P(G^q, G^t) = \sum_i P(\mathbf{a}_i^t|\mathbf{a}_i^q) P(\mathbf{a}_{ic}^t|\mathbf{a}_{ic}^q) \quad (1)$$

where $G^q = (\mathcal{V}^q, \mathcal{E}^q, \mathcal{A}^q)$ and $G^t = (\mathcal{V}^t, \mathcal{E}^t, \mathcal{A}^t)$ are their corresponding star graphs. $P(\cdot,\cdot)$ denotes the normalized distance metric value and is considered as probability. The feature similarity term $P(\mathbf{a}_i^t|\mathbf{a}_i^q)$ accounts for the similarity between features and the feature location correlation term $P(\mathbf{a}_{ic}^t|\mathbf{a}_{ic}^q)$ stands for the location correlation between two features.

We modify traditional ensemble matching [6] in several minor ways. First, similar to [1], we employ a two steps ensemble matching algorithm to accelerate the matching process. It first finds the most similar D target features $\{\mathbf{a}_j^t\}_{j=1}^{D}$ for each feature in the query (D is much smaller than the total feature amount in the target), then calculates location correlations only for these D features. The similarity between the query and the target is then:

$$P(G^q, G^t) = \sum_i \max_j P(\mathbf{a}_j^t|\mathbf{a}_i^q) P(\mathbf{a}_{jc}^t|\mathbf{a}_{ic}^q) \quad (2)$$

Second, the sum rule is employed to obtain the overall matching score (c.f. Equation (1)) instead of the product rule employed in [1], as it is proven to be the most resilient to estimation errors [3]. Third, we get rid of the center estimation and multi-scale matching, due to the single subject essence of the sketch images and pre-scale procedure for the sketches. And the center of the ensemble is set to the geometrical center of the sketch. Fourth, to ensure the matching score between two sketches is constant, two sides comparison is employed by swapping the query and the target. And a partial matching penalty factor is added to penalize the matching score according to how many points in the target is not matched. The final matching score is then:

$$P^f(G^q, G^t) = w_1 * P(G^q, G^t) + w_2 * P(G^t, G^q) \quad (3)$$

where, w_i is the proportion of points being matched in the current target.

For category filtering, we employ SVM classifiers to filter sketch categories prior to ensemble matching, therefore keep N categories closest to the query in term of local details other than holistic structure. More specifically, we represent a sketch by a n-dimensional BOF histogram $\mathbf{h}$. A set of SVM classifiers are trained with respect to the number of sketch categories in a training dataset. For a probe sketch image to be classified, the following voting function classifies a given probe sketch image into the ith category:

$$c^i(\mathbf{h}) = \sum_j w_j^i K(\mathbf{s}_j^i, \mathbf{h}) + b \quad (4)$$

where K is a kernel function, $\mathbf{s}_j^i$ are the support vectors, w_j^i are weights, and b is the bias. $c^i(\mathbf{h})$ is therefore the classification response measuring similarity between the probe and the ith category.

The conclusion is that by integrating ensemble matching into sketch recognition, the structure information of a sketch can be effectively represented using a star graph, and an unified ensemble matching with multi-SVM classification based category filtering benefits from both holistic structure and subtle local details of sketches.

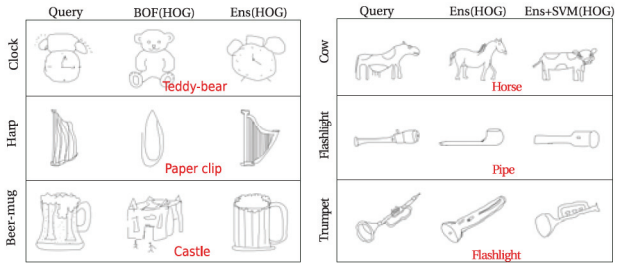


Figure 1: (a) qualitative comparisons of top retrieval results of three sketch probes on K nearest neighbors classification on bag-of-features (BOF(HOG)) and ensemble matching only method (Ens(HOG)). Ensemble matching preserves better holistic structure correspondence. (b) qualitative comparisons of top retrieval results of three sketch probes on ensemble matching only (Ens(HOG)) and ensemble matching with category filtering (Ens+SVM(HOG)). Category filtering helps to address subtle structural details.

[1] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJVC*, 74(1):17–31, 2007.

[2] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, pages 1025–1028, 2010.

[3] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.

[4] E. Mathias, H. Kristian, B. Tamy, and A. Marc. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 17(11): 1624 1636, 2011.

[5] E. Mathias, H. James, and A. Marc. How do humans sketch objects? *ACM TOG (Proceedings SIGGRAPH)*, 31(4):44:1–44:10, 2012.

[6] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, pages 1–8, 2007.

Learnable Stroke Models for Example-based Portrait Painting

Tinghuai Wang[1]
Tinghuai.Wang@surrey.ac.uk
John Collomosse[1]
J.Collomosse@surrey.ac.uk
Andrew Hunter[2]
Andrew.Hunter@hp.com
Darryl Greig[2]
Darryl.Greig@hp.com

[1] Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, UK
[2] Hewlett Packard Laboratories.
Stoke Gifford,
Bristol, UK

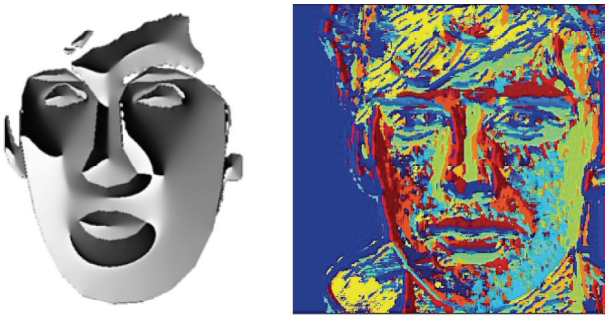
The stylization of photographs into high quality digital paintings remains a challenging problem in computer graphics. In recent years, sophisticated *painterly rendering* algorithms have been proposed that rely increasingly upon Vision to interpret structure and drive the rendering process. Although such algorithms generate a pleasing aesthetic for many image classes e.g. scenic shots, they typically perform poorly on portraits. The human visual system has a strong cognitive prior for portraits, and is particularly sensitive to distortion or loss of detail around facial features (Fig.1, right). Yet such artifacts are frequently observed when applying general purpose painterly algorithms to photographs of faces. High quality rendering of faces is important, as many scenarios for artistic stylization focus upon movie post-production, or consumer media collections, which predominantly contain people.



Figure 1: Portrait rendering using our algorithm (left), and a popular multi-resolution curved brush algorithm [2] (right). Facial details are preserved. Highlighting and stroke patterns were learned by example.

Our paper contributes a new stroke-based rendering (SBR) algorithm for stylizing photographs of faces into portrait paintings. SBR algorithms create paintings by compositing a sequence of curved spline strokes on a 2D canvas. In contrast to SBR algorithms that encode various rendering heuristics to target a particular artistic style, our algorithm learns the style of a human artist *by example*.

Given a photograph, and an ordered list of strokes (and related attributes) captured from a training session in which an artist paints that photograph, we are able to learn the artist's style and render previously unseen photographs of faces into portraits with a similar aesthetic. To enable our two core contributions; robust learning of styles, and the synthesis of high quality portraits, we harness Computer Vision to parse visual structure from the source image and drive our rendering process (Fig.2).



(a) (b)

Figure 2: Stroke parameters are learned local to each facial region according to visual structure.: (a) Orientation field and facial regions derived from ASM; (b) Codeword map inferred from dense SIFT and MRF.

Our algorithm is aligned with Image Analogies [1] and derivative techniques that learn non-parametric models of image filters from a pair of unfiltered and filtered greyscale images. Such systems are able to learn filters, including edge preserving filters reminiscent of a painterly effect, by sampling pairs of corresponding patches from the two images. The learned filter is applied by looking up patches from the new image. Our approach differs as we train at the level of the stroke, learning how the placement and appearance of each brush stroke is modulated according to underlying visual features in the training image. As such, our approach is specialized to the task of painting, enabling a wide variety of artistic styles. We specialize further to portraits by learning stroke models independently within semantic regions of the face, identified using an Active Shape Model (ASM) and Graph Cut. Image features are composed using a Markov Random Field (MRF) model to ensure spatial coherence of stroke style during learning and rendering. To the best of our knowledge our system is the first to explore portrait stylization by example at the level of the stroke.

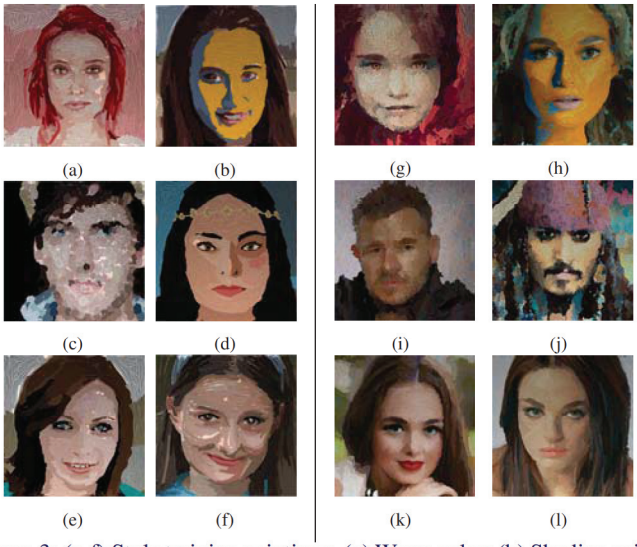


Figure 3: (a-f) Style training paintings: (a) Warm color; (b) Shading using complementary color; (c) Blobby strokes; (d) Thick, haphazard strokes; (e) Natural color, medium thick strokes; (f) Long strokes for hair. (g-l) Renderings using models learned from (a-f).

We are able to depict faces without the loss of salient detail exhibited by more general painterly methods [2] and without relying on a pre-painted arrangement of strokes to warp over the face as in the most closely aligned prior work by Zhao *et al.* [3]. In contrast to warping pre-captured stroke maps, we algorithmically place strokes as a function of image content presenting the first portrait painting algorithm to learn stroke style parameters by example (Fig.3).

[1] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *Proc. ACM SIGGRAPH*, pages 327–340, 2001.

[2] Aaron Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proc. SIGGRAPH*, pages 453–460, 1998.

[3] Mingtian Zhao and Song-Chun Zhu. Portrait painting using active templates. In *Proc. ACM NPAR*, pages 117–124, 2011.

Bayesian region selection for adaptive dictionary-based Super-Resolution

Eduardo Pérez-Pellitero[1, 2]
eduardo.perezpellitero@technicolor.com
Jordi Salvador[1]
jordi.salvador@technicolor.com
Javier Ruiz-Hidalgo[3]
j.ruiz@upc.edu
Bodo Rosenhahn[2]
rosenhahn@tnt.uni-hannover.de

[1] Image Processing Lab
Technicolor R&I
Hannover, Germany
[2] Institute for Information Processing
Leibniz Universität Hannover
Hannover, Germany
[3] Image Processing Group
Universitat Politècnica de Catalunya
Barcelona, Spain

The performance of learning-based Super-Resolution (SR) methods depends strongly on the content of the training dataset. In [3] the dictionary is built by randomly sampling raw patches from a large set of images regardless of the image to be recovered, hence relying on gathering sufficiently diverse patches so that they can generalize for any patch to be super-resolved. More recent follow-up works keep using the same strategy for the training, although these raw patches are compressed in a smaller number of patches through sparse coding techniques [4].

This paper proposes a novel sparse SR method, which focuses on adaptively selecting optimal patches for the dictionary training. Observing the fact that images usually contain a non-predictable group of different elements with their characteristic textures and edges (e.g. grass, rocks, fur, sand), we divide our training dataset into sub-image entities which we call regions, and extract descriptors in order to characterize them. The key idea is that, being these regions smaller, they have more consistent texture or edge content. For every patch to be super-resolved we find its best-fitting texture region from the training images by using the efficient local *Naive Bayes Nearest Neighbor* (NBNN), thus ensuring the obtained example pairs are highly correlated with the input low-resolution (LR) patches.

Each training image I_T is split in square regions R of size L_R. Given a patch y we find its training texture region R. Assuming a uniform region prior over R this can be achieved through a maximum likelihood decision rule:

$$\hat{R} = \underset{R}{\operatorname{argmax}}\, p(R \mid y) = \underset{R}{\operatorname{argmax}}\, p(y \mid R). \tag{1}$$

Let $\{f\} = f_1, f_2, \ldots, f_n$ denote the descriptors extracted from patch y. We use the Naive Bayes assumption, i.e. descriptors are independent, identically distributed:

$$\hat{R} = \underset{R}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(f_i \mid R). \tag{2}$$

In order to handle the high number of regions which a training set can contain, a local NBNN approach as the one in [2] is used.

Let R be some region and $\overline{R}$ the set of all other regions. The decision rule will be:

$$\hat{R} = \underset{R}{\operatorname{argmax}} \sum_{i=1}^{n} \log \frac{P(f_i \mid R)}{P(f_i \mid \overline{R})} + \log \frac{P(R)}{P(\overline{R})} \tag{3}$$

After finding a region R for every patch y, we will sample patches of size L_p with a certain overlap inside the selected regions and include them in HR and LR training sets T_l and T_h, which will train sparse adapted LR and high-resolution (HR) joint dictionaries D_l and D_h respectively.

In the paper, our method is compared with two well-known SR approaches: The original LLE-based method [1] and the sparse coding SR method of [4]. In our implementation, dictionaries are adaptively trained with the proposed method and used in [4]. The direct comparison is therefore explicitly showing the improvement by using our algorithm.

In order to prove the effectivity of our adaptive training scheme, two different testing scenarios and a single training set have been used. All the images have been upscaled by 2x and 3x magnification factors, computing the luminance peak-to-noise-ratio (PSNR) and the Structural Self Similarity (SSIM) as objective quality assesments.

Comparing our method to [1], the visual improvements are easily noticeable: our method is clearly super-resolving finer details resulting in sharper images. The objective PSNR and SSIM results support this qualitative visual impression.

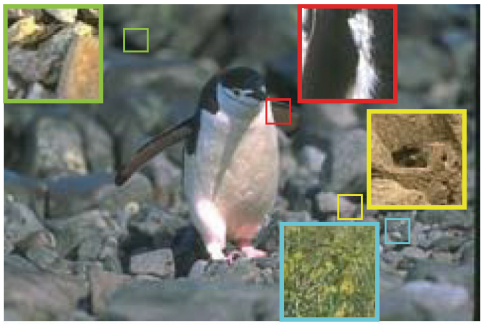


Figure 1: An example of how our method selects training regions based on Bayes rule. For every patch in the input image, a highly correlated region from the training images is found. The resulting set of regions is used to build a texture-adapted dictionary which will be used to recover the HR image.

With respect to the baseline method [4], careful visual inspection shows that ringing artifacts along the edges are mostly suppressed and certain edges are sharper. This is also suported in both PSNR and SSIM for the two magnification factors tested and in both testing set-ups. We observe that our method yields large improvement in performance when training and testing images are related (0.504 dB PSNR gain for 2x magnification factor), but there is also significant improvement for generic testing images (0.338 dB PSNR improvement for 2x magnification factor).

Our method improves performance for generic images, but also this improvement is accentuated when there exist training regions related to the testing images within the training dataset, therefore making our method specially adequate for applications where these conditions are met (e.g. video sequences, multiview scenarios, metadata-tagged images).

[1] H. Chang, D. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 275–282, 2004.

[2] S. McCann and D.G. Lowe. Local naive bayes nearest neighbor for image classification. *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3650–3656, 2012. ISSN 1063-6919.

[3] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR.*, pages 1–8, 2008. ISSN 1063-6919.

[4] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans. on Image Processing*, 19(11): 2861–2873, 2010.

Replicator Graph Clustering

Michael Donoser
http://vh.icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology, Austria

Motivation Data clustering, which aims at identifying a varying number of clusters and unique cluster assignments in a fully unsupervised manner, is an important topic in machine learning and computer vision. Many of the available methods in this field such as k-means or mean shift are based on a Euclidean assumption. In this work we overcome the shortcomings of such an assumption by considering the underlying similarity manifold using diffusion processes, which allows to handle non-metric data. Core idea of our novel approach is to combine an effective diffusion process, based on iteratively approaching evolutionary stable strategies from the field of game theory, with a provably optimal clustering step that analyzes a specific graph structure, that we denote as *Replicator Graph*.

Our clustering method (*Replicator Graph Clustering*) belongs to the field of pairwise or proximity-based based clustering approaches assuming that the input is an $N \times N$ affinity matrix $\mathbf{A} = (a_{ij})$, where each entry a_{ij} measures the similarity between two specific data points-to-be-clustered i and j. The goal of clustering is to uniquely assign each of the N data points to one of a set of clusters $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2, \ldots \mathcal{C}_C)$, where C is an automatically found number of clusters. Our method mainly consists of three subsequent steps (a) diffusing affinities by finding personalized evolutionary stable strategies of non-cooperative games (b) building a mutual k-nearest neighbor graph representing the underlying manifold and (c) applying a graph based clustering strategy which identifies the final clusters. Individual steps have low computational complexity which leads to an efficient clustering method, scaling well with an increasing number of data points.

Diffusing Affinities by Game Theory The first step of our method is to diffuse the provided affinities considering the underlying data manifold. For example, in [1] it was shown that such diffusion processes are able to significantly improve retrieval performance, and we aim at exploiting this property for our task of clustering. Our main idea is to consider each data point independently and to diffuse the affinities in relation to the current query data point, similar as in related semi-supervised learning [2] tasks. Query-specific diffusion means, that we aim at converting the query specific similarities $A_{i\cdot}$ (the i-th row of the given affinity matrix $\mathbf{A}$) into a new $N \times 1$ vector representation $\mathbf{x}^*$ by maximizing its agreement to the underlying similarity manifold spanned by the affinity matrix $\mathbf{A}$. For query-specific diffusion we adapt evolutionary dynamics from the field of game theory [4] to our specific requirements. Finally, all query-specific diffused affinities stacked together build the updated affinity matrix $\mathbf{A}^*$, which considers the similarity distribution on the underlying manifold. Algorithm 1 summarizes the diffusion process.

Algorithm 1: Algorithm for Affinity Diffusion

Input: $N \times N$ Affinity matrix $\mathbf{A}$ and number of neighbors S
Output: Updated $N \times N$ Affinity matrix $\mathbf{A}^*$

$\mathbf{A}^* = []$
Calculate S-nearest neighbors per query and build $\mathbf{A}_{SNN}$
for $i = 1$ **to** N **do**
  Apply replicator dynamics to $\mathbf{A}_{SNN}$ using $\overline{\mathbf{A}}_{i\cdot}$ as initialization and obtain evolutionary stable strategy $\mathbf{x}^{i*}$
  $\mathbf{A}^* = \begin{bmatrix} \mathbf{A}^* \\ [\mathbf{x}^{i*}]^T \end{bmatrix}$
return $\mathbf{A}^*$

Building Graph Structure and Clustering Based on the diffused affinity matrix $\mathbf{A}^*$, we build a specific graph structure denoted as mutual *kNN* graph (*mkNN*), which is surprisingly seldom used despite its interesting properties. The *mkNN* adds an additional constraint to the *kNN* graph, which requires that two connected data points belong to each others k-neighborhood, i. e. $j \in k\text{NN}(i)$ and $i \in k\text{NN}(j)$. In such a way we obtain our *Replicator Graph* by building the *mkNN* graph structure for the affinity matrix $\mathbf{A}^*$. Clustering is then done by analyzing th obtained *Replicator Graph* using a segmentation method proposed in [3], which precedes by merging clusters in decreasing order of the edge weights separating them. This is efficiently done by a variant of a minimum spanning tree method. The iterative merging stops once the final clustering is neither too fine (cluster should be merged due to high similarity) nor too coarse (cluster should be split due to boundary evidence). In such a way the algorithm obeys the global properties of being neither too fine nor too coarse and returns a global optimal (!) solution, although the algorithm only makes greedy decisions. For more details on the algorithm and proofs of the global optimality, see [3].

Experiments Experiments first analyze the influence of individual parameters and then demonstrate the efficacy of combining diffusion with the clustering approach. As final experiment we compare our proposed method on several data sets to related methods. We analyze three different data sets: (1) MPEG-7 CE Shape-1 Part-B (shape silhouettes, 1400 instances, 70 classes), (2) CALTECH 101 (object categories, 1010 instances, 101 clusters) and (3) UNIPEN (handwritten letters, 250 instances, 5 classes). We fixed the parameters for our method to $S = 35$ and $K = 10$. Table 1 shows results of the Normalized Mutual Information (NMI) score and runtimes on these data sets. As can be seen our proposed method provides competitive clustering results on all data sets in short computation time, e. g. being significantly faster than affinity propagation. Huge improvements against the k-Means baseline (+8.27% on MPEG-7, +15.21% on CALTECH 101 and +29.07% on UNIPEN) demonstrate the importance of considering the underlying manifold for clustering, which is implicitly done by the diffusion scheme.

Data	K-Means	Spectral	Aff.Prop	Proposed
MPEG	89.19 (**1.1s**)	95.99 (3.5s)	91.11 (14s)	**97.46** (2.7s)
CALT	41.19 (2.2s)	55.75 (1.3s)	55.65 (6.0s)	**56.40** (**0.7s**)
UNIP	42.59 (**0.1s**)	68.49 (0.1s)	59.58 (0.5s)	**71.66** (0.2s)

Table 1: Comparison of clustering quality (normalized mutual information – the higher the better) and runtime (in seconds) on several data sets.

Properties Our proposed clustering method has several important properties. First, we do not have any restrictions (like affinities have to satisfy metric properties) on the affinity matrix, e. g. negative and asymmetric affinity matrices can be handled. Second the diffusion step improves pairwise affinities due to considering the underlying manifold yielding improved clustering performance. Third, we are able to apply a provably optimal clustering method, which analyzes the widespread cluster criterium of high internal coherency and external incoherency. Fourth, we automatically identify the number of clusters using a single parameter to influence cluster granularity. Finally, we achieve improved clustering quality in comparison to state-of-the-art methods at low computation time in application fields like clustering object category images, handwritten letters and shape silhouettes.

[1] M. Donoser and H. Bischof. Diffusion processes for retrieval revisited. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[2] A. Erdem and M. Pelillo. Graph transduction as a noncooperative game. *Neural Computation*, 24(3):700–723, 2012.

[3] P. Felzenszwalb and F. Huttenlocher. Efficient graph-based image segmentation. *Intern. Journal of Computer Vision*, 59(2):167–181, 2004.

[4] P. Taylor and Jonker L.B. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40, 1978.

Of Gods and Goats: Weakly Supervised Learning of Figurative Art

Elliot J. Crowley
elliot@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk
Department of Engineering Science
University of Oxford

The objective of this paper is to automatically annotate the decorations on Greek vases with the gods and animals depicted there, given a dataset [1] of tens of thousands of images with associated text descriptions; such images often require expert labelling knowledge. Several papers have considered this 'words and pictures' problem [2] of automatically annotating image regions given only images and associated text. However, here there are additional challenges of: noisy supervision, non-naturalistic renderings, high intra-class variability, and low inter-class variability. Figure 1 shows a typical vase entry in the dataset.

Motivation. Automatic annotations of gods and animals is a very useful resource for archaeologists studying classical art, as assembling this type of material (e.g. all depictions of the god Zeus, aligned and size normalized from thousands of vases) manually would take quite some time. Apart from being a useful computer vision application, the method developed is directly applicable to other such art/archaeological collections with similar annotations, and the algorithm of progressive reduction of visual search space is useful in general for 'words and pictures' datasets.

Summary of method. To solve the annotation problem we propose a weakly supervised learning approach that proceeds in a number of stages. The key idea is that each stage strengthens the supervisory information available to allow for successful learning – this is akin to increasing the signal to noise ratio in signal processing. We proceed in three steps: (i) we use text mining methods to select sets of images that are visually consistent for a god depicted in a particular style (figure 2); (ii) we employ a form of multiple instance learning [5] to identify the image regions depicting the god in those images where he/she appears; (iii) the image regions are used to train a DPM sliding window detector [3] and all images in the dataset associated with the god can then be annotated by object category detection (figures 4 and 5).



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

A. Theseus and minotaur, with youths and women; B. Zeus seated, between women and onlookers;

Figure 1: **An example vase entry.** It consists of four images (a, b, c, d) and two text descriptions (A, B). Image (c) is a detail of (a) and both images correspond to description B. Similarly, image (d) is a detail of (b) and both correspond to description A. None of this information (correspondences, details) is provided. Affine transformations are automatically estimated between the images associated with each vase in order to determine and represent each vase by only the most zoomed images (c and d in this case).

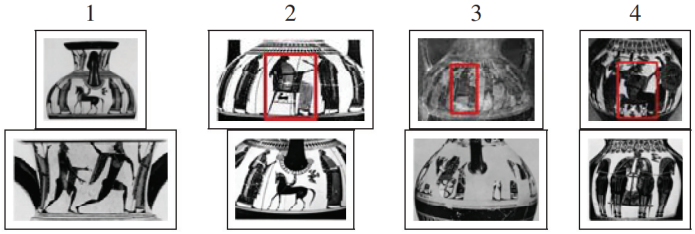


Figure 2: **Learning visually consistent clusters.** There is a large amount of variation in the depiction of each god. To overcome this difficulty we separate vase sets into groups corresponding to a similar depiction. We mine for verbs and non-person nouns in the text descriptions of each vase and greedily assign each vase to a cluster associated with such a word. A subset of the vases for the 'Zeus Seated' cluster can be seen above (4 vases shown from a cluster of 170 vases). Zeus, where present, is indicated by a red rectangle. Note that the data is noisy: if perfect, Zeus would be expected to be in one of the images for Vase no. 1 but he is not.

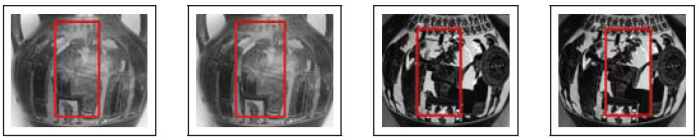
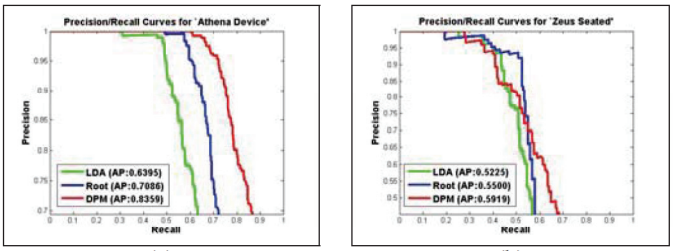


Figure 3: **Obtaining Candidate Regions**. To find the image regions within each visually consistent cluster containing the god we propose candidate regions and then determine which occur in other images within the cluster, but not outside of it by assessing the discriminative performance of an Exemplar-LDA based detector [4] trained on that region. The highest scoring regions are then re-ranked based on their visual consistency with each other and aligned. The windows corresponding to these top ranked detectors are suitable positive examples of the god. The regions obtained in this way for 'Zeus Seated' can be seen above.



(a) (b)

Figure 4: **Precision/Recall curves for (a) 'Athena Device' and (b) 'Zeus Seated'**. The green curves are for LDA models trained with the same positive examples as the DPM, the blue curves are for the DPM root-filters and the red curves are for the full DPM.



Figure 5: **Top Detections**. Above: Athena Device, Below: Zeus Seated.

[1] The Beazley Archive of Classical Art Pottery Database, July 2013. URL `http://www.beazley.ox.ac.uk/pottery`.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. 3:1107–1135, February 2003.

[3] P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. 2010.

[4] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Proc. ECCV*, 2012.

[5] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Proc. NIPS*, pages 570–576. MIT Press, 1998.

Multi-scale Joint Encoding of Local Binary Patterns for Texture and Material Classification

Xianbiao Qi[1]
qixianbiao@gmail.com
Yu Qiao[2]
yu.qiao@siat.ac.cn
Chun-Guang Li[1]
lichunguang@bupt.edu.cn
Jun Guo[1]
guojun@bupt.edu.cn

[1] Beijing University of Posts and Telecommunications, P.R. China
[2] Shenzhen Key Lab of CVPR, Shenzhen Institute of Advanced Technology, P.R. China

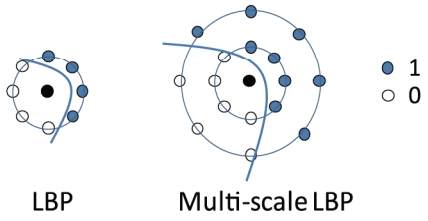


Figure 1: An illustration of LBP and Multi-scale LBP. Compared with LBP, LBPs in multiple scales jointly characterize stronger local structures. The LBP patterns in different scales have high correlation.

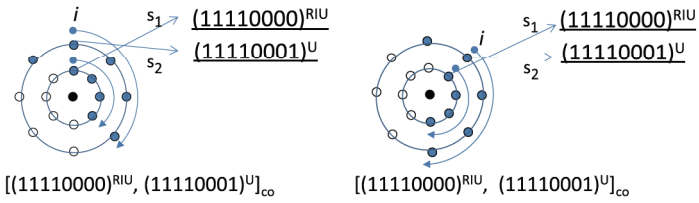


Figure 2: An illustration of the encoding method of MSJ-LBP and its rotation invariance. For the left image, we first compute its LBP^{RIU} pattern in the scale s_1 and determine the start point i which maximizes the binary sequence. According to the start point i, we can compute its LBP^{U} in the scale s_2. Thus, the joint pattern is $[(11110000)^{RIU}, (11110001)^{U}]$. Similarly, we can get the same joint pattern for the right joint pattern.

The Local binary pattern (LBP) [4] descriptor has achieved great success on texture and material classification due to its great computational efficiency and texture discrimination. Meanwhile, since its first publication, LBP has been widely applied to a lot of applications, such as face recognition, face detection, image retrieval, lip reading and many more [5]. A lot of LBP variants have been proposed in the past ten years. To achieve great rotation invariance, Ahonen et al. [1] propose an effective LBP histogram fourier (LBP-HF) features. Their method shows great robustness to image rotation. Besides of these two features, there are several powerful LBP variants, such as LTP, CLBP [3], LBPHF_S_M [7], LCP [2] and PRI-CoLBP [6], etc.

To depict texture information in different image resolutions, multi-scale strategy is introduced into texture and material classification. Firstly, single-scale LBP histogram features are extracted in each scale separately. Then, the histograms in each scale are concatenated into a final image representation. Similarly, the same multi-scale strategy are used by LCP, LBP-HF, DLBP and other LBP-based features. Since the multi-scale strategy always achieves much better performance than single scale, it is usually recognized as an indispensable means to achieve the state-of-the-art performance.

However, despite its effectiveness in texture and material classification, the classical multi-scale strategy ignores the correlation information between different scales. As shown on the left panel of Figure 1, each LBP pattern depicts a kind of local image structure. On the right panel, compared to single scale LBP, LBP patterns in multiple scales jointly depicts stronger local structure. In fact, texture patterns in different scales around the same central point usually have a strong correlation. Ignoring such correlation will lead to huge lose of discriminative information.

In this paper, we propose a Multi-Scale Joint encoding LBP (MSJ-LBP) feature to encode the joint distribution of LBP patterns in different scales for texture and material classification. Contrast to the classical multi-scale LBP (MS-LBP) that ignores the correlation between different scales, MSJ-LBP can effectively encode this kind of correlation. Compared to the single scale encoding, the multi-scale joint encoding strategy can depict stronger local structures. Meanwhile, the computational cost of MSJ-LBP is extremely low. In practice, the speed of MSJ-LBP is much faster than MS-LBP due to that MSJ-LBP uses fewer neighbors. In addition, the proposed feature is designed for rotation invariance, and excellent experimental results on the datasets with obvious image rotation demonstrate the robustness of the proposed feature to image rotation.

We have illustrated a detailed encoding strategy of the proposed MSJ-LBP in Figure 2. We can achieve the same co-occurrence encoding for the left and right pairs. In Figure 3, we have shown the experimental result of MSJ-LBP and several other methods on KTH_TIPS2a dataset.

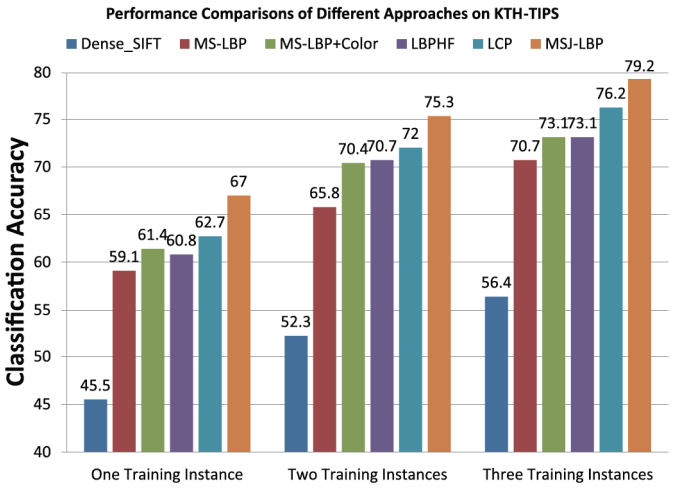


Figure 3: Experimental results of several methods on KTH-TIPS2a.

[1] Timo Ahonen, Jiří Matas, Chu He, and Matti Pietikäinen. Rotation invariant image description with local binary pattern histogram fourier features. In *Image Analysis*. 2009.

[2] Yimo Guo, Guoying Zhao, and Matti Pietikäinen. Texture classification using a linear configuration model based descriptor. In *BMVC*, 2011.

[3] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *TIP*, 2010.

[4] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 1996.

[5] M Pietikäinen, A Hadid, G Zhao, and T Ahonen. Computer vision using local binary patterns. Springer, 2011.

[6] Xianbiao Qi, Rong Xiao, Jun Guo, and Lei Zhang. Pairwise rotation invariant co-occurrence local binary pattern. In *ECCV*. 2012.

[7] Guoying Zhao, Timo Ahonen, Jiří Matas, and Matti Pietikainen. Rotation-invariant image and video description with local binary pattern features. *TIP*, 2012.

Spectral Imaging Using Basis Lights

Antony Lam[1]
http://research.nii.ac.jp/~antony
Art Subpa-Asa[2]
art@set.or.th
Imari Sato[1]
http://research.nii.ac.jp/~imarik
Takahiro Okabe[3]
http://www.pluto.ai.kyutech.ac.jp/~okabe
Yoichi Sato[4]
http://www.hci.iis.u-tokyo.ac.jp/~ysato

[1] National Institute of Informatics
Tokyo, Japan
[2] The Stock Exchange of Thailand
Bangkok, Thailand
[3] Kyushu Institute of Technology
Fukuoka, Japan
[4] The University of Tokyo
Tokyo, Japan

The spectral reflectance (SR) of objects provides innate information about material properties that have proven useful in applications such as classification, synthetic relighting, and medical imaging to name a few. General approaches to hyperspectral (HS) imaging include brute force capture of narrowband image stacks at consecutive wavelengths, use of conventional cameras with computational techniques for estimation of SR, use of specialized optics or filters, and active lighting. In particular, active lighting methods are fast and accurate. However, to our knowledge, past active lighting methods have not shown their illuminants to be optimal for SR recovery.

In our paper, we propose a new HS imaging method that utilizes optimally derived active illuminants with a standard monochrome camera. Our method is accurate and works even in the presence of unknown ambient lighting. Specifically, we make use of the well-known observation that spectra can be compactly represented by basis functions [1, 2, 3, 4]. We then show that spectral reflectance is analogous to a dot product between the light source spectrum and the surface's spectral reflectance distribution. This means that if we had a set of light sources with spectral distributions analogous to a set of basis vectors, the observed reflectance from projecting these lights onto the surface would give us coefficients. Such coefficients could then be used in conjunction with the basis vectors to recover the SR distributions of the surface.

Let us start with a definition of the reflectance model for a single surface point. Diffuse reflectance is modeled by the equation

$$I = \int s(\lambda) l(\lambda) d\lambda \tag{1}$$

where $s(\lambda)$ is the SR of the surface point at wavelength λ and $l(\lambda)$ is the illuminant at wavelength λ. From Eq. (1), we see that s and l are analogous to vectors and computation of its reflectance is analogous to a dot product. This means that if illuminant l were a basis vector, the observed reflectance I would be the projection of s onto l. In other words, I would be a basis coefficient.

So if we could project different illuminants l_m where each l_m would be equivalent to a different vector from the same basis, we could observe all the coefficients via reflectance. Having obtained all the coefficients, it would be possible to use the basis to reconstruct the SR s at the surface point. For the case of an entire surface, the same argument applies. If we could project such "basis lights" onto an entire surface and observe the reflected light using a camera, it would be possible to capture coefficients for all pixels in the scene. This would in turn permit us to recover the SR of all pixels.

The next question is, what conditions do the basis lights need to satisfy for optimal recovery using a camera? To answer this question we first define two relations.

The SR of a surface point can be expressed in terms of basis vectors as

$$s(\lambda) = \sum_{i=1}^{N} \sigma_n b_n(\lambda) \tag{2}$$

where $s(\lambda)$ is the SR at wavelength λ, $b_n(\lambda)$ is the n^{th} basis vector at wavelength λ, and σ_n is the coefficient associated with basis vector $\vec{b_n}$.

Next, we note that the the brightness $I_{p,m}$ of a given pixel p in an image taken using a grayscale camera and under lighting l_m follows the relation

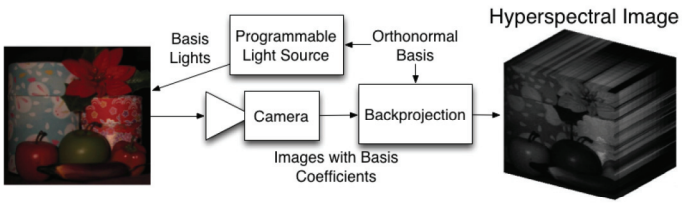


Figure 1: Flowchart of our Method

$$I_{p,m} = \int s_p(\lambda) c(\lambda) l_m(\lambda) d\lambda \tag{3}$$

where s_p is the spectral distribution at pixel p in the scene and c is the response of the camera. This is basically the same as Eq. (1) except the camera response is also taken into account.

If we substitute s_p from Eq. (3), with Eq. (2) we get the relation

$$I_{p,m} = \sum_{i=1}^{N} \sigma_n \int b_n(\lambda) c(\lambda) l_m(\lambda) d\lambda = \sum_{i=1}^{N} \sigma_n g_{n,m} \tag{4}$$

The intensities $\vec{I_p} = [I_{p,1} ... I_{p,M}]^T$ of pixel p under all lightings l_m from 1 to M can also be expressed in matrix form as

$$\vec{I_p} = \begin{bmatrix} g_{1,1} & \cdots & g_{N,1} \\ \cdot & & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & & \cdot \\ g_{1,M} & \cdots & g_{N,M} \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \cdot \\ \cdot \\ \cdot \\ \sigma_N \end{bmatrix} = G\vec{Q} \tag{5}$$

From Eq. (5), we see the basis coefficients $\vec{Q}$ can be solved for from the pixel intensities $\vec{I_p}$ and G. Ideally, if matrix G, has a condition number of 1, $\vec{Q}$ can be solved for reliably. We show in the paper that one way to satisfy this requirement is to use an orthonormal basis.

One issue is that an orthonormal basis would typically contain negative values, which are not physically plausible as light sources. In our paper, we show that negative light can be simulated and as an added benefit of this procedure, the effects of unknown ambient light can be canceled out. The experimental results also confirm that our method is able to accurately recover the spectral reflectance of entire scenes even under unknown ambient light.

[1] Te-Won Lee, Thomas Wachtler, and Terrence J. Sejnowski. The spectral independent components of natural scenes. In *First IEEE Intl. Workshop on Biologically Motivated Computer Vision*, BMVC, pages 527–534, London, UK, 2000. Springer-Verlag. ISBN 3-540-67560-4.

[2] Laurence T. Maloney. Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *JOSA A*, 3(10):1673–1683, Oct 1986.

[3] David H. Marimont and Brian A. Wandell. Linear models of surface and illuminant spectra. *JOSA A*, 9(11):1905–1913, Nov 1992.

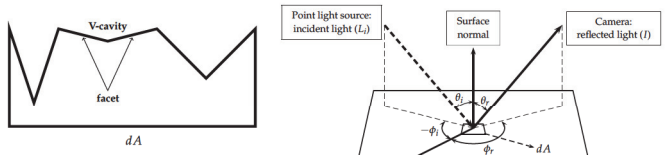
[4] J. P. S. Parkkinen, J. Hallikainen, and T. Jaaskelainen. Characteristic spectra of munsell colors. *JOSA A*, 6:318–322, 1989.

Generalised Perspective Shape from Shading with Oren-Nayar Reflectance

Yong Chul Ju[1]
ju@tu-cottbus.de
Silvia Tozza[2]
tozza@mat.uniroma1.it
Michael Breuß[1]
breuss@tu-cottbus.de
Andrés Bruhn[3]
bruhn@vis.uni-stuttgart.de
Andreas Kleefeld[1]
kleefeld@tu-cottbus.de

[1] Institute for Applied Mathematics and Scientific Computing
BTU Cottbus, Germany
[2] Department of Mathematics
Sapienza - University of Rome, Italy
[3] Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

In spite of significant advances in Shape from Shading (SfS) over the last years, it is still a challenging task to design SfS approaches that are flexible enough to handle a wide range of input scenes. In this paper, we address this lack of flexibility by proposing a novel model that extends the range of possible applications. To this end, we consider the class of modern perspective SfS models formulated via partial differential equations (PDEs) [5]. In contrast to existing approaches, however, we do not restrict our model by *either* choosing an advanced reflectance model [1, 3] *or* a general setup that allows an arbitrary position of the light source different from the camera centre [2, 7]. Instead, we propose a novel general model for perspective SfS that combines the advantages of both worlds.



(a) Facet model for surface patch dA consisting of many V-shaped Lambertian cavities.

(b) Diffuse reflectance for SfS with Oren-Nayar.

Figure 1: Sketch of the Oren-Nayar surface reflection model. From [3].

On the one hand, we consider the non-Lambertian reflectance model of *Oren* and *Nayar* [4]. By modelling rough surfaces via a Gaussian distribution of V-shaped cavities with standard deviation σ, this model allows us to handle advanced materials such as concrete, plaster, clay or cloth whose properties are considerably different from those of Lambertian ones (Fig. 1). On the other hand, we make use of a *spherical* coordinate system to parametrise the resulting brightness equation [2]. By placing the centre of this coordinate system at the position of the light source, this parametrisation allows us to consider realistic scenarios where the light source is not necessarily located at the centre of the camera (Fig. 2).

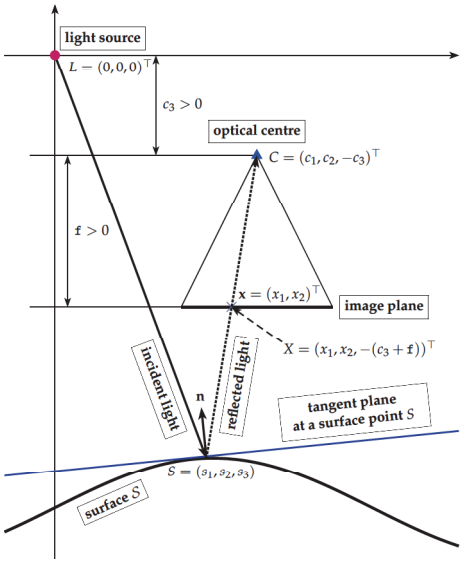
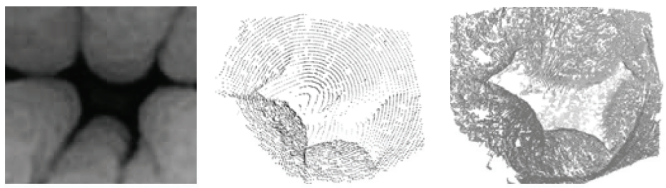


Figure 2: General SfS setup with arbitrary light source position. From [2].

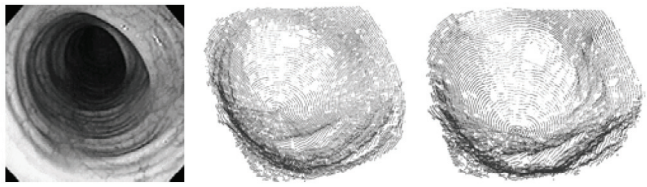
In contrast to the Lambertian case [2], the Hamilton-Jacobi equations (HJEs) that have to be solved for the Oren-Nayar model in spherical coordinates turn out to be significantly more complex. In fact, instead of solving a single HJE that is very compact, we have to distinguish *four* different cases with considerably more difficult equations. Nevertheless, following [2], we were able to derive an advanced numerical scheme of *fast marching* (FM) type [6] that allows for an efficient solution of the underlying HJEs. For the first time in the literature we thus succeed in modelling and solving an approach for perspective SfS that combines the advantage of *freely selecting* the position of the light source with the *robustness* of an advanced non-Lambertian reflectance model.

Experiments with medical real-world data demonstrate that our model offers the desired flexibility. As can be seen from Fig. 3 and Fig. 4, the quality of the reconstruction improves with finer grid sizes and the approach is stable under different choices of the roughness parameter σ.



(a) Input image. (b) $\delta_{\varphi\theta} = 0.025$. (c) $\delta_{\varphi\theta} = 0.0075$.

Figure 3: Reconstruction of gastric antrum with different grid sizes.



(a) Input image (211×208). (b) $\sigma = \frac{\pi}{6}$. (c) $\sigma = \frac{\pi}{2}$.

Figure 4: Reconstruction of duodenum with different roughnesses.

[1] A.H. Ahmed and A.A. Farag. A new formulation for shape from shading for non-Lambertian surfaces. In *Proc. CVPR*, 2006.

[2] S. Galliani, Y. C. Ju, M. Breuß, and A. Bruhn. Generalised perspective shape from shading in spherical coordinates. In *Proc. SSVM*, 2013.

[3] Y. C. Ju, M. Breuß, A. Bruhn, and S. Galliani. Shape from shading for rough surfaces: Analysis of the Oren-Nayar model. In *Proc. BMVC*, 2012.

[4] S.K. Nayar and M. Oren. Generalization of the Lambertian model and implications for machine vision. *IJCV*, 14(3):227–251, 1995.

[5] E. Prados and O.D. Faugeras. Perspective shape from shading and viscosity solutions. In *Proc. ICCV*, 2003.

[6] J.A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 2nd edition, 1999.

[7] C. Wu, S. Narasimhan, and B. Jaramaz. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *IJCV*, 86:(2-3):211–228, 2010.

High Frequency 3D Reconstruction from Unsynchronized Multiple Cameras

Yumi Kakumu Fumihiko Sakaue Jun Sato
yumi@cv.nitech.ac.jp,sakaue@nitech.ac.jp,junsato@nitech.ac.jp
Nagoya Institute of Technology, Japan

Kazuhisa Ishimaru
kazuhisa_ishimaru@soken1.denso.co.jp
NIPPON SOKEN, INC., Japan

Masayuki Imanishi
DENSO International America, Inc.

Stereo reconstruction generally requires image correspondence such as point and line correspondences in multiple images. Cameras need to be synchronized to obtain corresponding points of time-varying shapes. However, the image information obtained from synchronized multiple cameras is redundant, and has limitations with resolution. In this paper, we show that by using a set of unsynchronized cameras instead of synchronized one, we can obtain much more information on 3D motions, and can reconstruct higher frequency 3D motions than that with the standard synchronized cameras. As a result, we can attain super resolution 3D reconstruction from unsynchronized cameras.

In the standard reconstruction with multiple cameras, the cameras are synchronized, and observe the same set of M sequential points in 3D space as shown in Fig. 1 (a). Thus, the maximum frequency f_S of 3D motion, which can be recovered from camera images, is $f_S < \frac{1}{2}M$. If we observe the same 3D motion by using K unsynchronized cameras, we observe $K \times M$ different points in the 3D space as shown in Fig. 1 (b). As a result, all the $2KM$ observations from K cameras are independent of one another, unlike those from the synchronized cameras. Thus, we have a possibility of reconstructing 3D points up to $\frac{2}{3}KM$. Therefore, the maximum frequency f_U of recoverable 3D motion is $f_U < \frac{1}{3}KM$. Thus, we find that the unsynchronized cameras have a possibility of reconstructing $\frac{2}{3}K$ times higher frequency 3D motion than the synchronized cameras as follows:

$$f_U = \frac{2}{3}Kf_S \tag{1}$$

For example, if we have 3 unsynchronized cameras, we can recover 2 times higher frequency components than reconstruction with 3 synchronized cameras.

Unfortunately, the image points obtained from a set of unsynchronized cameras are not corresponding to each other as shown in Fig. 1 (b), and thus we cannot reconstruct 3D point by using the existing stereo methods. For reconstructing 3D motions from unsynchronized cameras, we consider a novel camera projection model, which projects 3D point in frequency space to image points in usual metric space.

Suppose the maximum frequency f of 3D motion Σ is less than $\frac{1}{2}N$. Then, 3D motion Σ can be fully represented by N frequency points $\mathbf{Z}_n = [X_n^f, Y_n^f, Z_n^f, \delta_n]^\top$ $(n = 0, \cdots, N-1)$. If we observe 3D motion Σ at M time instants, the observed 3D points $\mathbf{X}_t$ $(t = 0, \cdots, M-1)$ in real space can be represented by frequency points $\mathbf{Z}_n$ as follows:

$$\mathbf{X}_t = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{Z}_n\omega(t,n) \qquad (t = 0, \cdots, M-1) \tag{2}$$

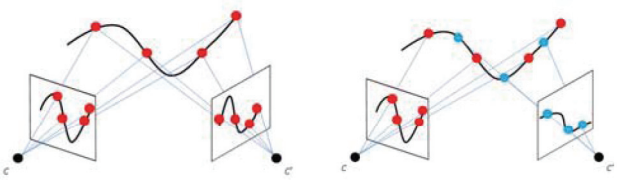
where, $\omega(t,n) = e^{2\pi jnt/N}$. Note, N and M are different in general, and we consider the case where the number of sampling, M, is less than the number of frequency components, N.

Now, if we observe 3D motion Σ by K different unsynchronized cameras at M time instants, their projections can be described as follows:

$$\lambda_t^i\mathbf{x}_t^i = \mathbf{P}^i\frac{1}{N}\sum_{n=0}^{N-1}\mathbf{Z}_n\omega(t+k_i,n) \qquad (i = 1, \cdots, K) \tag{3}$$

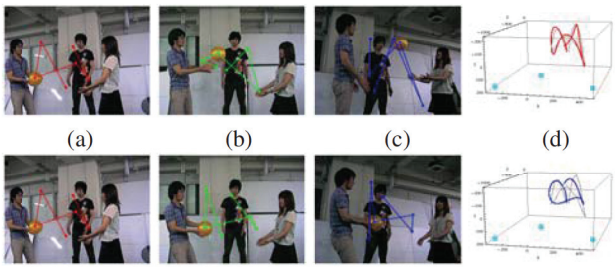
where, k_i denotes sampling delay of the ith camera with respect to the 1st camera. Thus, $k_1 = 0$. $\mathbf{P}^i$, $\mathbf{x}_t^i$ and λ_t^i are a camera projection matrix, a series of image points and their depth in the ith camera.

The equation (3) is a general projection model from 3D motion with N frequency components to K unsynchronized cameras with M time instants. The very important property of this projection model is that the identical 3D frequency point $\mathbf{Z}_n$ is projected to image points $\mathbf{x}_t^i$ in multiple cameras as shown in (3). This property exists in this projection model,



(a) synchronized cameras (b) unsynchronized cameras

Figure 1: Image projection in synchronized and unsynchronized cameras



(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 2: (a), (b) and (c) show images observed by 3 unsynchronized cameras, and (e), (f) and (g) show those observed by 3 synchronized cameras. The three people are passing a ball. The center of the ball is extracted in each camera at 6 time instants, and used for reconstruction. The points and lines show extracted points and their sequences. (d) and (h) show 3D points reconstructed from the proposed method and the standard method respectively.

since sampling delay $\omega(k_i,n)$ is separated from 3D frequency point $\mathbf{Z}_n$ in frequency space. As a result, 3D frequency point $\mathbf{Z}_n$ can be reconstructed from these image points.

By taking the vector product of (3) with $\mathbf{x}_t^i$, and reformulating it with respect to the 3D frequency components, $\mathbf{f} = [X_0^f, \cdots, X_{N-1}^f, Y_0^f, \cdots, Y_{N-1}^f, Z_0^f, \cdots, Z_{N-1}^f, 1]^\top$, we have the following system of linear equations:

$$\mathbf{M}\mathbf{f} = \mathbf{0} \tag{4}$$

where—$\mathbf{M}$ represents a $(2KM) \times (3N+1)$ matrix, whose components consist of $\mathbf{P}$, $\mathbf{x}$ and ω.

Assuming that sampling delays k_i and camera matrices $\mathbf{P}^i$ are known, 3D coordinates $\mathbf{f}$ of N frequency points $\mathbf{Z}_n$ can be computed from M image sequence of K unsynchronized cameras by solving the linear equation (4). Then, 3D point in real space can be recovered by transforming $\mathbf{Z}_n$ to $\mathbf{X}_t$ using (2).

Fig. 2 (d) shows 3D points reconstructed by the proposed method, and (h) shows the results from the standard method. As shown in these figures, the proposed method can fully recover original 3D motion, while the results from the standard method suffer from the aliasing problem and are imperfect.

[1] S. Agarwal, N. Snavely, I. Simon, S.M. Seitz, and R. Szeliski. Building rome in a day. In *Proc. ICCV*, 2009.

[2] O. D. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.

[3] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

Properties of Datasets Predict the Performance of Classifiers

Omid Aghazadeh
http://www.csc.kth.se/~omida
Stefan Carlsson
http://www.csc.kth.se/~stefanc

Computer Vision Group
Computer Vision and Active Perception Laboratory (CVAP)
Royal Institute of Technology (KTH)
Sweden

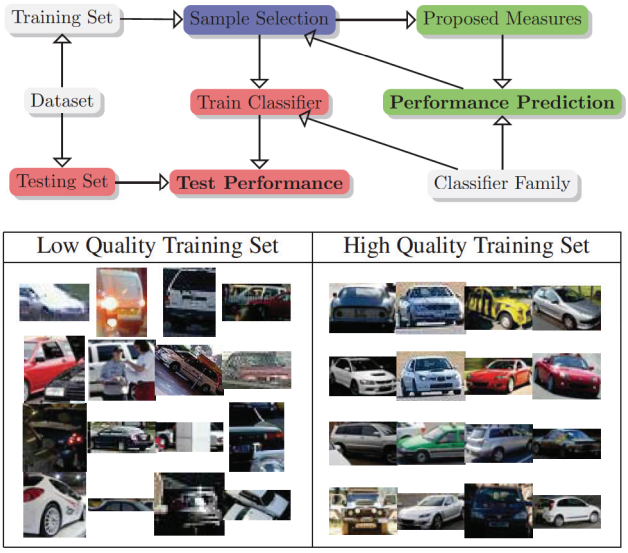


Figure 1: Top: illustration of the proposed procedure. The red boxes comprise the traditional training/testing procedure while the green boxes are proposed in this paper. Bottom: (right) illustration of automatic sample selection (the blue box) using the HOG feature. The low quality set (left) is intentionally generated for comparison. Both set are automatically generated from the "car" class of Pascal VOC 2007, using measures proposed in this paper.

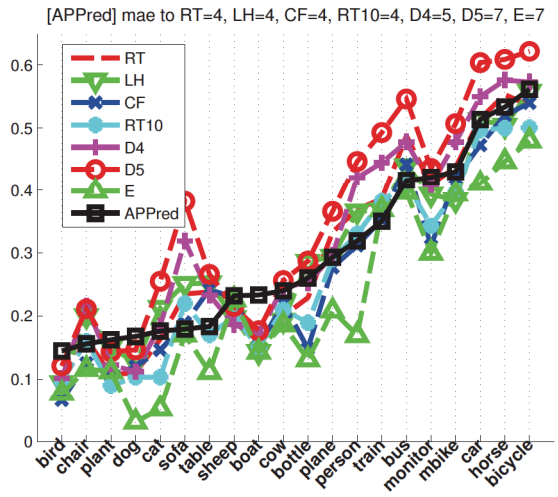


Figure 2: *Test Performance Prediction* of Pascal-VOC 2007 classes (APPred) and the performance of the reference methods. Best viewed electronically and in color.

It has been shown that the performance of classifiers depends not only on the number of training samples, but also on the quality of the training set [5, 6]. The purpose of this paper is to 1) provide quantitative measures that determine the quality of the training set, and 2) provide the relation between the test performance and the proposed measures.

The measures are derived from pairwise affinities between training exemplars of the positive class and they have a generative nature. We use the visual structural similarity measure $K^E_{\text{MMI}}(.,.)$ proposed in [1], which performs feature selection via discriminative reasoning. We aggregate the similarity measures between positive exemplars on local, semi-global and global scales. On each scale we compute the first and second order moments of the quantities in question, and come up with 8 data describing measures.

We show that the performance of the state of the art methods, on the test set, can be reasonably predicted based on the values of the proposed measures on the training set. We assume the training and test sets to be the outcomes of the same underlying distribution and model the test performance as a function of a description of the training set:

$$AP^{(\mathcal{C})}_{\mathcal{M}} = \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}\right) + \varepsilon_{\tilde{f}_{\mathcal{M}}} \tag{1}$$

where $AP^{(\mathcal{C})}_{\mathcal{M}}$ is the test performance of the classifier family $\mathcal{M}$ on class $\mathcal{C}$ and $\mu^{(\mathcal{C}_{TR})}$ is a vector describing the training set of the class $\mathcal{C}$. We then assume a sigmoid structure for $\tilde{f}_{\mathcal{M}}$ and model the test performance for a reference set of classifier families $\mathcal{R}$ as

$$\tilde{f}_{\mathcal{R}}(\mathbf{w}_{\mathcal{R}};\mathbf{v}) = \left(1+\exp\left\{-\mathbf{w}^T_{\mathcal{R}}\mathbf{v}\right\}\right)^{-1} \tag{2}$$

where $\mathbf{v}$ is a vector describing a training set. Given a data set $\mathcal{D} = \{\mathcal{C}_1,\ldots,\mathcal{C}_D\}$, we solve for $\mathbf{w}^{(\mathcal{C}_{CV})}_{\mathcal{R}} = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w},\mathcal{C}_{CV})$ where

$$\mathcal{L}(\mathbf{w},\mathcal{C}_{CV}) = \sum_{\mathcal{M}\in\mathcal{R}} \sum_{\mathcal{C}\in\mathcal{D}\setminus\{\mathcal{C}_{CV}\}} \|AP^{(\mathcal{C})}_{\mathcal{M}} - \tilde{f}_{\mathcal{R}}(\mathbf{w};\mathbf{v}^{(\mathcal{C})})\|^2 + \lambda\|\mathbf{w}\|^2 \tag{3}$$

Afterwards, $\mathbf{w}^{(\mathcal{C}_{CV})}_{\mathcal{R}}$ is used to predict the test performance for $\mathcal{C}_{CV}$ and this cross-validating procedure is performed for all $D = 20$ classes of Pascal VOC 2007 [2].

Figure 1 summarizes the proposed procedure. Our approach can be seen as a fundamentally revised version of the earlier approaches that estimate the complexity of the classification problems such as [4]. The results of test performance predictions are shown in figure 2. To summarize the results, we find out that 1) the size of the training set is not a good predictor of the test performance. This essentially means that the 'big data' should meet some quality requirements in order to improve recognition performance. 2) among our measures, the 'connected variation' measure correlates (positively) stronger with the test performances than our measure of 'instra-class variation' (negatively) correlates with them. This suggests that 'big connected data' might rectify the (negative) effects of intra-class variation.

To conclude, this study proposes data-describing measures that link the quality of the training set to the test performance of classifiers. This essentially quantifies the claim on "Unreasonable effectiveness of data" [3] and makes it possible to automatically measure the "cleanness of the data" [6]. This implies that it should be possible to devise rules for the automatic selection of training data that maximize the quality of the training set and consequently increase the test performance.

[1] Omid Aghazadeh, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Mixture component identification and learning for visual recognition. In *European Conference on Computer Vision*, 2012.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[3] Alon Y. Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 2009.

[4] S. Singh. Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[5] Antonio Torralba and Alexei A. Efros. Unbiased Look at Dataset Bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[6] Xiangxin Zhu, Carl Vondrick, Deva Ramanan, and Charless C. Fowlkes. Do we need more training data or better models for object detection? In *British Machine Vision Conference*, 2012.

Multi-view Pictorial Structures for 3D Human Pose Estimation

Sikandar Amin[1] sikandar.amin@in.tum.de
Mykhaylo Andriluka[2] andriluka@mpi-inf.mpg.de
Marcus Rohrbach[2] rohrbach@mpi-inf.mpg.de
Bernt Schiele[2] schiele@mpi-inf.mpg.de

[1] Intelligent Autonomous Systems Group
Technische Universität München, Germany

[2] Computer Vision and Multimodal Computing
Max Planck Institute of Informatics
Saarbrücken, Germany

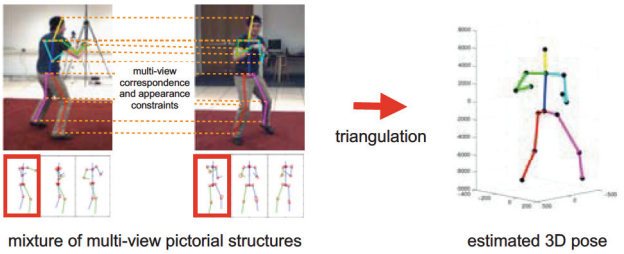
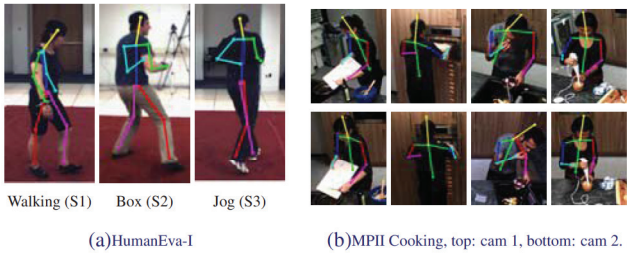


Figure 1: Overview of our approach to articulated 3D human pose estimation. Red boxes specify the selected component.



(a)HumanEva-I (b)MPII Cooking, top: cam 1, bottom: cam 2.

Figure 2: Example 3D pose estimation results from our approach (projected to 2D).

Pictorial structure models are the de facto standard for 2D human pose estimation. Numerous refinements and improvements have been proposed such as discriminatively trained body part detectors, flexible body models and local and global mixtures. While these techniques allow to achieve the state-of-the-art performance for 2D pose estimation, they have not yet been extended to enable pose estimation in 3D, instead this problem is traditionally addressed using 3D body models and involves complex inference in a high-dimensional space of 3D body configurations.

We formulate the articulated 3D human pose estimation problem as a joint inference over the set of 2D projections of the pose in each of the camera views. As a first contribution of this paper, we propose a 2D pose estimation approach that extends the state-of-the-art 2D pictorial structures model [6] with *flexible parts*, *color features*, *multi-modal pairwise terms*, and *mixtures of pictorial structures*. The second and main contribution is to extend this 2D pose estimation model to a multi-view model that performs joint reasoning over people poses seen from multiple viewpoints. The output of this novel model is then used to recover 3D pose.

We evaluate our multi-view pictorial structures model on HumanEva-I [8] and MPII Cooking [7] dataset. In comparison to related work for 3D pose estimation our approach achieves similar or better results while operating on single-frames only and not relying on activity specific motion models or tracking. Notably, our approach outperforms state-of-the-art for activities with more complex motions.

Single-view model: The pictorial structures model, originally introduced in [2, 3], represents the human body as a configuration $L = \{l_1, \ldots, l_N\}$ of N rigid parts and a set of pairwise part relationships E. The image position and absolute orientation of each part is given by $l_i = (x_i, y_i, \theta_i)$. We formulate the model as a conditional random field, and assume that the probability of the part configuration L given the image evidence I factorizes into a product of unary and pairwise terms:

$$p(L|I) = \frac{1}{Z}\prod_{n=1}^{N} f_n(l_n;I) \cdot \prod_{(i,j)\in E} f_{ij}(l_i,l_j), \qquad (1)$$

The part likelihood terms $f_n(l_n;I)$ are represented with boosted part detectors that rely on the encoding of the image using a densely computed grid of shape context descriptors [1]. We concatenate these shape context features with color features and learn a boosted part detector on top of this combined representation. Note that augmenting shape information with the color allows us to automatically learn the relative importance of both features at the part detection stage. The pairwise terms $f_{ij}(l_i,l_j)$ which encode the spatial constraints between parts are traditionally modeled with Gaussian distribution in the transformed space of the joint between two parts. We extend our model by introducing mixture models at the level of these pairwise part dependencies. To that end we replace the unimodal Gaussian with the term that maximizes over multiple modes and represent each mode with a Gaussian. Following [4, 5] we extend our approach to a mixture of pictorial structures models. We obtain the mixture components by clustering the training data with k-means and learning a separate model for each cluster. The components typically correspond to major modes in the data, such as various viewpoints of the person with respect to the camera. The index of the component is treated as a latent variable to be inferred at test time. We select the best component with the minimal uncertainty in the marginal posterior distributions of the body parts. In our experiments this approach worked slightly better compared to a trained holistic classifier that distinguishes the mixture component based on the contents of the person bounding box.

Multi-view model: To exploit the multi-view information we augument the model with appearance and spatial correspondence constraints across views. In order to estimate the 3D pose we proceed in two steps. In the first step we jointly infer the 2D projections of the 3D body joints across views exploiting multi-view constraints. In the second step, we recover the 3D pose by triangulation of the estimated 2D projections. For simplicity, we describe our multi-view model for the case of two views. For view m, let us denote the 2D body configuration as L_m and image evidence as I_m. According to Eq. 1 the single-view factors $f(L_1;I_1)$ and $f(L_2;I_2)$ representing the conditional posterior over body configurations decomposes into a product of unary and pairwise terms that define appearance and spatial constraints between parts independently for each view. The joint posterior over configurations in both views is given by

$$p(L_1,L_2|I_1,I_2) = \frac{1}{Z} f(L_1;I_1) f(L_2;I_2) \prod_n f_n^{app}(\mathbf{l}_n^1,\mathbf{l}_n^2;I_1,I_2) f_n^{cor}(\mathbf{l}_n^1,\mathbf{l}_n^2), \qquad (2)$$

The multi-view appearance factor f_n^{app} encodes the color and shape of the body part seen from multiple viewpoints. We define the joint appearance feature vector by concatenating the features from multiple views and train a boosted part detector using this representation. The multi-view correspondence factor f_n^{cor} encodes the constraint that part locations in each view should agree on the same 3D position. Given a pair of corresponding part locations $\mathbf{l}_n^1$ and $\mathbf{l}_n^2$ in each view and the projections of their reconstructed 3D point $\hat{\mathbf{l}}_n^1$ and $\hat{\mathbf{l}}_n^2$, we represent multi-view correspondence factor by $f_n^{cor}(\mathbf{l}_n^1,\mathbf{l}_n^2) = \exp(-(\|\mathbf{l}_n^1 - \hat{\mathbf{l}}_n^1\|^2 + \|\mathbf{l}_n^2 - \hat{\mathbf{l}}_n^2\|^2))$. When more than two views are available we connect the corresponding 2D body parts in all pairs of views. The posterior in Eq. 2 then includes multi-view appearance and correspondence factors for each pair of connected parts in all views as well as within-view spatial and appearance factors.

As for the single-view case, our multi-view model employs mixtures of pictorial structures, however, in this case the mixture components correspond to groups of poses consistent across views because they are learned from the projections of the same 3D poses.

[1] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, 2000.

[2] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[3] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[4] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[5] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[6] Andriluka Mykhaylo, Roth Stefan, and Schiele Bernt. Discriminative appearance models for pictorial structures. *IJCV*, 2011.

[7] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[8] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2), 2010.

Transitive Re-identification

Yulia Brand[1]
yulias@techunix.technion.ac.il
Tamar Avraham[2]
http://www.cs.technion.ac.il/~tammya
Michael Lindenbaum[2]
http://www.cs.technion.ac.il/~mic

[1] Electrical Engineering Department
[2] Computer Science Department

Technion - I.I.T.
Haifa 32000, Israel

Person re-identification (ReID) problem has lately received increasing attention especially due to its important role in surveillance systems, which should be able to keep track of people after they have left the field of view of one camera and entered the field of view of any overlapping or non-overlapping camera.

ReID accuracy can be significantly improved given a training set that demonstrates changes in appearances associated with the two non-overlapping cameras involved. Here we test whether this advantage can be maintained when directly annotated training sets are not available for all camera-pairs at the site. Given the training sets capturing correspondences between cameras A and B and a different training set capturing correspondences between cameras B and C, the Transitive Re-IDentification algorithm (TRID) suggested here provides a classifier for (A,C) appearance pairs; see Fig. 1.

TRID establishes a path between the non-directly trainable camera pair (A,C) by marginalization over the domain of possible appearances in camera B. Camera B plays the role of the 'connecting element' between cameras A and C. Our goal is to estimate the conditional probability $P(Y_{AC}|x_A,x_C)$ where the notations x_A and x_C refer to a feature vector describing the appearance observed by cameras A and C and the binary variable Y_{AC} gets the value 1 if and only if the appearances given in A and in C are of the same identity. When the feature vector is known to correspond to a particular individual of identity i, we denote it by x_A^i. Thus, the pair $\{(x_A^i,x_C^i)\}$ is a pair of feature vectors corresponding to the same person but to different cameras. Although a training set consisting of annotated pairs $\{(x_A^i,x_C^i)\}$ is not available, we can exploit the annotated sets $S_{AB}=\{(x_A^i,x_B^i)\}$, $i=1,...,n$ and $S_{BC}=\{(x_B^j,x_C^j)\}$, $j=n+1,...,n+m$. The desired conditional probability is expressed by:

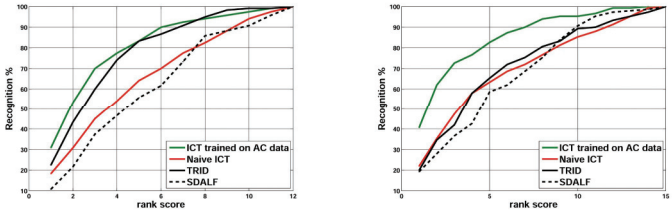
$$P(Y_{AC}|x_A,x_C)=\frac{\int_{x_B} P(Y_{AB}|x_A,x_B)P(Y_{BC}|x_B,x_C)f_{X_B}(x_B)dx_B}{1-\int_{x_B} P(\overline{Y_{AB}}|x_A,x_B)P(\overline{Y_{BC}}|x_B,x_C)f_{X_B}(x_B)dx_B} \quad . \tag{1}$$

In our experiments we found that the method is superior to that of a state-of-the-art camera invariant method, which used a fixed similarity measure (SDALF [2]). We compared the transitive TRID performance to direct learning from (A,C) examples using the ICT algorithm [1], which has shown state-of-the-art results for modeling the transfer of appearances associated with two specific cameras. We found that the TRID is also more accurate than a method that makes a naive use of training sets associated with (A,B) and (B,C) camera-pairs by just applying the ICT algorithm trained by the union of S_{AB} and S_{BC}; see Fig. 2. In certain conditions TRID even meets the performance achieved when direct (A,C) training data is available.



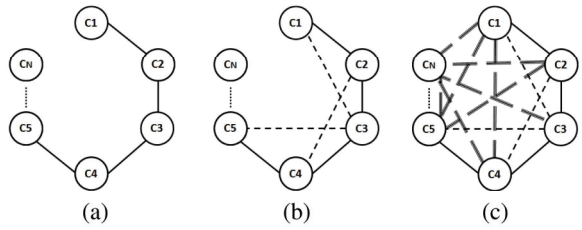
(a) (b)

Figure 1: The transitive ReID setup. Given the training sets S_{AB} ((a)left) capturing correspondences between people $1...n$ and a training set S_{BC} ((a)right) capturing correspondences between different people $n+1,...,m+n$, we would like to be able to classify (A,C) pairs (b) without an S_{AC} training set.



(a) [A B C]=[3 5 7] (b) [A B C]=[1 3 8]

Figure 2: CMC curves comparing the performance of TRID to that of other, non-transitive algorithms, over 2 combinations of camera-triplets from the SAIVT-SoftBio dataset [3] (see more results in the paper).



(a) (b) (c)

Figure 3: Illustration of a system consisting of N cameras $C_1,C_2,...,C_N$: (a) Inter-camera training sets are available only for $N-1$ *directly trainable pairs* (C_i,C_{i+1}), $i=1,...,N-1$. (b) The transitive algorithm infers classifiers for the *non-directly trainable pairs* (C_i,C_{i+2}). (c) Inter-camera classifiers for any other camera pair in the system can be deduced by recursively applying the transitive algorithm.

Given a camera network of N cameras, and *directly trainable pairs* (C_i,C_{i+1}), $i=1,...,N-1$ (see Fig. 3(a)), the proposed algorithm enables the inference of a classifier for the *non-directly trainable pair* (C_i,C_{i+2}); see Fig. 3(b). The inter-camera classifiers for any other camera pair can be deduced by recursively applying the transitive algorithm; see Fig. 3(c). Moreover, when adding camera $(N+1)$, only one inter-camera training set is required for establishing all correspondences. This approach significantly reduces the annotation effort inherent in a learning system, which goes down from $O(N^2)$ to $O(N)$.

To the best of our knowledge, this work is not only the first approach to transitivity in ReID but also, more generally, transitivity in domain adaptation. In our future work we intend to test the TRID algorithm on other domain adaptation tasks, and to examine alternative transitive approaches.

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *The 1st International Workshop on Re-Identification (Re-Id 2012), in conjunction with ECCV, LNCS*, pages 381–390, 2012.

[2] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117:130–144, 2013.

[3] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing: Techniques and Applications (DICTA 2012)*, pages 1–8, 2012.

Domain Adaptation for Upper Body Pose Tracking in Signed TV Broadcasts

James Charles[1]
j.charles@leeds.ac.uk
Tomas Pfister[2]
tp@robots.ox.ac.uk
Derek Magee[1]
d.r.magee@leeds.ac.uk
David Hogg[1]
d.c.hogg@leeds.ac.uk
Andrew Zisserman[2]
az@robots.ox.ac.uk

[1] School of Computing
University of Leeds
Leeds, UK

[2] Department of Engineering Science
University of Oxford
Oxford, UK

The objective of this work is to estimate upper body pose for signers in TV broadcasts. Given suitable training data, the pose is estimated using a random forest body joint detector (similar to that used by Shotton *et al*. [4], though without requiring depth measurements). However, obtaining such training data can be costly.

The novelty of this paper is a method of transfer learning that is able to harness existing training data and use it for new domains. Our contributions are: (i) a method for adapting existing training data to generate new training data by synthesis for signers with different appearances, and (ii) a method for personalising training data. In addition we add features to the tracker of Pfister *et al*. [3] to improve tracking performance. As a case study we show how the appearance of the arms for different clothing, specifically short and long sleeved clothes, can be modelled to obtain person-specific trackers.

Motivation. The main motivation for tracking the pose is to automatically learn to recognise sign language [1], where upper body layout [2] is of great importance. Tracking is non-trivial due to changing background (the signer is overlaid on the broadcast video, as shown in Figure 1), and also because the signer changes between broadcast and so there are variations in shape and clothing. Furthermore, large quantities of video data are required to learn even a modest number of signs, implying the tracker has to be reliable over long video sequences and for multiple signers, and require little or no human supervision.

Summary of method. The tracker in our previous work [3] was only designed for application on a domain of signers wearing long sleeved clothing as shown in Figure 1(a). Our target domain is signers wearing short sleeved clothes, shown in Figure 1(b). We generate training data for the target domain from a source domain of long sleeved clothes in two stages. Stage 1 of our method (output shown in Figure 2) uses material from the source domain to generate semi-synthetic training data of signers wearing sleeves of a specific length. With this, one can retrain multiple general upper body pose trackers for a particular sleeve length, as shown in Figure 1(c). Stage 2 (shown in Figures 3 and 4) re-synthesises the training data to become signer specific. Then the refined semi-synthetic data is used to train a personalised tracker tuned to a particular person's arm shape and sleeve length, as shown in Figure 1(d). Each step contributes a significant boost to pose estimation performance, as shown in Figure 1(e).

[1] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. CVPR*, 2009.

[2] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *IJCV*, 2011.

[3] T. Pfister, J. Charles, M. Everingham, and A. Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC*, 2012.

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. CVPR*, 2011.

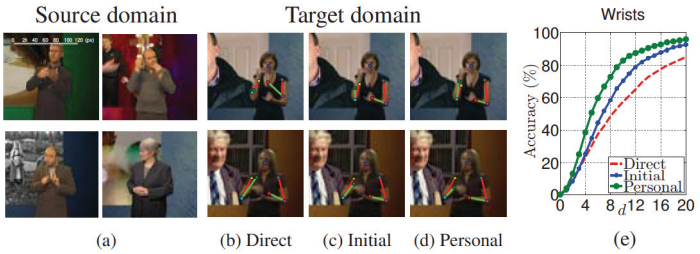


Figure 1: **Training an upper body joint detector using knowledge from the source domain.** In this example, the transfer is from a source domain (a) where signers wear long sleeves to a target domain (b)–(d) where signers wear short sleeves. The red and green lines on the signer show the output pose. The target domain results show: (b) the detector trained directly from the source domain; (c) trained from initial semi-synthetic images constructed using the source domain; and (d) from personal semi-synthetic images constructed from both source and target domain. (e) Accuracy of (b)–(d) in detecting the wrist joint in the target domain. Accuracy is percentage of predicted joints within a distance d pixels from ground truth (scale bar shown in top left of (a)). Note the improvement brought by the two stages – for example the accuracy almost doubles at 8 pixels.

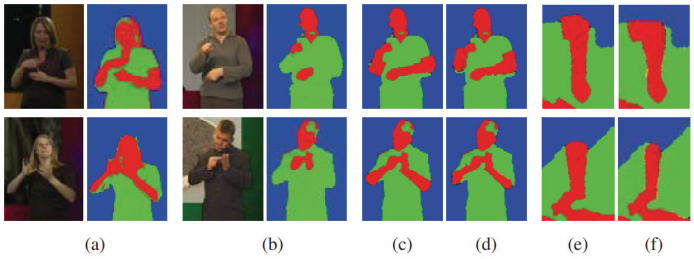


Figure 2: **Stages in synthesising training data.** Rows depict different sleeve length (sleeves in the top row are shorter than in the bottom row). (a) Raw RGB image and CP image counterpart of *short* sleeve signers. (b) Example *long* sleeve signers and CP images. (c) Synthetically produced CP images of short sleeves using CP images from (b) and of initial arm templates. (d) Personalised synthesis using learnt arm templates. For closer comparison, rotated left arms of the synthetic images in (c) and (d) are shown in (e) and (f) respectively.

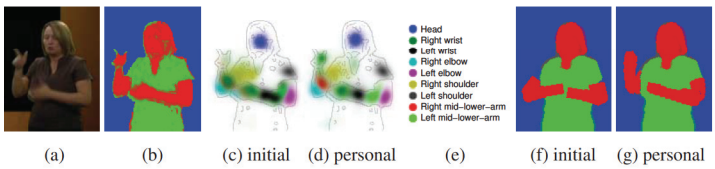


Figure 3: **Illustration of sample and verify procedure.** (a) Input signer and (b) CP image. (c) Shows body joint confidence map from initial forest, (d) confidence map produced with a personalised forest and (e) the body joint colour key. Most likely whole-image templates using initial (f) and personalised (g) arm templates, computed by sampling joint locations from (c) and (d) respectively. The personalised model produces both better joint samples (most notably for right wrist) and likelihood function.

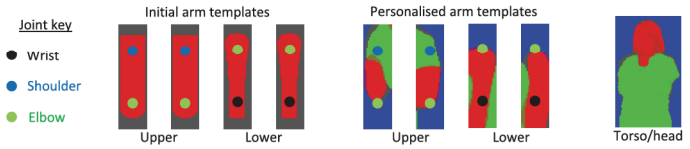


Figure 4: **Initial and personalised arm templates**. The personalised arm and torso/head templates are for the signer in Figure 3(a).

Multi-view Body Part Recognition with Random Forests

Vahid Kazemi
vahidk@csc.kth.se
Magnus Burenius
burenius@csc.kth.se
Hossein Azizpour
azizpour@csc.kth.se
Josephine Sullivan
sullivan@csc.kth.se

CVAP / KTH
The Royal Institute of Technology
Stockholm, Sweden

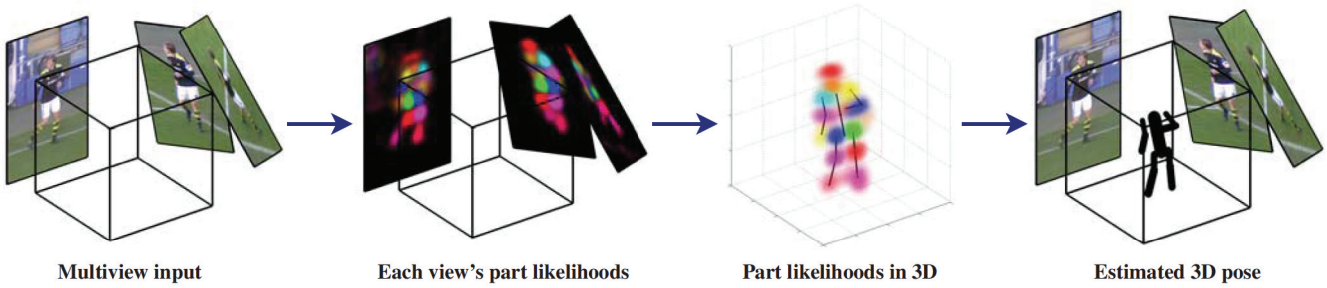


Figure 1: Overview of our 3D pose estimation framework. A random forest is first used to classify each pixel in each image as either belonging to a part or the background. The results are then back-projected to a 3D volume to get the 3D likelihood maps for each part. We find corresponding mirror symmetric parts across views by introducing a latent variable. Finally, a part-based model is used to estimate the 3D pose.

In this paper we address the problem of automatically estimating the 3D pose of a person seen from multiple calibrated cameras outside a studio environment [2]. Our particular focus is the estimation of the 3D pose of football players during a professional game. Football footage has several key characteristics some of which are shared between different sports. Most notably the images are commonly disturbed by motion blur because of the fast moving players and cameras. There is also a large variation in the players' 3D pose. On the other hand the variation in the players' clothing is limited and background clutter is not as severe as in less structured environments. Yet, low quality images and fast motion makes it hard to perform background subtraction reliably.

State of the art 2D part based models for human pose estimation rely on SVM classifiers applied to a HOG descriptor of an image patch[4]. However we opt to use a more efficient random forest approach for estimating the part likelihoods. We take our inspiration from the recent success of the *Kinect* system. Shotton *et al.* [3] use a random forest to estimate a person's 3D pose from a depth image. In this paper we consider ordinary visual images, as opposed to depth images, but similarly use a random forest to assign to every pixel a probability of being either a particular part or a background pixels. These probabilities form the basis for our part likelihood scores in 2D and 3D.

We create 3D part appearance likelihoods by aggregating the 2D likelihoods across all camera views. Care must then be taken regarding the correspondence of joints across the views. Because of the similar appearance of mirror symmetric parts, such as left and right arms and legs, and also the local nature of our part detectors, we can not directly distinguish the correct correspondences for each part. In this paper this issue is handled by introducing a latent variable into our model which represents the correspondence. At inference time we optimize for both the best pose and the best values of our hidden variable. We show that this approach is both feasible and effective and allows the 3D inference to perform well in the cases when the 2D likelihood maps are accurate.

To perform 3D pose estimation we follow the approach of [1] and discretize the search space. We assume that the person is within a bounding cube (fig. 1) and create a $64 \times 64 \times 64$ grid covering this cube. We evaluate our random forest based 3D appearance likelihoods (RF) for all grid points. We perform inference with and without the pose prior, imposing limb length and intersection constraints. We also perform inference with and without the latent variable, which handles the mirror ambiguity.

Figure 2 shows our estimated 3D poses (red) compared to the ground truth (blue), for four different frames. For this figure the inference was performed using the latent mirror variable but without any pose prior (uniform). The figure shows that our 3D appearance likelihoods accurately detect most of the body parts, even without imposing any pose prior. If we add the limb length and intersection constraints we are able to correct for some of the limited double counting that occurs for the lower legs.

[1] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013.

[2] T.B. Moeslund, A. Hilton, V. Krüger, and L. Sigal. *Visual Analysis of Humans: Looking at People*. Springer, 2011. ISBN 9780857299963.

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

[4] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

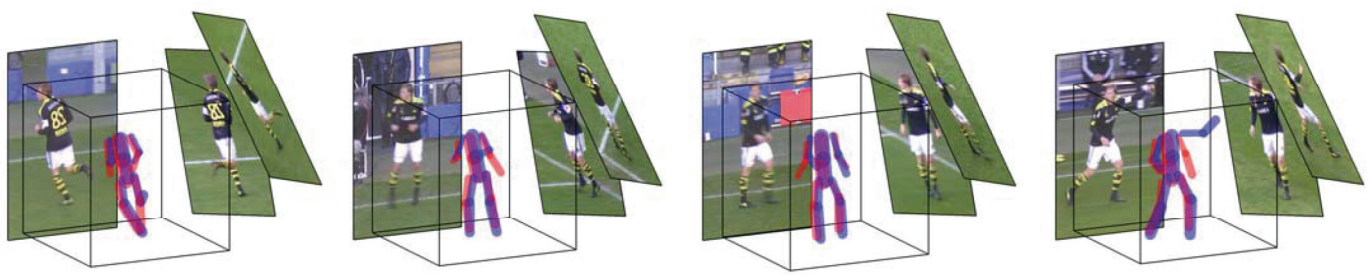


Figure 2: Example results. Final 3D poses obtained by taking, for each part independently, its most probable state over the grid. The mirror ambiguity is solved jointly.

KEYNOTE by Frank Dellaert *(Georgia Institute of Technology, USA)*

Factor Graphs for Fast and Scalable 3D Reconstruction and Mapping

Abstract: Simultaneous Localization and Mapping (SLAM) and Structure from Motion (SFM) are important and closely related problems in robotics and vision. I will show how both SLAM and SFM instances can be posed in terms of a graphical model, a factor graph, and that inference in these graphs can be understood as variable elimination. The overarching theme of the talk will be to emphasize the advantages and intuition that come with seeing these problems in terms of graphical models. For example, common computational tricks, such as the Schur complement trick in SFM, are simple choices about the order in which to eliminate the graph. In addition, while the graphical model perspective is completely general, linearizing the non-linear factors and assuming Gaussian noise yields the familiar direct linear solvers such as Cholesky and QR factorization. Based on these insights, we have developed both batch and incremental algorithms defined on graphs in the SLAM/SFM domain. In addition to direct methods, we recently worked on efficient iterative methods that use subgraphs of these factor graphs as pre-conditioners in a conjugate gradient scheme. Finally, we are now looking into how optimal control can be seamlessly integrated with the estimation algorithms for use in autonomous vehicles.

Biography. Frank Dellaert is a Professor in the School of Interactive Computing at the Georgia Institute of Technology. His research is in the areas of Robotics and Computer vision. He is particularly interested in graphical model techniques to solve large-scale problems in mapping and 3D reconstruction. You can find out more about his research and publications at the website http://www.cc.gatech.edu/~dellaert. The GTSAM toolbox which embodies many of the ideas his group has worked on in the past few years is available for download at http://tinyurl.com/gtsam

The Complete Rank Transform: A Tool for Accurate and Morphologically Invariant Matching of Structures

Oliver Demetz
demetz@mia.uni-saarland.de
David Hafner
hafner@mia.uni-saarland.de
Joachim Weickert
weickert@mia.uni-saarland.de

Mathematical Image Analysis Group
Saarland University
Saarbrücken, DE

Incorporating invariances is a popular tool to make computer vision systems more robust under real-world conditions. Since being invariant means ignoring something, every invariance has to come at the price of a loss of information. In our paper, we focus on severe illumination changes and descriptors that are invariant under such challenging conditions. The goal of our paper is to develop a novel descriptor that discards as little information as possible and carries the *maximally possible* amount of local image information under this invariance.

Intensity Order-based Transforms

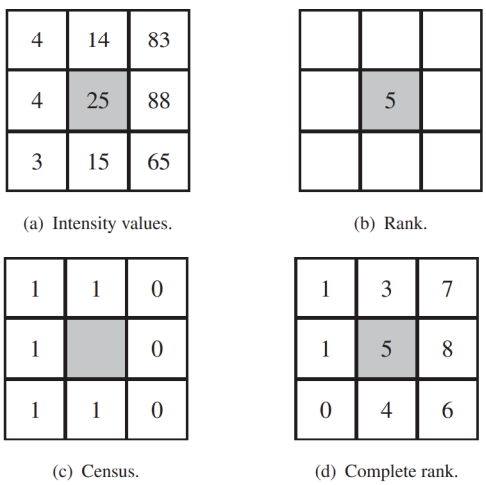
Before describing our novel patch-based descriptor, let us briefly revisit two closely related descriptors [2, 3] that form the basis of our proposal.

The *rank transform* (RT) encodes for each pixel the position of its grey value in the ranking of all grey values in its neighbourhood. Practically, this rank is determined by counting the number of neighbours with a smaller grey value than the reference pixel, cf. Figure 1(b).

The *census transform* (CT) can be seen as an extension of the rank transform: Besides encoding the rank, it adds a spatial component by expressing the relationship between the central pixel and each of its neighbours explicitly. In practice, one bit is stored for each pixel of the neighbourhood: If the neighbour is smaller than the reference pixel, the bit is 1, otherwise it is 0, cf. Fig. 1(c). Finally, all bits are concatenated in one signature.

We propose a novel transform that extracts as much local information as possible from the input image data while preserving the same desired invariance against monotonically increasing illumination changes. To this end, we make use of the following observation: Not only the rank of the central intensity value, but even the *complete intensity order* of the considered patch is morphologically invariant.

Hence, we extend the rank transform in such a way that it incorporates this novel idea of encoding the *complete* intensity order. The construction of our novel *complete rank transform* (CRT) is straightforward: First, we compute the rank of each pixel in the considered neighbourhood patch by determining the number of pixels with smaller intensity. Next, we concatenate these ranks as for the census transform to obtain the complete rank descriptor. Mathematically, for a patch size k, the complete rank transform is the mapping $\boldsymbol{s}_{\mathrm{CRT}} : \Omega \to \Pi_k \subset \{0,\dots,k-1\}^k$, where the co-domain Π_k is the set of all possible rankings of k elements including multiple occurrences of the same intensity value. The cardinality of this set is the k-th ordered Bell number ($OBN(k)$).



(a) Intensity values. (b) Rank. (c) Census. (d) Complete rank.

Figure 1: Illustration of the presented intensity order transforms ((b)–(d)) with a 3×3 neighbourhood patch ((a)), where the reference pixel is marked in grey.

Table 1: Comparison of the proposed intensity order transforms. The number of pixels in the considered neighbourhood is given by k.

transform	range D of one digit	signature length κ	spatial information	size of descriptor space
RT	$\{0,\dots,k-1\}$	1	–	k
CT	$\{0,1\}$	$k-1$	✓	2^{k-1}
CRT	$\{0,\dots,k-1\}$	k	✓	OBN(k)

Variational Optic Flow Model

As a proof of concept, we embed the novel signature in a generic TV–L^1-type variational model for optic flow [1] which assumes the signatures of corresponding pixels to coincide. We obtain the functional

$$E(u,v) = \int_\Omega \left(\Psi\left(\tfrac{1}{\kappa}|\boldsymbol{s}(\boldsymbol{x}+\boldsymbol{w}) - \boldsymbol{s}(\boldsymbol{x})|^2\right) + \alpha\cdot\Psi\left(|\nabla u|^2 + |\nabla v|^2\right)\right) \mathrm{d}\boldsymbol{x}, \tag{1}$$

whose minimiser is the sought flow field $(u,v)^\top : \Omega \to \mathbb{R}^2$. The first term (*data term*) penalises differences between the vector-valued signature $\boldsymbol{s} : \Omega \times [0,\infty) \to D^\kappa$ at position $\boldsymbol{x} := (x,y,t)^\top$ in the first frame and its corresponding one at $\boldsymbol{x}+\boldsymbol{w} := (x+u, y+v, t+1)^\top$ in the second frame. The second term (*smoothness term*) penalises the magnitude of the gradient $\nabla := (\partial_x, \partial_y)^\top$ of the flow field.

Results

Our novel CRT consistently outperforms the existing signature types. Figure 2 shows its unconditional invariance against γ-changes. On the Middlebury benchmark, our method ranked 43.5th, and on the KITTI Vision Benchmark it ranked 9th at time of submission.

[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV*. 2004.

[2] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. *Computer Vision Using Local Binary Patterns*. Springer, 2011.

[3] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. ECCV*. 1994.

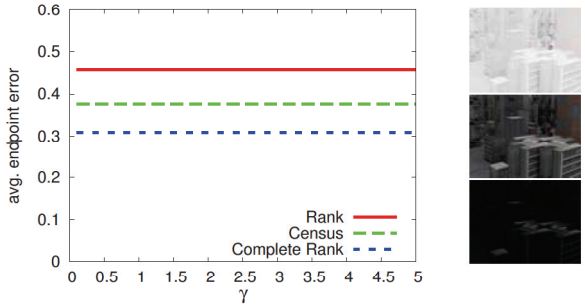


Figure 2: Behaviour under γ-changes. The plots show the results of our method under γ variations of the second frame of the *Urban2* sequence ($\gamma = \{0.1, 1, 3\}$).

Efficient Dense 3D Rigid-Body Motion Segmentation in RGB-D Video

Jörg Stückler
http://www.ais.uni-bonn.de/~stueckler
Sven Behnke
http://www.ais.uni-bonn.de/behnke

Computer Science Institute VI
University of Bonn
Bonn, Germany

Common motion is a fundamental grouping cue in video sequences. While for monocular and stereo image sequences, several approaches to motion segmentation have been investigated, it still remains a research problem to compute dense 3D motion segmentation efficiently. Many approaches match images sparsely at interest points and infer the groups of points with common 3D rigid-body motion (e.g., [1, 5]). Methods for dense 3D motion segmentation are still far from real-time performance (e.g., [6, 8]).

We propose in this paper an efficient approach to dense 3D motion segmentation in RGB-D video. We formulate an EM framework (see Fig. 1) that recovers a segmentation $\mathcal{L} = \{l_i = k\}_{i=1}^{N}$ into motion segments $\mathcal{M} = \{m_k\}_{k=1}^{M}$, estimates their 3D rigid-body motions $\Theta = \{\theta_k\}_{k=1}^{M}$, and also finds the number of segments M by optimizing

$$\arg\max_{\Theta} \sum_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref}) \ln p(I_{seg} \mid \Theta, I_{ref}, \mathcal{L}), \tag{1}$$

Our approach makes no difference between background and foreground objects and copes with camera motion as well as multiple moving objects.

We model the likelihood of a labelling $\mathcal{L}$ in a random field

$$p(\mathcal{L} \mid I_{seg}, \Theta, I_{ref}) \propto \prod_i p(z_i \mid \theta_{l_i}, I_{ref}) \prod_{j \in \mathcal{N}(i)} p(l_i, l_j \mid I_{seg}) \tag{2}$$

that incorporates the likelihood of the data at each site and pair-wise interaction terms between neighbors $\mathcal{N}(i)$ of site i. The data likelihood $p(z_i \mid \theta_{l_i}, I_{ref})$ quantifies the likelihood of the observation $z_i \in I_{seg}$ at a site under its label's motion estimate θ_{l_i}. For the pair-wise interaction terms we use a contrast-sensitive Potts model [2]. We incorporate additional pair-wise couplings to avoid multiple data associations of a site in the segmented image to the same site in the reference image.

Efficient graph-cuts [3] are then applied find a maximum likelihood labelling $\mathcal{L}_{ML} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid I_{seg}, \overline{\Theta}, I_{ref})$. To estimate motion, we apply approximations to arrive at

$$\arg\max_{\Theta} \sum_i \sum_{l_i} p(l_i \mid \mathcal{L}_{ML} \setminus \{l_i\}, I_{seg}, \overline{\Theta}, I_{ref}) \ln p(z_i \mid \theta_{l_i}, I_{ref}). \tag{3}$$

The weight of a site intuitively is the likelihood of belonging to a segment.

Model complexity is controlled using label costs [4]. We start with one segment and perform EM with one additional segment in each iteration until the number of segments keeps constant and EM has converged. For the sequential segmentation of an image towards a video we propose to initialize EM in each frame with the result from the previous frame. This way, the EM algorithm requires only few iterations.

The performance of our EM approach strongly depends on the underlying image representation. In principle, any representation is suitable that defines data likelihood $p(z_i \mid \theta_{l_i}, I_{ref})$, image site neighborhood $\mathcal{N}_S(i)$, and dissimilarity for the pair-wise interaction terms. To solve Eq. (3), an image registration technique is required that allows to incorporate individual weights for the image sites. Instead of segmenting the large number of pixels in the image, we represent RGB-D images compactly as multi-resolution surfel maps [7]. These maps capture noise characteristics in a local multi-resolution structure in which the maximum resolution adapts to the distance of the measurements. In effect, the content of an RGB-D image is compressed from 640×480 pixels to only several thousand voxels, making dense inference of labels in the map efficient.

In experiments on test sequences with ground-truth segmentations and rigid-body motions, we demonstrate that our approach efficiently identifies moving segments with high accuracy and recovers 3D rigid-body motion of the segments at good accuracy (see Fig. 2 for examples), also at real-time operation on a CPU.

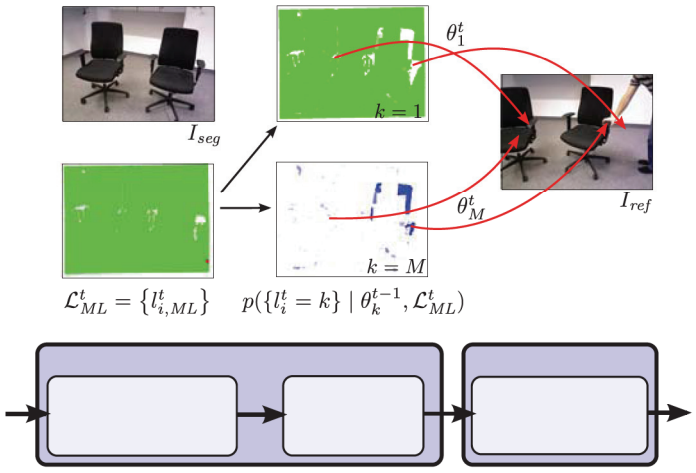


Figure 1: We segment motion in an RGB-D image I_{seg} towards a reference image I_{ref} in an efficient EM framework. In the E-step, we evaluate the likelihood of image site labels l_i under the latest motion estimates θ_k. Efficient graph cuts yield a maximum likelihood labelling $\mathcal{L}_{ML}$ given the motion estimates, which is then used to approximate the label likelihoods. In the M-step, motion is reestimated taking into account label likelihoods.

[1] M. Agrawal, K. Konolige, and L. Iocchi. Real-time detection of independent motion using stereo. In *Proc. of the IEEE Workshop on Motion*, 2005.

[2] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proc. of the IEEE Int. Conf. on Computer Vision*, 2001.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[4] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *Int. J. of Computer Vision*, 96 (1):1–27, 2012.

[5] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the EM algorithm. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[6] A. Roussos, C. Russell, R. Garg, and L. de Agapito. Dense multibody motion estimation and reconstruction from a handheld camera. In *Proc. of the IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*, 2012.

[7] J. Stückler and S. Behnke. Multi-resolution surfel maps for efficient dense 3D modeling and tracking. *J. of Visual Communication and Image Representation*, 2013.

[8] G. Zhang, J. Jia, and H. Bao. Simultaneous multi-body stereo and segmentation. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2011.

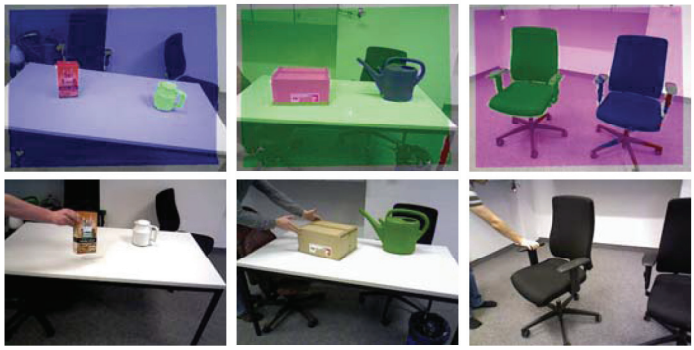


Figure 2: Example segmentations (outliers dark red).

Enhancing Action Recognition by Cross-Domain Dictionary Learning

Fan Zhu
fan.zhu@sheffield.ac.uk
Ling Shao
ling.shao@sheffield.ac.uk

Department of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD, UK

Our work is inspired by two facts of the human vision system. The first fact is that humans are able to learn tens of thousands of visual categories in their life, which leads to the hypothesis that humans achieve such a capability by accumulated information and knowledge. Another fact is that human's visual impressions towards the same action or the same object comes from a wide range, e.g., an action seen from 2D static images *vs.* the same action seen from 3D dynamic movies or an object seen from real-world scenes *vs.* the same object seen from low-resolution online images. These facts can be explained in the computer vision language as the human vision system possesses the ability of spanning the intra-class diversity of the original training instances through transferring prior knowledge. Motivated by the above two facts, we introduce a new action recognition framework that utilizes relevant actions from other domains as auxiliary knowledge (motivated by the first fact) to span the intra-class diversity of the original learning system (motivated by the second fact). In addition to manually annotated actions in the target domain, labeled actions from a different domain are provided as the source domain actions. Based on the recent success of dictionary learning methods in solving computer vision problems, we present a discriminative cross-domain dictionary learning (DCDDL) technique to learn a reconstructive, discriminative and domain-adaptive dictionary pair for data under different distributions. The flowchart of the proposed framework is shown in Figure 1.

The objective function:

$$\begin{aligned}\langle D_t, D_s, X_t, X_s\rangle = arg \min_{D_t,D_s,X_t,X_s} \|Y_t - D_t X_t\|_2^2 \\ +\|Y_s - D_s X_s\|_2^2 + \|X_t - f(Y_t, Y_s)X_s\|_F^2 \\ +\|X_s - f(Y_s, Y_t)X_t\|_F^2 \\ s.t. \forall i, [\ \|x_t^i\|_0, \|x_s^i\|_0\ ] \leq T,\end{aligned} \tag{1}$$

where the function $f(\cdot)$ computes the mapping of correspondence samples (i.e., samples that share the same class labels while being close to each other) across different domains. Thus, small values of $\|X_t - f(Y_t, Y_s)X_s\|_F^2$ and $\|X_s - f(Y_s, Y_t)X_t\|_F^2$ indicate that those data points close to each other are more likely to share the same class label in the new target feature domain and the new source feature domain respectively. Since we are only concerned with the smoothness within the target domain data, the last term in Equation (1) can be removed. According to the stated scenario, no manually annotated correspondences between the target domain data and the source domain data are available in the training phase, thus $f(\cdot)$ is computed using a category-specific searching method. We use the matrix $\mathcal{A}$ to represent the connections between the target domain data and the source domain data in the category-specific manner. In order to establish the correspondences across the target domain data and the source domain data, the maximum element in each column of $\mathbb{A}$ is preserved and set to 1 while the remaining elements are set to 0.

$$\mathbb{A}(i,j) = \begin{cases} 1, & if \quad \mathbb{A}(i,j) = max(\mathbb{A}(:,j)) \\ 0, & otherwise. \end{cases} \tag{2}$$

Assuming $\mathbb{A}$ leads to a perfect mapping across the sparse codes from both domains and the matched pair of samples in different domains possesses an identical representation after encoding, the Equation (1) can be rewritten as:

$$\begin{aligned}\langle D_t, D_s, X_t, X_s\rangle = arg \min_{D_t,D_s,X_t} \|Y_t - D_t X_t\|_2^2 \\ +\|(\mathbb{A}Y_s^T)^T - D_s X_t\|_2^2 \quad s.t. \forall i,\ \|x_t^i\|_0 \leq T.\end{aligned} \tag{3}$$

We further include a discriminative term to the objective function with respect to the optimal data distribution. Let the classifier $\mathcal{F}(x)$ satisfy the following equation:

$$\mathcal{P} = arg \min_{\mathcal{P}} \sum_i w_i \times \mathcal{L}\{h_i, \mathcal{F}(x_t^i, \mathcal{P})\} + \lambda_i \|\mathcal{P}\|_F^2, \tag{4}$$

where $\mathcal{L}$ is the classification loss function, h_i indicates the target domain labels of x_t^i, $\mathcal{P}$ denotes the classifier parameters and λ_i is a regularization parameter.

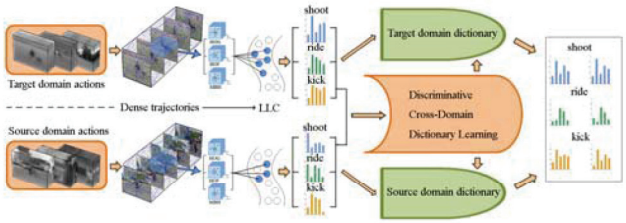


Figure 1: Flowchart of the proposed cross-domain action recognition framework.

Table 1: Performance comparison between DCDDL and other methods on the UCF YouTube dataset.

Algorithm	LLC	LLC	K-SVD	K-SVD	LC-KSVD	LC-KSVD	DCDDL
Learning	*N/A*	*N/A*	*UN*	*UN*	*SU*	*SU*	*SU*
Source data	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*Yes*
24 actors	86.67%	86.67%	82.22%	77.78%	86.67%	82.22%	**88.89%**
20 actors	75.42%	70.21%	68.75%	72.08%	75.42%	75.42%	**77.50%**
16 actors	70.88%	70.17%	63.96%	67.54%	72.08%	72.08%	**73.03%**
09 actors	61.41%	61.80%	55.70%	59.15%	65.25%	64.72%	**66.31%**
05 actors	54.10%	53.35%	50.05%	48.88%	56.55%	54.10%	**56.66%**

Where the HMDB51 dataset [1] is treated as the source domain. Specifically, 7 body movements, including ride bike, dive, golf, jump, kick ball, ride horse and shoot ball, are chosen from the HMDB51 dataset in correspondence with similar actions in the UCF YouTube dataset. Performance comparisons of the proposed method and other state-of-the-art methods are reported on both scenarios where the source domain data are included or excluded in Table 1 and Figure 2.

[1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision*. 2011.

[2] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

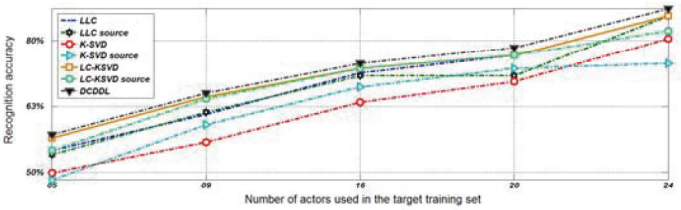


Figure 2: Performance comparison of the proposed DCDDL with other methods under different dataset partitions.

Unsupervised Object Discovery and Segmentation in Videos

Samuel Schulter[1]
schulter@icg.tugraz.at
Christian Leistner[2]
christian.leistner@microsoft.com
Peter M. Roth[1]
pmroth@icg.tugraz.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision
Graz University of Technology
Graz, Austria
[2] Microsoft Photogrammetry
Graz, Austria

Introduction: Unsupervised object discovery (UOD) is the task of finding repeating patterns and common visual concepts across an unsorted set of images without any human supervision. These concepts should describe objects, like pedestrians or cars, and stuff, like road or sky. Once discovered, this information opens several interesting applications like summarization and filtering of visual content, discovering novel or unusual visual patterns, or reducing human annotation effort. As the amount of visual data grows exponentially and human annotation becomes costly, such applications are getting more and more important.
However, existing UOD approaches typically build on still images and have to rely on prior knowledge to yield accurate results. In this work, we propose a novel *video-based approach*, allowing also for exploiting motion information, which is a strong and physically valid indicator for foreground objects, thus, tremendously easing the task (see Figure 1).

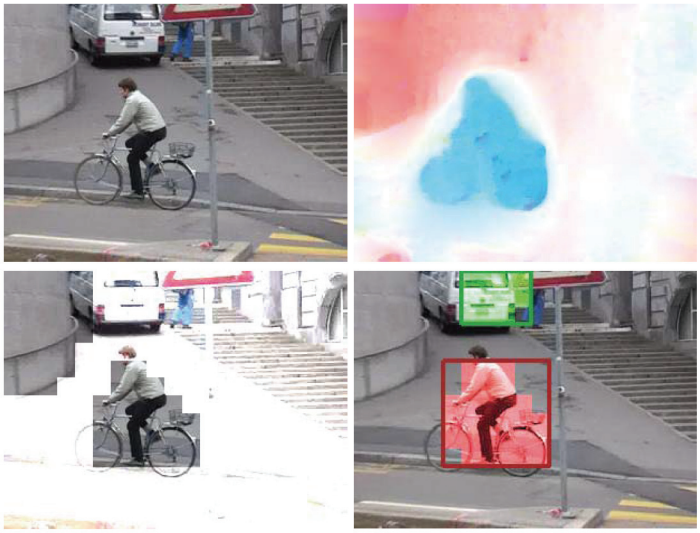


Figure 1: Given unlabeled video sequences *(top left)*, we first estimate the optical flow *(top right)* and calculate the motion segmentation *(bottom left)*. Then, using these segmentations we cluster the different object categories and learn appearance models. These models are finally used to discover objects in videos, where even static, non-moving objects as the car (green) can be found *(bottom right)*. Illustration for only one frame.

Approach: Our UOD approach consists of several building blocks as illustrated in Figure 2. The input is a set of videos and the number of categories k to be discovered, which is the only supervision necessary.
We start by calculating optical flow and perform a motion segmentation with our CRF formulation. This gives an initial set of tentative object proposals without any semantic information, yet. Our approach can then group those proposals into k semantically similar clusters and learn appearance-based object models for each cluster on two different abstraction levels (*local* and *holistic*) in a common framework. Given those object models, we can classify each frame of the input videos and calculate probability maps for each category. Given the additional semantic information from the learned classifiers, we formulate a joint CRF that allows for exploiting both, motion and semantic information. The CRF is now able (i) to confirm object proposals already identified via motion and to give them a semantic label, (ii) to remove outlier proposals from the first step, and (iii) also to discover new *static* objects having no motion information.

The output of our approach is thus a segmentation of each video frame into either background or one of k categories. Additionally, our approach also returns fully trained object detectors for each of the discovered categories, which can readily be applied to still images.

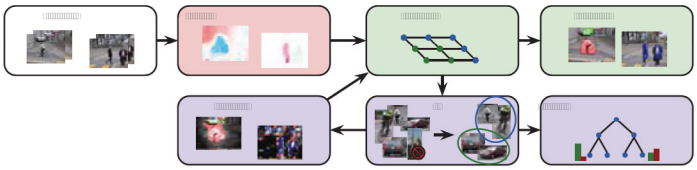


Figure 2: Building blocks of our Unsupervised Object Discovery approach from videos.

Experiments: To demonstrate the benefits of additionally using motion information for unsupervised object discovery, we run 3 different experiments: (i) Unsupervised object discovery in videos, where we show that our approach can successfully discover and discriminate between different categories. (ii) Video retrieval, where we demonstrate the generalization power of the learned appearance models. (iii) Object detection, where we show that our unsupervised learned models can even be applied to still images and give reasonable results compared to the fully supervised models.



Figure 3: Some qualitative results of our Unsupervised object discovery approach.

In Figure 3, we show some qualitative results from the first experiment on UOD. Our approach segments the main objects in the sequences (first block) and is also able to identify static objects (second block). The last block illustrates two failure cases, which, however, also demonstrate the power of the part-based Hough Forests, as it actually separates the person and the bicycle (bottom row).

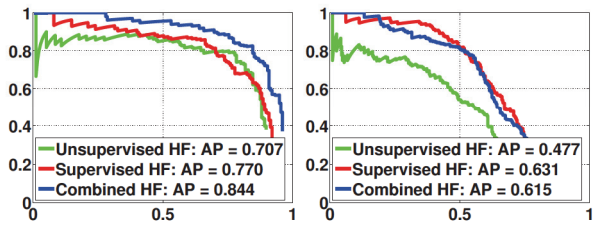


Figure 4: Precision-recall curves of all models on ETHZ-cars (left) and TUD-pedestrian (right).

In Figure 4, we compare our unsupervised trained models with the fully-supervised models and also a combination thereof on two different datasets. As can be seen from the average-precision curves, our unsupervised trained models yield reasonable results compared to their fully supervised competitors. On one data set, the combined approach can even outperform the supervised one.

Evaluating Superpixels in Video: Metrics Beyond Figure-Ground Segmentation

Peer Neubert
peer.neubert@etit.tu-chemnitz.de
Peter Protzel
protzel@etit.tu-chemnitz.de
Chemnitz University of Technology
Germany

There exist almost as many superpixel segmentation algorithms as applications they can be used for. Figure 1 shows two example superpixel segmentations. So far, the choice of the right superpixel algorithm for the task at hand is based on their ability to resemble human-made ground truth segmentations (besides runtime and availability). We investigate the equally important question of how stable the segmentations are under image changes as they appear in video data (**stability-criteria**). Further we propose a new quality measure that evaluates how well the segmentation algorithms cover relevant image boundaries (**discontinuity-criteria**). Instead of relying on human-made annotations, that may be biased by semantic knowledge, we present a completely data-driven measure that inherently emphasizes the importance of image boundaries. In detail, we exploit ground truth optical flow data provided by two recently published datasets for evaluation of optical flow algorithms (KITTI [3] and Sintel[2]) to evaluate the stability- and discontinuity-citeria related to questions a) and b) in Figure 1. Both criteria are discussed, formalized and used to compare several existing superpixel algorithms with available open source implementations. For further evaluation of other algorithms, we provide the results, a Matlab implementation of the metrics and functions to interface the datasets on our website.[1]

1 A Metric for the Stability-Criteria

At first evaluate the stability of superpixel segmentations in image sequences or video. While superpixel borders at considerable image gradients may constantly be detected, oversegmentation algorithms tend to create lots of spurious segment borders that strongly vary under image changes. Even slight changes of the image, e.g. a small camera motion, can cause substantial changes of the produced segmentation. For some applications this might be irrelevant, while others could benefit from a superpixel algorithm with more stable segmentations. For evaluation of the stability we define the motion undersegmentation error (MUSE). The key idea is to segment two images I_1, I_2 showing the same scene before and after some changes (e.g. dynamic objects, camera motion, illumination changes), resulting in label imagess L_1, L_2. Then apply ground truth optical flow data F to compute L_1^F, a transformation of the segmentation L_1 of the first image into the view of the second image to make them comparable. Finally, use the undersegmentation error metric to evaluate how well segmentation L_1^F can be reconstructed by segments of segmentation L_2 and vice versa. Undersegmentation error is a repeatedly used measurement for comparing superpixel segmentations. We use the parameter free equation of [4]. To compare two segmentations L_1^F and L_2, and being N the total number of pixels, we define MUSE to be computed as follows:

$$MUSE = \frac{1}{N}\left[\sum_{a\in L_1^F}\left(\sum_{b\in L_2 : a\cap b\neq\emptyset} min(b_{in}, b_{out})\right)\right] \quad (1)$$

Each segment a of segmentation L_1^F is reconstructed with segments b of L_2 that overlap with a. MUSE accumulates the error that is introduced by b when reconstructing a either when b is included in the reconstruction or not. Since this is not a symmetric metric (the error diverges whether comparing L_1^F to L_2 or vice versa) we compute the average of both cases.

2 A Metric for the Discontinuity-Criteria

Superpixel segmentation algorithms are typically compared based on human annotated ground truth segmentations. Although there may exist multiple manual segmentations for each image (like in BSDS [1]), the ground truth data depends on the *semantic interpretations* of objects and their boundaries by humans. Instead of asking humans, what relevant image boundaries are, we propose to define them as image areas whith differently moving pixels in their neighborhood. This allows a direct computation of motion discontinuities as high gradients in a ground truth optical flow field. Dependent on the application, it is important to have a segment boundary near positions with high motion gradients (e.g. 3D reconstruction). To avoid arbitrarily chosen thresholds to separate important high gradients from ignored low gradients, we propose to use the following error measure: Given F, a ground truth optical flow field from an image I to another image, B the boundary image of a segmentation of image I, and $D(B)$ the distance transform of B containing for each pixel the distance to the nearest segment boundary, we define the Motion Discontinuity Error (MDE) as follows:

$$MDE = \frac{1}{\sum_i \sum_j \|\nabla F(i,j)\|_2} \sum_i \sum_j \|\nabla F(i,j)\|_2 \cdot D(B(i,j)) \quad (2)$$

In one sentence this is the Frobenius inner product of the optical flow gradient magnitude and the distance transform of the boundary image of the segmentation, divided by the sum of all gradients. A more intuitive formulation is to accumulate over all image pixels a penalty, which is the product of the strength of motion discontinuity at this pixel and its distance to the next segment border. Finally, the measure is normalized by the total amount of motion in the image.

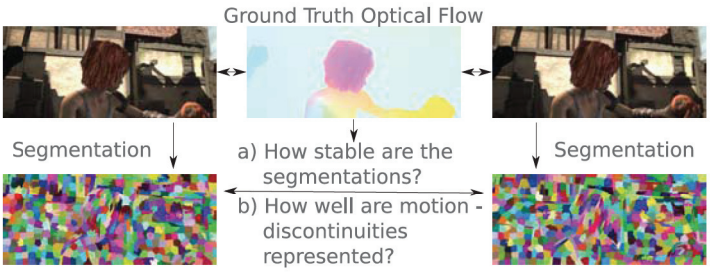


Figure 1: We propose to use ground truth optical flow fields to compare superpixel segmentation algorithms. On top left and right are two Sintel images with slight motion visualized by the optical flow field between them. Motion direction is coded by hue, saturations codes the motion magnitude. Beneath the images there are example superpixel segmentations (using *ERS*). While some object contours are visible, there seem to be a lot of spurious segment borders. In this work we provide metrics to answer questions a) and b).

3 Results of Superpixel Algorithm Comparison

We use the proposed criteria to compare several state of the art algorithms. With current algorithms, MDE and MUSE are somehow complementary measurements. Algorithms that perform well on one criteria often show problems with the other, in fact there is a lack of algorithms that produce stable segmentations and well resemble motion discontinuities. This opens space for further improvements and new superpixel segmentation algorithms. Based on the results on the present comparison and its previously published performance on figure-ground segmentations, the SLIC algorithm shows best balanced results.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33, 2011.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, 2012.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Comp. Vision a. Pattern Recog. (CVPR)*, 2012.

[4] P. Neubert and P. Protzel. Superpixel benchmark and comparison. In *Proc. Forum Bildverarbeitung*, 2012.

[1] http://www.tu-chemnitz.de/etit/proaut/forschung/superpixel.html

Solving Person Re-identification in Non-overlapping Camera using Efficient Gibbs Sampling

Vijay John
v.c.k.john@uva.nl
Gwenn Englebienne
g.englebienne@uva.nl
Ben Krose
b.j.a.krose@uva.nl

University of Amsterdam
Amsterdam, Netherlands

Person re-identification, or inter-camera data association, is the task of identifying people in different camera views in a network. With the recent interest in security and surveillance, camera networks are widely deployed with non-overlapping field of views (FOV). Typically, appearance-based methods are used for person re-identification using visual information corresponding to people extracted from the video sequences [1]. Appearance-based methods, though widely used, exhibit several challenging issues including spatio-temporal appearance variations, changing lighting conditions across different camera views, camera response variations and high computational complexity. The high computational complexity arises from the need to compare a unknown query person, or observation, with all other observations in the network [3]. In literature, researchers have sought to address the issues with illumination and camera response variation by using illumination invariant feature descriptors [1], by modeling the inter-camera variations in the network [2] or by colour calibration algorithms [4].

In our work, belonging to the second literature class, we model the appearance with the illumination variation and camera gain in the network. Unlike the works discussed so far [2], which model the inter-camera variations, we model the person's appearance in terms of the absolute illumination and gain associated with each camera in the network using a graphical model. To account for the high computational complexity, a novel constant time Gibbs sampling framework is proposed to perform person re-identification, and subsequently the person's appearance and camera properties are learnt in closed form. Our main contribution to literature are the following: modeling the absolute illumination variation and camera gain in the network; efficient constant time Gibbs sampling reducing the computational complexity.

Our problem is formulated as follows: given a set of observations $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ across multiple camera views, where each observation corresponds to a person's complete trajectory within a camera's FOV, person re-identification can be defined as the problem of identifying the set of indicator variables $\mathbf{z} = \{z_i\}_{i=1}^N$ identifying the person associated with each observation in $\mathcal{X}$. To identify the label $z_i \in [1,\dots,Z]$ associated with each trajectory, we use a combination of appearance features and transition probabilities between the cameras within a Bayesian inference framework. In our probablistic graphical model (Fig. 1), each observation $\mathbf{x}_i = \{l_i, e_i, t_i, \mathbf{a}_i\}$ consists of: $l_i \in [1,\dots,L]$ the camera that records the observation; the time of entry e_i in a camera's FOV; the time of leaving the camera's FOV t_i; and the observed appearance features $\mathbf{a}_i$ corresponding to the raw RGB colour values within the bounding box of the detected person averaged over the entire trajectory. We model the appearance of a person in a camera's view as a function of the camera's properties and the person's "absolute" appearance, and we model the transition of a person from one camera to the next in terms of transition probabilities between cameras and the time it takes to move from one camera to another. The likelihood $p(\{\mathbf{x}_i\}_{i=1}^N|\{z_i\}_{i=1}^N)$ is defined as

$$= \prod_{j=1}^{N} p(l_j|\{l_i\}_{i=1}^{j-1},\{z_i\}_{i=1}^{j})\, p(e_j|\{t_i\}_{i=1}^{j-1},\{z_i\}_{i=1}^{j})\, p(\mathbf{a}_j|z_j,l_j)\,. \qquad (1)$$

Here, $p(\mathbf{a}_j|z_j,l_j)$ is modelled as $\mathbf{a}_j = g_{l_j}(\mathbf{r}_{z_j} + \mathbf{w}_{l_j})$, where g_l is the multiplicative gain constant of camera l, $\mathbf{r}_z$ is the RGB color model, averaged over the entire trajectory, $\mathbf{w}_l$ is the illumination noise associated with camera l. It is clear from its structure that this model does not allow for efficient inference, since the Markov blanket of any observation is the complete set of observations and indicator variables preceding it. Yet if the latent indicator variables are known, the observations of a person become independent of all other persons, and the model becomes much simpler (see Fig. 1(b)). We therefore use Gibbs sampling to estimate $\mathbf{z}$,

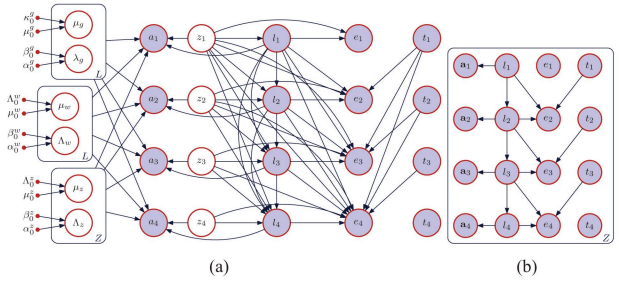


Figure 1: (a) Full graphical model of our probabilistic person re-identification algorithm and (b) Graphical model if the latent variables are known.

where each dimension of the sample $\mathbf{z}$ is sampled in alternation, according to the proposal distribution $p(z_i|\mathbf{z}_{\neg i},\mathcal{X})$ we use $\mathbf{z}_{\neg i}$ to denote $\mathbf{z}\backslash z_i$, the set of all labels except label i. This proposal distribution leads to accepting samples with a probability of one, thereby leading to a very efficient sampling mechanism. Typically, the proposal distribution is computed as follows.

$$p(z_i|\mathbf{z}_{\neg i},\mathcal{X}) = \frac{p(\mathcal{X}|\mathbf{z})\; p(\mathbf{z})}{\sum_{z_i=1}^{N} p(\mathcal{X}|\mathbf{z})\; p(\mathbf{z})}, \qquad (2)$$

Generally, $p(\mathcal{X}|\mathbf{z})$ can be computed in linear time of the number of observations, and the probability $p(z_i|\mathbf{z}_{\neg i},\mathcal{X})$ can also be computed in linear time. This leads to a scheme where the cost of obtaining a new sample is quadratic in the number of observations, since we need to look at all dimensions of $\mathbf{z}$. In our algorithm, we compute the prior probability over the object associations in constant time. Consequently, the conditional probability $p(z_i|\mathcal{X},\mathbf{z}_{\neg i})$ can also be computed in constant time with a little bookkeeping. Thus reducing the computational complexity.

For a given sample of latent indicator variables, efficient Bayesian inference can be performed in our graphical model. More specifically, we obtain closed form solutions for the camera gain, illumination variation and the appearance parameters which are then used to perform analytical updates in our graphical model.

We evaluated our method on two real-world datasets with multiple non-overlapping ceiling mounted cameras and compared our performance with the algorithm proposed by Pasula et al. [3]. Additionally, we evaluate the importance modeling the camera parameters by comparing it to direct modeling of the appearance. We show that our algorithm demonstrates better re-identification accuracy and computational complexity than [3]. We also show that our proposed appearance model with gain and illumination components performs better than the modified appearance model without the gain and illumination components.

[1] Federica Battisti, Marco Carli, Giovanna Farinella, and Alessandro Neri. Target re-identification in low-quality camera networks. In *Image Processing: Algorithms and Systems*, 2013.

[2] Andrew Gilbert and Richard Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *ECCV*, 2006.

[3] Hanna Pasula, Stuart Russell, Michael Ostland, and Ya'acov Ritov. Tracking many objects with many sensors. In *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1160–1171, 1999.

[4] Fatih Murat Porikli. Inter-camera color calibration by correlation model function. In *ICIP*, 2003.

Spacetime Forests with Complementary Features for Dynamic Scene Recognition

Christoph Feichtenhofer[1]
cfeichtenhofer@gmail.com
Axel Pinz[1]
axel.pinz@tugraz.at
Richard P. Wildes[2]
wildes@cse.yorku.ca

[1] Institute of Electrical Measurement and Measurement Signal Processing
Graz University of Technology, Austria

[2] Department of Electrical Engineering and Computer Science
York University
Toronto, Ontario, Canada

This paper presents spacetime forests defined over complementary spatial and temporal features for recognition of naturally occurring dynamic scenes. The present work makes three main contributions. (i) A novel descriptor is proposed that integrates complementary information from separate spatial and temporal orientation measurements for aggregation in spacetime pyramids. (ii) A random forest classifier is applied for the first time to dynamic scenes. This spacetime forest allows for automatic determination of the most discriminative features to separate the classes based on appearance and dynamics with computational efficiency. (iii) Video is processed in incremental temporal slices with scale matched preferentially to scene dynamics (in comparison to camera motion). This strategy allows for temporal alignment to be treated as latent in the classifier, efficient processing and robustness to large temporal variation across time (*e.g.* from camera motion), even while capturing intrinsic scene characteristics. Previous dynamic scene research has suffered in the presence of camera motion [1] and has provided little consideration of on-line processing concerns. Figure 1 overviews the framework.

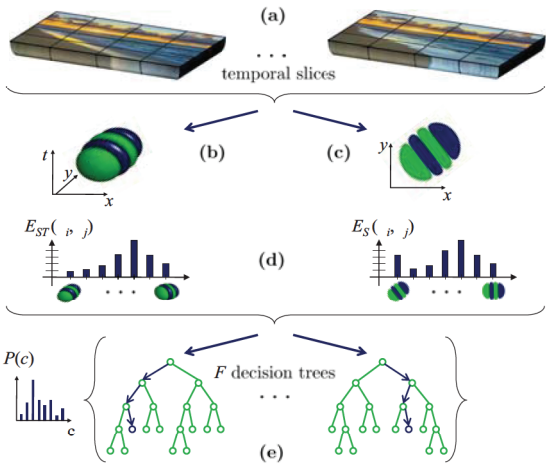


Figure 1: The proposed dynamic scene classification framework. (a) τ temporal slices are created to process the frames of a video in a sliding window approach, divided in spatiotemporal pyramid cuboids. (b,c) The cuboids are filtered by banks of multiscale, σ, oriented filters along equally separated directions, θ, in image spacetime (b) and space (c) to capture distributions (d) of both dynamic and static appearance information. (e) The class of each temporal slice is determined via likelihood voting, using a multi-class random forest. Subsequently, all slice-based classifications are incrementally combined across the entire input.

Complementary spacetime features. Spatial appearance and temporal dynamics information are extracted via applications of multiscale filter banks that are further tuned for spatial or spatiotemporal orientation. In the spatial domain, 2D Gaussian third derivative filters are applied (Fig. 1(c)) to yield a set of multiscale, multiorientation measurements. Similarly, dynamic information is extracted via application of 3D Gaussian third derivative filters (Fig. 1(b)). To yield measures of dynamic information independent of spatial appearance, the initial spatiotemporal orientation measurements are summed across all orientations consistent with a single frequency domain plane, where motion occurs as a plane through the origin. Previous spacetime filtering approaches to dynamic scene recognition tend to exhibit decreased performance when dealing with scenes captured with camera motion, in comparison to scenes captured with stationary cameras. A likely explanation for this result is that the approaches have difficulty in disentangling image dynamics that are due to camera motion vs. those that are intrinsic to the scenes. Here, it is interesting to note that camera motion often unfolds at coarser time scales (*e.g.*, extended pans and zooms) in comparison to intrinsic scene dynamics (*e.g.*, spacetime textures of water, vegetation, *etc.*); however, previous approaches have made their measurements using relatively coarse temporal scales and thereby failed to exploit this difference. In the present approach this difference in time scale is captured by making use of only fine scales during spatiotemporal filtering, so that they are preferentially matched to scene, as opposed to camera, dynamics. Note, however, that spatial filtering is performed at four coarse scales varying by octave. Furthermore, CIE-LUV colour channels are included to capture complementary chromatic information.

For the purpose of classification of the entire video into a scene class, the local descriptors are summed across time, t, within τ discrete units of equal duration to yield a set of temporally aggregated images, which are referred to as temporal slices. Temporal slicing is motivated by the desire for incremental processing that can allow for efficient, on-line operation. Use of short-term parcelling of the measurements also is well matched with the restriction to use of fine temporal scales during spatiotemporal filtering to favour scene over camera dynamics. Each temporal slice is hierarchically aggregated into histograms to form a spatiotemporal pyramid, analogous to that used previously in dynamic [1] scene analysis. At each level of the pyramid, each temporal slice is broken into 3D cuboids (see Fig. 1(a)), with filter measurements collapsed into histograms within each cuboid, as illustrated in Fig. 1(d). The support of the cuboid at any given level of the pyramid corresponds to the aggregation region in the filtering process. Moreover, the adjacency structure of the cuboids capture the overall scene layout. For each cuboid, the histograms are $L1$-normalized, to represent a distribution of chromatic, multiscale oriented spacetime energy as well as lack of oriented structure. The histograms for all cuboids are concatenated into a final feature vector that comprises the Complementary Spacetime Orientation descriptor (CSO) to characterize a temporal slice of the video.

Spacetime forests. A Random Forest (RF) classifier consisting of F decision trees is used for classification, based on the leaf node at which the corresponding CSO descriptor arrives (see Fig. 1(e)). During training, the temporal alignment of the video slices is treated as latent; correspondingly, each temporal slice of each training video generates its own feature vector. This approach allows leveraging of the high temporal diversity in the spatiotemporal patterns. As some classes may be better represented by specific feature types, the node optimization process in each tree is restricted to a single feature type. The input for the RF is first structured into the three complementary feature channels of the CSO descriptor. Then, the channels are used to train $\frac{F}{3}$ trees each, to best distinguish the classes, based on a particular channel only. Lastly, these complementary trees are merged, to obtain the spacetime forest. During classification, each temporal slice initially is classified individually, with the individual classifications subsequently combined to yield a final classification across all temporal slices available up to a given time.

The approach has been evaluated on two publically available datasets [1, 2]; results show that it achieves a new state-of-the-art in dynamic scene recognition.

[1] K. Derpanis, M. Lecce, K. Daniilidis, and R. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *Proc. CVPR*, 2012.

[2] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *Proc. CVPR*, 2010.

Hyperspectral Face Recognition using 3D-DCT and Partial Least Squares

Muhammad Uzair
uzair@csse.uwa.edu.au
Arif Mahmood
arif.mahmood@uwa.edu.au
Ajmal Mian
ajmal.mian@uwa.edu.au

Computer Science and Software Engineering
The University of Western Australia
35 Stirling Highway, Crawley, WA, Australia.

Hyperspectral imaging offers new opportunities for inter-person facial discrimination. However, due to the high dimensionality of hyperspectral data, discriminative feature extraction for face recognition is more challenging than 2D images. For dimensionality reduction and feature extraction most of the previous approaches just sub sampled the hyperspectral data [5, 6, 9] or used simple PCA [3]. In contrast, we propose the three dimensional Discrete Cosine Transform (3D-DCT) for feature extraction (Fig. 1). Exploiting the fact that hyperspectral data is usually highly correlated in the spatial and spectral dimensions, a transform such as DCT is expected to perform information compaction in a few coefficients by providing maximal decorrelation. DCT transform being an approximation of the KL-Transformation optimally compacts the signal information in a given number of transform coefficients. Moreover, compared to other transforms, such as the Fourier transform, the transformed coefficients are real and thus require less data to process.

The Discrete Cosine Transform (DCT) [1] expresses a discrete signal, such as a 2D image or a hyperspectral cube, as a linear combination of mutually uncorrelated cosine basis functions [4]. DCT generates a compact energy spectrum of the signal where the low-frequency coefficients encode most of the signal information. A compact signal representation can be obtained by selecting only the low-frequency coefficient as features.

The 2D-DCT of a 2D image $h(x,y)_{N_1\times N_2}$, and the 3D-DCT of a hyperspectral cube $H(x,y,\lambda)_{N_1\times N_2\times N_3}$ are given by

$$C(u,v)=\Omega_1(u)\Omega_2(v)\sum_{x=0}^{N_1-1}\sum_{y=0}^{N_2-1}h(x,y)\{cos[\frac{\pi(2x+1)u}{2N_1}]cos[\frac{\pi(2y+1)v}{2N_2}]\} \quad (1)$$

$$F(u,v,w)=\Omega_1(u)\Omega_2(v)\Omega_3(w)\sum_{x=0}^{N_1-1}\sum_{y=0}^{N_2-1}\sum_{\lambda=0}^{N_3-1}H(x,y,\lambda) \{cos[\frac{\pi(2x+1)u}{2N_1}]cos[\frac{\pi(2y+1)v}{2N_2}]cos[\frac{\pi(2\lambda+1)w}{2N_3}]\} \quad (2)$$

where $u=\{0,1,...,N_1-1\}$, $v=\{0,1,...,N_2-1\}$, $w=\{0,1,...,N_3-1\}$ and $\Omega_i(u)$ is defined as

$$\Omega_i(u)=\begin{cases}\sqrt{\frac{1}{N_i}} & \text{if } u\text{=0;}\\ \sqrt{\frac{2}{N_i}} & \text{otherwise}\end{cases} \quad (3)$$

The low frequency coefficients near the origin of $F(u,v,w)$ represent most of the energy of the hyperspectral cube (Fig. 1), therefore, the high-frequency coefficients can be discarded. In order to construct our proposed feature vector, we sample a frequency sub-cube $\Gamma(u,v,w)$ of dimensions $(\alpha\times\beta\times\gamma)$ by retaining only the low-frequency elements around the origin of $F(u,v,w)$ i.e., $\{(u,v,w)|u\leq\alpha,v\leq\beta,w\leq\gamma\}$ (Fig. 2). The sub-cube $\Gamma(u,v,w)$ is vectorized and normalized to unit magnitude to obtain the final feature vector $f\in\mathcal{R}^d$, where $d=(\alpha\beta\gamma)$, which is used for classification. For the purpose of classifying the 3D-DCT features, we propose Partial Least Square (PLS) regression.

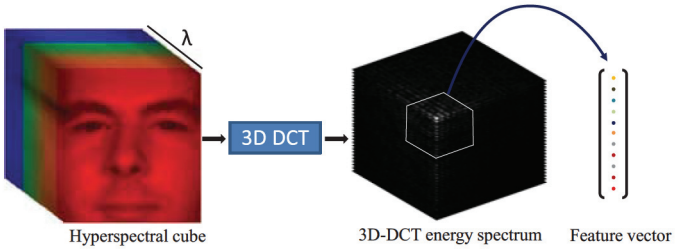


Figure 1: 3D-DCT based feature extraction. A hyperspectral cube (each band is rendered as RGB) and the corresponding 3D-DCT energy spectrum. It can be seen that only few coefficients around the origin contain most of the energy.

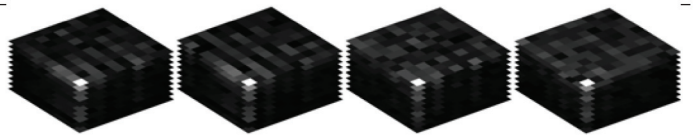


Figure 2: Sub-cubes of size $10\times10\times10$ sampled around the origin of the 3D-DCT energy spectrum of four different subjects of the UWA Hyperspectral Database.

We perform experiments on three standard hyperspectral face databases including the PolyU Hyperspectral [3, 7], CMU Hyperspectral [2] and UWA Hyperspectral databases. The results are compared with five existing hyperspectral face recognition algorithms. Table 1 shows that the proposed algorithm out-performed five existing hyperspectral face recognition algorithms on all three databases. We observe that PLS regression performed better than SRC. It is because PLS basis projects the feature vectors into a latent space in which feature vectors corresponding to the same subject are closer than the feature vectors corresponding to different subjects.

Table 1: Average recognition rates and standard deviations (%) for ten fold experiments on three databases.

Algorithm	PolyU Database	CMU Database	UWA Database
Hyperspectral			
Spectral Signature [5]	24.63±3.87	38.18±1.89	40.52±1.08
Spectral Angle [8]	25.49±4.36	38.16±1.89	37.95±4.15
Spectral Eigenface [6]	70.30±3.61	84.54±3.78	91.51±3.07
2D PCA [3]	71.11±3.16	72.10±5.41	83.85±2.42
3D Gabor Wavelets [9]	90.19±2.09	91.67±2.86	91.50±3.07
Proposed			
2D-DCT + SRC	75.86±2.92	97.44±1.24	97.00±1.29
2D-DCT + PLS	91.43±2.10	97.78±1.28	97.25±1.87
3D-DCT + SRC	87.02±1.72	98.10±0.69	98.00±1.84
3D-DCT + PLS	**93.00±2.27**	**99.00±0.85**	**98.00±1.39**

[1] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete Cosine Transform. *IEEE Transactions on Computers*, 1974.

[2] L. Denes, P. Metes, and Y. Liu. Hyperspectral face database. Technical Report CMU-RI-TR-02-25, 2002.

[3] Wei Di, Lei Zhang, David Zhang, and Quan Pan. Studies on hyperspectral face recognition in visible spectrum with feature band selection. *IEEE Trans. SMC-A*, 2010.

[4] S. A. Khayam. The Discrete Cosine Transform (DCT): Theory and application. Technical Report Michigan State University, 2003.

[5] Zhihong Pan, Glenn Healey, Manish Prasad, and Bruce Tromberg. Face recog. in hyperspectral images. *IEEE TPAMI*, 2003.

[6] Zhihong Pan, Glenn Healey, and Bruce Tromberg. Comparison of spectral-only and spectral/spatial face recog. for personal identity verification. *EURASIP JASP*, 2009.

[7] PolyU-HSFD. www4.comp.polyu.edu.hk/~biometrics/.

[8] Stefan A. Robila. Toward hyperspectral face recognition. In *Proceddings of SPIE ISTE Imaging*, 2008.

[9] L. Shen and S. Zheng. Hyperspectral face recognition using 3d gabor wavelets. In *ICPR*, 2012.

Merging Pose Estimates Across Space and Time

Xavier P. Burgos-Artizzu[1]
xpburgos@caltech.edu
David Hall[1]
dhall@caltech.edu
Pietro Perona[1]
perona@caltech.edu
Piotr Dollár[2]
pdollar@microsoft.com

[1] California Institute of Technology
Pasadena, CA, USA
[2] Microsoft Research
Redmond, WA, USA

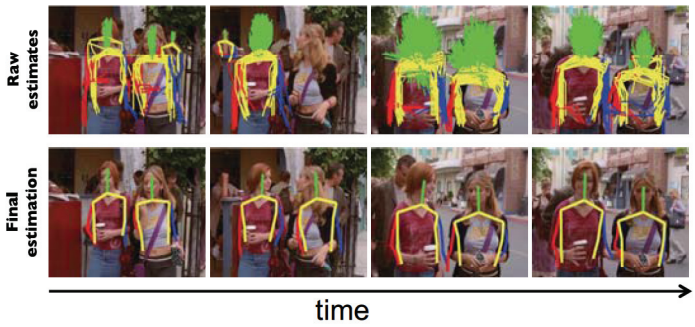


Figure 1: Our pose-NMS approach merges together pose estimates in space and time, outputting a final set of estimates more accurate than any of the individual ones. This is achieved by performing a robust average while simultaneously solving the correspondence problem.

Data driven approaches for pose estimation naturally output a set of pose hypothesis and rely on 'non-maximum suppression' (NMS) techniques to merge detections that are associated with the same objects. NMS is well developed for the case of object detection where the goal is to merge rigid object locations (bounding boxes) [5, 6]. However, it is still unclear how to extend it to flexible pose estimates. Applying standard NMS independently to each part location as in [7] fails to explicitly leverage the higher dimensionality of pose parameterization.

Our first contribution is a principled framework for merging multiple pose estimates in a single frame. This can be viewed as a generalization of NMS beyond bounding boxes. Our proposed approach makes minimal assumptions about the underlying method for pose estimation and generates a final set of pose estimates that are more accurate than any of the individual ones. We achieve this by performing a robust average while simultaneously solving the correspondence problem between pose estimates generated by multiple objects.

Our second contribution is to extend our approach to the multi-frame setting using the same mathematical framework, resulting in pose estimates that are further improved. While our approach is inspired by the recent success of 'tracking by detection' approaches [1, 2, 3, 4], we sidestep many of the inherent challenges associated with full tracking (e.g. objects entering and leaving a scene, extended periods of occlusion, etc.). Instead we present a principled, simple approach for merging multiple independent pose estimates across space and time and outputting both the number and pose of the objects present in a scene, see Fig. 1.

Pose-NMS is a versatile approach. It can be used to merge pose estimates in a single image or in an entire video, controlling the desired amount of temporal consistency. In scenarios where number of objects is fixed for long periods of time, (*e.g.* animals in a cage) it can be used to perform joint optimization over $K > 1$, improving joint reasoning. Pose-NMS can also be used to find all relevant trajectories when number of objects is variable.

We evaluate our approach on three different tasks: 1) Human body pose estimation, 2) Human face landmarks and 3) Animals. We collected 1,000 clips and manually annotated pose on the last frame of each clip, to measure how much pose estimation gets improved on frame T given previous $1 \leq t \leq T-1$ frames. Clip lengths vary from 1-10s.

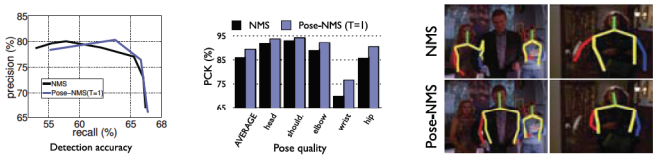


Figure 2: Results on (static) Buffy Stickmen dataset. Pose-NMS performs slightly better for detection but consistently improves the quality of pose estimates around 5%.

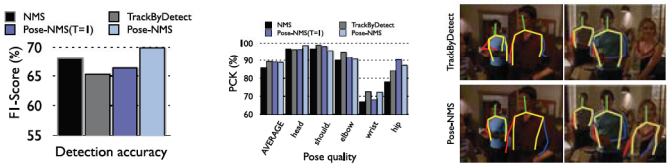


Figure 3: Results on the Video Buffy dataset. Single-frame Pose-NMS improves pose quality around 5% at similar detection accuracy compared with standard NMS. Full (multi-frame) Pose-NMS improves detection accuracy 6% while maintaining similar pose quality compared to running NMS prior to our tracking phase.

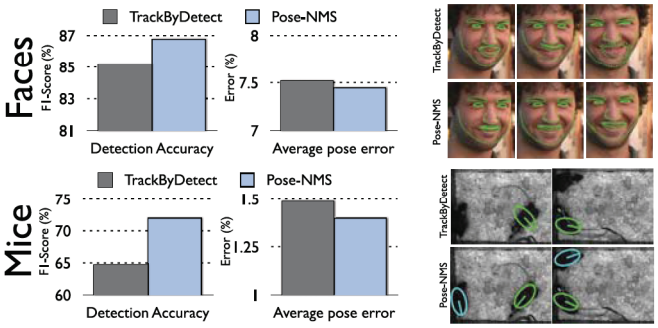


Figure 4: Results on Faces and Mice. Our approach improves detection by 3-7% and pose estimation quality between 2-7%. See text for details.

Single-frame Pose-NMS reaches 3% higher precision at similar recall rates compared to standard NMS, see Figure 2. More importantly, it consistently improves the quality of pose estimates by more than 5% on all body parts. This shows that Pose-NMS, unlike standard NMS, is capable of generating a final set of pose estimates that are more accurate than any of the original ones.

Full Multi-frame Pose-NMS improves detection accuracy between 3-7% and pose estimation quality between 1-7% when compared with running NMS prior to our multi-frame optimization (related to standard tracking-by-detection schemes), see Figures 3 and 4. Code can be downloaded from the authors' websites.

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

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

[3] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *PAMI*, 33(8):1619–1632, 2011.

[4] D. A. Forsyth D. Ramanan, and and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29(1):65–91, 2007.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

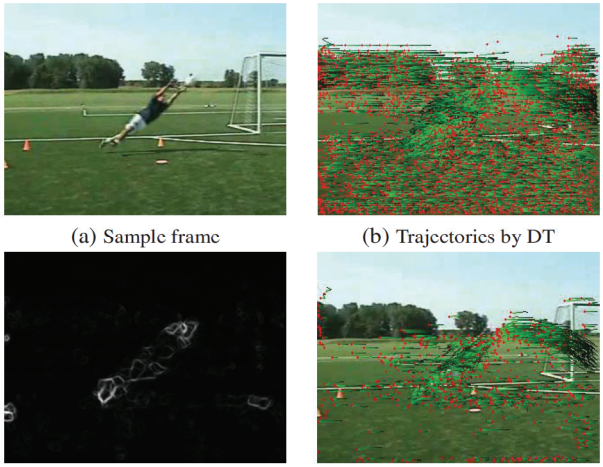
[6] C. Desai, D. Ramanan, and C.C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[7] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2013.

Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition

Xiaojiang Peng[1,2]
xiaojiangp@gmail.com
Yu Qiao[2,3]
yu.qiao@siat.ac.cn
Qiang Peng[1]
qpeng@home.swjtu.edu.cn
Xianbiao Qi[2]
qixiaobiao@gmail.com

[1] Southwest Jiaotong University, Chengdu, P.R. China
[2] Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] The Chinese University of Hong Kong



(a) Sample frame (b) Trajectories by DT (c) Motion boundary image (d) Trajectories by DT-MB

Figure 1: The trajectories captured by original DT and our DT-MB.

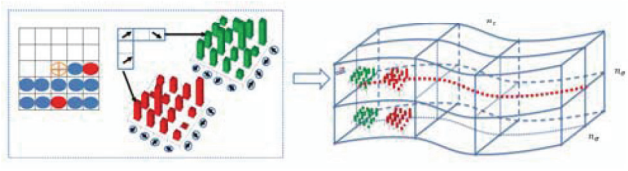
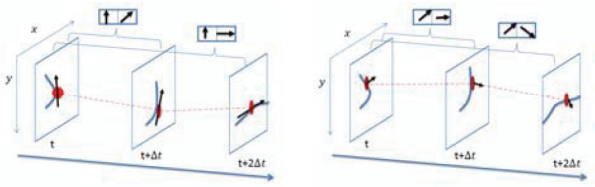


Figure 2: An example of spatial co-occurrence features with grid of size $n_\sigma \times n_\sigma \times n_\tau$.



(a) T-CoHOG and T-CoMBH (b) T-CoHOF

Figure 3: Temporal co-occurrence descriptors. (a): temporal pairs of gradient orientations in T-CoHOG or T-CoMBH. (b): temporal pairs of optical flow orientations in T-CoHOF.

The most important problem in action recognition is how to represent an action video. The approaches can be roughly divided into four categories: (1) human pose based approaches which utilize human structure information; (2) global action template based approaches which capture appearance and motion information on the whole motion body; (3) local feature based approaches which mainly extract valid space-time cuboids; (4) unsupervised feature learning based methods which learn the representation by hierarchical networks. Among these approaches, local feature with bag-of-features (BoF) framework is perhaps the most popular way for action recognition.

With the mentioned popular pipeline, Wang et al. [2] proposed dense trajectory (DT) based features for action video representation and achieved state-of-the-art performance on several action datasets recently. Though its great power, the DT method is expensive in memory storage and computation due to the large number of dense sampled points. In this paper, we improve the DT method in two folds. Firstly, we introduce a motion boundary based dense sampling strategy, called *DT-MB*, which greatly reduces the number of valid trajectories while preserves the discriminative power. Secondly, we develop a set of co-occurrence descriptors which describe the *spatial-temporal context* of motion trajectories.

Our DT-MB is partly implied by MBH descriptor [2] and motion boundary contour system (BCS) in neural dynamics of motion perception [1]. It constrains the sampled points on large magnitude regions of motion boundary image in the sampling step. A comparison with original DT method is illustrated in Fig.1. Our sampling approach removes a large number of points which are not on the motion foreground.

We propose spatial-temporal co-occurrence HOG, HOF and MBH to further enhance the performance of DT. The pipeline of spatial co-occurrence feature in a regularized spatial-temporal grid is shown in Fig.2, and the temporal one is depicted in Fig.3. The spatial co-occurrence HOG [3], HOF and MBH aim to capture complex spatial structures of appearance and motion. Our novel temporal co-occurrence descriptors depict clear motion and appearance changes from successive patches.

Our results of individual co-occurrence descriptors on three datasets are illustrated in Fig.4. It indicates that temporal context information for pure spatial feature is more effective, and spatial context information for pure temporal features is beneficial. Table 1 shows the combined results in detail.

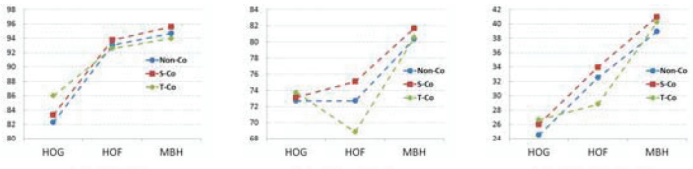
Table 1: Different combinations of descriptors using standard BOF.

Combination	KTH	YouTube	HMDB51
Trajectory+HOG+HOF+MBH	93.63	84.25	45.90
HOG+HOF+MBH	93.98	83.48	45.88
Trajectory+S-Co + T-Co	94.79	85.70	48.98
S-Co + T-Co	94.21	85.33	48.89
All combined	94.10	86.30	49.22
Best combined	95.60	86.56	49.22

[1] Stephen Grossberg and Ennio Mingolla. Neural dynamics of motion perception: direction fields, apertures, and resonant grouping. *Perception & psychophysics*, 53(3):243–278, 1993.

[2] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, pages 1–20, 2012.

[3] Tomoki Watanabe, Satoshi Ito, and Kentaro Yokoi. Co-occurrence histograms of oriented gradients for pedestrian detection. *Advances in Image and Video Technology*, pages 37–47, 2009.



(a) KTH (b) YouTube (c) HMDB51

Figure 4: Percentage accuracies of all the individual descriptors on three datasets. "Non-Co" corresponds to the original descriptors in [2].

Can Your Eyes Tell Me How You Think? A Gaze Directed Estimation of the Mental Activity

Laura Florea
http://alpha.imag.pub.ro/common/staff/lflorea/
Corneliu Florea
http://alpha.imag.pub.ro/common/staff/cflorea/CF_index.html
Ruxandra Vrânceanu
rvranceanu@alpha.imag.pub.ro
Constantin Vertan
http://alpha.imag.pub.ro/common/staff/vertan/cv_eng.html

Image Processing and Analysis Laboratory, LAPI
University "Politehnica" of Bucharest
Bucharest, Romania

We investigate the possibility of estimating the cognitive process used by a person when addressing a mental challenge by following the Eye Accessing Cue (EAC) model from the Neuro-Linguistic Programming (NLP) theory [1]. This model, showed in figure 1, describes the eye-movements that are not used for visual tasks (non visual movements) and suggests that the direction of gaze, in such a case, can be an indicator for the internal representational system used by a person facing a given query.

The actual EAC is thought to be identified by distinguishing between the relative position of the iris and the eye socket (lid edge). Our approach is to determine the four limits of the eye socket: the inner and outer corners, the upper and lower lids and the iris center and to subsequently analyze the identified region.

The entire method flowchart is presented in figure 2. The schematics of the method used for independently looking for the position of each **eye landmark** is described in figure 3. Given the face square position by Viola-Jones algorithm [4] and the eye centers given by the method from [3], we fuse information related to position, normalized luminance, template matching and shape constraining. For position and luminance, we construct priors over the training database, while for template matching we describe a patch by concatenation of integral and edge projections on horizontal and vertical directions. The score of how likely is a patch to be centered on the true landmark position is given by a Multi Layer Perceptron.

For the shape constrain, inspired by the CLM [2], we construct the probability density function in the eigenspace of the shapes in the training set. By ordering the landmarks according to a prior confidence (e.g. eye outer corners are more reliable than upper and lower eye boundaries) and by keeping all points fixed with the exception of the current least reliable, we build the likelihood of various current landmark positions. This information is fused with previous stages and we iteratively improve the landmark position.

The *final landmark position* is taken as the weighted center of mass of the convex combination between initial stages and shape likelihood.

To study the specific of the gaze direction we introduce **Eye-Chimera database**, which comprises 1172 frontal face images, grouped according to the 7 gaze directions, with a set of 5 points marked for each eye: the iris center and 4 points delimiting the bounding box.

Recognizing individual EACs. The recognition of the EAC case (gaze direction) is done by identifying the position of the iris center inside the eye socket complemented by the information of the interior of the eye delimited shape. The interior of the eye quadrilateral shape is described by the integral projections normalized to 32 samples. For the actual recognition we have trained a random forrest to take as input the EAC feature (landmarks positions and integral features).

We consider two types of recognition situations: three cases (looking left corresponding to constructed type of mental activity, looking straight - defocused and looking right - remembering) and 7 cases (described in figure 1). For the 3 case scenario, the approach upper limit is found using the *manual landmarks* where a recognition rate of *91.2%* is found. When applied on the actual data, the recognition rate was **78.57%** for three classes.

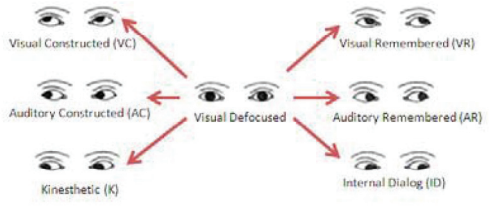


Figure 1: EACs as defined in the NLP theory. The model states that the specifics of the thinking mechanism is connected with a non-visual gaze orientation.

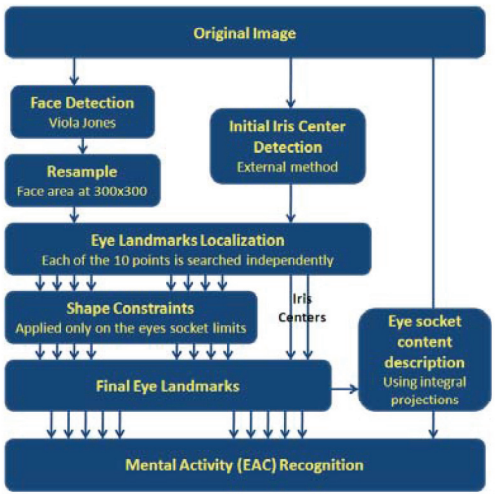


Figure 2: Algorithm flow chart. The method used for independently localizing the eye landmarks is presented in figure 3

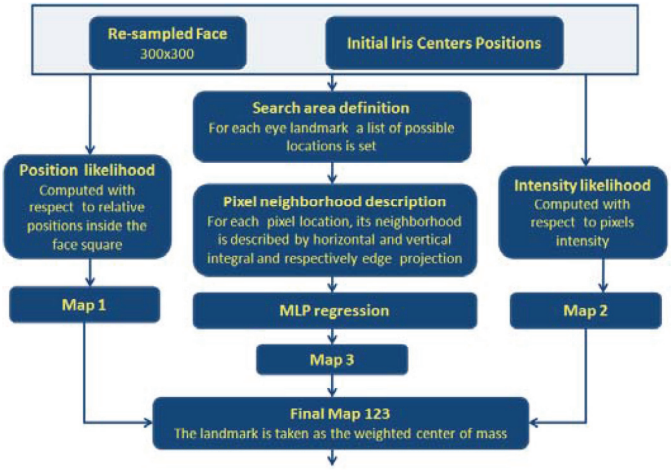


Figure 3: The flow chart of the method used for independently searching each eye landmark.

[1] R. Bandler and J. Grinder. *Frogs into Princes: Neuro Linguistic Programming*. Moab, UT: Real People Press, 1979.

[2] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, pages 929 – 938, 2006.

[3] L. Florea, C. Florea, C. Vertan, and R. Vranceânu. Zero-crossing base image projections encoding for eye localization. In *EUSIPCO*, pages 150 – 154, 2012.

[4] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2): 137–154, 2004.

Incremental Solid Modeling from Sparse and Omnidirectional Structure-from-Motion Data

Vadim Litvinov
vadim.litvinov@univ-bpclermont.fr
Maxime Lhuillier
http://maxime.lhuillier.free.fr

Institut Pascal
University of Blaise Pascal
Clermont-Ferrand, France

We introduce a new incremental 2-manifold surface reconstruction method. Compared to the previous works, its input is a sparse 3D point cloud estimated by a Structure-from-Motion (*SfM*) algorithm instead of a more common dense input. The main argument against such a method is that the lack of points implies an inaccurate scene surface. However, the advantages like point quality (thanks to the *SfM* machinery including bundle adjustment and interest point detection) and simplified resulting surface makes it worth of exploration.

Our algorithm is incremental since the surface is locally updated for every new camera pose (and its 3D points) estimated by *SfM*. This is an advantage compared to global methods like [5] or [2] for applications which require a surface while reading the video sequence. Compared to [6], our method avoids prohibitive time complexity in presence of loops in the camera trajectory. Last but not least, unlike other incremental methods like [3] the output surface is a 2-manifold, i.e. it is a list of triangles in 3D such that the neighborhood of every surface point is topologically a disk. This property is needed to define surface normal and curvature [1] and thus is used by many mesh processing and computational geometry algorithms.

Now we introduce notations. Let P be a set of 3D points on the unknown scene surface. The 3D Delaunay triangulation of P is a list T of tetrahedra which partition the convex hull of P. A list O of tetrahedra ($O \subseteq T$) represents the reconstructed object whose volume is $|O|$, the union of the O tetrahedra. Border δO is the list of triangles (tetrahedra faces) which are included in exactly one tetrahedra of O. The union of triangles $|\delta O|$ is our target surface and should be a 2-manifold. *SfM* also provides visibility knowledge R_i: every point $\mathbf{p}_i \in P$ is computed from camera locations $\mathbf{c}_j$ where $j \in R_i$. This implies that $|\delta O|$ should not intersect the rays (line segments) $\mathbf{c}_j\mathbf{p}_i, j \in R_i$ except at $\mathbf{p}_i$. The tetrahedra intersected by a ray are labeled *freespace*, the others are *matter*. Let F be the set of *free-space* tetrahedra. In practice, $|\delta F|$ is not a 2-manifold.

One iteration of our algorithm is shown in Fig. 1. At image (or time) $t+1$, we have the following input

- 3D Delaunay triangulation T_t;
- list F_t of *freespace* tetrahedra such that $F_t \subseteq T_t$;
- list O_t such that $O_t \subseteq F_t$ and $|\delta O_t|$ is manifold;
- list P_{t+1} of new *SfM* points, camera locations $\mathbf{c}_{t'}$ where $t' \leq t+1$.

and output

- 3D Delaunay triangulation T_{t+1} by adding P_{t+1} in T_{t+1};
- list F_{t+1} of *freespace* tetrahedra (by raytracing in T_{t+1}) such that $F_{t+1} \subseteq T_{t+1}$;
- list O_{t+1} such that $|\delta O_{t+1}|$ is manifold and O_{t+1} is as largest as possible in F_{t+1}.

Moreover, three assumptions are introduced for complexity reasons. First, $\mathbf{p}_i \in P_t$ and $t' \in R_i$ imply $t \in R_i$ and $t' \leq t$ (*SfM* does not update a reconstructed point once it is a vertex of T). Second, the length of all rays $\mathbf{p}_i\mathbf{c}_t$ are bounded by $r > 0$ (a consequence of standard *SfM* point filtering). Third, the diameters of the tetrahedra are bounded by $l > 0$ thanks to a large Cartesian grid of Steiner (extra) vertices. Then the addition of P_{t+1} is a local update of T included in ball B_{t+1} centered at $\mathbf{c}_{t+1}$ with radius $r+l$.

Assume that we add P_{t+1} in the Delaunay T_t as soon as the O_t computation is done. This destroys a list D of tetrahedra, which can contain tetrahedra in O_t (Fig. 1.f). The problem is the following: if we initialize $O_{t+1} = O_t \setminus D$, $|\delta O_{t+1}|$ can be non manifold, and there is no obvious method to update O_{t+1} such that $|\delta O_{t+1}|$ becomes manifold again.

Our idea is to compute D without updating T to ensure that $|\delta O_{t+1}|$ is always a manifold. In practice, the results are better if we replace D by list E of the tetrahedra included in B_{t+1} (then $D \subseteq E \subseteq T_t$). We initialize $O_{t+1} = O_t$ (Fig. 1.a) and we progressively remove tetrahedra from O_{t+1} such that $|\delta O_{t+1}|$ is maintained manifold. This shrinking process is stopped if $O_{t+1} \cap E = \emptyset$ (Fig. 1.b) or if O_{t+1} can not decrease.

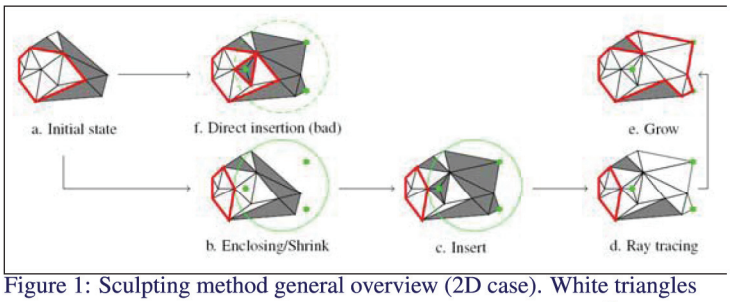


Figure 1: Sculpting method general overview (2D case). White triangles are freespace (in F), grey triangles are matter, red line is border δO, green dots are newly inserted points and the green circle encloses the tetrahedra that will be deleted by the addition of the green dots.

To add the points P_{t+1} to T_t, we initialize $T_{t+1} = T_t$ and $F_{t+1} = F_t$. For every point $\mathbf{p}$ in P_{t+1}, we compute the list $D(\mathbf{p})$ of tetrahedra in T_{t+1} which would be destroyed if we add $\mathbf{p}$ to T_{t+1}. Point $\mathbf{p}$ is added to T_{t+1} and $D(\mathbf{p})$ is removed from F_{t+1}, if and only if $D(\mathbf{p}) \cap O_{t+1} = \emptyset$ (Fig. 1.c). Since O_{t+1} is unchanged, $|\delta O_{t+1}|$ is still a 2-manifold.

Then we perform the ray tracing in T_{t+1} to increase the list F_{t+1}. First, we trace the rays of P_{t+1}, i.e. every ray $\mathbf{p}_i\mathbf{c}_j, j \in R_i$ of every point $\mathbf{p}_i$ in P_{t+1}, since this computation was not done before. Second, we collect in a list R every ray of $P_{t'}$ (where $t' \leq t$) which *can* intersect a tetrahedron of N and trace them (Fig. 1.d). List N is the list of the newly created tetrahedra. List R is computed by intersections of bounding boxes of N and $P_{t'}$.

Finally, O_{t+1} progressively grows in *freespace* F_{t+1} by adding tetrahedra such that $|\delta O_{t+1}|$ is maintained manifold (Fig. 1.e). As the finishing touch, incremental post-processing (handle removal and Laplacian smoothing) is performed on δO_{t+1} to improve the output surface quality.

We experiment on a large scale real urban scene using the Ladybug camera, as well as on synthetic images with ground truth. The trajectory of the real sequence is 2.5 km long and includes a large loop (2.3 km). 483k points are reconstructed [4] in 1306 keyframes, then the loop is closed thanks to a global bundle adjustment. The final surface has 528k triangles (`http://www.youtube.com/watch?v=w1AQfvhGx5I` or joint video). The synthetic sequence has 1553 images generated by ray-tracing with textures extracted from real images taken in a city. The trajectory is 621 m long and is a closed loop. 145k points are reconstructed in 346 keyframes. The estimated error between the final surface (182k triangles) and the ground truth is 0.76 m for 70% of the pixels.

Several improvements of the described method are subjects of future works. First, the choice of a better (smaller) enclosing area E would decrease the computation time. Second, our implementation and matching method should be improved. Last, several lines of investigation exist to enhance the quality of the reconstructed surface (curve reconstruction and integration in the Delaunay triangulation, surface denoising methods that make a better usage of the *SfM* properties and scene priors).

[1] M. Botsch, L. Kobbelt, M. Pauly, P. Alliez, and B. Levy. *Polygon mesh processing*. AK Peters, 2010.

[2] P. Labatut, J.P. Pons, and R. Keriven. Efficient multiview reconstruction of large-scale scenes using interest points, Delaunay triangulation, and graph cut. In *Proc. ICCV*, 2007.

[3] D. Lovi, N. Birkbeck, D. Cobzas, and M. Jagersand. Incremental free-space carving for real-time 3D reconstruction. In *Proc. 3DIMPVT*, 2012.

[4] E. Mouragnon, M.Lhuillier, M. Dhome, F. Dekeyer, and P. Sayd. Generic and real-time structure from motion. In *Proc. BMVC*, 2007.

[5] Q. Pan, G. Reitmayr, and T. Drummond. ProFORMA: probabilistic feature-based on-line rapid model acquisition. In *Proc. BMVC*, 2009.

[6] S. Yu and M. Lhuillier. Incremental reconstruction of a manifold surface from sparse visual mapping. In *Proc. 3DIMPVT*, 2012.

Egomotion Estimation Using Binocular Spatiotemporal Oriented Energy

Hao Zhong
zhhao31@gmail.com
Richard P. Wildes
wildes@cse.yorku.ca

Department of Electrical Engineering
and Computer Science
York University
Toronto, Ontario, Canada

Camera egomotion estimation is concerned with the recovery of a camera's motion (e.g., instantaneous translation and rotation) as it moves through its environment. It has been demonstrated to be of both theoretical and practical interest and therefore has been the subject of considerable research. In the light of previous work on egomotion estimation, the present paper makes three contributions. First, a novel algorithm for egomotion estimation based on binocularly matched orientation measurements in visual spacetime, (x,y,t), is presented. Basing the estimation on spacetime orientation measurements makes it possible to recover egomotion without the need to establish temporal correspondences or convert disparity into 3D world coordinates. Second, the algorithm is based on a formal analysis that relates binocularly matched orientations in visual spacetime to camera egomotion. It appears that this relationship has not been presented previously. Third, the estimation algorithm has been realized in software and evaluated quantitatively on a novel laboratory dataset with groundtruth as well as qualitatively on both indoor and outdoor real-world datasets. Performance is evaluated relative to comparable alternative algorithms. As part of the evaluation, a new binocular video dataset is introduced that includes groundtruth egomotion and is available to the community. Figure 1 overviews the proposed approach.

Technical approach. Binocularly matched, local measurements of spatiotemporal oriented energy (SOE) serve as the data on which the developed approach to egomotion estimation operates. SOEs provide an integrated way to capture spatial appearance and temporal characteristics of an image sequence [3]; therefore, they have the potential to support recovery of egomotion via consideration of temporal dynamics of spatial information as a function of egomotion. For present purposes, local SOE measurements are recovered separately in the left and right streams of an input binocular video via convolution with a bank of Gaussian second derivative filters and their Hilbert transforms, which are combined in quadrature to yield energy measurements. To establish correspondences between points in the left and right image sequences any reliable algorithm for establishing binocular correspondence could be applied on a framewise basis to the original image sequences. Here, since SOEs are available and previously have been shown useful for stereo video matching [3], that matching approach is applied to establish the needed left-right correspondences.

Given matched pointwise SOEs in the left and right image streams, egomotion estimation is based on an analysis that parameterizes binocularly corresponding orientations in visual spacetime, (x,y,t), in terms of camera egomotion. For these purposes, camera egomotion is defined in terms of instantaneous six degrees-of-freedom translation, $\mathbf{T}$, plus rotation, Ω, as the camera traverses an otherwise rigid 3D environment. By appealing to brightness constancy, it is further shown that orientations in visual spacetime that yield locally minimal energy are indicative of global egomotion. Correspondingly, global egomotion can be estimated by steering local SOE measurements to the direction that yields minimal energy as parameterized by $\mathbf{T}$ and Ω. Due to the nonlinear dependence of the resulting objective function on $\mathbf{T}$ and Ω, Gauss-Newton refinement is employed to obtain the solution. To facilitate efficient processing with large capture range, estimation is performed within a coarse-to-fine refinement scheme.

Empirical evaluation. The proposed approach to egomotion estimation has been realized in C++ for execution on standard PCs. On a 3.4GHz processor with 16GB RAM, the execution rate is ≈ 84 milliseconds/frame for 512×384 images, beyond the time required for SOE filtering and stereo matching. Significantly, previous research has shown that both SOE filtering and stereo matching can be done in real-time, *e.g.*, [3]. Thus, the overall approach has potential for real-time applications.

The implementation has been evaluated on three datasets. The first

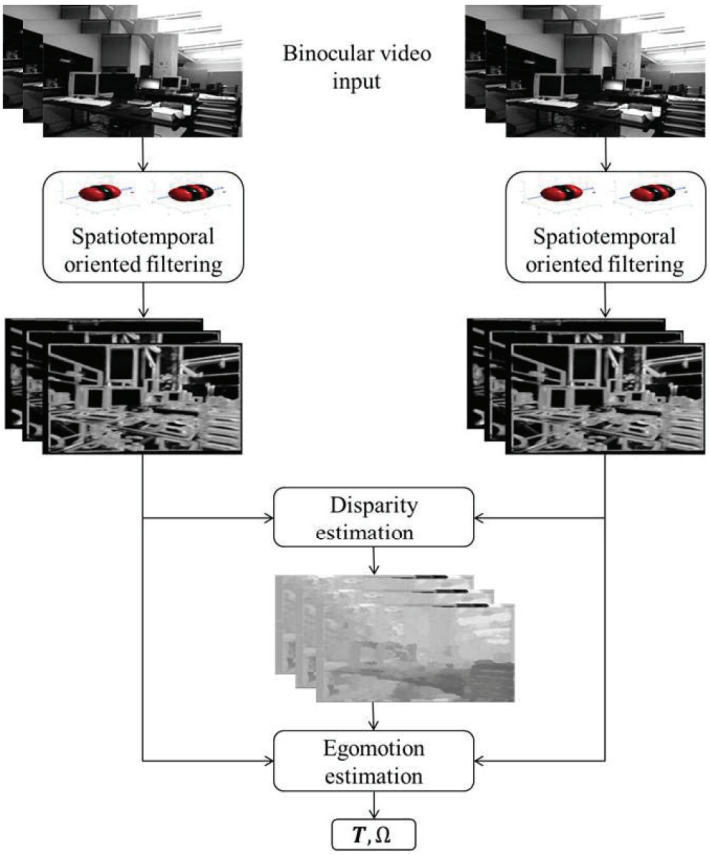


Figure 1: Overview of the proposed approach to egomotion estimation. Binocular video serves as input. The video streams are filtered with a bank of spatiotemporal filters that are tuned for 3D orientation in visual spacetime, (x,y,t). The filtered images are used as input to a stereo disparity estimator [3]. Recovered disparity is combined with the oriented filtering results to yield egomotion estimates as instantaneous translation, $\mathbf{T}$, and rotation, Ω.

was acquired in a calibrated laboratory setting and includes groundtruth egomotion. This dataset is available to the community. The second and third datasets were captured in more naturalistic settings, both indoor (an office) and outdoor (building exterior with foreground ground cover). For the sake of comparison, two alternative egomotion estimation algorithms [1, 2] also were executed on the same datasets. Both qualitative and quantitative results show that the developed approach is competitive with and even exceeds the accuracy of the alternative algorithms.

[1] K. Derpanis and P. Chang. Closed-form linear solution to motion estimation in disparity space. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 268–275, 2006.

[2] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 486–492, 2010.

[3] M. Sizintsev and R. Wildes. Spatiotemporal oriented energies for spacetime stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

Online Motion Agreement Tracking

Zheng Wu
wuzheng@bu.edu
Jianming Zhang
jmzhang@bu.edu
Margrit Betke
betke@bu.edu

Computer Science Department
Boston University
Boston, USA

We present a new appearance model for online multi-target tracking, where a patch-based representation of each target to be tracked is introduced along with a sequential update scheme, which we call "motion agreement tracking" (MAT). This method updates the appearance model of each target online by indirectly evaluating the motion consistency among its local patches. We show that a distance measure based on appropriately re-weighted local patches will successfully reduce tracker errors especially those lead to track fragmentation and track switching. We also demonstrate that the performance results of a well-designed online 2D tracker, like the MAT algorithm, can actually measure up to those of state-of-the-art offline algorithms on various popular pedestrian tracking benchmarks. Our competitive results are particularly appealing since our technique is very efficient. They also suggest that the role of a proper appearance model may be more important than researchers deemed for video-based tracking, where the majority of previous studies focused on motion dynamics.

In essence, we designed a person-specific appearance model with a collection of local image patches by dividing the bounding box of a detected person into a grid representation. Each local patch k is described by a 64-bin color histogram in HSV space. Each patch is associated with a weight w_k, which is set to be uniform when the tracker is initialized. We would like to assign a high weight to a stable patch that does not change significantly over time, and a low weight that the patch belongs to the background or represents a fast-changing part of the object.

When a detection is assigned to the tracker after solving the tracker-measurement assignment problem, a filtered estimate of the object's global motion vector v is computed by a Kalman filter. At the same time, the method estimates the local displacement v_k of each patch based on a similarity measure. Here, we compare two popular measures in our system: the maximum normalized cross-correlation and the minimum histogram intersection distance.

Given the local motion estimates, our method re-weights each patch by checking the agreement between v_k and the global motion v. The intuition is that if v_k is similar to v, then this local patch moves along with the pedestrian, so it is more likely to be a stable region that does not undergo significant appearance change. By focusing our effort on the most stable patches, we can construct a similarity measure that can distinguish between interacting objects. The discrete level g of the agreement is computed by our implementation as follows:

$$\begin{aligned} \theta_k &= \cos^{-1}\left(\frac{v_k \cdot v}{\|v_k\|\|v\|}\right) \\ g_k &= \begin{cases} 2, & \text{if } \theta_k < \frac{\pi}{4} \\ 1, & \text{if patch is in an occlusion relationship} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (1)$$

where θ_k is the angle between the two velocity vectors v and v_k. The motion score g is defined to be symmetric on both agreement and disagreement sides; the magnitude is not important here, as its contribution to the following weight update will be normalized. A score of $g_k = 1$ suggests a random guess. Using g_k, our algorithm updates the weight w_k associated with each patch k at time t as follows:

$$\begin{aligned} w_k(t) &= \frac{\alpha^{(t-1)} w_k^{(t-1)} + g_k}{\alpha^{(t-1)} + \sum_k g_k} \\ \alpha^{(t)} &= \frac{\alpha^{(t-1)} + \sum_k g_k}{2}, \end{aligned} \quad (2)$$

where α is a self-adaptive learning rate which controls how much the current estimates influence the update. A large value of α suggests a smooth update of weights at the current frame, which happens when most local motion estimates agree with the global motion model. Both α and w are non-negative, and w is always normalized.



Figure 1: Sample images from our tracking results. Colors and numbers indicate tracks corresponding to different people.

Finally, the distance function between an object i and a candidate detection j in the data association step is given as:

$$c_{i,j} = (1 - \frac{|b_i \bigcap b_j|}{|b_i \bigcup b_j|}) + \lambda \sum_k^{\#\text{patches}} w_k (1 - \sum_n^{\#\text{bins}} \min(h_i^n, h_j^n)), \quad (3)$$

where the first term evaluates the alignment of two bounding boxes i and j, and the second term measures the re-weighted difference of the appearance of i and j based on the histogram intersection distance. Parameter λ balances the contributions of the motion and appearance terms.

We integrated our appearance model into an online tracking system and tested it extensively on nine tracking benchmarks. The performance was evaluated by comparing our MAT method with a simple baseline, an online tracking algorithm that models the appearance holistically [5] and four offline tracking algorithms [1, 2, 3, 4] that solve the data association globally. Our results suggest that our online MAT algorithm indeed outperforms state-of-the-art algorithms for various benchmark videos. Given its efficiency and ease-of-use, our MAT algorithm is even valuable for tracking scenarios where its performance is expected to be inferior because the tracks it produces online may be used as valuable initializations for offline tracking algorithms.

[1] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR*, 2012.

[2] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking. In *CVPR*, 2011.

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

[4] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *CVPR*, 2012.

[5] J. Zhang, L. Lo Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *Proceeding of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2012.

A Topic Model Approach to Represent and Classify American Football Plays

Jagannadan Varadarajan[1]
vjagan@adsc.com.sg
Indriyati Atmosukarto[1]
indria@adsc.com.sg
Shaunak Ahuja[1]
shaunak@adsc.com.sg
Bernard Ghanem[12]
bernard.ghanem@kaust.edu.sa
Narendra Ahuja[13]
n-ahuja@illinois.edu

[1] Advanced Digital Sciences Center of Illinois, Singapore
[2] King Abdullah University of Science and Technology, Saudi Arabia
[3] University of Illinois at Urbana-Champaign, IL USA

Automated understanding of group activities is an important area of research with applications in many domains such as surveillance, retail, health care, and sports. Despite active research in automated activity analysis, the sports domain is extremely under-served. One particularly difficult but significant sports application is the automated labeling of plays in American Football ('football'), a sport in which multiple players interact on a playing field during structured time segments called plays. Football plays vary according to the initial formation and trajectories of players - making the task of play classification quite challenging.

We define a play as a unique combination of offense players' trajectories as shown in Fig. 1(e). This dynamic group activity (play) is labeled by play type (run or pass) and direction (left, middle, or right). A *run* play is simply an attempt to advance the ball via running with the ball in left/middle/right directions, while a *pass* play is defined as an attempt to advance the ball by throwing it to the left/middle/right third of the field. In this paper, we address the combined problem of automatic representation and classification of football plays as summarized in Fig. 1. Such analysis will allow broadcasters to analyze large collections of videos, index and retrieve them more efficiently. Furthermore, this will also help coaches to infer patterns and tendencies of opponents more efficiently, resulting in better strategy planning in a game. For instance, Team A's coach can input Team B's videos and quickly understand commonly chosen play types and underlying strategies of Team B.

Problem definition. Given a play (short video clip) consisting trajectories of six important offense players: quarterback (QB), running back (RB), wide receiver left (WR-L), wide receiver right (WR-R), tight end left (TE-L) and tight end right (TE-R), we consider two problems: 1) *representation*, *i.e.*, derive canonical templates (common routes taken by players) for each play type, and 2) *classification*, *i.e.*, predict a label from one of the six play types for a given play, as shown in Fig. 1(e).

More formally, we are given D plays (or clips) denoted by $\{v_d\}, 1 \leq d \leq D$, where a play v_d is represented as a set of trajectories from R players: $\{X_r^d(t) : 0 \leq t \leq T, 1 \leq r \leq R\}$, T is the maximum possible duration of a trajectory, R is the maximum number of tracked players, and each play, indexed by d, is also associated with a unique class label $y_d \in \mathcal{C} = \{1, 2, \cdots, C\}$ called a play type. Our objective is to learn a template τ_y corresponding to each class y, and a classifier function $F(v_d)$ that predicts a label $\hat{y}_d$ with high classification accuracy.

Approach. Our approach to modeling play types from trajectories builds on recent success of supervised topic models (STMs). Topic models in general, are used to capture dominant co-occurrent patterns, called "topics", in large collections of data. Intuitively, such patterns correspond to unique player interactions. For instance, run plays usually involve an intersection of QB and RB. STMs are effective in learning such salient co-occurring actions while achieving good classification performance. A particular example of STMs is the MedLDA model [2]. The MedLDA model learns salient topics with high classification accuracy by optimizing an objective function that combines the dual goals of maximizing inter-class margins and data log-likelihood simultaneously.

For our task of generating play type templates and classifying play types, we apply the MedLDA model on plays represented as a bag-of-words. In our case, a word in our vocabulary is defined by three different aspects of players' trajectories namely: motion direction, time and player role. More precisely, we use the following procedure to create words from plays: **(1)** motion directions coming from consequent trajectory observations are quantized into 18 equal sized directional bins of 20 degrees each; **(2)** temporal information is captured by grouping observations from every consecutive 10 frames into single time instance. Thus, each observation is also associated with a time-stamp; **(3)** a player's trajectory is associated with the player role, *e.g.*, QB, RB, etc. In our experiments, we use trajectories of six different players in creating our vocabulary. Therefore, a play with six different player roles occurring for a maximum duration of 250 frames can result in a vocabulary of $18 \times 25 \times 6 = 2700$ words.

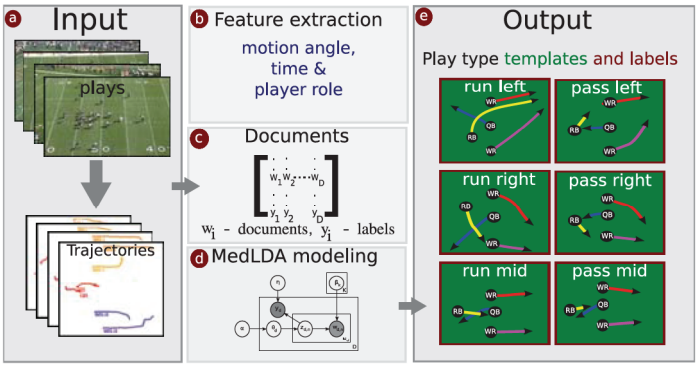


Figure 1: The goal of our work is to represent and classify play types in football videos. Our framework consists of a) player tracking, b) feature extraction, c) document creation and d) modeling using MedLDA [2]. Our outputs are: play type templates (e) and labels. Templates show are typical trajectories of four players (QB, RB, WR-L and WR-R).

Results. By applying MedLDA model on play clips, we obtain topics as well as a classifier that can be used to predict labels for test play clips. To qualitatively demonstrate that topics indeed capture salient features of play types, we associate each play type to a topic and generate trajectories from the respective topics. Quantitatively, we evaluate classification performance of our method by comparing against three other baselines on 6 class and 2 class (run vs pass) classification tasks for varying model sizes. We show that our approach based on MedLDA achieves 70% accuracy on 6-class and 88% on 2-classes, outperforming all three baselines for any number of topics. We also compare our results with the labels obtained from a sports expert who was asked to label plays based on only trajectory videos. By comparing the confusions of the expert with MedLDA, we show that MedLDA classifies marginally better (70%) than the expert while mostly agreeing on the classification performance for 4 out of 6 classes.

In addition to classifying plays, we extend our study to investigate which of the player roles is most discriminative in general and for a given play type in particular. From this analysis, we show that QB and RB give the highest classification accuracy as their actions largely determine if the play is a run or pass, and that WRs play an important role in pass plays than run plays. We validate our method using experiments on a large dataset comprising 271 play clips from real-world football games, which will be publicly available for future comparisons from [1].

[1] http://publish.illinois.edu/videoadsc/publications/football-trajectory-dataset/.
[2] Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: Maximum margin supervised topic models. *JMLR*, 13:2237–2278, 2012.

Sketch Retrieval via Dense Stroke Features

Chao Ma[1]
chaoma@sjtu.edu.cn
Xiaokang Yang[1]
xkyang@sjtu.edu.cn
Chongyang Zhang[1]
sunny_zhang@sjtu.edu.cn
Xiang Ruan[2]
gen@omm.ncl.omron.co.jp
Ming-Hsuan Yang[3]
mhyang@ucmerced.edu

[1] Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
[2] Omron Coporation
Kyoto, Japan
[3] Electrical Engineering and Computer Science
University of California, Merced
California, USA

Sketch retrieval aims at retrieving most similar sketches from a large database based on one hand-drawn query. As sketches are hand-drawn with different styles to represent objects, sketch retrieval is challenging due to several factors. First, there exist large intra-class differences, as a result of experiential and cognitive differences among individuals. Second, there exist small inter-class differences, due to their loss of visual information (i.e., texture and appearance). Thus, successful retrieval hinges on an effective representation of sketch images and an efficient search method.

In this paper, we propose an effective representation scheme which takes sketch strokes into account with local features, thereby facilitating efficient retrieval with codebooks. The proposed algorithm consists of three components: selecting most representative stroke points, describing stroke features using a quantized histogram of gradients, and representing sketch images using a hierarchical vocabulary tree for matching. Figure 1 shows the main steps of the proposed method.

In this work, we use local stroke features to represent sketches instead of contour segments [1, 3, 6]. As sketch images consists of strokes with no textural information, it is essential to select the most representative stroke points for local features. Since corners and end pixels of strokes always encode important geometric information of a sketch, they are used as anchor positions for dense sampling to encode shape information properly. We compute the corner response of a sketch image I by:

$$E(x,y) = \sum_{u,v} w(u,v)[I(x+u,y+v) - I(x,y)]^2, \tag{1}$$

where $w(u,v) = \exp(-(u^2+v^2)/\sigma^2)$ is the Gaussian kernel. We use these corners as anchors and add a number of points (e.g., twice the desired number of points) randomly sampled on the strokes. We next remove those points, other than the anchors, that are too close to each other, in order to spare the points evenly (i.e., points with large spreads are preferred).This stroke point detection method is summarized in Algorithm 1.

Inspired by [2], we compute the histogram of orientations on the dense gradient field $\mathcal{G}$ from a sketch image, where $\mathcal{G}$ can be approximated by convolution through $\Delta\mathcal{G}$ and sketch Image I [5]:

$$\mathcal{G}(x,y) = \sum_{u,v} I(x,y)\Delta\mathcal{G}(x-u,y-v), \tag{2}$$

where $\Delta\mathcal{G}$ denotes Laplace of Gaussian operator,

$$\Delta\mathcal{G}(x,y) = -\frac{1}{\pi\sigma^2}[1 - \frac{x^2+y^2}{2\sigma^2}]e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{3}$$

Then, we adapt the Poisson-based HOG (PHOG) by first discarding the central distance weights (i.e., we do not compute the distance voting to each grid cell center) and then computing a histogram with coarse quantized orientations (e.g., $d = 4$) with an anti-alias function as follows:

$$\cos(x - \alpha_i)^3 > 0, \tag{4}$$

where x denotes the gradient orientation weighted by its magnitude, and α_i denotes the i-th bin center.

Finally, we use a hierarchical tree to train a codebook in spirit similar to the vocabulary tree [4]. A codebook is organized in a hierarchical vocabulary tree, which maintains structural information of visual words and enables efficient retrieval in sub-linear time. A hierarchical tree can be defined by two parameters, K and L, where K is the number of cluster center and L denotes the depth of tree. We iteratively use the K-means clustering algorithm at each level until the tree grows to the pre-defined level L. The nodes of the tree represent the cluster centers, and each local PHOG feature descriptor of a sketch image can be effectively represented by a path from the root node to a leaf node (See Figure 1). Thus, the histograms of all the paths of local PHOG descriptors is the signature of a sketch. Similar to the inverted indexing scheme, we assign each leaf node a list with image identities (labels) which contain the same PHOG feature descriptor (See Figure 1).

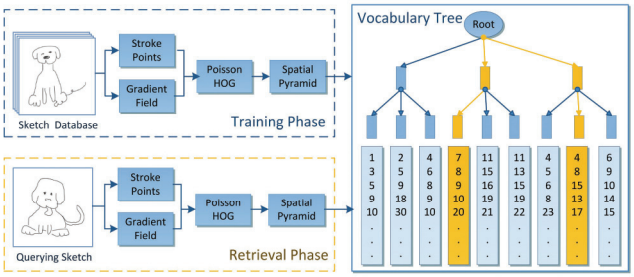


Figure 1: Main steps of the proposed local feature based sketch retrieval: the vocabulary tree is constructed offline for retrieval. Inverted training image identities are indexed below the tree leaves. Each query sketch is retrieved by the hits of the training identities (orange).

Given a query, we find the closest sketches via the sketch descriptor based on their distance using the χ^2 kernel [7]. Given a sketch pair, I_q and I_r, and their corresponding sketch descriptors h_q as well as h_r, we compute their distance as follows:

$$D(h_q,h_r) = \frac{1}{2}\sum_{i=1}^{n}\frac{[h_q(i) - h_r(i)]^2}{h_q(i) + h_r(i)}. \tag{5}$$

Thus, for each query sketch I_q, we use D_Λ to denote the distance matrix of I_q to a subset Λ of the training sketch images, in which each training image has local PHOG features in the same bin as the query sketch in the hierarchical tree. Thus, distance matrix computed on the subset Λ and retrieval can be performed in sub-linear time. The rank k retrievals are:

$$rank(k) = \arg\min_{1,\ldots,k} D_\Lambda. \tag{6}$$

Experimental results on three data sets with more than 20,000 sketch images demonstrate the merits of the proposed algorithm for effective and efficient sketch retrieval, which performs favorably against state-of-the-art methods in terms of retrieval accuracy and execution time.

[1] V. Ferrari, T. Tuytelaars, and L.V. Gool. Object detection by contour segment networks. In *ECCV*, pages 14–28, 2006.

[2] R. Hu, M. Barnard, and J. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, 2010.

[3] M. Liu, O. Tuzel, A. Veeraraghavan, and R. Chellappa. Fast directional chamfer matching. In *CVPR*, pages 1696–1703, 2010.

[4] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[5] P. Perez, M. Gangnet, and A. Blake. Poisson image editing. *Proc. SIGGRAPH*, pages 313–318, 2003.

[6] P. Srinivasan, Q. Zhu, and J. Shi. Many-to-one contour matching for describing and discriminating object shape. In *CVPR*, 2010.

[7] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, pages 3539–3546, 2010.

PedCut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues

Fabian Flohr[1,2]
fabian.flohr@daimler.com
Dariu M. Gavrila[1,2]
www.gavrila.net

[1]Environment Perception Department, Daimler R&D, Ulm, Germany
[2]Intelligent Systems Laboratory, Univ. of Amsterdam, The Netherlands

Person segmentation is a key computer vision problem in a number of application domains, such as image editing, surveillance and intelligent vehicles. This paper presents an iterative, EM-like framework for accurate pedestrian segmentation, combining generative shape models and multiple data cues. It is able to cope with a large variation of pedestrian appearances across cluttered backgrounds. In the E-step, shape priors are introduced in the unary terms of a Conditional Random Field (CRF) formulation, joining other data terms derived from color, texture and disparity cues. In the M-step, the resulting segmentation is used to adapt an Active Shape Model (ASM) [2]. The EM process alternates until the CRF-based segmentation does not appreciably change any more or a maximum number of iterations is reached. Fig. 1 illustrates our framework.

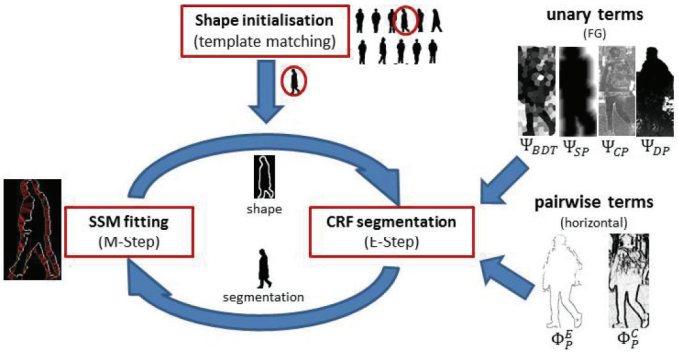


Figure 1: Our EM-like segmentation framework, alternating CRF segmentation (E-step) and SSM fitting (M-step), given shape initialisation.

Shape initialisation Input is an image region of interest (i.e. a bounding box) provided by a pedestrian detector front-end. As ASMs can get stuck in local minima, we obtain our initial shape by matching a set of pedestrian shape exemplars from a training set. We use chamfer matching differentiated by gradient direction (four discretization intervals, not encoding the gradient sign), as in [3]. The best matching shape exemplar is converted to its Statistical Shape Model (SSM) representation (we use several SSMs to model various pose clusters [4], e.g. feet apart/closed); it acts as a shape prior in the following CRF segmentation step.

CRF segmentation Let $\boldsymbol{I}$ and $\boldsymbol{D}$ be the color and disparity values of the image region. We use Semi Global Matching [5] for disparity computation. Furthermore, let $\boldsymbol{S}$ be the superpixel feature vectors of the region.

We specify four unary potentials for the CRF: 1) the sigmoid converted output (Ψ_{BDT}) of a Boosted Decision Tree ensemble trained with dense SIFT and Texton features on the image region, 2) a shape potential (Ψ_{SP}) calculated on a distance transformation obtained from the current shape contour Ω, 3) a GMM-based color potential (Ψ_{CP}) similar to the GrabCut framework [6] based on the current segmentation and 4) a GMM-based disparity potential (Ψ_{DP}) based on the median disparity over the current segmentation.

We further specify two pairwise potentials, which take the form of generalized Potts models [1]: 1) a color-sensitive potential (Φ_P^C), specified such, that it increases the costs of an edge inversely proportional to the color difference in Lab color space of two neighbored pixels i and j, and 2) a contour-sensitive potential (Φ_P^E), which increases the cost inversely proportional to the edge magnitude between pixels i and j, weighted based on the disparity information.

Given the unary and pairwise terms we minimize an energy functional defined on the index set $\mathcal{V}$ with an eight-connected edge neighborhood $\mathcal{E}$, of the following form:



Figure 2: Results after four iterations. First row: input images with initial/final SSM fit (red/white). Second row: correct/missing/excessive segmentation (white/red/cyan). a)-c): Penn-Fudan dataset (BDT+SP+CP); d)-f): Our dataset (BDT+SP+CP+DP)

$$E(\boldsymbol{x},\Omega,\boldsymbol{I},\boldsymbol{D},\boldsymbol{S},\boldsymbol{\omega}) = \sum_{i\in\mathcal{V}} \omega_{BDT}\Psi_{BDT}(x_i,\boldsymbol{S}) + \omega_{SP}\Psi_{SP}(x_i,\Omega) + \omega_{CP}\Psi_{CP}(x_i,\boldsymbol{I}) + \omega_{DP}\Psi_{DP}(x_i,\boldsymbol{D}) + \sum_{i,j\in\mathcal{E}} \omega_P^C\Phi_P^C(x_i,x_j,\boldsymbol{I}) + \omega_P^E\Phi_P^E(x_i,x_j,\boldsymbol{I}.\boldsymbol{D}). \quad (1)$$

Main CRF parameters $\boldsymbol{\omega}$ are the weights for the specified unary and pairwise terms (ω_{BDT}, ω_{SP}, ω_{CP}, ω_{DP}, ω_P^C and ω_P^E). As our pairwise terms stay submodular, we can perform inference with Graph Cut [1] methods.

SSM fitting We use an ASM approach [2] for fitting the SSM model to the obtained CRF segmentation. Point correspondences between SSM and image are given by chamfer matching [3]. As in shape initialisation, we can differentiate chamfer matching by gradient direction. Since we have a binary segmentation, we can here utilize information about the gradient sign to improve matching (i.e. eight discretization intervals for gradient direction).

Results We show the benefit of different cue combinations and the ability of the framework to improve results with each additional cue. On the public Penn-Fudan dataset [7] (Fig. 2 a-c), we outperform the state-of-the-art by more than 5% on foreground accuracy while remaining ahead on background accuracy. Further we provide results on our own pedestrian dataset (Fig. 2 d-f), captured from on-board a vehicle, which includes disparity data. This dataset is made public for non-commercial research purposes to facilitate benchmarking.

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE PAMI*, 23(11):1222–1239, 2001.
[2] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995.
[3] D. M. Gavrila. A Bayesian, exemplar-based approach to hierarchical shape matching. *IEEE PAMI*, 29(8):1408–1421, 2007.
[4] J. Giebel and D. M. Gavrila. Multimodal shape tracking with point distribution models. *Pat. Rec.*, pages 1–8, 2002.
[5] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE PAMI*, 30(2):328–341, 2008.
[6] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *Proc. ACM TOG (SIGGRAPH)*, 23(3):309–314, 2004.
[7] L. Wang, J. Shi, G. Song, and I.-F. Shen. Object detection combining recognition and segmentation. In *Proc. ACCV*, pages 189–199. Springer, 2007.

Fast Approximation of Distance Between Elastic Curves using Kernels

Hedi Tabia[1]
http://perso-etis.ensea.fr/tabia
David Picard[1]
http://perso-etis.ensea.fr/~picard
Hamid Laga[2]
http://people.unisa.edu.au/Hamid.Laga
Philippe-Henri Gosselin[1,3]
http://perso-etis.ensea.fr/~gosselin/

[1] ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051, France
[2] Phenomics and Bioinformatics Research Centre, University of South Australia, Australia
[3] INRIA Rennes Bretagne Atlantique, France

Elastic shape analysis on non-linear Riemannian manifolds provides an efficient and elegant way for simultaneous comparison and registration of non-rigid shapes. In such formulation, shapes become points on some high dimensional shape space. A geodesic between two points corresponds to the optimal deformation needed to register one shape onto another. The length of the geodesic provides a proper metric for shape comparison [2, 3, 4]. Joshi et al. [2] and Srivastava et al. [7] proposed the Square Root Velocity Function parametrization (SRVF) that allow to compute geodesic distances between closed curves $\mathbb{R}^n$. The distance is invariant to different shape preserving geometric transformation including translation, rotation and re-parametrization. The computation of geodesics, and therefore the metric, is computationally very expensive as it involves a search over the space of all possible rotations and re-parameterizations. This problem is even more important in shape retrieval scenarios where the query shape is compared to every element in the collection to search.

In this paper, we propose a fast approximation of this metric, using Kernel functions, while keeping its nice properties such as the invariance to geometric transformations. We search a subspace, equipped with the standard dot product, and in which the discriminative power of the original distance between two curves is retained. The key idea is to design a kernel function $k(\cdot,\cdot)$ associated with the elastic metric, and build a mapping function such that the dot product between mapped elements is as close as possible to the original kernel $k(x,y)$. This mapping is based on the Nyström method for kernel approximation [1]. The advantage of this formulation is that the heavy-computational metric becomes now a dot product in the new subspace with very low dimensions. This reduces significantly the computation time needed to compare one shape to all the elements of the collection to search. We further show that the approximated distance preserves the invariance properties and achieves retrieval performance that is competitive with the original metric.

Given the elastic distance between two curves $d(\beta_1,\beta_2)$, we consider a kernel function $k(\beta_1,\beta_2)$ associated with it. The approximation we propose finds a mapping function P to some specific space such that the dot product between mapped elements $P(\beta_1)^\top P(\beta_2)$ is as close as possible to the original kernel $k(\beta_1,\beta_2)$. We first consider the expression of the triangular kernel [5], $k(\beta_1,\beta_2)$ associated to $d(\beta_1,\beta_2)$ as follows:

$$k(\beta_1,\beta_2) = 1 - \frac{d(\beta_1,\beta_2)^2}{2} \tag{1}$$

With this expression of k, we derive the projection into a lower dimensional space that preserves most of the metric proprieties.

Let us consider a training set $\mathcal{A} = \{\beta_i\}$ used for the training of the projection $P(\cdot)$. We want P such that $\forall \beta_i,\beta_j \in \mathcal{A}$, $P(\beta_i)^\top P(\beta_j) = k(\beta_i,\beta_j)$. The Gram matrix of the kernel k on $\mathcal{A}$ is given by $K = [k(\beta_i,\beta_j)]_{\beta_i,\beta_j\in\mathcal{A}}$. We propose to approximate the matrix K by a low rank version and to compute the corresponding projection. This is known as the Nyström approximation for kernels, and has been used to speed up large scale kernel based classifiers in machine learning [1, 8]. Since k is a Mercer kernel and thus positive semi-definite, we can compute its eigen-decomposition: $K = V\Lambda V^\top$. The non-linear projection $P(\cdot)$ is then given by:

$$P(\beta) = \Lambda^{-\frac{1}{2}} V^\top [k(\beta_i,\beta)]_{\beta_i\in\mathcal{A}} \tag{2}$$

Where $[k(\beta_i,\beta)]_{\beta_i\in\mathcal{A}}$ is the vector of entry-wise computation of the kernel between β and the elements of the training set. Let us consider the matrix Y of the projected elements of $\mathcal{A}$: $Y = [P(\beta)]_{\beta\in\mathcal{A}}$. The Gram matrix in the projection space using the linear kernel is thus:

$$Y^\top Y = (\Lambda^{-\frac{1}{2}} V^\top K)^\top \Lambda^{-\frac{1}{2}} V^\top K = KV\Lambda^{-1}V^\top K = KK^{-1}K = K \tag{3}$$

Furthermore, the Euclidean distance between mapped elements perfectly fits to the original distance on the training set, $\forall \beta_i,\beta_j \in \mathcal{A}$:

$$\begin{aligned}
\|P(\beta_i) - P(\beta_j)\|^2 &= P(\beta_i)^\top P(\beta_i) + P(\beta_i)^\top P(\beta_i) - 2P(\beta_i)^\top P(\beta_j) \\
&= k(\beta_i,\beta_i) + k(\beta_j,\beta_j) - 2k(\beta_i,\beta_j) \\
&= 1 - \frac{d(\beta_i,\beta_i)^2}{2} + 1 - \frac{d(\beta_j,\beta_j)^2}{2} - 2 + 2\frac{d(\beta_i,\beta_j)^2}{2} \\
&= d(\beta_i,\beta_j)^2.
\end{aligned}$$

The more $\mathcal{A}$ is a good sampling of the original space, the closer the linearization P along with the dot product are to the original kernel k and thus the better the distance $d(\cdot,\cdot)$ is approximated. Some of the eigenvalues of the Kernel matrix are often very small compared to the other ones and can be safely discarded to further gain in computational efficiency. In this case, the reconstruction of the Gram matrix (respectively the distance matrix) on the training set is not perfect. However, the associated projectors often encode noise in the data, and discarding them can act as a de-noising process, as a PCA would do with linear projections. Since we are using a non-linear mapping, this process is analog to Kernel PCA [6].

We illustrate the effectiveness of the proposed method using three different experiments: shape retrieval from two databases and 3D face recognition application. The experiments show that approximated distances proposed in this paper achieves a very similar performance to the original elastic metric. The most important advantage is that the distance computing is reduced to a simple dot product on some projection space. This drastically reduces the running time of the computing process, since the non-linear computations have a complexity increasing linearly with the size of the collection (instead of quadratic with the original distance). Several applications can benefit from this work. This includes 3D face recognition, 3D biometrics and 3D shape retrieval and matching.

[1] Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

[2] S.H. Joshi, E. Klassen, A. Srivastava, and I. Jermyn. Removing shape-preserving transformations in square-root elastic (sre) framework for shape analysis of curves. In *EMMCVPR*, pages 387–398, 2007.

[3] Y. Laurent. Computable elastic distance between shapes. *SIAM Journal of Applied Mathematics*, 58(2):565–586, 1998.

[4] P. W. Michor and D. Mumford. Riemannian geometries on spaces of plane curves. *J. Eur. Math. Soc.*, 8:1–48, 2006.

[5] B Schölkopf. The kernel trick for distances. TR MSR 2000-51, Microsoft Research, Redmond, WA, 2000. Advances in Neural Information Processing Systems, 2001.

[6] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Advances in kernek methods - Support vector learning*, pages 327–352. MIT Press, 1999.

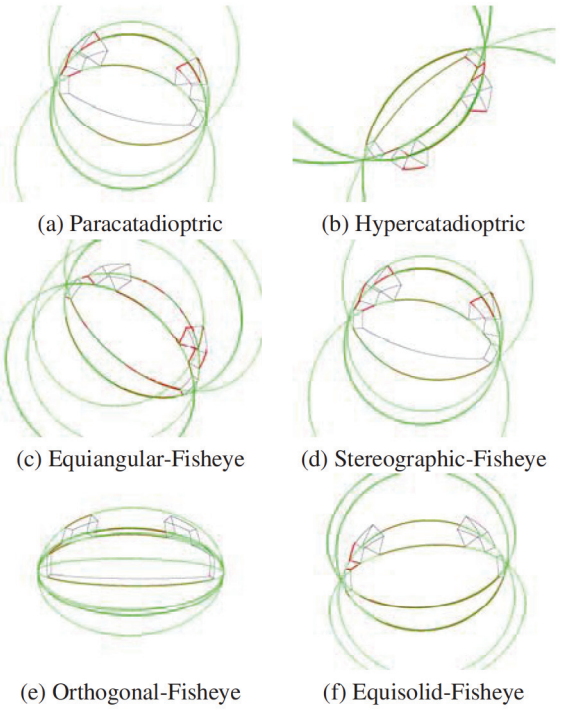
[7] Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7), 2011. ISSN 0162-8828.

[8] Chris Williams and Matthias Seeger. Using the nystroem method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

Line extraction in uncalibrated central images with revolution symmetry

Jesus Bermudez-Cameo
bermudez@unizar.es
Gonzalo Lopez-Nicolas
gonlopez@unizar.es
Jose J. Guerrero
josechu.guerrero@unizar.es

Instituto de Investigación en Ingeniería de Aragón (I3A)
Departamento de Informática e Ingeniería de Sistemas
Universidad de Zaragoza
C/ Maria de Luna 1, 50018, Zaragoza, Spain



(a) Paracatadioptric (b) Hypercatadioptric

(c) Equiangular-Fisheye (d) Stereographic-Fisheye

(e) Orthogonal-Fisheye (f) Equisolid-Fisheye

Figure 1: Extraction example on simulated images.

In omnidirectional cameras, straight lines are projected onto curves called line-images. The shape of these curves is strongly dependent of the particular camera configuration. The great diversity of omnidirectional camera systems makes harder the line-image extraction in a general way. Therefore, it is difficult to design uncalibrated general approaches, and existing methods to extract lines in omnidirectional images require the camera calibration. In this paper, we present a novel method to extract line-images in uncalibrated images which is valid for radially symmetric central systems.

In central systems the projection surface of a 3D line is a plane which has two degrees of freedom (DOF). That means we only can recover two of the four degrees of freedom of the line from a single projection. If a line-image is a curve which has more than two degrees of freedom, the extra DOFs implicitly encodes the projection model and the line-image can be used to compute the calibration [2, 3].

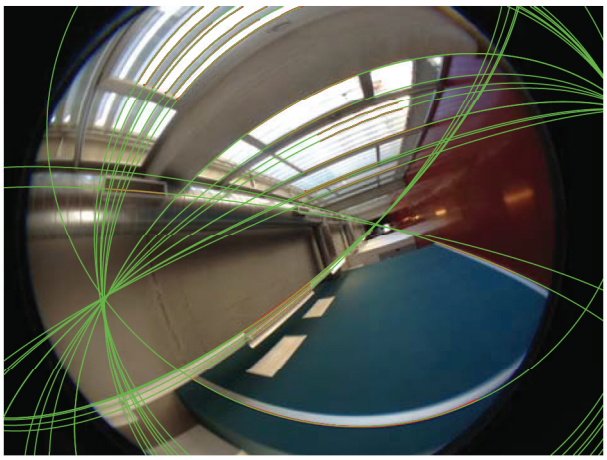


Figure 2: Extraction example on real equiangular-fisheye image

Perspective	$(l_1+l_2+l_3)=0$
Para catadioptric	$\hat{r}_{vl}=\sqrt{\frac{l_1\hat{r}_1^2+l_2\hat{r}_2^2+l_3\hat{r}_3^2}{l_1+l_2+l_3)}}$
Hyper catadioptric	$\hat{r}_{vl}=\sqrt{\left(\frac{l_1\sqrt{\hat{r}_1^2+f^2}+l_2\sqrt{\hat{r}_2^2+f^2}+l_3\sqrt{\hat{r}_3^2+f^2}}{l_1+l_2+l_3}\right)^2-f^2}$
Equiangular fisheye	$l_1\hat{r}_1\cot\left(\frac{\pi}{2}\frac{\hat{r}_1}{\hat{r}_{vl}}\right)+l_2\hat{r}_2\cot\left(\frac{\pi}{2}\frac{\hat{r}_2}{\hat{r}_{vl}}\right)+l_3\hat{r}_3\cot\left(\frac{\pi}{2}\frac{\hat{r}_3}{\hat{r}_{vl}}\right)=0$
Stereographic fisheye	$\hat{r}_{vl}=\sqrt{\frac{l_1\hat{r}_1^2+l_2\hat{r}_2^2+l_3\hat{r}_3^2}{l_1+l_2+l_3}}$
Orthogonal fisheye	$l_1\sqrt{\hat{r}_{vl}^2-\hat{r}_1^2}+l_2\sqrt{\hat{r}_{vl}^2-\hat{r}_2^2}+l_3\sqrt{\hat{r}_{vl}^2-\hat{r}_3^2}=0$
Equisolid fisheye	$l_1\frac{\hat{r}_1^2-\hat{r}_{vl}^2}{\sqrt{2\hat{r}_{vl}^2-\hat{r}_1^2}}+l_2\frac{\hat{r}_2^2-\hat{r}_{vl}^2}{\sqrt{2\hat{r}_{vl}^2-\hat{r}_2^2}}+l_3\frac{\hat{r}_3^2-\hat{r}_{vl}^2}{\sqrt{2\hat{r}_{vl}^2-\hat{r}_3^2}}=0$

Table 1: Three points line-image constraint for different central projection systems with revolution symmetry.

In our proposal we exploit the line-image description

$$n_x\hat{x}\pm n_y\hat{y}-n_z\hat{\alpha}(\hat{r})=0\,, \tag{1}$$

presented in [1], to develop the unified line-image constraint which is described as

$$l_1\hat{\alpha}_1+l_2\hat{\alpha}_2+l_3\hat{\alpha}_3=0. \tag{2}$$

The unified main calibration parameter $\hat{r}_{vl}$ encodes calibration information in different types of camera systems, dioptric and catadioptric. The line-image constraint (2) is analytically solved for each camera system (Table 1) allowing us to extract simultaneously the projection plane of the line and the main calibration parameter from three points of the line-image. Using in addition gradient-based information we recover the uncalibrated line-image from a minimum of two image points,

$$-\nabla I_y n_x\pm\nabla I_x n_y+n_z\frac{\partial\hat{\alpha}}{\partial\hat{r}}\frac{1}{\hat{r}}\left(\hat{x}\nabla I_y-\hat{y}\nabla I_x\right)=0, \tag{3}$$

by linearly approximating the distortion function α around the main calibration parameter $\hat{r}_{vl}$

$$\hat{\alpha}(\hat{r})=\frac{\partial\hat{\alpha}(\hat{r}_{vl})}{\partial\hat{r}}(\hat{r}-\hat{r}_{vl})\,. \tag{4}$$

This scheme is used in a line-image extraction algorithm to obtain lines from uncalibrated omnidirectional images without any assumption about the scene.

The algorithm is evaluated with synthetic images for a collection of different catadioptric and dioptric systems (Fig. 1). It has been also tested with real images and sequences for hyper-catadrioptic and equiangular-fisheye systems (Fig. 2) showing good performance.

[1] J. Bermudez-Cameo, G. Lopez-Nicolas, and J. J. Guerrero. A unified framework for line extraction in dioptric and catadioptric cameras. In *ACCV, Daejeon, Korea*, 2012.

[2] F. Devernay and O. Faugeras. Straight lines have to be straight. *Machine Vision and Applications*, 13(1):14–24, 2001.

[3] J.P. Tardif, P. Sturm, and S. Roy. Self-calibration of a general radially symmetric distortion model. *ECCV 2006*, pages 186–199, 2006.

Tree-like Shapes Distance Using the Elastic Shape Analysis Framework

Alejandro Mottini[1]
alejandro.mottini_d_oliveira@inria.fr
Xavier Descombes[1]
xavier.descombes@inria.fr
Florence Besse[2]
Florence.Besse@unice.fr

[1] INRIA CRI-SAM
2004 route des Lucioles, 06902 Sophia Antipolis Cedex, France
[2] IBV
Faculté des Sciences, Parc Valrose, 06108 Nice Cedex 2, France

The analysis and comparison of tree-like shapes is of great importance since many structures in nature can be described by them. In the field of biomedical imaging, trees have been used to describe structures such as neurons, blood vessels and lung airways. Since it is known that their morphology provides information on their functioning and allows the characterization of pathological states, it is of paramount importance to develop methods to analyze their shape and to quantify differences in structures. At present, there are a number of methods for the comparison of trees. Some only take into account the topology of the trees (such as [1], [2]), while others consider both topological and geometrical information (such as [3], [4]).

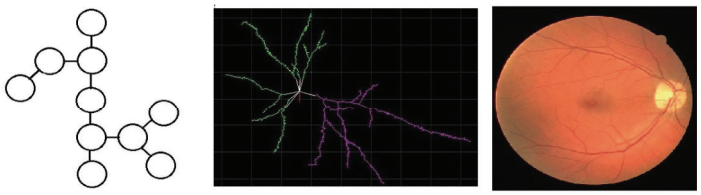


Figure 1: Diagram of a rooted tree (left) and examples of tree-like structure found in nature: neuron (middle) and blood vessels (right). Taken from http://neuromorpho.org/ and http://www.isi.uu.nl/Research/Databases/DRIVE/

In this paper, we present a new method for comparing tree-like shapes that takes into account both topological and geometrical information. It is based on the Elastic Shape Analysis Framework, a framework originally designed for comparing shapes of 3D closed curves in Euclidean spaces [5]. The idea behind this approach is the following. A 3D open curve is parameterized in a special way called SRVF (squared root velocity function). The set of all the SRVF parameterized curves in $\mathbb{R}^3$ defines a manifold called the preshape space. In order to only compare the shape of the curves, all shape preserving transformations are removed from the representation (translation, rotation, scale and parametrization). This forms a quotient space of the original manifold called the shape space. This framework was later extended to include additional information in the shape analysis. This is achieved by including an additional function $\beta_c : [0,1] \to \mathbb{R}^k$ (where k is an arbitrary dimension) to the original representation of the curve. The two components are combined to form a new curve in $\mathbb{R}^{3+k}$. Finally, one can define geodesics in the resulting space that can later be used to define distances and to compute averages of sets of curves.

In this paper we extend the Elastic Shape framework to define a metric between trees. We consider two trees T_1 and T_2, each consisting of a main curve and several branches (and possibly sub branches). All are represented by 3D open curves in $\mathbb{R}^3$.

A matching function that assigns the branches of one tree to the other is defined, along with a branch function C which indicates, for a given point on curve β, how many branches remain after it. We only take into account branches which have a match in the other tree. Finally, we define the distance between two trees T_1, T_2 as:

$$D(T_1,T_2) = \min_M [d((\beta_1(t),C1(t,M)),(\beta_2(t),C2(t,M))) + \\ + \sum_{(i,j)} \alpha_{i,j} M(i,j) D(T_1(i),T_2(j))] \quad (1)$$

where β_k is the main curve of tree k, C_k its branch function, M the matching function, $\alpha_{i,j}$ a weight parameter and $D(T_1(i),T_2(j))$ the distance between the matched branches (sub trees) of the two trees.

The first term of the distance definition embeds some geometrical information on the main branch by matching segments between branching points. Topological information is taken into account through the C function. Sub branches are taken into account in a recursive way in the second term. Note that the coefficients weight the different order of sub branches with respect to the main branch. Moreover, $\alpha_{i,j}$ controls how non matched branches ($M(i,0) = 1$ or $M(0,j) = 1$) are penalized. Typically, $\alpha_{i,j}$ equals a_1 if $i \neq 0$, $j \neq 0$ and it equals a_2 otherwise.

The formulation for the distance between the two main curves, two matched branches and unmatched branches is described in the paper.

As a first application, we used our method for the comparison of axon morphology. We have labeled single neurons within intact adult Drosophila fly brains and have acquired 3D fluorescent confocal microscopy images of their axonal trees. Both normal neurons and neurons in which a gene related to neurological diseases in humans was inactivated have been imaged (20 normal and 24 mutant). Each 3D image contains a single axonal tree that was manually segmented by an experienced biologist (see Figure 2) and the result converted to 3D open curves in $\mathbb{R}^3$.

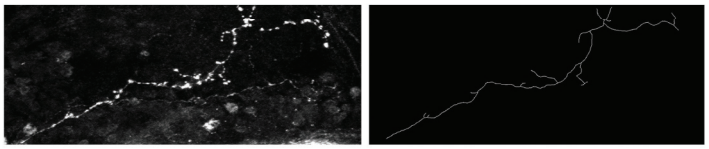


Figure 2: Original confocal microscopy image of a normal axonal tree (left) and its manual tracing (right) (maximum intensity projections).

For the validation of our method we have compared our results with the ones obtained by RTED [6], an efficient solution for the tree edit distance problem. We have calculated inter and intra class distances between normal and mutant populations and implemented a basic classification scheme. Results showed that the proposed method better distinguishes between the two populations than a pure topological metric.

Furthermore, our method allows us to compute mean shapes between two axonal trees by taking the middle point along the geodesic between the two trees (Figure 3). This is something that TED and other methods are unable to do. This will allow us to derive more complex classification scheme such as the K-means algorithm.

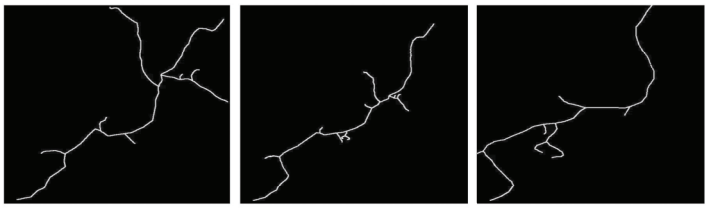


Figure 3: Two axonal trees (left and right) and their mean shape (middle) (2D projections).

[1] H. Heumann and G. Wittum. The tree-edit-distance, a measure for quantifying neuronal morphology. *Neuroinformatics*, 7(3):179–190, 2009.

[2] T. A. Gillette, K. M. Brown, and G. A. Ascoli. The diadem metric: comparing multiple reconstructions of the same neuron. *Neuroinformatics*, 9(2,3):233–245, 2011.

[3] A. Feragen, F. Lauze, P. Lo, M. de Bruijne, and M. Nielsen. Geometries on spaces of treelike shapes. In *Proc. 10th Asian conference on Computer vision, Volume Part II, 160-173*, 2010.

[4] S. Basu, B. Condron, and S. T. Acton. Path2path: Hierarchical path-based analysis for neuron matching. In *Proc. IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011.

[5] S. H Joshi, E. Klassen, A. Srivastava, and I. Jermyn. An efficient representation for computing geodesics between n-dimensional elastic shapes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[6] M. Pawlik and N. Augsten. Rted: A robust algorithm for the tree edit distance. In *PVLDB 5(4):334-345*, 2011.

Image Blur Classification and Parameter Identification using Two-stage Deep Belief Networks

Ruomei Yan
r.yan@sheffield.ac.uk
Ling Shao
ling.shao@sheffield.ac.uk

Department of Electronic and Electrical Engineering,
The University of Sheffield,
Sheffield, UK

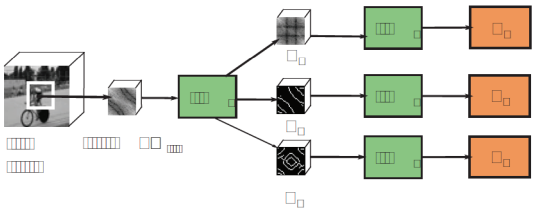


Figure 1: The TDBN architecture: DBN_1 is the first stage for blur type classification, which has 3 output labels. DBN_2 is the blur PSF parameter identification, which has different output labels for each blur type. P_1, P_2, and P_3 are the estimated parameter labels.

Image blur kernel classification and parameter estimation are critical for blind image deblurring. Current dominant approaches use handcrafted blur features [5, 6] that are optimized for a certain type of blur, which is not applicable in real blind deconvolution where the Point Spread Function (PSF) of the blur is unknown. Inspired by the successful applications of deep learning techniques to object recognition and image processing [2, 4], in this paper, a Two-stage system using Deep Belief Networks (TDBN) is proposed to first classify the blur type and then identify its parameters.

In this paper, we intend to design a patch-based blur type classification and parameters identification method to better solve the realistic blur analysis problem. Deep Belief Network (DBN) [3] is chosen for accomplishing the feature extraction and final classification in this system. A two-stage framework is proposed: first, for the input image patches with different blurs, the DBN is used for identifying the blur type; second, different samples with the same blur type will be sent to the corresponding DBN blocks for further parameter estimation. The DBN is trained in a semi-supervised way: the unsupervised training of the DBN is done by a greedy layer-wise pre-training before the supervised backpropagation for the fine-tuning. The unsupervised process helps the feature learning, and the backpropagation helps to construct the discriminative information.

In a word, our contributions are threefold: 1) To the best of our knowledge, this is the first time that deep belief network has been applied to the problem of blur analysis; 2) A discriminative feature, derived from edge extraction on Fourier Transform coefficients, has been proposed to preprocess blurred images before they are fed into the DBN; 3) A two-stage framework is proposed to estimate the blur type and parameters for any given image degraded by spatially invariant blur of an unknown type.

To begin with, for the input blurred patches, the logarithmic spectra are used as a type of feature for the identification of the blur pattern, which is shown in the following Fig. 2. Three types of blur have been considered in our paper, which are the Gaussian blur, the motion blur and the defocus blur. They can be formulated as:

$$q(\mathbf{x},\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x_1^2+x_2^2}{2\sigma^2}), \quad \mathbf{x}\in R \tag{1}$$

where σ is the blur radius to be estimated, and R is the region of support. R is usually set as $[-3\sigma,3\sigma]$, because it contains 99.7% of the energy in a Gaussian function [1].

$$q(\mathbf{x}) = \begin{cases} \frac{1}{M}, & if(x_1,x_2)\begin{pmatrix} sin(\omega) \\ cos(\omega) \end{pmatrix} = 0 \quad and \quad x_1^2 + x_2^2 \le M^2/4 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where M describes the length of motion in pixels and ω is the motion direction with its angle to the x axis.

$$q(\mathbf{x}) = \begin{cases} \frac{1}{\pi R^2}, & \sqrt{x_1^2+x_2^2} \le R \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where the blur radius R is proportional to the extent of defocusing.

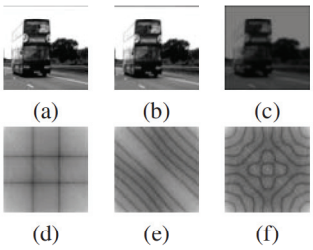


Figure 2: The blur images and their logarithmic spectra. (a) Image with Gaussian blur. (b) Image with motion blur. (c) Image with out-of-focus blur. (d) Logarithmic spectrum of Gaussian blur ($\sigma = 2$). (e) Logarithmic spectrum of motion blur ($M = 9, \omega = 45$). (f) Logarithmic spectrum of out-of-focus blur. ($R = 30$)

In the second stage, an edge detector is applied to the motion blur and defocus blur. Gaussian blur is not affected by this step because its Fourier transform of the PSF barely changes with the blur radius. The processed vectors serve as the input for the individual DBN in the next level.

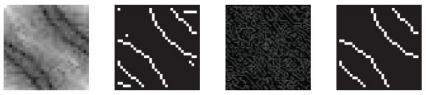


Figure 3: Comparison of the three edge detection methods applied to a training sample. From left to right: (1) the logarithmic power spectrum of a patch; (2) the edge detected by Canny detector (automatic thresholds); (3) the edge detected by the improved Canny detector using scale multiplication; (4) the edge detected by our method.

The final stage is the parameter estimation. For Gaussian blur, we estimate the blur radius of the Gaussian distribution. For motion blur, the length and angle of the motion are estimated. For out-of-focus blur, the blur radius is identified.

Our proposed method has been compared with existing blur classification algorithms based on both handcrafted features and neural networks. It outperforms state-of-the-art approaches significantly (up to 4%) due to the advantages of our selected features and the proposed TDBN structure.

[1] F. Chen and J. Ma. An empirical identification method of gaussian blur parameter for image deblurring. *IEEE Transactions on Signal Processing*, 57(7):2467–2478, 2009.

[2] A. Ciancio, A. Costa, E. Silva, A. Said, R. Samadani, and P. Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, 2011.

[3] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[4] V. Jain and H. Seung. Natural image denoising with convolutional networks. In *Proc. NIPS*, pages 769–776, 2008.

[5] R. Liu, Z. Li, and Jia J. Image partial blur detection and classification. In *Proc. CVPR*, pages 23–28, 2008.

[6] B. Su, S. Lu, and C. Tan. Blurred image region detection and classification. In *Proc. ACM Multimedia*, pages 1397–1400, 2011.

Discriminative tree-based feature mapping

Miroslav Kobetski
http://www.csc.kth.se/~kobetski/
Josephine Sullivan
http://www.csc.kth.se/~sullivan/

Computer Vision Group, CVAP
KTH, Royal Institute of Technology
Teknikringen 14,
114 28 Stockholm

We approach the classification and detection problem by viewing each step in the classification pipeline as a feature mapping. Good discriminative features $x(I)$ (such as HOG or LBP) can be seen as functions that map from pixel intensities I to a feature space where classes are more easily separable. Mixture components and parts [1] can be thought of as an additional feature mapping layer $\Phi(x,z)$ where the mapping function is also parametrised by component and part positions z. Also a non-linear kernel-based classifier can be seen as a linear classifier in an implicitly mapped space $\Phi(x)$. Finite explicit approximations to a number of kernels have been proposed by other works [3, 4, 5].

Rather that trying to approximate a kernel, we present explicit approximate mappings of tree-based classifiers learnt specifically for a particular problem and feature set. Tree-based classifiers such as boosted ensembles or random forests are powerful and fast, but are not easy to integrate with other well-developed methods that rely on explicit feature representations. In this paper we propose a more direct approach for finding a low-dimensional feature mapping $\Phi(x)$ to a space where the object and background classes are more easily separated by a linear hyperplane; having a well-generalizing underlying classifier is important for this separability to translate to unseen examples. Rather than finding a kernel with these properties and approximating it, we directly learn a non-linear decision boundary using boosted decision trees. We then induce the new feature dimensions from the decision rules of the trees. We present two of our methods for inducing $\Phi(x)$ from trees. We show that they increase discriminative performance compared to our baseline, at a small cost of evaluating the $\Phi(x)$ function. Linear SVMs in the tree-mapped space even outperform the original tree-based classifiers.

One main reason for inducing a feature mapping rather than using the trees directly is that many existing methods and frameworks (many generative models, clustering methods, detection frameworks) require an explicit feature representation. Hard negative mining is another example of a very important technique that is well defined for linear SVMs [1], but more of an unstable art-form for tree ensembles.

Each feature dimension $z_j = \phi_j(x)$ of our mapped feature vector $z = \Phi(x)$ encodes the decision path S_j taken to one leaf node of the tree. Starting at the root, the path to a leaf node can be seen as a sequence of decisions based on a number of feature values x_i and thresholds α_i, where x_i is the i:th dimension of feature vector x. We define

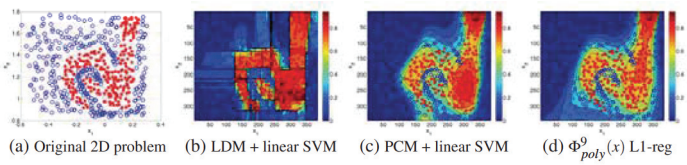
$$S_j(x) = \prod_{k=1}^{n} \delta_{jk}(x_{j_k}, \alpha_{j_k}), \tag{1}$$

where $\delta_{jk} \in \{\delta^+, \delta^-\}$ encode binary decisions $\delta^+(z_1,z_2) = [z_1 \geq z_2]$ and $\delta^-(z_1,z_2) = [z_1 < z_2]$, described using Iverson bracket notation. Such rules are inherently binary, so we enrich our feature mappings by also considering the distance to the decision boundary or rule margin $m_j(x)$. Introducing the concept of rule margin adds some ambiguity to the mapping - since for a rule j, $m_j(x)$ is the accumulated score of a number of decisions and can be envisioned and implemented in a number of ways. We present two different ones - *Leaf node decision mapping* (LDM), that most closely resembles the space partitioning of the underlying trees and *Polynomial cross-term mapping* (PCM) that can be related to sparse polynomial cross terms.

LDM is the most straight-forward addition of $m_j(x)$ to the otherwise discrete tree mapping. For a leaf node of depth n, the mapping simply becomes

$$\phi_j(x) = S_j(x)(x_{j_n} - \alpha_{j_n}). \tag{2}$$

The non-linearity is encoded through the binary decisions made following the path $S_j(x)$. An example receives value zero if it fails any of the binary decisions. If all decisions are true the example gets a value signifying its signed distance to the decision threshold made just prior to the leaf node. Each added dimension can be seen as a partition of space where similar examples are ordered according to feature x_{j_n} and dissimilar examples (as they have failed earlier similarity tests) are ignored. Due to this partitioning property LDM closely resembles the decision rules from which it is induced.



(a) Original 2D problem (b) LDM + linear SVM (c) PCM + linear SVM (d) $\Phi^9_{poly}(x)$ L1-reg

Figure 1: 2D toy problem where the 2D points have been mapped using tree-based feature mapping, after which a linear SVM was learned and projected back to 2D. $\Phi^d_{poly}(x)$ is the feature map of a d-degree polynomial kernel where each dimension is a polynomial term. The color values represent the scores of the linear SVM.

Mapping	mAP	improvement
Linear	64.85	0.00
χ^2-mapped	69.10	4.25
Polynomial kernel	70.09	5.24
Boosted Trees	72.68	7.83
LDM	72.76	7.91
PCM	72.88	8.02
CTM	74.01	9.16

Table 1: VOC2007 patch classification performance for a single HOG template and linear SVM with various feature mappings. The proposed mappings LDM, PCM and CTM outperform the other methods, including the boosted tree classifier they were induced from.

PCM accumulates the signed distance to each of the splits along $S_j(x)$

$$\phi_j(x) = \prod_{k=1}^{n} (2\delta_{j_k} - 1)(x_{j_k} - \alpha_{j_k}). \tag{3}$$

This can be seen as selecting a subset of all cross terms of a polynomial kernel, where the depth of the tree determines the degree of the polynomial. This mapping is also much sparser than a polynomial kernel since only the cross terms of features that jointly produce discriminative decisions for the particular class are selected.

We perform image patch classification experiments on the VOC2007 and INRIAPerson datasets and image classification on VOC2007. Table 1 shows that both PCM and LDM mappings improve patch classification results significantly and that the best performance is obtained by combining the tree mappings (CTM). For image classification the kernel-approximating mappings (χ^2 and Hellinger) that we used are theoretically well suited, which is reflected in the results - χ^2 improves mAP by 5.13% where while the combined CTM mapping reaches 4.68% improvement.

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), September 2010.

[2] J.H. Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28, 2000.

[3] Subhrasnu Maji. Linearized smooth additive classifiers. In *eccv, Workshop on Web-scale Vision and Social Media*, 2012.

[4] Subhrasnu Maji, Alexander C. Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[5] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3), 2011.

Unsupervised (parameter) learning for MRFs on bipartite graphs

Boris Flach
http://cmp.felk.cvut.cz/~flachbor
Tomas Sixta
http://cmp.felk.cvut.cz/~sixtatom

Center for Machine Perception
Czech Technical University
Prague, Czech Republic

This paper considers unsupervised (parameter) learning for general MRFs on bipartite graphs. That means that we assume training samples which consist of i.i.d. realisations of the field of visible variables only. The variables indexed by the vertices of the other graph part are assumed to be latent. The corresponding learning task is non-trivial because the (log) likelihood is a non-concave function of the model parameters, and, what is worse, its gradient is not tractable. A common approach is to calculate an approximation of the gradient by applying a stochastic gradient method known as "Persistent Contrastive Divergence" [5]. Another option, discussed in [3], is to marginalise over the latent variables (what can be done up to the unknown partition sum) and to maximise the pseudo-likelihood for the resulting higher order model directly. Notice however, that the resulting objective function is then no longer concave.

The main contribution of this paper is to introduce an alternative learning approach – a modified EM-algorithm with pseudo-likelihood estimator in the M-step, which is tractable on account of the bipartiteness of the model graph. In principle, such a modified EM-algorithm can be applied for parameter learning of arbitrary MRFs [6]. The resulting algorithm will however remain to be intractable, because so is the computation of the posterior pairwise marginal probabilities in the E-step. It is the bipartiteness of the graph, which ensures that the E-step *and* the M-step of the EM-algorithm are both tractable, if the maximum likelihood estimator in the M-step is replaced by the pseudo-likelihood estimator.

MRFs on bipartite graphs can be described as follows. Let (V,E) be an undirected bipartite graph and V_1, V_2 denote its parts. Let X be a collection of K_1-valued random variables indexed by vertices of V_1. That is, $X = \{X_i \mid i \in V_1\}$, where each X_i is a K_1-valued random variable. Similarly, Y denotes a collection of K_2-valued random variables indexed by vertices of the second part V_2. Both co-domains K_1 and K_2 are assumed finite. We denote realisations of the random field (X,Y) by (x,y), *i.e.* $x\colon V_1 \to K_1$ and $y\colon V_2 \to K_2$.

The joint p.d. of an MRF on (V,E) can be written as an exponential family (assuming strictly positive probability mass)

$$p_u(x,y) = \frac{1}{Z(u)} \exp \sum_{ij\in E} \langle \boldsymbol{\varphi}(x_i,y_j), \boldsymbol{u}_{ij} \rangle, \tag{1}$$

where $\boldsymbol{\varphi}\colon K_1 \times K_2 \to \mathbb{R}^{|K_1||K_2|}$ designates the vector valued indicator mapping $\varphi_{lm}(k,k') = \delta_{lk}\delta_{mk'}$ and $u = \{\boldsymbol{u}_{ij} \mid ij \in E\}$ denotes the set of model parameters. This model class includes Restricted Boltzmann Machines [2] which are often used in the context of deep learning [1].

We assume from here on that the variables X_i, $i \in V_1$ are visible, whereas the variables Y_j, $j \in V_2$ are latent and consider the task of parameter estimation given an i.i.d. sample $\mathcal{T}_\ell$ of ℓ realisations of the field X. It is assumed that the realisations were generated by $p_u(x) = \sum_{y\in\mathcal{Y}} p_u(x,y)$ with unknown u. If the maximum likelihood estimator is used, the task is

$$\frac{1}{\ell} \sum_{x\in\mathcal{T}_\ell} \log \sum_{y\in\mathcal{Y}} p_u(x,y) \to \max_u, \tag{2}$$

where $\mathcal{Y}$ denotes the set of all possible realisations of the field Y. A common method to maximize the likelihood (2) in the presence of incomplete data is the EM-algorithm.

E-step: Calculate posterior probabilities

$$\beta^{(t)}(y \mid x) := p_{u^{(t)}}(y \mid x) \tag{3}$$

for each realisation $x \in \mathcal{T}_\ell$ using the current parameter estimate $u^{(t)}$. This task is feasible for the considered model class, because the conditional p.d. $p_u(y \mid x)$ factorises.

M-step: Given the current $\beta^{(t)}$ maximise the log-likelihood for complete information

$$L_c(u) = \frac{1}{\ell} \sum_{x\in\mathcal{T}_\ell} \sum_{y\in\mathcal{Y}} \beta(y \mid x) \log p_u(x,y) \to \max_u. \tag{4}$$

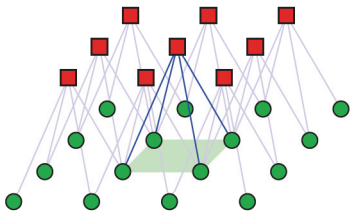


Figure 1: MRF on a translational invariant bipartite graph. Visible variables depicted as green circles, latent variables depicted as red squares. Edges and receptive field are highlighted for one of the latent variables.

This objective function is concave in u but computing the components of the gradient requires to compute pairwise marginal statistics of the model $p_u(x,y)$ and is therefore not tractable.

Following the interpretation given by one of the authors of the EM-algorithm [4], the task to be solved in the M-step is itself a (parameter) learning task, now in presence of complete data. The model parameters u must be estimated given the "observed" distribution $p^*(x,y) = \beta(y \mid x)p^*(x)$, where $p^*(x)$ is the empirical distribution associated with the sample $\mathcal{T}_\ell$. This implies that i.i.d. samples from p^* can be easily generated. The key idea is therefore to replace the maximum likelihood estimator in the M-step by any consistent *and* tractable estimator. A reasonable choice is the pseudo-likelihood estimator.

Let us denote by $\mathcal{T}^*$ an i.i.d. sample of realisations (x,y) generated from $p^*(x,y)$. The pseudo-likelihood estimator for MRFs on bipartite graphs reads

$$L_p(u) = \sum_{(x,y)\in\mathcal{T}^*} \Big[\log p_u(y \mid x) + \log p_u(x \mid y)\Big] \to \max_u. \tag{5}$$

This objective function is concave and has a tractable gradient. Optimization task in the M-step then reads to maximise the pseudo-likelihood (5) e.g. by using a gradient ascend algorithm.

In the rest of the paper we discuss implementation details of the proposed algorithm and provide experimental comparison with the PCD. We show, that despite the same per iteration time complexity our algorithm converges faster (by an order of magnitude) and more stable. On the other hand, we have no proof that the sequence of likelihood values $L(u^{(t)})$ is increasing. This should be true in the limit of an infinite training sample because the pseudo-likelihood estimator is known to be consistent.

[1] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[2] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[3] Benjamin Marlin, Kevin Swersky, Bo Chen, and Nando de Freitas. Inductive principles for restricted Boltzmann machine learning. *Journal of Machine Learning Research*, 9:509–516, 2010.

[4] Michail I. Schlesinger. The interaction of learning and self-organization in pattern recognition. *Cybernetics and Systems Analysis*, 4(2):66–71, 1968.

[5] T. Tieleman. Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM New York, NY, USA, 2008.

[6] Muneki Yasuda, Junya Tannai, and Kazuyuki Tanaka. Learning algorithm for boltzmann machines using max-product algorithm and pseudo-likelihood. *Interdisciplinary Information Sciences*, 18(1):55–63, 2012.

Generic Object Crowd Tracking by Multi-Task Learning

Wenhan Luo
http://www.iis.ee.ic.ac.uk/~whluo
Tae-Kyun Kim
http://www.iis.ee.ic.ac.uk/~tkkim

Department of Electrical and Electronic Engineering,
Imperial College, London, UK

Multiple Object Tracking (MOT) is an important problem for its various applications. In general, approaches for MOT can be categorised into two types, sequential ones and batch ones. Sequential ones utilise observations from frames up to the time while batch ones use observations from all frames of a video. For the significant progresses achieved in the pedestrian detection field, most existing work for MOT follows the batch fashion (i.e. it employs a pedestrian detector to carry out detection in each frame in advance, and then handles MOT as a data association problem by treating the detection responses of all the frames as observations). However, little attention has been paid to detection and tracking of multiple objects of an arbitrary type. In this paper, we tackle the problem of tracking multiple objects without limitation of the type of the objects. Furthermore, we show how this problem can be formulated within the Multiple Task Learning (MTL) framework [3].

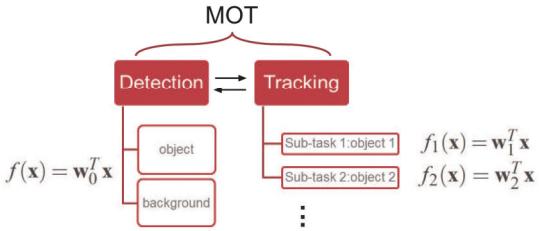


Figure 1: Task decomposition of our MOT problem.

By adopting the strategy of tracking-by-detection, we decompose our problem into two main tasks, detection and tracking, and further decompose the tracking task into multiple sub-tasks, each of which corresponds to tracking an individual object. Fig. 1 shows the decomposition of our task. In the stage of fulfilling the detection task, a binary detector is learnt to detect objects in images. For the tracking task, multiple trackers are learnt on top of the detector to trace detected objects in subsequent frames. To prevent trackers from drifting away from targets, the detector is utilised to anchor the trackers by the proposed Mean Regularised Joint Feature Learning algorithm. At the same time, the trackers are jointly learnt by sharing common features to capture the relatedness among multiple tasks. To further improve the performance, we use a smoothness term which globally considers all the labelled and unlabelled data. In the following, details of detection and tracking will be given.

The detection task is completed by posing it as a linear Laplacian SVM [2] optimisation problem. We firstly reject some sliding windows which are impossible to be objects by the objectness measurement [1] to accelerate the whole procedure. Then we construct a graph treating all the samples $\mathbf{X}$ as vertices and similarities among objects as edges. Let us write the detector as $f(\mathbf{x}) = \mathbf{w}_0^T\mathbf{x}$, then we have the following object function to minimise,

$$\begin{aligned} \min_{\mathbf{w}_0,\varepsilon_i} \quad & \gamma_1\|\mathbf{w}_0\|^2 + \gamma_2\mathbf{w}_0^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}_0 + \gamma_3\sum_{i=1}^{n_l}\varepsilon_i \\ s.t. \quad & y_i\mathbf{w}_0^T\mathbf{x} \geq 1-\varepsilon_i,\ i=1,2,...,n_l \\ & \varepsilon_i \geq 0,\ i=1,2,...,n_l \end{aligned} \tag{1}$$

where $\mathbf{L}$ is the Laplacian matrix calculated from the graph and $\{\varepsilon_i,\ i=1,2,...,n_l\}$ are slack variables. Following the primal-dual formulation, we derive a quadratic optimisation problem as,

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^{n_l}} \quad & \sum_{i=1}^{n_l}\alpha_i - \frac{1}{2}\alpha^T\mathbf{Q}\alpha \\ s.t. \quad & 0 \leq \alpha_i \leq \gamma_3,\ i=1,2,...,n_l \end{aligned} \tag{2}$$

where $\mathbf{Q} = \mathbf{Y}^T\mathbf{J}^T\mathbf{X}^T(2\gamma_1\mathbf{I}+2\gamma_2\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{J}\mathbf{Y}$, $\mathbf{J} = [\mathbf{I}\ \mathbf{0}]^T$ is a $n\times n_l$ matrix with $\mathbf{I}$ as the $n_l\times n_l$ identity matrix, $\mathbf{Y} = diag(y_1,...,y_{n_l}) \in \mathbb{R}^{n_l\times n_l}$ and $\alpha = [\alpha_1,...,\alpha_{n_l}]^T \in \mathbb{R}^{n_l}$ are Lagrangian multipliers. Solving this problem we can obtain α, and the detector is $\mathbf{w}_0 = (2\gamma_1\mathbf{I}+2\gamma_2\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{J}\mathbf{Y}\alpha$.

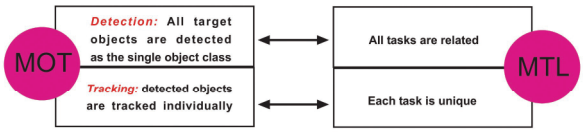


Figure 2: Formulation of the MOT problem into MTL.

To track multiple objects, we formulate the problem within MTL framework. The motivation for it can be illustrated by Fig 2. Writing each tracker as $f_t(\mathbf{x}) = \mathbf{w}_t^T\mathbf{x}$ and all the trackers $[\mathbf{w}_1,\cdots,\mathbf{w}_T]$ as a matrix $\mathbf{W}\in\mathbb{R}^{d\times T}$, we propose the Mean Regularised Joint Feature Learning algorithm which minimises the following objective function with regard to $\mathbf{W}$,

$$\begin{aligned} \min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{J}_t^T\mathbf{X}_t^T\mathbf{w}_t - \mathbf{Y}_t\|^2 + \rho_1\|\mathbf{W}\|_{2,1} + \frac{\rho_2}{2}\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}_0\|^2 \\ + \frac{\rho_3}{2}\sum_{t=1}^{T}\mathbf{w}_t^T\mathbf{X}_t\mathbf{L}_t\mathbf{X}_t^T\mathbf{w}_t \end{aligned} \tag{3}$$

In the above function, the first term is the loss from labelled data of each task. In this term $\mathbf{X}_t$ is the combination of labelled samples and unlabelled samples for a sub-task t, $\mathbf{J}_t$ is a matrix which chooses only the labelled data to calculate the loss. $\mathbf{Y}_t$ is the label vector of the task t (we give the neutral label 0 to the unlabelled data). The second term is the joint feature learning term. We learn the features shared by multiple tasks via the penalty of $\|\mathbf{W}\|_{2,1}$, the $\ell_{2,1}$ norm of $\mathbf{W}$. This regularisation term can result in that only some rows of $\mathbf{W}$ are non-zero, which correspond to the features shared by all sub-tasks. The third term is the regularisation to relate the two main tasks. This regularisation term benefits the trackers in two aspects. Firstly, as $\|\mathbf{w}_t\|^2 = \|\mathbf{w}_t-\mathbf{w}_0+\mathbf{w}_0\|^2 \leq \|\mathbf{w}_t-\mathbf{w}_0\|^2+\|\mathbf{w}_0\|^2$, and we have minimised $\|\mathbf{w}_0\|^2$ in the detection stage, thus minimising $\|\mathbf{w}_t-\mathbf{w}_0\|^2$ equals minimising $\|\mathbf{w}_t\|^2$, further improving the generalisation ability of each tracker. Secondly, this term can prevent trackers from drifting to the background as we enforce each tracker to be close to the detector. The last term is the combination of smoothness for each tracker/task. $\mathbf{L}_t$ is the Laplacian matrix associated with the graph of the task t. This smoothness term enables the tracker to view the labelled and unlabelled samples (candidates) together. This composite optimisation problem can be solved by adopting the Accelerated Gradient Method (AGM) [4].

We test our approach on four challenging data sets (including a publicly available data set) to evaluate its performance. Results compared with our self baselines and some other counterparts show that the proposed method significantly outperforms the state-of-the-art methods.

Our conclusion is that by decomposing our problem into two main tasks and representing their relation via the proposed Mean Regularised Joint Feature Learning algorithm, we can effectively derive the desired list of detected and tracked objects in frames.

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[3] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[4] Y. Nesterov. Gradient methods for minimizing composite objective function. core discussion papers 2007076, université catholique de louvain. *Center for Operations Research and Econometrics (CORE)*, 2007.

Discriminative Generative Contour Detection

Chao Zhang[1]
bjzchao@cn.ibm.com
Xiong Li[2]
lixiong@sjtu.edu.cn
Xiang Ruan[3]
gen@omm.ncl.omron.co.jp
Yuming Zhao[2]
arola_zym@sjtu.edu.cn
Ming-Hsuan Yang[4]
mhyang@ucmerced.edu

[1] IBM China Research, China
[2] Shanghai Jiao Tong University, China
[3] Omron Corporation, Japan
[4] University of California Merced, United States

Object contour is of prime importance as it contains essential visual information, such as shape and identity that finds numerous applications.Contour detection is a fundamental problem in computer vision which is closely related to other tasks, e.g., segmentation, shape analysis, pose estimation, visual saliency, and object recognition, to name a few.

In this work, we propose a learning algorithm for contour detection based on multi-level visual cues. We extract pixel-level features that integrate both local and global visual information. Basic point (pixel) features have been widely used for representing edge information of grayscale and color images such as image gradients, texture inhibition, brightness and color gradients as well as compass operators.As these features capture different visual information, we extract and combine these features as local visual cues for contour detection at different scales. Global information extracted from visual saliency is then incorporated to provide cues of salient objects in the scenes. These features are integrated to form effective pixel-level features to represent object contours.

In addition, segment-level features are extracted to exploit structural information of contours. Structural cues such as segments contain important information more than pixelwise evidence.Toward this goal, we compute superpixels to extract structural segments with the SLIC algorithm [1], which has been verified to perform well terms of efficiency and effectiveness. Point features are used to describe edge pixels on the line fragment, and segment-level features are then extracted by computing their mean value, variance and differences in this work.

To extract discriminative information from features, we propose a mapping method based on the posterior divergence (PD) measure over Gaussian mixture model (GMM). We transform the vectors formed by point-level and segment-level features based this mapping to obtain more discriminative information for contour detection. Let $\mathbf{x} \in \mathrm{R}^D$ be the observed random variable, i.e., the combination of multi-scale features. Let $\mathbf{z} = \{z_1, \cdots, z_K\}$ be the hidden variable, where $z_k = 1$ if the k-th mixture center is selected to generate a sample and $z_k = 0$ otherwise; $\mathbf{a} = (a_1, \cdots, a_K)^\top$ be the mixture prior satisfying; $\mathbf{u}_k$ and Σ_k respectively be the mean and variance matrix of the k-th mixture center. For any observed sample $\mathbf{x}^t$, similar to [3], we assume that the posterior distribution of $\mathbf{z}$ takes the same from with its prior $P(\mathbf{z})$ but with different parameter $\mathbf{g}^t = (g_1^t, \cdots, g_K^t)^\top$, $Q^t(\mathbf{z}) = \prod_{k=1}^K {g_k^t}^{z_k}$. Let $\boldsymbol{\theta}$ be the model estimated from a set of $N-1$ training samples $\mathcal{X} = \{\mathbf{x}^i\}_{i=1}^{N-1}$, and $\boldsymbol{\theta}_{+t}$ be the model estimated from a set of N samples $\mathcal{X} \cup \{\mathbf{x}^t\}$. PD derives three measures, where the posterior divergence Φ^{pd} measures how much $\mathbf{x}$ affects the model, the fitness function Φ^{fit} measures how well $\mathbf{x}$ fits the model, the entropy function Φ^{ent} measures how uncertain the fitting is. The mappings according to the variable $\mathbf{x}$ can be derived as follows,

$$\Phi_x^{pd} = \sum_{i\neq t}^{N}\sum_{k=1}^{K}\sum_{d=1}^{D} g_k^i\left(-\frac{(x_d^t-u_{d,+t})^2}{2\delta_{d,+t}^2}+\frac{(x_d^t-u_d)^2}{2\delta_d^2}+\delta_{d,+t}^{D/2}-\delta_d^{D/2}\right) = \sum_{d=1}^{D}\Phi_{x_d}^{pd},$$

$$\Phi_x^{fit} = \sum_{k,d=1}^{K,D} g_k^t\left(-\frac{(x_d^t-u_{d,+t})^2}{2\delta_{d,+t}^2}+\delta_{d,+t}^{D/2}\log\sqrt{2\pi}\right) = \sum_{d=1}^{D}\Phi_{x_d}^{fit},$$

The feature mappings according to the hidden variable $\mathbf{z}$ are,

$$\Phi_z^{pd} = \sum_{i\neq t}^{N}\sum_{k=1}^{K} g_k^i \log\frac{a_{k,+t}}{a_k} = \sum_{k=1}^{K}\Phi_{z_k}^{pd},$$

$$\Phi_z^{fit} = \sum_{k=1}^{K} g_k^t \log a_{k,+t} = \sum_{k=1}^{K}\Phi_{z_k}^{fit}, \quad \Phi_z^{ent} = \sum_{k=1}^{K} g_k^t \log g_k^t = \sum_{k=1}^{K}\Phi_{z_k}^{ent}.$$

Therefore for input $\mathbf{x}^t$, we obtain a set of feature mapping:

$$\Phi^t = \mathrm{vec}(\{\Phi_{x_d}^{pd}, \Phi_{x_d}^{fit}, \Phi_{z_k}^{pd}, \Phi_{z_k}^{fit}, \Phi_{z_k}^{ent}\}_{d,k}). \tag{1}$$

With the features mapping, various classifiers within the random forest [2] learning framework can be constructed based on different representations. A random forest is an ensemble classifier consisting of numerous decision trees where the class label is determined based on the mode of the outputs by individual trees.

We evaluate the proposed algorithm on the Berkeley segmentation data set. Figure 1 shows the precision-recall curves and the F-measures with different thresholds.

Figure 2 shows one contour detection results using features with and without using the proposed feature extraction method via posterior divergence.

Experimental results bear out feature selection from multi-scale visual cues via posterior divergence with a random forest classifier facilitates effective contour detection in natural images.

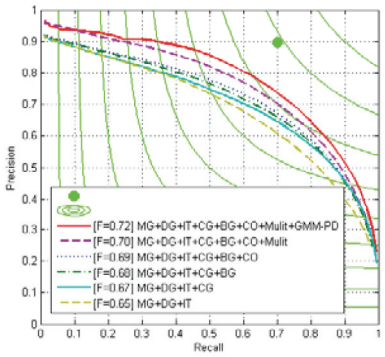


Figure 1: Precision-recall curves on BSDS500: different feature combinations.

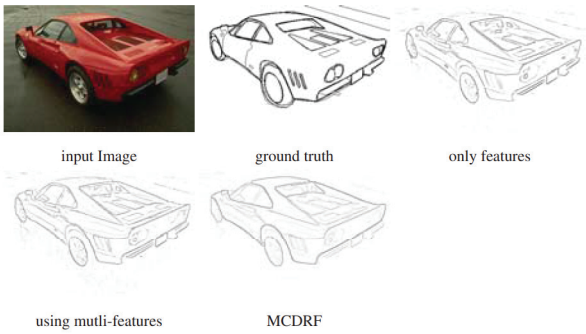


Figure 2: Contour detection results with and without feature extraction.

[1] Radhakrishna Achanta, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine SÍźsstrunk. SLIC Superpixels. Technical Report 149300, EPFL, 2010.

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

[3] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Sail. An introduction to variational methods for graphical models. In *Machine Learning*, volume 37, pages 183–233, 1999.

Accurate and Computationally-inexpensive Recovery of Ego-Motion using Optical Flow and Range Flow with Extended Temporal Support

Graeme A. Jones
dircweb.kingston.ac.uk/graeme/

Digital Imaging Research Centre
School of Computing and Information Systems, Kingston University,
Penrhyn Road, Kingston upon Thames, UK, KT1 2EE

Introduction

Recovering the ego-motion of a moving camera within a static scene supports many applications in robotics and computer vision. The presented work is motivated by *pre-vis* applications in the film industry; specifically the ability to render digital assets into the scene during production in real-time. A low-cost commodity depth camera can be easily mounted on and calibrated to a high quality production cameras and used to extract changes in sensor pose from the induced motion of the rigid scene. This work explores the effectiveness of the computationally efficient *range flow* technique to generate this real time pose information directly from the depth stream of a Kinect sensor.

A number of challenges within the approach are addressed. First, an iterative version of the *small rotations* motion estimator is developed to ensure the most accurate inter-frame estimates. Second, the substantial issue of *drift* is addressed - the accumulated error between true and estimated sensor pose as motion estimates are temporally integrated. *Anchor frames* which enjoy significant overlap with subsequent frames are stored and used to provide additional temporal range flow constraint within the estimation process. Where there are loops in the data sequence, it is advantageous to select anchors from previously seen parts of the scene. Finally, in some scene configurations, there is insufficient constraint from the depth images. We exploit the availability of registered intensity images to further constrain the sensor motion using the *optical flow* framework.

Analogous to optical flow, *range flow* is a per-pixel constraint on the 3D displacement of an imaged 3D point given its local spatio-temporal depth derivatives. These must be combined across a region or an image to provide sufficient constraint to extract 3D motion[1, 2].

The Optical Flow and Range Flow Constraints

The *constant brightness equation* relates how the luminance at a 3D scene point is captured in temporally separated intensity images. A 3D point $\mathbf{X}$ is imaged at pixel position $\mathbf{x} = (x,y)^T$ in the intensity map I_t. This point undergoes a 3D motion $\Delta\mathbf{X} = (\Delta X, \Delta Y, \Delta Z)^T$ which results in an image motion $\Delta\mathbf{x}$ between frames t and τ and reprojects with the same intensity at the new image location $\mathbf{x}+\Delta\mathbf{x}$. Thus the *optical flow* constraint is defined as

$$I_\tau(\mathbf{x}+\Delta\mathbf{x}) = I_t(\mathbf{x}) \tag{1}$$

Analogous to the *constant brightness equation*, the following range constraint relates how a 3D point is captured in temporally separated depth images. A 3D point $\mathbf{X}$ (measured in the depth camera's coordinate system) is captured at pixel position $\mathbf{x} = (x,y)^T$ in the depth map Z_t. This point undergoes a 3D motion $\Delta\mathbf{X}$ which results, first, in an image motion $\Delta\mathbf{x}$ between frames t and τ, and second, in a change of the depth ΔZ of the 3D point captured at this new image location $\mathbf{x}+\Delta\mathbf{x}$. Thus the *range flow* constraint is formulated as

$$Z_\tau(\mathbf{x}+\Delta\mathbf{x}) = Z_t(\mathbf{x}) + \Delta Z \tag{2}$$

Motion Model

Scene displacement is induced by the movement of the camera whose rotation ω and translation $\mathbf{t}$ motion parameters can be modelled using the useful linear *small rotation* formulation when the movement between consecutive frames is relatively small. Here the 3D displacement $\Delta\mathbf{X} = (\Delta X, \Delta Y, \Delta Z)^T$ of a 3D scene point $\mathbf{X} = (X,Y,Z)^T$ is given by

$$\Delta\mathbf{X} = \omega \times \mathbf{X} + \mathbf{t} = \begin{bmatrix} Z\omega_y - Y\omega_z + t_x \\ X\omega_z - Z\omega_x + t_y \\ Y\omega_x - X\omega_y + t_z \end{bmatrix} = M(\mathbf{X})\mathbf{a} \tag{3}$$

where $\mathbf{a}$ is a concatenation of the motion parameters $\mathbf{a} = (\omega, \mathbf{t})^T$.

Minimising drift using Anchor Frames

Simply integrating between-frame motion estimates over time will inevitably result in *drift i.e.* the accumulated error between true and estimated sensor pose. To minimise this, additional temporal constraint can be included. Specifically we introduce the concept of an *anchor frame*. In addition to recovering a parameter update for the motion between the current frame and the previous frame, the same update is also constrained by the motion between the current frame and its anchor. A depth frame is an anchor to all subsequent frames with which it retains a significant degree of overlap. Once the amount of overlap falls below a threshold, the last frame is promoted as the next anchor. Updates to the motion are now computed from two sources of range flow constraint.

When there are loops in the sequence, it would be advantageous to select anchors from previously seen data rather than using the last frame. Such constraint from early frames makes a significant impact on the degree of drift. To this end, a list of all anchor frames is maintained. When a new anchor is required, this list is searched for the earliest anchor overlapping the current frame. However, a consequence of this approach, is the linear growth in storage requirements for these anchors and in the computational cost of searching through these anchors as the length of the video sequence grows.

Evaluation

The recently published TUM RGB-D Benchmark [3] is used to evaluate the motion estimator. This resource provides Kinect depth and registered RGB sequences with synchronized ground truth of the sensor pose for extensive set of sequences of varying lengths and differing levels of challenge such as large visual velocities or large dominant planar surfaces. Our study uses nine Freiburg1 (FR1) sequences in which the Kinect sensor is moved within a typical indoor environment. The resource also provides an evaluation tool that computes the root mean square error (RSME) between an estimated trajectory and the associated ground truth once these have been aligned. Specifically we use the translation and rotation RMSE measures, T-RMSE and R-RMSE respectively, and add the maximum value of the *absolute trajectory error* (MATE) which identifies the maximum sensor positional error anywhere along the estimated trajectory.

Conclusion

Pose accuracy can reasonably be judged as commensurate with SLAM approaches (as represented by the RGB-D SLAM system) but available at a fraction of the computational cost. Indeed a real-time implementation is running on an old Samsung P460 Notebook. Given the low computational cost of generating reasonably accurate pose estimates, the presented approach could usefully bootstrap more computationally expensive techniques.

[1] John L. Barron and Hagen Spies. "The Fusion of Image and Range Flow"). In *Proceedings of the 10th International Workshop on Theoretical Foundations of Computer Vision: Multi-Image Analysis*, pages 171–189, London, UK, UK, 2001.

[2] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill. "3D Pose Tracking with Linear Depth and Brightness Constraints". In *International Conference on Computer Vision*, pages 206–213, 1999.

[3] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. "A Benchmark for the Evaluation of RGB-D SLAM Systems". In *Int. Conf. on Intelligent Robot Systems*, October 2012.

Emotion recognition from facial images with arbitrary views

Xiaohua Huang
huang.xiaohua@ee.oulu.fi
Guoying Zhao
gyzhao@ee.oulu.fi
Matti Pietikäinen
mkp@ee.oulu.fi

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

Facial expression recognition has been predominantly utilized to analyze the emotional status of human beings. In practice nearly frontal-view facial images may not be available. Therefore, a desirable property of facial expression recognition would allow the user to have any head pose. We show in our paper a new method to recognize arbitrary-view facial expressions by using discriminative neighborhood preserving embedding and multi-view concepts.

Given n training images with C classes, denoted as $\mathbf{X} = [\vec{x}_1, \ldots, \vec{x}_n] \in R^{\mathrm{D}\times \mathrm{n}}$, based on class labels, we construct the intra-class and inter-class sets Ω_p^{wi} and Ω_p^{bw} of $\vec{x}_p$, where $p = 1, \ldots, \mathrm{n}$. The former contains intra-class $\vec{x}_q$ ($q = 1, \ldots, \mathrm{k}^{\mathrm{wi}}$) nearest neighboring to $\vec{x}_p$, while the latter contains inter-class $\vec{x}_r$ ($\mathrm{r} = 1, \ldots, \mathrm{k}^{\mathrm{bw}}$) nearest neighboring to $\vec{x}_p$. We further design an intrinsic graph $\mathbf{G}^{\mathrm{wi}} = \{\mathbf{X}, \mathbf{V}^{\mathrm{wi}}\}$ that preserves the intrinsic structure of intra-class samples, and a penalty graph $\mathbf{G}^{\mathrm{bw}} = \{\mathbf{X}, \mathbf{V}^{\mathrm{bw}}\}$ that describes the margin across inter-class boundaries.

(1) To preserve the similarity of intra-class samples, the similarity matrix $\mathbf{V}^{\mathrm{wi}}$ for the intrinsic graph $\mathbf{G}^{\mathrm{wi}}$ can be obtained by minimizing the following formulation,

$$\varepsilon^{\mathrm{wi}}(\mathbf{X}) = \sum_p \| \vec{x}_p - \sum_{\vec{x}_q \in \Omega_p^{\mathrm{wi}}} v_{p,q}^{\mathrm{wi}} \vec{x}_q \|^2, \tag{1}$$

with the constraint $\sum_q v_{p,q}^{\mathrm{wi}} = 1$, where $v_{p,q}^{\mathrm{wi}}$ in $\mathbf{V}^{\mathrm{wi}}$ is the weight of the edge from $\vec{x}_p$ to $\vec{x}_q$.

Given $\mathbf{U} \in R^{\mathrm{D}\times \mathrm{d}}$, the sample $\vec{x}_p$ is transformed to this space via $\vec{y}_p = \mathbf{U}^T \vec{x}_p \in R^{\mathrm{d}\times 1}$, where d is the dimensionality of this space. Therefore, the sample $\vec{y}_p$ can be represented as a linear combination of its neighbors with the corresponding coefficients v_{pq}^{wi} as follows,

$$\varepsilon^{\mathrm{wi}}(\mathbf{Y}) = \sum_p \| \vec{y}_p - \sum_{\vec{y}_q \in \Omega_p^{\mathrm{wi}}} v_{p,q}^{\mathrm{wi}} \vec{y}_q \|^2 = \mathbf{U}^T \mathbf{S}^{\mathrm{wi}} \mathbf{U}, \tag{2}$$

where $\mathbf{S}^{\mathrm{wi}} = \mathbf{X}(\mathbf{I} - \mathbf{V}^{\mathrm{wi}})(\mathbf{I} - \mathbf{V}^{\mathrm{wi}})^T \mathbf{X}^T$ represents the local geometric structure of intra-class samples.

(2) The penalty similarity matrix $\mathbf{V}^{\mathrm{bw}}$ can be obtained by minimizing the following formulation,

$$\varepsilon^{bw}(\mathbf{X}) = \sum_p \| \vec{x}_p - \sum_{\vec{x}_r \in \Omega_p^{\mathrm{bw}}} v_{p,r}^{\mathrm{bw}} \vec{x}_r \|^2, \tag{3}$$

with the constraint $\sum_r v_{p,r}^{\mathrm{bw}} = 1$, where $v_{p,r}^{\mathrm{bw}}$ in $\mathbf{V}^{\mathrm{bw}}$ represents the weight of the edge from $\vec{x}_p$ to $\vec{x}_r$ with different class labels.

To maximize the boundary of samples of different class labels, $\mathbf{U}$ can be obtained by maximizing the following function,

$$\varepsilon^{bw}(\mathbf{Y}) = \sum_p \| \vec{y}_p - \sum_{\vec{y}_r \in \Omega_p^{\mathrm{bw}}} v_{p,r}^{\mathrm{bw}} \vec{y}_r \|^2 = \mathbf{U}^T \mathbf{S}^{\mathrm{bw}} \mathbf{U}, \tag{4}$$

where $\mathbf{S}^{\mathrm{bw}} = \mathbf{X}(\mathbf{I} - \mathbf{V}^{\mathrm{bw}})(\mathbf{I} - \mathbf{V}^{\mathrm{bw}})^T \mathbf{X}^T$ represents the local geometric structure of inter-class samples.

Maximum margin criterion [3] was proposed to maximize the margin between classes after dimensionality reduction. Additionally, it does not suffer from the small sample size problem. $\mathbf{U}$ can be obtained accordingly by the following function,

$$\mathbf{U} = \arg\max \{\mathbf{U}^T \mathbf{S}^{\mathrm{bw}} \mathbf{U} - \eta \mathbf{U}^T \mathbf{S}^{\mathrm{wi}} \mathbf{U}\}, \tag{5}$$

with the constraint $\mathbf{U}^T \mathbf{X} \mathbf{X}^T \mathbf{U} = 1$, where η is the balancing factor that adjusts the second term to ensure a positive objective function.

In practice, multi-view facial expression recognition aims to recognize facial expressions with arbitrary views. Here, we simply suppose that there are N views for each sample. Given the samples with the ith view ($i \in \{1, \ldots, \mathrm{N}\}$), they are denoted as $\mathbf{X}_i$. From Eqn.5, we can obtain the formula $\mathbf{A}_i = (\mathbf{S}_i^{\mathrm{bw}} - \eta_i \mathbf{S}_i^{\mathrm{wi}})$ of the ith view with respect to $\mathbf{X}_i$. Thus Eqn.5 can be revised as follows,

$$[\mathbf{U}_1, \ldots, \mathbf{U}_{\mathrm{N}}] = \arg\max \sum_{i=1}^{\mathrm{N}} \mu_i \mathbf{U}_i^T \mathbf{A}_i \mathbf{U}_i, \tag{6}$$

with constraints $\sum_i \mathbf{U}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{U}_i = 1$, where the positive term μ_i is included to bring the balance between multiple objectives. Eqn.6 embeds the respective objective function of each view into one common function of all views, but still loses the correlation of samples with distinct views.

Multiset canonical correlation analysis (MCCA) [4] was proposed to search the correlation vector for multiple sets. According to MCCA and the formulation of closeness in [1], we define the objective function that maximizes the correlation between samples with the same expression label yet on all views and meanwhile minimizes the covariance of samples with the same view label, as follows,

$$[\mathbf{U}_1, \ldots, \mathbf{U}_{\mathrm{N}}] = \arg\max \sum_{i=1}^{\mathrm{N}} \sum_{j=1, j\neq i}^{\mathrm{N}} \mathbf{U}_i^T \mathbf{M}_i \mathbf{M}_j^T \mathbf{U}_j, \tag{7}$$

with the constraint $\sum_{i=1}^{\mathrm{N}} \mathbf{U}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{U}_i = 1$, where $\mathbf{M}_i$ is the class mean matrix of samples on the ith view.

Based on two objective functions (6) and (7), we can obtain the completed formulation as follows,

$$[\mathbf{U}_1, \ldots, \mathbf{U}_{\mathrm{N}}] = \arg\max \sum_{i=1}^{\mathrm{N}} \{\mu_i \mathbf{U}_i^T \mathbf{A}_i \mathbf{U}_i + \sum_{j=1, j\neq i}^{\mathrm{N}} \alpha_{i,j} \mathbf{U}_i^T \mathbf{M}_i \mathbf{M}_j^T \mathbf{U}_j\}, \tag{8}$$

with the constraint $\sum_{i=1}^{\mathrm{N}} \beta_i \mathbf{U}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{U}_i = 1$, where μ_i, $\alpha_{i,j}$ and β_i are balancing parameters in which adjust the importance of terms in the objective function and constraint item. Through this formula, we can obtain discriminative feature space of facial expression $\mathbf{U}_i$ in each view.

Emotion Classification: Our aim is to match two face images with the same or different facial expression label in different views. Partly motivated by cross-view classification [2], we design the mean-correlation maximization classifier to classify the sample $\vec{x}$ as follows,

$$h(\vec{x}) = \max_c (\mathrm{mean}_c (\max_i \{\mathrm{corr}(\mathbf{U}_i^T \mathbf{X}_{i,c}, \mathbf{U}_i^T \vec{x})\}_{i=1}^{\mathrm{N}})), \tag{9}$$

where $\mathbf{X}_{i,c}$ represents training samples of the cth facial expression label with the ith view, *corr* represents Pearson's linear correlation coefficient operator, N is the number of views, *mean* and *max* are the mean and maximum value operator, respectively.

Conclusion: Our method firstly captures the discriminative property of inter-class samples. In addition, it explores the closeness of intra-class samples with arbitrary view in a low-dimensional subspace.

[1] S. Abhishek, K. Abhishek, and D. Hal. Generalized multiview analysis: a discriminative latent space. In *Proc. CVPR*, pages 2160–2167, 2012.

[2] A. Li, S. Shan, and W. Gao. Coupled bias-variance trade off for cross pose face recognition. *IEEE Trans. Image Processing*, 21(1):305–315, 2012.

[3] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Proc. NIPS*, pages 157–165, 2003.

[4] A. Nielsen. Multiset canonical correlation analysis and multispectral, truly multitemporal remote sensing data. *IEEE Trans. Image Processing*, 11(3):293–305, 2002.

RPM: Random Points Matching for Pair-wise Face-Similarity

M. Saquib Sarfraz[1]
saquib.sarfraz@kit.edu
Muhammad Adnan Siddique[2]
siddique@ifu.baug.ethz.ch
Rainer Stiefelhagen[1]
rainer.stiefelhagen@kit.edu

[1] Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany

[2] Swiss Federal Institute of Technology (ETH) Zurich
Zurich, Switzerland

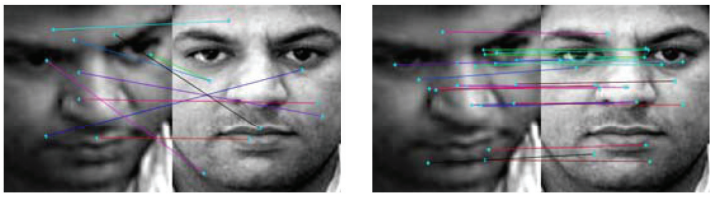
Pair-wise face verification strives to determine whether two probe images belong to the same person. With the increasing demand of face recognition/verification in real-world applications, the challenges confronted by the verification algorithms include not only dealing with large pose and lighting variations, but also very different imaging conditions e.g. different imaging device or resolution, sensor noise, varied expressions, lens distortion and occlusions.

Face verification algorithms generally require image representation followed by matching. For the representation, the current methods typically rely on aligning the two images such that the effect of pose and other geometric deformations are minimized. A feature analysis is then performed for matching, either globally on the whole face, or locally on some points typically found by a fiducial point detector. The over restrictive requirement of image alignment, especially for global feature analysis, is not trivial to meet. Besides the need to know the pose a priori, it requires 2D warps or 3D model-fitting which in turn depends on reliably detected fiducial points. Therefore, the performance of a verification algorithm is inevitably limited by the possibility of detecting reliable fiducial points.

In general, even the state-of-the-art fiducial or key point detectors do not guarantee to find all the required fiducial points. SIFT and similar detectors may fail to provide useful points for matching, especially when the probe images exhibit wide imaging differences. The points found do not guarantee to correspond to the same physical location in the two images and thus result in poor matching correspondence. Figure 1 a. depicts this problem by showing the matches found by employing the SIFT key point detector and matching engine. The image pair shown is comprised of a high resolution frontal gallery image and the probe face detected from a poor resolution surveillance video of the same individual. Due to the underlying imaging differences, the SIFT detector has failed to provide good corresponding points belonging to the same physical location on the two images. Fiducial point detectors pose similar behaviour, particularly in case of wide pose differences (e.g. when matching a frontal and profile view).

In this paper we shift from this over reliance of key point detectors and propose a novel method to automatically obtain well-matched points across the two images, subsequently referred to as the homologous points. we argue that this key/fiducial point detector may not guarantee a good correspondence in the presence of large pose and other imaging variations especially when the underling image modalities are different due to , for example, different image sensors. This is largely because these detectors try to find the specific regions on the face and thus become image specific. We on the other hand propose to find the corresponding point pair (homologous points) that falls on the same physical location from a dense set of randomly initiated points. These pairs can directly provide useful image similarity information. We use these points to establish direct pair-wise similarity, nonetheless, all of the current methods may benefit by using these newly established correspondences.

The procedure is based on randomly initializing enough points on the detected face windows and then follow up with a powerful mutual information assisted matching strategy. The approach lifts the restrictions, noted in the preceding paragraph, of the current key/fiducial point detectors as it is more probable to yield well matched pairs of points (on the same physical location on the face) from the randomly initialized points by using a meaningful iterative matching strategy. The procedure is such that it automatically finds pairs of points that not only correspond well but are also meaningful in establishing the similarity of the face in the image pair. To finally match the image pair we propose a new similarity metric that combines the useful parameters found during random point matching and the similarity scores computed by using a local descriptor on these homologous points.



(a) (b)

Figure 1: Point features matching in a pair of positive images: (a) with SIFT detector and descriptor; (b) with RPM (our approach).

Our approach is to match the candidate regions 'C' centred on each of the random points in the pair by using normalized Mutual Information (nMI).

$$m^i = \arg\max_j \text{nMI}\left(C_A^i, C_B^j\right) \tag{1}$$

where m^i is the matched point. From among the set of these matched points the one that offers the highest MI is retained.

$$\begin{gathered} i_r = \arg\max_{i_k} \text{nMI}\left(C_A^{i_k}, C_B^{m^i}\right) \\ i_r \rightarrow m^{i_r} \end{gathered} \tag{2}$$

We then use a recursive model fitting and an area based optimization to refine these estimates.

$$\Delta T_{xy}^{i_r} = \arg\max_{(\Delta T_{xy})} \text{nMI}\left(C_A^{i_r}, C_B^{m^i+\Delta T_{xy}}\right) \tag{3}$$

The matched pairs retained at the end are the homologous points. We then combine the number of these homologous points 'n', the variance of the model parameters σ (obtained during matching) and the local similarities 'L' computed using a local descriptor on these points, in a novel similarity metric referred to as Q-score.

$$Q = \exp\left(\frac{(L+1)\times n}{\sigma + n}\right) \tag{4}$$

We tested our approach on images that exhibit wide pose and other imaging differences using two challenging datasets, FacePix and LFW. Our proposal shows an improvement over state-of-the-art methods that computes a direct pair-wise similarity.

The RPM formulation, to the best of our knowledge, is the first attempt at generating reliable homologous points among a pair of face images from completely random points, obviating the need to invoke a key or fiducial point detector as traditionally required before performing a local or global pair-wise feature analysis. It lends us useful metrics for pair-wise face similarity verification, such as the proposed Q-score which performs better than some of the state-of-the-art methods. Moreover, our work helps in setting the new perspective that the non-trivial requirement of image alignment (for an appropriate representation prior to matching) can be circumvented.

We aim to extend this work in two directions. In this paper, the different metrics originating from the RPM formulation have been used to build a direct measure of similarity. We intend to investigate the potential use of these metrics in training a classifier as well. Secondly, as the RPM formulation provides well-matched homologous points even in case of wide variations in imaging conditions, it can be used for pose-estimation and image alignment itself. We expect that the RPM formulation may provide a general solution to various machine vision problems.

Segmentation Driven Low-rank Matrix Recovery for Saliency Detection

Wenbin Zou[1]
zouwb2008@gmail.com
Kidiyo Kpalma[1]
kidiyo.kpalma@insa-rennes.fr
Zhi Liu [2]
liuzhisjtu@163.com
Joseph Ronsin [1]
joseph.ronsin@insa-rennes.fr

[1] IETR / UMR CNRS 6164,
INSA Rennes,
Université Européenne de Bretagne
Rennes, France

[2] School of Communication and Information Engineering,
Shanghai University
Shanghai, P.R. China

Low-rank matrix recovery (LRMR) model, aiming at decomposing a matrix into a low-rank matrix and a sparse one, has shown the potential to address the problem of saliency detection, where the decomposed low-rank matrix naturally corresponds to the background, and the sparse one captures salient objects. This is under the assumption that the background is consistent and objects are obviously distinctive. Unfortunately, in real images, the background may be cluttered and show low contrast with objects. Thus directly applying the LRMR model to the saliency detection has limited robustness. This paper proposes a novel approach that exploits bottom-up segmentation as a guidance cue for the matrix recovery. This method is fully unsupervised, yet obtains higher performance than the supervised LRMR model.

A key and distinguishing element of this model is the use of proposed *segmentation prior* integrating to the low-rank matrix recovery. Firstly, let us take a look at the images and their coarse-grained (CG) segmentations in Figure 1. Salient objects locate at diversity of positions: center, bottom, left, right and corner. Both background and objects are typically segmented into several regions, thus, the bottom-up segmentation can not be expected to totally separate objects from the background. However, the segmented regions of background have very high probability of connecting with the border of the image, while very few regions of objects link to it. Even if an object is truncated on the border, like the bike and the child of the two right-most images, border regions of object are small compared to the whole object in the image. In contrast, the border regions of the background are usually large, as the background appears more uniform, like sky, road, tree, wall, etc. This observation implies that objects can be roughly separated from the background by the bottom-up segmentation. Therefore, we propose the segmentation prior according to the connectivity between each region and image border. Let r_m be a segmented region of image $\mathbf{I}$, the segmentation prior of region r_m is defined as

$$h_m = \exp\left(-\frac{\|r_m \cap C\|}{\sigma \psi_m}\right) \tag{1}$$

where $\|\cdot\|$ denotes the length of intersection, C is the border of image $\mathbf{I}$, ψ_m is the outer perimeter of region r_m, and σ is a balance parameter which is set to 0.3 in our experiments. Clearly, if a region touches the image border, its prior value is in the range of $(0,1)$, otherwise it is equal to 1. In other words, the segmentation prior gives a small weight to the region touching the image border. Using (1), segmentation priors of all regions can be computed, and form the prior of the input image.

In Figure 1, one might observe that, on one hand, there are still some regions of the background without connection with the image border, on the other hand, some regions of objects are inevitably merged with the background. Indeed, such a strategy can not perfectly separate the objects from the background. However, the segmentation prior can serve as a guidance cue for LRMR model to address the task of saliency detection.

Suppose an input image $\mathbf{I}$ is segmented into N superpixels, and represented by a feature matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N]$. Let $\mathbf{H}_c = [h_1^c, h_2^c, \cdots, h_N^c]$ denote a set of CG segmentation prior values of the superpixels. In order to recover well salient objects with the LRMR model, the feature matrix $\mathbf{A}$ is firstly modulated by the CG segmentation prior $\mathbf{H}_c$

$$\mathbf{B} = [h_1^c \mathbf{a}_1, h_2^c \mathbf{a}_2, \cdots, h_N^c \mathbf{a}_N]. \tag{2}$$

Then, the modulated feature matrix $\mathbf{B}$ is used as the input of the standard LRMR model

$$\begin{aligned} \min \quad & \|\mathbf{U}\|_* + \lambda \|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{B} = \mathbf{U} + \mathbf{E} \end{aligned} \tag{3}$$

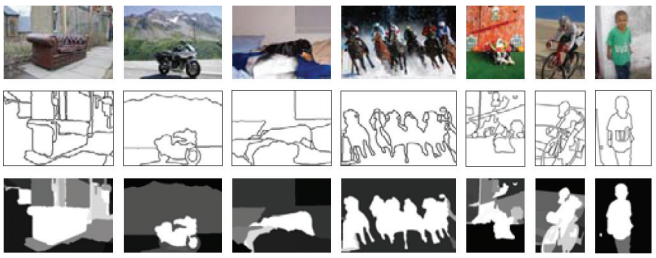


Figure 1: Examples of segmentation prior. First row: input images; second row: segmentation results; last row: segmentation prior.

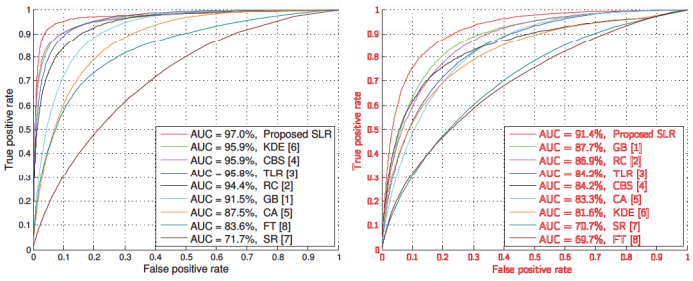


Figure 2: ROC curves and AUC scores of different models on MSRA-1000 (left) and PASCAL-1500 (right) datasets.

where λ is a coefficient to balance $\mathbf{U}$ and $\mathbf{E}$, $\|\cdot\|_*$ denotes the nuclear norm of matrix $\mathbf{U}$ (the sum of singular values of $\mathbf{U}$), and $\|\cdot\|_1$ indicates l_1-norm which ensures to produce a sparse matrix $\mathbf{E}$. With the optimal sparse matrix $\mathbf{E}$, the saliency of a superpixel is given by the l_1 energy of corresponding vectors in $\mathbf{E}$.

The success of the proposed segmentation prior modulated low-rank matrix recovery model (SLR) is mainly due to two reasons. On one hand, real images typically possess high redundancy in the feature space, and object pixels tend to be salient compared to the background, which enables to discover objects from the sparse matrix $\mathbf{E}$. On the other hand, as the segmentation prior assigns small weights to most of background feature vectors in $\mathbf{B}$, the l_1 energies of the corresponding vectors in the recovered matrix $\mathbf{E}$ are inclined to be small. Therefore, objects can be captured more effectively from the matrix $\mathbf{E}$.

The proposed SLR model is evaluated on two datasets (2500 images in total): the widely used MSRA-1000 dataset and the newly introduced but more challenging PASCAL-1500 dataset. The performance of saliency detection is measured by receiver operator characteristic (ROC) curve and the area under the curve (AUC). As shown in Figure 2, the proposed model outperforms 8 state-of-the-art models [1-8] on both datasets, with 1.1% and 3.7% improvement in terms of AUC score on MSRA-1000 and PASCAL-1500, respectively, compared to the best one among the reference models.

[1] Harel et al. Graph-based visual saliency. *Adv. Neural Inf. Process. Syst.*, 2007.
[2] Cheng et al. Global contrast based salient region detection. *CVPR*, 2011.
[3] Shen and Wu. A unified approach to salient object detection via low rank matrix recovery. *CVPR*, 2012.
[4] Jiang et al. Automatic salient object segmentation based on context and shape prior. *BMVC*, 2011.
[5] Goferman et al. Context-aware saliency detection. *CVPR*, 2010.
[6] Liu et al. Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut. *TMM*, 2012.
[7] Hou and Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007.
[8] Achanta et al. Frequency-tuned salient region detection. *CVPR*, 2009.

Action Attribute Detection from Sports Videos with Contextual Constraints

Xiaodong Yu[1]
xiaodong_yu@cable.comcast.com
Ching Lik Teo[2]
cteo@cs.umd.edu
Yezhou Yang[2]
yzyang@cs.umd.edu
Cornelia Fermüller[2]
fer@cfar.umd.edu
Yiannis Aloimonos[2]
yiannisg@cs.umd.edu

[1] Comcast Corporation
Washington DC, USA
[2] University of Maryland,
College Park, MD, USA

In this paper, we study the problem of detecting action attributes from sport videos. Action attributes include atomic components of action classes (such as the motion patterns of human limbs and body), contextual components of action classes (such as the objects and scenes involved in the action), and non-semantic attributes, a.k.a data-driven attributes [2]. A common property of action attributes is that they can be generalized into different action classes. This is especially true in sports videos. For example, *bend*, as an action attribute that describes the motion of human body, is present in the action `tennis serve`, `bowling`, `snatch`, etc. That is why they can be learned even from training sets that contain only a few examples for each action class (*one-shot* learning) or even no example for some action classes (*zero-shot* learning). We focus on the action attributes related to motion patterns of human body in this paper; however, our model can be easily extended to detect the other types of action attributes as well.

The concept of action attribute was first introduced in [2]. However, the approach proposed in [2] has several severe limitations that restrict the applicability of action attributes. One of the most noticeable problems is that action attributes are labelled at the level of an action class, instead of being labelled at the level of a frame or at the level of a video. As a result, all videos from the same action class are assumed to have the same set of action attributes, regardless of the exact content of a specific video, and the temporal structure of the action attributes is totally discarded. But in reality, not every video belonging to the same action class have the same set of action attributes; more often than not, real-world videos would have some exceptions. For example, in a video of the `snatch` activity, the athlete may not be able to completely lift the barbell above his head at the end so we cannot say this video has an action attribute *two arms raise pose*. Furthermore, the temporal structure is a unique property of action attributes and we lose lots of descriptive capacity if we ignore it. For example, given a video of `basketball layup`, a description *"the athlete starts with a slow run and lasts for half a second, then jumps forward with single leg in the next second, finally jumps up and throws the ball (into the basket), and maintains a slow run at the end of the video"* will be more useful than simply saying *"there are slow running, jumping forward, jumping up, throwing in this video"*. The goal of this paper is thus to detect the key action attributes at each frame from a given video so that we can generate video descriptions at a much finer granularity than those from previous work.

While having great advantages as discussed above, it is obviously a much more challenging task to locate the temporal occurrences of every action attributes in a given video. As a high-level semantic concept, a particular action attribute may exhibit significant variability due to viewpoint changes, photometric measurements, intra-class variability (e.g. attribute *two arms open* can have different opening angles between two arms, *jump up* can have different height and velocity, etc). Naive detectors that rely entirely on local features will easily be overwhelmed by a large number of false positives and/or false negatives. So we must take into account the contextual constraints in both the temporal and semantic domains, thereby reducing the noise in the local feature space to produce more reliable results.

We take into account two types of contextual constraints in this paper: the temporal context and the semantic context. To promote agreement between the different attribute labels at different frames, we model the contextual constraints with a conditional random field (CRF) as shown in Figure 1. The features extracted from T frames in a video are denoted as a vector of T local observations, $\mathbf{X} = \{x_1, x_2, ..., x_T\}$; and at each local observation at frame t, by $\mathbf{x}_t$, the histogram of visual words as discussed in the preceding section. At each frame t, we wish to detect the presence of A action attributes $\mathbf{y}_t = \{y_t^1, y_t^2, ..., y_t^A\}$, which are the states in the CRF model. In the literature, this is also known as a factorial CRF [4]. In a factorial CRF, we have multiple states at each time point, $\mathbf{y}_t = \{y_t^1, y_t^2, ..., y_t^A\}$, and there are edges between every pair of y_t^i and y_t^j, $(i, j) \in \{1, 2, ..., A\}$[1]. To avoid clutter, we only show two attributes at each time point in Figure 1(b). In the experimental dataset used, there are 24 attributes at each time point. The between-chain edges are designed to promote agreement between different attributes at the same time point. The intuition is that some attributes tend to occur together, e.g.*two arms oscillate* and *fast run* while others don't, e.g. *slow run* and *fast run*. Thus the between-chain edges take into account the co-temporal correlation among attributes. The within-chain edges are designed to promote agreement between the states of adjacent time point based on the Markovian assumption and we enforce agreement between states in adjacent time points after accounting for the correlation between the neighboring temporal states. In this paper, the effectiveness of our methods are clearly illustrated by the experimental evaluations.

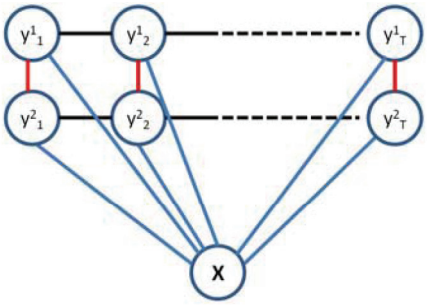


Figure 1: A factorial CRF model. To avoid clutter, we only show two attributes at each time points. In reality, we can have as many as attributes at each time points and they are fully connected to each other.

[1] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[2] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing Human Actions by Attributes. In *CVPR*, 2011.

[3] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Conditional Models for Contextual Human Motion Recognition. *CVIU*, 104:210–220, 2006.

[4] Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. *Journal of Machine Learning Research*, 8:693–723, 2007.

[1] For clarity, we call the edges between states of the same time points as *between-chain edges* and the edges between states of adjacent time points as *within-chain edges*. In Figure 1, the former are colored in red and the latter in black

Decomposed Learning for Joint Object Segmentation and Categorization

Yi-Hsuan Tsai
ytsai2@ucmerced.edu
Jimei Yang
jyang44@ucmerced.edu
Ming-Hsuan Yang
mhyang@ucmerced.edu

Electrical Engineering and Computer Science
University of California
Merced, USA

We present a learning algorithm for joint object segmentation and categorization that decomposes the original problem into two sub-tasks and admits their bidirectional interaction. In the first stage, in order to decompose the output space, we train category-specific segmentation models to generate multiple figure-ground hypotheses. In the second stage, by taking advantage of object figure-ground information, we train a multi-class segment-based categorization model to determine the object class. A re-ranking strategy is then applied to classified segments to obtain the final category-level segmentation results (see Figure 1).

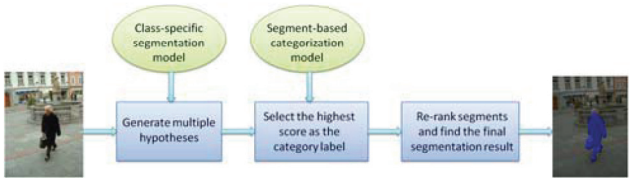


Figure 1: Overview of the algorithm.

Category-specific segmentation. The first step is to train the two-class pylon models [5] based on the segmentation tree generated by the gPb method [1]. Suppose the image I can be partitioned into hierarchical regions $\mathbf{S} = \{S_1, S_2, ..., S_{2N-1}\}$ and a figure-ground label, $f_i = 1$ or 2, respectively is assigned to each region, we formulate the conventional CRF energy function:

$$E(\mathbf{f}) = \sum_{i=1}^{2N-1} U(f_i) + \sum_{(i,j)\in\mathcal{N}} V(f_i, f_j), \tag{1}$$

where $U(f_i)$ indicates the cost of assigning f_i to the segment S_i and $V(f_i, f_j)$ is the smoothness term of the boundary cost for two neighboring segments S_i and S_j. Furthermore, we define the unary energy:

$$U(f_i) = \begin{cases} |S_i| \cdot <\mathbf{w}_1, \mathbf{h}(\mathrm{S_i})>, & \text{for } f_i = 1, \\ |S_i| \cdot <\mathbf{w}_2, \mathbf{h}(\mathrm{S_i})>, & \text{for } f_i = 2, \end{cases} \tag{2}$$

where $\mathbf{h}(S_i)$ denotes the feature vectors and the weighting factor $|S_i|$ is the size of the segment, which encourages the model to prefer larger regions.

In the testing stage, with the learned energy function for $E(\mathbf{f})$, instead of finding only the MAP solution, we introduce parametric min cut [4] into our unary function in Equation 1:

$$U(f_i, \lambda) = \begin{cases} U(f_i) + |S_i| \cdot \lambda, & \text{for } f_i = 1, \\ U(f_i) - |S_i| \cdot \lambda, & \text{for } f_i = 2. \end{cases} \tag{3}$$

Different values of λ provide our model a bias to generate parametrized results, so we can adjust the hyperplane $\mathbf{w}$ and generate multiple segmentation hypotheses by solving a series of graph cuts.

Segment-based categorization. Given an image, we can apply each category-specific model to obtain a set of segmentation hypotheses $B_i, i = 1, 2, ..., K$ for K classes. We first divide our hypothesis set into positive and negative bags, denoted by B_i^+ and B_i^- respectively. The positive bag consists of the hypotheses generated with the positive segmentation classifier w_i^+, and likewise negative bags contain examples from negative segmentation classifiers w_i^-.

For learning the categorization model, we use standard SVM model and select the best segmentation among a positive bag and the ground truth segmentation as positive samples x^+. In the meanwhile, we use all negative samples x^- from all the negative bags to reduce the chances of false positives (see Figure 2).

Given a test image, a bag of segmentation hypotheses from each segmentation model is generated as the training process. We then first determine the image category label by selecting the highest classification score.

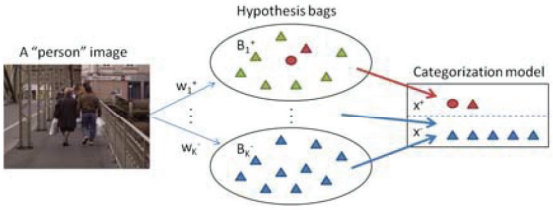


Figure 2: Training for object categorization. The red circle is the ground truth and the red triangle is the best segmentation hypothesis. All samples from B_i^- are negative samples, denoted as blue triangles.

To produce the final segmentation result, we re-rank all the hypotheses in the predicted category bag. The ranking process can be carried out by the class-wise Support Vector Regression and their scores measured by the overlapping ratio between the segment and the ground truth.

Our algorithm enjoys bidirectional interactions between segmentation and categorization. In the segmentation phase, category information facilitates breaking down the multi-class segmentation problem into class-wise sub-problems such that high-quality figure-ground separation can be generated in a reduced labeling space. In the categorization phase, segmentation information helps identifying object locations, shapes as well as context, and hence objects can be precisely represented in the feature space and improve the categorization performance. For concreteness, we demonstrate the merits of the proposed algorithm on the Graz-02 and Caltech 101 data sets.

Table 1: Graz-02 segmentation results using intersection/union overlap metric.

Method	Background	Bicycle	Car	Person	mean
[7]	82.32	46.18	36.49	38.99	50.99
[3]	77.97	55.60	41.51	37.26	53.08
Proposed	91.20	64.95	59.60	60.49	69.06

Table 2: Part of results for Caltech-101 classification. MFea denotes multiple features. Geo denotes the geometric information.

	Method	30 training
MFea + Geo	Gu et al. [2]	77.7
	SvrSegm [6]	82.3
	Proposed	84.2 ± 0.3

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011. ISSN 0162-8828.

[2] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009.

[3] A. Jain, L. Zappella, P. McClure, and R. Vidal. Visual dictionary learning for joint object categorization and segmentation. In *ECCV*, 2012.

[4] V. Kolmogorov, Y. Boykov, and C. Rother. Applications of parametric maxflow in computer vision. In *ICCV*, 2007.

[5] V. Lempitsky, A. Vedaldi, and A. Zisserman. A pylon model for semantic segmentation. In *NIPS*, 2011.

[6] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010.

[7] D. Singaraju and R. Vidal. Using global bag of features models in random fields for joint categorization and segmentation of objects. In *CVPR*, 2011.

Action Chart: A Representation for Efficient Recognition of Complex Activities

Hyung Jin Chang
hj.chang@imperial.ac.uk

Department of
Electrical and Electronic Engineering,
Imperial College London, London, UK

Jiyun Kim, Jungchan Cho, Songhwai Oh,
Kwang Moo Yi, Jin Young Choi
{jiyun07,cjc83,songhwai,kmyi,jychoi}@snu.ac.kr

Department of Electrical
and Computer Engineering, ASRI
Seoul National University, Seoul, Korea

Action recognition has been widely studied for decades and there are many successful approaches to recognize relatively simple actions [2]. Recently, more realistic and complex activity recognition tasks have been dealt with, but the current status of the research on complex activities is in its initial phase and far from the recognition ability of human. In our work, we are interested in recognizing temporally very long, complex and diverse action streams. Reliable and efficient recognition methods of this kind of complex action streams can be useful for various tasks, such as complex activity categorization, long video abstraction and similar activity-based video retrieval in YouTube.

In this paper, we propose a pipelined motion-information embedding structure from a high dimensional local feature flow to a low dimensional attentional motion spot flow in order to recognize long and complex action streams efficiently. Figure shows an overall scheme of the proposed method. Each step of the proposed method is focused on extracting distinctive motion information and filtering out noise. In order to reduce high dimensional action video sequences (>640x480x6000 frames) into simple representations while retaining the necessary information, we propose a new composite motion feature generation method by combining various conventional low-level local features. The composite features characterize local, holistic, and sequential motion changes with small memories. The 21-dimensional composite feature sequences are embedded into one-dimensional feature sequences with preserving motion characteristics by proposing a hierarchical embedding method. The one-dimensional embedded feature sequence is utilized to catch distinctive motions. The distinctive motion instances are referred to as attentional motion spots (AMSs), which are automatically determined in our scheme. The AMSs appear in similar feature space-time locations for the same activity classes. We model the distribution of AMSs as a weighted Gaussian mixture model (GMM) using expectation maximization (EM) in embedded feature space-temporal domain. The sketch of this model looks like a music chart, thus we name our representation as *Action Chart*, which is used for action class recognition.

Composite Motion Feature Flow: The composite motion features are newly defined in this subsection by manipulating the low-level local feature information. The composite motion features extracted in each frame form a temporally sequential flow through the whole video frames, referred to as composite motion feature flow (CMFF). The CMFF ($\mathcal{M} = \{\mathcal{M}(t)|t = 1,...,N\}$) is composed of holistic ($\mathcal{M}_H$) and local ($\mathcal{M}_L$) motion features. The holistic motion feature $\mathcal{M}_H$ is composed of five measurements; motion intensity (m_I), motion extent (m_E), motion speed (m_S), motion change (m_C) and motion diversity (m_D). The motion intensity, extent, and speed represent quantitative motion property, and the change and diversity reflect qualitative motion property. At each t frame, the five measurements are obtained independently. The local motion feature ($\mathcal{M}_L(t) \in R^{16}$) represents the relative location distributions of local feature points, $p_{(x,y)}$, using a concentric 16-bin histogram method whose center is placed in the center of human body. We compose the CMFF by concatenating the holistic and local features; $\mathcal{M}(t) = [\mathcal{M}_H(t), \mathcal{M}_L(t)] = [m_I(t), m_E(t), m_S(t), m_C(t), m_D(t), m_L^{[1,...,16]}(t)]^T$ ($\mathcal{M}(t) \in R^{21}$).

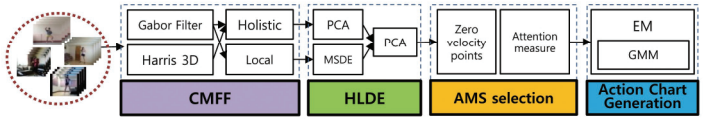


Figure 1: Overall scheme of building *Action Chart*. The proposed method is composed of four steps: (1) composite motion feature flow generation using low-level local features, (2) hierarchical embedding of the feature flow into 1-dimensional (1-D) feature sequence, (3) attentional motion spot selection in the 1-D sequence, and (4) activity modeling and recognition using the attentional motion spots.

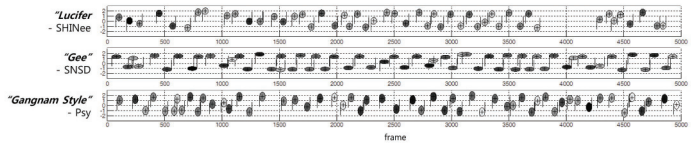


Figure 2: Generated *Action Charts* for the Pop-Dance dataset.

Hierarchical Low Dimensional Embedding: To avoid the curse of dimensionality in analyzing motion streams, it is necessary for the CMFF $\mathcal{M}$ to be embedded to a lower dimensional space. However $\mathcal{M}$ consists of two different information groups; $\mathcal{M}_H$ and $\mathcal{M}_L$. Each dimension of $\mathcal{M}_H$ implies independent and distinctive motion information, while the 16 dimensions of $\mathcal{M}_L$ represent only one motion information as a combination. In other words, the importance of each dimension is different. To handle this problem, we propose a hierarchical low dimensional embedding (HLDE) method.

AMS Selection: By mimicking the human perception mechanism, we propose a method to catch and focus on distinctive instances along the motion feature flow X. We define the distinctive instances as AMS, and we use velocity (the first derivative) of X to find the AMS, which is similar to the human mechanism of using motion changes as a clue for segmentation. The number of zero-velocity points is determined automatically. To avoid the false detection problem that other zero-velocity based methods [3] suffer from, we introduce an attention measure η at j^{th} zero-velocity point z_j, and the η is used for filtering out noisy zero-velocity points by thresholding. The number of attentional points determined automatically.

Action Chart Generation and Recognition: Each AMS set $\mathbf{Y}^c$ follows a weighted GMM distribution in the feature space-time domain. It can then be written as $p(\mathbf{Y}^c|\theta^c) = \sum_{m=1}^{k^c} \omega_m^c p(\mathbf{Y}^c|\theta_m^c)$, where $\omega_1^c, ..., \omega_{k^c}^c$ ($\omega_m^c \geq 0, m = 1, ..., k^c$, and $\sum_{m=1}^{k^c} \omega_m = 1$) are the *weights of each component*, each θ_m^c is a set of Gaussian parameters $\theta_m^c = \{\mu_m^c, \Sigma_m^c\}$ defining m^{th} component, and $\Theta^c = \{\theta_1^c, ..., \theta_{k^c}^c, \omega_1^c, ..., \omega_{k^c}^c\}$. We define the Θ^c as *Action Chart* of action class c. To estimate each Θ^c, the EM algorithm is used. We adopt the Figueiredo and Jain [1] algorithm for unsupervised parameter estimation. The class of the test action stream X^{test} is determined through maximum log-likelihood. The AMS set of X^{test} is obtained and represented as $Y^{test} = \{y_1^{test}, ..., y_{n^{test}}^{test}\}$. The class recognition is performed by matching the Y^{test} and the trained *Action Chart* Θ^c one by one.

In order to validate the proposed method, we generated a new complex action dataset; the Pop-Dance dataset. The experimental results showed that the *Action Chart* could give a promising recognition performance with a very low computational load. Furthermore it could be used for abstracting a long video sequence aims. Our method can contribute to recognizing repetitive sequential activities (e.g. workplace safety, retail fraud detection or sweethearting, and product quality assurance) and sequentially combined action tasks (e.g. sign language and cooking menu), which are our future research.

[1] Mario A.T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002.

[2] J.K.Aggarwal and M.S.Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), April 2011.

[3] Daniel Weinland. *Action Representation and Recognition*. PhD thesis, Institut National Polytechnique De Grenoble, oct 2008.

Simultaneous Pose, Focal Length and 2D-to-3D Correspondences from Noisy Observations

Adrian Penate-Sanchez, Eduard Serradell,
Juan Andrade-Cetto and Francesc Moreno-Noguer

Institut de Robòtica i Informàtica Industrial
CSIC-UPC, Barcelona, Spain.

Matching methods based on RANSAC rely on having a small outlier rate. This way, the probability of obtaining a correct minimal set of points is high enough to be achieved in fixed time. If the percentage of outliers increases, the number of RANSAC loops needed to obtain a correct set of minimal points grows exponentially. We propose a matching algorithm that performs robust matching under the presence of a large percentage of outliers. Our approach is a generalization of the work in [1] to deal with uncalibrated cameras and noisy 3D information.

We are given a reference model made up of M 3D points $\mathcal{X} = \{\mathbf{x}_i\}$, with their 2D correspondences on a reference image, and a set of N 2D points $\mathcal{U} = \{\mathbf{u}_j\}$ on an input image, acquired with an uncalibrated camera. Let us denote by $\mathbf{p}$ a 6-dimensional vector, parameterizing the pose, and let f be the unknown camera focal length. As the 3D-to-2D correspondences are unknown, they need to be retrieved together with the pose and focal length parameters.

$$\underset{\mathbf{p},f}{\text{minimize}} \sum_{i=1}^{M} \mathsf{Inlier}(\|\mathsf{Proj}(\mathbf{x}_i;\mathbf{p},f) - \mathsf{Match}(\mathbf{x}_i;\mathcal{U})\|) \qquad (1)$$

where $\mathsf{Proj}(\mathbf{x}_i;\mathbf{p},f)$ returns the 2D perspective projection $\tilde{\mathbf{u}}_i$ of a 3D point $\mathbf{x}_i$ given the pose and focal length parameters; $\mathsf{Match}(\mathbf{x}_i;\mathcal{U})$ returns the $\mathbf{u}_j \in \mathcal{U}$ that is closest to $\tilde{\mathbf{u}}_i$; and

$$\mathsf{Inlier}(d) \begin{cases} d & \text{if } d < \mathit{Max_distance_inlier} \\ \mathsf{Penalty_outlier} & \text{otherwise} \end{cases}$$

is a function that penalizes points whose reprojection error is above a *Max_distance_inlier* threshold to avoid local minima.

We build a hyper-box in the 6-dimensional pose space using the pose priors, which is then subsampled using Montecarlo. Expectation Maximization is run over these samples to compute the N_p Gaussian priors on the pose, defined by a set of mean poses $\mathbf{p}_k$, $k = 1,\ldots,N_p$, and a set of 6×6 covariance matrices $\Sigma_k^{\mathbf{p}}$. The range of feasible focal lengths, is split into N_f Gaussian priors, defined by mean values f_l and the corresponding one-dimensional variances σ_l^f, $l = 1,\ldots,N_f$.

Our approach lets us handle uncertain 3D models, such as those obtained from a Kinect camera Fig. 1. Each 3D model point $\mathbf{x}_i$ is assigned a covariance $\Sigma_i^{\mathbf{x}}$, computed considering the depth variations of their neighboring points. Those points with larger uncertainties will have less weight in the computation of the solution. We also assign an uncertainty $\Sigma_j^{\mathbf{u}}$ to each 2D measurement. Given the sets $\mathcal{X}$ and $\mathcal{U}$, the pose and focal length priors, and the 3D and 2D uncertainties, we proceed to the optimization of Eq. 1 by progressively exploring each pair of priors $\{\mathbf{p}_k,\Sigma_k^{\mathbf{p}}\};\{f_l,\Sigma_l^f\}$. To limit the number of potential 2D match candidates for each 3D point $\mathbf{x}_i$, we project them onto the image plane and compute the uncertainty in the projection assuming independent contribution from all three sources: 3D point uncertainty, pose uncertainty, and focal length uncertainty. The result is a Gaussian distribution with mean $\tilde{\mathbf{u}}_i$ and covariance $\Sigma_i^{\tilde{\mathbf{u}}}$:

$$\begin{aligned} \tilde{\mathbf{u}}_i &= \mathsf{Proj}(\mathbf{x}_i;\mathbf{p}_k,f_l) \\ \Sigma_i^{\tilde{\mathbf{u}}} &= \mathbf{J}_{\mathbf{x}}\Sigma_i^{\mathbf{x}}\mathbf{J}_{\mathbf{x}}^{\top} + \mathbf{J}_{\mathbf{p}}\Sigma_k^{\mathbf{p}}\mathbf{J}_{\mathbf{p}}^{\top} + \mathbf{J}_f\sigma_l^f\mathbf{J}_f^{\top}\,, \end{aligned} \qquad (2)$$

where $\mathbf{J}_g = \frac{\partial \mathsf{Proj}(\mathbf{x}_i;\mathbf{p}_k,f_l)}{\partial g}$ is the Jacobian of the projection function with respect to each of the uncertain parameters $g = \{\mathbf{x},\mathbf{p},f\}$. Using the Gaussian distribution $\{\tilde{\mathbf{u}}_i,\Sigma_i^{\tilde{\mathbf{u}}}\}$, we can define a search region for the point $\mathbf{x}_i$, and consider as potential candidates $\mathcal{PC}(\mathbf{x}_i)$ all points $\mathbf{u}_j \in \mathcal{U}$ whose Mahalanobis distance is below a threshold $\mathsf{Max_Mah}$, i.e:

$$\mathcal{PC}(\mathbf{x}_i) = \left\{\mathbf{u}_j \in \mathcal{U} \text{ s.t. } (\mathbf{u}_j - \tilde{\mathbf{u}}_i)^{\top}(\Sigma_i^{\tilde{\mathbf{u}}})^{-1}(\mathbf{u}_j - \tilde{\mathbf{u}}_i) < \mathsf{Max_Mah}^2\right\} \cup \{\emptyset\} \qquad (3)$$

where $\emptyset$ denotes the possibility that $\mathbf{x}_i$ is in fact an outlier and does not have a 2D image correspondence.

This work has been partially funded by Spanish Ministry of Economy and Competitiveness under project PAU+ DPI2011-27510; by the EU project ARCAS FP7-ICT-2011-28761 and by the ERA-Net CHISTERA project VISEN.

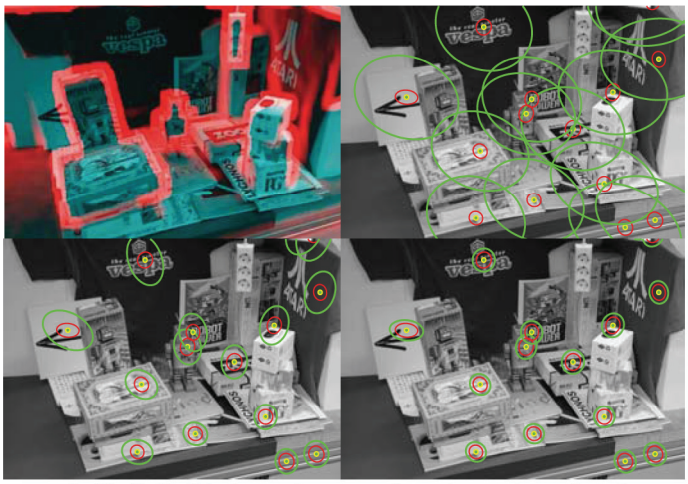


Figure 1: **Up Left:** We detect the uncertain regions –shown in red– computing depth covariances within local neighborhoods. We define a search region to limit the number of candidates **Up Right:** Search region obtained after first projection of the three terms of Eq. 2 independently. **Bottom:** Refinement of the search space, after establishing correspondences.

Once we have defined the set of potential 2D candidates for all the 3D points, we start a hypothesize and test strategy. After hypotheses have been made, we use a Kalman filter formulation to shrink the size of the Gaussian regions associated to the pose and focal length, to further reduce and guide the set of potential candidates in each iteration Fig. 1. We initialize this step choosing the least ambiguous point

$$\mathbf{x}_i^* = \underset{\mathbf{x}_i \in \mathcal{X}}{\arg\min} |\mathcal{PC}(\mathbf{x}_i)|\,, \qquad (4)$$

i.e, the 3D point with the lowest number of potential candidates. In doing so we start with a 3D point with low uncertainty, since these are the ones with smaller search regions for potential matches, in the Mahalanobis sense. We then hypothesize the match $\{\mathbf{x}_i^*,\mathbf{u}_j^*\}$, where $\mathbf{u}_j^*$ is the 2D candidate within $\mathcal{PC}(\mathbf{x}_i^*)$ that is closest to $\tilde{\mathbf{u}}_i^*$ in terms of Mahalanobis distance. We then use standard Kalman filter equations to update the pose and focal length and reduce their associated covariances:

$$\begin{aligned} \mathbf{p}_k^+ &= \mathbf{p}_k + \mathbf{K}_{\mathbf{p}}(\mathbf{u}_j^* - \tilde{\mathbf{u}}_i^*) & f_l^+ &= f_l + \mathbf{K}_f(\mathbf{u}_j^* - \tilde{\mathbf{u}}_i^*) \\ \Sigma_k^{\mathbf{p},+} &= (\mathbf{I} - \mathbf{K}_{\mathbf{p}}\mathbf{J}_{\mathbf{p}})\Sigma_k^{\mathbf{p}} & \sigma_l^{f,+} &= (1 - \mathbf{K}_f\mathbf{J}_f)\sigma_l^f \end{aligned}\,.$$

This process is repeated until the Kalman update terms become negligible, usually in less than five iterations. Upon convergence, we project the remaining 3D points onto the image and match them to the nearest 2D feature point. 3D points whose nearest neighbor distance is larger than *Max_distance_inlier* are classified as outliers. Using both the inlier and outliers points, we compute the error of Eq. 1 and stop the algorithm for the current prior set $\{\mathbf{p}_k,\Sigma_k^{\mathbf{p}}\};\{f_l,\Sigma_l^f\}$ if the error falls below a given threshold. If not, we backtrack through the list of 3D-to-2D matches to change the assignments and repeat the guided search and refinement process. When no more assignments are available, we repeat the process with a different pose and focal length prior.

By progressively exploring these priors we are able to efficiently prune the potential number of 3D-to-2D matches, while reducing the uncertainty of the pose and focal length estimates. The method is shown to be highly resilient to clutter and noise on the image features and in the 3D model. The latter is especially suited for dealing with 3D models obtained from noisy range sensors, such as the Kinect or Time of Flight cameras.

[1] F. Moreno-Noguer, V. Lepetit, and P. Fua. Pose priors for simultaneously solving alignment and correspondence. In *ECCV*, 2008.

Discriminative Tensor Sparse Coding for Image Classification

Yangmuzi Zhang[1]
ymzhang@umiacs.umd.edu
Zhuolin Jiang[2]
zhuolin.jiang@huawei.com
Larry S. Davis[1]
lsd@umiacs.umd.edu

[1] University of Maryland, College Park, USA
[2] Noah's Ark Lab, Huawei Technologies

Sparse models have been successfully applied to many problems in image processing, computer vision, and machine learning. Many algorithms [3, 12] have been proposed to learn an over-complete and compact dictionary based on such models. In general, the input feature representations to these approaches are based on traditional vector descriptors. As pointed out in recent work [2, 7], vectorizing the original data structure, however, may destroy some inherent ordering information in the data. In computer vision, the region covariance feature, introduced in [9], is an image descriptor that captures natural correlations amongst multiple features. Hence, there has been growing interest in the development of sparse coding for positive definite descriptors. In [10], the problem of sparse coding within the space of symmetric positive definite matrices is tackled by embedding Riemannian manifolds into kernel Hilbert spaces. [7] proposed tensor sparse coding on positive definite matrices, which keeps descriptors in their original space and uses a set of randomly selected training samples as the dictionary. It successfully extended sparse coding techniques to the space of positive definite matrices. However, little research has been done to learn a discriminative and compact dictionary over such spaces.

We present a discriminative dictionary learning method for tensor sparse coding. Rather than simply using a subset of region covariance descriptors for training images as the dictionary [7], we learn a discriminative dictionary from the training set. A structural incoherence term is introduced into the dictionary learning process to regularize the incoherence between different sub-dictionaries, which increases the discriminativeness of the learned dictionary. We further incorporate classification error into the objective function to make the learned dictionary effective for classification tasks. Instead of learning multiple classifiers for each pair of classes [4, 5, 11], a linear multi-class classifier can be easily obtained during the training process. Unlike [8], which focuses on the reconstructive capability of a dictionary, the dictionary learned by our approach has both good reconstruction and discrimination capabilities. Based on this learned high-quality dictionary, we are able to obtain discriminative tensor sparse representations. Classification can be efficiently performed on these representations using the learned multi-classifier as it only involves matrix multiplication.

The quality of the dictionary influences the discriminativeness of the tensor sparse representations. Updating each dictionary atoms separately does not result in sufficient discriminating information in the sub-dictionaries. Following [1, 6], we introduce structural incoherence into sub-dictionary atoms. Incoherence will promote dictionary atoms from different classes to be independent from each other; thus it leads to sparse and discriminating representations for images. Given a training data set $S = \{S_l\}_{l=1}^N$, we will learn a dictionary $A = \{A^i\}_{i=1}^K$, with sub-dictionary $A^i = [\mathbf{a}_1^i, \mathbf{a}_2^i, ..., \mathbf{a}_{K_i}^i]$ for class i. The problem is formulated as:

$$\begin{aligned} min_{A,X} \quad & \Sigma_{l=1}^N D_{ld}(\mathbf{x}_l \otimes A, S_l) + \lambda||\mathbf{x}||_1 + \eta\Sigma_{i\neq j,s,t}||(\mathbf{a}_s^j)^T\mathbf{a}_t^i||_F^2 \quad (1) \\ s.t. \quad & \mathbf{x}_l \geq \mathbf{0} \quad \forall l \\ & \mathbf{a}_t^i,\ \mathbf{a}_s^j \succeq 0 \quad \forall i,t,j,s \\ & 0 \preceq \mathbf{x}_l \otimes A \preceq S_l \quad \forall l \end{aligned}$$

The first two terms are the reconstruction error and the sparsity regularization. The last term sums up the Frobenius norms between every two dictionary atoms $\mathbf{a}_s^j, \mathbf{a}_t^i$ which belong to different sub-dictionaries A^j and A^i. λ, η are penalty parameters balancing reconstruction error, sparsity, and dictionary structural incoherence.

To make the learned dictionary to be more adaptive to classification tasks, we minimize the classification error in the objective function of dictionary learning as pointed out in [3]. A linear multi-classifier $f(x;W) = Wx$ is used for classification. W denotes the linear classifier's parameters. Hence, the classification error can be explicitly included in the objective function during the dictionary learning. The classifier will be learned through the training process, as well. The objective function is formulated as below:

$$\begin{aligned} min_{A,X,W} \quad & \Sigma_{l=1}^N D_{ld}(\mathbf{x}_l \otimes A, S_l) + \lambda||\mathbf{x}||_1 + \eta\Sigma_{i\neq j,s,t}||(\mathbf{a}_s^j)^T\mathbf{a}_t^i||_F^2 \\ & +\gamma||H - WX||_2^2 \quad (2) \\ s.t. \quad & \mathbf{x}_l \geq \mathbf{0} \quad \forall l \\ & \mathbf{a}_t^i,\ \mathbf{a}_s^j \succeq 0 \quad \forall i,t,j,s \\ & 0 \preceq \mathbf{x}_l \otimes A \preceq S_l \quad \forall l \end{aligned}$$

where the term $||H - WX||_2^2$ represents the classification error. $H = [h_1, h_2, ..., h_N]$ $R^{K\times N}$ denotes the label matrix. The column vector $h_i = [0, 0, ...1...0, 0]^T \in R^K$ is a label vector for sample i. The position of 1 indicates its class index. γ controls the contribution of the classification error regularization term in the optimization process.

We have tested our approach on three types of public datasets, including texture, digit, and face database. Experiment results demonstrate the effectiveness of our approach.

[1] C. Chen, C. Wei, and Y. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition, 2012. *CVPR*.

[2] T. Hazan, S. Polak, and A. Shashua. Sparse image coding using a 3d non-negative tensor factorization, 2005. *ICCV*.

[3] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictioanry for sparse coding via label consistent k-svd, 2011. *CVPR*.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis, 2008. *CVPR*.

[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning, 2009. *NIPS*.

[6] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structural incoherence and shared features, 2010. *CVPR*.

[7] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Tensor sparse coding for region covariances, 2010. *ECCV*.

[8] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Positive definite dictionary learning for region covariances, 2011. *ICCV*.

[9] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification, 2006. *ECCV*.

[10] O. Tuzel, F. Porikli, and P. Meer. Human detectio via classification on riemannian manifolds, 2007. *CVPR*.

[11] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding, 2010. *CVPR*.

[12] Y. Zhang, Z. Jiang, and L. Davis. Learning structured low-rank representations for image classification, 2013. *CVPR*.

Hierarchical Scene Annotation

Michael Maire[1]
mmaire@caltech.edu
Stella X. Yu[2]
stellayu@icsi.berkeley.edu
Pietro Perona[1]
perona@caltech.edu

[1] California Institute of Technology
1200 East California Blvd
Pasadena, CA 91125
[2] UC Berkeley / ICSI
1947 Center St. Ste. 600
Berkeley, CA 94704

Supervised datasets play a central role as standards against which to benchmark long-term progress in computer vision. Over the past decade, the PASCAL [2] and Berkeley segmentation datasets (BSDS) [3] have filled these roles for the object detection and image segmentation tasks, respectively. The type of annotation available for each dataset determines the particular visual subtasks to which it is applicable. Object bounding boxes can benchmark detection algorithms, but are of limited use for training or evaluating segmentation. Segmented objects are more widely useful, but more time-consuming to annotate. What visual tasks are most important and what level of annotation detail is appropriate?

We present an alternative to thinking about dataset annotation in terms of a restricted set of visual tasks. Our key observation is that a hierarchical groundtruth representation, in the form of a ***doubly ordered region tree***, allows one to subsume disparate aspects of image labeling into a single framework. Specifically, we capture a nearly complete description of any scene in terms of objects, parts, object-part containment, segmentation, and figure-ground or occlusion ordering. Figure 1 illustrates the type of detail our annotation model encompasses for a typical scene.

Our unifying abstraction regards a scene as a set $S = \{R_1, R_2, ..., R_n\}$ where each $R_i \subseteq I$ is a region in the image I. In general, $R_i \cap R_j$ may be nonempty. It then organizes regions $\{R_i\}$ into a tree T. Let $N(R_i)$ denote the node of T corresponding to region R_i. $N(R_j)$ is set to be the parent node of $N(R_i)$ iff: (1) $R_j \supset R_i$, (2) R_j and R_i have an object-part relationship, and (3) $\nexists R_k : R_j \supset R_k \supset R_i$ and R_j, R_k, R_i have an object-part-subpart relationship, respectively. If for R_i, no region R_j satisfies all three conditions, then we set $N(R_i)$ to be a child of the root node. Simply stated, T decomposes the the scene into a multilevel object-part hierarchy.

We exploit one additional degree of freedom within T: the order $O(\cdot)$ in which nodes appear beneath a common parent encodes local occlusion relationships. Given sibling regions R_i and R_j such that $R_i \cap R_j \neq \emptyset$, then R_i occludes R_j if $O(R_i) < O(R_j)$, and R_j occludes R_i if $O(R_i) > O(R_j)$. If $R_i \cap R_j = \emptyset$, then they do not occlude one another and we disregard their relative ordering. Given T and O, preorder tree traversal uses the object-part hierarchy to translate local occlusions into a global figure-ground ordering. It also recovers a groundtruth ultrametric contour map (UCM) [1] weighting visible boundaries by the structural importance (object, part, subpart) of the regions they enclose.

Figure 2 shows a partial scene tree for a self-occluding object. Our full paper describes extensions for representing loops (shirt both behind and in front of arm). To mitigate the cost of creating rich groundtruth, we introduce a ***web-based annotation tool*** with a graphical interface for managing the region hierarchy. Our software eliminates tedious tracing of region boundaries through a dynamic paintbrush that snaps to the shape of underlying superpixels in a precomputed oversegmentation. Combined with a touch-up mode, it guides fast creation of pixel-perfect regions.

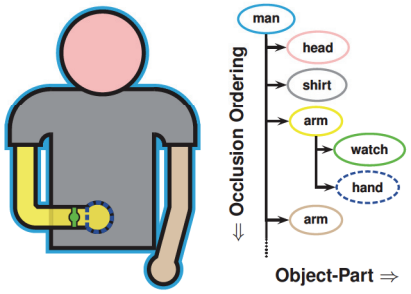


Figure 2: Our model of a scene or object groups pixels into regions and maps regions to nodes in a tree. Parent-child links denote region containment and semantic object-part relationships. Relative ordering of sibling nodes resolves occlusion ambiguities.

We use our annotation tool to groundtruth a dataset of 97 photographs previously used in experiments measuring the importance of objects in complex scenes [4]. Similar in size to the original BSDS test set, but with greater diversity in scene type and wider range of object scale, our labeled image set provides an object-centric context in which to benchmark segmentation algorithms. Instead of binary groundtruth boundary maps, as used in the BSDS, we have ***real-valued hierarchical groundtruth*** in the form of UCMs generated by scene tree traversal.

This permits direct evaluation of the degree to which the hierarchical output of state-of-the-art machine segmentation algorithms, such as gPb-UCM [1], respects the groundtruth object-part hierarchy. Specifically, we examine the order in which boundaries are recalled as one varies an algorithm's boundary detection threshold. The ideal algorithm would recover the occlusion boundaries between the top-level objects in the scene first, followed by large-scale object-part boundaries, followed by finely detailed part-subpart boundaries.

Our ***benchmark*** reveals a mismatch between gPb-UCM and the groundtruth object-part hierarchy. This performance gap serves as a call for further research into object-centric hierarchical scene segmentation and underscores the importance of our annotation tool and dataset.

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[2] Mark Everingham, Luc van Gool, Chris Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 2010.

[3] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2001.

[4] Merrielle Spain and Pietro Perona. Measuring and predicting object importance. *IJCV*, 2010.

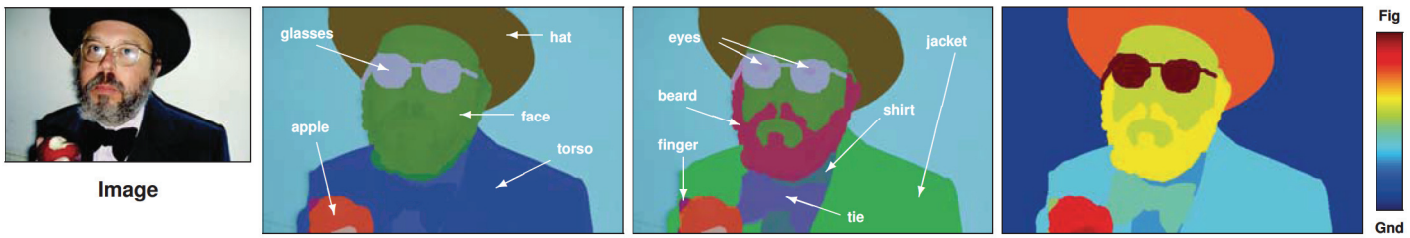


Figure 1: At the coarsest level, this scene contains two objects: a man standing in front of a wall. Looking in detail, we subdivide the man into regions for his hat, glasses, face, torso, and the apple he holds. In even more detail, his face consists of eyes, skin, and beard, and his torso is covered by a shirt, tie, and jacket. Traditional segmentation datasets label a single two-dimensional region partition. Our annotation model captures the object-part-subpart decomposition, as well as the occlusion relationships (*e.g.* apple in front of jacket, glasses in front of face) present in the scene.

Blockwise Linear Regression for Face Alignment

Enrique Sánchez-Lozano[1]
esanchez@gts.uvigo.es
Enrique Argones-Rúa[2]
eargones@gradiant.org
José Luis Alba-Castro[1]
jalba@gts.uvigo.es

[1] Multimedia Technologies Group
AtlantTIC Research Center
University of Vigo
Vigo, Spain

[2] Multimodal Information Area
Gradiant
Vigo, Spain

In this paper, we present a deeper analysis of linear regression as an efficient method for face alignment. Linear regression for face alignment [2] consists of learning a mapping matrix between image features and shape parameters displacements. Typically, shape parameters are projections of a set of points which follow a Point Distribution Model onto the subspace generated by applying Principal Component Analysis to a set of manually landmarked training images. Often, image features are extracted within patches around each point, and concatenated into a column vector $\mathbf{d}$. We call the features belonging to the i-th patch with $\mathbf{d}^i$. We can generate examples for each image j by systematically perturbing the ground-truth parameters $\mathbf{p}_0^j$ with $\delta\mathbf{p}$. If we rearrange the training features into the matrix $\mathbf{D}$, and their corresponding perturbations into the matrix $\mathbf{P}$, the mapping matrix is obtained through least squares:

$$\mathbf{R} = \min_{\mathbf{R}} \|\mathbf{P} - \mathbf{R}\mathbf{D}\|_F^2 = \mathbf{P}\mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}. \quad (1)$$

A closer look to Eqn. (1) reveals what is inside $\mathbf{D}\mathbf{D}^T$. Let us consider each sample patch i, from the k-th training image, $\mathbf{d}_k^i$, as a contribution of three terms: the mean feature vector (intrinsic for all faces) $\mathbf{d}_0^i$, the subject-dependent features $\widetilde{\mathbf{d}}_k^i$, and a sample noise (which may be produced by the landmarking process) ε_k^i, which is assumed zero-mean gaussian. Then, $\mathbf{D}\mathbf{D}^T$ holds all the cross-products between patches along the training set. However, the main contribution of these products comes from the mean part. Fig. 1 (left) shows an example of this matrix. Let us have a closer look to the covariance matrix (right), which consists of the outer products of the subject-dependent features and the sample noise, since they can not be separated in practice. As can be seen, most of the information lie on the diagonal, although some remaining information is found outside it. Thus, we propose to find which patches are related, and cluster them, in order to remove all the remaining noise. Once patches are clustered, and the noise is removed, a "clean" mapping matrix is obtained.

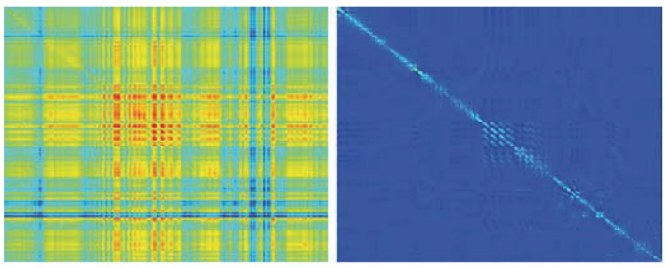


Figure 1: **Left:** Cross products between samples across the training set (best viewed in color). **Right:** Covariance matrix.

Correlation: For measuring the relation between two patches, we define the *modified Pearson correlation coefficient* as:

$$\gamma^{ij} = \frac{\|(\widetilde{\mathbf{D}}^i)(\widetilde{\mathbf{D}}^j)^T\|_F^2}{\|(\widetilde{\mathbf{D}}^i)(\widetilde{\mathbf{D}}^i)^T\|_F \|(\widetilde{\mathbf{D}}^j)(\widetilde{\mathbf{D}}^j)^T\|_F}, \quad (2)$$

where $\widetilde{\mathbf{D}}^i$ are the zero-mean features belonging to patch i. $\gamma^{ij} \in [0,1]$ is a measure of correlation between structures of independent variables. Two patches are related if $\gamma^{ij} \geq \theta$, where θ is a threshold.

Clustering: When two patches are related with a third one, but not with themselves, they must be clustered together. Two patches are clustered together if, and only if, there exists some patch that is related with both patches. If $\mathcal{P}$ denotes a patch, the clustering operation can be summarized as:

$$P^i \sim P^j \iff \{\exists P^h \in \mathcal{P} \,|\, \gamma^{ih} \geq \theta\,, \gamma^{jh} \geq \theta\}. \quad (3)$$

A fast way to see whether two patches are related or not, as well as whether two patches belong to the same cluster is through a binary matrix $\mathbf{Z}$, where the element (i,j) denotes whether two patches (i and j) meet $\gamma^{ij} \geq \theta$ (1), or not (0). An algorithm for clustering the data from the matrix $\mathbf{Z}$ and some examples can be found in the paper.

Regression: Once clusters are obtained, a regression matrix is learned for a modified version of each one. Instead of working with the whole matrices, we remove non-related products, by working with $z^{ij}\widetilde{\mathbf{D}}^i\widetilde{\mathbf{D}}^j$ (thorough mathematics can be found in the paper). Let $\widetilde{\mathcal{D}}_l$ be the features belonging to the l-th cluster, and $\widetilde{\mathcal{M}}_{ll,\theta} = \{z^{ij}\widetilde{\mathbf{D}}^i\widetilde{\mathbf{D}}^j | P^i, P^j \in \mathcal{C}_l\}$. Each regression matrix is then obtained as follows:

$$\mathcal{R}_l = \mathbf{P}(\widetilde{\mathcal{D}}_l)^T(\widetilde{\mathcal{M}}_{ll,\theta})^{-1}. \quad (4)$$

Fitting: Fitting is now calculated as follows. Consider the input data $\mathbf{d}$, which is divided into c clusters $\mathbf{d}_l$, and the mean feature vector $\mathbf{d}_0$, also divided into clusters $(\mathbf{d}_{10} \ldots \mathbf{d}_{c0})$. Shape parameters are now computed as

$$\delta\mathbf{p} = \sum_{l=1}^{c} \mathcal{R}_l(\mathbf{d}_l - \mathbf{d}_{l0}), \quad (5)$$

Experiments: We have trained our algorithm using a mixture of the Multi-PIE [3] and LFPW [1] databases, and tested it on the BioID [4]. We have measured the typical me_{17} error, by combining the eyes detection with the fitting done with our algorithm. Fig. 2 shows the results obtained for different values of θ. As can be seen, the best performance occurs when considering some relations among patches, although it may vary depending on the database.

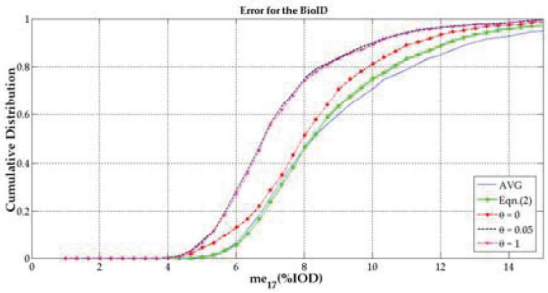


Figure 2: Fitting results obtained for the BioID database, varying θ.

[1] P.N. Belhumeur, D. Jacobs, D.J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, 2011.

[2] T.F. Cootes and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(6): 680–689, 2001.

[3] R. Gross, I. Matthews, J.F. Cohn, T. Kanade, and S. Baker. Multi-pie. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG'08)*, 2008.

[4] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *Int'l Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'01)*, pages 90–95, 2001.

Fast and Robust ℓ_1-averaging-based Pose Estimation for Driving Scenarios

German Ros[1,2]
gros@cvc.uab.es
Julio Guerrero[3]
juguerre@um.es
Angel D. Sappa[1]
asappa@cvc.uab.es
Daniel Ponsa[1,2]
daniel@cvc.uab.es
Antonio M. López-Peña[1,2]
antonio@cvc.uab.es

[1] Computer Vision Center
Edifici O, Campus UAB, 08193
Bellaterra (Barcelona), Spain
[2] Computer Science Dept.
Universitat Autònoma de Barcelona
Campus UAB, Bellaterra (Barcelona), Spain
[3] Dept. de Matemática Aplicada
FIUM, Universidad de Murcia
Campus de Espinardo, Murcia, Spain

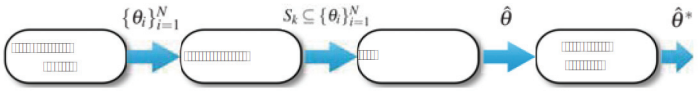


Figure 1: Main pipeline stages of the presented technique.

Robust camera-pose estimation is a fundamental stage of many computer vision problems, being specially important for Visual Simultaneous Localization and Mapping (*VSLAM*) [1][6] and Visual Odometry (*VO*)[4][5] systems. Here, the capability of estimating correct models in the presence of noise and high level of outliers is a fundamental requirement. During the last decades, many approaches have been proposed to solve these problems, being RANSAC [2] one of the most accepted and used. However, with the arrival of new challenges, such as large driving scenarios for autonomous vehicles, along with the improvements in the data gathering frameworks, new issues must be considered. One of these issues is the capability of a technique to deal with very large amounts of data while meeting the real-time constraint.

The use of large amounts of data to perform model estimation has proven to be beneficial for improving model accuracy, as stated in [7]. In the past, the amount of input information was very limited, but now modern front-ends are able to extract thousands of features from a set of images and match them in real-time in a standard CPU. Nevertheless, serious issues, as the presence of outliers within the flow of data still need to be carefully addressed. RANSAC-like methods are affected by this "excess" of information, what produces an increment on the time dedicated to evaluate and rank the generated models. In order to avoid this drawback, real-time implementations opt to use just a part of the available data, therefore discarding a great amount of information and penalizing the accuracy of the resultant models. To address these problems, in the presented work we propose a novel technique to perform robust camera-pose estimation that is specially suitable to deal with large flows of data, something that is common in urban scenarios, where we successfully tested the approach.

The main idea behind our proposal is to combine a very fast hypotheses assessment that produces sub-optimal evaluation and a procedure capable of combining partially incorrect hypotheses into a new and accurate hypothesis. The procedure is depicted by Fig. 1, and starts by creating N candidate models $\theta_i \in \mathbb{SE}(3)$ from the available data $\mathcal{X} = \{(x_{l,p}, x_{r,p}, x_{l,c}, x_{r,c})^{(i)}\}_{i=1}^D$, which represent matched points in the four views of a moving stereo-rig (i.e., at two different time instants). This step is carried out by optimizing Eq. 1 for $M = 3$ correspondences while accounting for the manifold structure of the model.

$$\text{argmin}_\psi \sum_{i=1}^M \left\| \mathrm{K}\Pi_3 \left(\mathbf{exp}_r(\psi)\hat{X}_{l,p}^{(i)}\right) \times \hat{x}_{l,c}^{(i)} \right\|_{\ell_2}^2 + \left\| \mathrm{K}\left(\Pi_3 \left(\mathbf{exp}_r(\psi)\hat{X}_{l,p}^{(i)}\right) - \vec{B}\right) \times \hat{x}_{r,c}^{(i)} \right\|_{\ell_2}^2 \quad (1)$$

After generating the set of models $\{\theta_i\}_{i=1}^N$, they are assessed by F_{coarse} (Eq. 2), an evaluation function that has been designed to be extremely fast. This is achieved thanks to the use of a *Reduced Measurement Matrix* (*RMM*) [3], an algebraical reduction of the input data $\mathcal{X}$ that creates a compact equivalent $\boldsymbol{M}$ under the ℓ_2-norm. The advantage of this reduction is that $\boldsymbol{M}$ can be efficiently computed even for very large collections of data and this has to be done just once, at the beginning of the process.

$$F_{coarse}(\theta) = \sum_{i=1}^D \left\| \mathrm{K}\left(\Pi_3 \theta \hat{X}_{l,p}^{(i)}\right) \times \hat{x}_{l,c}^{(i)} \right\|_{\ell_2}^2 + \left\| \mathrm{K}\left(\Pi_3 \theta \hat{X}_{l,p}^{(i)} - \vec{B}\right) \times \hat{x}_{r,c}^{(i)} \right\|_{\ell_2}^2 =$$
$$\sum_{i=1}^D \left\| W_l^{(i)} \check{\theta} \right\|_{\ell_2}^2 + \left\| W_r^{(i)} \check{\theta} \right\|_{\ell_2}^2 = \left\| \boldsymbol{W_l} \check{\theta} \right\|_{\ell_2}^2 + \left\| \boldsymbol{W_r} \check{\theta} \right\|_{\ell_2}^2 = \check{\theta}^T (\boldsymbol{M_l} + \boldsymbol{M_r}) \check{\theta} \quad (2)$$

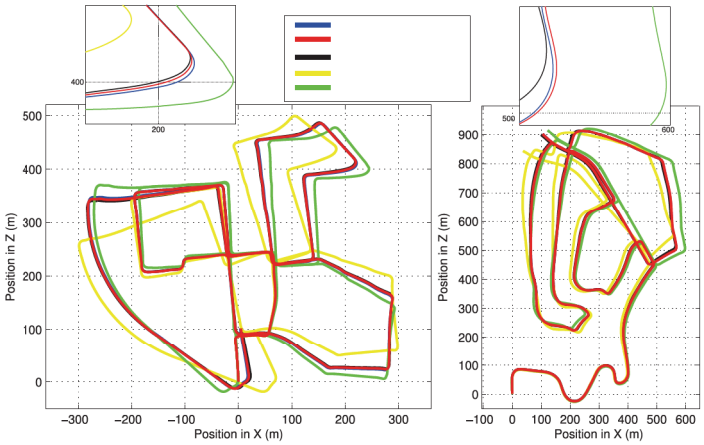


Figure 2: Results of the *VO* experiments for KITTI sequences 00 (left) and 02 (right). Notice that, although all robust methods lead to similar trajectories, C-Avg and P-Avg remain closer to the ground truth.

However, F_{coarse} is not robust to noise and outliers, what will lead to the wrong evaluation of some models. Fortunately, a thorough analysis of real data sequences showed that models with a high number of inliers (the expected good models) produce low residuals for F_{coarse}, even when $\boldsymbol{M}$ contains outliers. On the other hand, models corresponding with a low number of inliers present random values for F_{coarse}, producing low residuals just occasionally. Our approach exploits this weak property by selecting a subset $S_k \subseteq \{\theta_i\}_{i=1}^N$ of partially corrupted models according to the decision of F_{coarse}.

Afterwards, all the models are combined in a robust and fast way to generate a final and more accurate model $\hat{\theta}$. This is done by using the Weiszfeld algorithm to perform ℓ_1-averaging on $\mathbb{SE}(3)$ in order to respect the structure of the camera pose. We show that, under the appropriate conditions, the resultant model $\hat{\theta}$ is very accurate, at the level of models estimated by RANSAC, as shown in Fig. 2. The benefits of using this approach against RANSAC are shown in the paper.

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE TPAMI*, 2007.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6), 1981.

[3] R. Hartley. Minimizing algebraic error in geometric estimation problems. In *Proceedings of the IEEE ICCV*, Washington, DC, USA, 1998.

[4] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of the IEEE IV Symposium*, San Diego, CA, USA, 2010.

[5] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the IEEE CVPR*, volume 1, Washington, DC, USA, 2004.

[6] G. Sibley, C. Mei, I. Reid, and P. Newman. Vast-scale outdoor navigation using adaptive relative bundle adjustment. *International J. of Robotics Research*, 2010.

[7] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Visual SLAM: Why filter? *Image Vision Comput.*, 30(2):65–77, February 2012.

Learning Speed Invariant Gait Template via Thin Plate Spline Kernel Manifold Fitting

Sheng Huang[1]
https://sites.google.com/site/siteshuang/
Ahmed Elgammal[2]
http://www.cs.rutgers.edu/~elgammal
Dan Yang[1]

[1] College of Computer Science, Chongqing University
[2] Department of Computer Science, Rutgers, The State University of New Jersey

Similar to the solutions of general gait recognition approaches, there are two ways to handle the cross-speed gait recognition issues. The first approach is model-based approach which is to model the walking action using static or dynamic body parameters [6]. The second class of cross-speed gait recognition approach is model free approach, also known as feature or appearance-based approach. It focuses on directly extracting holistic gait features from gait sequences [4, 5]. Our proposed method, Speed Invariant Gait Template (SIGT), belongs to the second category.

According to some recent studies, the walking action is considered as residing on a manifold which is topologically equivalent to a unit circle [1]. Thus, we can format the gait feature extraction issue as a gait manifold fitting issue. To address the cross-speed gait recognition issue, Thin Plate Spline (TPS) kernel based RBF interpolation is used to fit the gait manifold, since TPS has a desirable property [7] that it separates the mapping coefficients into an affine component and a non-affine component. And a natural assumption, that the affine component depicts the manifold deformation caused by walking and the non-affine component preserves the static features irrelevant to motion, can be given. Consequently, the non-affine component is a gait feature which is robust to the walking speed variation and we name this new feature *Speed Invariant Gait Template* (SIGT) (see Figure 1).

We show in this paper how to use our proposed TPS-based manifold model to extract SIGTs from the full cyclic gait silhouette sequences. Firstly, The implicit function-based representation is used to represent each gait silhouette, since this representation is robust to the noise and silhouette fragmentation. Each pixel x of a gait silhouette is represented as 0 when it is on the contour, as $d_c(x)$ when it is inside the contour and as $-d_c(x)$ when it is outside the contour, where $d_c(x)$ denotes the distance to the closest point on the contour.

Next, TPS Kernel based RBF interpolation is used to learn the mapping between the manifold embedding space and the represented gait space. Let $Y = \{y_i \in R^l, i = 1, \cdots, N\}$ be a gait sequence in the gait space and $X = \{x_i \in R^2, i = 1, \cdots, N\}$ be the corresponding points in the embedding space. $T = \{t_j \in R^2, i = 1, \cdots, N_t\}$ denotes N_t equally spaced centers in the embedding space. We can solve for multiple TPS kernel interpolants $f^k : R^2 \rightarrow R$ where k is the kth element (dimension) of the gait vector (the vectorizated represented gait image) in the gait space and f^k is a RBF interpolant. We minimize a regularized risk criteria to learn nonlinear mapping from the embedding space to each individual dimension in the gait space that satisfies $y_i^k = f^k(x_i)$. From the representer theorem [3], such a function admits a representation of the form of linear combination of basis functions around arbitrary points (centers). Therefore, to the kth dimension of the input, the form of function $f^k(x)$ is as follows:

$$f^k(x) = p^k(x) + \sum_{i=1}^{N_t} d_i^k \phi(||x - t_i||_2) \tag{1}$$

where $\phi(u) = u^2 log(u)$ is a thin plate spline function and $p^k(x) = [1, x^T] \cdot s^k$ denotes the TPS smoothness term as a linear polynomial function with coefficients s^k. The matrix form of interpolation is as follows:

$$f(x) = W \cdot \Phi(x) \tag{2}$$

where W is a $l \times (N_t + 3)$ matrix with the kth row $[d_1^k, \cdots, d^k{}_{N_t}, s^{kT}]$ and $\Phi(x)$ is a vector that $\Phi(x) = [\phi(||x - t_i||_2), \cdots, \phi(||x - t_{N_t}||_2), 1, x^T]$. The matrix W represents the mapping coefficients which are the l nonlinear mappings from the embedding space to gait space. In order to make the problem be well posed and insure the orthogonality, an additional constraint should be added:

$$\sum_{i=1}^{N_t} d_i p_j(x_i) = 0, j = 1, 2, 3 \tag{3}$$



(a) (b)

Figure 1: The visualization of SIGTs, (a) The sparsely sampled gait silhouettes of half cycle from OU-ISIR Treadmill Dataset A, (b) The extracted SIGTs from the gait sequences.

where p_j is the linear basis of the polynomial part $p(x)$.

Thus, the mapping coefficients W can be obtained by directly solving the following linear systems:

$$\begin{pmatrix} C & P_x \\ P_t{}^T & 0_{3\times 3} \end{pmatrix} W^T = \begin{pmatrix} C & P_x \\ P_t{}^T & 0_{3\times 3} \end{pmatrix} \begin{pmatrix} D & S \end{pmatrix}^T = \begin{pmatrix} Y \\ 0_{3\times l} \end{pmatrix} \tag{4}$$

where C is a $N \times N_t$ matrix with $C_{ij} = \phi(||x_i - t_j||_2), i = 1, \cdots, N, j = 1, \cdots, N_t$, P_x is a $N \times 3$ matrix with ith row $[1, x_i{}^T]$, P_t is a $N_t \times 3$ matrix with ith row $[1, t_i{}^T]$. $D = [d_1, \cdots, d_{N_t}]$ is a $l \times N_t$ coefficient matrix of TPS based interpolants and $S = [s^T]$ is a $l \times 3$ coefficient matrix of the smoothness term. According to the property of TPS, the static feature of gait manifold, which is robust to the speed variation, is embedded in the $l \times 3$ matrix S. Then, the proposed *Speed invariant Gait Template* (SIGT) can be obtained by vectorizing this matrix.

Finally, in order to speed up the recognition and avoid the curse of dimensionality, an recent Improved Locality Preserving Projections (LPP) method named Globality-Locality Preserving Projections (GLPP) [2] is applied to reduce the dimensionality of SIGT. Compared to the state-of-the-arts methods such as LDA, PCA and LPP, GLPP obtains a much better recognition performance via taking geometric structures of both samples and classes into consideration.

The implementations of our method and related algorithms are described in this paper in detail. Our conclusion is that a novel gait template for cross-speed gait recognition is proposed and these gait templates are extracted by fitting the gait manifold via TPS kernel based RBF interpolation. The main contribution of this work is that it provides a natural way to separate the dynamic features and static features and such separation is very general to other computer vision issues.

[1] A. Elgammal and C-S. Lee. Homeomorphic manifold analysis (hma): Generalized separation of style and content on manifolds. *Image and Vision Computing*, 2013.

[2] S. Huang, D. Yang, F. Yang, Y. Ge, X. Zhang, J. Zhou, and J. Lu. Face recognition via globality locality preserving projections. *Neurocomputing*, (under review), 2013.

[3] G.S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[4] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Gait recognition across various walking speeds using higher order shape configuration based on a differential composition model. *IEEE T-MSCB*, pages 1654–1668, 2012.

[5] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE T-PAMI*, 28(6):863–876, 2006.

[6] A. Tsuji, Y. Makihara, and Y. Yagi. Silhouette transformation based on walking speed for gait identification. In *CVPR*, pages 717–722, 2010.

[7] G. Wahba. *Spline models for observational data*, volume 59. Society for industrial and applied mathematics, 1990.

Parsing Clothes in Unrestricted Images

Nataraj Jammalamadaka
nataraj.j@research.iiit.ac.in
Ayush Minocha
ayush.minocha@students.iiit.ac.in
Digvijay Singh
digvijay.singh@students.iiit.ac.in
C. V. Jawahar
jawahar@iiit.ac.in

Center for Visual Information Technology
IIIT Hyderabad
India, 500032

Cloth parsing involves locating and describing all the clothes (e.g., T-shirt, shorts) and accessories (e.g, bag) that the person is wearing. The main challenges in solving this include the large variety of clothing patterns that have been developed across the globe by different cultures. Occlusions from other humans or objects, viewing angle and heavy clutter in the background further complicates the problem.

In the recent past, Yamaguchi *et al*. [4] have proposed a method to parse clothes for fashion photographs where the image settings are simple with no clutter or occlusion. In our work, we aim to segment clothes in unconstrained settings by modelling the cloth to its body part vicinity in a CRF framework. Poselets [2] are adapted to obtain body part locations, as alternatives like human pose estimation algorithms frequently fail and give wrong pose estimates under occlusions and clutter.

Given an image, first the superpixels and body joint locations are computed. These superpixels form the vertices V of the CRF. Two superpixels which share a border are considered adjacent and are connected by an edge $e \in E$. The best labeling using the CRF model is given by the equation,

$$\hat{L} = argmax_L P(L|Z,I), \tag{1}$$

where L is the label set, Z is a distribution of the body joint locations and I is the image.

The MAP configuration of CRF probability function given by the equation 1 is computationally expensive to compute and is usually a NP-hard problem. We thus make a simplifying assumption that at most two vertices in the graph form a clique thus limiting the order of a potential to two. Thus the CRF factorizes into unary and pair-wise functions and the log probability function is given by,

$$\begin{aligned} \ln P(L|Z,I) \equiv \sum_{i \in V} \Phi(l_i|Z,I) \quad &+ \quad \lambda_1 \sum_{(i,j) \in E} \Psi_1(l_i,l_j) \\ &+ \quad \lambda_2 \sum_{(i,j) \in E} \Psi_2(l_i,l_j|Z,I) - \ln G, \end{aligned} \tag{2}$$

where V is the set of nodes in the graph, E is the set of neighboring pairs of superpixels, and G is the partition function.

The unary potential function Φ models the likelihood of a superpixel s_i taking the label l_i. First, using the estimated pose $Z = (z_1, \ldots, z_P)$ and the superpixel s_i, a feature vector $\phi(s_i,Z)$ is computed. Using the pre-trained classifier $\Phi(l_i|\phi(s_i,Z)) = \Phi(l_i|Z,I)$ for label l_i, a score is computed.

For the pairwise potentials, we use the definitions from [4]. Pairwise potential, defined between two neighboring super-pixels, model the interaction between them. The pair-wise potential is defined in equation 3 as sum of two functions (called factors) $\Psi(l_i,l_j)$ and $\Psi(l_i,l_j|Z,I)$. The pairwise potential function Ψ_1 models the likelihood of two labels l_i,l_j being adjacent to each other and Ψ_2 models the likelihood of two neighboring sites s_i,s_j taking the same label given by the features $\phi(s_i,Z)$ and $\phi(s_j,Z)$ respectively. The function Ψ_1 is simply a log empirical distribution and Ψ_2 is model learnt over all the label pairs respectively. The pairwise potential functions are given by,

$$\Psi_1(l_i,l_j), \quad \Psi_2(l_i,l_j|Z,I) \equiv \Psi_2(l_i,l_j|\psi(s_i,s_j,Z)) \tag{3}$$

where $\psi(s_i,s_j,Z)$ is defined as,

$$\psi(s_i,s_j,Z) \equiv \left[(\phi(s_i,Z)+\phi(s_j,Z))/2, |(\phi(s_i,Z)-\phi(s_j,Z))/2|\right]. \tag{4}$$

Given images and cloth labels which include background, Logistic regression is used for $\Phi(l_i|Z,I)$ and $\Psi_2(l_i = l_j|\psi(s_i,s_j,Z))$. Given a new image, the super-pixels, poselets and feature vector ϕ are computed. For each super-pixel, the unary potential and pairwise potential values are computed using the feature vector and the learnt models. The best label is inferred using the belief propagation implemented in libDAI [3] package. The parameters λ_1, λ_2 in the equation 3 are found by cross validation. Our experiments on the complex H3D dataset [2] indicates that the proposed algorithm significantly outperformed the previous work [4] while on a relatively simple Fashionista dataset [4] it is on par.

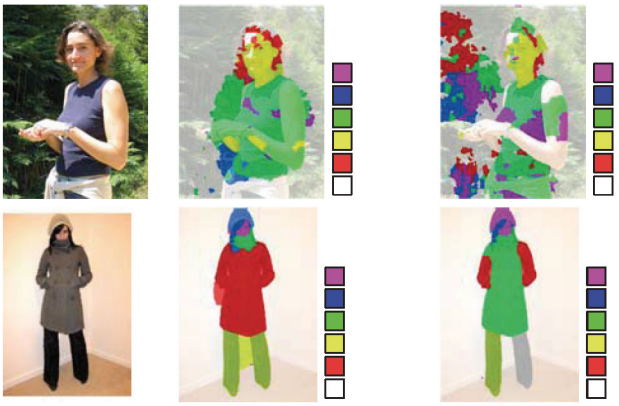


Figure 1: For the input image in column 1, our results are displayed in column2 and results of [4] are displayed in column 3.

Using the labelling obtained from the above method, interesting cloth and color co-occurrences can be mined using apriori algorithm [1].



Figure 2: **Cloth co-occurrences (Row 1):** The first three images display Cardigan-Dress co-occurrence and the next three images display Top-Skirt co-occurrence. **Color co-occurrence (Row 2):** The first three images display blue-blue co-occurrence and the next three images display white-blue co-occurrence.

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *CVPR*, 2009.

[3] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010. URL `http://www.jmlr.org/papers/volume11/mooij10a/mooij10a.pdf`.

[4] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

Image Fusion Using Gaussian Mixture Models

Katie Heaps[1]
heaps006@umn.edu
Josh Koslosky[1]
kosl0052@umn.edu
Glenn Sidle[2]
sidleg@duq.edu
Stacey Levine[2]
sel@mathcs.duq.edu

[1] Department of Mathematics
University of Minnesota
Minneapolis, MN, USA

[2] Department of Mathematics and Computer Science
Duquesne University
Pittsburgh, PA, USA

Recently, a number of works have show that patch-based image features can lead to more robust models than their pixel based counterparts. Critical to their success is that natural image patches can be expressed sparsely in appropriately defined dictionaries which can be tuned to a variety of applications. Yu, Sapiro and Mallat [1] demonstrated that estimating image patches from multivariate Gaussians is equivalent to finding sparse representations in a structured overcomplete PCA-based dictionary and can be solved using a straightforward piecewise linear estimator (PLE). In this work we show how a similar PLE can be formulated to fuse images with various linear degradations and different levels of additive noise.

We consider the degradation model $y = Uf + w$, in which a given image f has undergone some linear degradation U and is corrupted by additive noise $w \sim \mathcal{N}(0, \sigma^2)$. Decomposing f into overlapping $\sqrt{n} \times \sqrt{n}$ *vectorized* patches $f_i \in \mathbb{R}^n$, $i = 1..,I$ and noting that each patch may have a unique linear degradation U_i and unique additive noise w_i, we can express the degraded image patches as $y_i = U_i f_i + w_i$ for $i = 1,..,I$. Recovering f from y then becomes the problem of recovering f_i from y_i and rebuilding f from $\{f_i\}_{i=1}^I$.

In this work we propose a model for recovering f from J images, $y^j = U^j f + w^j$, $j = 1,...,J$, where the linear degradations $U^j$and noise levels $w^j \sim \mathcal{N}(0, \sigma_j^2)$ may vary amongst images. Following the single image recovery model proposed in [1], we show that given a fixed number of multivariate Gaussians parametrized by their means $\mu_k \in \mathbb{R}^n$ and covariance matrices $\Sigma_k \in \mathbb{R}^{n \times n}$, $k = 1,...,K$, each patch f_i (for $i = 1,...,I$) can be estimated by maximizing the log *a posteriori* probability $p(f_i | y_i^1, ..., y_i^J, \Sigma_k)$, which is equivalent to solving

$$(\hat{f}_i, \hat{k}_i) = \arg\min_{f,k} \sum_{j=1}^{J} \left(\frac{1}{\sigma_j^2} ||U_i f - y_i^j||^? \right) + f^T \Sigma_k^{-1} f + \log|\Sigma_k|. \quad (1)$$

The best estimate from each Gaussian is found using a simple linear filter

$$f_i^k = \sum_{j=1}^{J} W_{k,i}^j y_i^j, \; k = 1,...,K \quad (2)$$

where $W_{k,i}^j = \Sigma_k \left(\sum_{l=1}^{J} \frac{1}{\sigma_l^2} (U_i^l)^T U_i^l \Sigma_k + Id \right)^{-1} (U_i^j)^T$.

For each patch, the solution $\hat{k}_i$ of (1) is found by simply evaluating the functional at f_i^k (from (2)) for $k = 1,...,K$, and choosing $\hat{k}_i$ to be the one that yields the smallest value. Then the minimizing patch for (1) is

$$\hat{f}_i = f_i^{\hat{k}_i}. \quad (3)$$

The patches $\{\hat{f}_i\}_{i=1}^I$ are overlapping and averaged at each location to produce the final result, $\hat{f}$. So each pixel will typically be estimated by a mixture of Gaussians.

Once the model selection $\hat{k}_i$ and signal estimation $\hat{f}_i$ are known for each patch, setting $C_k = \{i | \hat{k}_i = k\}$ for $k = 1,...,K$, the new Gaussian parameters are computed by

$$\hat{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \hat{f}_i \text{ and } \hat{\Sigma}_k = \frac{1}{|C_k|} \sum_{i \in C_k} (\hat{f}_i - \hat{\mu}_k)^T (\hat{f}_i - \hat{\mu}_k). \quad (4)$$

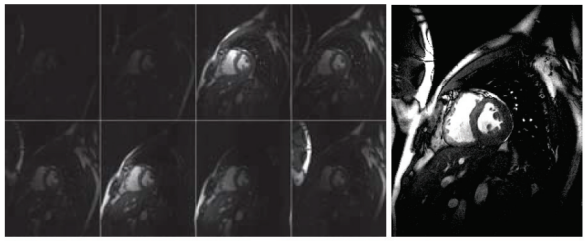
The steps (1) and (4) are iterated until convergence, thus the initialization of the parameters $(\mu_k, \Sigma_k)_{k=1,...,K}$ is critical. We follow the initialization proposed in [1] based on clustering edges at fixed orientations.

For each $k \in \{1,...,K\}$, the solution f_i^k in (2) can also be interpreted as a sparse patch-based representation. Specfically, $f_i^k = B_k a_i^k$ where

$$a_i^k = \arg\min_{a \in \mathbb{R}^n} \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} ||U_i^j B_k a - y_i^j||^2 + \sum_{m=1}^{n} \frac{|a[m]|^2}{\lambda_m^k} \right). \quad (5)$$

Here B_k is the orthonormal PCA basis diagonalizing the covariance matrix Σ_k corresponding to the k^{th} multivariate Gaussian, and the λ_m^k are the corresponding eigenvalues in decreasing order. The weighted l_2 norm in (5) favors dominant eigenvectors, and since the eigenvalues rapidly decrease in magnitude [1], it induces sparsity. Furthermore, the PCA basis B_k and the eigenvalues incorporate information from the data, so this model promotes collaborative filtering where the dictionary is learned from the data.

The model is flexible, and the level of smoothing can also be adapted to each patch. A number of experimental results demonstrate the success of the model, some of which are included below.



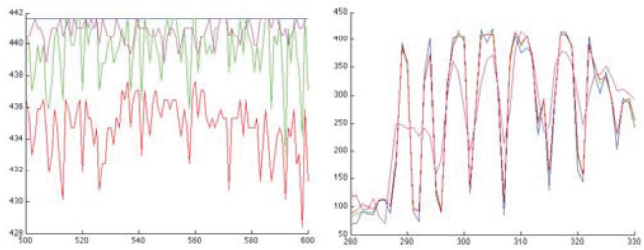
(a) 8 coil MRI data, co. Mark Griswold. (b) Our fusion result.



(c) blurry image, PSNR=27.5454 (d) noisy image, PSNR=29.2852



(e) original image (f) fusion result, PSNR=35.7534



(g) flat region of the sky (h) lettering on the plane

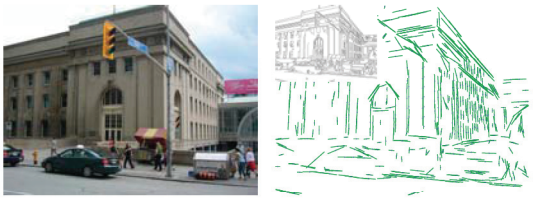
Figure 1: (a-b) fusing 8 coil MRI data; (c-f) fusing a noisy/blurry pair; (g-h) typical cross-sections of the magnitude of the intensity for the blurry/noisy pair (c)-(d): blue=original, green=fused result (PSNR=35.7534), red=denoised only (PSNR=33.4498), magenta=deblurred only (PSNR=31.3776).

[1] Guoshen Yu, Guillermo Sapiro, and Stephan Mallat. Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. *IEEE Trans. Image Process.*, 21(5): 2481– 2499, 2012.

Real Projective Plane Mapping for Detection of Orthogonal Vanishing Points

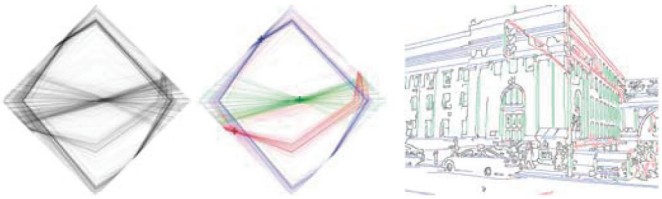
Markéta Dubská
http://www.fit.vutbr.cz/~idubska
Adam Herout
http://www.fit.vutbr.cz/~herout

Graph@FIT
Brno University of Technology
Czech Republic



(a) Input image (b) Extracted edges

Figure 1: Detection of orthogonal vanishing points. Input image and edglets corresponding to edge points.



(a) Diamond space (right: color-coded contributions) (b) Color-coded edges

Figure 3: Accumulated "diamond space" and the affiliation of edges/contributions to different orthogonal VPs by color.

This paper deals with the detection of orthogonal vanishing points in the Manhattan world. We are using a modified scheme of the Cascaded Hough Transform where only one Hough space is accumulated – the space of the vanishing points. The parameterization of the VPs is based on the PClines line parameterization and it is defined as a mapping of the whole real projective plane to a finite space (the "diamond space").

Our algorithm operates directly on edgelets (Fig. 1), skipping the common step of grouping edges into straight lines or line segments. The parameterization of VPs is in all aspects linear; it involves no goniometric or other non-linear operations and thus it is suitable for implementation in embedded chips and circuitry (Fig. 3). The iterative search scheme allows for finding orthogonal triplets of VPs with high accuracy and low computational costs.

The algorithm builds upon a parameterization of lines for the Hough transform presented by Dubská *et al.* [3] and the Cascaded HT [6].

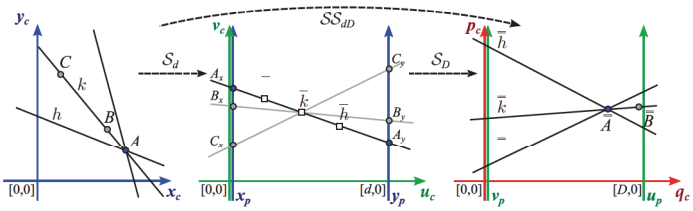


Figure 2: Cascaded PClines transformations via the straight $\mathcal{S}$ spaces (referred as $\mathcal{SS}_{dD}$ mapping). **left:** Original image space with points and lines. **middle:** The same objects in parallel coordinates **right:** Second transformation to parallel coordinates.

Mapping $\mathcal{SS}_{dD}$ (Fig. 2) is transformation of one infinite space to another infinite space. In the case of line detection, the infinite space can be replaced by two finite dual spaces [3, 6]. Dubská *et al.* [3] flip the $\mathbf{y}_p$ axis, put it in $-d$ distance and form a twisted $\mathcal{T}$ space. The CHT based on the PClines parameterization can be done by using all four combinations of the mappings. For each quadrant, a different transformation ($\mathcal{SS}, \mathcal{ST}, \mathcal{TS}$ or $\mathcal{TT}$) is used and mapped to a finite part. These four parts can be attached because images of the axes $\mathbf{x}_c, \mathbf{y}_c$ and the ideal line always lie on the borders of two segments (Fig. 4).

The point mappings between the original plane and the joined diamond space are:

$$[x,y,w]_o \rightarrow [-dDw, -dx, \mathrm{sgn}(xy)x + y + \mathrm{sgn}(y)dw]_d \quad (1)$$

$$[x,y,w]_d \rightarrow [Dy, \mathrm{sgn}(x)dx + \mathrm{sgn}(y)Dy - dDw, x]_o, \quad (2)$$

However, in the joined space, the image of a straight line is not a line anymore. The result of the mapping is a polyline whose number of segments depends on the number of quadrants the line passes. The sequence of endpoints defining the polyline corresponding to line (a,b,c) is in Eq. (3).

$$\alpha = \mathrm{sgn}(ab), \quad \beta = \mathrm{sgn}(bc), \quad \gamma = \mathrm{sgn}(ac)$$
$$\left[\frac{\alpha a}{c+\gamma a}, \frac{-\alpha c}{c+\gamma a}\right], \left[\frac{b}{c+\beta b}, 0\right], \left[0, \frac{b}{a+\alpha b}\right], \left[\frac{-\alpha a}{c+\gamma a}, \frac{\alpha c}{c+\gamma a}\right] \quad (3)$$

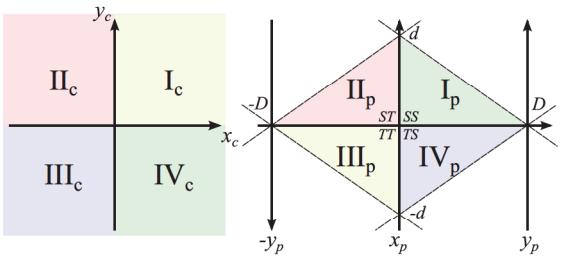


Figure 4: Quadrants of the original infinite Cartesian space. *right:* Quadrants of the PClines space (two attached spaces of parallel coordinates).

The detection accuracy is evaluated on the York Urban Database [2], consisting of 102 images, each with three orthogonal ground truth vanishing points. Our algorithm yields **98.04 %** success rate at 10° angular error tolerance with average error **1.41 %**.

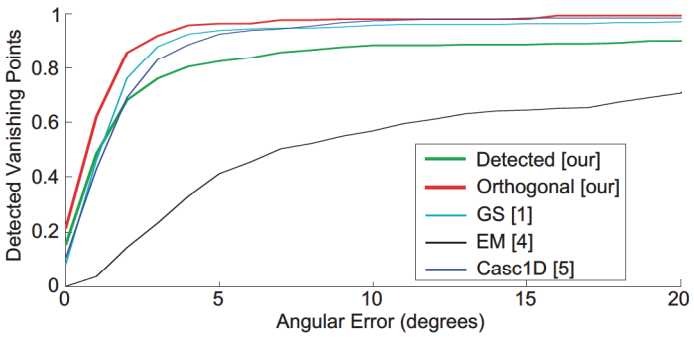


Figure 5: Cumulative histogram of the correctly detected VPs. *Horizontal axis:* angular error of the detected VPs from the ground truth. *Vertical axis:* fraction of VPs detected with the given error tolerance. **green:** Our algorithm without the orthogonalization. **red:** Our detection with the orthogonalization. **GS**, **EM** and **Casc1D** are algorithms used in [1, 4, 5].

[1] S. T. Barnard. Interpreting perspective images. *Artif. Intell.*, 21, November 1983.

[2] P. Denis, J. H. Elder, and F. J. Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *ECCV 2008*.

[3] M. Dubská, A. Herout, and Havel J. PClines – line detection using parallel coordinates. In *CVPR 2011*.

[4] J. Košecká and W. Zhang. Video compass. In *ECCV 2002*.

[5] B. Li, K. Peng, X. Ying, and H. Zha. Vanishing point detection using cascaded 1D hough transform from single images. *Pattern Recognition Letters 2012*.

[6] T. Tuytelaars, M. Proesmans, L. Van Gool, and E. Mi. The cascaded Hough transform. In *Proceedings of ICIP 1998*.

Efficient Shape Matching using Vector Extrapolation

Emanuele Rodolà[1,2]
emanuele.rodola@in.tum.de
Tatsuya Harada[1]
harada@mi.t.u-tokyo.ac.jp
Yasuo Kuniyoshi[1]
kuniyosh@isi.imi.i.u-tokyo.ac.jp
Daniel Cremers[2]
cremers@tum.de

[1] Information Science and Technology, The University of Tokyo
[2] Computer Vision Group, Technical University of Munich

Shape matching is a pervasive problem in computer vision. It concerns, in its general form, the problem of determining a map $f : X \to Y$ among two given shapes X and Y in such a way that their geometrical properties are preserved by the transformation. Arguably one of the most adopted formulations for shape matching takes form as a NP-hard quadratic assignment problem (QAP), where a quadratic term in the objective function encodes a measure of pairwise association among a set $C \subseteq X \times Y$ of putative matches. In particular, let $\mathbf{x} \in [0,1]^{|C|}$ be a (soft) indicator vector representing a set of point-to-point matches among the two shapes, and let $\mathbf{S} \in \mathbb{R}^{|C|\times|C|}$ be a *similarity matrix* among pairs of matches. Then the matching problem can be cast as the quadratic program

$$\max_{\mathbf{x}} \ \mathbf{x}^{\mathrm{T}}\mathbf{S}\mathbf{x} \quad \text{s.t. } \Pi(\mathbf{x}) \preceq \mathbf{c}\,, \tag{1}$$

where $\preceq$ denotes element-wise inequality, $\mathbf{c} \in \mathbb{R}^{|C|}$ is a fixed vector, and Π is a (possibly nonlinear) function specifying the mapping constraints on the correspondence (*e.g.*, one-to-one, one-to-many).

Several attempts at finding good local optima for variants of the QAP have been proposed in computer vision literature. A well-known approach consists in solving a modified problem in which $\|\mathbf{x}\|_2^2 = 1$ replaces the mapping constraints appearing in (1). The method manifests a tendency to assign matches to each point, therefore bringing incorrect correspondences in the final solution even in presence of moderate noise. The more recent techniques based on game theory [1] replace this unit norm constraint by the L_1 counterpart $\|\mathbf{x}\|_1 = 1$, promoting stable, yet very sparse solutions. In order to strike a balance between the two, and therefore capture a broader family of matching scenarios, in this paper we first propose to consider the following family of relaxations for the QAP:

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^{\mathrm{T}}\mathbf{S}\mathbf{x} \\ \text{s.t.} \quad & (1-\alpha)\|\mathbf{x}\|_1 + \alpha\|\mathbf{x}\|_2^2 = 1\,,\ \mathbf{x} \succeq \mathbf{0},\ \alpha \in [0,1]\,. \end{aligned} \tag{2}$$

The convex combination of L_1 and L_2 penalties appearing above takes the name of *elastic net* [4]. This particular formulation gives rise to the so called *grouping effect*, *i.e.*, the joint selection of entire clusters of highly similar matches. This is a desirable feature in most relaxed QAP scenarios, in which one typically seeks only high-precision correspondences in a situation where there is huge ambiguity in the candidate set. A local optimum to problem (2) can be determined by following a projected gradient approach. The optimization process is governed by the equations

$$\mathbf{x}^{(t+1)} = P_\alpha\left(\mathbf{x}^{(t)} + \delta\mathbf{S}\mathbf{x}^{(t)}\right), \tag{3}$$

where P_α is a projection operator taking a solution back to the feasible set. Projected gradient is a natural choice for this family of problems, nevertheless its application in real-world settings is often rendered prohibitive by the large scale of the data.

The major contribution of this paper derives from the observation that, in most non-convex matching scenarios, the intermediate sequence of solutions generated by the optimization algorithm has a tendency to exhibit a certain degree of smoothness. This suggests the possibility to infer the general direction of convergence from previous iterates. To this end, we look at a family of techniques coming under the umbrella term of *vector extrapolation* [3]. These techniques have found application in the context of fixed-point iterative methods for solving linear and nonlinear systems of equations; until now, limited attempts have been made to adopt such methods in computer vision problems [2].

Consider the vector sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ as generated by the non-linear process of Eq. (3). It is often the case that the iterative process

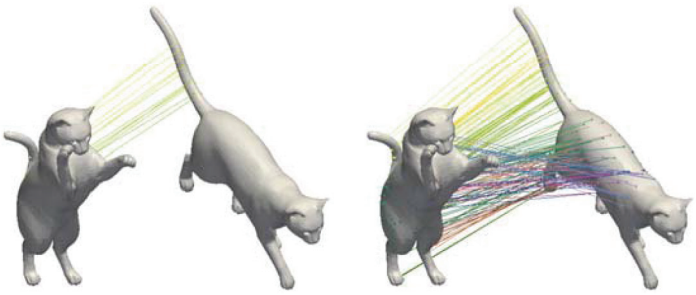


Figure 1: Solutions obtained via vector-extrapolated projected gradient in a problem of non-rigid matching under elastic net constraints. A very selective correspondence is given by $\alpha = 0.1$ (left); a denser, but similarly accurate solution can be obtained with $\alpha = 0.85$ (right).

requires many iterations to reach good accuracy, or that the individual terms $\mathbf{x}^{(t)}$ themselves are expensive to compute due to the action of P_α. We are thus looking for a means to give an estimate of the limit point $\mathbf{s} = \lim_{t\to\infty} \mathbf{x}^{(t)}$ using as few terms as possible. As we show in the paper, such an estimate can be expressed as the linear combination

$$\mathbf{s}^* = \sum_{i=0}^{k} \gamma_i \mathbf{x}_i\,, \tag{4}$$

with $\sum_{i=0}^{k} \gamma_i = 1$. In other words, an approximation to the limit point can be given as a linear combination of $k+1$ past non-linearly generated iterates. In practice, solving for the optimal set of weights γ_i simply amounts to minimizing a least-squares problem subject to linear constraints.

In order to assess the validity of the method, we performed a wide range of experiments on three related computer vision tasks, namely rigid matching of point clouds, non-rigid matching of three-dimensional shapes, and feature matching for multiple-view stereo. From an optimization point of view, these matching scenarios differ in their specific definitions of the similarity term, and typically exhibit energy landscapes with rather different characteristics. This allows us to assess to what extent vector extrapolation may be adopted as a means to accelerate convergence in difficult settings, and whether its introduction into the matching process may lead to premature convergence and thus poor local optima. We note that the general improvement is almost one order of magnitude if compared with an optimization carried out with no vector extrapolation at all. This large increase in performance suggests that there is little reason not to adopt vector extrapolation for this class of problems, and even in those cases where the advantage is limited, vector extrapolation may still be adopted whenever a higher degree of accuracy is required.

[1] A. Albarelli, S. Rota Bulò, A. Torsello, and M. Pelillo. Matching as a non-cooperative game. In *Proc. IEEE Intl. Conf. on Comput. Vis.*, pages 1319 –1326, 29 2009-oct. 2 2009.

[2] G. Rosman, M. M. Bronstein, A. M. Bronstein, and R. Kimmel. Nonlinear dimensionality reduction by topologically constrained isometric embedding. *Int. J. Comput. Vision*, 89(1):56–68, August 2010.

[3] D. A. Smith, W. F. Ford, and A. Sidi. Extrapolation methods for vector sequences. *SIAM Review*, 29(2):pp. 199–233, 1987.

[4] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Incremental Line-based 3D Reconstruction using Geometric Constraints

Manuel Hofer
hofer@icg.tugraz.at
Andreas Wendel
wendel@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

Generating accurate 3D models of man-made objects and urban scenery from an image sequence is a challenging task. Traditional Structure-from-Motion (SfM) approaches often fail because of the high amount of untextured objects and wiry structures present. At the very least, these objects are poorly represented in the resulting point clouds. Since most man-made objects can be approximated by line segments, line-based 3D reconstruction techniques can be used as an alternative. While common appearance-based approaches usually deliver accurate results for a wide range of urban scenery, they cannot be directly applied to wiry structures. Since the resulting matching scores are based on the surroundings of a line segment, explicit matching would fail for these structures due to the changing background. To overcome such limitations, methods which do not rely on appearance-based line segment matching can be applied [1, 3]. Those algorithms, which assume known cameras, are usually based on generating a large set of possible 3D line segment hypotheses, gradient based scoring, and spatial clustering. These are very time consuming steps, which have to be computed offline after all cameras are available and oriented correctly.

We propose a novel line-based 3D modelling approach, which extends the principles presented in [1, 3] by incremental hypotheses clustering and geometric verification steps, without the need of time consuming scoring in the image space. We demonstrate how fusing this approach with an incremental point-based SfM [2] leads to an online 3D reconstruction method, which is able to cover wiry- and repeated structures, as well as solid objects.

To perform incremental SfM we need to have an initial geometry involving at least two views. Therefore, given two images I_1 and I_2 and their respective sets of 2D line segments, we create an initial set of 3D line segment hypotheses H by computing all possible line segment matches between the two images. To limit the number of potential matches, we exploit epipolar constraints using the corresponding cameras C_1 and C_2. For each putative match we compute a 3D line segment K_h by triangulating the corresponding 2D line segments from I_1 and I_2. Each match results in a new hypothesis $h \in H$. Each hypothesis has a score $s(h)$ and a corresponding camera $C^*(h)$ defined as

$$s(h) = 1 - min\left\{\left|\left\langle \frac{\overrightarrow{K_h}}{\|\overrightarrow{K_h}\|}, \frac{\overrightarrow{C_i}}{\|\overrightarrow{C_i}\|} \right\rangle\right|\right\}, \quad C^*(h) = \underset{C_i}{\operatorname{argmax}}(s(h)) \qquad (1)$$

where $\overrightarrow{K_h}$ is the directional vector of K_h, and $\overrightarrow{C_i}$ denotes the camera ray of camera C_i, and $\langle \cdot, \cdot \rangle$ is the inner product.

To perform incremental hypothesis merging for further incoming images, we need to define a spatial grouping radius r_{space}, which we derive from the image space dynamically to be scale invariant. Therefore, we define a maximum uncertainty σ in the image space. To bring this value to 3D space, we first compute a specific grouping radius $r_{space}(h)$ for each hypothesis $h \in H$. Therefore, we project the 3D line segment K_h back into the two supporting images. We then shift the resulting 2D line segments in the same orthogonal direction by σ, and triangulate them to obtain a shifted 3D line segment $\hat{K}_h$. The radius $r_{space}(h)$ is defined as the maximum distance between $\hat{K}_h$, and the infinite line passing through K_h. To be robust against imprecise triangulation, we compute a characteristic grouping radius $r_{space}(C_i)$ for each view, by using the median of all referenced radii. This allows us to adapt the system to severe viewpoint changes.

When a new image I_i is available we integrate it into our existing reconstruction, based on the current set of previously computed images. Therefore we have to define a set of neighboring views $N(I_i)$ for I_i. As above, we compute all possible matches for each line segment l, with the segments in $N(I_i)$. For each possible correspondence we try to add l to an existing hypothesis. We create a triangulated 3D line segment and compute the spatial distances to all candidate hypotheses. If we find a hypothesis h for which the spatial distance is below r_{space}, we consider l to be a part of h, and re-compute the score $s(h)$ and update the corresponding view $C^*(h)$. The estimated 3D line segment K_h has to be adapted as well, incorporating the newly triangulated line segment. For each possible match, for which we cannot find an existing hypothesis to be added to, we create a new hypothesis in the same way as during initialization.



(a) (b) (c)

Figure 1: (a) An example image from the *Pylon* sequence by [1] (106 images). (b) Reconstruction obtained by [1] (67 minutes, lines only). (c) Our reconstruction, $M = 10$, $\lambda = 4$ (9 minutes, including SfM).

After all line segments have been matched, we compute the current inlier set. Therefore, we sort the hypotheses $h \in H$ descending by the number of supporting line segments. If two hypotheses have the same number of line segments, we order them according to their reprojection error. We compute the current inlier set by iterating over the sorted hypotheses set. For h to be an inlier the following criteria have to be fulfilled: the number of supporting line segments has to be at least λ, and the score $s(h)$ has to be higher than 0.5. If this holds, the hypothesis is valid and all other hypotheses related to any of the segments referenced by h, are skipped during the iteration. If the validity criteria are not satisfied, the hypothesis is considered to be an outlier. This incremental grouping procedure prevents evaluation of a very large set of hypotheses at the end of the algorithm. However, using our proposed method might as well produce a huge hypotheses set for large image sequences. Hence, we need to remove unpromising hypotheses from time to time. To achieve this, we evaluate the number of supporting line segments in a hypothesis h compared to the number of views, which have been matched with the corresponding view $C^*(h)$. If there are less than λ line segments that agree on h, and $C^*(h)$ has been matched with at least $2 \cdot \lambda$ views, then hypothesis h is permanently removed from the hypothesis set.

Figure 1 shows results for a wiry structure, using the *Pylon* sequence from our previous work [1]. As we can see, our new approach has even less outliers due to the improved scoring approach and the automatic grouping radius selection. Additionally we manage to reconstruct the scene significantly faster, even though we also perform pose estimation, while in [1] the cameras are assumed to be known beforehand. For a more extensive evaluation about the parameters and additional testcases, we kindly refer to the full paper.

[1] M. Hofer, A. Wendel, and H. Bischof. Line-based 3D reconstruction of wiry objects. *Computer Vision Winter Workshop*, 2013.

[2] C. Hoppe, M. Klopschitz, M. Rumpler, A. Wendel, S. Kluckner, H. Bischof, and G. Reitmayr. Online feedback for structure-from-motion image acquisition. *British Machine Vision Conference*, 2012.

[3] A. Jain, C. Kurz, T. Thormaehlen, and H. Seidel. Exploiting global connectivity constraints for reconstruction of 3D line segments from images. *Conference on Computer Vision and Pattern Recognition*, 2010.

Spline Fusion: A continuous-time representation for visual-inertial fusion with application to rolling shutter cameras

Steven Lovegrove
slovegrove@gwu.edu
Alonso Patron-Perez
apatron@gwu.edu
Gabe Sibley
gsibley@gwu.edu

Department of Computer Science
The George Washington University
Washington, DC, USA

In this paper, we describe a method for performing SLAM and visual-inertial calibration robustly using inexpensive sensors such as rolling shutter CMOS cameras and MEMS IMUs. We make use of a continuous-time model for the trajectory of the camera that naturally allows us to fuse information from many unsynchronized and potentially high-rate sensors whilst limiting state size. We model the rolling shutter of a camera explicitly and can form errors generatively on inertial measurements. This model is not limited to visual-inertial SLAM and may also simplify integration of other sensors such as spinning SICK Laser rangers.

At the heart of our approach lies a continuous trajectory representation similar to the one presented in [2]. We chose a formulation which offers:

- Local control, allowing the system to function online as well as in batch.
- C^2 continuity, to enable us to predict IMU measurements.
- A good approximation of minimal torque trajectories.

Cubic B-Splines are a well-known representation for trajectories in $\mathbb{R}^3$, but are not so easily applied when dealing with 3D rotations, such as interpolation in $\mathbb{SO}3$. For example, C^2 continuity is not necessarily preserved [3]. We choose to parameterize a continuous trajectory using cumulative basis functions formed using the Lie Algebra, equivalent to that proposed in [1]. Using cumulative B-Spline basis functions were first proposed for quaternion interpolation in [4] in the context of computer animation. This representation is not only C^2 continuous, but it also provides a very simple second derivative formulation.

The cumulative B-Spline parameterization enables the computation of analytical time derivatives at any point in the spline. This allows us to trivially synthesize accelerometer and gyroscope measurements, which we can in turn use to form direct errors on observed measurements.

Another advantage of a continuous-time framework is in dealing with rolling shutter cameras. Although the projective geometry of a rolling shutter camera remains the same as that of a global shutter camera, every line of the image is exposed for a different period, each one more delayed than the last. When the camera is in motion, this can cause the image to appear distorted and skewed (Figure 2a). Using a continuous-time model for the motion of the camera, we are free to treat every line of the image as its own exposure as shown in Figure 1.

Given generative models for visual and inertial data, we can solve for spline and camera parameters in batch or over a window by minimizing an objective function formed from the difference of measured to predicted observations. By using a continuous-time formulation, reprojection errors and inertial errors can be treated uniformly, weighted by their respective information matrices computed from device specifications or calibration.

We have conducted experiments in both simulation and with real data to evaluate our flexible continuous-time approach. We present sliding window visual odometry results on a simulated monocular rolling shutter dataset (Figure 2) and then continue by demonstrating our system on real data for joint visual-inertial SLAM and self-calibration (Figure 3).

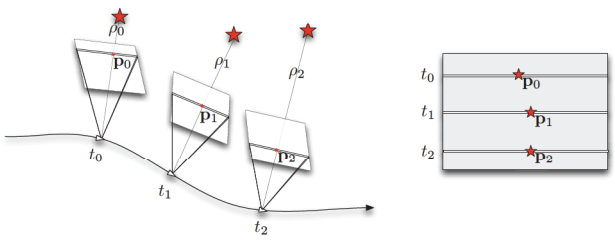
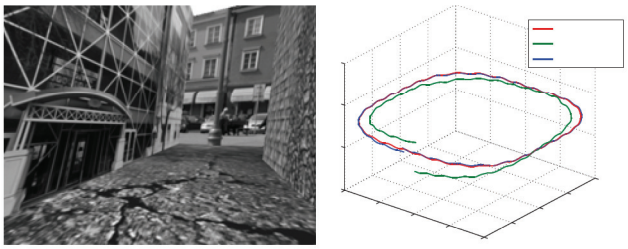


Figure 1: Rolling shutter cameras are easily modeled with continuous-time SLAM. Landmarks observed at pixel locations $\mathbf{p}_i$ are represented by their inverse depth, ρ_i and time of measurement, t_i. Each scanline is effectively a single push-broom camera (left); such scanline-camera measurements are captured over time, which defines the image returned by the actual sensor (right).



(a) (b)

Figure 2: Results from simulated monocular rolling shutter experiment. (a) Still from simulated sequence exhibiting large rolling shutter warp. (b) Comparison of ground truth trajectory (red) against estimated trajectories when modelling (blue) and not modelling (green) rolling shutter.

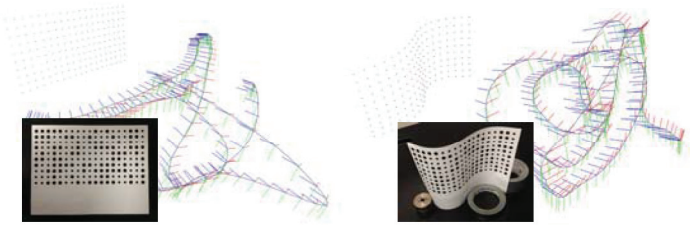


Figure 3: Example estimates computed for full-SLAM and self-calibration including camera-to-IMU, IMU biases, gravity vector, camera intrinsics (with distortion), 3D landmark locations and sensor trajectory. Again, the structure and scale in these figures are accurately estimated without any prior knowledge of the scene. This being our first experimental analysis, full and robust front-end feature tracking has not been completed, and the grid is only used to simplify data-association. This still allows evaluation and demonstration of full SLAM and visual-inertial self-calibration.

[1] P. Crouch, G. Kun, and F. Silva Leite. The de casteljau algorithm on lie groups and spheres. *Journal of Dynamical and Control Systems*, 5(3):397–429, July 1999.

[2] P. Furgale, T. D. Barfoot, and G. Sibley. Continuous-time batch estimation using temporal basis functions. In *International Conference on Robotics and Automation*, 2012.

[3] M-J. Kim, M.-S. Kim, and S. Shin. A c^2-continuous b-spline quaternion curve interpolating a given sequence of solid orientations. In *Computer Animation*, 1995.

[4] M-J. Kim, M-S. Kim, and S. Shin. A general construction scheme for unit quaternion curves with simple high order derivatives. In *SIGGRAPH*, 1995.

Incremental Surface Extraction from Sparse Structure-from-Motion Point Clouds

Christof Hoppe[1]
hoppe@icg.tugraz.at
Manfred Klopschitz[2]
manfred.klopschitz@siemens.com
Michael Donoser[1]
michael.donoser@tugraz.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision,
Graz University of Technology,
Graz, Austria
[2] Imaging and Computer Vision
Research Group Video Analytics
Corporate Technology
Siemens AG Austria, Graz

In this paper we propose a new method to incrementally extract a surface from a consecutively growing Structure-from-Motion (SfM) point cloud in real-time. Our method is based on a Delaunay triangulation (DT) on the 3D points. The core idea is to robustly label all tetrahedra into free- and occupied space using a random field formulation and to extract the surface as the interface between differently labeled tetrahedra.

For this reason, we propose a new energy function that achieves the same accuracy as state-of-the-art methods but reduces the computational effort significantly. Furthermore, our new formulation allows us to extract the surface in an incremental manner, i. e. whenever the point cloud is updated we adapt our energy function. Instead of minimizing the updated energy with a standard graph cut, we employ the dynamic graph cut of Kohli et al. [1] which enables efficient minimization of a series of similar random fields by re-using the previous solution. In such a way we are able to extract the surface from an increasingly growing point cloud nearly independent of the overall scene size.

Energy Function for Surface Extraction Our method formulates surface extraction as a binary labeling problem, with the goal of assigning each tetrahedron either a *free* or *occupied* label. For this reason, we model the probabilities that a tetrahedron is free- or occupied space analyzing the set of rays that connect all 3D points to image features. Following the idea of the truncated signed distance function (TSDF), which is known from voxel-based surface reconstructions, a tetrahedron in front of a 3D point X has a high probability to be free space, whereas a tetrahedron behind X is presumably occupied space. We further assume that it is very unlikely that neighboring tetrahedra obtain different labels, except for pairs of tetrahedra that have a ray through the face connecting both. Such a labeling problem can be elegantly formulated as a pairwise random field and since our priors are submodular, we can efficiently find a global optimal labeling solution e. g. using graph cuts. In contrast to existing methods like [2], our energy depends only on the visibility information that is directly connected to the four 3D points that span the tetrahedra V_i. Hence a modification of the tetrahedral structure by inserting new points has only limited effect on the energy function. This property enables us to easily adopt the energy function to a modified tetrahedral structure.

Incremental Surface Extraction To enable efficient incremental surface reconstruction, our method has to consecutively integrate new scene information (3D points as well as visibility information) in the energy function and to minimize the modified energy efficiently.

Integrating new visibility information, i. e. adding rays for newly available 3D points, affects only those terms of the energy function that relate to the 3D point, which are typically only a few. Integrating additional 3D points is more complicated, since it affects the Delaunay triangulation, i. e. new tetrahedra are created and others are deleted. For the update of the energy this implies that certain terms are deleted and new terms are added. Since our energy depends only on rays that are directly connected to a certain tetrahedron, the new costs can be calculated efficiently.

The consecutive integration of additional information leads to a series of energies that typically only differ by a few terms. Instead of solving the labeling each time from scratch using a standard graph cut, we employ the dynamic graph cut proposed by Kohli et al. [1]. In our experiments, we show that the complexity of the minimization does not depend on the overall number of terms but on the number of terms that differ between two subsequent energies. Hence, our surface extraction algorithm is nearly independent of the overall scene size.

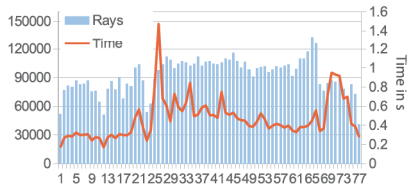


Figure 1: Runtime analysis for integrating 1000 new 3D points per step. Although the mesh gets larger, computation time remains independent of the overall scene size.

Experiments In the experimental evaluation, we first compare our approach to the state-of-the-art in static surface extraction from sparse point clouds and demonstrate competitive performance while reducing the computational effort by more than 50%. We furthermore demonstrate that our method is able to incrementally extract the surface from a growing sparse point cloud robustly in real-time while being nearly independent of the overall scene size (Figure 1). Our approach used in an incremental manner requires on average 440 ms to integrate 1000 new 3D points into the existing surface which is up to 20 times faster than the state-of-the-art. Figures 2 and 3 illustrate how the mesh evolves over time for two different point clouds. Our promising prove the applicability of our method for efficiently obtaining surface meshes from consecutively growing sparse point clouds, as e. g. might be provided by a keyframe-based SLAM method.

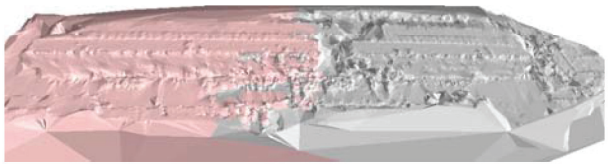


Figure 2: Surface evolution over time. The gray part of the mesh has been extracted from 40 000 3D points. The red triangles are created by integrating 20 000 further points.



(a) (b) (c)

Figure 3: (a) Input image for SfM (b) Surface mesh after integration of 1000 3D points (c) Final surface mesh

[1] P. Kohli and P.H.S. Torr. Dynamic graph cuts for efficient inference in markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(12):2079–2088, December 2007.

[2] P. Labatut, J.P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *International Conference on Computer Vision (ICCV)*, 2007.

Dense, Auto-Calibrating Visual Odometry from a Downward-Looking Camera

Jacek Zienkiewicz
j.zienkiewicz12@imperial.ac.uk
Robert Lukierski
r.lukierski12@imperial.ac.uk
Andrew Davison
a.davison@imperial.ac.uk

Department of Computing
Imperial College London
London, UK

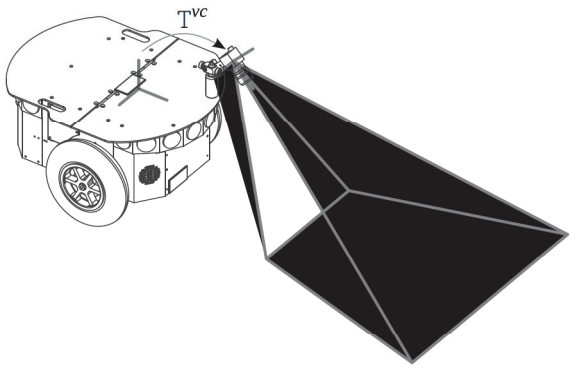


Figure 1: Robot and camera set-up; the camera view frustum intersects the ground plane.

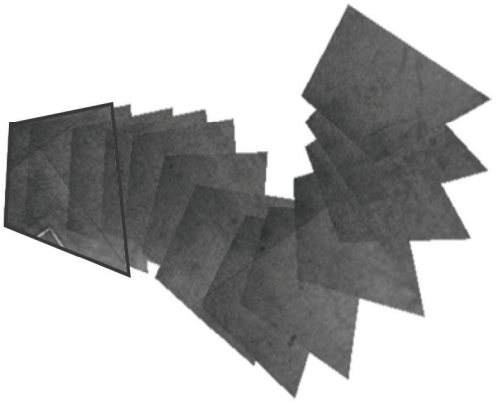


Figure 3: Keyframes registered into a local mosaic.

Thursday 09:40

Computer vision is now widely considered a suitable technology for mobile robot navigation, but few practical mass-market solutions rely solely on cameras for outward-looking sensing. In this paper we present a technique whereby a single camera can be used as a high precision visual odometry sensor in a range of practical settings using simple, computationally efficient techniques. Taking advantage of the local planarity of common floor surfaces, we use real-time dense alignment of a 30Hz video stream as the camera looks down from a fast-moving robot. By using every image pixel rather than sparse features to contribute to a global motion estimate, we find that almost every surface has trackable texture when observed from close enough, and we are able to show robust and accurate real-time performance on a range of normal poorly-textured surfaces such as carpet, vinyl and wood.

The main part of our method uses a whole image alignment approach evolved from the iterative technique introduced by Lucas and Kanade [1]. We define an energy function F which measures the discrepancy between a reference image $\mathcal{I}^r$, and a live image $\mathcal{I}^l$ warped by a plane-induced homography H^{lr}. The homography is parameterised by the incremental planar motion of a robot:

$$F(\mathbf{x}) = \frac{1}{2}\|\mathcal{I}^l(\boldsymbol{\pi}(\mathrm{H}^{lr}(\mathbf{x})\mathbf{p})) - \mathcal{I}^r(\boldsymbol{\pi}(\mathbf{p}))\|_2^2 \,. \tag{1}$$

In order to minimise $F(\mathbf{x})$ we perform an iterative optimisation based on Efficient Second-order Minimisation (ESM) [2, 3]. An example plot of the cost function with respect to the translational degrees of freedom of the robot can be seen in Fig. 2.

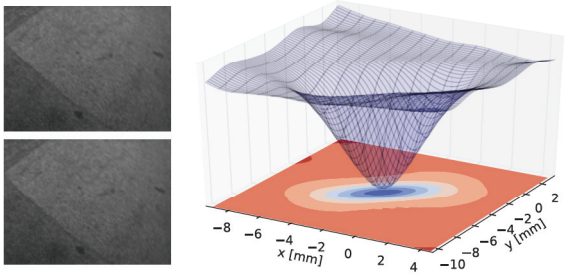


Figure 2: Typical dense tracking cost function.

Our key novelty, and crucial to the practicality of this approach, is rapid and automatic calibration for 6DoF camera extrinsics relative to the robot frame. This procedure has two main steps. First, based purely on the image sequence and the knowledge that the sequence is obtained from planar motion, we can determine the normal of the ground plane relative to the camera, which is equivalent to recovering its elevation and roll angles. An example of a dense local map created as part of the auto-calibration process is shown in Fig. 3. Second, we use relative motion estimates from an external source to calibrate for the remaining parameters, which describe the position of the camera relative to the robot: its yaw angle and metric 3DoF translational position.

Our real-time visual odometry is implemented almost entirely in parallel and achieve a mean frame-rate of about 270–300 FPS on a desktop GPU and about 60–65 FPS on a mid-range laptop GPU.

We experimentally evaluated our visual odometry approach against accurate odometry by making extended runs of the robot with different camera heights on a variety of different surfaces (Fig. 4). Our experiments show unbiased estimation of robot motion and a good performance of the algorithm most of the time. At high linear and angular velocities we see some tendency to underestimation of the velocity resulting mostly from the reduced inter-frame overlap and motion blur. Tracking failure usually only happens though when these effects are combined with disadvantageous other issues such as shadows cast by the robot.

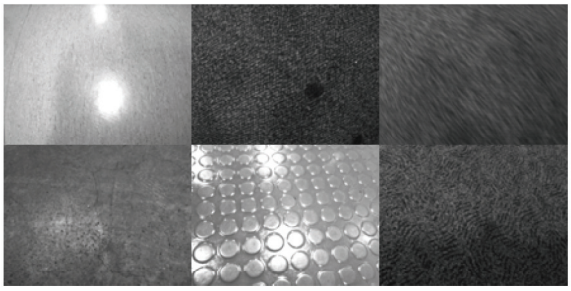


Figure 4: Snapshots of different surfaces used for evaluation.

[1] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1981.

[2] E. Malis. Improving vision-based control using efficient second-order minimization techniques. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2004.

[3] C. Mei, S. Benhimane, E. Malis, and P. Rives. Efficient Homography-Based Tracking and 3-D Reconstruction for Single-Viewpoint Sensors. *IEEE Transactions on Robotics (T-RO)*, 24(6):1352–1364, 2008.

A Novel Trilateral Filter based Adaptive Support Weight Method for Stereo Matching

Dongming Chen
http://liris.cnrs.fr/membres?idn=dchen
Mohsen Ardabilian
http://liris.cnrs.fr/membres?idn=mardabil
Liming Chen
http://perso.ec-lyon.fr/liming.chen/index.htm

LIRIS UMR 5205
Ecole Centrale de Lyon
Lyon, France

The performance of local stereo matching algorithm highly depends on the support window selected, in which the cost are aggregated. A variety of cost aggregation approaches (proposed before 2008) were comprehensively analyzed in [7] and these approaches attempt to seek an optimal support window for each pixel by changing the window size, shape and center offset. The ideal optimal window should satisfy the rule that all pixels in this window lie on the same disparity with the center pixel.

Recent years have witnessed a great deal of attention focused on the Adaptive Support Weight (ASW) based methods [1, 4, 6, 8], proposed firstly by Yoon and Kweon in [8]. The ASW methods assign an adaptive weight to each pixel of the support window, depending on how it is likely to lie on the same disparity with the center pixel. Essentially, the assignment of an adaptive weight amounts to changing the support window in terms of size, shape or center offset. In ASW methods, the weight function is very important, because it directly decides the support window. The weight function proposed in [8] is based on bilateral filter. Following this pioneering work, various weight function were proposed, including in particular the segmented bilateral filter weight function [6], the geodesic weight function [1], the guided filter weight function [4]. Thus, which weight function is the most accurate one? Recently, Hosni *et al.* [2] carried out a comprehensive comparative study to fairly evaluate various weight functions while fixing the preprocessing, matching cost function and post-processing. Their conclusion is that both bilateral filter weight function [8] and guided filter weight function [4] are the best, since bilateral filter weight function performs better on the average rank while guided filter weight function produces a lower average error.

We revisit the bilateral filter weight function [8], which obeys two rules that, given a support pixel, (1) if its color is similar to the center pixel's and (2) if it is spatially close to the center pixel, it is likely to lie on the same disparity with the center pixel. Therefore, the bilateral filter weight function consists of two parts, color similarity term and spatial proximity term, defined as:

$$w_{bf}(p,q) = e^{-\frac{\Delta c_{pq}}{\gamma_c}} e^{-\frac{\Delta g_{pq}}{\gamma_g}}, \tag{1}$$

where q is a pixel within the support window centered at pixel p. The color similarity Δc_{pq} represents the Euclidean distance between the color of these two pixels, measured in the CIELab color space as

$$\Delta c_{pq} = \sqrt{\sum_{j\in(L,a,b)} (I_j(p) - I_j(q))^2} \tag{2}$$

and the geometric proximity Δg_{pq} is the Euclidean distance between their coordinates (x,y) as

$$\Delta g_{pq} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}. \tag{3}$$

The parameter γ_c and γ_g are set by user to adjust the color similarity term and geometric proximity term respectively.

While these two rules can handle most depth ambiguities within a support window, it unfortunately fail to resolve the ambiguity induced by nearby pixels at different disparities but with similar colors as illustrated in Figure 1. (a) is the reference image and we focus on the regions within the red box, zoomed in (b). In (b), there are two nearby planks with similar colors but in different disparities. An imaginary situation is presented in (c): these two planks are substituted by one cross-shaped plank in the same disparity. The bilateral filter weight of pixel p and q in (b) is equal to that in (c), because their color similarity and spatial distance are the same. But obviously these two weights should not be equal because the two pixels are in the same disparity in (c) but not in (b). We can observe

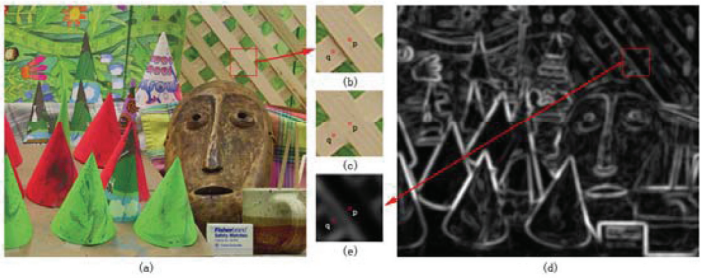


Figure 1: (a) reference image. (b) zoomed real situation. (c) zoomed imaginary situation. (d) boundary cue (e) zoomed boundary cue.

that the depth discontinuity between pixel p and q in (b) induces a color discontinuity between these two planks and results in color boundaries nearby, as shown in (e). Therefore, the boundary cue is helpful to obtain a more faithful weight. As a result, we propose in this paper a trilateral filter weight function which extends the bilateral filter weight function by a new boundary strength term that measures the boundary strength between two pixels, defined as

$$w_{tf}(p,q) = e^{-\frac{\Delta c_{pq}}{\gamma_c}} e^{-\frac{\Delta g_{pq}}{\gamma_g}} \left(e^{-\frac{\Delta E_{pq}}{\gamma_e}} + e^{-\frac{\Delta c_{pq}}{\gamma_c}} e^{-\frac{\Delta g_{pq}}{\gamma_g}} \right), \tag{4}$$

where $exp(-\Delta E_{pq}/\gamma_e)$ is the boundary strength term and parameter γ_e is set by user. The boundary strength at pixel p is defined as [3],

$$E(p) = \sum_{\theta} \sqrt{(I(p) * F_{\theta,odd})^2 + (I(p) * F_{\theta,even})^2}, \tag{5}$$

where $*$ denotes the convolution operator; the odd-phase filter $F_{\theta,odd}$ and even-phase filter $F_{\theta,even}$ are a pair of quadrature filters in orientation θ.

Finally, we evaluated our method on the Middlebury benchmark [5], using four pairs of standard data sets, and fairly compared the proposed trilateral filter weight function with the guided filter weight function [4] and bilateral filter weight function [8]. The proposed method ranks 15^{th} out of 150 submissions and is the current most accurate local stereo matching algorithm on the benchmark.

[1] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann. Local stereo matching using geodesic support weights. In *ICIP*, pages 2069–2072, 2009.

[2] A. Hosni, M. Bleyer, and M. Gelautz. Secrets of adaptive support weight techniques for local stereo matching. *CVIU*, 117:620–632, 2013.

[3] M.C. Morronea and R.A. Owens. Feature detection from local energy. *PRL*, 6:303–313, 1987.

[4] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, pages 3017–3024, 2011.

[5] D. Scharstein and R. Szeliski. Middlebury stereo evaluation - version 2. http://vision.middlebury.edu/stereo/.

[6] F. Tombari, S. Mattoccia, and L. Di Stefano. Segmentation-based adaptive support for accurate stereo correspondence. In *PSIVT*, volume 1, pages 427 – 438, 2007.

[7] F. Tombari, S. Mattoccia, L. Di Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. In *CVPR*, 2008.

[8] K. Yoon and I. Kweon. Adaptive support-weight approach for correspondence search. *PAMI*, 28(4):650–656, 2006.

Exploring Cross-Channel Texture Correlation for Color Texture Classification

Xianbiao Qi[1]
qixianbiao@gmail.com
Yu Qiao[2]
yu.qiao@siat.ac.cn
Chun-Guang Li[1]
lichunguang@bupt.edu.cn
Jun Guo[1]
guojun@bupt.edu.cn

[1] Beijing University of Posts and Telecommunications, P.R. China
[2] Shenzhen Key Lab of CVPR, Shenzhen Institute of Advanced Technology, P.R. China

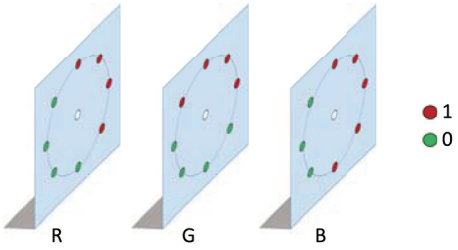


Figure 1: An illustration of LBPs in RGB channels.

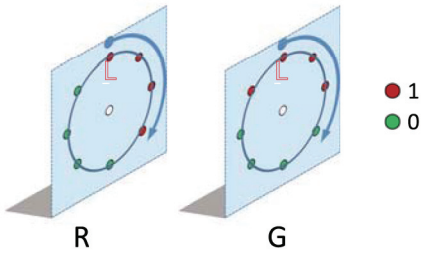


Figure 2: An illustration of encoding strategy between two different color channels.

Color and Texture are two important aspects of natural images. It is widely recognized that they provide strong complementary cues to each other in a lot of computer vision applications, such as object recognition, flower recognition, texture classification, material recognition, content-based image retrieval, color texture segmentation and many more.

Recently, Local Binary Pattern (LBP) descriptor [4] and its variants [1, 2, 5, 7] have been widely applied on texture relevant tasks due to their computational efficiency and great texture discriminative ability. In [5], Timo et al. systematically introduced the LBP operator and its several variants including uniform LBP and rotation invariant LBP. Since then, LBP has been widely applied to a lot of tasks, such as face recognition, image retrieval, and many more [6]. Recently, in [1], Zhenhua et al. proposed a completed LBP operator, which further enriches the original LBP feature.

The simplest and most direct method to incorporate color cue to the texture classification task is to extract LBP histograms from each color channel ,and then concatenate the histograms into a final image representation. In the literature, this method has been used in several works [3, 8]. In [3], Maenpaa et al. used color LBPs (different color space) for texture classification, and in [8], Zhu et al. proposed to use color LBPs for object recognition.

However, the above color and texture fusing approach approach just treats each channel independently, and ignores the cross-channel texture correlation information. In fact, there exists strong correlation between texture patterns of different color channels. Such a correlation reflects the common character and the difference between patterns from different channels. Fig. 1 shows a typical structure in RGB color channels. The LBP sequences in RGB channels are greatly correlative to each other. Encoding the correlation effectively can greatly boost the discriminative ability of the feature.

In this paper, we propose a novel method to encode the cross-channel texture correlation to conduct color texture classification task. Firstly, we quantitatively study the texture correlation between different color channels using LBP as texture descriptor and using the Shannon's information entropy as correlation measurement. Based on the study, we find that (R, G) channel pair have stronger texture correlation than (R, B) and (G, B) channel pairs. Secondly, we propose a novel approach to encode the cross-channel texture correlation. The proposed method captures well the relative variance of the texture patterns between different channels. In practice, we divide three color channels into three pairs and encode any one pair each time. The histograms of three pairs are finally concatenated into a final image representation. The proposed method is computationally efficient and robust to image rotation. Its computational cost is similar to the classical multi-channel color LBP. Meanwhile, the novel encoding strategy of the proposed method can guarantee rotation invariance.

Encoding Strategy: Take (R,G) channels as an example, as shown in Fig. 2, the Cross-Channel LBP (*CCLBP*) can be defined as follows:

$$CCLBP_{R,G}(n,r|c_R,c_G) = [LBP^{riu}_R(n,r|c_R), LBP^{u}_G(n,r|c_G,i)]_{co}, \quad (1)$$

Where

$$i = \arg\max_i \{ROR(LBP_R(n,r,|c_R),i) \mid i = 0,1,....,P1-1\}, \quad (2)$$

According to Eq. 1, the dimension of joint encoding of two channels is $10 \times 59 = 590$. In this paper, we use dense sampling to calculate the histogram. For each pixel in an image, we accumulate their joint patterns into the corresponding bin of histogram. Finally, we normalize the histogram feature with L_1 norm. The histograms of (R, B) and (G, B) can be built similarly. Thus, by concatenating these histograms of pairs of crossing-channel features, we obtain the whole *CCLBP* representation for the image as:

$$CCLBP_{R,G,B} = [CCLBP_{R,G}, CCLBP_{R,B}, CCLBP_{G,B}]. \quad (3)$$

in which the dimension of the $CCLBP_{R,G,B}$ is $590 \times 3 = 1770$.

[1] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *TIP*, 19(6): 1657–1663, 2010.

[2] Zhenhua Guo, Lei Zhang, and David Zhang. Rotation invariant texture classification using lbp variance (lbpv) with global matching. *Pattern Recognition*, 43(3):706–719, 2010.

[3] Topi Mäenpää and Matti Pietikäinen. Classification with color and texture: jointly or separately? *Pattern Recognition*, 37(8):1629–1640, 2004.

[4] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[5] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002.

[6] M Pietikäinen, A Hadid, G Zhao, T Ahonen, M Pietikäinen, A Hadid, G Zhao, and T Ahonen. Computer vision using local binary patterns. 2011.

[7] Xianbiao Qi, Rong Xiao, Jun Guo, and Lei Zhang. Pairwise rotation invariant co-occurrence local binary pattern. In *ECCV*, pages 158–171. Springer, 2012.

[8] Chao Zhu, C-E Bichot, and Liming Chen. Multi-scale color local binary patterns for visual object classes recognition. In *ICPR*, pages 3065–3068, 2010.

Thursday 10:20

Depth really Matters: Improving Visual Salient Region Detection with Depth

Karthik Desingh[1]
http://researchweb.iiit.ac.in/~karthik.d/
K. Madhava Krishna[1]
http://www.iiit.ac.in/~mkrishna/
Deepu Rajan[2]
http://www.ntu.edu.sg/home/ASDRajan/
C.V. Jawahar[1]
http://www.iiit.ac.in/~jawahar/

[1] IIIT - Hyderabad
Hyderabad, India
[2] Nanyang Technological University
Singapore

Depth information has been shown to affect identification of visually salient regions in images. In this paper, we investigate the role of depth in saliency detection in the presence of (i) competing saliencies due to appearance, (ii) depth-induced blur and (iii) centre-bias. Having established through experiments that depth continues to be a significant contributor to saliency in the presence of these cues, we propose a 3D-saliency formulation that takes into account structural features of objects in an indoor setting to identify regions at salient depth levels. Computed 3D-saliency is used in conjunction with 2D-saliency models through non-linear regression using SVM to improve saliency maps. Experiments on benchmark datasets containing depth information show that the proposed fusion of 3D-saliency with 2D-saliency models results in an average improvement in ROC scores of about 9% over state-of-the-art 2D saliency models.

The main contributions of this paper are: (i) The development of a 3D-saliency model that integrates depth and geometric features of object surfaces in indoor scenes. (ii) Fusion of appearance (RGB) saliency with depth saliency through non-linear regression using SVM. (iii) Experiments to support the hypothesis that depth improves saliency detection in the presence of blur and centre-bias. The effectiveness of the 3D-saliency model and its fusion with RGB-saliency is illustrated through experiments on two benchmark datasets that contain depth information. Current state-of-the-art saliency detection algorithms perform poorly on these datasets that depict indoor scenes due to the presence of competing saliencies in the form of color contrast. For example in Fig. 1, saliency maps of [1] is shown for different scenes, along with its human eye fixations and our proposed saliency map after fusion. It is seen from the first scene of Fig. 1, that illumination plays spoiler role in RGB-saliency map. In second scene of Fig. 1, the RGB-saliency is focused on the cap though multiple salient objects are present in the scene. Last scene at the bottom of Fig. 1, shows the limitation of the RGB-saliency when the object is similar in appearance with the background.

Effect of depth on Saliency: In [4], it is shown that depth is an important cue for saliency. In this paper we go further and verify if the depth alone influences the saliency. Different scenes were captured for experimentation using Kinect sensor. Observations resulted out of these experiments are (i) Humans fixate on the objects at closer depth, in the presence of visually competing salient objects in the background, (ii) Early attention happens on the objects at closer depth, (iii) Effective fixations are high at the low contrast foreground compared to the high contrast objects in the background which are blurred, (iv) Low contrast object placed at the center of the field of view, gets more attention compared to other locations. As a result of all these observations, we develop a 3D-saliency that captures the depth information of the regions in the scene.

3D-Saliency: We adapt the region based contrast method from Cheng *et al.* [1] in computing contrast strengths for the segmented 3D surfaces or regions. Each segmented region is assigned a contrast score using surface normals as the feature. Structure of the surface can be described based on the distribution of normals in the region. We compute a histogram of angular distances formed by every pair of normals in the region. Every region R_k is associated with a histogram H_k. Contrast score C_k of a region R_k is computed as the sum of the dot products of its histogram with histograms of other regions in the scene. Since the depth of the region is influencing the visual attention, the contrast score is scaled by a value Z_k, which is the depth of the region R_k from the sensor. In order to define the saliency, sizes of the regions i.e. the number of the points in the region, have to be considered. We find the ratio of the region dimension to the half of the scene dimension. Considering n_k as the number of 3D points in the region R_k, the constrast score becomes

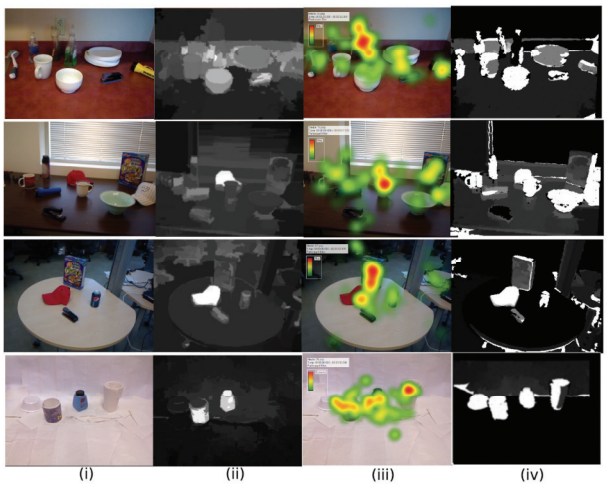


Figure 1: Four different scenes and their saliency maps; For each scene from top left (i) Original Image, (ii) RGB-Saliency map using RC [1], (iii) Human fixations from eye-tracker and (iv) Fused RGBD-saliency map

$$C_k = \frac{2Z_k n_k \sum_{j \neq k} D_{kj}}{\sum_j n_j} \quad (1)$$

where D_{kj} is the dot product between histograms H_k and H_j. The region with less C score is considered to be the one that is unique in the 3D scene. Hence saliency of the region R_k becomes $S_k = 1 - C_k/C_{max}$, where C_{max} is the maximum contrast score in the scene for a region. With the obtained 3D-saliency map, we fuse saliency maps given by the state-of-the-art algorithms to obtain the RGBD-saliency map.

Fusion process: The obtained 3D-saliency map is fused with the saliency maps of existing visual saliency models to create RGBD-saliency. We perform this operation using SVM regression model trained on set of sample images. This fusion process also includes additional local features of the 3D region in the training. They are *Color histogram*, *Contour compactness*, *Dimensionality*, *Perspective score*, *Discontinuities with neighbours*, *Size and location*, and *Verticality*. Inclusion of these features improves the performance by more than 5%. Results are quantified across three datasets, University of Washington RGB-D dataset [3], Berkely 3D dataset [2] and our own dataset.

Our conclusion is that the depth is an important cue for the detection of salient region in a scene. With a novel formulation we propose 3D-saliency which when fused with saliency of the existing models enhances their performance.

[1] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, and S.M. Hu. Global Contrast based Salient Region Detection. In *CVPR*, 2011.

[2] A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko, and T. Darrell. A Category-Level 3D Object Dataset: Putting the Kinect to Work. In *ICCV Workshop*. 2011.

[3] K. Lai, L. Bo, X. Ren, and D. Fox. A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In *ICRA*. 2011.

[4] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth Matters: Influence of Depth Cues on Visual Saliency. In *ECCV*. 2012.

Oriented pooling for dense and non-dense rotation-invariant features

Wan-Lei Zhao[1]
http://www.cs.cityu.edu.hk/~wzhao2
Hervé Jégou[1]
http://people.rennes.inria.fr/Herve.Jegou
Guillaume Gravier[2]
http://people.irisa.fr/Guillaume.Gravier

[1] INRIA
Rennes, France
[2] CNRS/IRISA
Rennes, France

Bag-of-words has been extended in various ways. This paper enriches the representation of each descriptor with the (quantized) characteristic scale and dominant orientation [2] associated with the region of interest. This additional information is exploited by a Hough-like voting procedure [1] to favor the images that have been scaled and rotated consistently.

In this paper, we depart from most existing works by using *rotation-invariant dense features*. In image search, most systems rely on key-points or region detectors. A description relying on regular dense features achieves good performance but at the cost of loosing invariance to orientation, which is not desirable in many applications. Pooling on dominant orientations of the local features have been explored in the context of object classification [5], however they do not enforce the relative orientation to be preserved. In our work, we aim at obtaining both the discriminative power conveyed by dense descriptors and invariance to orientation.

The vector of locally aggregated descriptors (VLAD) [4] is an encoding technique that produces a fixed-length vector representation v from a set $\mathcal{X} = \{x_1, ..., x_m\}$ of m local d-dimensional descriptors (*e.g.*, SIFT, $d = 128$), which have been extracted from a given image. The VLAD computation procedure relies on a visual vocabulary $\mathcal{C} = \{c_1, \cdots, c_k\}$ where the dictionary (size k) is trained offline with k-means algorithm. It is used by a quantization function $q : \mathbb{R}^d \rightarrow \mathcal{C}$ that associates x_i to its Euclidean nearest neighbor in the vocabulary $\mathcal{C}$, as $q(x) = \arg\min_{c \in \mathcal{C}} \|x - c\|$. VLAD is a $d \times k$ vector, where each component is indexed by both the indices i and j associated to the quantization indexes and sift components, respectively. A component of VLAD vector $v = [v_{1,1}, \ldots, v_{i,j}, \ldots, v_{k,d}]$ associated with $\mathcal{X}$ is obtained as

$$v_{i,j} = \sum_{x \in \mathcal{X} : q(x) = c_i} x_j - c_{i,j}, \tag{1}$$

where x_j and $c_{i,j}$ are the j^{th} components of descriptor x and visual word c_i, respectively. As a post-processing, the vector v is ℓ_2-normalized.

To boost the performance of VLAD, several pre- and post-processing operations have been proposed, such as pairwise square-root on SIFT, rotating SIFT with PCA and applying pairwise power law on VLAD. VLAD undergone these series operations is denoted as VLAD*.

In the regular VLAD, features with different orientation are aggregated, loosing the possibility to estimate any geometrical transformation between the aggregated features. To alleviate this problem, we propose to pool features according to some characteristic geometrical quantities, more specifically the characteristic scales and dominant orientations [6] obtained as a byproduct of the descriptor computation stage. We aggregate features having similar characteristic scales or dominant orientations, to obtain a new pooling strategy termed Covariant-VLAD (CVLAD).

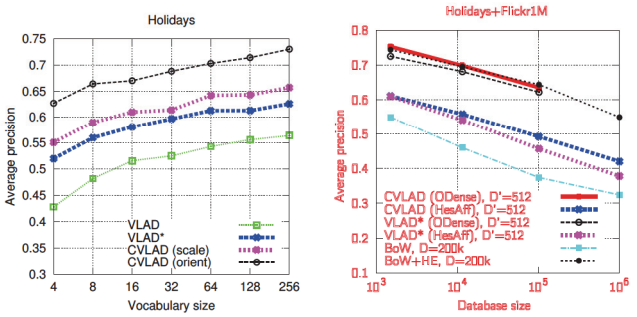
Take pooling on dominant orientation as an example. Let denote by θ the dominant orientation associated with a given feature x, and let

$$b_B(\theta) = \left\lfloor B \frac{\theta}{2\pi} \right\rfloor \tag{2}$$

be the quantization function used to quantize angles with B equally sized bins. Our pooling strategy modifies Equation 1 as

$$p_{b,i,j} = \sum_{x \in \mathcal{X} : q(x) = c_i \wedge b_B(\theta) = b} x_j - c_{i,j}. \tag{3}$$

In Eqn. 3, the pooling of the feature x is controlled by both its quantization index $q(x)$ and its quantized dominant angle $b_B(\theta)$. Another way to see the CVLAD construction procedure is to consider that $\mathbf{P} = [\mathbf{P}_1, \ldots, \mathbf{P}_B]$ is a concatenation of B VLAD $k \times d$-dimensional vectors, each of which encodes the features having the same quantized dominant orientation. This produces a vector B times longer than VLAD. Similar pooling is feasible with the characteristic scale. Note that the series pre- and post-processing adopted in VLAD* is also applied on each sub-vector of CVLAD.



(a) CVLAD *vs* VLAD and VLAD* (b) Search in large-scale

Figure 1: Performances of CVLAD in Holiday datasets. (a) Performances of pooling with dominant orientation (orient) and characteristics scale (scale) in comparison with conventional VLAD and VLAD*. (b) Performance of CVLAD in large-scale image search as a function of database size. The performance of VLAD has been compared with CVLAD*, BoW and BoW+HE [3]. Hessian-Affine+SIFT and dense SIFT have been adopted in experiment (a) and (b) respectively.

The similarity $s(.,.)$ between two CVLAD vectors $\mathbf{P}^i$ and $\mathbf{P}^j$ is defined on the basis of VLAD sub-vectors as

$$s(P^i, P^j) = \underset{\Delta t \in 0 \ldots B-1}{\operatorname{argmax}} \sum_{t=0}^{B-1} \cos\left(\mathbf{P}^i_t, \mathbf{P}^j_{\bmod(t+\Delta t, B)}\right) \tag{4}$$

which amounts to selecting the orientation maximizing the similarity between the two vectors. This process is comparable to estimating the dominant rotation transformation between two feature sets in WGC [3], however here it is done directly on the aggregated vectors.

Eqn. 4 performs a circulant matching between two sets of VLAD sub-vectors. The matching shifts the VLAD sub-vector in P^j circularly to search for the best match between two groups of VLAD.

We presented CVLAD which pools on dominant orientations of local features. This approach builds upon the recent VLAD descriptor, but offers a new trade-off with respect to geometrical invariance by implicitly selecting the rotation (likewise scale) to maximize the similarity for a given image pair.

[1] Paul V. C. Hough. Method and means for recognizing complex patterns. US patent 3069654, Dec. 1962.

[2] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, Oct. 2008.

[3] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, Feb. 2010.

[4] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local descriptors into compact codes. In *Trans. PAMI*, Sep. 2012.

[5] Piotr Koniusz, Fei Yan, and Krystian Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer Vision and Image Understanding*, 17 (5):479–492, 2013.

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

Efficient Discriminative Convolution Using Fisher Weight Map

Hideki Nakayama
http://www.nlab.ci.i.u-tokyo.ac.jp/

Graduate School of Information Science and Technology
The University of Tokyo, JAPAN

Abstract

Convolutional neural networks (CNNs) have been studied for a long time, and recently gained increasingly more attention. Deep CNNs have especially achieved remarkably high performance on many visual recognition tasks due to their high levels of flexibility. However, since CNNs require numerous parameters to be tuned via iterative operations through layers, their computational cost is immense. Moreover, they often require a huge number of training samples and technical tricks, such as unsupervised pretraining and heuristic tuning, to successfully train the system.

In this work, we present a very simple method of layer-wise convolution. We can obtain discriminative filters by using a Fisher weight map (FWM) [8], which well separates convolved images between categories. This operation can be deterministically solved as a simple eigenvalue problem and no back propagation or hyper-parameters are needed. Because our method is layer-wise and based on a simple eigenvalue problem, it is computationally efficient. We demonstrate the promising performance of our method in extensive experiments with two datasets.

Network architecture

Let $\boldsymbol{x}^{(k)}_{(x',y')}$ denote a vector of stacked features in a receptive filed in the k-th layer (see Figure 1), and $\boldsymbol{X}$ denote a matrix that contains all $\boldsymbol{x}^{(k)}$ from an instance in its columns. Also, we let $\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{w}$ denote a convolved image vector via projection $\boldsymbol{w}$. FWM finds discriminative projections by maximizing between-class distance of $\boldsymbol{z}$. The optimal weights can be analytically obtained by solving the following eigenvalue problem.

$$\Sigma_B \boldsymbol{w} = \lambda \Sigma_W \boldsymbol{w}. \tag{1}$$

Here, Σ_W and Σ_B are within- and between-class scatter matrices, namely

$$\Sigma_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (\boldsymbol{X}_i^{(j)} - \bar{\boldsymbol{X}}^{(j)})(\boldsymbol{X}_i^{(j)} - \bar{\boldsymbol{X}}^{(j)})^T, \tag{2}$$

$$\Sigma_B = \sum_{j=1}^{C} N_j (\bar{\boldsymbol{X}}^{(j)} - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(j)} - \bar{\boldsymbol{X}})^T, \tag{3}$$

where C is the number of categories and N_j is the number of training samples in class j.

Moreover, we found that the key to achieving the best performance was to use the above method with appropriate methods of pooling and rectification, which is another contribution we made. More specifically, it is crucial to use our convolutional layer with subsequent average-pooling (AP) and rectified linear units (ReLU) operations. We observed that ReLU exploiting both positive and negative activations are particularly effective.

$$R_2(x) = \begin{pmatrix} \max(0, x) \\ \max(0, -x) \end{pmatrix}. \tag{4}$$

Results and discussion

We tested our methods on two benchmarks, *i.e.*, the STL-10 [2] and MNIST [7] datasets (Figure 2). The experimental results revealed that our convolution layer could reasonably improve the performance of the original descriptors. Moreover, our method used together with appropriate pooling methods and ReLU operations achieved remarkably high levels of performance on both datasets that were comparable or better than those of state-of-the-art networks.

Tables 1 and 2 summarize the scores of our method and those from previous work for STL-10 and MNIST, respectively. We implemented K-means features for STL-10 and random features for MNIST as the low-level inputs of the networks. Our best model outperformed all previously published scores in the literature. Superior results on STL-10 especially indicated that our method could stably learn from a limited number of training examples (100 per class). This is probably because our method is based on a simple eigenvalue problem and is capable of densely learning from high-dimensional descriptors without dropping connections between neurons.

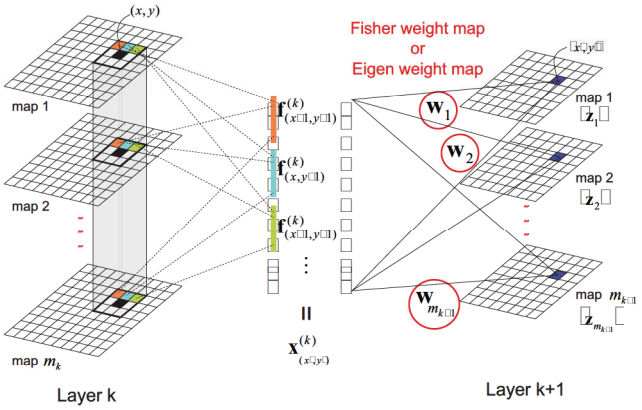


Figure 1: Convolution with weight map techniques.



Figure 2: Images from the STL-10 (top) and MNIST (bottom) datasets.

Table 1: Comparison of classification rates on STL-10 (%).

1-layer Vector Quantization [4]	54.9 ± 0.4
1-layer Sparse Coding [4]	59.0 ± 0.8
3-layer Learned Receptive Field [3]	60.1 ± 1.0
Discriminative Sum-Product Network [5]	62.3 ± 1.0
Ours, $K_m(9, 1000)$-$MP(8, 4)$-$C_F(3, 100)$-R_2-$AP(4, 3)$	**66.0 ± 0.7**

Table 2: Comparison of classification errors on MNIST (%). We compared our method with previous work using raw training dataset.

Large CNN (unsup. pretraining) [6]	0.53
3-layer CNN + Stochastic Pooling [9]	0.47
Multi-Column Deep Neural Network [1]	0.46*
Ours, $Rand(5, 1000)$-R-$AP(3, 2)$-$C_F(3, 500)$-R_2-AP_q	**0.44**

[1] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Proc. IEEE CVPR*, 2012.

[2] A. Coates, H. Lee, and A. Ng. An analysis of single-layer networks in unsupervised feature learning. In *Proc. AISTATS*, 2011.

[3] A. Coates and A. Ng. Selecting receptive fields in deep networks. In *Proc. NIPS*, 2011.

[4] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proc. ICML*, 2011.

[5] R. Gens and P. Domingos. Discriminative learning of sum-product networks. In *Proc. NIPS*, 2012.

[6] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. Lecun. What is the best multi-stage architecture for object recognition? In *Proc. IEEE ICCV*, 2009.

[7] Y. LeCun. The MNIST database of handwritten digits. URL `http://yann.lecun.com/exdb/mnist/`.

[8] Y. Shinohara and N. Otsu. Facial expression recognition using Fisher weight maps. In *IEEE FG*, pages 499–504, 2004.

[9] M. D. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. In *arXiv preprint*, 2013.

Single-Image Super-Resolution via Adaptive Joint Kernel Regression

Chen Huang
yach23@gmail.com
Xiaoqing Ding
dingxq@tsinghua.edu.cn
Chi Fang
fangchi@tsinghua.edu.cn

State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Single image super-resolution (SR) methods can be broadly categorized into three classes: interpolation-based methods, reconstruction-based methods [7], and example-based methods [2, 3, 6]. The reconstruction-based methods often incorporate prior knowledge to regularize the ill-posed problem. For example, Zhang *et al*. [7] assembled the Steering Kernel Regression [5] (SKR)-based local prior and Nonlocal Means [1] (NLM)-based nonlocal prior. The example-based methods strongly rely on the chosen dictionary for satisfactory results. This paper focuses on learning good image priors and robust dictionaries for SR reconstruction. Among the extensively studied natural image priors, we choose to exploit the *local structural regularity* prior and *nonlocal self-similarity* prior in a coherent framework.

We propose in this paper an Adaptive Joint Kernel Regression (AJKR)-based prior to simultaneously exploit both image statistics. Our approach differs from others in several ways: 1) we combine a set of NLM-generalized local kernel regressors, which are more consistent with our nonlocal collaborative framework; 2) the proposed *regional redundancy* measure introduces higher-order statistics at the region level for each regression group, making the overall framework more adaptive (see Fig. 1(a)); and 3) an adaptive PCA-based dictionary learning scheme is adopted to bridge the gap of dictionaries learned online and offline by mixing them, and more importantly such a scheme together with its induced sparsity prior can adapt to the AJKR process in response to the regional redundancy measure (see the block diagram in Fig. 1(b)).

The imaging model for SR (assume $\mathbf{Y} \in \mathbb{R}^m$ is low resolution (LR) image, $\mathbf{X} \in \mathbb{R}^n$ is high resolution (HR) image) is usually expressed as

$$\mathbf{Y} = \mathbf{DHX} + \mathbf{V}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{m\times n}$ and $\mathbf{H} \in \mathbb{R}^{n\times n}$ are the downsampling matrix and blurring matrix respectively, and $\mathbf{V} \in \mathbb{R}^m$ is the Gaussian noise. To incorporate the AJKR prior, we first generalize NLM in the kernel regression framework

$$\hat{\mathbf{a}}_i = \arg\min_{\mathbf{a}} \|\mathbf{Y}_i - \mathbf{\Phi a}\|^2_{\mathbf{W}_i^N}, \tag{2}$$

where $\mathbf{Y}_i$ is a patch of neighboring pixels around location $\mathbf{x}_i$, and $\mathbf{\Phi}$ and $\mathbf{a}$ are the regression bases and coefficients. $\mathbf{W}_i^N = diag\left[w_{i1}^N, w_{i2}^N, \ldots, w_{iL}^N\right]$ is the patch similarity-based kernel weight matrix with

$$w_{ij}^N = exp\left(-\frac{\|\mathbf{Y}_i - \mathbf{Y}_j\|^2_{\mathbf{W}_G}}{h_n^2}\right). \tag{3}$$

This procedure paves the way for our framework "harmonization" in a complete nonlocal sense. Then we combine all the local regressors over similar patches found in the nonlocal range $\mathcal{P}(\mathbf{x}_i)$ (sufficiently large image region), and weight them by their mutual similarities w_{ij}^N

$$\hat{\mathbf{a}}_i = \arg\min_{\mathbf{a}} \sum_{j\in\mathcal{P}(\mathbf{x}_i)} w_{ij}^N \|\mathbf{Y}_j - \mathbf{\Phi a}\|^2_{\mathbf{W}_j^N}. \tag{4}$$

Such a scheme allows us to exploit both local/nonlocal image priors collaboratively, and can act as a regularization term for SR reconstruction. We further propose a regional redundancy measure $R_i = \sum_{j\in\mathcal{P}(\mathbf{x}_i)} \left(w_{ij}^N\right)^2$ to determine the confidence of these regression groups and thus control their relative effects of regularization using $\mathbf{R} = diag\left[R_1, R_2, \ldots, R_n\right]$. This leads to the matrix form of our SR optimization function

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{DHX}\|_2^2 + \lambda \|(\mathbf{I} - \mathbf{K})\mathbf{X}\|^2_{\mathbf{R}}, \tag{5}$$

where $\mathbf{K}$ packs the *equivalent regression kernels* derived from the solutions of Eq. (4). Although previous models [6, 7] also considered the unified priors, ours is much more coherent and adaptive with the redundancy measure that introduces higher-order statistics.

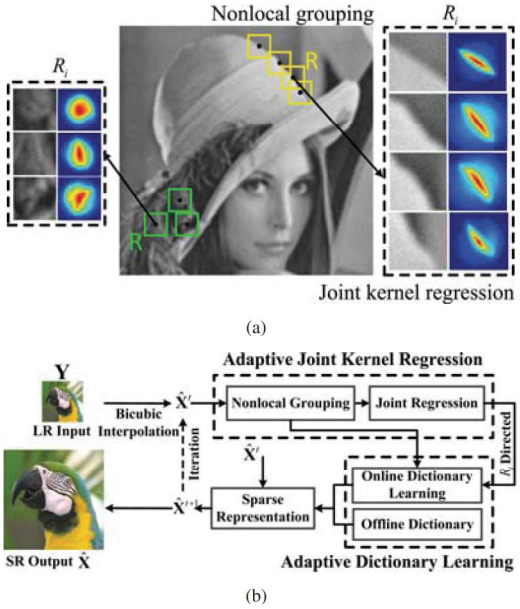


Figure 1: (a) Graphical illustration of the AJKR process; (b) Block diagram of our SR algorithm.

The AJKR framework can further benefit from dictionary-based methods. We follow the adaptive PCA strategies in [7] and [2] to learn the offline and online dictionaries $\mathbf{B}_0$ and $\mathbf{B}_1$, respectively. By combining the two dictionaries ($\mathbf{B} = [\mathbf{B}_0\ \mathbf{B}_1]$) and enforcing sparsity under the dictionary representation (α), we can reach to our final optimization function

$$\hat{\alpha} = \arg\min_{\alpha} \|\mathbf{Y} - \mathbf{DHB}\circ\alpha\|_2^2 + \lambda \|(\mathbf{I} - \mathbf{K})\mathbf{B}\circ\alpha\|^2_{\mathbf{R}} + \beta \|\alpha\|_1. \tag{6}$$

Compared with similar schemes for combined dictionaries (*e.g.* [4]), our combined dictionary together with its induced sparsity prior distinguishes itself by offering the ability to interact with the regression prior in response to the redundancy measure in a unified framework.

Details of SR implementation of this method is described in this paper. The large variety of experiments show that our method achieves state-of-the-art performance and is robust with minimum artifacts even under large amounts of noise and blur. The extension to image deblurring also produces outstanding results as compared with other methods.

[1] A. Buades, B. Coll, and J. M. Morel. A non-local algorithm for image denoising. In *Proc. CVPR*, pages 60–65, June 2005.

[2] P. Chatterjee and P. Milanfar. Clustering-based denoising with locally learned dictionaries. *IEEE TIP*, 18(7):1438–1451, July 2009.

[3] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE TIP*, 20(7):1838–1857, 2011.

[4] Daniele Perrone, Avinash Ravichandran, René Vidal, and Paolo Favaro. Image priors for image deblurring with uncertain blur. In *Proc. BMVC*, pages 114.1–114.11, 2012.

[5] H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. *IEEE TIP*, 16(2):349–366, 2007.

[6] Haichao Zhang, Jianchao Yang, Yanning Zhang, and Thomas S. Huang. Non-local kernel regression for image and video restoration. In *Proc. ECCV*, pages 566–579, 2010.

[7] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE TIP*, 21(11):4544–4556, 2012.

Anatomical Structure Sketcher for Cephalograms by Bimodal Deep Learning

Yuru Pei[1]
peiyuru@cis.pku.edu.cn
Bin Liu[1]
liubin@cis.pku.edu.cn
Hongbin Zha[1]
zha@cis.pku.edu.cn
Bing Han[2]
orthohanks@sina.com
Tianmin Xu[2]
tmxuortho@gmail.com

[1] Key Laboratory of Machine Perception (MOE)
Peking University
Beijing, China
[2] Stomatology Hospital
Peking University
Beijing, China

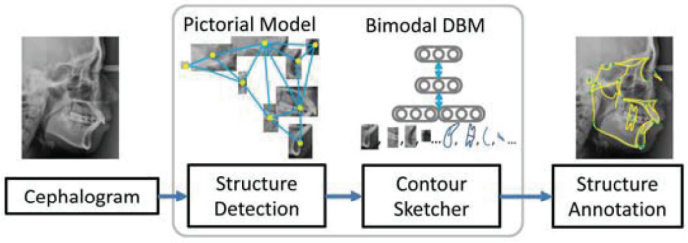


Figure 1: Flowchart of our cephalogram sketcher system.

Lateral cephalogram X-ray (LCX) images are essential to provide patient-specific morphological information of anatomical structures. The automatic annotation of anatomical structures in cephalograms has been performed in the biomedical engineering for nearly twenty years. Most systems only handle a portion of salient craniofacial landmark set [1, 2, 3]. Although model-based methods can produce a full set of markers [5, 7], the pattern fitting can fail to converge in blurry images. It is challenging to annotate LCX images with high fidelity.

In this work, we propose a novel cephalogram sketcher system as shown in Fig. 1 for the automatic anatomical-structure annotation, especially for the blemished images due to structure overlappings and device-specific distortions during projection. Firstly, we introduce an hierarchical extension of a pictorial model to detect anatomical structures. Secondly, the bimodal deep Boltzmann machine (DBM) is employed to sketch the structure contours. Specifically, the contour sketcher takes advantages of the path in the DBM to extract the contour definitions from the patch textures by alternating Gibbs sampling.

Given a cephalogram I, the structure definition S, and the parameters $\Theta = (\Theta_q, \Theta_r)$ with respect to the intra- and inter-layer correlations, the posterior probability distribution according to the Bayes rule is defined as $P(S|I,\Theta) \propto P(I|S,\Theta)P(S|\Theta)$, where $P(S|\Theta)$ is a shape prior distribution. $P(I|S,\Theta)$ is the image likelihood given the hierarchical architecture and the model parameters. The likelihood can be factorized as a product of likelihoods of local structures.

$$\ln P(I|S,\Theta_q,\Theta_r) = \sum_{l=0}^{D-1} \sum_{s_i^l \in L_l} \ln \phi(s_i^l) + \sum_{l=0}^{D-2} \{ \sum_{C_r} \ln \varphi(s_i^l, s_j^{l+1}|\Theta_{r,ij}) + \sum_{C_q} \ln \varphi(s_i^{l+1}, s_j^{l+1}|\Theta_{q,ij}) \}. \quad (1)$$

$\phi(s_i)$ is potential of local structures computed based on the output of SVM classifiers as in [4], and $\phi(s_i) = (1+\exp(A_i f_i(s_i) + B_i))^{-1}$, where f_i is the output of the linear SVM classifier for the i_{th} kind of structures s_i. A linear SVM is trained for each kind of structures respectively, where the parameters, Λ_i and B_i, are predefined.

When given multimodal data, the deep architecture can build a joint representation by virtue of hidden layers from each modality. The shared attributes can be transferred from one modality to the other via the joint layer. Considering the undirected RBM, the observed data of one modality take a role in training the parameters related to layers of the other modality, which is the main difference from the multimodal DBN [6].

The bottom layers $(v_m, h_m^{(1)}, \ldots, h_m^{(K-1)} | m = p,t)$ are related to the image patches and contour shapes respectively, where K is the number of hidden layers and set at 3 in our experiments. The top one $h^{(K)}$ is the joint hidden layer connecting both modalities. Given two modalities, the patches v_p and shape contours v_t, the joint distribution

$$p(v_p, v_t|\vartheta) = \sum_{h_p^{(K-1)}, h_t^{(K-1)}, h^{(K)}} \exp\left(-E_j\right) \cdot \prod_{m=p,t} \sum_{h_m} \left(\exp\left(-E_g\left(v_m, h_m^{(1)}\right) - \ldots - E_b\left(h_m^{(K-2)}, h_m^{(K-1)}\right)\right)\right). \quad (2)$$

The partition functions are removed for clarity. The term E_j is the energy between the joint hidden layer $h^{(K)}$ and the upper hidden layers with respect to the patch and contour modalities, $h_p^{(K-1)}$ and $h_t^{(K-1)}$, and

$$E_j = -\sum_i^{N_{h,K}} a_i^K h_i^{(K)} - \sum_m \left(\sum_j^{N_{h,K-1}^m} a_{m,j}^{K-1} h_{m,j}^{(K-1)} + \sum_i^{N_{h,K-1}^m} \sum_j^{N_{h,K}} h_{m,i}^{(K-1)} W_{ij}^{K,m} h_j^{(K)} \right). \quad (3)$$

It is intractable to learn the parameters by maximizing the above likelihood. Alternatively, the problem can be solved by mean-fields, which minimized the KL-divergence $p(h|v_p, v_t)$ [6]. The model parameters are initialized by learning the layer-wise stacking of RBMs.

For the purpose of contour sketching, $p(v_t|v_p,\vartheta)$ needs to be solved. In DBM, when given the observed modality, v_p, the missing modality v_t can be generated by alternating Gibbs sampling. Specifically, v_p serves as input, while all the hidden units are initialized randomly (e.g. set at zeros in our case). In order to decide when to stop the iteration, we measure a score of the contour predication by the distance between the input image patches and the reconstructed ones.

$$P(v_t|v_p,\vartheta) \propto \exp\left(-\left\| v_p - v_p^{model} \right\|\right), \quad (4)$$

where v_p^{model} is the reconstructed image patch features, i.e. the HOG histograms in our system, and v_p is the input.

Implementation of this method is described in the paper. Our conclusion is that the sketcher as an integration of the hierarchical pictorial model and the deep learning can infer the positions and contours of anatomical structures effectively, and robust to noisy data.

[1] D. Giordano, R. Leonardi, F. Maiorana, G. Cristaldi, and M. Distefano. Automatic landmarking of cephalograms by cellular neural networks. *Artificial Intelligence in Medicine*, pages 333–342, 2005.

[2] A. Innes, V. Ciesielski, J. Mamutil, and S. John. Landmark detection for cephalometric radiology images using pulse coupled neural networks. In *Proc. Int. Conf. on Artificial Intelligence*, volume 2, 2003.

[3] R. Leonardi, D. Giordano, and F. Maiorana. An evaluation of cellular neural networks for the automatic identification of cephalometric landmarks on digital images. *Journal of Biomed and Biotec*, 2009.

[4] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[5] AA Saad, A. El-Bialy, AH Kandil, and AA Sayed. Automatic cephalometric analysis using active appearance model and simulated annealing. *Int J on GVIP*, 6:51–67, 2006.

[6] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Proc. NIPS*, 2012.

[7] W. Yue, D. Yin, C. Li, G. Wang, and T. Xu. Automated 2-d cephalometric analysis on x-ray images by a model-based approach. *IEEE Trans. on Biomedical Engineering,*, 53(8):1615–1623, 2006.

Reconstructing Faces from their Signatures using RBF Regression.

Alexis Mignon
alexis.mignon@unicaen.fr
Frédéric Jurie
frederic.jurie@unicaen.fr

GREYC – CNRS UMR 6072,
University of Caen Basse-Normandie,
Caen, France

Faces recognition technologies are now omnipresent on the Internet (Google, Facebook, etc. now use face recognition to automatically tag people).

One important question related to privacy is to know if the information encoded by the face signatures (or *face templates* as they are called in the biometrics literature) might be reverse engineered, revealing the identities of the persons in case the database is stolen or made publicly available by mistake. In other words, do face signatures (or face templates, we interchangeably use the two expressions) contain personal information?

This is also a topic of interest for biometric applications where the way templates are encoded is critical [6]. Indeed, stolen templates should not allow reconstruction of images that can be used to break in a system. As face representation and face verification technologies have progressed a lot during the last five years, we believe it is important to know, if and to what extent, people can be recognized from their face templates once their face is reconstructed.

In this context, the objective of this paper is twofold: (i) to propose a novel simple but efficient method for reconstructing face images from face templates with unknown identities; (ii) to show that these reconstructed images allow the recognition of people and can fool modern face recognition algorithms. We believe that the proposed approach could be used each time a new method for representing faces is proposed, to evaluate how anonymity is preserved by the encoding.

In practice, we chose to use the I-LPQ approach of [7] – which is, at the moment, the state-of-the-art unsupervised approach on the LFW [2] and FERET [5] datasets – as a baseline approach for computing face templates. Another reason for using this approach is that it is a good representative of modern approaches for face recognition, most of them being based on histograms of LBP/SIFT/Gabor-like features computed on a regular grid [1, 4, 8].

The proposed reconstruction method is based on radial basis function regression in the face eigenspace. During a training phase, face/signature pairs are used to learn a non linear mapping between signatures and images. This mapping minimizes the reconstruction error of the eigenface representations of the training images. The mapping can then be used directly to reconstruct faces from their signatures.

We consider the case where N image/signature pairs are available. We denote $\mathbf{x}_i$ ($i = 1 \dots N$) the N signatures and, to simplify the notations, we consider the face images as 1-dimensional vectors $\mathbf{y}_i$ of P pixels. We can then define the following quantities:

$$
\begin{aligned}
Y &= [\mathbf{y}_1 \dots \mathbf{y}_N]^T \\
\Phi &= \begin{pmatrix} \phi(\|\mathbf{x}_1 - \mathbf{x}_1\|) & \dots & \phi(\|\mathbf{x}_1 - \mathbf{x}_N\|) \\ \vdots & \ddots & \vdots \\ \phi(\|\mathbf{x}_N - \mathbf{x}_1\|) & \dots & \phi(\|\mathbf{x}_N - \mathbf{x}_N\|) \end{pmatrix} \\
W &= \begin{pmatrix} w_{11} & \dots & w_{1P} \\ \vdots & \ddots & \vdots \\ w_{N1} & \dots & w_{NP} \end{pmatrix}
\end{aligned}
$$

where W are weights to be estimated and $\phi(\|\mathbf{x}_i - \mathbf{x}_j\|) = \sqrt{s^2 + \|\mathbf{x}_i - \mathbf{x}_j\|^2}$ is the multi-quadric function with scale parameter s. We call $H = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ the centering operator where $\mathbf{1}_N$ is the N-dimensional vector with all elements equal to 1. Let $\bar{Y} = USv^T$ be the singular value decomposition of $\bar{Y} = HY$, and $\bar{\Phi} = H\Phi H$ be the centered version of Φ. We solve:

$$\min_W \|\bar{\Phi}W - v\|_F^2 \tag{1}$$

where $\|.\|_F$ is the Frobenius norm. This gives $W = \Phi^+ v$ where Φ^+ is the pseudo invert of Φ.

Given a set of M signature $\mathbf{x}'_i$, to reconstruct the corresponding images, we need the matrix Φ' where $\Phi'_{ij} = \phi(\|x'_i - x_j\|)$ and its centered version with respect to Φ, which is $\bar{\Phi}' = (\Phi' - \frac{1}{N}\mathbf{1}_M\mathbf{1}_N^T\Phi)(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)$. The M corresponding images, stacked as the rows of a $M \times P$ matrix Y', are then reconstructed as:

$$Y' = \left(\bar{\Phi}'(\bar{\Phi}^+ vv^T) + \frac{1}{N}\mathbf{1}_M\mathbf{1}_N^T\right)Y \tag{2}$$



Figure 1: Reconstructing LFW images. Top rows are real faces while bottom ones are faces reconstructed with a decoder trained with 3000 training images.

We evaluated our algorithms on two face databases: FERET [5] and LFW [2]. Qualitative results on the LFW database are shown in figure 1. We also performed quantitative evaluations and observed that our method consistently outperforms the method of Mohanty *et al*. [3] based on multi-dimensional scaling (MDS) on both datasets.

Acknowledgments. This work was realized as part of the SecuLar program funded by the French National Research Agency (ANR), grant reference ANR-12-CORD-0014.

[1] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA*, 2010.

[2] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[3] P. Mohanty, S. Sarkar, and R. Kasturi. From scores to face templates: A model-based approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2065–2078, 2007.

[4] H. V. Nguyen and L. Bai. Cosine Similarity Metric Learning for Face Verification. In *ACCV*. LNCS, 2010.

[5] J. P. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[6] Christian Rathgeb and Christoph Busch. *Multi-Biometric Template Protection: Issues and Challenges, New Trends and Developments in Biometrics*. InTech, 2012.

[7] S. ul Hussain, T. Napoléon, and F. Jurie. Face Recognition using Local Quantized Patterns. In *Proceedings of the 23rd British Machine Vision Conference, Surrey, England*, 2012.

[8] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA*, 2011.

Isowarp and Conwarp: Warps that Exactly Comply with Weak-Perspective Projection of Deforming Objects

Daniel Pizarro
dani.pizarro@gmail.com
Adrien Bartoli
adrien.bartoli@gmail.com
Toby Collins
toby.collins@gmail.com

ISIT-UMR 6284 CNRS
Universite d'Auvergne
Clermont Ferrand, France

Deformable surface reconstruction from monocular views has been intensively studied during the last few years [1, 6]. In the template-based case, the goal is to obtain the 3D shape of the deformed template from a single image. This problem is ill-posed without constraints on how the surface deforms. Shape inference methods have been proposed for a variety of deformation constraints: low-rank shape priors [4, 6], temporal deformation smoothness [7], isometric deformations [1, 2, 5] and conformal deformations [1]. Isometric and conformal constraints accurately model deformations in many real cases and they are described with simple differential constraints. Isometric reconstruction from perspective camera views has attracted much of the attention [1, 5, 6].

We study surface reconstruction with isometric and conformal deformations and weak-perspective projection. In real scenarios, cameras with large focal lengths and shallow scene depth produce images close to affine projection conditions. In those conditions the solution is ambiguous [3] and the state-of-art with perspective cameras provide unstable results. This paper gives analytical solutions and shows that the registration between the template and the input image is not independent of the reconstruction. We define registration warps that comply with the weak-perspective projection of isometric and conformal surfaces, ensuring analytical solutions to compute surface's shape. We show that this method is more accurate than state-of-art methods based on perspective models, when the camera conditions are close to affine.

Figure (1) shows the different functional relationships between the template and the deformed surface. The template is represented as a surface $\mathcal{T} \subset \mathbb{R}^3$ embedded in 3D. We assume that $\mathcal{T}$ has disk topology, admitting a flat template denoted as $\mathcal{P} \subset \mathbb{R}^2$. We denote the flattening as the invertible mapping function $\Delta \subset \mathbb{C}_1$.

The input image $\mathcal{I}$ shows the projection of the deformed template $\mathcal{S} \subset \mathbb{R}^3$. The camera model Π is a weak-perspective projection. We define $\eta \in \mathbb{C}_1$ as the warp mapping a point $\mathbf{p} \in \mathcal{P}$ in the flattened template to a corresponding point $\mathbf{q} = \eta(\mathbf{p}) \in \mathcal{I}$ in the image. Function $\psi \in \mathbb{C}_1$ models deformations between the template $\mathcal{T}$ and the deformed surface $\mathcal{S}$. Conformal and isometric deformations are described with conditions on the first derivatives of ψ, namely J_ψ:

$$J_\psi^\top J_\psi = \mu \mathbf{I}_{3\times 3}, \qquad \text{where} \qquad \mu = \begin{cases} 1 & \text{Isometric} \\ \mu \in \mathbb{C}_1 \text{ unknown} & \text{Conformal.} \end{cases} \tag{1}$$

We define $\varphi = (\psi \circ \Delta) \in \mathbb{C}_1$ which maps points $\mathbf{p} \in \mathcal{P}$ in the 2D template to points $Q = \varphi(\mathbf{p}) \in \mathcal{S}$ in the deformed surface. Using the properties of the conformal flattening and the conditions (1), the following differential properties are derived:

$$J_\varphi^\top J_\varphi = \lambda \mathbf{I}_{2\times 2}, \qquad \text{where} \qquad \lambda = \begin{cases} \sqrt{\det(J_\Delta^\top J_\Delta)} & \text{Isometric} \\ \lambda \in \mathbb{C}_1 \text{ unknown} & \text{Conformal.} \end{cases} \tag{2}$$

Reconstruction with Isometric and Conformal Deformations

The reconstruction problem with Isometric or Conformal deformations is equivalent to find a solution to the following system of PDEs:

$$\begin{cases} \chi_x + s^2 \left(\frac{\partial \varphi_z}{\partial x}\right)^2 = \chi_y + s^2 \left(\frac{\partial \varphi_z}{\partial y}\right)^2 = \lambda s^2 \\ \xi + s^2 \left(\frac{\partial \varphi_z}{\partial x}\right)\left(\frac{\partial \varphi_z}{\partial y}\right) = 0, \end{cases} \tag{3}$$

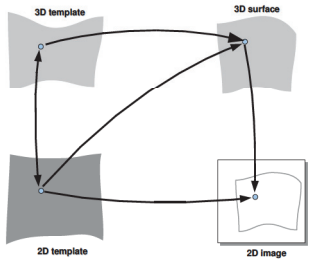


Figure 1: General model of template-based 3D reconstruction

where λ is given by equation (2) and:

$$\chi_x = \frac{\partial \eta}{\partial x}^\top \frac{\partial \eta}{\partial x} \qquad \chi_y = \frac{\partial \eta}{\partial y}^\top \frac{\partial \eta}{\partial y} \qquad \xi = \frac{\partial \eta}{\partial x}^\top \frac{\partial \eta}{\partial y}. \tag{4}$$

The isowarp

We define the isowarp as a function η that represents the weak-perspective projection of an isometric deforming surface. If η is an isowarp, system (3) has an exact solution and directly solves the isometric reconstruction problem. An isowarp η satisfies the following system of PDEs:

$$\begin{aligned} (s^2\lambda - \chi_x)(s^2\lambda - \chi_y) - \xi^2 = 0 \\ (s^2\lambda - \chi_y)\left(s^2\frac{\partial \lambda}{\partial y} - \frac{\partial \chi_x}{\partial y}\right)^2 - (s^2\lambda - \chi_x)\left(s^2\frac{\partial \lambda}{\partial x} - \frac{\partial \chi_y}{\partial x}\right)^2 = 0, \end{aligned} \tag{5}$$

where $s^2 \in \mathbb{R}^+$

The conwarp

We define the conwarp as a function η that represents the weak-perspective projection of a conformal deforming surface. A conwarp satisfies the following PDE:

$$(s^2\lambda - \chi_y)\left(s^2\frac{\partial \lambda}{\partial y} - \frac{\partial \chi_x}{\partial y}\right)^2 - (s^2\lambda - \chi_x)\left(s^2\frac{\partial \lambda}{\partial x} - \frac{\partial \chi_y}{\partial x}\right)^2 = 0, \tag{6}$$

where $\lambda = \frac{1}{2s^2}\left(\chi_x + \chi_y + \sqrt{(\chi_x - \chi_y)^2 + \xi^2}\right)$

[1] A. Bartoli, Y. Gérard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *IEEE CVPR, 2012*, pages 2026–2033. IEEE, 2012.

[2] F. Brunet, R. Hartley, A. Bartoli, N. Navab, and R. Malgouyres. Monocular template-based reconstruction of smooth and inextensible surfaces. *Computer Vision–ACCV 2010*, pages 52–66, 2011.

[3] T Collins and A Bartoli. Locally affine and planar deformable surface reconstruction from video. In *VMV 2010*, pages 339–346, 2010.

[4] A. Del Bue, X. Llad, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *IEEE CVPR 2006*, volume 1, pages 1191–1198. IEEE, 2006.

[5] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *IJCV*, 95(2):124–137, 2011.

[6] M. Salzmann and P. Fua. Linear local models for monocular reconstruction of deformable surfaces. *IEEE PAMI*, 33(5):931–944, 2011.

[7] M. Salzmann, R. Hartley, and P. Fua. Convex optimization for deformable surface 3-d tracking. In *ICCV 2007*, pages 1–8. IEEE, 2007.

Point Light Source Estimation based on Scenes Recorded by a RGB-D camera

Bastiaan J. Boom[1]
http://homepages.inf.ed.ac.uk/bboom/
Sergio Orts-Escolano[2]
http://www.dtic.ua.es/~sorts/
Xin X. Ning[1]
avener1446@hotmail.com
Steven McDonagh[1]
http://homepages.inf.ed.ac.uk/s0458953/
Peter Sandilands[1]
http://homepages.inf.ed.ac.uk/s0569500/
Robert B. Fisher[1]
http://homepages.inf.ed.ac.uk/rbf/

[1] Institute of Perception, Action and Behaviour
University of Edinburgh
Edinburgh, UK

[2] Computer Technology Department
University of Alicante
Alicante, Spain



Figure 1: Rendering a synthetic object (dragon) into a real scene recorded with the Kinect sensor, where both the shading and shadows are based on an estimated light source position determined using only the information provided by the Kinect sensor.

Goal: To realistically render objects in a scene taking into account the illumination of the scene. Our approach achieves this by estimating the position of a point light source that explains the illumination observed in the scene. The advantage of a point light source is that it can be used easily by even simple rendering software. This allows us to render synthetic objects in real scenes (see Figure 1), improving the augmented reality experience. In this paper, the only sensor that is used for the estimation of the point light source is an RGB-D camera (Kinect sensor). No probe objects, extra fish-eye cameras or other devices are used to estimate the illumination conditions. In order to estimate the point light source position in the scene our method uses the following assumptions: 1) The Lambertian reflectance model describes roughly the appearance of objects in the scenes, where the normals of these objects can be obtained from the depth camera. The normals constrain the position of a single point light source. 2) Segments in the RGB image with a similar colour that are contiguous tend to have the same albedo.

Contributions: *Point light source estimation method:* A method that is able to estimate a point light source that explains the illumination in the scene recorded by the Kinect. This method only uses the image and depth map provided by the Kinect sensor for the estimation, where all other methods use also other information, for instance a probe sphere to determine the illumination condition in a scene. *Simulation experiments to verify the method:* A simulation based on a rendered scene is performed to verify the methodology. Our methodology used the Lambertian reflectance model to estimate the illumination of a scene. In our simulation, we are able to verify what to expect in the cases of specular reflections and ambient lighting. *Scene with measured point light position:* In order to verify the performance of the methodology under real conditions, different scenes illuminated with a single light source are recorded by the Kinect sensor, where the position of the light source and the camera are known allowing us to estimate the accuracy of our methodology.

Method: Using the two assumptions mentioned above, the Lambertian reflectance model can be defined as follows:

$$I_o(\mathbf{p}) = \rho(\mathbf{p}) \min(\mathbf{n}(\mathbf{p})\mathbf{s}(\mathbf{p})^T i, 0) \quad (1)$$

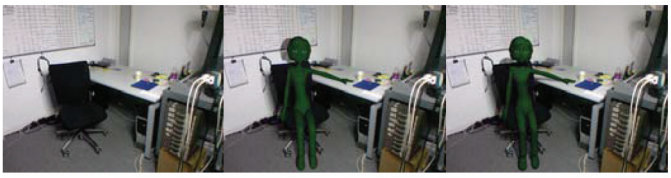


Figure 2: The rendering of a synthetic person in the scene where the shadows are different. Humans can however not observe what the correct light position was given the small difference in angle. (Left: original image, middle: true light position, right: estimated light position)

The original image I_o at pixel $\mathbf{p} = \{x, y\}$ according to the Lambertian reflectance model consists of the albedo ρ, the surface normal $\mathbf{n}$ of the objects in the scene at $\mathbf{p}$ and the direction $\mathbf{s}$ (normalised vector) and intensity i of the light. Given the output of the Kinect sensor, both the image intensity I and the surface normals $\mathbf{n}$ are obtained, where the surface normals are computed from the depth map. However, both the albedo $\rho(\mathbf{p})$ and light parameters $\mathbf{s}(\mathbf{p}), i$ are unknown. Using the assumption that the albedo in segments R_j of the image are the same: for instance, a red billiard ball in an image, the RGB values of this ball in the image are not the same, however the albedo of this ball is the same where the variation is caused by the reflectance model. Given the arbitrary light parameters $\mathbf{s}_r(\mathbf{p}), i_r$ (given by the search method), first the albedo values ρ in the segment R_j can be computed

$$\rho(\mathbf{p}) = \frac{I_o(\mathbf{p})}{\mathbf{n}(\mathbf{p})\mathbf{s}_r(\mathbf{p})^T i_r}, \mathbf{p} \in R_j \quad (2)$$

To obtain the same albedo value for the positions in segment ρ_{R_j}, the mean or median of the albedo values $\rho(\mathbf{p}), \mathbf{p} \in R_j$ can be used. Because some segments contain normals $\mathbf{n}$ that are close to being perpendicular to the light direction $\mathbf{s}$, the median is a more stable function for obtaining the albedo of the segment ρ_{R_j}. A reconstructed image I_r for a given segment R_j can be computed using the following equation:

$$I_r(\mathbf{p}) = \rho_{R_j} \min(\mathbf{n}(\mathbf{p})\mathbf{s}(\mathbf{p})^T i, 0), \mathbf{p} \in R_j \quad (3)$$

$$E = \sum_{\mathbf{p} \in P} ||I_o(\mathbf{p}) - I_r(\mathbf{p})|| \quad (4)$$

By performing this for all segments, we obtain a reconstructed image I_r. An error E between the original image I_o (given in Equation 1) and the reconstructed image I_r obtained in Equation 3, can be calculated as follows: The error E (Equation 4) indicates to which extent the reconstructed image I_r explains the observed scene I_o with respect to the gradient normally observed in object appearance due to the illumination and the surface of the object. This operation was performed for an arbitrary point light source (light parameters $\mathbf{s}_r(\mathbf{p}), i_r$), where search methods are used to find the point light source with that minimise the error E.

Discussion: In Figure 2, the scene is shown rendered using the true light source and the estimated light source position, where humans humans find it difficult to determine which is the more realistic rendering. In practise, the illumination in the scene can be very complex, however for the goal of rendering synthetic objects estimations of the illumination conditions are convincing.

Closed form solution for radial distortion estimation from a single vanishing point

Horst Wildenauer
horst.wildenauer@gmail.com
Vienna University of Technology
Vienna, Austria

Branislav Micusik
branislav.micusik@aic.ac.at
AIT Austrian Institute of Technology
Vienna, Austria

Radial lens distortion found in real, most notably off-the-shelf medium to wide angle optics can be quite severe. A-priori calibration remedies the problem, but requires access to the camera. Alternative approaches make use of correspondences in multiple images of a moving camera [3], relying on sufficiently overlapping views.

Our paper fits in the category of techniques, where straight scene lines are used to determine distortion parameters from a single image. These plumbline methods [2] critically depend on the existence of long edges and get thrown of easily if a significant portion originates from non-linear structures. Here, the use of vanishing points provides an advantage: When a large number of image edges groups *w.r.t.* to a common vanishing point, it is very likely that they stem from parallel scene lines. Recent methods for vanishing point detection either ignore radial distortion completely, or assume only weak distortion, which is accounted for after vanishing point estimation [1, 4]. If strong distortion is present the effects are detrimental to the extraction of vanishing points and such algorithms are bound to fail. Furthermore, when dealing with radial distortion effects in refinement stages, the error is assessed in the undistorted image. This creates significant bias towards estimates shrinking the undistorted image [5]

Contributions: To overcome this limitation, we suggest a simple RANSAC procedure for vanishing point detection by clustering of distorted lines. The proposed technique builds upon: a) Our newly devised closed-form solution for simultaneously estimating a vanishing point and radial distortion from three distorted image lines. b) A consistency measure which avoid bias by quantifying the error of lines *w.r.t.* a vanishing point in the distorted image.

Specifically, our work bases on the division model [3], mapping a distorted image point $\mathbf{x} = (x,y)^\mathsf{T}$ to the undistorted point $\tilde{\mathbf{x}}$ by

$$C: \quad \tilde{\mathbf{x}} = \mathbf{x}/(1+\lambda r^2). \tag{1}$$

Here, $r = \sqrt{x^2+y^2}$ is the distance of the distorted point to the distortion center, which we fix at the center of the image. Under this model, straight line segments are distorted into circular arcs [5]. Such arcs serve as basic features for our approach. We extract them using a Canny edge detector followed by circle fitting. For each detected arc smoothed estimates of the arc's midpoint $\mathbf{x} = (x,y)^\mathsf{T}$ and the normal $\mathbf{n} = (u,v)^\mathsf{T}$ of the tangent at that point are computed from a fitted circle.

Closed-form solution: If we undistort $\mathbf{x}$ and map $\mathbf{n}$ accordingly to its undistorted image, the resulting transformed tangent line

$$\tilde{\mathbf{t}} = \begin{pmatrix} u \\ v \\ -ux-vy \end{pmatrix} + \lambda \begin{pmatrix} ux^2+2vxy-uy^2 \\ vy^2+2uxy-vx^2 \\ 0 \end{pmatrix}. \tag{2}$$

coincides with the line generating the circle, passing through the undistorted vanishing point $\tilde{\mathbf{v}} = (\tilde{v}_1, \tilde{v}_2, \tilde{v}_3)^\mathsf{T}$. Stacking the equations $\tilde{\mathbf{t}}_i^\mathsf{T}\tilde{\mathbf{v}}_i = 0,\ i = 1,2,3$ of three tangent lines, we obtain the generalized eigenvalue problem

$$(\mathsf{D}+\lambda\mathsf{E})\,\tilde{\mathbf{v}} = \left(\begin{bmatrix} d_1 & d_2 & d_3 \\ d_4 & d_5 & d_6 \\ d_7 & d_8 & d_9 \end{bmatrix} + \lambda \begin{bmatrix} e_1 & e_2 & 0 \\ e_3 & e_4 & 0 \\ e_5 & e_6 & 0 \end{bmatrix} \right) \tilde{\mathbf{v}} = \mathbf{0}. \tag{3}$$

The characteristic polynomial of the problem is a quadratic $\lambda^2 c_2 + \lambda c_1 + c_0 = 0$, which is can be easily solved. Once λ is obtained, the undistorted vanishing point $\tilde{\mathbf{v}}$ can be found by plugging into (3).

Consistency measure: The consistency of a circular arc with midpoint $\mathbf{x}$ and normal $\mathbf{n}$ *w.r.t.* a vanishing point $\tilde{\mathbf{v}}$ and λ is computed as

$$\mathrm{dist}(s,\tilde{\mathbf{v}},\lambda) = \frac{l_s}{2}\sin\angle(\mathbf{n},\mathbf{n}'). \tag{4}$$

Here, $\mathbf{n}'$ is the normal of the arc "corrected" to be compatible with $\tilde{\mathbf{v}}$:

$$\mathbf{n}' = \begin{pmatrix} 1+\lambda r^2-2\lambda x^2 & -2\lambda xy \\ -2\lambda xy & 1+\lambda r^2-2\lambda y^2 \end{pmatrix} \begin{pmatrix} \tilde{v}_3 \frac{y}{1+\lambda r^2} - \tilde{v}_2 \\ \tilde{v}_1 - \tilde{v}_3 \frac{x}{1+\lambda r^2} \end{pmatrix} \tag{5}$$

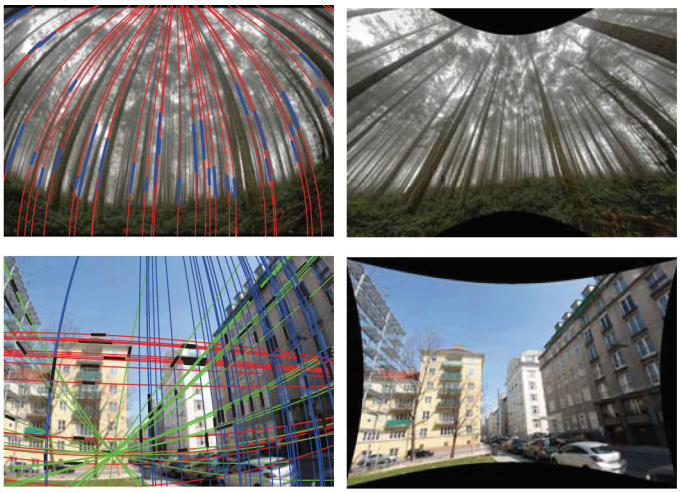


Figure 1: Exemplary results for one (top row) and three vanishing points (bottom row). Left: Circular arcs grouped *w.r.t.* to vanishing points, indicated by converging circles. Right: Automatically undistorted images.

The weighting factor l_s is the arc length in pixels and represents a means of allowing less angular deviation for longer arcs. Furthermore, (4) gracefully handles the case when no distortion ($\lambda = 0$) is present, degenerating to the line segment-based consistency employed in [6].

In Fig. 1(top row) a typical result obtained with our RANSAC-based method to estimate distortion from a single vanishing point is depicted. One can see that even images of natural scenes may be used to compute visually pleasing results. Unfortunately, many images have more than one vanishing point and if the dominant one is close to the image center, radial distortion cannot by reliably estimated. To remedy this problem, we suggest to augment our method with the vanishing-point based calibration scheme introduced in [6]. Using two additional circular arcs (for a total of five), we arrive at a unified camera calibration approach, which in addition to stably estimating radial distortion, computes the camera's focal length and up to three orthogonal vanishing points. The algorithm was tested on two data sets containing over 200 images of urban environments taken with calibrated cameras. For an example, see Fig. 1(bottom row). As detailed in the paper, our approach exhibits high accuracy in terms of focal length and radial distortion estimates, handling strong radial distortion just as well as negligible one.

Our conclusion is that it is possible to directly estimate vanishing points in images with strong radial distortion with only little more effort than in the standard, undistorted case. Furthermore, our algorithm may be used to detect vanishing points in images unaffected by radial distortion without the danger of overfitting.

[1] Christian Bräuer-Burchardt and Klaus Voss. Automatic correction of weak radial lens distortion in single views of urban scenes using vanishing points. In *ICIP*, 2002.

[2] Frederic Devernay and Olivier D. Faugeras. Straight lines have to be straight. *Mach. Vis. Appl.*, 13(1):14–24, 2001.

[3] Andrew W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *CVPR*, pages 125–132, 2001.

[4] Lazaros Grammatikopoulos, George Karras, and Elli Petsa. An automatic approach for camera calibration from vanishing points. *ISPRS Journal of Photogrammetry and Remote Sensing*, pages 64–76, 2007.

[5] Rickard Strand and Eric Hayman. Correcting radial distortion by circle fitting. In *BMVC*, 2005.

[6] Horst Wildenauer and Allan Hanbury. Robust camera self-calibration from monocular images of manhattan worlds. In *CVPR*, 2012.

Robust Scene Stitching in Large Scale Mobile Mapping

Filip Schouwenaars[1]
filip.schouwenaars@student.kuleuven.be
Radu Timofte[12]
radu.timofte@esat.kuleuven.be
Luc Van Gool[12]
luc.vangool@esat.kuleuven.be

[1] VISICS, ESAT-PSI / iMinds
KU Leuven
Leuven, Belgium
[2] Computer Vision Lab, D-ITET
ETH Zürich
Zürich, Switzerland

This paper presents a solution for the loop closure problem in an image-based mobile mapping context. A van equipped with stereo cameras collects recordings in an urban environment, simultaneously monitoring GPS information. Using Structure-from-Motion, the position of the van and the surroundings are retrieved. The determination of the translation and orientation of the van's position is recursive: a slight drift can gradually build up to flawed localizations. One can rely on the GPS information to perform adjustments, but its accuracy is not adequate to yield a model with high precision. Yet, visual loop-closing – recognizing that a location is revisited – may help mitigate the issue. The current system does not take into account possible reoccurrences of identical features in distant recordings. This paper adds such loop closure.

Local feature matching in two stages detects when a particular site is revisited, in order to enforce correspondences between images, that may have been taken with large time lapses in between. Our system relies on GPS but does not use odometric information. We extend the original image-to-image matching approach to a pose-to-pose matching approach, combining several images and achieving robust scene matching results. Parameter optimization is followed by extensive experiments. Our pipeline, which facilitates parallel execution, reaches matching rates higher than those reported for typical state-of-the-art algorithms. We also demonstrate robustness to odometric inconsistencies resulting from poor prior model build-up.

Loop closure is crucial for high-accuracy models. The current state-of-the-art in topological mapping are FABMAP [3] and CAT-SLAM [4], but limit themselves to a binary decision, *i.e.* whether or not the location was visited already. In the envisioned application however, it is desirable to have actual image point correspondences to facilitate bundle adjustment. An approach closer to this goal, by directly attempting to match local features among images, is described in [5]. The utilized epipolar constraint to cope with false positives is, however, not error-free. The devised approach method implements more robust error dismissal.

The approach consists of two major steps. First, the issue of detecting revisited sites over time is encountered by clustering GPS information and taking its inaccuracy into account. Next, in every *route* of such a cluster, two poses are selected that are expected to contain common elements, based on a Naive Bayesian matching framework with severely downscaled images. When a so-called *cross-route pose pair* is obtained, re-occurrences of the same physical points are tracked in the associated images. This problem is treated in two steps: *single pose matching* and *cross-route image matching*. The former finds matches and deducts corresponding 3D points using the SURF [1] detector and descriptor among views taken from the same van position. Since camera calibration is available, an epipolar consistency check is straightforward. Using the surviving matches, for every van pose a point cloud results. The latter step of cross-route image matching attempts to match the images from different van poses again using SURF. This practice establishes a link between the two earlier constructed point clouds. PROSAC [2], a prioritized RANSAC algorithm, is applied to robustly calculate the transformation between the two point clouds in a time-optimal way while pruning out false positives. Figure 1 provides an illustration.

This paper has two main contributions. First, our novel loop-closure technique does not depend on single image pairs for correspondences. Instead, a cloud is constructed around two van poses and these clouds are fitted together; an image-to-image approach is extended to a pose-to-pose approach. Second, our method is able to detect matches in challenging wide baseline conditions, were other systems tend to fail.

Since it concerns a system-specific application, a specialized dataset is devised that comprises a substantial amount of images from an urban environment. Separate datasets were used for parameter tuning and subsequent validation. Several time-decreasing techniques that impact on the SURF scheme and the inherent properties of the algorithm allow a balance between timing and performance. An adaptive approach of retrieving 3D information even increases performance while reducing computation significantly. The different steps can be carefully monitored through computation break-down experiments. Fetching the image data from memory constitutes a large part of execution time, followed by SURF detection and description.



Figure 1: Illustration of cross-route image matches after pruning of false positives by means of PROSAC. Note that these are not the only correspondences found; also for other image combinations matches are tracked.

#	#DNB	Descriptor	INC	Matching Rate	Comp. Time
13	100	USURF-64	✗	48/49 (98.0%)	1259 s

Table 1: Concise statement of results for the preferred setup on a 5531 poses dataset. *#DNB* stands for number of descriptors used for Naive Bayes Matching. *Descriptor* provides type and dimension of the descriptor. *INC* shows if the incremental approach is enabled.

Typically, a matching that exceeds 85% is achieved, *i.e.* when a cross-route pose pair is selected, the system is successful in finding correspondences between distant recordings. The number of false positives is zero. Information and results related to the preferred setup are summarized in Table 1.

Mock-up examples to mimic poor prior model build-up show robustness of the technique and prove the applicability in a wide context. Finally, the resulting correspondences between images that associate with distant recordings are embedded in the current systems. Further work is necessary to report on expected accuracy increase following from the promising results.

Our conclusion is that the novel pose-to-pose matching approach for robust scene stitching in a problem-specific context shows promising results, as proven by thorough experimentation. The performance is however not straightforwardly comparable to current state-of-the-art techniques due to the utilization of specialized datasets.

[1] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 2008.

[2] Ondrej Chum and Jiri Matas. Matching with PROSAC - Progressive Sample Consensus. In *CVPR*, 2005.

[3] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 2008.

[4] Will Maddern, Michael Milford, and Gordon Wyeth. CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory. *Int. J. Rob. Res.*, 2012.

[5] Stephen Se, David Lowe, and Jim Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The International Journal of Robotics Research*, 2002.

Local Zernike Moment Representation for Facial Affect Recognition

Evangelos Sariyanidi[1]
e.sariyanidi@eecs.qmul.ac.uk
Hatice Gunes[1]
hatice@eecs.qmul.ac.uk
Muhittin Gökmen[2]
gokmen@itu.edu.tr
Andrea Cavallaro[1]
andrea.cavallaro@eecs.qmul.ac.uk

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

[2] Department of Computer Engineering
Istanbul Technical University,
Istanbul, Turkey

Local representations became popular for facial affect recognition as they efficiently capture the image discontinuities, which play an important role for interpreting facial actions. We propose to use Local Zernike Moments (ZMs) [4] due to their useful and compact description of the image discontinuities and texture. Their main advantage in comparison to well-established alternatives such as Local Binary Patterns (LBPs) [5], is their flexibility in terms of the size and level of detail of the local description. We introduce a local ZM-based representation which involves a non-linear encoding layer (quantisation). The functionality of this layer is mapping similar facial configurations together and increasing compactness. We demonstrate the use of the local ZM-based representation for posed and naturalistic affect recognition on standard datasets, and show its superiority to alternative approaches for *both* tasks.

Contemporary representations are often designed as frameworks consisting of three layers [2]: (Local) feature extraction, non-linear encoding and pooling. Non-linear encoding aims at enhancing the relevance of local features by increasing their robustness against image noise. Pooling describes small spatial neighbourhoods as single entities, ignoring the precise location of the encoded features, and increasing the tolerance against small geometric inconsistencies. In what follows, we describe the proposed local ZM-based representation scheme in terms of this three-layered framework.

Feature Extraction – Local Zernike Moments: The computation of (complex) ZMs can be considered equivalent to representing an image in an alternative space. As shown in Figure 1-a, an image is decomposed onto a set of basis matrices (ZM bases), which are useful for describing the variation at different directions and scales. ZM bases are orthogonal, therefore there is no overlap in the information conveyed by each feature (ZM coefficient). ZMs are usually computed for the entire image, however in this case, ZMs cannot capture the local variation due to ZM bases lacking localisation [3]. In contrary, when computed around local neighbourhoods across the image, they become an efficient tool for describing the image discontinuities which are essential to interpreting facial activity.

Non-linear Encoding – Quantisation: We perform quantisation via converting local features into binary values. Such coarse quantisation increases compactness and allows us to code each local block only with a single integer. Figure 1-b illustrates the process of obtaining the Quantised Local ZM (QLZM) image. Firstly, local ZM coefficients are computed across the input image (LZM layer) — each image in the LZM layer (LZM image) contains the features that are extracted through a particular ZM basis. Next, each LZM image is converted into a binary image by quantising each pixel via the signum(·) function. Finally, the QLZM image is obtained by combining all of the binary images. Specifically, each pixel in a particular location of the QLZM image is an integer (QLZM integer), computed by concatenating all of the binary values in the corresponding location of all binary images. The QLZM image is similar to an LBP-transformed image, in the sense that it contains integers of a limited range. Yet, the physical meaning of the information encoded by each integer is quite different. LBP integers describe a circular block by considering only the values along the border, neglecting the pixels that remain inside the block. Therefore, the efficient operation scale of LBPs is usually limited to 3-5 pixels [1, 5]. QLZM integers, on the other hand, describe blocks as a whole, and provide flexibility in terms of operation scale without major loss of information.

Pooling – Histograms: Our representation scheme pools encoded features over local histograms. Figure 1-c illustrates the overall pipeline of the proposed representation scheme. Firstly, the QLZM image is computed through the process that is illustrated in detail in Figure 1-b. Next,

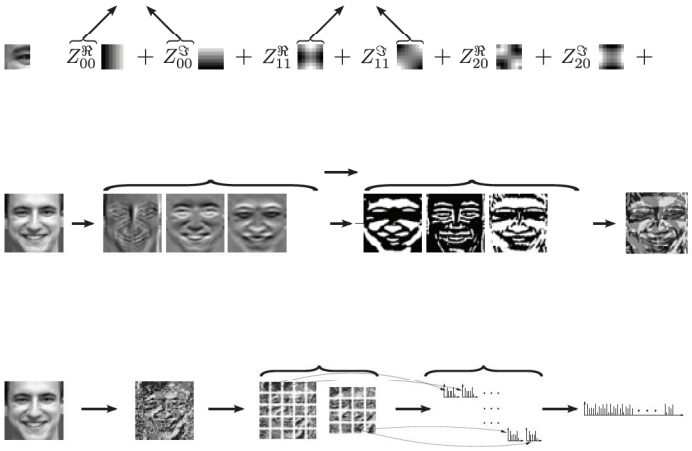


Figure 1: Illustration of the proposed representation scheme. (a) Local ZM features (ZM coefficients) and the ZM bases. (b) Illustration of the process of computing the QLZM image. (c) Illustration of the overall representation pipeline.

the QLZM image is divided into uniform regions, and the local histograms that count the QLZM integers are computed for each region. The final representation is obtained by concatenating all local histograms.

Our experimental results show that the local ZM based representation is superior to its well-established alternatives. By experimenting both on datasets of posed and spontaneous affective behaviour, we show that the suitable level of detail to encode differs for these two different contexts, and the contribution of quantisation is more pronounced on spontaneous data. While fine-grained representation is suitable for the idealised context of posed affective behaviour, coarser representation via quantisation is more suitable for spontaneous behaviour. This is likely to be due to the better generalisation capability that quantisation provides. This improvement suggests that the full potential of alternative non-linear encoding schemes such as quantisation with finer resolutions or non-linear functions with continuous output should be explored further.

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12): 2037–2041, 2006.

[2] Kevin Jarrett, Koray Kavukcuoglu, Marc' Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2146–2153, 2009.

[3] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[4] Evangelos Sariyanidi, Volkan Dagli, Salih Cihan Tek, Birkan Tunc, and Muhittin Gökmen. Local Zernike Moments: A new representation for face recognition. In *Proceedings of the IEEE International Conference on Image Processing*, pages 585–588, 2012.

[5] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803 – 816, 2009.

An Adaptive Subdivision Scheme for Quadratic Programming in Multi-Label Image Segmentation

Marko Rak
rak@isg.cs.ovgu.de
Tim König
tim.koenig@isg.cs.ovgu.de
Klaus-Dietz Tönnies
toennies@isg.cs.ovgu.de

Department of Simulation and Graphics
University of Magdeburg, Germany

Convex quadratic optimization is one of the most widely used concepts in image segmentation. It facilitates a wide range of information sources, such as edge, intensity, texture and shape. The problem is especially challenging for the multi-label case, even being NP-hard in its most general setting. Therefore, fast "solutions", as the α-expansion of [1], are limited to local optimality. Addressing this problem, several approaches relax the labeling integrality condition, resulting in quadratic programs (QPs) like in [2] and in [4], which can be solved in polynomial time.

Although this is efficient in a theoretical sense, large-scale QPs that arise from typical multi-label tasks can rarely be used for image segmentation directly due to either time or space constraints, or both. We address this issue by an adaptive domain subdivision scheme, reducing the problem to a short sequence of spatially smoothed medium-scale QPs, which subsequently better approximate the large-scale program. Our scheme is globally optimal in terms of the approximated problem.

Putting our main focus on the subdivision, we restrict ourselves to minimization of the popular but rather simple piecewise constant Mumford-Shah functional. Therefore, we seek for a labeling that trades off the length of the labeling border and the approximation of image intensity u by known reference intensitites u_i for each label i. For discrete domains the associated energy can be written as

$$E = \sum_{j\in\mathcal{P}} \sum_i \alpha_i (u-u_i)^2 x_i^j + \gamma \sum_{(j,k)\in\mathcal{N}} \sum_i (x_i^j - x_i^k)^2, \tag{1}$$

with $\mathcal{P}$ being the set of domain elements and $\mathcal{N}$ being the set of neighborhood relations. The binary variable x_i^j indicates whether domain element j mutually exclusively belongs to label i or not.

Circumventing the hardness of general convex quadratic labelings, we remove the binarity constraint while keeping the rest of the problem fixed. For any quadratic labeling energy the problem then transforms into a $[0,1]$-relaxed binary QP $(2-4)$ in $\mathbf{x} = (x_1^1 x_1^2 \dots x_l^{p-1} x_l^p)^\mathrm{T}$.

$$\mathbf{x}^\mathrm{T}\mathbf{H}\mathbf{x} + \mathbf{f}^\mathrm{T}\mathbf{x} \quad \to \min! \tag{2}$$

$$\mathbf{A}\mathbf{x} = \mathbf{1} \tag{3}$$

$$\mathbf{0} \preceq \mathbf{x} \preceq \mathbf{1} \tag{4}$$

Here, the element-wise mutual label exclusiveness degrades to affinity, i.e., $\forall j : \sum_i x_i^j = 1$, which is reflected by Eq. (3). For the Mumford-Shah example, matrix $\mathbf{H}$ and vector $\mathbf{f}$ subsume the neighborhood relations and the intensity approximation terms of Eq. (1), respectively.

Clearly, the above problem is of large scale, because $\mathbf{x}$ contains $p \cdot l$ variables, where p is the number of domain elements and l is the number of labels. Treating this issue, we seek for an adaptive domain subdivision that gradually becomes finer near borders of the (unknown) labeling. If such a domain subdivision would be known beforehand then we could set up l label variables for each of the r subdivision regions and arrange them altogether into a vector $\mathbf{z} = (z_1^1 z_1^2 \dots z_l^{r-1} z_l^r)^\mathrm{T}$. Furthermore, we could set up a projector $\mathbf{P}$ which by

$$\mathbf{x} = \mathbf{P}\mathbf{z} \tag{5}$$

superimposes the regional basis in $\mathbf{z}$ to the element basis in $\mathbf{x}$ according to the subdivision. Inserting the superimposition (5) into the QP $(2-4)$, we would obtain a spatially smoothed approximation of the large-scale program. The approximate QP in $\mathbf{z}$ would be of medium scale, because $r \ll p$ unless the mild assumption of spatial labeling coherence is violated.

Certainly, we do not know the subdivision beforehand,which is why we have to (re-)construct it alongside the labeling. To attain this goal we draw on a hierarchical domain subdivision via quadtree, octree and their generalization to higher dimensions. The key idea is to create a short sequence $(0, 1, \dots, i, \dots, c)$ of superimpositions

$$\mathbf{x} = \mathbf{P}_i \mathbf{z}_i \tag{6}$$



Figure 1: Comparison of our adaptive subdivision scheme to the large-scale quadratic program on standard image data. First column: input images, second column: labeling of the large-scale quadratic program, third column: labeling of our adaptive subdivision scheme for $c = 5$

that stem from subsequent adaptive subdivision refinement steps. More precisely, we initially decompose the domain up to a preset tree level c above ground, i.e., above the 0-level where regions equal domain elements. We then alternate $c+1$ times between two steps. In the first step, we set up and solve the current approximate QP

$$\mathbf{z}_i^\mathrm{T}\mathbf{P}_i^\mathrm{T}\mathbf{H}\mathbf{P}_i\mathbf{z}_i + \mathbf{f}^\mathrm{T}\mathbf{P}_i\mathbf{z}_i \quad \to \min! \tag{7}$$

$$\mathbf{A}\mathbf{P}_i\mathbf{z}_i = \mathbf{1} \tag{8}$$

$$\mathbf{0} \preceq \mathbf{z}_i \preceq \mathbf{1}, \tag{9}$$

based on the tree leaves known so far. Its solution is then transformed from regional to element basis via Eq. (6) and "thresholded" to a hard labeling, i.e., taking the most certain label for each element. In the second step, leaves of the tree are subdivided if they touch a border of the hard labeling. The refined subdivision tree is then passed into the next alternation.

As the alternation elapses, the number of regions r_i increases, improving on the problem approximation. Due to the adaptivity of the subdivision, we keep the medium-scale property, i.e., $r_i \ll p$ for each QP $(7-9)$ of the sequence. Keeping space requirements at its lowest, we can assemble $\mathbf{P}_i^\mathrm{T}\mathbf{H}\mathbf{P}_i$, $\mathbf{f}^\mathrm{T}\mathbf{P}_i$ and $\mathbf{A}\mathbf{P}_i$ directly without ever constructing the large-scale quantities $\mathbf{H}$, $\mathbf{f}$ and $\mathbf{A}$ explicitly.

We drew an experimental comparison of our scheme to the large-scale QP based on minimization of the Mumford-Shah functional (1). We used synthetic image data with strong noise as well as standard images from the Berkeley Segmentation Dataset [3], some examples of which are given in Figure 1. Experiments show high quality labelings widely independent of the initial tree level. Our scheme even introduces additional regularization, preventing the emergence of spurious segments. Depending on the complexity of the multi-label problem, our scheme outperformed the large-scale program by, at least, factors between two and seven. The gain in performance even allowed for near-interactive multi-label segmentation in our experiments on medium-sized problems.

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23: 1222–1239, 2001.

[2] O. Dalmau-Cedeno, M. Rivera, and P. P. Mayorga. Computing the alpha-channel with probabilistic segmentation for image colorization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1–7, 2007.

[3] 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. IEEE Int. Conf. Comput. Vis.*, pages 416–423, 2001.

[4] M. Rivera, O. Dalmau-Cedeno, and J. Tago. Image segmentation by convex quadratic programming. In *Proc. Int. Conf. Pattern Recognit.*, pages 1–5, 2008.

Visual Tracking via Subspace Motion Model

Jun Wang[1]
jwang1004@gmail.com
Fan Zhong[1]
zhongfan@sdu.edu.cn
Guofeng Wang[1]
wangguofeng13@163.com
Qunsheng Peng[2]
peng@cad.zju.edu.cn
Xueying Qin[1]
qxy@sdu.edu.cn(Correspondence)

[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] The State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China

The art of visual tracking has been widely studied in the past decades [6]. While most of researches focus on exploring new methods to represent object appearance, little attention has been paid on the description of object motion. In this paper we propose a novel motion model for visual tracking, and in comparison with previous methods, it can better parameterize instantaneous image motion caused by both object and camera movements.

Our approach is inspired by the subspace theory of image motion, that is, for a rigid object imaged by a projective camera, the displacements matrix of its trajectories over a short period of time should approximately lie in a low-dimensional subspace with a certain rank upper bound [2, 5]. We adopt this subspace as the state transition space in particle filtering (PF) [3]. This differs from affine model in two ways: first, the dimension number as well as the sampling weight for each dimension at each moment can be determined by the rank of the subspace automatically; second, the subspace motion model can naturally represent the disparity brought by object or camera rotation. We will show that when compared with the affine model, the subspace motion model is superior in accuracy.

Figure 1 illustrates the procedures of our method. To estimate the motion model, some 2D feature points of the object are first tracked by the standard KLT approach [4]. Assuming that k successive frames have been tracked before the current frame I_t, then the displacements matrix can be built as:

$$U = \begin{bmatrix} u_{11} & u_{22} & \dots & u_{n1} \\ u_{12} & u_{22} & \dots & u_{n2} \\ & & \vdots & \\ u_{1k} & u_{2k} & \dots & u_{nk} \end{bmatrix}, V = \begin{bmatrix} v_{11} & v_{22} & \dots & v_{n1} \\ v_{12} & v_{22} & \dots & v_{n2} \\ & & \vdots & \\ v_{1k} & v_{2k} & \dots & v_{nk} \end{bmatrix} \quad (1)$$

where (u_{ij}, v_{ij}) denotes the displacement of point (x_i, y_i) from the reference frame I to frame I_j. Each row in U and V corresponds to a single frame, and each column corresponds to a single point. U, V can be stacked vertically as $[\frac{U}{V}]_{2k\times n}$. Then, according to the subspace theory, the displacements matrix can be expressed as a bilinear product of matrices:

$$\begin{bmatrix} U \\ V \end{bmatrix}_{2k\times n} = \begin{bmatrix} M_U \\ M_V \end{bmatrix}_{2k\times r} P_{(r\times n)} \quad (2)$$

where P is made up of point-dependent column vectors involving only points positions parameters, while M_U and M_V are made up of frame-dependent row vectors involving only camera motion parameters. The decomposition can be performed by SVD [1], from which the actual value of r is determined by the singular values.

We always deem a moving camera with a static object and describe the initial state of the PF in subspace as:

$$\mathbf{x}_{t-1} = \begin{bmatrix} (M_U)_{k-1} \\ (M_V)_{k-1} \end{bmatrix}_{2\times r} \quad (3)$$

where $(M_U)_{k-1}$ and $(M_V)_{k-1}$ are the last rows of M_U and M_V respectively, corresponding to the frame-dependent vectors of frame I_{t-1}.

The prediction stage of PF is implemented as a first-order autoregressive process:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta\mathbf{m} \quad (4)$$

where $\Delta\mathbf{m}$ is a two-row weighted Gaussian random sample drawn from the normal distribution. The sample weight of each column of $\Delta\mathbf{m}$ is decided by the weight of each subspace dimension.

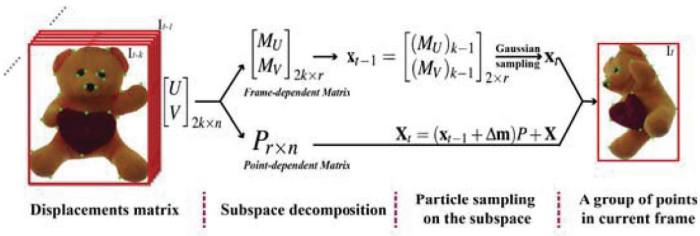


Figure 1: Flow chart of the subspace motion model

Each particle $\mathbf{x}$ in the subspace is corresponding to a group of feature points positions $\mathbf{X}_t$ in the current frame I_t. Let $\mathbf{X}$ be the set of feature points positions in the reference frame I, $\mathbf{X}_t$ is computed by:

$$\mathbf{X}_t = (\mathbf{x}_{t-1} + \Delta\mathbf{m})P + \mathbf{X} \quad (5)$$

The observation state is taken as the point tracking results $\mathbf{Y}_t$ in the current frame I_t. The importance weights w_t^i of each particle is estimated by:

$$w_t^i = \frac{1}{W_t} \exp(-\frac{1}{\eta}\|\mathbf{X}_t^i - \mathbf{Y}_t\|_F^2) \quad (6)$$

where $W_t = \sum_{i=1}^{N} \exp(-\frac{1}{\eta}\|\mathbf{X}_t^i - \mathbf{Y}_t\|_F^2)$ is the normalization term, and η the smooth term. Then the tracking result $\mathbf{Z}_t$ and tracking error can be represented as:

$$\mathbf{Z}_t = \sum_{i=1}^{N} w_t^i \mathbf{X}_t^i, \quad err_t = \frac{1}{N}\|\mathbf{Z}_t - \mathbf{Y}_t\|_F^2 \quad (7)$$

The subspace motion model also can facilitate the handling of object appearance change. According to the subspace theory, the rank of the displacement matrix will remain unchanged when a new point belonging to the object is added into the matrix. Based on this idea the appearance model can be updated more robustly by considering only the points that would not change the rank of the displacement matrix.

Comparative experiments of this method are described in the paper, qualitatively and quantitatively. Our conclusion is that, the subspace motion model is capable to describe the image deformation brought by motion of non-plane-like object, and the particle distribution from the PF process could cover the probable object motion states more accordantly. The feasibility of our approach has been effectively demonstrated.

[1] G. H. Golub and C. F. Van Loan. *Matrix Computations 3rd Ed.* Johns Hopkins University Press, 1996.
[2] M. Irani. Multi-frame correspondence estimation using subspace constraints. *Int. J. Computer Vision*, 48(3):173–194, 2002.
[3] M. Isard and A. Blakes. Condensation-conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.
[4] J. Shi and C. Tomasi. Good features to track. In *Proc. CVPR*, pages 593–600, 1994.
[5] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137–154, 1992.
[6] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):Article No.13, 2006.

Place recognition from disparate views

Rob Frampton
http://www.cs.bris.ac.uk/~frampton
Andrew Calway
http://www.cs.bris.ac.uk/~andrew

Visual Information Laboratory
University of Bristol



Figure 1: Illustration of our geometric place model

Visual place recognition methods which use image matching techniques have shown success in recent years [1, 2, 3, 6, 8, 9], with some systems operating in real-time on very large databases [4, 7], however their reliance on local features restricts their use to images which are visually similar and which overlap in viewpoint.

In this paper we suggest that a semantic approach to the problem would provide a more meaningful relationship between views of a place and so allow recognition when views are disparate and database coverage is sparse. As initial work towards this goal we present a system which uses detected street objects such as traffic lights and signs as basic features, which are used to generate a 2D geometric place model. We then score the similarity of a pair of models by extracting features which characterise the pair, and use distributions learned from training examples to compute the probability that they depict the same place. We also gain some semantic understanding of the relationship between the two views by estimating the position of each object, the ground plane and the relative pose of the cameras.

Geometry estimation. We use a geometry model inspired in part by the work of Hoiem *et al*. [5], which assumes the relative world height of each object class is known, from which we estimate the depths of the objects as well as the plane they sit on. Ultimately we generate a model like that shown in Figure 1 - essentially a top-down 2-dimensional view of the scene. This is the place model which we use to perform place recognition.

Hypothesis scoring. If the place model is a good approximation of the scene geometry, it should be the case that when we are given two images of the same place, their models look very similar; indeed, if we had perfect measurements, they would be related by just a rotation and translation. With this in mind, we extract a set of features $\mathcal{F}$ from the place models and compute the probability $p(C|\mathcal{F})$, where C is the event that the two images depict the same place. The problem is now treated as a machine learning problem; distributions $p(C|\mathcal{F}_i)$ for each feature are estimated from training data and are used to compute the final probability during testing.

Since we do not know the correspondence of objects between images, we must evaluate each correspondence hypothesis separately. Due to the risk of generating an intractable number of hypotheses, we only consider 5-object correspondences between views and only use the top 10 ranking objects according to the detector confidence scores. We evaluate $p(C|\mathcal{F})$ for all possible 5-object correspondences and use the highest hypothesis probability as the probability that the images depict the same place.

We use five features to score each hypothesis. Two features are based on geometric measurements of the place models. We also compute a rectified ground plane image from the source images and extract edges to quantify their visual similarity, which is used as a feature. The last two features are the residual on the ground plane estimate and the object confidence scores given by the detector. Probabilities are estimates from distributions learned during a training stage, and a simple naive Bayes classifier is used to obtain the final probability.

Experimental results. We evaluated the system by performing a standard place recognition experiment on a dataset of 40 urban locations, with about three viewpoints per location. The system achieved a recognition rate of up to 73.1%. We also observed that some discriminative ability of the system is provided by the different object classes, and whilst this is a legitimate place recognition scenario, we wanted to observe the discriminative ability of the features alone. Thus, we also tested the system on a subset of the dataset with 30 locations, all of which contained the same two object classes, meaning that almost every image was capable of valid object correspondences with every other image. The system achieved a recognition rate of 67.9% in this harder scenario. Figure 2 shows an example of a pair of images correctly identified by our method as representing the same place, as well as the estimated geometry given by the system.

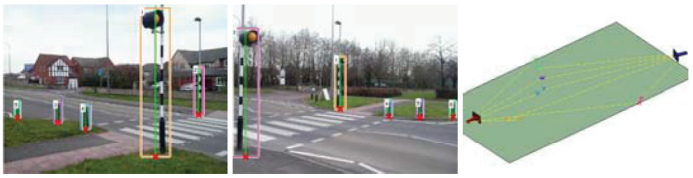


Figure 2: An example of successful output from our system

Due to the lack of similar work against which to compare our result, we also designed a place recognition task for human participants with the aim of providing a benchmark against which to compare our system. Unsurprisingly we found that humans are in general able to perform better (Figure 3), however about a third of participants performed worse than our system - indicating that the problem is not trivial to solve. The experiment also indicated that humans are more likely to use other distinctive semantic cues to discriminate between places rather than the purely geometric approach of our system, which will help to guide future work.

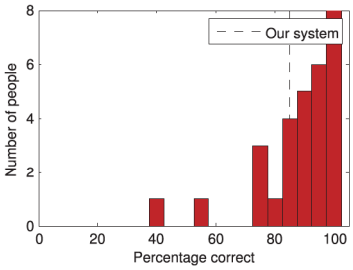


Figure 3: The results of the human experiment for humans and our system.

[1] Adrien Angeli, David Filliat, Stephane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, October 2008.

[2] S. Bazeille and D. Filliat. Combining odometry and visual loop-closure detection for consistent topo-metrical mapping. *RAIRO - Operations Research*, 44(4):365–377, January 2011.

[3] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, June 2008.

[4] Mark Cummins and Paul Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Robotics Science and Systems*, pages 1–8, Seattle, USA, 2009.

[5] Derek Hoiem, Alexei a. Efros, and Martial Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, April 2008.

[6] K. Konolige, J. Bowman, J. D. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua. View-based maps. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.

[7] M. Labbe and F. Michaud. Appearance-based loop closure detection for online large-scale and long-term operation. *Robotics, IEEE Transactions on*, PP(99):1–12, 2013.

[8] David Nistér and Henrik Stewénius. Scalable recognition with a vocabulary tree. In *CVPR (2)*, pages 2161–2168, 2006.

[9] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. *CVPR*, pages 1–7, June 2007.

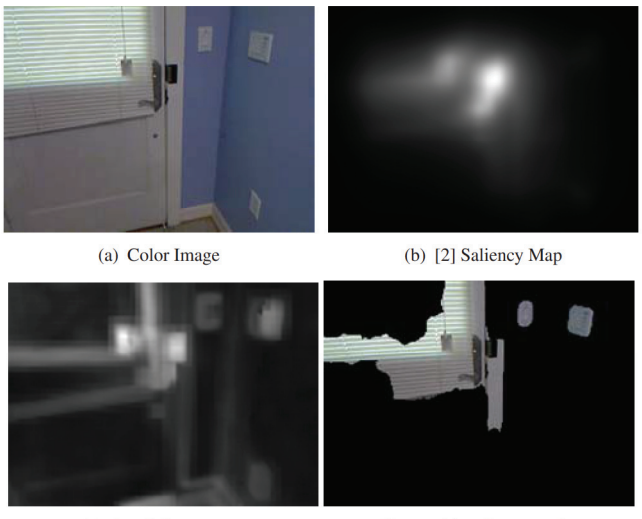
An In Depth View of Saliency

Arridhana Ciptadi
arridhana@gatech.edu
Tucker Hermans
thermans@cc.gatech.edu
James M. Rehg
rehg@gatech.edu

Center for Robotics and Intelligent Machines
School of Interactive Computing
Georgia Institute of Technology

Visual saliency is a computational process that identifies important locations and structure in the visual field. Most current methods for saliency rely on cues such as color and texture while ignoring depth information, which is known to be an important saliency cue in the human cognitive system. We propose a novel computational model of visual saliency which incorporates depth information. We compare our approach to several state of the art visual saliency methods and we introduce a method for saliency based segmentation of generic objects. We demonstrate that by explicitly constructing 3D layout and shape features from depth measurements, we can obtain better performance than methods which treat the depth map as just another image channel. Our method requires no learning and can operate on scenes for which the system has no previous knowledge. We conduct object segmentation experiments on a new dataset of registered RGB-D images captured on a mobile-manipulator robot.

Standard approaches to saliency use color, gradient, and intensity differences to distinguish unique regions from the rest of the visual field. We propose a novel method which incorporates depth measurements into the computation of visual saliency. Human subject studies have shown that depth is an important cue in determining salient regions in human visual processing [5, 6]. Depth measurements make it possible to separate objects which may be similar in appearance. In addition, shape information can be recovered from the depth channel and used to improve the discriminability of scene elements.



(a) Color Image (b) [2] Saliency Map

(c) Our Saliency Map (d) Resulting Segmentation

Figure 1: An example saliency map and resultant segmentation produced from the color and depth image pair. Note the two light switches in the upper right of the image, which our approach highlights as salient regions.

One motivation for saliency research is the development of generic object segmentation and detection capabilities [1, 3, 4, 7]. As an example, consider a personal robot that can move through a home environment and manipulate objects. During a clean-up task, the robot should be able to handle unfamiliar objects for which it has no prior experience. Saliency provides a basic mechanism for characterizing an unfamiliar scene and generating hypotheses about potential object locations. We use the task of generic object segmentation to quantify the effectiveness of our depth-based saliency method. We present a novel segmentation approach that incorporates saliency in an MRF model defined over superpixels. We demonstrate improved performance over several previous approaches.

Consider the images presented in Figure 1 captured on a robot operating in a home environment. The scene has multiple foreground objects of potential interest. Our saliency method scores all the relevant objects (door handle, light switches, window blind) more highly than the background elements of the scene. The resulting saliency-based segmentation successfully separates the foreground objects from the background producing a useful set of proto-objects for a robot to explore or manipulate.

We have collected a new dataset of color and depth images, which form the basis for our experimental evaluation. The dataset was collected using a mobile-manipulator robot in a real-world home environment. The dataset includes ground truth pixel-level segmentations of salient objects that we will release publicly. In summary, this work makes four contributions:

- We introduce a new method for estimating visual saliency, which combines color and depth measurements.
- We demonstrate that explicit 3D layout and shape features from depth measurements produce more informative saliency maps than approaches which simply treat depth as another channel of the image.
- We present an approach to saliency-based segmentation of generic objects based on a superpixel MRF and show promising results.
- We introduce a new dataset for depth-based saliency, including ground truth pixel-level segmentations of salient objects

[1] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, 2010.

[2] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Neural Information Processing Systems*, 2006.

[3] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[4] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[5] Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth Matters: Influence of Depth Cues on Visual Saliency. In *Proceedings of the European Conference on Computer Vision*, 2012.

[6] Jan Theeuwes, Paul Atchley, and Arthur F. Kramer. Attentional control within 3-d space. *Journal of Experimental Psychology: Human Perception and Performance*, 24(5):1476–1485, 1998.

[7] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

Insulin Granule Segmentation in 3-D TEM Beta Cell Tomograms

David Nam[1]
D.Nam@bristol.ac.uk
Judith Mantell[2,3]
J.Mantell@bristol.ac.uk
David Bull[1]
Dave.Bull@bristol.ac.uk
Paul Verkade[2,3,4/shared last author]
P.Verkade@bristol.ac.uk
Alin Achim[1]
Alin.Achim@bristol.ac.uk

[1] Visual Information Laboratory
University of Bristol, UK
[2] Wolfson Bioimaging Facility
University of Bristol, UK
[3] School of Biochemistry
University of Bristol, UK
[4] School of Physiology and Pharmacology
University of Bristol, UK

To obtain accurate volumetric information from 3-D transmission electron microscopy tomography data of beta cells, we propose an automated segmentation algorithm for insulin granules. Due to the image acquisition procedure, granule segmentation is especially challenging. Typically, insulin granules are described as organelles containing a dense core, surrounded by a halo and an enclosing membrane Fig. 1. Analysis of insulin granules in transmission electron microscopy (TEM) beta cell images is usually done manually. Currently, it is not feasible to analyse large numbers of 3-D TEM beta cell images —as manual segmentation is very time consuming.

3-D electron microscopy tomograms of beta cells are obtained using a similar procedure as for 2-D tomograms. Firstly, the sample has to be fixed using either a chemical or high pressure freezing fixation method. The imaging procedure gives relatively poor contrast and has the greatest effect on granule halos, which is demonstrated in Fig. 1. We tackle the problem of segmenting granules using active contours. We use a region-based active contour segmenting cores and halos. We chose a region-based approach to cope with weak boundaries within the images; the active contour incorporates region variances [3]. For halo segmentation we introduce a novel boundary prior to the Bayesian region-based active contour. To prevent the contour from moving into the core we propose the following; given a level set function ϕ, our boundary prior energy functional is as follows:

$$E_{boun} = -log\left(1 - H\left(\int (C_{core} - H_\varepsilon(\phi))C_{core}\,dx\right)\right). \quad (1)$$

Where C_{core} is the granule core, H is the Heaviside function and H_ε is a smoothed Heaviside function, as shown in [1]. We include our boundary term with the region-based energy term as well as smoothing and signed distance regularisers, to produce the following energy functional:

$$E(\phi) = \sum_{i=1}^{2} \frac{1}{2} \int \left(log(\sigma_i^2) + \left(\frac{(I-\mu_i)^2}{\sigma_i^2} \right) \right) M_i(\phi)\,dx$$
$$+\nu \int |\nabla H_\varepsilon(\phi)|\,dx + E_{boun} + \lambda \int \frac{1}{2}(|\nabla\phi| - 1)^2\,dx. \quad (2)$$

Where I is the image, $M_1(\phi) = H_\varepsilon(\phi)$, $M_2(\phi) = 1 - H_\varepsilon(\phi)$, ν and λ are constants and $\mu_i(\phi)$ and $\sigma_i^2(\phi)$ are the mean and variances —of the regions inside and outside the contour— respectively.

To help prevent the contour from getting stuck in local minima during its evolution, we utilise the dual active contour framework [4]. The dual energy term is an internal energy, which means it is not influenced by image information. This results in an uniform attraction between the contours. This would be ideal if the granule membrane is always located in the centre of the two contour initialisations, however this is not always the case. In some cases, granules have a membrane which is slightly visible. We introduce a gradient based term to help take advantage of this information and use it to slow the evolution of a contour around the granule membrane:

$$g = \frac{1}{\left(1 + |\nabla G_{\sigma_s} * I|^4 C_{coread\,j}\right)}, \quad (3)$$
$$C_{coread\,j} = 1 - \frac{G_{\sigma_s} * C_{core}}{max(G_{\sigma_s} * C_{core})}.$$

Where G_{σ_s} is a Gaussian smoothing kernel with standard deviation σ_s. Given two contours initialised outside and inside the granule, with corresponding level set functions ϕ and ψ respectively, we can define our dual energy functional as:

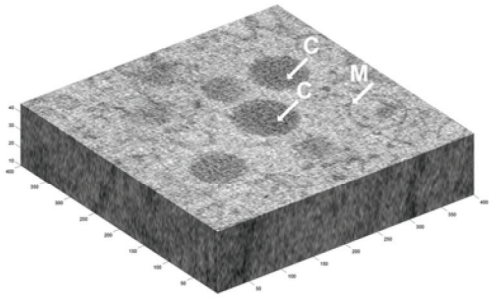


Figure 1: Section of a 3-D TEM data set of beta cell, $401 \times 401 \times 36$ pixels. The dark circles are the granule cores *C*. Note the weak membrane information *M* presented in the image.

$$E_{dual}(\phi,\psi) = E(\phi) + E(\psi) + \tau g \int [H_\varepsilon(\phi) + H_\varepsilon(\psi)]^2\,dx, \quad (4)$$

where τ is a weighting constant for the dual term.

Insulin granule segmentation begins with core segmentation, it then uses the results of the core segmentation as a starting point for halo segmentation. We first do 2-D core segmentation on each slice, to get an estimate of the core locations and shape. We utilise the region-based active contour with shape regulariser in [2] to do the core segmentations. The roundness at the vertices is not captured by doing the 2-D segmentation, also it does not produce a smoothly connected core. In order to correct for this and encourage continuity, we extend 2-D dual region-based active contour in to 3-D. Before we do active surface segmentation preprocessing is done to remove incorrect segmentations and connect disjoint cores. The result of this active surface is a smoother, more accurate, core segmentation. We then use the core as a starting point for halo segmentation. We utilise the dual active surface in (4). The core is used as the inner surface and a dilated version of the core is used at the outer surface.

Our proposed method is able to segment granule cores and membranes from 3-D TEM images of beta cells. Our novel active contour, when compared to other state-of-the-art active contours, gives a higher Dice's coefficient when compared to manual segmentations. With this automated method biologists will be able to acquire larger numbers of accurate and unbiased data of insulin granules.

[1] T.F. Chan and L.A. Vese. Active contours without edges. *IEEE Trans. Image Process.*, 10(2):266–277, Feb. 2001. ISSN 1057-7149. doi: 10.1109/83.902291.

[2] D. Nam, J. Mantell, D. Bull, P. Verkade, and A. Achim. Active contour based segmentation for electron microscopy images of insulin granule cores. *in Proc. EMBC, Annual International Conference of the IEEE, San Diego, California*, August. 2012.

[3] M. Rousson and R. Deriche. A variational framework for active and adaptative segmentation of vector valued images. *in Proc. Workshop Motion Vid. Computi.*, 2002.

[4] G. Zhu, Q. Zeng, and C. Wang. Dual geometric active contour for image segmentation. *Opt. Eng.*, 45(8):080505–1–3, 2006.

Plane-based Odometry using an RGB-D Camera

Carolina Raposo
http://arthronav.isr.uc.pt/~carolina
Miguel Lourenço
http://arthronav.isr.uc.pt/~mlourenco
Michel Antunes
http://www.isr.uc.pt/~michel
João P. Barreto
http://www.deec.uc.pt/~jpbar

Institute of Systems and Robotics,
Faculty of Sciences and Technology,
University of Coimbra,
3030 Coimbra, Portugal

Overview The advent of commodity RGB-D cameras led to intense research towards developing pipelines for estimating self-motion using simultaneously depth and vision cues [1, 2, 3]. State-of-the-art methods [2] are able to provide accurate motion estimations in real-time for the case of short displacement between consecutive frames. However, and since they use a differential registration approach, the performance degrades in the presence of dynamic scenes and/or wide baselines. Situations of wide baseline often arise because of high sensor velocity or increase in image resolution, with the frames taking more time to acquire and/or process.

This article describes a new odometry algorithm for RGB-D cameras that explores the fact of indoor environments being dominated by plane surfaces. The estimation of the relative pose between frames is cast as a plane registration problem that can be elegantly formulated in the dual projective 3D space as the successive estimation of a global rotation and projective scaling. The relative pose between image pairs is then refined by minimizing the photometric error in a set of planar patches.

While state-of-the-art methods use temporal information and motion priors to enforce smooth trajectories, we recover the camera motion by putting together the pairwise estimates without using any global optimization step. Several experiments with real data show that this naive approach is as accurate as the recent DVO algorithm [2] when the baseline is small, and provides substantially better motion estimates for the case of large baselines and/or dynamic foreground. These results suggest that plane features can be an advantageous alternative to point features for the purpose of visual odometry using RGB-D sensors.

Description of the pipeline Given two corresponding sets of 3 planes it is possible to compute the relative pose in a closed-form manner. The problem can be cast as finding the rotation R and translation $\mathbf{t}$ such that

$$\Pi_s^{(i)} \sim \begin{bmatrix} \mathsf{R} & \mathbf{0} \\ -\mathbf{t}^\mathsf{T}\mathsf{R} & 1 \end{bmatrix} \Pi_f^{(i)}, i = 1,2,3 \qquad (1)$$

verifies, where $\Pi_f^{(i)}$ and $\Pi_s^{(i)}$ are planes in the first and second reference frames, respectively, in homogeneous representation. Knowing that points and planes are dual entities in 3D, equation (1) can be seen as a projective transformation in $\mathcal{P}^{3*}$ that maps points $\Pi_f^{(i)}$ into points $\Pi_s^{(i)}$. It can be shown that R and $\mathbf{t}$ can be determined separately.

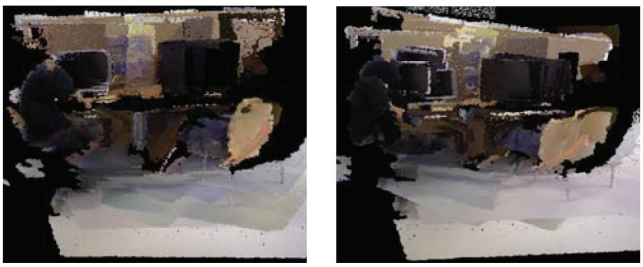
Planes are initially segmented and reconstructed using simultaneously depth and color information. The depth information is used in a RANSAC framework for suggesting plane hypotheses, followed by a plane merging procedure [4]. We reject planes corresponding to small segments, and re-run RANSAC for obtaining the final plane estimates.

Instead of explicitly matching planes between frames, we use an hierarchical scheme where corresponding sets with more planes are selected first. At least 3 non-parallel planes are necessary for computing R and $\mathbf{t}$. Situations where this is not verified are detected and, in case of ambiguity, the strictly necessary point matches are used to obtain an initial estimate.

Given the initial pose estimation, we use a photo-consistency framework to refine the camera motion and plane parameters. We optimize the motion and plane parameters by minimizing the cost function

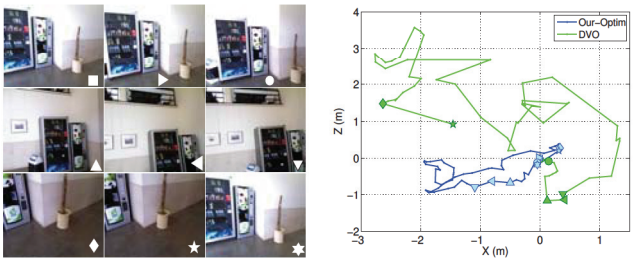
$$\varepsilon = \sum_{\mathbf{q}_f \in \mathcal{N}} \left[\mathtt{I}_s(\mathbf{w}(\mathbf{q}_f; \mathbf{p} + \delta\mathbf{p})) - \mathtt{I}_f(\mathbf{q}_f) \right]^2, \qquad (2)$$

where $\mathbf{w}(\mathbf{q}_f; \mathbf{p})$ is a 2D warping function and $\mathbf{p}$ the warping parameter vector that encodes 3 parameters for camera rotation, 3 for translation, and 3 for the plane structure. For robustness against a noisy camera motion initialization, we use a coarse-to-fine registration framework.



(a) Our-Optim (b) DVO [2]

Figure 1: 3D reconstruction for a dynamic dataset. The figure shows the 3D reconstruction computed with 1(a) our method with optimization, and 1(b) the DVO method. The quality of the 3D reconstruction indicates that our algorithm is more robust than DVO for scenes with dynamic motion.



(a) Our-Optim (b) DVO [2]

Figure 2: Loop closing experiment with $N = 59$ images. 2(a) shows some images of the sequence and 2(b) shows the estimated trajectories. Our algorithm enables to keep a reliable trajectory estimation, with a consistent smooth transition between frames. The DVO method diverges after the first couple of frames, providing an erroneous trajectory.

Conclusion We conclude that plane-based registration for relative camera motion estimation is advantageous with respect to point-based registration because: (i) plane-primitives have a more global character, which helps avoiding local minima issues, (ii) scenes are often dominated by large planes, allowing correspondences between wide-baseline frames, (iii) plane primitives are typically in the static background, improving odometry robustness to possible dynamic foreground, and (iv) the fact that the number of plane-features is much smaller than point-features, favors faster correspondence and scalability under increasing image resolution. We show that our algorithm outperforms [2] for large baselines, while keeping similar performance for small baseline sequences.

[1] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard. In *IEEE-ICRA*, 2012.

[2] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for rgb-d cameras. In *IEEE-ICRA*, 2013.

[3] F. Steinbruecker, J. Sturm, and D. Cremers. Real-time visual odometry from dense rgb-d images. In *IEEE-ICCV Workshop on Live Dense Reconstruction with Moving Cameras*, 2011.

[4] Camillo J. Taylor and Anthony Cowley. Parsing indoor scenes using rgb-d imagery. In *Robotics: Science and Systems*, 2012.

Reflective Symmetry Detection by Rectifying Randomized Correspondences

Zhongwei Tang
tang@cmla.ens-cachan.fr
Mariano Tepper
mariano.tepper@duke.edu
Guillermo Sapiro
guillermo.sapiro@duke.edu

Department of Electrical and Computer Engineering
Duke University
130 Hudson Hall
Box 90291
Durham, NC 27708, USA

Images often contain multiple reflective symmetries. We propose a method to detect multiple reflective symmetries at different scales and viewpoints based on the J-linkage framework [5], which combines ideas from RANSAC (robustness against outliers) and the Hough transform (multiple models detection through voting schemes).

How J-linkage works in symmetry detection ? Given N reflective matches $\mathcal{M} = \{m_i\}_{i=1,\cdots,N}$ in one image, we estimate K symmetries by randomly sampling K valid minimal seed sets from $\mathcal{M}$. We will see that two matches are sufficient to define a minimal seed set. We thus obtain K symmetries, each with an associated consensus set (the subset of compatible with each symmetry). A binary $N \times K$ matrix is thus built, where the entry (i, j) is 1 if the i-th match is in the consensus set of the j-th symmetry, and 0 otherwise. Each row of this matrix indicates which symmetries are preferred by each match and is considered as a binary feature vector for that match. Using these features, agglomerative hierarchical clustering based on the Jaccard distance is used to cluster the matches. Finally, each, large enough, cluster corresponds to a local symmetry.

How to rectify the distorted symmetry from two matches ? In practice, the symmetry can be observed in a slanted view and thus undergoes some perspective distortion. In this more general case, $m_1 = \{\mathbf{p}_1, \mathbf{p}_1'\}$ and $m_2 = \{\mathbf{p}_2, \mathbf{p}_2'\}$ intersecting at the vanishing point $\mathbf{v}$ are necessary to determine the symmetric axis (see Fig. 1(a)). We decompose the homography, which rectifies the distorted symmetry, into three parts:

$$\mathbf{H} = \mathbf{ARH}_0 = \mathbf{AR}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ h_{31} & h_{32} & 1 \end{bmatrix}. \tag{1}$$

where $\mathbf{H}_0$ and $\mathbf{A}$ are a projective transform and a shear, respectively, and $\mathbf{R}$ is a rotation. We will compute $\mathbf{H}$ and use it to bring the symmetry into a fronto-parallel setting. This process is depicted in Fig. 1(a). $\mathbf{H}_0$ is computed such that it is as close as possible to an identity matrix and sends the vanishing point to the infinity. Then, $\mathbf{R}$ and $\mathbf{A}$ further transform the symmetry to the fronto-parallel view.

The homography $\mathbf{H}$ is then applied to the keypoints of all the matches $m_i \in \mathcal{M}$. In this rectified plane, all the matches compatible with the reflective symmetry defined by m_1, m_2 should be parallel with the seed line segments $\overline{m}_1, \overline{m}_2$ and have their midpoint on the symmetric axis. This observation is used to define a simple criteria to select inliers to build the consensus set (see Fig. 1(b)).

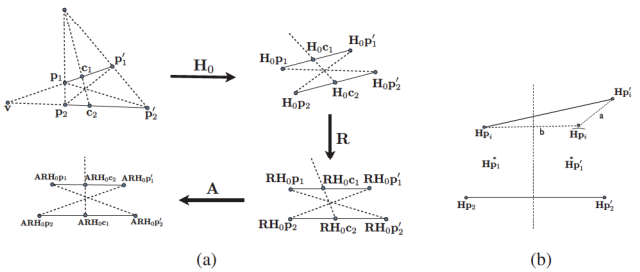


Figure 1: (a) Finding the rectification homography ($\mathbf{H} = \mathbf{ARH}_0$) from a pair of matches. (b) m_i is an inlier match if $\frac{a}{b}$ is smaller than some threshold η .

Adaptively sampling seed matches. One of the keys for the success of the J-linkage algorithm is choosing a proper non-uniform sampling. The rationale behind this is the intent to oversample the *true* symmetries in the image, thus obtaining stable row features that robustify the clustering process. We use the following adaptive non-uniform sampling strategy. The first match is sampled according to the following mixed probability, $\forall m_i = \{\mathbf{p}_i, \mathbf{p}_i'\} \in \mathcal{M}$,

$$\Pr(m_i) = \frac{1}{Z_1}\exp\left(-\frac{1}{\sigma_d^2}(\|\mathrm{desc}(\mathbf{p}_i) - \mathrm{desc}(\mathbf{p}_i')\| - d_0)^2 - \frac{1}{\sigma_l^2}(\|\mathbf{p}_i - \mathbf{p}_i'\| - l_0)^2\right) \tag{2}$$

where d_0 and l_0 indicate the scale in the descriptor domain and image domain at which we prefer to detect the symmetry, $\mathrm{desc}(\mathbf{p}_i)$ is the descriptor of keypoint $\mathbf{p}_i$, Z_1 is a normalization factor, and, finally, σ_d and σ_l decide how strict the preferences are.

Since an image can contain multiple symmetries at different scales, we need to update the parameters of Eq. (6) to find all possible symmetries. Each time two seed matches are sampled and its consensus set $\mathcal{M}_0$ is computed, we decrease the probability of all these matches by a factor κ close to 1, followed by the renormalization of the probability:

$$\Pr(m_i) \leftarrow \kappa \Pr(m_i),\ m_i \in \mathcal{M}_0. \tag{3}$$

Once the first match is sampled following the above adaptive non-uniform sampling, we sample the second match according to the conditional probability:

$$\Pr(m_j|m_i) = \frac{1}{Z_2}\exp\left(-\frac{1}{\sigma_c^2}(\|\mathbf{c}_i - \mathbf{c}_j\| - c_0)^2\right),\ m_i \in \mathcal{M},\ m_j \in \mathcal{M} \setminus \{m_i\}, \tag{4}$$

where Z_2 is a normalization factor such that $\sum_j \Pr(m_j|m_i) = 1$, $\mathbf{c}_i$ is the midpoint of line segment $\overline{m}_i$, and σ_c, c_0 control the shape of the probability function.

If the current consensus set $\mathcal{C}$ is big enough, i.e., $|\mathcal{C}| > T$ for some T, we switch to uniformly sample K_u seeds *inside* $\mathcal{C}$. This ensures that good models are oversampled.

	Synthetic Single				Synthetic Multiple			
	LE [3]	LHXS [2]	CL [1]	Proposed	LE [3]	LHXS [2]	CL [1]	Propc
TP/GT	92%	62%	100%	**100%**	35%	28%	77%	**67**
FP/GT	15%	0%	15%	**0%**	4%	8%	33%	**10**

	Real Single				Real Multiple			
	LE [3]	LHXS [2]	CL [1]	Proposed	LE [3]	LHXS [2]	CL [1]	Propc
TP/GT	84%	29%	94%	**97%**	43%	18%	68%	**65**
FP/GT	68%	3%	69%	**39%**	44%	0%	17%	**16**

Table 1: Performance comparison of several methods on the PSU dataset [4]. TP, FP, and GT respectively denote the number of true positives, false positives, and ground truth symmetries. The percentage of the methods are taken from [4] and [1].

Experiments. We compare the proposed method with three recent ones [1, 2, 3] on the PSU dataset [4], which is composed of 88 images. The results in Table 1 show that our method is much better than Loy and Eklundh's (LE [3]) and Liu *et al.*'s (LHXS [2]). Compared with Cho *et al.*'s (CL [1]) method, we have better performance for single symmetry detection. For multiple symmetries detection, our method is more conservative, sometimes detecting fewer symmetries, which explains the decrease in both the true positive and false positives rates. This conservative strategy can be explained by the strict criterion used to only create precise consensus sets. Nonetheless, the proposed method results are highly competitive.

[1] M. Cho and K. Mu Lee. Bilateral symmetry detection via symmetry-growing. In *BMVC*, 2009.
[2] Y. Liu, J. Hays, Y-Q. Xu, and H-Y. Shum. Digital papercutting. Technical Sketch. In *SIGGRAPH*, 2005.
[3] G. Loy and J-O. Eklundh. Detecting symmetry and symmetric constellations of features. In *ECCV*, 2006.
[4] S. Lee M. Park, P-C. Chen, S. Kashyap, A. Butt, and Y. Liu. Performance evaluation of state-of-the-art discrete symmetry detection algorithms. CVPR, 2008.
[5] R. Toldo and A. Fusiello. Robust multiple structures estimation with j-linkage. *ECCV*, 2008.

An Iterative 5-pt Algorithm for Fast and Robust Essential Matrix Estimation

Vincent Lui

Tom Drummond

Department of Electrical & Computer Systems Engineering, Monash University

The essential matrix, first introduced by Longuet-Higgins [5], is a 3×3 matrix encoding the relative pose information between two views. Conventional approaches for relative pose estimation is to solve a system of linear equations. The 5-pt algorithm [6] is the current state-of-the-art algorithm in relative pose estimation. It is a minimal-set direct solver which solves the essential matrix as a system of polynomial equations.

We show in this paper an iterative method which provides robust and real-time essential matrix estimation, capable of 30Hz, frame-rate performance. The benefit of an iterative approach lies in its simplicity and speed. The use of high degree polynomials may lead to ill-conditioning [1] and are often difficult to solve, leading to alternative methods which sacrifice speed for simplicity [1, 3, 4]. Although convergence is not guaranteed, when used within RANSAC, more hypotheses can be evaluated in the same block of time, yielding improved performance. While iterative solvers which minimizes the algebraic epipolar reprojection error have previously been proposed [2, 7], our parametrization is based on a novel geometric error which incorporates the half-plane constraint [8] and thus enforces orientation consistency between points.

Figure 1 illustrates the concept of our iterative 5-pt algorithm. A coordinate frame is chosen such that the z-axis e_z joins the two camera centres. In this frame, vectors $\hat{v}_i$ and $\hat{v}'_i$ and e_z are coplanar. The goal is to find a rotation for each of the two cameras that maps from their internal coordinate frame to that of Figure 1. We parametrize the image for each camera as a unit 2-sphere, mapping image points in normalized camera coordinates $[x, y, 1]^T$ to unit vectors by dividing by $\sqrt{x^2+y^2+1}$.

At each iteration, the normalized point correspondences $\hat{u}_i \leftrightarrow \hat{u}'_i$ are left multiplied with the rotations R and R', giving the rotated point correspondences $\hat{v} = R\hat{u}, \quad \hat{v}' = R'\hat{u}'$. This rotates the two unit spheres, changing the direction of the epipoles. By rotating the unit spheres such that the z-axis e_z is aligned with the epipoles, *i.e.* $e_z = Re = R'e'$, $\hat{v}_i \leftrightarrow \hat{v}'_i$ become coplanar with the epipoles e, e'. The epipoles can then be computed as

$$e = R^T e_z, \quad e' = R'^T e_z. \tag{1}$$

The error function to be minimized is based on the idea that if R and R' are incorrect, there will be a non-zero angle between the plane defined by $\hat{v}_i$ and e_z and the plane defined by $\hat{v}'_i$ and e_z. We compute this angle by projecting $\hat{v}_i$ and $\hat{v}'_i$ onto the x-y plane, using the atan2() function:

$$A(\hat{v}_i) = \text{atan2}(y_i, x_i), \tag{2}$$

where $[x_i\ y_i\ z_i]^T$ is the 3-vector of a rotated point $\hat{v}_i$. The residual error associated with each rotated point correspondence is then given as

$$r_i = A(\hat{v}_i) - A(\hat{v}'_i), \tag{3}$$

We constrain R' to have no component of rotation about its z-axis as two rotations provide 6 degrees of freedom (DOF) but the essential matrix has only 5DOF—there is a rotational gauge freedom about the joint z-axis e_z. This gives 5 parameters α_j, where α_{1-3} represents R and $\alpha_{4,5}$ represents R'. At each iteration, the rotational parameters are updated as

$$R^{k+1} = e^{\sum_{i=1}^{3} \alpha_i G_i} R^k, \quad R'^{k+1} = e^{\alpha_4 G_1 + \alpha_5 G_2} R'^k. \tag{4}$$

where G_i represents the SO(3) Lie group generators. To update the rotation parameters, the Jacobian of the error relative to α_{1-5} has to be computed. Representing the projected point p_i as the 2-vector $[x_i, y_i]$, the Jacobian is given as

$$J = \frac{(x_i)((G_j\hat{v}_i)[1]) - (y_i)((G_j\hat{v}_i)[0])}{x_i^2 + y_i^2}. \tag{5}$$

Using the LM algorithm, the change to be applied to the rotational parameters α_j is $\delta = (J^T J + \lambda I)^{-1} J^T r$. This process is repeated iteratively until the algorithm converges. Using our iterative 5-pt algorithm as a hypothesis generator within RANSAC, the essential matrix can then be recovered simply from the best hypothesis, with the translation $\hat{t}$ given by the second epipole e', and the rotation matrix given by $R'^T R$.

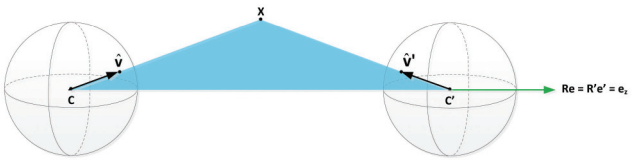


Figure 1: Visual concept of the iterative 5-pt algorithm. Image planes are represented as unit spheres, constraining the epipole to lie on the sphere's surface. Here, the rotated point pairs are co-planar with the epipoles.

In our approach, detected inliers can be easily triangulated using cylindrical coordinates. The vectors $\hat{v}_i$ and $\hat{v}'_i$ can be transformed into cylindrical coordinates by dividing by $\sqrt{x_i^2+y_i^2}$ and $\sqrt{x_i'^2+y_i'^2}$ respectively. In this coordinate frame, $||x_i, y_i|| = 1$ and $||x'_i, y'_i|| = 1$, so the inliers should have $x_i = x'_i$ and $y_i = y'_i$ since inliers have $[x_i, y_i]$ pointing in the same direction as $[x'_i, y'_i]$. The disparity $|z_i - z'_i|$ now encodes the cylindrical radial distance to the point, so setting

$$P_i = \frac{\hat{v}_i}{|z_i - z'_i|} \quad \text{and} \quad P'_i = \frac{\hat{v}'_i}{|z_i - z'_i|} \tag{6}$$

gives the coordinates of the triangulated point in the coordinate frame of Figure 1. This calculation guarantees that $p_i = p'_i \pm e_z$ because the vector e_z represents the motion between the two cameras in the frame of $\hat{v}_i$ and $\hat{v}'_i$. The coordinates of the triangulated point in the original camera frames can then be recovered by simply multiplying P_i and P'_i by R^T and R'^T, *i.e.* $Q_i = R^T P_i$ and $Q'_i = R'^T P'_i$.

Full details of our iterative 5-pt algorithm, including the derivation of the Jacobian, implementation details, and comprehensive experiments. In conclusion, the simplicity of the method, coupled with the ability to easily triangulate inlier points, makes it a useful method for relative pose estimation.

[1] Dhruv Batra, Bart Nabbe, and Martial Hebert. An Alternative Formulation for Five Point Relative Pose Problem. In *Proc. WMVC*, pages 21–27, 2007.

[2] Uwe Helmke, Knut Hüper, Pei Yean Lee, and John Moore. Essential Matrix Estimation Using Gauss-Newton Iterations on a Manifold. *IJCV*, 74(2):117–136, January 2007.

[3] Li Hongdong and Richard Hartley. Five-Point Motion Estimation Made Easy. In *Proc. ICPR*, pages 630–633, 2006.

[4] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Polynomial Eigenvalue Solutions to the 5-pt and 6-pt Relative Pose Problems. In *Proc. BMVC*, 2008.

[5] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–134, 1981.

[6] David Nistér. An efficient solution to the five-point relative pose problem. *TPAMI*, 26(6):756–77, June 2004.

[7] Edward Rosten, Gerhard Reitmayr, and Tom Drummond. Improved RANSAC performance using simple, iterative minimal-set solvers. 2010.

[8] Tomas Werner and Oxford Ox. Constraint on Five Points in Two Images. In *Proc. CVPR*, pages 2–7, 2003.

Reduced Dimensionality Extended Kalman Filter for SLAM

Dinesh Gamage
http://eng.monash.edu.au/ecse/about/people/profile/dineshg
Tom Drummond
http://eng.monash.edu.au/ecse/about/people/profile/twd
Department of Electrical and Computer Systems Engineering, Monash University, Australia.

Computational complexity of the Kalman filter grows at least quadratically with the number of dimensions in the filter. This is a particular problem for applications like monocular simultaneous localization and mapping (SLAM) where it is not possible to run a single filter on a large map with many thousands of landmarks. The filtering approach for SLAM, maintains only the current camera pose with all landmarks of interest as the state [2].

This paper presents a method for reducing the computational complexity of the Kalman filters by reducing the dimensionality as information is acquired. The method reduces the dimensionality of the extended Kalman filter (EKF) for SLAM by identifying dominant modes of the filter. This can be used in general to reduce the dimensionality of the EKF irrespectively of its application, without being limited to SLAM.

For a filter with zero process noise, the mean of this distribution can be represented as a point in a nD space with its uncertainty as a hyper ellipse. Further information will reduce this uncertainty along some directions. After some time uncertainty along many directions becomes comparatively small, making further information redundant. So the covariance matrix can be decompose into certain and uncertain dimensions.

$$\Sigma_{t_1} = UDU^T \tag{1}$$

where U is the set of singular vectors of Σ_{t_1} and D is the diagonal matrix of singular values. The singular vectors represent de-correlated uncertainty directions with variances proportional to their singular values. We partition D into D_s the significant (large) set and D_i the insignificant (small) set. U is also partitioned into U_s vectors corresponding to D_s and U_i vectors corresponding to D_i. With this Σ_t can now be written in block form:

$$\Sigma_{t_1} = \begin{bmatrix} U_s \mid U_i \end{bmatrix} \left[\begin{array}{c|c} D_s & 0 \\ \hline 0 & D_i \end{array}\right] \left[\begin{array}{c} U_s^T \\ \hline U_i^T \end{array}\right] \tag{2}$$

So the original covariance matrix becomes:

$$\Sigma_{t_1} = U_s D_s U_s^T + U_i D_i U_i^T \tag{3}$$

As the second term of the above expression is comparatively small the column space of U_s can be used as the reduced space ignoring small singular vectors. Let the significant vectors U_s extracted at $t = t_1$ be U_{st_1} and x'_t be the reduced state which is Gaussian distributed in the reduced space with a mean μ'_t at $t(> t_1)$. The reduced mean μ'_t relates to the original state mean μ_t as:

$$\mu_t = \mu_{t_1} + U_{st_1}\mu'_t \tag{4}$$

This new state x'_t represents the variations of the original state around the point μ_{t_1} along the directions of the column space of U_{st_1}. Initially we start with μ_{t_1} being a zero vector, indicating our knowledge about uncertainty is zero along corresponding singular vectors in U_{st_1}. Subsequent observations can be projected onto the derived reduced space to gather information about the new state x'_t. The projected covariance matrix Σ_{t_1} is obtained as:

$$\Sigma'_{t_1} = U_{st_1}^T \left(\Sigma_{t_1} - U_i D_i U_i^T\right) U_{st_1} \tag{5}$$

The dimensionality of x'_t can be kept quite small, compared to the original state x_t. For all time steps $t(\geq t_1)$, information can be collected to update x'_t by changing its mean and the covariance. This makes the reduced state time dependent. To obtain the prediction equation in the reduced space, the linearized EKF states at time steps t and $t-1$ can be decomposed according to the equation 4. As we are assuming zeros process noise, the process model becomes the identity. Substituting the decomposed states into the process equation yields:

$$\mu_{t_1} + U_{st_1}\mu'_t = \left(\mu_{t_1} + U_{st_1}\mu'_{t-1}\right) \tag{6}$$

making the predicted state same as the previous state:

$$\mu'_t = \mu'_{t-1} \tag{7}$$

Similarly, if the measurement is z_t with the model Jacobian H_t and measurement noise v_t, update equation, after substitution becomes:

$$z_t = H_t\left(\mu_{t_1} + U_{st_1}\mu'_t\right) + v_t \tag{8}$$

which can be modified as:

$$z_t - H_t\mu_{t_1} = H_t U_{st_1}\mu'_t + v_t \tag{9}$$

Here $z_t - H_t\mu_{t_1}$ becomes the modified observation and $H_t U_{st_1}$ the projected Jacobian. The dimensionality reduction described so far works only when the state is static. In a SLAM setup, the camera pose keeps changing requiring a slightly different approach.

The computational savings achieved by reducing the dimensionality of the filter can be spent by admitting more variables into the filter to be measured (thus increasing its dimensionality again). A new variable l, can be directly added to the reduced state to obtain an augmented state. It has to be noted that this augmentation increases the dimensionality of the original space as well by the same number of dimensions. So the froze mean μ_{t_1} has to be augmented with zeros to get a modified mean $\hat{\mu}_{t_1}$.

With continuous augmentations the reduced state $x'_{k|k}$ also will start growing. To keep the dimensionality of $x'_{k|k}$ manageable, we decompose it continuously by distributing newly learned knowledge over the froze mean μ_{t_1} and select a new reduced basis.

To model the desired variation while retaining standard operations on the camera, here we perform reduction only upon landmark states. As the first step, landmarks observed up to time $t = t_1$ are decomposed, keeping the camera state intact by directly transferring it into the reduced state.

When landmarks are static, the prediction involves estimating the camera and its covariance with cross-covariance [1]. As we transfer the camera state directly in to the reduced state x'_t, camera parameters can be predicted in the usual manner while keeping x'_{y_t} unchanged to get the predicted state $x'_{t|(t-1)}$. The covariance block in the reduce covariance Σ'_t corresponding to the camera state has to be added with process noise to get the predicted covariance $\Sigma'_{t|(t-1)}$.

State update is done by projecting the measurement model onto the reduced space. Measurement model Jacobian H_t, can be projected onto reduced space by a right multiplication. If we drop the subscript t and denote the Jacobian as H:

$$H_s = HU_{st_1} \tag{10}$$

Using these construction, in this paper we introduce a dimensionality reduction technique to handle the complexity growth of the extended Kalman filter for SLAM. Though the Kalman filter is not the state of the art any more with current sparse matrix methods, we believe still the Kalman filter has a considerable potential generally in computer vision.

[1] J. Civera, A. Davison, and J. Martínez Montiel. 1-point ransac. *Structure from Motion using the Extended Kalman Filter*, pages 65–97, 2012.

[2] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, 2007.

Learning Smooth Pooling Regions for Visual Recognition

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

From the early HMAX model to Spatial Pyramid Matching [2, 4], spatial pooling has played an important role in visual recognition pipelines. By aggregating local statistics, it equips the recognition architectures with a certain degree of robustness to translation and deformation yet preserving spatial information. Despite of its predominance in current recognition systems, we have seen little progress to fully adapt the pooling strategy to the task at hand, and this critical decision is most prominently based on hand-crafted layouts.

We propose in this paper a flexible parameterization that allows for a richer set of possible pooling regions and show results on classification tasks using two different pipelines [1, 3]. The higher-level pooling representation is learned jointly with the classifier to support the recognition task. In order to deal with the increased flexibility of the model, we investigate different regularizers and efficient learning schemes. In particular, we propose a smoothness regularizer that yields the strongest performance improvements in our experiments.

The simplest form of the spatial pooling is computing histogram over the whole image. This can be expressed as $\Sigma(U) := \sum_{j=1}^{M} u_j$, where $u_j \in \mathbb{R}^K$ is an encoded patch extracted from the image (out of M such codes) and an index j refers to the spatial location that the code originates from[1]. Another popular pooling scheme that has been proven successful [5] is max-pooling: $\mathbb{M}(U) := \max_{j=1}^{M} u_j$. Since the pooling approach looses spatial information of the codes, Lazebnik et al. [2] proposed to first divide the image into subregions, and afterwards to create pooled features by concatenating histograms computed over each subregion. There are two problems with such an approach: first, the division is largely arbitrary and in particular independent of the data; second, discretization artifacts occur as spatially nearby codes can belong to two different regions as the 'hard' division is made.

In our paper we address both problems by using a parameterized version of the pooling operator

$$\Theta_w(U) := \rho_{j=1}^{M}(w_j \circ u_j) \tag{1}$$

where $a \circ b$ is the element-wise multiplication, and $\rho \in \{\max, \sum\}$ is a pooling function. Moreover, we have investigated a few regularization terms on the pooling weights showing that smooth pooling regions are crucial.

Consider a sampling scheme and an encoding method producing M codes each K dimensional. Every coordinate of the code is an input layer for the multilayer perceptron. Then we connect every j-th input unit at the layer k to the l-th pooling unit a_l^k via the relation $w_{lj}^k u_j^k$. Since the receptive field of the pooling unit a_l^k consists of all codes at the layer k, we have $a_l^k := \sum_{j=1}^{M} w_{lj}^k u_j^k$ or $a_l^k := \max_{j=1}^{M} w_{lj}^k u_j^k$. Next, we connect all pooling units with the classifier allowing the information to circulate between the pooling layers and the classifier (Fig. 1). The latter has an access to the class membership and can use this information back to the pooling stage to shape better pooling regions. For the training purpose we have derived the backpropagation rules used for the weights' update.

To make the whole approach more scalable towards bigger dictionaries we introduce two approximations. The first one, called pre-pooling, uses standard pooling scheme to aggregate the codes over small neighbourhood before our joint training of the pooling regions together with the classifier's parameters is applied. The second approximation (batches) divides a K dimensional code into $\frac{K}{D}$ batches, each D dimensional. The latter enables embarrassingly parallel training of the model with sizable dictionaries. Our implementation of the proposed method is publicly available at `http://www.d2.mpi-inf.mpg.de/datasets`.

We have evaluated our method on two classification datasets CIFAR-10 and CIFAR-100 following Coates and Ng [1] pipeline, and UIUC sports events following Li-Jia et al. [3].

[1] That is $j = (x, y)$ where x and y refer to the spatial location of the center of the extracted patch.

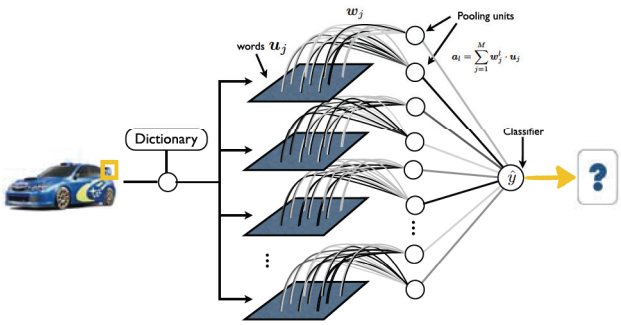
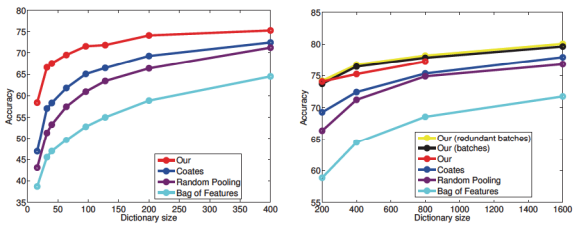


Figure 1: Sketch of our architecture. We encode the patches extracted from the images using popular encoding method. Next, we couple every position of such encoded patches with the classifier via the pooling weights. Our method learns both the pooling weights and classifier's parameters at the same time by using the backpropagation rule.

Figures 2(a) and 2(b) show the classification accuracy of our model against the baseline [1] on CIFAR-10. Our method outperforms the approach of Coates by 10% for dictionary size 16 (our method achieves the accuracy 57.07%, whereas the baseline only 46.93%). For bigger dictionaries (1600) with an accuracy for the batched model of 79.6% we outperform the Coates baseline by 1.7%. On CIFAR-100 our model achieves 56.29% outperforming the baseline by 4.63%. Lastly, we investigate events recognition on the UIUC Sports database based on object bank features [3]. Our learnable pooling strategy achieves accuracy 79.4%, about 3.1% higher then the hand-crafted scheme used in Li-Jia et al. [3]. Our experiments show the importance of the optimized pooling strategy.



(a) (b)

Figure 2: Figure 2(a) shows accuracy of the classification with respect to the number of dictionary elements on smaller dictionaries. Figure 2(b) shows the accuracy of the classification for bigger dictionaries when our approximation is used (batches, and the redundant batches).

[1] A. Coates and A. Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[3] L. Li-Jia, S. Hao, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[4] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2009.

[5] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges

Philipp Werner[1]
Philipp.Werner@ovgu.de
Ayoub Al-Hamadi[1]
Ayoub.Al-Hamadi@ovgu.de
Robert Niese[1]
Robert.Niese@ovgu.de
Steffen Walter[2]
Steffen.Walter@uni-ulm.de
Sascha Gruss[2]
Sascha.Gruss@uni-ulm.de
Harald C. Traue[2]
Harald.Traue@uni-ulm.de

[1] Institute for Information Technology and Communications, Otto-von-Guericke University, Magdeburg, Germany
[2] Department for Psychosomatic Medicine and Psychotherapy, University of Ulm, Ulm, Germany

Pain is what the patient says it is. But what about these who cannot utter? Automatic pain monitoring opens up prospects for better treatment, but accurate assessment of pain is challenging due to the subjective nature of pain. To facilitate advances, we contribute a new dataset, the BioVid Heat Pain Database which contains videos and physiological data of 90 persons subjected to well-defined pain stimuli of 4 intensities. We propose a fully automatic recognition system utilizing facial expression, head pose information and their dynamics. The approach is evaluated with the task of pain detection on the new dataset, also outlining open challenges for pain monitoring in general. Additionally, we analyze the relevance of head pose information for pain recognition and compare person-specific and general classification models. For a visual overview, see Fig. 1.

BioVid Heat Pain Database: We introduce a newly collected experimental pain research dataset, named BioVid Heat Pain Database. It was recorded in a study with 90 participants aged from 18 to 65 years. In contrast to existing databases known to the authors, it does not only provide video information, but also physiological data. Thus, it promotes advancements in both, observational and physiological pain measurement and facilitates the fusion of both in one automatic recognition system. Other key features include well-defined pain stimulation through a thermode, stimulation of four pain intensity levels, recording of high resolution video from different points of view, recording of depth information, measurement of pain threshold and tolerance, posed and spontaneous emotion, and filled in questionnaires about the participants' personality and psychosomatic health. The database will be made available for non-commercial research purposes.

Pain Recognition System: A new approach to recognize pain from facial expression and head pose information is proposed. It utilizes dynamics through a time window descriptor which is calculated from frame level features. To the best knowledge of the authors, our pain recognition system is the first which takes up the challenges arising with fully automatic feature extraction when permitting non-frontal head poses. Lucey *et al.* [2] rely on manually labeled key-frames. Werner *et al.* [4] assume that 3D landmarks are available. Hammal and Kunz [1] and Niese *et al.* [3] rely on frontal view on the face. Our approach is also the first utilizing head pose information for classification. The head pose is estimated from depth data, and facial landmarks are detected from color images, both employing methods known from literature. Next, for each image frame a set of distance and gradient features is extracted. These are selected to capture pain related facial actions. Facial distances are calculated in 3D to uncouple them from head pose. To utilize dynamics, a time window consisting of multiple frames is considered. Each frame level feature forms a discrete-time signal which is condensed in a time window descriptor through statistical parameters of the signal and its derivations. The descriptors are used for classification with a support vector machine.

Experiments: We conduct experiments applying our recognition system to the BioVid Database. Whereas the others evaluate their systems based on observational measures (e.g. facial actions) or do not even specify the source of their ground truth labels, we compete with the gold standard of pain assessment, i.e. self-report. Since our focus is on recognizing pain, not only facial expressions of pain, we try to classify the stimulus. This does not only reveal strengths and weaknesses of the recognition

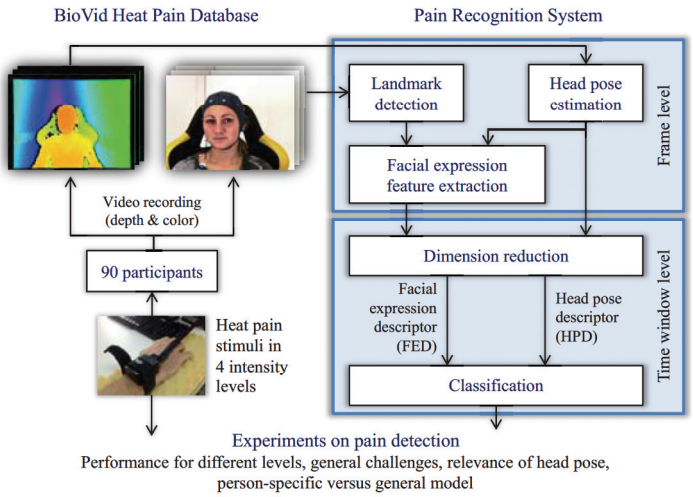


Figure 1: Overview on the new database, the recognition concept and the conducted experiments.

system, but also challenges caused by the nature of pain. First, we investigate which level of pain can be detected by our system. We obtain good results for the two highest intensities, whereas they are better for younger age. Lower pain intensities yield less observable reaction and are much harder to detect. In addition, there are large differences in expressiveness between persons. Whereas some persons already show considerable facial expressions at low pain intensities, other seem not to react at all, even when stimulated with high temperature at their pain tolerance level. Further, we discuss age related differences and the need for very accurate and robust landmark detection. In another experiment, we analyze the role of head pose for pain recognition. It is shown that head movement is not only a personal trait. When combined with facial expression information it can also improve predictive performance of a general model.

Future work includes research in how to handle the diversity of persons, and multi-modal pain recognition integrating physiological information.

[1] Z. Hammal and M. Kunz. Pain monitoring: A dynamic and contcxt-sensitive system. *Pattern Recognition*, 45(4):1265–1280, April 2012.

[2] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.

[3] R. Niese, A. Al-Hamadi, A. Panning, D. Brammen, U. Ebmeyer, and B. Michaelis. Towards pain recognition in post-operative phases using 3D-based features from video and support vector machines. *Int. Journal of Digital Content Technology and its Applications*, 2009.

[4] P. Werner, A. Al-Hamadi, and R. Niese. Pain recognition and intensity rating based on comparative learning. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2313–2316, Orlando, 2012.

Overlapping camera clustering through dominant sets for scalable 3D reconstruction

Massimo Mauro[1]
http://www.ing.unibs.it/~massimo.mauro/
Hayko Riemenschneider[2]
http://www.vision.ee.ethz.ch/~rhayko/
Luc Van Gool[2]
http://www.vision.ee.ethz.ch/~vangool/
Riccardo Leonardi[1]
http://www.ing.unibs.it/~leon/

[1] Department of Information Engineering
University of Brescia
Brescia, Italia

[2] Computer Vision Lab (CVL),
Swiss Federal Institute of Technology (ETH)
Zurich, Switzerland

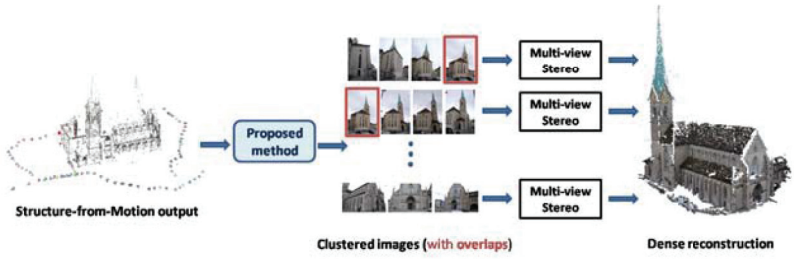


Figure 1: The overall scheme of a reconstruction using camera clustering.

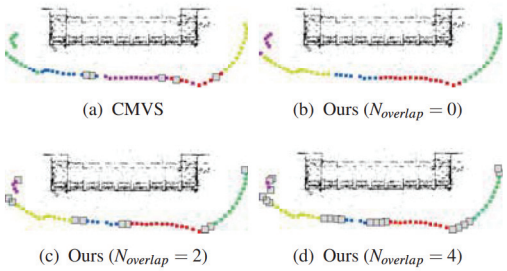


Figure 2: Camera clusters on the *Hall* dataset with overlaps in gray.

Scalability is a great issue in modern large-scale 3D reconstruction pipeline [1, 2]. Recently, *image clustering* [3, 5] and *image selection* [4, 7] methods have been developed for scaling both Structure-from-Motion (SfM) and Multi-View Stereo (MVS) algorithms. This paper focuses on MVS scalability and proposes a novel camera clustering method. Our technique produces a set of overlapping clusters suited to be processed independently by the consequent MVS (see Fig. 1). Overlaps are important to avoid the creation of holes between clusters and their benefit for the final reconstruction is analyzed in the paper.

Our clustering algorithm groups together cameras pointing to similar parts of the scene. We model our problem as a graph, where the nodes are the cameras and the edges represent the similarity between views. Given a set of N cameras and a 3D sparse point cloud obtained from SfM we define a $N \times N$ symmetric matrix W of camera similarities between as

$$w_{ij} = \frac{\sum_{p \in (V_i \cap V_j)} w_{\alpha_{ijp}}}{|\, V_i \cap V_j \,|} \tag{1}$$

where $w_{\alpha_{ijp}} = exp(-\frac{\alpha_{ijp}^2}{\sigma^2})$ depends on the angle α between the viewing directions. See full paper for details of the method.

Once the graph is constructed, we adopt the game-theoretic model of dominant sets for clustering [6]. Dominant set clustering works by iteratively evaluating the coherency of the graph nodes based on the similarity matrix W_0 (where W_0 is equal to the W matrix without any self-loops, i.e. with zeros along the main diagonal). The assignment of each camera is determined by the *participation vector* $\mathbf{x}$, which is of length N as the number of cameras. It contains the level of participation - in the range $[0, 1]$ - of each camera into clusters. The algorithm for finding a dominant set (a cluster) is the following:

1. **Initialize** all elements of $\mathbf{x}$ to $1/N$;
2. **Evolve** the system

$$x_i(t+1) = x_i(t) \frac{(W_0 \mathbf{x}(t))_i}{\mathbf{x}(t)^T W_0 \mathbf{x}(t)} \tag{2}$$

3. **Stop** when

$$\mathbf{x}(t+1)^T W_0 \mathbf{x}(t+1) - \mathbf{x}(t)^T W_0 \mathbf{x}(t) < \varepsilon \tag{3}$$

When the algorithm terminates, the vector $\mathbf{x}$ has zero values for non-relevant nodes and values above zero for nodes in the dominant set. We obtain a multi-cluster division by iteratively running the method on the remaining set of cameras, until no further separation is possible.

Dominant set clustering can be implemented with few lines of code and executed efficiently. A further main advantage of using dominant sets clustering is that additional constraints are naturally integrated, such as

- *minimum size constraint*: every cluster must contain at least three cameras, for a robust dense stereo.
- *maximum size constraint*: every cluster must be smaller than a specified N_{size}. This constraints allows to run memory expensive reconstructions on machines with limited memory capabilities.
- *overlap constraint*: every cluster must define a number $N_{overlap}$ of overlapping cameras, which lie at the boundaries of other clusters.

Contrary to related work, our method is explicitly developed with an easy integration of selection in mind: paying attention to not remove overlapping cameras (Fig. 2), any image selection technique can be computed independently on every cluster without worrying about the creation of holes between clusters. See the paper for the details about dominant sets clustering and constraints satisfaction.

In our experimental evaluation we show both quantitative (measuring the reconstruction quality and the runtimes) and qualitative results (illustrating the compactness of the clusters and the overlaps configuration). Our method leads to significant speedup factors (up to 6) for the dense multi-view reconstruction with respect to considering all images at once, while maintaining high reconstruction quality. When compared to the state-of-the-art CMVS clustering method [3], we show our clusters to have a cleaner separation and better overlaps (Fig. 2).

In summary, our method provides an alternative formulation of view clustering, which directly incorporates overlapping cameras and speeds up the subsequent Multi-View Stereo methods by providing cleaner yet well-connected clusters for parallelization.

Acknowledgments. This work was supported by the European Research Council (ERC) under the project VarCity (#273940).

[1] S. Agarwala, Y. Furukawaa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building rome in a day. *Commmunications of ACM*, 54(10), 2011.

[2] J-M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *ECCV*, 2010.

[3] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010.

[4] A. Hornung, B. Zeng, and L. Kobbelt. Image selection for improved multi-view stereo. In *CVPR*, 2008.

[5] X. Li, C. Wu, C. Zach, S. Lazebnik, and J-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, 2008.

[6] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *PAMI*, 29(1):167–172, 2007.

[7] N. Snavely, S. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *CVPR*, 2008.

Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues

Isarun Chamveha[1]
isarunc@iis.u-tokyo.ac.jp
Yusuke Sugano[1]
sugano@iis.u-tokyo.ac.jp
Yoichi Sato[1]
ysato@iis.u-tokyo.ac.jp
Akihiro Sugimoto[2]
sugimoto@nii.ac.jp

[1] The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

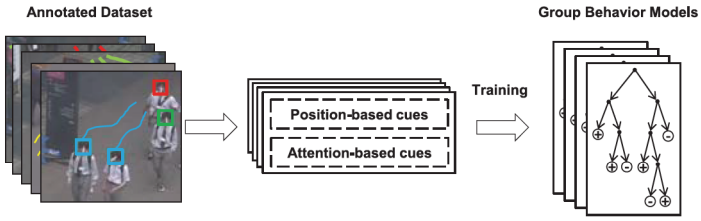


Figure 1: The framework of our proposed approach. Given annotated dataset of pedestrian states, set of measurements based on attention-based and position-based cues is created for each pair of pedestrians. Estimator is then applied to construct set of decision trees representing group behavior models for social group discovery task.

Table 1: Accuracy of estimates with of our dataset. Accuracy is measured as average accuracy between two classes to avoid bias problems in the test samples. **[6]** presented results using the approach of Yamaguchi *et al.* [6]. **Proposed** indicates results with our proposed approach. Note that the frame rate in the datasets is 30 fps.

Dataset	Approach	N_{past}					
		0	30	60	120	240	∞
UT Surveillance	[6]	74.9	75.8	76.5	75.8	76.0	76.4
	Proposed	**76.4**	**76.6**	**77.9**	**77.9**	**80.3**	**81.2**
Town Centre	[6]	68.2	67.5	69.8	69.3	68.5	68.5
	Proposed	**68.3**	**73.9**	**75.4**	**75.2**	**81.4**	**81.8**

Abstract

This paper presents an approach to discover social groups in surveillance videos by incorporating attention-based cues to model group behaviors of pedestrians in videos. Group behaviors are modeled as a set of decision trees with the decisions being basic measurements based on position-based and attention-based cues. Rather than enforcing explicit models, we apply tree-based learning algorithms to implicitly construct the decision tree models. The experimental results demonstrate that incorporating attention-based cues significantly increased the estimation accuracy compared to the conventional approaches that used position-based cues alone.

1 Introduction

Group discovery is an important problem in computer vision as group information can be used to aid pedestrian tracking in low frame-rate videos [5, 6] and to analyze human behavior [1]. Position-based cues, such as the relative position and velocity of pedestrians has been proved to be useful in solving the problem of social group discovery. However, attention-based cues, *i.e.*, how people pay attention to one another, have not yet been taken into account there. We aim to combine attention-based cues and position-based cues to discover social groups in this work. We propose a data-driven approach to discover social groups using statistics of both attention-based and position-based cues over pedestrian trajectories.

2 Proposed Framework

We define social group discovery as a pairwise problem to determine whether two people belong to the same group given information on their past states. The framework for our approach is outlined in Figure 1. We model the social group behaviors of pedestrians in a scene from two types of cues: *attention-based cues* and *position-based cues*. Given a set of pair-wise pedestrian states, several measurements of both cues are calculated at each time step, and are aggregated into histograms to evaluate the frequencies of each behavior over the entire trajectory. Using an annotated dataset of pedestrian states, a tree-based learning algorithm is then applied to construct the decision trees representing group behavior models for the social group discovery task.

Two types of attention-based cues are exploited in this work. The first cue is the *gaze exchange* between pedestrians. In order to perform group events, *e.g.*, conversation events, pedestrians in the same group often exchanged their gazes and fixed their attentions at one another. The second cue is the *mutual attention* of pedestrians in the same group. This is based on the observation that pedestrians often pay attention to the same object of interest. Measurements of position-based cues are derived from trajectories of two pedestrians based on the approach by Yamaguchi *et al.* [6].

To train decision trees, we calculate a set of measurements for every pair of pedestrians in the training set with overlapping existent. These measurements are aggregated into histograms. The decision trees could be implicitly constructed by using tree-based learning algorithms with the combination of histogram bins as decisions. A random trees classifier [3] was used in our approach to construct the trees. The histograms for each measurement were concatenated to represent the feature vector for that sample, and these vectors were used to train the random trees.

3 Experimental Results

We conducted experiments using two sequences: the UT-Surveillance sequence used by [4] and the Town Centre sequence used by [2] with varying number of available information of past frames N_{past}. We compared our method with the one by Yamaguchi *et al.* [6], who proposed to solve the same problem in the similar setting to our approach The results are listed in Table 1. It can be seen that our approach improves the accuracy of social group discovery tasks in every case in both datasets.

[1] L. Bazzani, M. Cristani, G. Paggetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: A social signaling perspective. In *Video Analytics for Business Intelligence*, pages 271–305. 2012.

[2] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *ICCV*, 2011.

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

[4] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto. Appearance-based head pose estimation with scene-specific adaptation. In *ICCV Workshops*, 2011.

[5] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2010.

[6] K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg. Who are you with and where are you going? In *CVPR*, 2011.

RLBP: Robust Local Binary Pattern

Jie Chen
Vili Kellokumpu
Guoying Zhao
Matti Pietikäinen

Center for Machine Vision Research
University of Oulu
Finland

Local binary pattern (LBP) is perhaps the best performing dense descriptor and it has been widely used in various applications, such as texture classification, human detection and face recognition [5]. However, one issue of LBP is that it is not so robust to the noise present in images when the gray-level changes resulting from the noise are not monotonic, even if the changes are not significant. To this end, we propose a new descriptor based on LBP, i.e., robust local binary pattern (RLBP). The idea is to locate the possible bit in LBP pattern changed by the noise and then revise the changed bit of the LBP pattern. The idea is very simple, but it works very well.

An example of how to compute RLBP from LBP is shown in Fig. 1. The neighbors of the pixel x_c would give the LBP pattern string (11010011). However, the pixel value of x_2=124 here is of high probability of being noisy since it results in a (101) substring. If we change the corresponding bit of x_2 in LBP string from 0 to 1, the new LBP string of this pixel (11110011) would denote a local corner, which is a more meaningful pattern for the texture representation.

Let us consider the LBP in the case that P=8 and R=1, i.e., LBP(8,1). In general, any neighboring three-bit substring in an eight-bit LBP pattern string is one in the set Y={y_1=(000), y_2= (001), y_3= (010), y_4= (011), y_5= (100), y_6= (101), y_7= (110), y_8= (111)}. Following the idea of the example in Fig. 1, we assume that the cases of y_3 and y_6 are noisy and change them to a new sub-string: y'_3= (000), and y'_6= (111). Our motivation for RLBP is to develop a descriptor to be robust to point noise (e.g., Gaussian). According to the statistical results [3], the neighboring pixels in an image should be smooth (e.g., LBP=00000000) or disturbed by edges (e.g., LBP=00001111). If one pixel is a noise, the bit in LBP pattern might be changed, e.g, LBP=00001101, where the seventh bit is different from its two neighboring bits, i.e., we have a substring y_6= (101). However, for other patterns, e.g., y_2= (001), the middle bit takes the same bit value as one of its neighboring bits at least, which might describe an edge (e.g., LBP=00001111). In addition, we only mapped those which were non-uniform patterns. In [3], all the non-uniform patterns are mapped to the same bin when building the histogram. By our method, we improve the discriminability of LBP.

Experimental results on the Brodatz [1] and UIUC [2] texture databases show that RLBP impressively outperforms the other widely used descriptors (e.g., SIFT, Gabor, MR8 and LBP) and other variants of LBP (e.g., completed LBP), especially when we add noise in the images (see Fig. 2). In addition, experimental results on human face recognition also show a promising performance comparable to the best known results on the Face Recognition Grand Challenge (FRGC) face dataset [4]. One example face from FRGC data set is shown in Fig.3.

Specifically, the experimental results over Brodatz texture dataset and the noisy Brodatz textures show perfect results. For example, RLBP achieves the accuracy of 98.9% for the dataset without noise and 98.6% with high levels of noise, while LBP gets the accuracy of 91.4% for the dataset without noise and 23.1% with high levels of noise (see Fig. 4). Experimental results on the UIUC texture database also show that RLBP impressively outperforms LBP. In addition, experimental results on FRGC face recognition also show a promising performance comparable to the best known results and reasonably good results for the dataset with added white Gaussian noise. For example, when log(1/*SNR*) =0.04, *RLBP* achieves the accuracy of 70% compared to 63% of Gabor, 48% of CLBP and 44% of LBP.

[1] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*, Dover Publications, New York, 1966.
[2] S. Lazebnik, C. Schmid, and J. Ponce. A Sparse Texture Representation Using Local Affine Regions. *IEEE TPAMI,* 2005.
[3] T. Ojala, M. Pietikäinen, and T. Mäenpää, Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns, IEEE *TPAMI*, 2002.
[4] P. J. Phillips, P. J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, Overview of the Face Recognition Grand Challenge, *CVPR*, 2005.
[5] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. *Computer Vision Using Local Binary Patterns*. Springer, 2011.

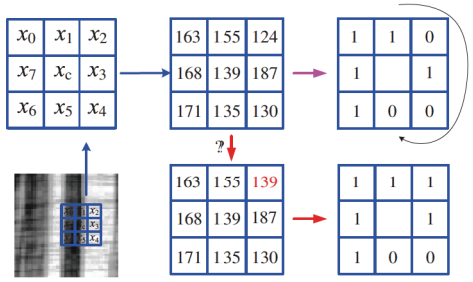


Fig. 1. Robust local binary pattern

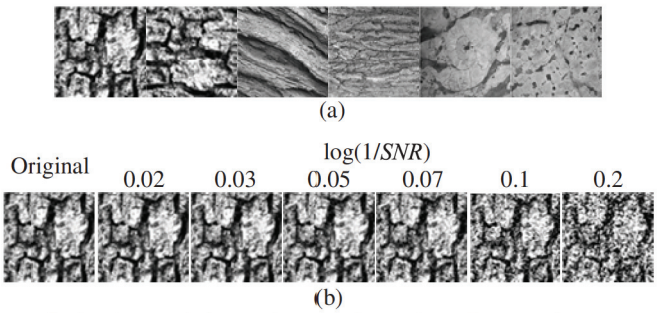


Fig. 2. (a) Example images from Brodatz and UIUC texture datasets. (b) An example from Brodatz by adding white Gaussian noise.

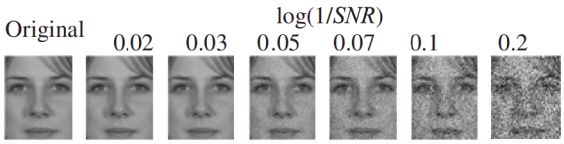


Fig. 3. An example from FRGC data set, with added white Gaussian noise.

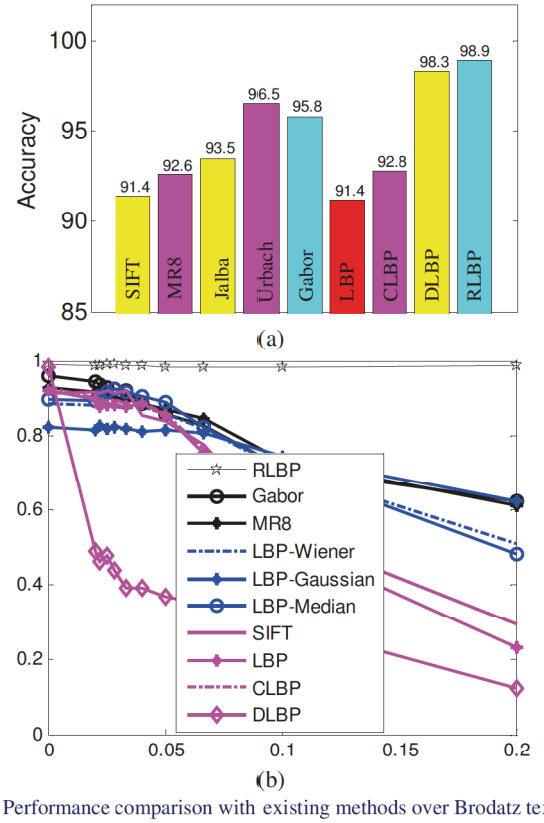


Fig. 4. Performance comparison with existing methods over Brodatz textures, (a) the original textures, (b) white Gaussian noise.

3D Deformable Shape Reconstruction with Diffusion Maps

Lili Tao
lltao@uclan.ac.uk
Bogdan J. Matuszewski
bmatuszewski1@uclan.ac.uk

Applied Digital Signal and Image Processing Research Centre
University of Central Lancashire, UK

Motivation This paper presents a method for recovering deformable shape and motion from uncalibrated 2D video sequence in the presence of missing data. Considering that the data dimensionality may not represent the true complexity of the problem, we suggest that the shapes can be well-modelled in a low dimensional manifold. However, building a dense representation of the manifold requires a large amount of training data which is not feasible in many real applications [3]. The main contribution of this paper is to propose a novel approach for estimating accurate 3D reconstructions using manifold learning technique, namely Diffusion maps, from a relatively small number of training samples. The problem is addressed by grouping shapes into evolving clusters, with the shapes in each cluster represented in the linear subspace, estimated based on the observations and the prior learned manifold.

Assuming that a set of image feature points have been tracked in the 2D image sequence viewed by an orthographic camera, the problem consists of shapes $\mathcal{S} = \{\mathbf{S}_1, \ldots \mathbf{S}_f\}$ and camera rotation $\mathcal{R} = \{\mathbf{R}_1, \ldots, \mathbf{R}_f\}$, recovery from 2D observations $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_f\}$. According to the shape basis assumption, shape $\mathbf{S}_t$ can be represented as a linear combination of n unknown but fixed basis shapes $\mathbf{B}_l$, $\mathbf{S}_t = \sum_{l=1}^{n} \theta_{tl}\mathbf{B}_l$, while the shape coefficients θ_{tl} are adjustable over time. Therefore the measurement can be arranged as:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_f \end{bmatrix} = \begin{bmatrix} \theta_{11}\mathbf{P}\cdot\mathbf{R}_1 & \cdots & \theta_{1n}\mathbf{P}\cdot\mathbf{R}_1 \\ \vdots & \ddots & \vdots \\ \theta_{f1}\mathbf{P}\cdot\mathbf{R}_F & \cdots & \theta_{fn}\mathbf{P}\cdot\mathbf{R}_F \end{bmatrix} \begin{bmatrix} -\mathbf{B}_1- \\ \vdots \\ -\mathbf{B}_n- \end{bmatrix} = \mathbf{MB} \quad (1)$$

$\mathbf{P}$ represents a known orthographic camera projection matrix.

The initial shapes and motion are calculated based on linear approach.

Shape embedding Having a shape $\mathbf{S}_t$ not present in the training set $\mathbf{X}$, an embedding $\mathbf{S}_t \mapsto (\hat{\Psi}_1(\mathbf{S}_t), \cdots, \hat{\Psi}_K(\mathbf{S}_t))$ of this new shape is calculated from the Nyström extension [1]:

$$\hat{\Psi}_k(\mathbf{S}_t) = \sum_{\mathrm{X}_j \in \mathbf{X}} p(\mathbf{S}_t, \mathrm{X}_j)\varphi_k(\mathrm{X}_j) \quad (2)$$

where $p(\mathbf{S}_t, \mathrm{X}_j)$ is from diffusion operator [2].

Shape update Once the initial shapes have been embedded into a lower dimensional space, finding their inverse mapping (the pre-image problem) can help to update shapes. Suppose we have an embedded point $\mathbf{b}_t \in \mathbb{R}^n$, a Delaunay triangulation can be computed in n dimensional reduced space, enabling selection of $n+1$ nearest neighbours $\mathbf{x}_{tl}$ of $\mathbf{b}_t$. Each point $\mathbf{b}_t$ can be represented as $\mathbf{b}_t = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{x}_{tl}$, where the coefficients $\theta = \{\theta_{t1}, \ldots, \theta_{t(n+1)}\}$ are the barycentric coordinates of $\mathbf{b}_t$ and the inverse mapping can be formulated as,

$$\hat{\Psi}^{-1}(\mathbf{b}_t) = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{X}_{tl} \quad with \quad \sum_{l=1}^{n+1} \theta_t = 1, 0 \le \theta_t \le 1 \quad (3)$$

where training sample $\mathbf{X}_{tl}$ is the pre-image of $\mathbf{x}_{tl}$.

Shape clustering We stipulate that the points in the reduced space belong to the same Delaunay simplex (i.e. cluster), can be modelled by the same linear subspace embedded in $\mathbb{R}^N$, and therefore all corresponding reconstructed shapes (represented by that cluster) can be approximated by a linear combination of the same set of unknown but fixed basis shapes. Thus all the shapes in the cluster i can be represented as $\mathbf{S}_t = \sum_{l=1}^{n+1} \theta_{tl}\mathbf{B}_l^i, \forall t \in \mathcal{T}_i$, where a set of basis shapes $\mathcal{B}^i = \{\mathbf{B}_1^i \ldots \mathbf{B}_{n+1}^i\}$ is spanning the tangent linear subspace representing all the shapes from the cluster i.

The reconstructed shapes are often different from the training samples, therefore cannot be perfectly mapped into the manifold $\mathcal{M}$. As the result we relax the constraint for the basis shapes, only "encouraging" them to be close to the basis shapes spanning the tangent subspace, instead of being exactly the same. The additional constraint applied to the i^{th} set of basis shapes is,

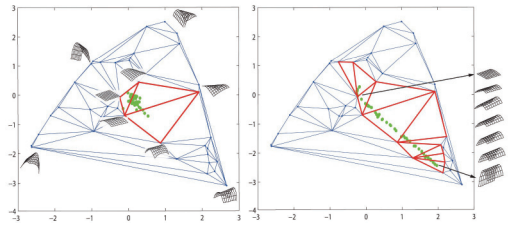


Figure 1: Delaunay triangulations (blue line) in the reduced space; Embedded initial shapes (left) and reconstructed shapes(right) (green dots) and the actual used triangles (red line)

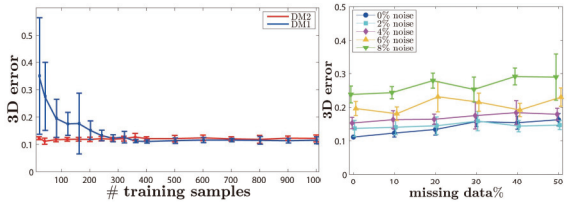


Figure 2: Results for the *cardboard* data. Left: 3D error as function of the number of training samples. Right: Varying levels of missing data and 5 levels of noise.

$$\varepsilon_{bs}^i = \sum_{l=1}^{n+1} \left\| \mathbf{B}_l^i - \mathbf{X}_l^i \right\|^2, \mathbf{X}_l^i \in \mathcal{X} \quad (4)$$

Fig. 1 illustrates an example of how the initial shapes are redistributed in the reduced space after algorithm has converged.

Non-linear refinement The parameters $\theta_{tl}, \mathbf{B}_l^i$ and $\mathbf{R}_t$ are optimised simultaneously by minimising the 2D re-projection error with additional constraints on basis shapes and rotation matrices. The cost function can be written as,

$$E(\mathbf{R}_t, \mathbf{B}_l^i, \theta_{tl}) = \sum_{t \in \mathcal{T}_i} \left\| \mathbf{Y}_t - \mathbf{P}\cdot\mathbf{R}_t \sum_{l=1}^{n+1} \theta_{tl}\mathbf{B}_l^i \right\|^2 + \lambda_B \varepsilon_{bs}^i + \lambda_R \sum_{t \in \mathcal{T}_i} \varepsilon_{rot} \quad (5)$$

where $\varepsilon_{rot} = \left\| \mathbf{R}_t\mathbf{R}_t^T - \mathbb{I} \right\|$ enforces orthonomality of all $\mathbf{R}_t$. λ_B and λ_R are regularisation constants. A non-linear optimisation based on bundle adjustment using Levenberg-Marquardt algorithm was applied to minimize this cost function.

Evaluation Fig.2(Left) shows the effect of the number of training shapes on the reconstruction accuracy for [3] (DM1) and the proposed one (DM2). The average reconstruction errors with the standard deviation calculated over 10 trials, each using different data subset for training. In real cases, missing data and measurement noise are distorting the observations in the same time. Fig.2(Right) shows the relevant results.

Real data result shows in Fig.3.

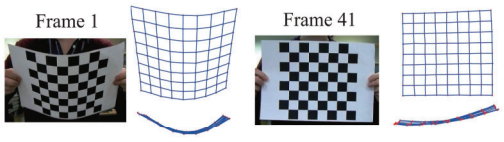


Figure 3: Selected 2D frames from paper bending video sequence. Front and top views of the corresponding 3D reconstructed results.

[1] P. Arias, G. Randall, and G. Sapiro. Connecting the out-of-sample and pre-image problems in kernel methods. In *CVPR*, 2007.

[2] R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[3] L. Tao and B.J. Matuszewski. Non-rigid structure from motion with diffusion maps prior. In *CVPR*, pages 1530–1537, 2013.

Towards a minimal solution for the relative pose between axial cameras

Francisco Vasconcelos
fpv@isr.uc.pt
João P. Barreto
jpbar@deec.uc.pt

Institute for Systems and Robotics
University of Coimbra
Coimbra, Portugal

An axial camera is a particular case of a non-central camera where every back-projection ray intersects a line in 3D (the axis). The axial camera can be used to model vision systems and imaging situations of practical interest. Examples include any catadioptric system that combines a revolution mirror with a central camera for which the viewpoint is aligned with the mirror axis (e.g. a pinhole looking at a spherical mirror) [8]; the situation of a perspective camera looking through multiple flat refractive mediums [1]; or a multi-camera rig composed by two or more pinhole cameras with collinear optical centers [3].

This paper addresses the problem of estimating the translation $\mathbf{t}$ and the rotation R between two axial cameras using point correspondences. Pless showed that this problem can be linearly solved from a minimum of 17 point correspondences using a DLT like approach [4]. Later in [3, 7] it was observed that for the case of axial cameras this linear estimation could be accomplished from 16 point correspondences.

The relative pose problem has 6 unknowns meaning that in theory 6 point correspondences provide enough information for determining the relative rotation and translation of the axial camera. Stewenius et al. proposed in [6] a minimal solution for the relative pose between generalized cameras. However, their algorithm is complex, provides a large number of possible solutions (up to 64), and, as reported in [3], it degenerates for most axial camera configurations. This article does not provide a minimal solution for the relative pose between axial cameras, but shows how the motion can be computed using as few as 10 point correspondences. Our 10-point method is an advance with respect to the previous 16-point [3, 7].

Given that all back-projection rays of an axial camera intersect its axis, they belong to a linear line congruent of dimension 4 [5]. This means that all rays can be represented by 5 dimensional coordinate vectors λ_i that are a linear combination of 5 base lines aligned with the axes $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ in Fig. 1(a).

Given a set of intersecting ray correspondences (λ_i, λ_i'), we can establish linear relations with the form

$$\lambda_i^{\mathsf{T}} \Phi \lambda_i' = 0 \tag{1}$$

with Φ being a 5×5 matrix that encodes the 4 essential matrices displayed in Fig. 1(b)



$$\mathsf{E}_1 = \Phi^{\{1:3,1:3\}} = [\mathbf{t}]_\times \mathsf{R} \tag{2}$$

$$\mathsf{E}_2 = \Phi^{\{1:3,3:5\}} = [\mathsf{R}\mathbf{v} + \mathbf{t}]_\times \mathsf{R}\mathsf{W} \tag{3}$$

$$\mathsf{E}_3 = \Phi^{\{3:5,1:3\}} = [\mathsf{W}^{\mathsf{T}}(\mathbf{t} - \mathbf{v})]_\times \mathsf{W}^{\mathsf{T}}\mathsf{R} \tag{4}$$

$$\mathsf{E}_4 = \Phi^{\{3:5,3:5\}} = [\mathsf{W}^{\mathsf{T}}(\mathsf{R}\mathbf{v} + \mathbf{t} - \mathbf{v})]_\times \mathsf{W}^{\mathsf{T}}\mathsf{R}\mathsf{W} \tag{5}$$

The matrix Φ has 17 free parameters, and therefore can be linearly estimated from 16 correspondences.

Additionally, the following family of matrices

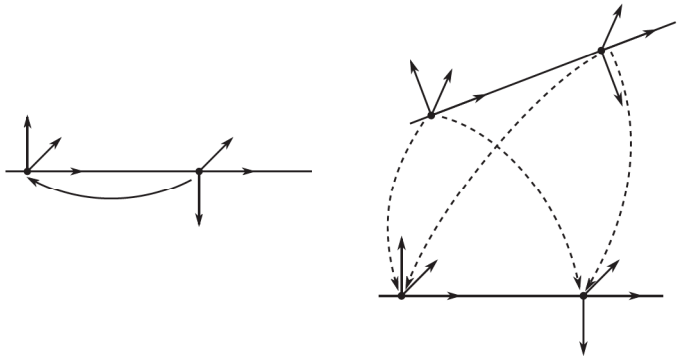
$$\mathsf{E}_i = \alpha\mathsf{E}_1 + \beta\mathsf{E}_2\mathsf{W}^{\mathsf{T}} + \gamma\mathsf{W}\mathsf{E}_3 + \delta\mathsf{W}\mathsf{E}_4\mathsf{W}^{\mathsf{T}}, \quad \forall \alpha,\beta,\gamma,\delta \in \Re \tag{6}$$

must have the properties of an essential matrix, and therefore verify the following nonlinear constraints

$$2\mathsf{E}_i\mathsf{E}_i^{\mathsf{T}}\mathsf{E}_i - tr(\mathsf{E}_i\mathsf{E}_i^{\mathsf{T}})\mathsf{E}_i = 0 \tag{7}$$

$$\det \mathsf{E}_i = 0 \tag{8}$$

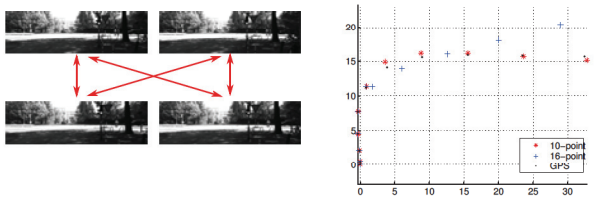
This makes us able to solve the problem using just 10 correspondences (λ_i, λ_i') by first generating a 7 dimensional linear subspace for Φ and then solving a system of cubic equations in 6 variables, with the action matrix technique [2].



(a) Representation of reference frames along the camera axis $\mathbf{B}$

(b) The four 3×3 essential matrices encoded by Φ

Figure 1: A new parameterization for axial cameras



(a) Input data

(b) Trajectory

Figure 2: Performance comparison between 10-point algorithm and 16-point algorithm with real data.

The algorithm is validated and compared against the 16-point algorithm [3] for estimating the relative pose between stereo camera pairs, using both synthetic and real input data (Fig. 2), and showing that our algorithm has a superior performance.

Our long-term goal, however, is to reach a 6-point minimal algorithm, which will require a more in-depth study of the non-linear relations between the essential matrices described in this paper.

[1] A. Agrawal, S. Ramalingam, Y. Taguchi, and V. Chari. A theory of multi-layer flat refractive geometry. In *CVPR*, 2012.

[2] Martin Byröd, Klas Josephson, and Kalle Åström. Fast and stable polynomial equation solving and its application to computer vision. *Int. J. Comput. Vision*, 84(3):237–256, September 2009. ISSN 0920-5691.

[3] Jae-Hak Kim, Hongdong Li, and Richard Hartley. Motion estimation for nonoverlapping multicamera rigs: Linear algebraic and L_∞ geometric solutions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(6): 1044–1059, June 2010.

[4] R. Pless. Using many cameras as one. In *CVPR*, 2003.

[5] H Pottmann and J Wallner. *Computational line geometry*. Springer Verlag, Berlin, 1 edition, 2001.

[6] H. Stewénius, D. Nistér, M. Oskarsson, and K. Åström. Solutions to minimal generalized relative pose problems. In *OMNIVIS*, 2005.

[7] P. Sturm. Multi-view geometry for general camera models. In *CVPR*, 2005.

[8] Peter Sturm, Srikumar Ramalingam, Jean-Philippe Tardif, Simone Gasparini, and João Barreto. Camera models and fundamental concepts used in geometric computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(1-2):1–183, 2011.

Combining Features and Intensity for Wide-Baseline Non-Rigid Surface Registration

Jim Braux-Zin[1,2]
jim.braux-zin@cea.fr
Romain Dupont[1]
romain.dupont@cea.fr
Adrien Bartoli[2]
adrien.bartoli@gmail.com

[1] CEA, LIST
91 191 Gif-sur-Yvette
France
[2] ISIT, Université d'Auvergne
63 000 Clermont-Ferrand
France

Non rigid surface registration consists in estimating the deformation of a known surface between two images, usually by fitting a warp such as a Thin-Plate Spline or a Free-Form Deformation. Common techniques are split in two categories: *feature-based surface detection* i.e. estimation of a potentially important deformation from an image and a flat source template, and *pixel-based surface tracking* where important deformations can be estimated over a video sequence as long as the frame to frame steps are small. Our contribution consists in bridging the two worlds by introducing a new data term robustly merging feature and pixel-based costs in a pyramidal variational approach. By using a robust estimator we achieve an implicit optimal filtering of features and automatic balancing between the two terms.

Our goal is to directly estimate a deformation between an image $\mathcal{I}$ and a given flat template $\mathcal{I}_0$. This deformation is parametrized by the displacements $\mathbf{u}$ of the control points of a Free-Form Deformation warp [4]. The image in $\mathcal{I}$ of the point $\mathbf{q}$ from $\mathcal{I}_0$ is $\mathcal{W}(\mathbf{q},\mathbf{u})$. For the sake of clarity and to allow easier comparison with feature-based methods, we adopt a feature-filtering model for our cost function:

$$\varepsilon(\mathbf{u},\mathcal{F},\mathcal{I},\mathcal{I}_0) = \underbrace{\varepsilon_f(\mathbf{u},\mathcal{F})}_{\text{features cost}} + \underbrace{\varepsilon_b(\mathbf{u})+\varepsilon_s(\mathbf{u})+\varepsilon_d(\mathbf{u},\mathcal{I},\mathcal{I}_0)}_{\text{priors}} \quad (1)$$

where $\mathcal{F}$ is the set of feature matches provided as input. This function is optimized over all control points using a Gauss-Newton approximation embedded in a coarse-to-fine framework.

The feature-based cost function is similar to the ones used in [5] and [1]:

$$\varepsilon_f(\mathbf{u},\mathcal{F}) = \lambda_f \sum_{(\mathbf{f}_0,\mathbf{f})\in\mathcal{F}} \iint_\Omega \omega(\mathbf{q},\mathbf{f}_0)\Psi_\sigma\left(\|\mathbf{u}-(\mathbf{f}-\mathbf{f}_0)\|_2\right)\mathrm{d}\mathbf{q} \quad (2)$$

where ω is a weight function based on the distance to feature only, and $\Psi_\sigma(x) = \frac{x^2}{\sigma+x^2}$ is the Geman McClure estimator, providing an implicit filtering of outliers. By avoiding any assumption on the feature used (such as the availability of a confidence measure), we leave open the possibility of using any current and future feature matchers.

We use three priors to constrain and regularize the warp produced by the feature matches. First the bending energy regularizer

$$\varepsilon_b(\mathbf{u}) = \lambda_b \iint_\Omega \left(\frac{\partial\mathcal{W}(\mathbf{q};\mathbf{u})^2}{\partial^2 q_x}\right) + \left(\frac{\partial\mathcal{W}(\mathbf{q};\mathbf{u})^2}{\partial q_x \partial q_y}\right) + \left(\frac{\partial\mathcal{W}(\mathbf{q};\mathbf{u})^2}{\partial^2 q_y}\right)\mathrm{d}\mathbf{q} \quad (3)$$

that enforces a smooth displacement field. Then a *shrinker* term [3] preventing the warp from folding in self-occluded areas:

$$\varepsilon_s(\mathbf{u}) = \lambda_s \sum_{\mathbf{q}\in\Omega}\sum_{\mathbf{d}\in\mathcal{D}}\sum_{c\in\{x,y\}} \gamma\left\{\left(D^{(l)}_{\mathbf{d}}(\mathbf{q};\mathbf{u})\right)_c \left(D^{(r)}_{\mathbf{d}}(\mathbf{q};\mathbf{u})\right)_c\right\} \quad (4)$$
$$\gamma(x) = 0 \text{ if } x \geq 0 \text{ and } x^2 \text{ otherwise}$$

where $\mathcal{D}$ is a discretized set of directions and the function γ penalizes points whose right and left derivatives have opposite signs.

The flexible robust-estimator-based framework allows us to use the brightness constancy assumption (common in dense optical flow estimation) as an additional prior to fully take advantage of all the available data:

$$\varepsilon_d(\mathbf{u},\mathcal{I},\mathcal{I}_0) = \lambda_d \sum_{\mathbf{q}\in\Omega} (1-\mathcal{P}_{\mathrm{SO}})(\mathcal{I}_0(\mathbf{q}) - \mathcal{I}(\mathcal{W}(\mathbf{q};\mathbf{u})))^2 \quad (5)$$

where $\mathcal{P}_{\mathrm{SO}}$ is the self-occlusion probability. It is computed by noticing that on self-occluded areas, the derivative of the warp vanishes in one direction.

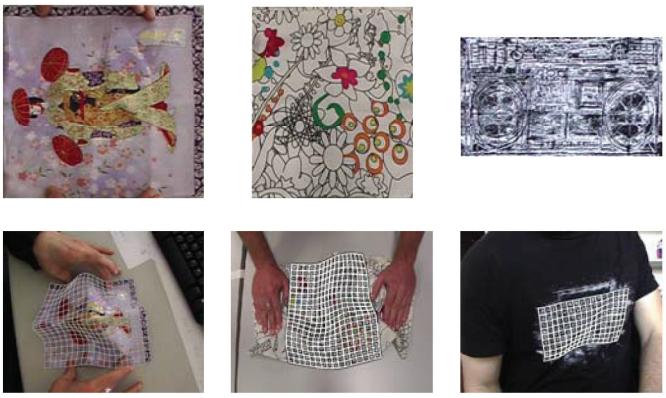


Figure 1: Qualitative results on [7, 8]. First row: templates, second row: deformed surface with estimated deformation.

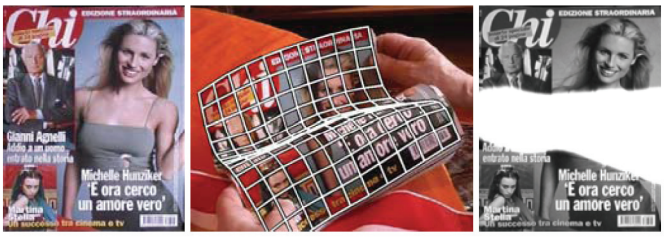


Figure 2: Challenging case from [2]. From left to right: template, estimated deformation and estimated self-occlusions (in white).

Most state of the art methods [6, 9] for wide-baseline surface registration without tracking operate in separate steps: feature filtering, warp fitting using features and optional refining using a pixel-wise data term. We demonstrate that our joint warp estimation and implicit feature filtering using the brightness constancy prior outperforms methods based on separate steps. Qualitative and quantitative evaluations using different priors on synthetic data and public datasets validate this claim and are available in the full paper.

[1] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *PAMI*, 2011.

[2] V. Ferrari, T. Tuytelaars, and L. Gool. Simultaneous object recognition and segmentation by image exploration. In *ECCV*. 2004.

[3] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of nonrigid registrations with image-based self-occlusion reasoning. *PAMI*, 2010.

[4] S. Lee, G. Wolberg, K. Chwa, and S. Shin. Image metamorphosis with scattered feature constraints. *Visualization and Computer Graphics*, 1996.

[5] J. Pilet, V. Lepetit, and P. Fua. Fast Non-Rigid Surface Detection, Registration and Realistic Augmentation. *IJCV*, 2008.

[6] D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. *IJCV*, 2012.

[7] M. Salzmann, R. Hartley, and P. Fua. Convex optimization for deformable surface 3-d tracking. In *ICCV*, 2007.

[8] Mathieu Salzmann and Pascal Fua. Reconstructing sharply folding surfaces: A convex formulation. In *CVPR*, 2009.

[9] Q. Tran, T. Chin, G. Carneiro, M. Brown, and D. Suter. In defence of RANSAC for outlier rejection in deformable registration. In *ECCV*, 2012.

Single Image Shadow Detection using Multiple Cues in a Supermodular MRF

Tomás F. Yago Vicente
tyagovicente@cs.stonybrook.edu
Chen-Ping Yu
cheyu@cs.stonybrook.edu
Dimitris Samaras
samaras@cs.stonybrook.edu

Image Analysis Lab
Computer Science Dept.
Stony Brook University
Stony Brook, NY, USA

In this work we present a complete methodology for single image shadow detection based on the learned appearance of shadows. Our main contributions are:

- A novel single region classifier with a multikernel model specifically tailored for shadow region classification with SVM that outperforms the state of the art.
- A novel boundary classifier for shadow boundaries cast over surfaces with the same material, and two improved paired regions classifiers.
- We pose shadow detection as an MRF binary labelling problem that combines our accurate single region classifier with both boundary cues and illumination relationships between pairs of regions of the same material obtained from precise classifiers.

We propose a single region shadow classifier based on a multikernel SVM. Our multikernel model is a linear combination of χ^2 and Earth Mover's Distance(EMD)[5] kernels that operate on texture and color histograms disjointly. This single region classifier already outperforms the more complex state of art methods, without performing MRF/CRF optimization.

The local appearance of a single region is often ambiguous. Even for a human observer it can be hard to discern if a region is in shadow or not, without considering its context. Hence, it is sensible to look beyond the boundaries of a single region to decide its shadow label [1] [6]. In contrast to previous work we strive to use such contextual information sparingly. For MRF optimization reasons we prefer that most of the work is handled by the single region classifier (unary MRF potentials), with sparse pairwise connections that smooth the label changes across regions. We build on the work of [1] to propose our own improved pairwise classifiers but constrained to adjacent regions: for pairs of regions sharing the same material and same illumination condition, and for same material pairs viewed under different illumination (first lit, second in shadow). We also propose a shadow boundary classifier. Since shadow boundaries often overlap with reflectance changes confounding the effects of the illumination change, our classifier focuses on boundaries of shadows cast over surfaces with the same underlying material.

We integrate our single region classifier, our pairwise classifiers, and our boundary classifier using an MRF. Confident positive predictions of the pairwise and boundary classifiers are used to define the pairwise potentials and the graph topology of the MRF. The unary potentials are defined based on the single region classifier. We want to minimize the following functional:

$$E(\mathbf{x}) = \sum_{i \in \mathbf{R}} \phi(x_i) + \sum_{i,j \in \Omega^s} \psi_s(x_i,x_j) + \sum_{i,j \in \Omega^d} \psi_d(x_i,x_j) + \sum_{i,j \in \beta} \psi_b(x_i,x_j)$$

where Ω^s, Ω^d denote pairs of regions of the same material confidently predicted to have same and different illumination respectively, and β is the set of confidently predicted shadow boundaries. The key idea is to combine a very strong region classifier with a set of precise boundary and paired region classifiers that define the smoothing in the MRF. In order to achieve sparsity and increase reliability we favor precision over recall in the outcome of the context based classifiers. The same material/same illumination relationships induce submodular potentials (ψ_s). Potentials due to the boundary classifier (ψ_b) are supermodular. The directionality of the same material different/illumination classifier introduces asymmetric potentials (ψ_d). We minimize the MRF energy using QPBO [3] [4], the asymmetric potentials led to a minor modification of the initial reparameterization stage of QPBO.

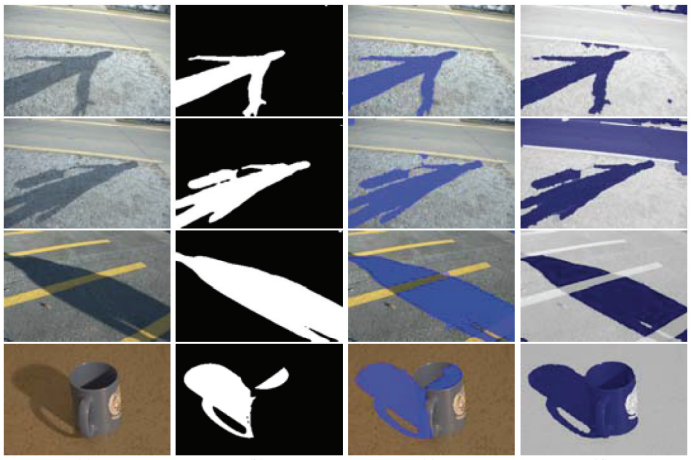


Figure 1: Comparison of shadow detection results. a) Input image. b) Ground truth shadow pixel mask. c) Our results overlaid in ligh blue d). Guo *et al.* [2] results in dark blue.

	Acc.	BER	Sha.	Non
Guo *et al.* [2]	0.891	0.166	0.716	0.952
single region	0.916	0.124	0.795	0.957
MRF ψ_b	0.926	0.101	0.840	0.958
MRF $\psi_s \psi_d$	0.937	0.087	0.862	**0.963**
MRF	**0.941**	**0.078**	**0.880**	0.962

Table 1: Shadow detection quantitative results.

Experimental results on the dataset collected by [1] show that our method clearly outperforms the state of the art. Our single region classifier reduces the balance error rate (BER) by a 25% with respect to [2], correctly detecting 11% more shadow pixels. Moreover, with our MRF optimization we further reduce the BER to a 7.8%. That is a 53% decrease compared to the state of the art. Our total pixel accuracy is 94.1% (88% on shadow pixels).

[1] Ruiqi Guo, Qieyun Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2033 –2040, june 2011. doi: 10.1109/CVPR.2011.5995725.

[2] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints):1, 2012. ISSN 0162-8828.

[3] Vladimir Kolmogorov and Carsten Rother. Minimizing non-submodular functions with graph cuts - a review. Technical report, TPAMI, 2007.

[4] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary mrfs via extended roof duality. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007.

[5] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. pages 59–66, 1998.

[6] Jiejie Zhu, K.G.G. Samuel, S.Z. Masood, and M.F. Tappen. Learning to recognize shadows in monochromatic natural images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 223 –230, june 2010.

Recognizing Humans in Motion: Trajectory-based Aerial Video Analysis

Yumi Iwashita[1]
yumi@ieee.org
M. S. Ryoo[1]
mryoo@jpl.nasa.gov
Thomas J. Fuchs
Thomas.Fuchs@jpl.nasa.gov
Curtis Padgett
curtis.w.padgett@jpl.nasa.gov

Jet Propulsion Laboratory,
California Institute of Technology
Pasadena, CA, USA

Aerial surveillance images cover a wide area at low resolution [1, 2]. In order to detect objects (e.g., pedestrians) from such videos, conventional methods either utilize appearance information from raw videos or extract blob information from background subtraction results. However, people seen in low resolution images have less appearance information, and hence are very difficulty to classify based on their appearance or blob size. In addition, due to heavy camera movements caused by aerial vehicle ego-motion and wind, the system is expected to generate many noisy false detections including parallax. Figure 1 shows an example aerial image and a magnified area of the image. As we are able to observe in Figure 1 (b), appearance of people (yellow dotted circles) and rocks (red cricles) show very similar visual characteristics.

This paper proposes a novel method to detect/recognize people by classifying object candidates in low resolution aerial images. The idea is to detect and classify objects from aerial videos based on their motion: we analyze a trajectory of each object candidate, deciding whether it is a person-of-interest or simple noise based on how it moved. After objects are tracked by a Kalman filter-based tracking, we represent their motion as multi-scale histograms of 'orientation changes', which efficiently captures movements displayed by objects. Finally, random forest classifiers are applied to our new representation to make the decision.

Our histogram is based on an idea that tracks originated by humans participating in activities will contain movements completely different from tracks generated by noisy false object detections. The histogram is calculated using relative orientation changes between every pair of vectors, which are defined using three consecutive observed points with a constant frame difference as shown in Figure 2. Figure 3 shows examples of calculated histograms for a person, a car, and a false detection caused stabilization errors. In addition, we propose a multi-magnitude version of our histogram of orientation changes, which is descriged to capture movement magnitudes as well as orientation changes of the objects.

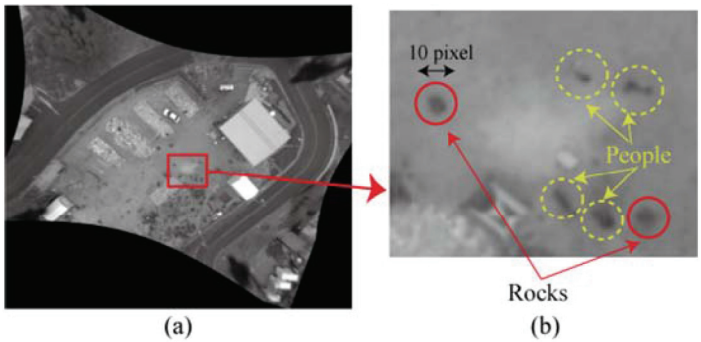


Figure 1: (a) An example of rectified and stabilized images, (b) an enlarged area including rocks and people whose appearances are similar.

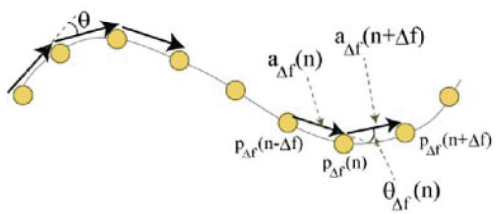


Figure 2: An example of orientation changes.

[1]These authors contributed equally to the paper.

The experimental results illustrate that our approach recognizes objects-of-interest (i.e., humans) even when there exist a large number of false detection/tracking. Figures 4 (a) ∼ (c) show example tracking results, classification results of our approach, and the ground truths. We are able to observe that our approach is able to successfully recognize humans from low resolution aerial videos with heavy ego-motion, by discarding many false detections while maintaining true detections based on our histogram representation.

The contributions of this paper are (i) the introduction of a new concept that objects can be better recognized using their motion information particularly in aerial video and (ii) our novel feature representation to capture object motion effectively and efficiently.

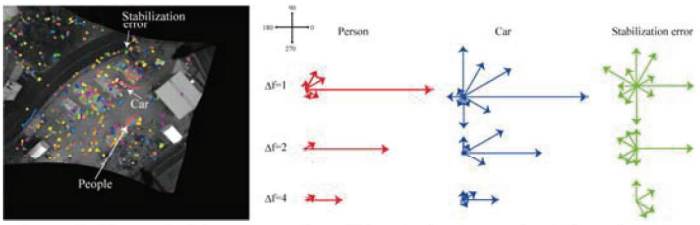


Figure 3: Example histograms describing trajectory orientation changes.

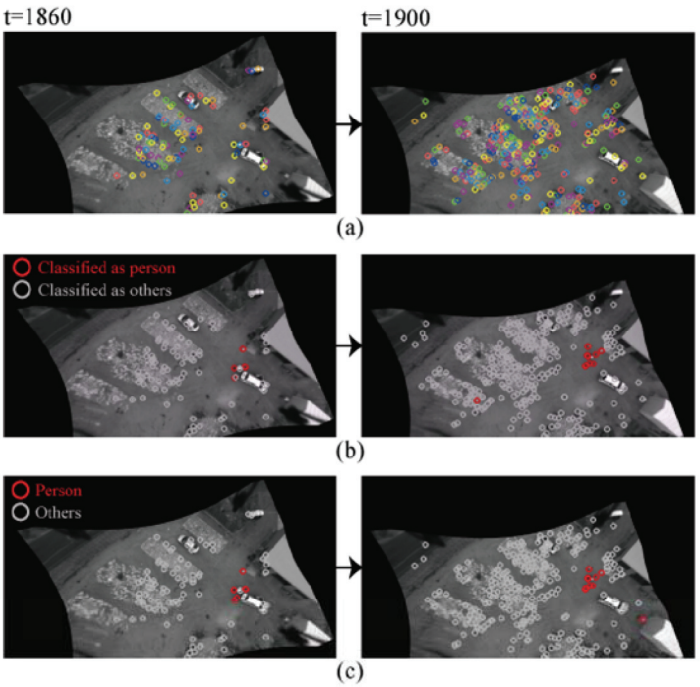


Figure 4: (a) Example results of object tracking, (b) classification results by the proposed method, and (c) ground truth. The video contains the activity of 'a group of person unloading an object from the vehicle', and we are able to observe that our approach correctly recognizes human tracks as opposed to other noise.

[1] V. Reilly, H. Idrees & M. Shah, *Detection and tracking of large number of targets in wide area surveillance*, ECCV 2010.

[2] T. Pollard & M. Antone, *Detecting and tracking all moving objects in wide-area aerial video*, CVPRW 2012.

Combining Local and Global Cues for Closed Contour Extraction

Vida Movahedi
http://www.cse.yorku.ca/~vida
James H. Elder
http://elderlab.yorku.ca/~elder/

Centre for Vision Research
York University
Toronto, Canada

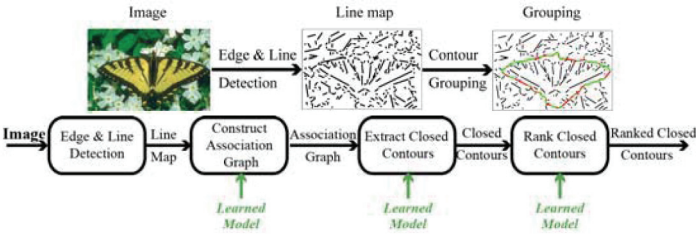


Figure 1: Overview of algorithm. The output consists of multiple non-intersecting closed polygons comprised of visible line segments (green) connected by linear interpolants (red).

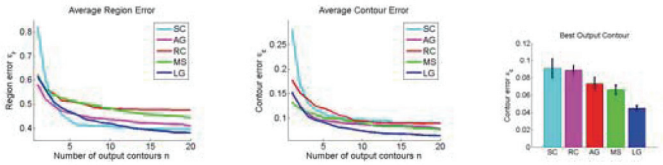


Figure 2: Quantitative evaluation on SOD test dataset. (a) Minimum region error and (b) Minimum contour error as a function of the number of output contours allowed. (c) Minimum contour error over all output contours.

We address the problem of extracting the simple closed contours that bound the salient objects in a natural image[4]. The method proposed in this paper falls in the category of *contour grouping methods*, which search for the optimal cycle of local oriented primitives (*e.g.* edgels, line segments) forming the boundary.

The first stage of processing in our method extracts from the image a sparse graph representation of the oriented structure in the image. Local oriented edges detected in the image are grouped locally into subpixel-localized line segments of variable length. As in prior work, each of the resulting segments forms a vertex in our association graph, and each edge in this graph represents a grouping hypothesis between specified endpoints of two segments.

Contour grouping algorithms are generally based upon classical Gestalt cues such as proximity, good continuation and similarityWhile prior methods [1, 2] have approximated these local cues as independent and modelled the marginals parametrically, we use EM to learn Gaussian mixture models for the 4-dimensional likelihoods, which allows us to capture small dependencies between the cues. Our sparse graph is then formed by representing only the k outgoing edges with highest likelihood for each vertex.

Algorithms for computing closed contours are generally based upon these local Gestalt cues relating pairs of oriented elements, and a Markov assumption to then group these elements into chains. Without additional global constraints, these algorithms generally do not perform well on general natural scenes. Such global cues could include symmetry, shape priors or global colour appearance. A key challenge is to combine these local and global cues in a statistically optimal way. A main contribution of our paper is a novel, effective method for combining local and global cues, both at the stage of forming new closed contour hypotheses, and at the stage of evaluating and ranking these hypotheses.

We develop and test this method using global colour contrast cues in both a greedy constructive search for closed contours, and in the final ranking of closed contours. We adopt a regression approach: learn a regressor that will predict the distance of a candidate path from ground truth based on both local and global cues. We take a nonparametric approach to the regression problem, binning the local cue $f_l(c_k)$ and global cue $f_g(c_k)$, and then estimating the mean $\hat{\varepsilon}_l$ and $\hat{\varepsilon}_g$ and variance σ_l^2 and σ_g^2 of predicted error for each. Assuming independence between the local and global cues, the least-squares error prediction ε_{ML} (ML for normal distributions) is then given by

$$\hat{\varepsilon}_{ML} = \frac{\hat{\varepsilon}_l/\sigma_l^2 + \hat{\varepsilon}_g/\sigma_g^2}{\sigma_l^{-2} + \sigma_g^{-2}} \quad (1)$$

A second challenge faced by contour grouping algorithms stems from the exponential nature of the search space, which tends to yield many very similar high probability hypotheses, while neglecting other more diverse hypotheses with slightly lower probability. This tendency lowers performance as partial hypotheses that look slightly less promising are weeded out too early. A second contribution of the present paper is a novel method for promoting diversity in the formation and ranking of contour hypotheses, leading to substantial improvements in performance.

We evaluate our results on a standard public dataset, the Salient Object Dataset (SOD) [6], which in turn is based upon the Berkeley Segmentation Dataset (BSD). We compared our local/global (LG) method with the Ratio Contour (RC) algorithm of Stahl & Wang [7], the Adaptive Grouping (AG) method of Estrada & Jepson [3], the Multiscale (MS) method of Estrada & Elder [2], and the Superpixel Closure (SC) method of Levinshtein & Dickinson [5], in each case using the parameters recommended by the authors.

We used two evaluation measures: i) A region error ε_r based on the popular intersection to union measure, and a contour error ε_c based on a normalized version of the contour mapping (CM) measure [6], defined as the average distance between corresponding pixels on contours A and B, normalized by the perimeter of the ground truth boundary B.

Figure 2 shows the quantitative results. In panels (a) and (b) we report the average region error ε_r and contour error ε_c, respectively, as a function of the number of contours n the algorithm is allowed to report. We see that our LG method outperforms other methods by the region measure when the number of output contours is 12 or greater, and by the contour measure when the number of output contours is 2 or greater.

Panel (c) shows how the best contours computed by each method compare, using the contour error measure ε_c. Our LG method performs better than other methods when the number of hypotheses is not limited. This suggests that further improvements to the LG algorithm can potentially be obtained through better methods for ranking the closed contours extracted by the algorithm.

[1] J. H. Elder and R. M. Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4): 324–353, 2002.

[2] F. J. Estrada and J. H. Elder. Multi-scale contour extraction based on natural image statistics. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06*, page 183, 2006.

[3] F. J. Estrada and A. D. Jepson. Robust boundary detection with adaptive grouping. In *Computer Vision and Pattern Recognition Workshop, CVPRW '06*, volume 2006, pages 184–184, 2006.

[4] F. Ge, S. Wang, and T. Liu. Image-segmentation evaluation from the perspective of salient object extraction. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:1146–1153, 2006.

[5] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Optimal contour closure by superpixel grouping. *Proceedings of European Conference on Computer Vision*, 2010.

[6] V. Movahedi and J.H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 49 –56, 2010.

[7] J. S. Stahl and S. Wang. Edge grouping combining boundary and region information. *IEEE Transactions on Image Processing*, 16(10): 2590–2606, 2007.

Lagrangian Strain Tensor Computation with Higher Order Variational Models

Alexander Hewer[1]
hewer@mia.uni-saarland.de
Joachim Weickert[1]
weickert@mia.uni-saarland.de
Tobias Scheffer[2]
t.scheffer@mx.uni-saarland.de
Henning Seibert[2]
h.seibert@mx.uni-saarland.de
Stefan Diebels[2]
s.diebels@mx.uni-saarland.de

[1] Mathematical Image Analysis Group
Saarland University, Campus E1.7
66041 Saarbrücken, Germany

[2] Chair of Applied Mechanics
Saarland University, Campus A4.2
66041 Saarbrücken, Germany

1 Motivation

In specific application areas, obtaining higher order motion information is of great interest. An example of such information is the Lagrangian strain tensor [3] that plays a vital role in mechanical engineering. Since this tensor is computed by means of first-order motion derivatives, it is tempting to estimate the optical flow field with a highly accurate variational model and compute its derivatives afterwards. However, this procedure yields in general not very good results because of the following reasons: First, these derivatives have to be obtained via numerical differentiation. Differentiation is a classical ill-posed problem where small fluctuations in the data can cause large deviations in the result [1]. Secondly, most high-accurate models make use of a first-order smoothness assumption that causes the results to be biased towards translational motion; see e.g. [5].

2 Contribution

In this paper, we propose a novel approach for deriving higher order variational models that offer the following advantages: They directly estimate regularised versions of the motion derivatives, which circumvents the problems of numerical differentiation. Moreover, they deploy higher order smoothness assumptions that do not suffer from translational bias.

3 Our Recursive Approach

First we consider a *baseline model* that we want to extend. For instance, a variational approach for optical flow computation can be based on a Horn and Schunck model [2] with a data term that can handle large displacements. It computes the motion flow field $\mathbf{u}$ between two images f_1 and f_2 defined on $\Omega \subset \mathbb{R}^2$ as minimiser of the energy functional

$$E_1(\mathbf{u}) = \int_\Omega \Big(\big(f_2(\mathbf{x}+\mathbf{u}) - f_1(\mathbf{x})\big)^2 + \alpha \underbrace{\sum_{i=1}^{2} \|\nabla u_i\|^2}_{=:S_1(\mathbf{u})} \Big)\, \mathrm{d}\mathbf{x}.$$

This method suffers from the aforementioned drawbacks: It only estimates the motion between the images, and the first-order smoothness term S_1 prefers solutions biased towards translational motion.

As a remedy, we derive a *new model* that replaces S_1 by two new terms. It minimises the energy

$$E_2(\mathbf{u},\mathbf{A}) = \int_\Omega \Big(\big(f_2(\mathbf{x}+\mathbf{u}) - f_1(\mathbf{x})\big)^2 + \alpha_1 \Big(\underbrace{\sum_{i,j=1}^{2} \big(a_{i,j} - \partial_{x_j} u_i\big)^2}_{=:M(\mathbf{u},\mathbf{A})} + \alpha_2 \underbrace{\sum_{i,j=1}^{2} \|\nabla a_{i,j}\|^2}_{=:S_2(\mathbf{A})} \Big) \Big)\, \mathrm{d}\mathbf{x}.$$

The new terms have the following effects: The similarity term M causes the matrix-valued function $\mathbf{A} = [a_{i,j}]_{2\times 2}$ to contain estimates of the first-order motion derivatives that can be used for the strain tensor computation. Furthermore, the first-order smoothness term S_2 ensures that these estimates are regularised. Finally, combining M and S_2 yields a second-order smoothness assumption for the motion $\mathbf{u}$, which no longer causes the translational bias.

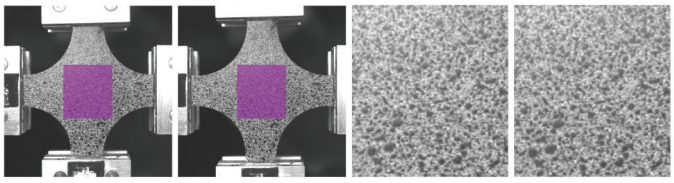


Figure 1: **(a) Left:** Two frames of the biaxial tensile experiment with the areas of interest colored in purple. **(b) Right:** Enlarged versions of the respective areas of interest.

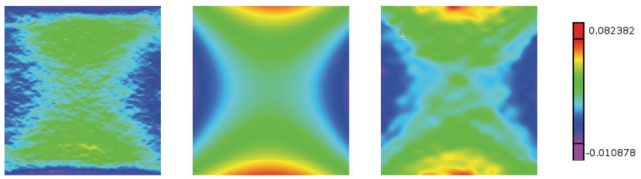


Figure 2: Results for the strain in y-direction obtained from different methods. **(a) Left:** Numerical differentiation of optical flow field. **(b) Centre:** Higher order model obtained from presented approach. **(c) Right:** Commercial software Vic-2D.

This procedure may be repeated by using E_2 as the baseline model and replacing S_2, which indicates that this approach is of generic nature and therefore can be used to derive models of arbitrary order n.

4 Experiment

In order to assess the performance of our new approach, we consider image data of a biaxial tensile experiment that is often conducted in mechanical engineering [4]. We consider the area of interest depicted in Figure 1. In Figure 2, we observe that a higher order model obtained from the presented approach gives much better results than a direct numerical differentiation of the optical flow field. Last but not least, our model is also able to outperform the commercial software Vic-2D that belongs to the leading tools for evaluating data of such experiments.

References

[1] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, August 1988.

[2] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[3] R. W. Ogden. *Non-linear Elastic Deformations*. Dover, New York, 1997.

[4] L. R. G. Treloar. *The Physics of Rubber Elasticity*. Oxford University Press, Oxford, 2005.

[5] W. Trobin, T. Pock, D. Cremers, and H. Bischof. An unbiased second-order prior for high-accuracy motion estimation. In G. Rigoll, editor, *Pattern Recognition*, volume 5096 of *Lecture Notes in Computer Science*, pages 396–405. Springer, Berlin, 2008.

Object Recognition in Multi-View Dual Energy X-ray Images

Muhammet Baştan[1]
mubastan@gmail.com, mbastan@turgutozal.edu.tr
Wonmin Byeon[2]
byeon@rhrk.uni-kl.de
Thomas M. Breuel[2]
tmb@cs.uni-kl.de

[1] Department of Computer Engineering, Turgut Özal University, Ankara, Türkiye (work done while at IUPR)
[2] Image Understanding and Pattern Recognition Group (IUPR), Technical University of Kaiserslautern, Germany

Object recognition in X-ray images is an interesting application of machine vision that can help reduce the workload of human operators of X-ray scanners at security checkpoints. However, automatic inspection systems using machine vision techniques are not yet commonplace for generic threat detection in X-ray images. Moreover, this problem has not been well explored by machine vision community due to the lack of publicly available X-ray image datasets. This paper aims to fill in this gap.

We first present a comprehensive evaluation of image classification and object detection in X-ray images using standard local features in a BoW framework with (structural) SVMs. Then, we extend the features to utilize the extra information available in dual energy X-ray images. Finally, we propose a multi-view branch-and-bound algorithm for multi-view object detection. Through extensive experiments on three object categories (laptops, guns, bottles), we show that the classification and detection performance substantially improves with the extended features and multiple views.

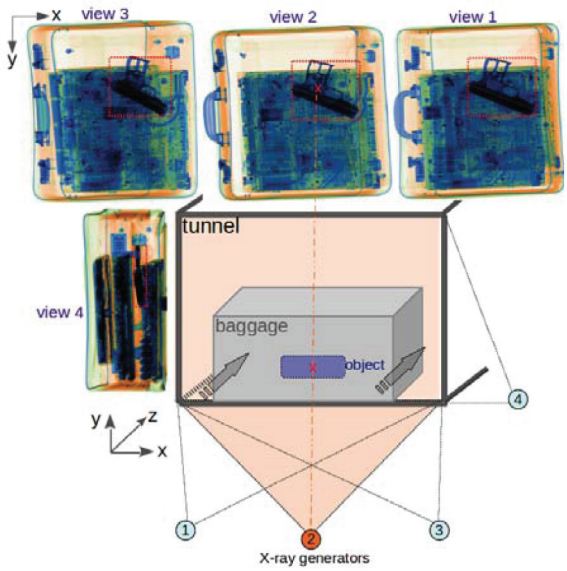


Figure 1: Simplified model of a 4-view X-ray scanner. As the baggage goes through the tunnel (in z direction), 4 X-ray generators scan one slice of the baggage ($x-y$ plane) to generate two energy images for each of the 4 views. The energy images are combined through a look-up table to obtain the pseudo color RGB images for better viewing for the machine operators. The pseudo colors encode the material information (e.g., blue: metals, orange: organic). We make use of two main properties of these X-ray images to boost the object recognition performance: (1) material information, (2) multiple views.

X-ray images contain much ***less texture*** compared to regular photographic images, with the implication of harder object recognition. X-ray images may get extremely ***cluttered***, with many overlapping, possibly high density objects, making the inspection/recognition impossible even for the human screeners (in which case manual inspection is required). Objects usually undergo in– and out-of-plane ***rotations***, which is another major difficulty for recognition. On the other hand, the problem of change of object ***scale*** and ***illumination*** in regular images is not a big issue in X-ray images, since the machine geometry and energy levels are fixed. More importantly, the material information provided by ***dual energy imaging*** and multiple view images provided by ***multi-view imaging*** are both very helpful for improving the performance of object recognition.

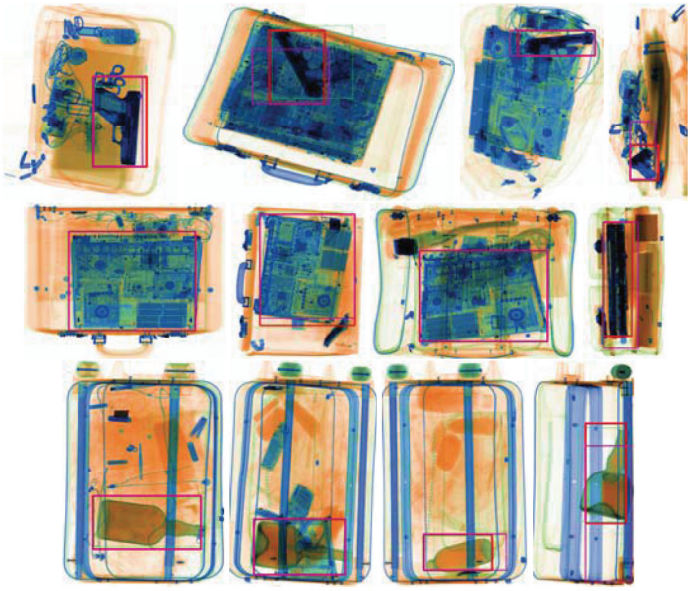


Figure 2: Sample single/multi-view object (handgun, laptop, glass bottle) detections. Green box (thin): ground truth, purple box (medium thickness): single view detection, red box (thick): multi-view detection.

We show that using multiple local color and texture features (different point detectors + descriptors) improves classification and detection performance. Furthermore, utilizing the material information, encoded in the dual energy images or pseudo color images, with extended SPIN descriptors (ESPIN, CSPIN) significantly improves the performance. For textureless object classes (e.g., bottles), contour and super pixel based features are better than sparse texture features. We present the results of extensive experiments with combinations of various features.

Multi-view imaging helps the screeners examine the baggage contents better from multiple viewing angles. We show that multi-view object detection improves over single view object detection [1, 3], especially when the single view detection does not work well. More specifically, we propose a multi-view branch-and-bound search algorithm to estimate the 3D location of an object passing through the scanner tunnel. To do so, we approximate the 3D bounding box of an object with 5 parameters (instead of 6, so that the B&B search converges in reasonable time) by making use of the machine geometry. We project the 3D box into 2D views and use the local features from 2D view images directly.

We show that our single wiew detector with multiple local features (BoWs + structural SVM + single view B&B search) works better than the HOG-based single view detector (HOG + linear SVM + sliding-rotating windows) proposed in [2]. Our multi-view object detector also works better than the multi-view detector of [2], which fuses the HOG-based single view detections in 3D to reinforce the geometrically consistent detections while suppressing the false detections.

[1] M.B. Blaschko and C.H. Lampert. Learning to Localize Objects with Structured Output Regression. In *ECCV*, 2008.
[2] T. Franzel, U. Schmidt, and S. Roth. Object Detection in Multi-view X-Ray Images. In *DAGM*, 2012.
[3] C.H. Lampert, M.B. Blaschko, and T. Hofmann. Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

Relatively-Paired Space Analysis

Zhanghui Kuang
http://i.cs.hku.hk/~zhkuang/
Kenneth K.Y. Wong
http://i.cs.hku.hk/~kykwong/

Department of Computer Science
The University of Hong Kong
Hong Kong

It is very common that an object can have very different presentations in different modalities. For instance, printed and hand-written forms of the same character can look very different, so are face photo and face sketch of the same person. Humans have little problem in recognizing objects across different modalities (e.g., matching face sketches to face photos). In contrast, conventional machine learning methods, such as k-NN classifiers, perform poorly in cross-modality pattern recognition since they assume both the training data and test patterns are randomly sampled from the same distribution (which is not the case in cross-modality pattern recognition) [9].

There exist a number of research studies in the literature targeting at cross-modality pattern recognition, which can be roughly classified into one of the three main approaches. The first approach consists of transforming one modality into another in a preprocessing step [2, 10]. The second approach is by extracting modality-invariant features to represent an object [6, 12]. A major limitation of these two approaches is that methods based on these approaches are usually tailor-made for each different modality pair involved in different recognition tasks. The third approach is to find an underlying latent common space shared between different modalities [3, 4, 7, 8, 9]. Unlike the first two approaches, the third approach does not depend on task-dependent knowledge. Methods based on the third approach are therefore general frameworks that can be applied to different applications. Existing methods of the third approach often require absolutely-paired observations as training data. We refer to them as *Absolutely-Paired Space Analysis* (APSA). These methods assume the projections of paired observations being dependent in the latent space, and can only represent a binary relationship between observations (i.e., either paired observations or non-paired observations).

In this paper, we propose a general framework named *Relatively-Paired Space Analysis* (RPSA) which works on relatively-paired observations. Note that RPSA is *not* a trivial extension of APSA as they are based on completely different models. APSA methods are often based on generative models [1, 3, 7] which either explicitly or implicitly assume the distributions of model parameters and noise (e.g., Gaussian distribution). The final estimation will be unreliable when real data do not fit the assumption. As opposed to APSA, our method is based on a discriminative model that has no distribution assumption. Besides, APSA methods learn a projection function for each modality by exploring the statistics dependence of the projections of absolutely-paired observations in the latent common space. This one-to-one absolute-pairing requirement makes them not suitable for relatively-paired observations. In our proposed framework, we compute the projection functions by preserving the relative proximities of observations in the latent common space (i.e., if observations a and b are more-likely-paired than observations a and c, then the distance between the projections of a and c in the latent common space is assumed to be longer than that between a and b).

Consider a set of M modalities $\{\Omega_1,\Omega_2,\ldots,\Omega_M\}$ with dimensions $\{d_1,d_2,\ldots d_M\}$ respectively, and a training dataset of N observations $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ with a corresponding flag set $\{t_1,t_2,\ldots,t_N\}$ such that $t_i\in\{1,\ldots,M\}$ indicates that $\mathbf{x}_i$ comes from Ω_{t_i}. Let the relative-pairing knowledge of the observations be represented by a set of triplets $S=\{(i,j,k)\}$, where each triplet (i,j,k) encodes that $\mathbf{x}_i$ and $\mathbf{x}_j$ are more-likely-paired than $\mathbf{x}_i$ and $\mathbf{x}_k$. We have

$$\|\mathbf{W}_{\Omega_{t_i}}\mathbf{x}_i-\mathbf{W}_{\Omega_{t_j}}\mathbf{x}_j\|^2\le\|\mathbf{W}_{\Omega_{t_i}}\mathbf{x}_i-\mathbf{W}_{\Omega_{t_k}}\mathbf{x}_k\|^2, \tag{1}$$

where $\mathbf{W}_{\Omega_m}$ is the linear projection matrix for each modality Ω_m. Let $\mathbf{W}=[\mathbf{W}_1\ \mathbf{W}_2\ \ldots\mathbf{W}_M]$, $\mathbf{A}=\mathbf{W}^{\mathrm{T}}\mathbf{W}$, and $\mathbf{A}_{\Omega_m}$ be a $\sum d_n\times d_m$ matrix with all elements being zero except for row $\sum_{n<m}d_n+1$ to row $\sum_{n\le m}d_n$ being an identity matrix, such that $\mathbf{W}_{\Omega_m}=\mathbf{W}\mathbf{A}_{\Omega_m}$. Substituting this into (1) gives

$$\mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,k})-\mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j})\ge 0\quad\forall(i,j,k)\in S, \tag{2}$$

where $\mathrm{Tr}(.)$ gives the trace of a matrix, and

$$\mathbf{C}_{i,j}=(\mathbf{A}_{\Omega_{t_i}}\mathbf{x}_i-\mathbf{A}_{\Omega_{t_j}}\mathbf{x}_j)(\mathbf{A}_{\Omega_{t_i}}\mathbf{x}_i-\mathbf{A}_{\Omega_{t_j}}\mathbf{x}_j)^{\mathrm{T}}. \tag{3}$$

Let $\mathbf{C}_{i,j,k}=\mathbf{C}_{i,k}-\mathbf{C}_{i,j}$. By introducing a positive slack variable to each relative proximity constraint (for improving robustness against noise) and a regularization term, learning the projection matrix for each modality can be reformulated into an SVM style [11] energy function, given by

$$\begin{aligned}&\min\quad\|\mathbf{A}\|_{\mathrm{F}}^2+\gamma\textstyle\sum\xi_{i,j,k}\\&s.t.\quad\mathrm{Tr}(\mathbf{A}\mathbf{C}_{i,j,k})\ge 1-\xi_{i,j,k},\ \mathbf{A}\succeq 0\ \text{and}\ \xi_{i,j,k}\ge 0,\ \forall(i,j,k)\in S,\end{aligned} \tag{4}$$

where $\|\mathbf{A}\|_{\mathrm{F}}$ is the Frobenius norm of $\mathbf{A}$, and γ controls the relative weights of the regularization and loss terms.

We optimize (4) efficiently by maximizing its dual problem alternatively with eigenvalue decomposition and off-the-shelf first order Newton algorithm such as L-BFGS-B [5]. After getting the optimum $\mathbf{A}^*$, we obtain $\mathbf{W}$ by eigenvalue decomposition.

In conclusion, relative-pairing can explore more general semantic relationships between observations than absolute-pairing, and allows easy integration of label information. Theoretically, RPSA is a discriminative model which does not assume any parameter or noise distribution, and is a general framework which can be used in any cross-modality pattern recognition. We have evaluated the performance of RPSA by applying it to cross-pose face recognition and feature fusion. Experimental results show that RPSA outperforms other state-of-the-art techniques, some of which are tailored for the particular problems. We have made the code available online (http://i.cs.hku.hk/~zhkuang/Software.html).

[1] F.R. Bach and M.I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005.

[2] V. Blanz, P. Grother, P.J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *CVPR*, volume 2, pages 454–461, 2005.

[3] M. Borga, H. Knutsson, and T. Landelius. Learning canonical correlations. In *SCIA*, volume 1, pages 1–8, 1997.

[4] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, pages 13–26, 2006.

[5] D.C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

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

[7] S.J.D. Prince, J. Warrell, J.H. Elder, and F.M. Felisberti. Tied factor analysis for face recognition across large pose differences. *PAMI*, 30(6):970–984, 2008.

[8] T. Sun, S. Chen, J. Yang, and P. Shi. A novel method of combined feature extraction for recognition. In *ICDM*, pages 1043–1048, 2008.

[9] J.B. Tenenbaum and W.T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[10] B. Xiao, X. Gao, D. Tao, Y. Yuan, and J. Li. Photo-sketch synthesis and recognition based on subspace learning. *Neurocomputing*, 73 (4-6):840–852, 2010.

[11] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In *NIPS*, pages 2214–2222, 2009.

[12] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520, 2011.

Sparse-Coded Features for Image Retrieval

Tiezheng Ge
getzh@mail.ustc.edu.cn
University of Science and Technology of China
Hefei, China

Qifa Ke
qke@microsoft.com
Microsoft Bing
Sunnyvale, CA, US

Jian Sun
jiansun@microsoft.com
Microsoft Research Asia
Beijing, China

The bag-of-features(BOF) image representation [7] is popular in large-scale image retrieval. With BOF, the memory to store the inverted index file and the search complexity are both approximately linearly increased with the number of images. To address the retrieval efficiency and the memory constraint problem, besides some improvement work based on BOF, there come alternative approaches which aggregate local descriptors in one image into a single vector using Fisher Vector [6] or Vector of Local Aggregated Descriptor (VLAD) [1]. It has been shown in [1] that with as few as 16 bytes to represent an image, the retrieval performance is still comparable to that of the BOF representation.

In this paper, we illustrate that Fisher Vector, VLAD and BOF can be uniformly derived in two steps: **i Encoding** – separately map each local descriptor into a code, and **ii Pooling** – aggregate all codes from one image into a single vector. Motivated by the success of these two-step approaches, we propose to use sparse coding(SC) framework to aggregate local feature for image retrieval. SC framework is firstly introduced by [10] for the task of image classification. It is a classical two-step approach:

Step 1: Encoding. Each local descriptor $\mathbf{x}$ from an image is encoded into an N-dimensional vector $\mathbf{u} = [u^1, u^2, ..., u^N]$ by fitting a linear model with sparsity (L_1) constraint:

$$\min_{\mathbf{u}} \quad \|\mathbf{x} - \mathbf{u}\mathtt{V}\|_2^2 + \lambda\|\mathbf{u}\|_1, \qquad (1)$$
$$\text{subject to} \quad \mathbf{u} \succeq 0$$

Here $\mathtt{V} = [\mathbf{v}_1, ..., \mathbf{v}_N]^\top$ is the codebook, i.e., a set of over-complete bases. It's learned beforehand by alternative optimization [10]. The second term is the L_1 penalty term to enforce sparsity on the vector $\mathbf{u}$, and λ is the parameter to control the sparsity.

Step 2: Pooling. For a given image with T descriptors $\mathtt{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, after obtaining their corresponding sparse codes $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_T\}$ in Step 1, we now pool them into a single N-dimensional vector $\mathbf{y} = [y^1, ..., y^i, ...y^N]$. There are two often-used pooling methods—average pooling and max pooling:

$$\text{average pooling:} \quad y^i = \sum_{t=1}^{T} u_t^i \qquad (2)$$
$$\text{max pooling:} \quad y^i = \max\{u_t^i \,|\, t = 1, ..., T\} \qquad (3)$$

The pooled vector $\mathbf{y}$ is subsequently normalized by $\mathbf{y} := \mathbf{y}/\|\mathbf{y}\|_2$ to generate the final single-vector representation. In practice we adopt max pooling, for it generally works better in experiment.

Despite some implementation differences, the SC framework can be easily applied for image retrieval. And as expected, it works well in experiments–be slightly superior to Fisher Vector and VLAD given the same local features. We take it to be our first contribution.

As our second contribution, we propose the strategy to combine multiple local features, which can greatly improve search accuracy while not hurt efficiency. In particular, we propose the following *multiple feature pooling and compression*:

1. Extract multiple local descriptor sets from a given image (one set for each type of descriptors);
2. Apply sparse coding to each descriptor set to derive a single aggregated vector;
3. Jointly compress all sparse-coded aggregated vectors to derive the final representation in a compact vector.

More formally, if we extract K sets of local descriptors and apply sparse coding and max pooling on each set independently to obtain aggregated vectors $\{\mathbf{y}_1, ..., \mathbf{y}_K\}$. Then, the final image representation is a concatenated vector:

$$\mathbf{y}^* = [\mathbf{y}_1^\top, ..., \mathbf{y}_K^\top]^\top. \qquad (4)$$

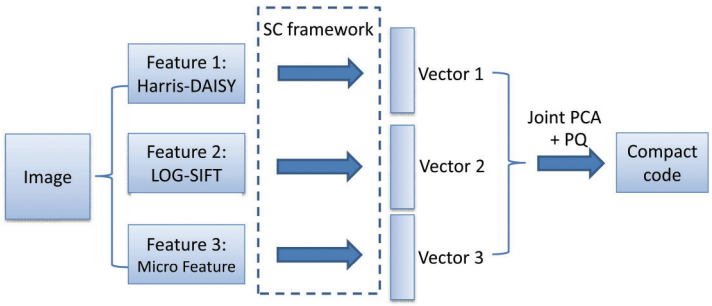


Figure 1: System pipeline

we found PCA performs well in reducing the dimension and achieve further compression (*e.g.*, to 20-40 bytes) with product quantization(PQ) [3].

To find effective and complementary local feature combination, we explored different local feature detectors [4](*e.g.*, LOG, Harris, MSER) associated with different types of descriptors(*e.g.*, SIFT, DAISY [8]). We tried various combinations of multiple detectors and descriptors, and empirically found that the best two-type combination was Harris-DAISY(HD) and LOG-SIFT(LS).

As our third contribution, we further propose a kind of novel color feature(called *micro feature*) which can well complement existing local invariant features. It is inspired by the work of bag-of-colors(BOC) [9] and can be seen as its extension.

The *micro feature* is quite simple: First, we densely sample the image on a grid with some predefined step size s. Then, at each sampled point, we extract a vector consisting of color values of a small $k \times k$ color patch in the CIE-Lab color space, resulting in a $3k^2$ dimensional vector, the *micro feature*. Finally, we feed all extracted vectors into the same sparse coding framework to produce our *sparse-coded micro features*.

The pipeline of complete system is depicted in Fig. 1.

We use two standard datasets (Holidays [2] and UKB [5]) to evaluate the performance of sparse-coded local invariant features, micro features, and their combinations.

[1] H. Jégou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[2] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[3] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011.

[4] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[5] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[6] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[7] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[8] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *CVPR*, 2008.

[9] Christian Wengert, Matthijs Douze, and Herve Jégou. Bag of colors for improved image search. In *ACM Multimedia*, 2011.

[10] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classication. In *CVPR*, 2009.

Combination of paths for interactive segmentation

Julien Mille
http://liris.cnrs.fr/~jmille

Sébastien Bougleux
https://bougleux.users.greyc.fr

Laurent D. Cohen
https://www.ceremade.dauphine.fr/~cohen

Université de Lyon, CNRS
Université Lyon 1, LIRIS, UMR5202
F-69622, Villeurbanne, France
Université de Caen-Basse Normandie, CNRS, GREYC, UMR6072
F-14050, Caen, France
Université Paris Dauphine, CNRS
CEREMADE, UMR7534
F-75016, Paris, France

Boundary-based interactive segmentation methods aim to build a closed contour, very often using minimum cost paths linking user-provided ordered landmark points. While discrete methods rely on a graph modeling of the image, continous models address the problem by means of curve optimization. Among them, the minimal path method [1] find curves of minimal length defined from an edge-based cost function $P = g + w$, where g is a decreasing function of the image gradient magnitude defined over image domain $\mathcal{D}$, and w is a regularizing constant. Given two points a and b, the *geodesic path* between them,

$$\gamma_{a,b} = \underset{\mathcal{C}\subset\mathcal{D}}{\operatorname{argmin}} \left\{ L[\mathcal{C}] = \int_0^1 P(\mathcal{C}(u)) \left\| \mathcal{C}'(u) \right\| \mathrm{d}u \right\} \text{ s.t. } \begin{cases} \mathcal{C}(0) = a \\ \mathcal{C}(1) = b \end{cases} \quad (1)$$

can be obtained by considering the geodesic distance map, or *minimal action map*, $U_a : \mathcal{D} \to \mathbb{R}^+$, efficiently computable by the Fast Marching (FM) method [3]. The minimal path between a and b can be extracted by a gradient descent on U_a, starting from b until a is reached. This can fail to extract the desired curve, for instance when P is too noisy or when the length of the target curve is too important. To address this issue, several approaches build a *piecewise-geodesic curve*, i.e. a concatenation of geodesic curves connecting pairs of successive landmark points or *vertices*. Among them, the geodesically linked active contour (GLAC) model [2] is generated by joining end-to-end geodesic paths built from a set of vertices $\mathcal{V} = \{\mathbf{v}_i\}_{1 \le i \le n}$. The GLAC approach consists in finding the sequence of n vertices which generates a piecewise-geodesic curve minimizing a weighted sum of an edge-fitting term and a region homogeneity term. While this model allows to blend the benefits of minimal paths and region-based terms and is relatively robust to local minima, it turns out to have a significant drawback, as its initial state is not necessarily a simple closed curve. This can occur for instance when the initial vertices are unevenly distributed around the target boundary. In this case, geodesics gather on particular sides of the target boundary, overlapping each other.

Hence, we propose a more relevant contour construction preserving the advantages of piecewise-geodesic curves. Assuming that a set of several possible relevant paths is available for each pair of successive vertices, we may select a single path from each set and combine these paths in order to build the best boundary curve. Let $\mathcal{A}_i$ be a set of K_i *admissible paths* linking the two vertices,

$$\mathcal{A}_i = \{\gamma_{i,j}\}_{1 \le j \le K_i},$$

which we refer to as *admissible set*. Curve $\gamma_{i,1}$ is the geodesic path between $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$. To generate supplemental paths $\gamma_{i,j}, 2 \le j \le K_i$, we propose an approach based on the extraction of *saddle points*. When propagation is performed from two source points a and b, yielding the *combined action map* $U_{a,b} = \min(U_a, U_b)$, the two propagation fronts meet at the **saddle points**, which are the mountain passes on the different roads travelling from a to b. For each pair of successive vertices $(\mathbf{v}_i, \mathbf{v}_{i+1})$, the combined action map is propagated simultaneously from $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$, generating two monotonically advancing fronts. Among the locations where the two fronts collide, those which are also local minima of $U_{\mathbf{v}_i,\mathbf{v}_{i+1}}$ in the direction orthogonal to the propagation direction are kept as the K_i saddle points. Two gradient descents are performed both sides apart from each saddle point and the resulting assembled paths make up the admissible set $\mathcal{A}_i$. The combined action map, the detected saddle points and corresponding paths are shown in Fig. 1 (top row, left).

The computation of an admissible closed contour consists in picking up one path per admissible set $\mathcal{A}_i$ such that the resulting concatenation of selected paths minimizes energy E, designed to penalize contours that are not simple, poorly fitting to image edges or enclosing regions with overlapping color distributions:

$$E[\Gamma] = E_{\text{simplicity}}[\Gamma] + \omega_{\text{edge}} E_{\text{edge}}[\Gamma] + \omega_{\text{region}} E_{\text{region}}[\Gamma].$$

Basically, $E_{\text{simplicity}}$ is a novel term involving measures of self-overlapping and self-intersection. The edge term sums up a decreasing function of the gradient magnitude along the contour, whereas the region term is based on the Bhattacharya coefficient between color distributions of inner and outer regions. To determine the best sequence of labels $\{x_1, \ldots, x_n\}$, instead of an exhaustive search, we propose a greedy search in $O(n^2 K_{\max})$, where $K_{\max}$ is the maximum number of admissible paths over all sets $\mathcal{A}_i$. In each admissible set $\mathcal{A}_i$, paths are sorted according to increasing *exteriority* $\mathcal{X}$, i.e. the signed area, calculated with Green's theorem, formed by a given path $\mathcal{C}$ and the line segment from $\mathcal{C}(1)$ returning to $\mathcal{C}(0)$. The vertices being located clockwise, admissible paths are sorted from the innermost to the outermost (see Fig. 1, middle and bottom rows). Starting from an initial labelling corresponding to the most interior configuration, labels are changed according to a local search, by iteratively testing candidate labellings. At each iteration, candidate sequences are tested that differ from a single label from the current labelling, by increasing labels only. The candidate sequence leading to the smallest energy is kept along the iterations.

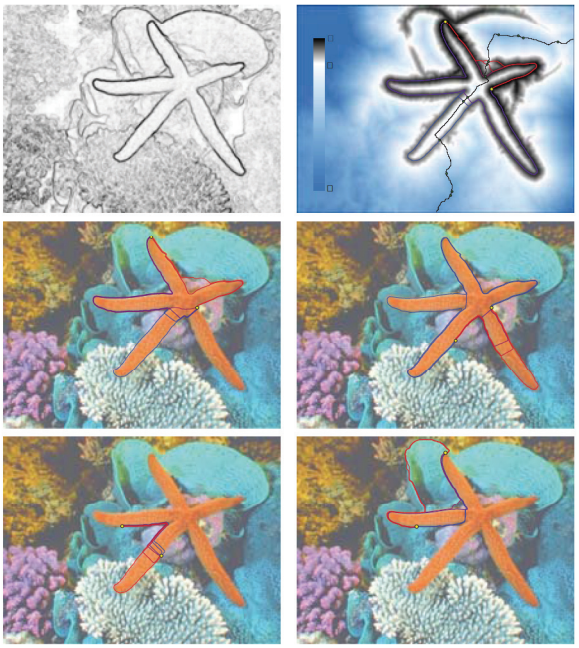


Figure 1: Top row: potential P. Medial curve (black) and saddle points (green) with corresponding paths drawn over combined action map. Middle and bottom row: Admissible set for each pair of successive vertices with $n = 4$. Paths are sorted according to their exteriority

[1] L. Cohen and R. Kimmel. Global minimum for active contour models: a minimal path approach. *IJCV*, 24(1):57–78, 1997.

[2] J. Mille and L. Cohen. Geodesically linked active contours: evolution strategy based on minimal paths. SSVM, volume 5567 of *LNCS*, pages 163–174, Voss, Norway, 2009. Springer.

[3] J.A. Sethian. A fast marching level set method for monotonically advancing fronts. *PNAS*, 93(4):1591–1595, 1996.

Collaboratively Regularized Nearest Points for Set Based Recognition

Yang Wu
mm.media.kyoto-u.ac.jp/members/yangwu
Michihiko Minoh
mm.media.kyoto-u.ac.jp/members/minoh
Masayuki Mukunoki
mm.media.kyoto-u.ac.jp/members/mukunoki

Academic Center for Computing and Media Studies
Kyoto University
Kyoto, 606-8501, Japan

Set based recognition (*i.e*, using a set of instances together for recognition) has been attracting more and more attention in recent years, benefitting from two facts: the difficulty of collecting sets of images for recognition fades quickly, and set based recognition models generally outperform the ones for single instance based recognition. In the past few years various approaches have been proposed, which were reviewed in [4] and [7].

In this paper, we propose a novel model called collaboratively regularized nearest points (CRNP) which inherits the merits of simplicity, robustness, and high-efficiency from the very recently introduced regularized nearest points (RNP) method [7] on finding the set-to-set distance using the l_2-norm regularized affine hulls. Meanwhile, CRNP makes use of the powerful discriminative ability induced by collaborative representation, following the same idea as that in sparse recognition for classification (SRC) for image-based recognition [3] and collaborative sparse approximation (CSA) for set-based recognition [5]. However, CRNP uses l_2-norm instead of the expensive l_1-norm for coefficients regularization, which makes it much more efficient.

Given the test/query set $\mathbf{Q}$ and all the training/gallery sets $\mathbf{X}_i, i \in \{1,\dots,n\}$, CRNP solves the following optimization problem:

$$\begin{aligned} &\min_{\alpha,\beta} \left\{ \|\mathbf{Q}\alpha - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\alpha\|_2^2 + \lambda_2 \|\beta\|_2^2 \right\}, \\ &s.t. \quad \textstyle\sum_k \alpha_k = 1, \sum_{i=1}^{n} \sum_j \beta_i^j = 1, \end{aligned} \tag{1}$$

where $\mathbf{X} = [\mathbf{X}_1,\dots,\mathbf{X}_n]$ denotes all the training/gallery sets together; $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T,\dots,\boldsymbol{\beta}_n^T]^T$ are the corresponding coefficients for these sets; λ_1 and λ_2 are trade-off parameters.

This problem inherits the distance finding model from RNP, however, it performs collaborative set-to-sets distance finding using all the training/gallery sets instead of the independent set-to-set distance finding in RNP as illustrated in Figure 1.

The optimization problem (1) with equality constraints can be transformed to the following unconstrained optimization problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \left\|\mathbf{z} - \hat{\mathbf{Q}}\boldsymbol{\alpha} - \hat{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}, \tag{2}$$

where $\mathbf{z} = [\mathbf{0}_{1,m}, \sqrt{\gamma_1}, \sqrt{\gamma_2}]^T$ with m denoting the dimensionality of the image feature space and γ_1 and γ_2 denoting the Lagrangian multipliers. $\hat{\mathbf{Q}} = [\mathbf{Q}^T, \sqrt{\gamma_1}\mathbf{1}_{N_q,1}, \mathbf{0}_{N_q,1}]^T$ in which N_q is the number of samples in $\mathbf{Q}$, and $\hat{\mathbf{X}} = [-\mathbf{X}^T, \mathbf{0}_{N_x,1}, \sqrt{\gamma_2}\mathbf{1}_{N_x,1}]^T$ where N_x is the number of samples in $\mathbf{X}$. $\mathbf{0}_{i,j}$ and $\mathbf{1}_{i,j}$ denote the $i \times j$ zero matrix and the $i \times j$ dimensional matrix of ones, respectively.

We follow [7] on alternatively optimizing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which avoids the time-consuming matrix inverse operation of an integrated matrix containing both $\mathbf{Q}$ and $\mathbf{X}$ for each test/query set $\mathbf{Q}$. More concretely, when $\boldsymbol{\alpha}$ is fixed, $\boldsymbol{\beta}$ has a closed-form solution

$$\boldsymbol{\beta}^* = \mathbf{P}_x \left(\mathbf{z} - \hat{\mathbf{Q}}\boldsymbol{\alpha}\right), \tag{3}$$

where $\mathbf{P}_x = \left(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \lambda_2\mathbf{I}\right)^{-1}\hat{\mathbf{X}}^T$ (with $\mathbf{I}$ denoting the identity matrix) only depends on $\mathbf{X}$, so it can be pre-computed. When $\boldsymbol{\beta}$ is fixed, $\boldsymbol{\alpha}$ also has a closed-form solution

$$\boldsymbol{\alpha}^* = \mathbf{P}_q \left(\mathbf{z} - \hat{\mathbf{X}}\boldsymbol{\beta}\right), \tag{4}$$

where $\mathbf{P}_q = \left(\hat{\mathbf{Q}}^T\hat{\mathbf{Q}} + \lambda_1\mathbf{I}\right)^{-1}\hat{\mathbf{Q}}^T$.

As claimed in [7], the objective function in Formula (2) has a lower bound of 0 and it is jointly convex w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Since in the alternative optimization, each step on updating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ decreases the objective, the iteration will converge to the global optimal solution. In our experiments to be presented, the iteration usually terminates in no more than 10 steps.

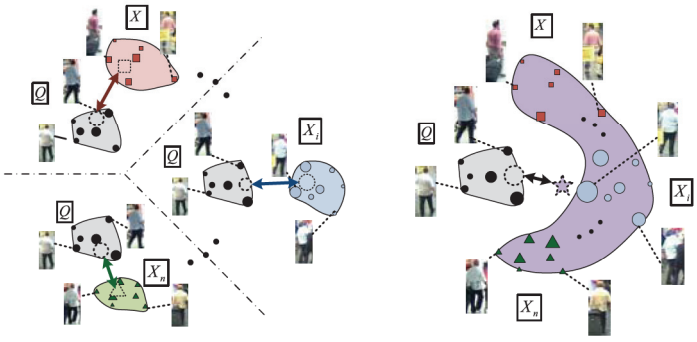


Figure 1: Motivation illustration. (a) The set-to-set distances generated by traditional independent distance finding approaches; (b) the set-to-sets distance got by collaborative distance finding methods.

Besides that, CRNP has a different classification model from that of RNP which makes use of the discriminative coefficients generated by collaborative distance finding. We define the dissimilarity between $\mathbf{Q}$ and $\mathbf{X}_i, i \in \{1,\dots,n\}$ as

$$d_{CRNP}^i = \left(\|\mathbf{Q}\|_* + \|\mathbf{X}_i\|_*\right) \cdot \left\|\mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}_i^*\right\|_2^2 / \left\|\boldsymbol{\beta}_i^*\right\|_2^2, \tag{5}$$

where $\|\mathbf{Q}\|_*$ is the nuclear norm of $\mathbf{Q}$, i.e. the sum of the singular values of $\mathbf{Q}$. Then, $\mathbf{Q}$ is classified by

$$C(\mathbf{Q}) = \arg\min_i \left\{ d_{CRNP}^i \right\}. \tag{6}$$

Extensive experiments on five benchmark datasets for face recognition and person re-identification demonstrate that CRNP is not only more effective but also significantly faster than other state-of-the-art methods.

Table 1: Computational cost comparison with all the related methods on all of the recognition tasks. The results are averaged over 10 random trials if applicable, and we report them in the "milliseconds per sample" manner to eliminate the influence of dataset size variation. The best results for the methods excluding the less robust "CRC" model are shown in bold.

Dataset	MPD[2]	SRC[3]	CRC[8]	CHISD[1]	SANP[6]	SBDR[4]	CSA[5]	RNP[7]	CRNP
Honda/UCSD (50)	3.2	1.2×10^3	*0.28*	77.7	19.6	259	17.4	11.5	**0.32**
Honda/UCSD (100)	6.4	4.1×10^3	0.55	330	17.3	97.8	32.6	14.5	**0.46**
CMU MoBo (50)	12.4	7.6×10^3	*0.94*	89.0	47.2	85.0	29.0	3.5	**2.1**
CMU MoBo (100)	71.4	2.7×10^4	*1.8*	394	53.0	79.3	39.1	5.9	**2.5**
iLIDS-MA	3.9	741	*0.51*	58.7	121	N/A	9.6	24.5	**3.3**
iLIDS-AA	9.9	2337	*1.2*	150	344	N/A	36.8	83.4	**7.2**
CAVIAR4REID	**3.8**	214	*0.35*	55.3	249	N/A	15.8	30.8	8.0

[1] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009.

[4] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao. Set based discriminative ranking for recognition. In *ECCV*, volume 7574, pages 497–510. 2012.

[5] Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao. Collaborative sparse approximation for multiple-shot across-camera person re-identification. In *AVSS*, pages 209 –214, 2012.

[6] Y. Hu and A. S. Mian and R. Owens. Sparse Approximated Nearest Points for Image Set Classification. In *CVPR*, pages 121–128, 2011.

[7] M. Yang, P. Zhu, L.V. Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *FG*, 2013.

[8] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: which helps face recognition? In *ICCV*, 2011.

***WORKSHOP KEYNOTE by Gabriel Brostow* (UCL, UK)**

If Only We Had Tracked Something Like This Before

Abstract: We want algorithms that can tell us what went where in a video. Tracking is hard because each situation is different, featuring a different camera operator, and subject(s) whose appearance, motion, and setting are in a novel configuration. While each video may be pixel-wise unique, we hypothesize that the observed motions are quite similar, at least for tracking purposes. We propose that better tracking can be achieved by learning to automatically associate different videos (or parts) with different algorithms. Instead of seeking an elusive one-size-fits-all tracking strategy (often in the form of an energy function), we advocate keeping multiple strategies, but recognizing when/where to use each. We demonstrate this approach for the problems of optical flow and interest-point tracking.

Biography. Dr Gabriel Brostow is an assistant professor (Lecturer) in Computer Science at the University College London, UK. His group explores research problems relating to Computer Vision and Computer Graphics. His focus is on "Smart Capture" for analysis and synthesis applications. Previously, Gabriel was a visiting researcher at ETH Zürich's CVG Group, and a Marshall Sherfield Fellow in the Computer Vision & Robotics Group at Cambridge University.

WORKSHOP KEYNOTE by Edward Rosten *(Computer Vision Consulting)*

Optimized Corner and Object Detection: a Completely Non-unified Approach

Abstract: Many problems in computer vision involve optimization. Choosing what to optimize can be difficult; firstly because optimization of the appropriate objective may be intractably difficult and secondly because even the correct choice of objective may not be clear. This talk is about optimization in three areas of computer vision: corner detection, object detection and biological optical microscopy. A corner detector should repeatable detect the same corners between images, and ideally should operate efficiently. These objectives can be quantified, and I demonstrate a method for generating optimized corner detectors. In object detection, the definition of a detection versus a misdetection or missed detection is not obvious. On this subject, I will present an object detection system for detecting small objects. This system introduces a new family of features, and detectors optimized for several different definitions of what 'a detection' really is. The third part of this talk is about about using factorial hidden Markov model analysis as an object detection strategy to break the resolution barrier in biological optical microscopy. By optimizing the correct model--an ensemble of fluorescent protein positions – a resolution of up to four times higher than the theoretical resolution limit for this technique can be achieved.

AUTHOR INDEX

Achim , Alin 124
Aghazadeh , Omid 55
Ahuja , Narendra 75
Ahuja , Shaunak 75
Alba-Castro , JoseLuis 96
Alcantarilla , Pablo 23
Al-Hamadi , Ayoub 130
Aloimonos , Yiannis 90
Altwaijry , Hani 25
Amin , Sikandar 56
Andrade-Cetto , Juan 93
Andriluka , Mykhaylo 56
Antunes , Michel 125
Ardabilian , Mohsen 107
Argones-Rúa , Enrique 96
Atmosukarto , Indriyati 75
Avraham , Tamar 57
Azizpour , Hossein 59
Barreto , Joao 125
Barreto , João 135
Bartoli , Adrien 23
Bartoli , Adrien 115
Bartoli , Adrien 136
Bastan , Muhammet 141
Beetz , Michael 29
Behnke , Sven 62
Belhumeur , Peter 24
Belongie , Serge 22
Belongie , Serge 24
Belongie , Serge 25
Bermudez-Cameo , Jesus 79
Besse , Florence 80
Betke , Margrit 74
Beyan , Cigdem 32
Bischof , Horst 30
Bischof , Horst 64
Bischof , Horst 103
Bischof , Horst 105
Boom , Bas 116
Bougleux , Sébastien 144
Brand , Yulia 57
Braux-Zin , Jim 136
Breuel , Thomas 141
Breuss , Michael 53
Brostow, Gabriel 146
Brucker , Ludovic 38
Bruhn , Andres 53
Bull , David 124
Burenius , Magnus 59
Burgos-Artizzu , Xavier 69
Byeon , Wonmin 141
Calway , Andrew 122
Carlsson , Stefan 55
Cavallaro , Andrea 119
Chamveha , Isarun 132
Charles , James 31
Charles , James 58
Charpiat , Guillaume 38
Chen , Dongming 107
Chen , Jie 133
Chen , Liming 107
Cho , Jungchan 92
Christiansen , Eric 22
Chul Ju , Yong 53
Ciptadi , Arridhana 123
Coates , Adam 11
Cohen , Laurent 144
Collins , Toby 115
Collomosse , John 47
Cremers , Daniel 102
Crowley , Elliot 50
Csurka , Gabriela 43
D. Miller , Stephen 35
Daftry , Shreyansh 30
Davis , Larry 94
Davison , Andrew 106
Dellaert, Frank 60
Demetz , Oliver 61
Descombes , Xavier 80
Desingh , Karthik 109
Diebels , Stefan 140
Ding , Xiaoqing 112
Dollar , Piotr 69
Donoser , Michael 49
Donoser , Michael 105
Drummond , Tom 127
Drummond , Tom 128
Dubout , Charles 39
Dubská , Markéta 101
Dupont , Romain 136
Edelkamp , Stefan 29
Elder , James 139
Elgammal , Ahmed 98
Englebienne , Gwenn 66
F Yago Vicente , Tomas 137
Fan , Bin 26
Fang , Chi 112
Feichtenhofer , Christoph 67
Fei-Fei , Li 35
Feng , Jiashi 19

AUTHOR INDEX

Fermuller , Cornelia 90
Ferrari , Vittorio 42
Fisher , Robert 32
Fisher , Robert 116
Fitzgibbon , Andrew 14
Flach , Boris 83
Fleuret , François 39
Flohr , Fabian 77
Florea , Corneliu 71
Florea , Laura 71
Frampton , Rob 122
Fritz , Mario 129
Fuchs , Thomas 138
Furlan , Axel 35
G. Sorrenti , Domenico 35
Gamage , Dinesh 128
Gavrila , Dariu 77
Ge , Tiezheng 143
Ghanem , Bernard 75
Gökmen , Muhittin 119
Gong , Shaogang 46
Gosselin , Philippe-Henri 78
Grauman , Kristen 41
Gravier , Guillaume 110
Greig , Darryl 47
Gruss , Sascha 130
Guan , Wei 45
Guerrero , J.J. 79
Guerrero , Julio 97
Gunes , Hatice 119
Guo , Jun 51
Guo , Jun 108
Haberdar , Hakan 16
Hafner , David 61
Hall , David 69
Hall , Peter 34
Han , Bing 113
Harada , Tatsuya 102
Heaps , Katie 100
Hermans , Tucker 123
Herout , Adam 101
Hertzmann , Aaron 14
Hewer , Alexander 140
Hofer , Manuel 103
Hogg , David 58
Hong , Yi 27
Hoppe , Christof 105
Huang , Chen 112
Huang , Sheng 98
Huang , Xiaohua 87
Hunter , Andrew 47
Imanishi , Masayuki 54
Ishimaru , Kazuhisa 54
Iwashita , Yumi 138
Jammalamadaka , Nataraj 99
Jawahar , CV 36
Jawahar , CV 99
Jawahar , CV 109
Jegou , Herve 110
Jiang , Zhuolin 94
Jin Chang , Hyung 92
John , Vijay 66
Jones , Graeme 86
Jurie , Frederic 20
Jurie , Frederic 114
Kakumu , Yumi 54
Kanatani , Kenichi 37
Kazemi , Vahid 59
Ke , Qifa 143
Kellokumpu , Vili 133
Kim , Jiyun 92
Kim , Tae-Kyun 84
Kleefeld , Andreas 53
Klopschitz , Manfred 105
Kobetski , Miroslav 82
König , Tim 120
Koslosky , Joshua 100
Kpalma , Kidiyo 89
Krause , Andreas 12
Kriegman , David 22
Kriegman , David 24
Krishna , Madhava 109
Krose , Ben 66
Kuang , Zhanghui 142
Kuettel , Daniel 42
Kuipers , Benjamin 40
Kuniyoshi , Yasuo 102
Kwak , Iljung 24
Laga , Hamid 78
Lam , Antony 52
Larlus , Diane 43
Learned-Miller , Erik 28
Leistner , Christian 64
Leonardi , Riccardo 131
Levine , Stacey 100
Lhuillier , Maxime 72
Li , Chun-Guang 51
Li , Chun-Guang 108
Li , Feng 44
Li , Xiong 85

AUTHOR INDEX

Li , Yi 46
Lik Teo , Ching 90
Lindenbaum , Michael 57
Litvinov , Vadim 72
Liu , Bin 113
Liu , Hong 21
Liu , Zhi 89
López , Antonio 97
Lopez-Nicolas , Gonzalo 79
Lourenco , Miguel 125
Lovegrove , Steven 104
Lu , Hongtao 19
Lui , Vincent 127
Lukierski , Robert 106
Luo , Wenhan 84
Ma , Chao 76
Magee , Derek 58
Mahmood , Arif 68
Maire , Michael 95
Malinowski , Mateusz 129
Mantell , Judith 124
Marchesotti , Luca 17
Matsushita , Yuichi 37
Matuszewski , Bogdan 134
Maurer , Michael 30
Mauro , Massimo 131
McCandless , Tomas 41
McDonagh , Steven 116
Mears , Benjamin 28
Menze , Bjoern 38
Mian , Ajmal 68
Micusik , Branislav 117
Mignon , Alexis 114
Mille , Julien 144
Minocha , Ayush 99
Minoh , Michihiko 145
Moreno-Noguer , Francesc 93
Mottini , Alejandro 80
Movahedi , Vida 139
Mukunoki , Masayuki 145
Murillo , Ana 24
Nakayama , Hideki 111
Nam , David 124
Neubert , Peer 65
Niese , Robert 130
Ning , Xi 116
Nuevo , Jesus 23
Oh , Songhwai 92
Okabe , Takahiro 52
Orts-Escolano , Sergio 116
Padgett , Curtis 138
Parkhi , Omkar 18
Patron-Perez , Alonso 104
Pei , Yuru 113
Penate-Sanchez , Adrian 93
Peng , Qiang 70
Peng , qunsheng 121
Peng , Xiaojiang 70
Pérez-Pellitero , Eduardo 48
Perona , Pietro 69
Perona , Pietro 95
Perronnin , Florent 15
Perronnin , Florent 17
Perronnin , Florent 43
Pfister , Tomas 31
Pfister , Tomas 58
Picard , David 78
Pietikainen , Matti 87
Pietikainen , Matti 133
Pinz , Axel 67
Pizarro , Daniel 115
Ponsa , Daniel 97
Pons-Moll , Gerard 14
Porikli , Fatih 44
Protzel , Peter 65
Qi , Xianbiao 51
Qi , Xianbiao 70
Qi , Xianbiao 108
Qiao , Yu 51
Qiao , Yu 70
Qiao , Yu 108
Qin , Xueying 121
Rabaud , Vincent 22
Rajan , Deepu 109
Rak , Marko 120
Raposo , Carolina 125
Riemenschneider , Hayko 131
Rodola , Emanuele 102
Rodriguez , Jose 15
Rohrbach , Marcus 56
Ronsin , joseph 89
Ros , German 97
Rosenhahn , Bodo 48
Rosten, Edward 147
Roth , Peter 64
Ruan , Xiang 76
Ruan , Xiang 85
Ruiz-Hidalgo , Javier 48
Ryoo , Michael 138
Sakaue , Fumihiko 54

AUTHOR INDEX

Salvador , Jordi 48
Samaras , Dimitris 137
Sánchez-Lozano , Enrique 96
Sandilands , Peter 116
Sapiro , Guillermo 126
Sappa , Angel 97
Saquib Sarfraz , M. 88
Sariyanidi , Evangelos 119
Sato , Imari 52
Sato , Jun 54
Sato , Yoichi 52
Sato , Yoichi 132
Savarese , Silvio 35
Scheffer , Tobias 140
Schiele , Bernt 56
Schouwenaars , Filip 118
Schulter , Samuel 64
Seibert , Henning 140
Serradell , Eduard 93
Sevilla-Lara , Laura 28
Shah , Shishir 16
Shao , Ling 63
Shao , Ling 81
Sharma , Gaurav 20
Shi , Xun 33
Shotton , Jamie 14
Sibley , Gabe 104
Siddique , Muhammad Adnan 88
Sidle , Glenn 100
Simonyan , Karen 18
Singh , Digvijay 99
Sixta , Tomas 83
Song , Yi-Zhe 46
Stiefelhagen, Rainer 88
Stommel , Martin 29
Stückler , Jörg 62
Subpa-Asa , Art 52
Sugano , Yusuke 132
Sugaya , Yasuyuki 37
Sugimoto , Akihiro 132
Sullivan , Josephine 59
Sullivan , Josephine 82
Sun , Jian 143
Sun , Qianru 21
Tabia , Hedi 78
Tang , Zhongwei 126
Tao , Lili 134
Tarabalka , Yuliya 38
Taylor , Jonathan 14
Tepper , Mariano 126
Timofte , Radu 118
Tönnies , Klaus-Dietz 120
Tozza , Silvia 53
Traue , Harald 130
Tsai , Grace 40
Tsai , Yi-Hsuan 91
Tsotsos , John 33
Uzair , Muhammad 68
Van Gool , Luc 118
Van Gool , Luc 131
Varadarajan , Jagannadan 75
Vasconcelos , Francisco 135
Vedaldi , Andrea 18
Verkade , Paul 124
Verma , Yashaswi 36
Vertan , Constantin 71
Vranceanu , Ruxandra 71
Walter , Steffen 130
Wang , Bo 33
Wang , Guofeng 121
Wang , Jun 121
Wang , Tinghuai 47
Wang , Zhenhua 26
Weickert , Joachim 61
Weickert , Joachim 140
Wendel , Andreas 30
Wendel , Andreas 103
Werner , Philipp 130
Wiedemeyer , Thiemo 29
Wildenauer , Horst 117
Wildes , Richard 67
Wildes , Richard 73
Wong , Kenneth 142
Wu , Fuchao 26
Wu , Qi 34
Wu , Yang 145
Wu , Zheng 74
Xie , Saining 19
Xu , Tianmin 113
Yan , Ruomei 81
Yan , Shuicheng 19
Yang , Dan 98
Yang , Jimei 91
Yang , Ming-Hsuan 76
Yang , Ming-Hsuan 91
Yang , Xiaokang 76
Yang , Yezhou 90
Yang, Ming-Hsuan 85
Yi , Kwang 92
You , Suya 45

AUTHOR INDEX

Young Choi , Jin	92
Yu , Chen-Ping	137
Yu , Stella	95
Yu , Xiaodong	90
Zha , Hongbin	113
Zhang , Chao	85
Zhang , Chongyang	76
Zhang , Jianming	74
Zhang , Yangmuzi	94
Zhao , Guoying	87
Zhao , Guoying	133
Zhao , Wan-Lei	110
Zhao , Yuming	85
Zhong , Fan	121
Zhong , Hao	73
Zhu , Fan	63
Ziegler , Andrew	22
Zienkiewicz , Jacek	106
Zisserman , Andrew	13
Zisserman , Andrew	18
Zisserman , Andrew	31
Zisserman , Andrew	50
Zisserman , Andrew	58
Zou , Wenbin	89

NOTES