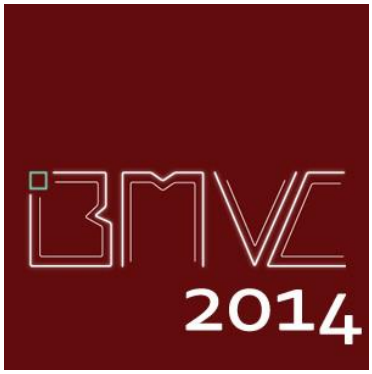


25th British Machine Vision Conference

1st-5th September 2014
Nottingham,
United Kingdom

Abstract Book

Contents

Foreword

Welcome to the 25th British Machine Vision Conference. In its Silver Jubilee year BMVC has appropriately returned to the Jubilee Campus of the University of Nottingham, where it was last held in 1999. BMVC remains one of the strongest events in the computer vision community's calendar, and this year again attracted over 430 submissions by authors from around the world. While BMVC has always been a key meeting for the British vision community, some 74% of this year's accepted papers are from outside the UK, reflecting the conference's growth in stature since the first BMVC in Oxford in 1990. In 2014, 26% of the accepted papers are from a UK-based institute, 36% from Europe (excluding UK), 15% from Asia, 22% from North America and 1% from Australia.

BMVC 2014's international submission profile is matched by its international panel of reviewers and Area Chairs (ACs). Producing three independent reviews and two meta-reviews of each paper is a substantial amount of work, and we are indebted to our team of 174 referees and 43 Area Chairs (ACs). Each reviewer assessed at least 4 papers, with each AC responsible for at least 15 papers. Following vigorous discussion between the reviewers and ACs of each paper, consensus was reached and final decisions were made. The final programme comprises 33 oral presentations and 98 posters, covering a variety of computer vision techniques and problems, which we hope you will enjoy. These figures give BMVC 2014 a podium acceptance rate of 7.5% and an overall acceptance rate of 30%.

In addition to a full programme of submitted papers, BMVC 2014 is proud to include invited talks from Prof Luc Van Gool (ETH Zurich) and Dr Fei-Fei Li (Stanford). This year's tutorial, *Image representations, from shallow to deep* will be given by Dr Andrea Vedaldi of the University of Oxford. We would like to thank them, along with all our other presenters, for their contribution to the conference. We are also grateful to Qualcomm, Microsoft, NVIDIA, Springer and all other sponsors for their financial support of BMVC 2014.

BMVC2014 has been organised by members of the Computer Vision Laboratory (CVL) of the School of Computer Science, University of Nottingham. We would particularly like to thank Susie Lydon, of the

Centre for Plant Integrative Biology, University of Nottingham, for her hard work as Local Arrangements Chair, Mike Pound for creating and maintaining the conference website, Peter Blanchfield for his work as Workshop Chair and Debbie Pitchfork and Felicia Knowles of the School of Computer Science for their invaluable help with administrative matters. The conference would not have run as smoothly without our team of enthusiastic student helpers, and we thank them too. Finally, we would like to thank the BMVA Executive for their support, and the BMVC 2013 team for answering so many questions so quickly.

We hope you find BMVC 2014 and your stay in Nottingham both enjoyable and rewarding.

Michel Valstar, Andrew French, Tony Pridmore

Nottingham, September 2014

Tutorial: Monday 1st September
Andrea Vedaldi - Image representations, from shallow to deep

In this tutorial, I will focus on image recognition using image-based models. The key to the successful application of machine learning methods to image understanding tasks is devising appropriate representations of images. Here, I will review four different representation flavours sampling from the past fifteen years of research: handcrafted features, kernel embeddings, representation obtained by learning a discriminative metric, and, lastly, representations from deep learning. I will stress three key factors of good representations. The first one is power, that is their ability to achieve strong recognition performance in applications. The second one is speed, which for example impacts our ability to learn new visual concepts on the fly, upon a user's request. The last one is compactness, which determines whether large datasets can be stored in small amounts of memory, for example in RAM for fast access. Example applications to large scale indexing, recognition, and retrieval will be demonstrated.

Biography. Andrea Vedaldi is Associate Professor in Engineering Science at the University of Oxford since 2012. His research focuses on the automatic interpretation of images and related problems in machine learning and large scale optimisation. He is author of more than forty papers in major computer vision and machine learning conferences and journals, as well as leading author of the VLFeat computer vision library. From 2008 to 2012 he was postdoctoral researcher and junior research fellow at the University of Oxford, supported by the Glasstone Research Fellowship in Science and the New College W. W. Spooner Fellowships. He is the recipient of the PhD and MSc degrees in Computer Science from the University of California at Los Angeles in 2008 and 2005 respectively (outstanding PhD and MSc thesis awards), and of the BSc degree in Information Engineering by the University of Padua in 2003.

Keynote: Tuesday 2nd September
Luc Van Gool - Winner-uses-all

We tend to go after the best possible algorithm to do X. But we also observe that some methods to do X are better at dealing with certain cases, and others with different cases. This begs the question whether we really need to arrive at one, single 'ueber-algorithm' for X? We explore examples where some relatively cheap pre-processing allows us to select a method that is probably best at handling the particular image / case at hand, and then to apply that specific one method among several alternatives. The extra cost is limited. As memory gets cheaper, the cost of keeping the code of several alternative methods available is affordable. Such winner-uses-all approach is also efficient, as one only needs to run the cheap pre-computation to select the appropriate method and then to apply the one selected method. We give examples from diverse areas, as diverse as texture synthesis, super-resolution, and 3D reconstruction.

Biography. Luc Van Gool is full professor for computer vision at ETH Zurich and KU Leuven. He has worked on different aspects of computer vision. His experience includes texture analysis and synthesis, 2D and 3D object (class) recognition, action recognition and gesture analysis, and passive and active 3D and 4D shape acquisition. He has been a member of the editorial boards of several major computer vision journals, incl. the Int. J. of Computer Vision and the IEEE Trans. on Pattern Analysis and Machine Intelligence. He currently is on the boards of Machine Vision and Applications and the ACM J. on Computing and Cultural Heritage. He also is an editor-in-Chief of the Journal Foundations and Trends in computer Graphics and Vision.

He is a co-founder of several spin-off companies, including Eyetronics (3D modeling, mainly for the movie and games industry, e.g. for James Bond, Lara Croft, and many other movies), kooaba (recognition of landmarks, labels, press articles via photos taken with a mobile phone, acquired by Qualcomm), and GeoAutomation (mobile mapping for 3D measurements, mainly in urban environments).

Keynote: Wednesday 3rd September

Fei-Fei Li - Computer Vision: A Quest for Visual Intelligence

More than half of the human brain is involved in visual processing. While it took mother nature billions of years to evolve and deliver us a remarkable human visual system, computer vision is one of the youngest disciplines of AI, born with the goal of achieving one of the loftiest dreams of AI. The central problem of computer vision is to turn millions of pixels of a single image into interpretable and actionable concepts so that computers can understand pictures just as well as humans do, from objects, to scenes, activities, events and beyond. Such technology will have a fundamental impact in almost every aspect of our daily life and the society as a whole, ranging from e-commerce, image search and indexing, assistive technology, autonomous driving, digital health and medicine, surveillance, national security, robotics and beyond. In this talk, I will give an overview of what computer vision technology is about and its brief history. I will then discuss some of the recent work from my lab towards large scale object recognition. I will particularly emphasize what we call the "three pillars" of AI in our quest for visual intelligence: data, learning and knowledge. Each of them is critical towards the final solution, yet dependent on the other. This talk draws upon a number of projects ongoing at the Stanford Vision Lab.

Biography. Fei-Fei Li is an associate professor at the Computer Science Department and the director of the Vision Lab at Stanford. Fei-Fei's main research interest is in vision, particularly high-level visual recognition. In computer vision, Fei-Fei's interests include image and video classification, retrieval, and understanding. Some of the most recent work in her lab relates to fundamental technological problems related to large-scale Internet data, mobile computing, machine learning and artificial intelligence. In human vision, she has studied the interaction of attention and natural scene and object recognition, and decoding the human brain fMRI activities that are known as "mind reading" of the brain. Fei-Fei is a recipient of the 2011 Alfred Sloan Faculty Award, 2012 Yahoo Labs FREP award, 2009 NSF CAREER award, the 2006 Microsoft Research New Faculty Fellowship and a number of Google Research awards.


Jubilee Campus
& University of Nottingham Innovation Park (UNIP)
MAIN ENTRANCE
24hr Security Gatehouse
Western Boulevard A6514
Crown Island
Melton Hall
PD
G
SC
Wollaton Road A609
Middleton Boulevard A6514
To University Park Campus (0.5 miles)
Nottingham 1
Ilkeston Road A609
To city centre
Main conference venue
The Exchange
The University of Nottingham
Jubilee Campus
Wollaton Road
Nottingham
NG8 1BB
The Sir Harry & Lady Djanogly Learning Resource Centre
Triumph Road
Faraday Road
Southwell Hall
Accommodation
Manor Villages
Sports Centre
BC
Dearing Building
Newark Hall
Railway
YANG Fujia Building
Raleigh Park
Amenities Building
IPD
Sir Colin Campbell Building
The University of Nottingham
Innovation Park
NG7 2TU
National College for Teaching and Leadership
Welcome drinks
Sports ground
Tennis courts
Riverside Point
Pedestrian/cycle route to campus
To University Park Campus (0.5 mile)
Derby Road A6200
24-hour ambulance/fire/police
(0115) 951 8888
24-hour security contact
(0115) 951 3013
0 metres 100
The University of Nottingham
UNITED KINGDOM • CHINA • MALAYSIA
04/2014
© Crown Copyright Licence no. 100030223
Academic schools and departments
Computer Science 4
Engineering 13/16/18
Education 2/6/10/15
International Office 10
Nottingham University Business School 1/7/10
Si Yuan Centre 9
Other services
Auditorium 8
Banks/Retail 2
Cafes 2/5/7/11/30
Careers and Employability Service 10
Faith/Prayer rooms 11
Graduate Centre 11
Libraries 3/7
Sports 30
Student Services Centre 11
Students' Union 10
The University of Nottingham Innovation Park
Aerospace Technology Centre 18
Energy Technologies Building 16
Institute of Mental Health 17
Innovation Park Reception 12
Nottingham Geospatial Building 13
Romax Technology Centre 14
Academic buildings
Residences
Off campus student residences
Other services
Under construction
The University of Nottingham Innovation Park (UNIP)
Footpath
Pay & Display visitor parking
UNIP Pay & Display visitor parking
Blue-badge parking
Gatehouse
Barrier-access control
Secure cycle parking
Hopper bus stop
Public bus stop
Public/Hopper bus stop
Building public entrances
Aspire sculpture

Programme

All talks, except the tutorial, are in Exchange Building, Lecture Theatre 3 (LT3).
Posters and refreshments are in C3 and C33 (one floor below LT3)

Monday 1st September

Time	Event
13.30	Registration opens (Exchange building foyer)
15:30-17:30	Tutorial: Image representations, from shallow to deep (LT2) Andrea Vedaldi
17:30-20:00	Welcome drinks, and registration at NCT&L

Tuesday 2nd September

Time	Event
8:30	Registration opens and coffee
9:00-10:00	**Keynote**: Luc Van Gool: Winner-uses-all
	Person detection and identification
10:00-10:20	Re-id: Hunting Attributes in the Wild Ryan Layne, Tim Hospedales, Shaogang Gong
10:20-10:40	Evidential combination of pedestrian detectors Philippe Xu, Franck Davoine, Thierry Denoeux
10:40-11:10	**Coffee**
	Machine Learning
11:10-11:30	Mining Structure Fragments for Smart Bundle Adjustment Luca Carlone, Pablo Fernandez Alcantarilla, Han-Pang Chiu, Zsolt Kira , Frank Dellaert
11:30-11:50	Distributed Non-Convex ADMM-inference in Large-scale Random Fields Ondrej Miksik, Vibhav Vineet, Patrick Pérez, Phillip Torr
11:50-12:10	Boosted Cross-Domain Categorization Fan Zhu, Ling Shao, Jun Tang
12:10-12:30	Return of the Devil in the Details: Delving Deep into Convolutional Nets Ken Chatfield, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman
12:30-12:50	Transductive Multi-label Zero-shot Learning Yanwei Fu, Yongxin Yang, Tim Hospedales, Tao Xiang, Shaogang Gong
12:50-13:30	**LUNCH**
13:30-14:45	Poster Session I
	Video and Structure From Motion
14:45-15:05	Optimal Representation of Multi-View Video Marco Volino, Dan Casas, John Collomosse, Adrian Hilton
15:05-15:25	Unsupervised Spatio-Temporal Segmentation with Sparse Spectral Clustering Mahsa Ghafarianzadeh, Matthew Blaschko, Gabe Sibley
15:25-16:00	**Coffee**
16:00-16:20	Depth Extraction from Videos Using Geometric Context and Occlusion Boundaries Syed Raza, Omar Javed, Aveek Das, Harpreet Sawhney, Hui Cheng, Irfan Essa
16:20-16:40	Non-Rigid Shape-from-Motion for Isometric Surfaces using Infinitesimal Planarity Ajad Chhatkuli, Daniel Pizarro, Adrien Bartoli
16:40-17:00	Virtual Insertion: Robust Bundle Adjustment over Long Video Sequences Ziyan Wu, Zhiwei Zhu, Han-Pang Chiu

Wednesday 3rd September

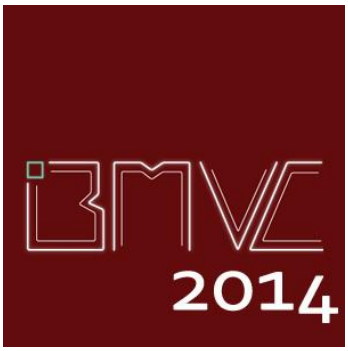
Time	Event
8:30	Registration opens and coffee
9:00-10:00	**Keynote**: Fei-Fei Li: Computer Vision: A Quest for Visual Intelligence
	Faces
10:00-10:20	Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization Adria Ruiz, Joost Van de Weijer, Xavier Binefa
10:20-10:40	Expression-Invariant Age Estimation Fares Alnajar, Zhongyu Lou, Jose Alvarez, Theo Gevers
10:40-11:10	**Coffee**
	3D and Stereo
11:10-11:30	A Stochastic Cost Function for Stereo Vision Christian Unger, Slobodan Ilic
11:30-11:50	Fusing Multiple Features for Shape-based 3D Model Retrieval Takahiko Furuya, Ryutarou Ohbuchi
11:50-12:10	Unsupervised RGB-D image segmentation using joint clustering and region merging Md Abul Hasnat, Olivier Alata, Alain Trémeau
12:10-12:30	CP-Census: A Novel Model for Dense Variational Scene Flow from RGB-D Data David Ferstl, Gernot Riegler, Matthias Rüther, Horst Bischof
12:30-12:50	Is 2D Information Enough For Viewpoint Estimation? Amir Ghodrati, Marco Pedersoli, Tinne Tuytelaars
12:50-13:30	**Lunch**
13:30-14:45	Poster session II
	Segmentation and object detection
14:45-15:05	Discrete Multi Atlas Segmentation using Agreement Constraints Stavros Alchatzidis, Aristeidis Sotiras, Nikos Paragios
15:05-15:25	Video Object Segmentation by Non-Local Consensus voting Alon Faktor, Michal Irani
15:25-16:00	Coffee
16:00-16:20	Embedding Geometry in Generative Models for Pose Estimation of Object Categories Michele Fenzi, Jörn Ostermann
16:20-16:40	Cracking BING and Beyond Qiyang Zhao, Zhibin Liu, Baolin Yin
16:40-17:00	How good are detection proposals, really? Jan Hosang, Rodrigo Benenson, Bernt Schiele
18:00-23:00	Conference banquet at the National Space Centre

Thursday 4th September

Time	Event
8:30	Registration opens and coffee
	Tracking
9:00-9:20	Part Context Learning for Visual Tracking Guibo Zhu, Jinqiao Wang, Chaoyang Zhao, Hanqing Lu
9:20-9:40	Simultaneous Mosaicing and Tracking with an Event Camera Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, Andrew Davison
9:40-10:00	Deformable Template Tracking in 1ms David Joseph Tan, Stefan Holzer, Nassir Navab, Slobodan Ilic
10:00-10:20	Learn++ for Robust Object Tracking Feng Zheng, Ling Shao, James Brownjohn, Vitomir Racic
10:20-10:40	L_0-Regularized Object Representation for Visual Tracking Jinshan Pan, Jongwoo Lim, Zhixun Su , Ming-Hsuan Yang
10:40-11:10	**Coffee**
	Image classification
11:10-11:30	You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, Walterio Mayol-Cuevas
11:30-11:50	Hierarchical Cascade of Classifiers for Efficient Poselet Evaluation Bo Chen, Pietro Perona, Lubomir Bourdev
11:50-12:10	Regularized Max Pooling for Image Categorization Minh Hoai
12:10-12:30	Discriminative Embedding via Image-to-Class Distances Xiantong Zhen, Ling Shao, Feng Zheng
12:30-14:00	**Lunch**
14:00-16:00	Social: Guided tour of Wollaton Hall

Friday 5th September

Time	Event
9:00-16:00	UK Doctoral Consortium workshop (details provided separately)



BMVC 2014

Oral abstracts

Re-id: Hunting Attributes in the Wild

Ryan Layne
r.d.c.layne@qmul.ac.uk
Timothy M. Hospedales
t.hospedales@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Computer Vision Group
Queen Mary University of London
London E1 4NS.
http://qml.io/vision

Re-identification research breaks down into two main areas; developing effective representations that are discriminative for identity whilst invariant to lighting and viewpoint change [2] and development of learning methods trained to discriminate identities [1, 3]. Feature-centric approaches [2, 4] suffer from the problem that it is extremely challenging to obtain features that are discriminative enough to distinguish people reliably, while simultaneously being invariant to all the practical covariates such as motion blur, clutter, view angle and pose change, lighting and occlusion. In contrast, learning approaches [3] better use a given set of features, by discriminatively training models to maximise re-identification performance, for example metric learning [3] and support vector machines (SVM) [1].

In this paper we address these issues by automatically constructing a bottom-up attribute ontology, and learning an effective representation by large-scale mining noisy but abundant content from social photo sharing sites. We discover attributes automatically by clustering photo tag and comment data. These clusters are used to train a large bank of detectors, resulting in a number of visually detectable attributes[1]. This process is significantly more scalable than manually annotating data per surveillance site for attribute learning and the greater volume and diversity of data used to train these automatically discovered attributes results in a more generalisable attribute representation than conventional approaches on surveillance datasets. We validate our contribution and obtain excellent results on a set of four of the most challenging re-identification datasets.

We first apply self-tuned spectral clustering based on the BOW tf-idf metatext representations with a vocabulary of $\approx 5,000$ bigrams. We calculate the similarity between the frequency of the unigrams and bigrams rather than using the Levenshtein distance on the second gram within each bigram. Our intuition is that in our case it is the co-occurrence of the grams that is semantically relevent, not the similarity to other bigrams. Spectral clustering performs well regardless of the spatial arrangement of the underlying clusters, making it suitable for our needs. We extract bounding boxes of people from this extremely varied collection of photos; after conservatively thresholding person detection confidence, we are left with 69,000 person crops with corresponding meta-text features. We train an independent LDA model for each of the $N_a = 200$ discovered attribute clusters. Finally we build a representation for any person's image X in an internet-attribute semantic-space by stacking the positive-class posteriors from each detector into a N_a dimensional vector: $IA(X)$. To compensate for the differences in image quality between internet and surveillance data, we align the two datasets, using domain adaptation.

The attributes obtained thus far are trained directly from discovered text clusters so there is variability in their reliability and their usefulness for re-identification. We therefore address learning a linear weighting $\mathbf{w}$ to rescale the attributes IA such that they are weighted according to their maximum utility for re-identification.

[1]This is in contrast to expert defined ontology, which while intuitive to experts, may correspond to properties not possible to detect reliably with current vision techniques.

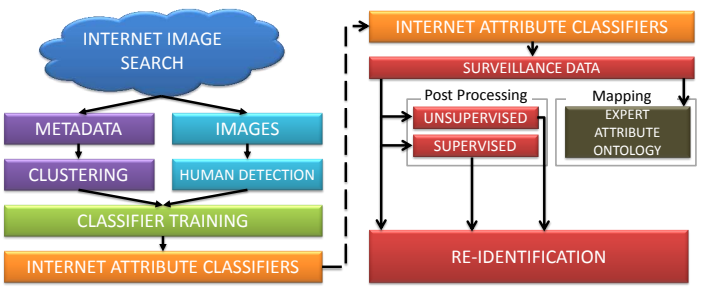


Figure 1: Schematic overview of our pipeline; Post-Processing modules such as distance-metric learning or domain-adaptation can be applied depending on the level of supervision available in order to boost "rank 1" or overall system performance as needed

Figure 2: Our FUSIA internet attribute (IA) representation provides a distributed representation of conventional expert-defined attributes such as "red shirt" (right), meaning that it can be mapped to them to allow query in terms of existing expert attribute ontologies (EAO) built for other surveillance data (SD).

We wish to enforce both a strong early-rank score, and good overall performance. To achieve this, we maximise the *product* of the CMC curve values $\hat{p}(k)$ at each rank k

$$\hat{P}_{\mathbf{w}}(k) = CMC_{\mathbf{w}}(k) = \frac{1}{n}\sum_{p=1}^{n} \mathbf{1}\left(k_p \leq k\right) \quad (1)$$

where k_p is the distribution of the ranks based on NN re-identification using $L1$ distances $D(IA_p, IA_g)$ between each attribute encoded probe $IA_p \in \mathcal{P}$ and all gallery member, $IA_g \in \mathcal{G}, g = 1,...,n$. Specifically we use greedy search to select the weight $\mathbf{w}$ that maximises the following metric when used to scale each dimension/attribute a:

$$\min_{\mathbf{w}} \prod_{k=1}^{n} \hat{P}_{\mathbf{w}}(k) \quad (2)$$

Finally, we integrate our representation with metrics based on other low-level features. Specifically, we fuse BR-SVM [1] (trained on ELF features), SDALF [2] and our weighted internet attributes after further discriminative training [3]. The resulting pseudo-metric's fusion weights *beta* can be trivially selected with standard optimisation methods:

$$D(X_p, X_g) = d_{KISS}(IA(X_p), IA(X_g)) \quad (3)$$
$$+ \beta_{SDALF} \cdot d_{SDALF}(X_p, X_g) \quad (4)$$
$$+ \beta_{BRELF} \cdot d_{BRELF}(X_p, X_g). \quad (5)$$

We perform nearest-neighbour re-identification with the above metric, obtaining state of the art re-identification performance (Figure 3). Our FUSIA representation also provides a distributed representation of conventional expert-defined attributes. It can be mapped to them, thus allowing queries in terms of existing attribute ontologies (Figure 2).

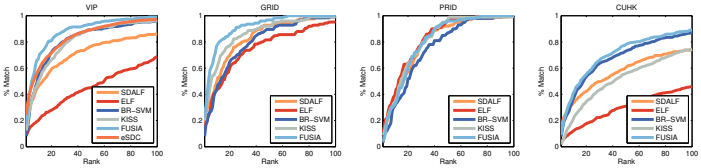


Figure 3: Overall re-identification performance of our FUSIA representation versus alternatives (CMC Curves)

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning Implicit Transfer for Person Re-identification. In *European Conference on Computer Vision, International Workshop on Re-identification*, 2012.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[3] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012.

[4] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based Re-Identification. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-Identification*. Springer London, 2013.

Evidential combination of pedestrian detectors

Philippe Xu[1]
https://www.hds.utc.fr/~xuphilip
Franck Davoine[12]
franck.davoine@gmail.com
Thierry Denœux[1]
https://www.hds.utc.fr/~tdenoeux

[1] UMR CNRS 7253, Heudiasyc
Université de Technologie de Compiègne
Compiègne, France

[2] CNRS, LIAMA
Beijing, P. R. China

The importance of pedestrian detection in many applications has led to the development of many algorithms. In this paper, we address the problem of combining the outputs of several detectors. A pre-trained pedestrian detector is seen as a black box returning a set of bounding boxes (BB) with associated scores. We conducted our experiments using the Caltech Pedestrian Detection Benchmark [2]. More than 30 state-of-the-art detectors were tested on this dataset and their outputs are publicly available.

To illustrate the potential gain from combining multiple detectors, we show in Fig. 1 (a) some detection statistics for the Caltech dataset. We can see that, at one False Positive Per Image (FPPI), more than 95% of the pedestrians in the "Reasonable" scenario were detected by at least one detector. The "Reasonable" scenario corresponds to pedestrians over 50 pixels tall and with an occlusion rate lower than 35%. As a comparison, the currently best performing algorithm has a recall rate of about 80% at 1 FPPI. Similarly, in the "Overall" scenario where all the pedestrians were considered, about 60% of the pedestrians were detected by at least one detector. The currently best algorithm hardly reached a 40% recall rate. The potential gain of combining in a proper way all those detectors is thus fairly significant.

In order to combine the outputs of different detectors, the BBs returned by the detectors need to be associated. In a sliding windows approach, a single pedestrian is often detected at several nearby positions and scales. A non-maximal suppression (NMS) step is often needed in order to select only one BB per pedestrian. In our context, the same issue occurs but, instead of having multiple detections from a single detector, they are returned by several ones. We formulated the NMS problem as a simple hierarchical clustering where the distance between two BBs is defined as their area of overlap. The clustering was done greedily by defining the distance between two clusters as the distance between their respective highest-scored BBs. This implies that the outputs from the different detectors are comparable.

To handle this issue, a calibration step was used to transform the scores into calibrated probabilities. Fig. 1 (b) illustrates the calibration results obtained from a logistic regression and an isotonic one. One particularity of object detection is the relatively high false positive rate. For example with the 'HOG' algorithm [1], more than 99% of the detections have a score less than 0.1 and less than 0.1% of these detections are true positives. As a result, most detections have an associated probability lower than 0.1. From a Bayesian perspective, multiple sources of information returning low probabilities would actually lead to an even lower one. This would go against the idea that multiple detections should lead to increased confidence.

The theory of belief function [3] was used to handle this issue. We interpreted the output $q \in [0,1]$ of a calibration function as a simple mass function $m = \{1\}^{1-q}$ defined over the frame of discernment $\Omega = \{0,1\}$ as

$$m(\{1\}) = q, \qquad m(\{0,1\}) = 1-q. \tag{1}$$

To combine two mass functions $\{1\}^{\alpha_1}$ and $\{1\}^{\alpha_2}$, three combination rules were considered in our experiments: Dempster's rule, the cautious rule and a triangular norm-based rule. They are defined, respectively, as

$$\{1\}^{\alpha_1} \oplus \{1\}^{\alpha_2} = \{1\}^{\alpha_1\alpha_2}, \tag{2}$$

$$\{1\}^{\alpha_1} \textcircled{\wedge} \{1\}^{\alpha_2} = \{1\}^{\min(\alpha_1,\alpha_2)}, \tag{3}$$

$$\{1\}^{\alpha_1} \textcircled{\top}_p \{1\}^{\alpha_2} = \{1\}^{\alpha_1 \top_p \alpha_2}, \tag{4}$$

where

$$\alpha_1 \top_p \alpha_2 = \begin{cases} \alpha_1 \wedge \alpha_2 & \text{if } p = 0, \\ \alpha_1 \alpha_2 & \text{if } p = 1, \\ \log_p\left(1 + \frac{(p^{\alpha_1}-1)(p^{\alpha_2}-1)}{p-1}\right) & \text{otherwise.} \end{cases} \tag{5}$$

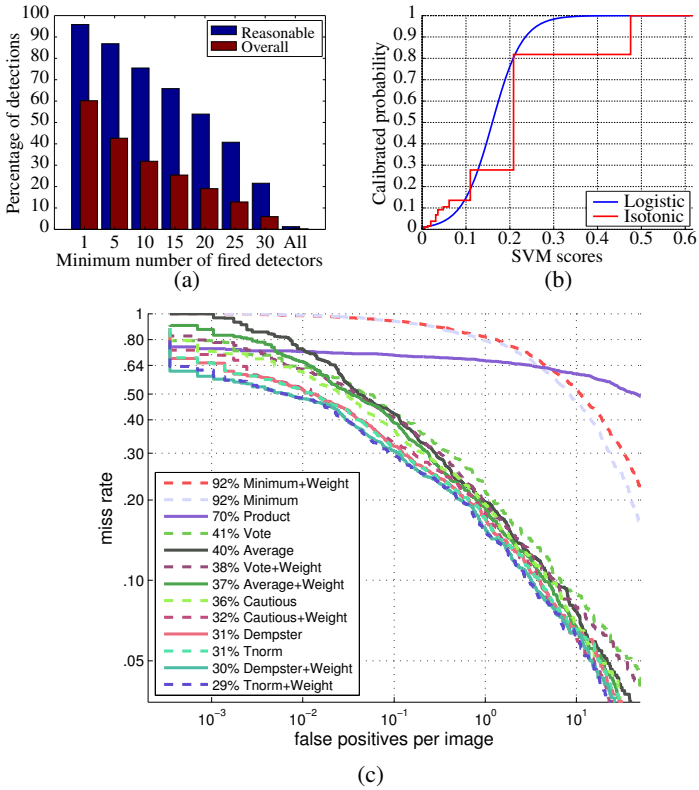


Figure 1: (a) Percentage of detected pedestrians by at least $k \in \{1,5,\ldots,34\}$ detectors at 1 FPPI. (b) Logistic and isotonic calibration of the scores from the 'HOG' pedestrian detector [1]. (c) Results of different combination strategies using a logistic regression calibration on the "Reasonable" scenario.

The triangular norm-based rule was used to better handled the dependencies among detectors. The detectors were first grouped using a hierarchical clustering and the parameter $p \in [0,1]$ of the triangular norm was then optimized for each pairwise combination.

In our experiments, we compared probabilistic combination rules (product, average, min and max) to evidential ones. Figure 1 (c) shows the results obtained from a logistic calibration on the "Reasonable" case scenario. We can see that the product and minimum rules performed very poorly. The average rule performed better than the majority vote. The cautious rule, which is equivalent to the maximum rule, performed better than all the other probabilistic rules but worse than Dempster's rule and the t-norm based rule. Using an additional weight led to better results for all combination methods except the minimum combination rule. Similar conclusions were reached by using an isotonic calibration. Compared to the best single pedestrian detector, the logistic weighted t-norm led to an improvement of 9% in terms of log-average miss rate and 6% for the isotonic one. The weighted average only led to 1% improvement. All the other probabilistic combination rules led to a decrease in performance.

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, San Diego, USA, 2005.

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[3] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

Mining Structure Fragments for Smart Bundle Adjustment

Luca Carlone[1]
luca.carlone@gatech.edu
Pablo Fernandez Alcantarilla[2]
pablo.alcantarilla@crl.toshiba.co.uk
Han-Pang Chiu[3]
han-pang.chiu@sri.com
Zsolt Kira[4]
Zsolt.Kira@gtri.gatech.edu
Frank Dellaert[1]
dellaert@cc.gatech.edu

[1] Georgia Institute of Technology, College of Computing, USA
[2] Toshiba Research Europe, Cambridge Research Laboratory, UK
[3] SRI International, Center for Vision Technologies, USA
[4] Georgia Tech Research Institute, ATAS Laboratory, USA

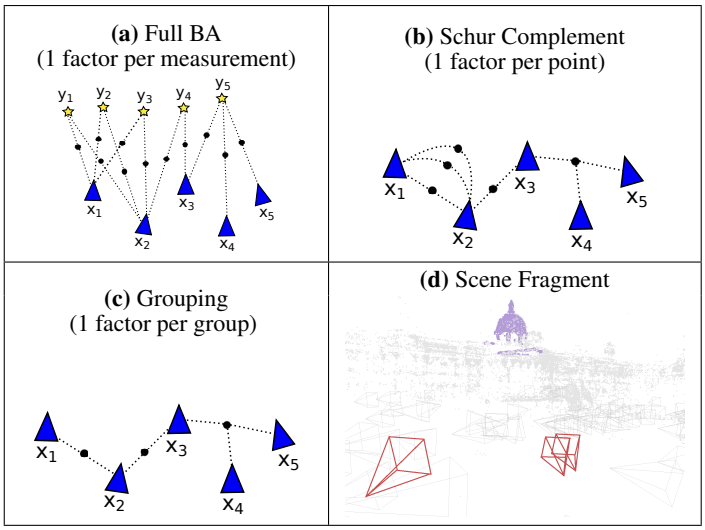


Table 1: (a) Factor graph $\mathcal{G}$ corresponding to standard BA. Points are shown as yellow stars, cameras are blue triangles, and factors are denoted with black dots. (b) Factor graph obtained after eliminating points from $\mathcal{G}$. (c) Factor graph obtained by grouping factors corresponding to points that are co-visible by the same patter of cameras. (d) A scene *fragment* is a group of N points that are visible in $M < N$ cameras. For those groups it is convenient to use an *explicit* representation for the Schur complement, as this reduces the computational cost of each conjugate gradient iteration.

Efficient bundle adjustment (BA) is an important prerequisite to a number of practical applications, ranging from 3D modeling and photo tourism, to hand-eye calibration, augmented reality, and autonomous navigation.

BA and Conjugate Gradient. BA estimates camera parameters and scene structure via nonlinear optimization. State-of-the-art approaches are based on successive linearizations: the nonlinear cost is linearized around the current estimate and a local update is computed by minimizing a quadratic approximation of the cost. Computing the local update requires solving a linear system, which is expensive for large problems.

Conjugate Gradient (CG) has been shown to be an effective linear solver for large-scale BA. The number of CG iterations can be reduced by *preconditioning*, or by using the *truncated Newton method*, which trades off accuracy of the solution for computational efficiency.

Recent work [2] provides the key insight that, in the CG method, the Schur complement trick can be applied without the explicit computation of the *reduced camera matrix*. This *implicit* representation is shown to be convenient (storage and computation-wise) with respect to *explicit* representations, in which a large (but sparse) square matrix has to be formed.

Contribution. In this paper, rather than proposing strategies to reduce the *number* of CG iterations, we propose an insight that reduces the complexity of *each* CG iteration. We adopt a factor graph perspective, and interpret BA in terms of inference over a factor graph (Table 1a). We show that the elimination of a single point induces a *factor* (i.e., a probabilistic constraint) over the cameras observing the point (Table 1b). The elimination of all points leads to the standard Schur complement, while in our approach we never build the reduced cameras system explicitly.

Reasoning in terms of factor graphs allows the solver to choose the best representation (i.e., implicit vs explicit) for each factor. A factor produced by the elimination of a single point provides a low-rank constraint on the cameras, and the use of an explicit representation is not efficient for those. However, we show that "grouping" factors corresponding to many points that are co-visible by the same set of cameras produces a single *grouped* factor, for which the explicit representation can be convenient (Table 1c): when a group of N points is visible in $M < N$ cameras, the *explicit* representation has a smaller computational cost in each conjugate gradient iteration. We call these groups of points "fragments" (Table 1d).

The grouping can be done in a grounded way: the problem is formally equivalent to well studied problems in data mining (e.g., *frequent items mining* [4]). The computational cost of grouping is negligible in BA.

In summary, our BA solver computes the fragments using data mining techniques and uses an explicit representation for the corresponding groups of factors, while the remaining factors are kept in implicit form.

Results. We tested our approach in the Bundle Adjustment in the Large benchmarking datasets [2]. Our method is implemented in C++ and released in [3]. We compare our technique against one using an implicit representation for all factors, and against a state-of-the-art solver (the *iterative Schur Solver*) available in the `Ceres` optimization suite [1].

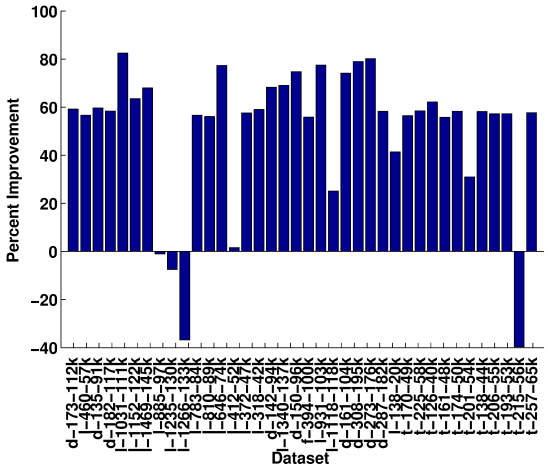


Figure 1: Total reduction in the optimization time, comparing the proposed approach against `Ceres`. The dataset *Dubrovnik* with 150 cameras and 95821 points is denoted with d-150-96k. Similar labels are used for *Ladybug*(l), *Trafalgar*(t), and *Final*(f).

Fig. 1 shows the reduction in the optimization time, comparing our approach against `Ceres`. Grouping leads to a time reduction of 50% (averaged across all datasets), with peaks reaching 80%; only in few cases `Ceres` was faster, due to early termination in CG iterations. The paper and the supplemental material include further comments and results, to show that this advantage is consistent across a large variety of tests.

[1] S. Agarwal, K. Mierle, and Others. Ceres solver, `https://code.google.com/p/ceres-solver/`.

[2] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Bundle adjustment in the large. In *European Conf. on Computer Vision (ECCV)*, pages 29–42, 2010.

[3] F. Dellaert et al. GTSAM: Georgia Tech Smoothing And Mapping, `https://borg.cc.gatech.edu`.

[4] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

Distributed Non-Convex ADMM-inference in Large-scale Random Fields

Ondrej Miksik[1]
http://www.miksik.co.uk
Vibhav Vineet[1]
vibhav.vineet@gmail.com
Patrick Pérez[2]
patrick.perez@technicolor.com
Philip H. S. Torr[1]
http://www.robots.ox.ac.uk/~tvg/

[1] Department of Engineering Science
University of Oxford
Oxford, UK
[2] Technicolor Research & Innovation
Cesson Sévigné, FR

We propose a parallel and distributed algorithm for solving discrete labeling problems in large scale random fields. Our approach is motivated by the following observations: i) very large scale image and video processing problems, such as labeling dozens of million pixels with thousands of labels, are routinely faced in many application domains; ii) the computational complexity of the current state-of-the-art inference algorithms makes them impractical to solve such large scale problems; iii) modern parallel and distributed systems provide high computation power at low cost. At the core of our algorithm is a tree-based decomposition of the original optimization problem which is solved using a non convex form of the method of alternating direction method of multipliers (ADMM). This allows efficient parallel solving of resulting sub-problems. We evaluate the efficiency and accuracy offered by our algorithm on several benchmark low-level vision problems, on both CPU and Nvidia GPU. We consistently achieve a factor of speed-up compared to dual decomposition (DD) approach and other ADMM-based approaches.

Probabilistic graphical models such as the Markov Random Fields (MRF) and Conditional Random Fields (CRF), and related energy minimization based techniques have become ubiquitous in computer vision and image processing. They have been proven especially useful to solve a variety of important, high-dimensional, discrete inference problems. Examples include per-pixel object labelling, image denoising, image inpainting, disparity and optical flow estimation, etc. [2]. Their use nonetheless implies computational costs that are often not compatible with very large scale problems met today in many applications. This concern is at the heart of present work.

We first define a discrete random field $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$ attached to the N nodes of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Each random variable takes a label from a discrete space $\mathcal{L}$ of size L. We define $\mathcal{Y} = \mathcal{L}^N$ the set of all possible label assignments. This random field is a pairwise Markov Random Field (MRF) if there exists an energy function of the form

$$E(\mathbf{Y}) := \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j), \tag{1}$$

composed of unary and pairwise potentials. Finding the lowest cost labeling of the energy over $\mathcal{Y}$ is an NP-hard combinatorial problem which can be written as the Integer Linear Program (ILP)

$$\begin{aligned} \textbf{ILP}-\textbf{MRF}: \ \text{minimize} \quad & \sum_{i \in \mathcal{V}} \theta_i \cdot p_i + \sum_{(i,j) \in \mathcal{E}} \theta_{ij} \cdot q_{ij} \\ \text{with respect to} \quad & (p,q) \in \text{Marg}(\mathcal{G}). \end{aligned} \tag{2}$$

where θ_i, θ_{ij} are vectors of unary and pairwise potentials and p, q are corresponding binary indicators.

Following [1], we split the original graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into S sub-graphs $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, $s = 1 \dots S$ and associate to each one auxiliary variables $p^s = \{p_i^s\}_{i \in \mathcal{V}_s}$ and $q^s = \{q_{ij}^s\}_{(i,j) \in \mathcal{E}_s}$, and potential parameters $\{\theta_i^s,\ i \in \mathcal{V}_s\}$ and $\{\theta_{ij}^s,\ (i,j) \in \mathcal{E}_s\}$, such that:

$$\sum_{s: i \in \mathcal{V}_s} \theta_i^s = \theta_i,\ \forall i \in \mathcal{V}; \qquad \sum_{s:(i,j) \in \mathcal{E}_s} \theta_{ij}^s = \theta_{ij},\ \forall (i,j) \in \mathcal{E}. \tag{3}$$

This implies that each node and each edge of the original graph must be covered by at least one sub-graph and that the sub-graphs can share freely nodes and edges and that the potentials on all shared vertices or edges of the sub-graphs sum to that of the original graph.

Given sub-graphs and associated parameters, we aim to replace the difficult inference problem (2) by a set of sub-problems that can be solved in parallel, while consistency between them is enforced in some way. Within the ADMM framework, there are several ways to achieve this goal. We choose to rely on "master" variables $p = \{p_i\}_{i \in \mathcal{V}}$ at the node level only. Thanks to constraints (3), it is easy to see that the original ILP-MRF problem can be written as

$$\begin{aligned} \textbf{DIP}-\textbf{MRF}: \ \text{minimize} \quad & \sum_{s=1}^{S} \left(\sum_{i \in \mathcal{V}_s} \theta_i^s \cdot p_i^s + \sum_{(i,j) \in \mathcal{E}_s} \theta_{ij}^s \cdot q_{ij}^s \right) \\ \text{with respect to} \quad & (p^s, q^s) \in \text{Marg}(\mathcal{G}_s),\ \forall s \\ & p_i \in \{0,1\}^L,\ \forall i \in \mathcal{V} \\ \text{subject to} \quad & p^s = p_{|s},\ \forall s \end{aligned} \tag{4}$$

where $p_{|s} = \{p_i\}_{i \in \mathcal{V}_s}$ denotes the sub-vector of p containing variables only for nodes of $s-$th sub-graph.

This problem can be turned into an unconstrained minimization problem by introducing the *augmented* Largrangian:

$$\begin{aligned} L_\rho(\{(p^s, q^s)\}, p, \{\lambda^s\}) = \sum_{s=1}^{S} \Big(& E_s(p^s, q^s; \theta^s) + \sum_{i \in \mathcal{V}_s} \lambda_i^s \cdot (p_i^s - p_i) + \\ & \frac{\rho}{2} \sum_{i \in \mathcal{V}_s} \|p_i^s - p_i\|_2^2 \Big) \end{aligned} \tag{5}$$

where $E_s(p^s, q^s; \theta^s) = \sum_{i \in \mathcal{V}_s} \theta_i^s \cdot p_i^s + \sum_{(i,j) \in \mathcal{E}_s} \theta_{ij}^s \cdot q_{ij}^s$, $p \in \{0,1\}^P$, $(p^s, q^s) \in \text{Marg}(\mathcal{G}_s)$ and $\lambda^s = \{\lambda_i^s\}_{i \in \mathcal{V}_s} \in R^{L \times |\mathcal{V}_s|}$. This is a consensus problem in that we essentially have multiple copies of the same variable that should take the value of the master.

Vector λ is the dual variable as in classic Lagrangian duality and ρ is a positive parameter. While the additional penalty destroys the separability as compared to classic Lagrangian, it helps solving dual problem efficiently. The ADMM approach conducts the joint optimization of augmented Lagrangian by alternating the following three steps:

$$(p^s, q^s)^{(t+1)} := \underset{(p^s, q^s) \in \text{Marg}(\mathcal{G}_s)}{\text{argmin}} \ L_\rho(\{(p^s, q^s)\}, p^{(t)}, \{\lambda^{s(t)}\}), \quad \forall s. \tag{6}$$

$$p_i^{(t+1)} := \mathcal{P}_{\text{Marg}(\mathcal{G})} \left(\frac{1}{|\mathcal{I}(i)|} \sum_{s \in \mathcal{I}(i)} \left(p_i^{s(t+1)} + \frac{1}{\rho} \lambda_i^{s(t)} \right) \right), \qquad \forall i \in \mathcal{V}, \tag{7}$$

$$\lambda_i^{s(t+1)} := \lambda_i^{s(t)} + \rho \left(p_i^{s(t+1)} - p_i^{(t+1)} \right), \quad \forall s,\ \forall i \in \mathcal{V}_s. \tag{8}$$

In this paper, we show how to solve such problem efficiently on a modern GPU. Our approach is easy to implement, since each sub-problem requires one call to a dynamic programming solver, and is highly suitable for modern GPUs with thousands of CUDA cores. Finally, we show empirically that our approach rapidly converges to a good quality estimates and is able to return a solution at any point in practice, which is important when developing interactive systems.

[1] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):531–552, 2011.

[2] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappena, and C. Rother. A comparative study of energy minimization methods for markov random fields. *TPAMI*, 30 (6), 2007.

Boosted Cross-Domain Categorization

Fan Zhu[1]
fan.zhu@sheffield.ac.uk
Ling Shao[1]
ling.shao@ieee.org
Jun Tang[2]
tangjunahu@gmail.com

[1] Departmetn of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD, UK

[2] School of Electronics and Information Engineering
Anhui University
Hefei, 230601, China

We introduce a boosted cross-domain categorization (BCDC) framework that utilizes labeled data from other domains as the source data to span the intra-class diversity of the original learning system. In addition to the manually annotated information in the target domain, partially labeled data from another visual domain are provided as the source domain. In comparison, the proposed learning framework shares the same basic principle of sequentially updating the impacts of training instances; yet our learning framework attempts to sequentially update the data representations of those "dis-similar" samples instead of simply weighting less on them. The learning function is formulated as:

$$\begin{aligned}&\langle D_t, D_s, X_t, \Phi, \mathcal{P}\rangle \\ &\quad = arg \min_{D_t, D_s, X_t, \Phi, \mathcal{P}} \|Y_t - D_t X_t\|_2^2 \\ &\quad +\alpha\|Q - \Phi X_t\|_2^2 + \beta\|\mathcal{H} - \mathcal{P}X_t\|_2^2 \\ &\quad +\|Y_s\mathbb{A}^T - D_s X_t\|_2^2 \quad s.t. \forall i,\ \|x_t^i\|_0 \leq T.\end{aligned} \tag{1}$$

In order to distinguish the "dissimilar" data from the smooth data, we include the weighted discriminative sparse codes into the learning function. Specifically, $q_i = [q_i^1, q_i^2, \cdots, q_i^K]^T = [0, \cdots, w_i, w_i, \cdots, 0]^T \in \Re^K$, where the non-zeros occur at those indices where $y_t^i \in Y_t$ and $X_t^k \in X_t$ share the same class label. Given $X_t = [x_1, x_2, \cdots, x_6]$ and $Y_t = [y_1, y_2, \cdots, y_6]$, and assuming x_1, x_2, y_1 and y_2 are from class 1, x_3, x_4, y_3 and y_4 are from class 2, x_5, x_6, y_5 and y_6 are from class 3, Q is then defined with the following form:

$$\begin{pmatrix} w_1 & w_2 & 0 & 0 & 0 & 0 \\ w_1 & w_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & w_3 & w_4 & 0 & 0 \\ 0 & 0 & w_3 & w_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_5 & w_6 \\ 0 & 0 & 0 & 0 & w_5 & w_6 \end{pmatrix}, \tag{2}$$

Since predictions are made with respect to the data distribution of X_t, w_i is included in each item of $\mathcal{H}$. Thus $\mathcal{H}$ can be defined as follows according to the same example in Equation (2)

$$\begin{pmatrix} w_1 & w_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & w_3 & w_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_5 & w_6 \end{pmatrix}. \tag{3}$$

By defining $Y = (Y_t^T, (Y_s\mathbb{A}^T)^T, \sqrt{\alpha}Q^T, \sqrt{\beta}\mathcal{H}^T)^T$ and $D = D_t^T, D_s^T, \sqrt{(\alpha)}\Phi^T, \sqrt{(\beta)}\mathcal{P}^T)^T$, where column-wise L_2 normalization is applied to D, the objective function in equation 1 can be solved through sequentially updating dictionary atoms and sparse codes as in [8].

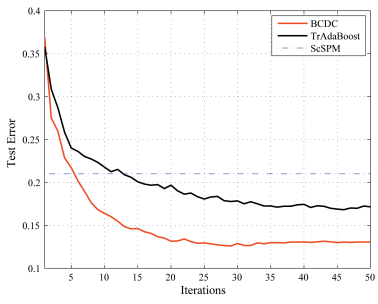


Figure 1: Error rate comparison of the proposed BCDC method with TrAdBoost and ScSPM on the Caltech-101 dataset.

Algorithm 1 Boosted cross-domain dictionary learning

Input the labeled target domain data $\mathcal{D}_l^t$ and the source domain data $\hat{\mathcal{D}}^s$, the maximum number of iterations $Max.iter$ and the Weak Learner.
Output a "strong" classifier $\mathcal{F}(\cdot)$ and updated representations of the source domain instances.
Initialize the data distribution as uniform, i.e., the initial weights $w^1 = (w_1^1, w_2^1, \cdots, w_{N+M}^1)$ have an identical value. Cross-domain discriminative dictionary learning is applied to both target domain and source domain data under the initialized uniform distribution, so that $\mathcal{D}_l^t$ and $\hat{\mathcal{D}}^s$ can be represented by X_t and X_s^1 respectively.
for $j = 1$ to $Max.iter$ **do**

1. Set data distribution $p^j = \frac{w^j}{\sum_{i=1}^{N+M} w_i^j}$
2. Update X_s^j as the new representation of $\hat{\mathcal{D}}^s$ under data distribution p^j with cross-domain discriminative dictionary learning.
3. Compute the hypothesis $h_t^j : X_t \to l(X_t)$ and $h_s^j : X_s^t \to l(X_s)$, providing that p^j is over both $\mathcal{D}_l^t$ and $\hat{\mathcal{D}}^s$.
4. Calculate the error ε^j of h_t^j:
$$\varepsilon^j = \sum_{i=1}^{N} \frac{w_i^j \times |h_t^j(x_i) - l(x_i)|}{\sum_i^N w_i^j},$$
where ε^j is required to be less than 0.5.
5. Set $\beta_t^j = \frac{\varepsilon^j}{1-\varepsilon^j}$ and $\beta_s = \frac{1}{1+\sqrt{2\ln M / Max.iter}}$
6. Update the new weight vector:
$$w_i^{j+1} = \begin{cases} w_i^j \beta_t^{j^{-|h_t^j(x_i) - l(x_i)|}}, & 1 \leqslant i \leqslant N \\ w_i^j \beta_s^{|h_s^j(x_i) - l(x_i)|}, & otherwise. \end{cases}$$

end for

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(1):4311–4322, 2006.

[2] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[3] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, 2008.

[4] Julien Mairal, Marius Leordeanu, Francis Bach, Martial Hebert, and Jean Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *ECCV*, 2008.

[5] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *NIPS*, 2009.

[6] Jianchao Yang, Kai Yu, and Thomas Huang. Supervised translation-invariant sparse coding. In *CVPR*, 2010.

[7] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, 2010.

[8] Fan Zhu and Ling Shao. Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1-2):42–59, 2014.

Return of the Devil in the Details: Delving Deep into Convolutional Nets

Ken Chatfield
ken@robots.ox.ac.uk
Karen Simonyan
karen@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
Oxford, UK

The latest generation of Convolutional Neural Networks (CNN) have been shown to achieve impressive results in challenging benchmarks on image recognition and object detection, significantly raising the interest of the community in these methods. Nevertheless, it is still unclear how different CNN methods compare with each other and with previous state-of-the-art shallow representations such as the Bag-of-Visual-Words and the Improved Fisher Vector (IFV). This paper conducts a rigorous evaluation of these new techniques, exploring different deep architectures and comparing them on a common ground, identifying and disclosing important implementation details in a similar vein to our previous work on shallow encoding methods [1].

We identify several useful properties of CNN-based representations, including the fact that the dimensionality of the CNN output layer can be reduced significantly without having an adverse effect on performance. We also identify aspects of deep and shallow methods that can be successfully shared. In particular, we show that the data augmentation techniques commonly applied to CNN-based methods can also be applied to shallow methods, and result in an analogous performance boost.

Evaluation over multiple standard benchmark datasets is presented (PASCAL VOC 2007 and 2012, Caltech-101, Caltech-256 and ILSVRC-2012) and our best CNN-based method achieves performance comparable to state-of-the-art over all four (refer to Table 1). We also present a variety of other configurations, each striking a different trade-off in the balance between performance, computation speed and compactness.

As with our previous work, source code and CNN models to reproduce the experiments presented in the paper are available from the project webpage[1] in the hope that it would provide common ground for future comparisons, and good baselines for image representation research.

1 CNN-based Methods

Our **Fast (CNN-F)** method provides the fastest computation time, and is similar in architecture to the one used by Krizhevsky *et al*. [3], our **Medium (CNN-M)** method strikes balance between being relatively fast to compute and greater performance, being loosely based on the architecture of Zeiler and Fergus [7]. Finally, our **Slow (CNN-S)** method focuses on maximum performance, and is similar architecturally to the 'accurate' network from the OverFeat package [6]. We further investigate the impact of: (a) different data augmentation strategies, (b) reducing the output dimensionality of the output layer and (c) the performance boost (if any) possible by fine-tuning the networks to the target dataset.

2 Compared to Shallow Methods

By applying data augmentation techniques similar to with CNN-based methods to IFV, we obtain a performance boost to 68.0% on the PASCAL VOC 2007 benchmark. We further investigate the impact of: (a) different IFV normalisation and spatial information encoding strategies, (b) adding colour information to shallow features, or removing it from CNN-based methods and (c) combining IFV with CNN-based methods into a single fused representation.

3 Performace Evolution over PASCAL VOC 2007

A comparative plot of the evolution in the performance of the methods evaluated in this paper, along with a selection from our earlier review of shallow methods [1] is presented in Fig. 1. Classification accuracy over PASCAL VOC was 54.48% mAP for the BoVW model in 2008, 61.7% for the IFV in 2010 [1], and 73.41% for DeCAF [2] and similar [4, 5] CNN-based methods introduced in late 2013. Our best performing CNN-based method (CNN-S with fine-tuning) achieves 82.42%, comparable to the most recent state-of-the-art.

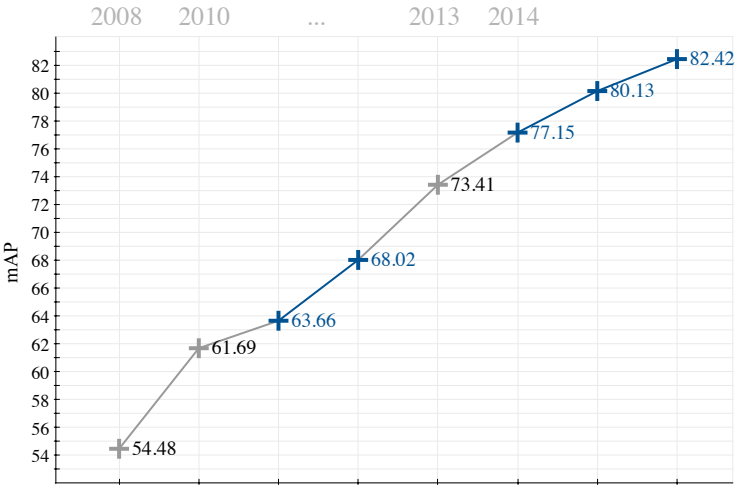


Figure 1: **Evolution of Performance on PASCAL VOC-2007 over the recent years.** Refer to Table 2 in the paper for details and references.

	ILSVRC-2012 (top-5 error)	VOC-2007 (mAP)	VOC-2012 (mAP)	Caltech-101 (accuracy)	Caltech-256 (accuracy)
FK IN	-	65.4	–	–	–
FK IN +aug	-	68.0	–	–	–
CNN F	16.7	77.4	79.9	–	–
CNN M	13.7	79.9	82.5	87.15 ± 0.80	77.03 ± 0.46
CNN S	**13.1**	79.7	82.9	87.76 ± 0.66	**77.61** ± **0.12**
CNN S TN	**13.1**	**82.4**	**83.2**	**88.35** ± **0.56**	77.33 ± 0.56

Table 1: **Sample of key results from the paper** on ILSVRC2012, VOC2007, VOC2012, Caltech-101, and Caltech-256. 'TN' – dataset-specific fine-tuning. For IFV, '+aug' indicates full data-augmentation.

[1] 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.

[2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[4] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proc. CVPR*, 2014.

[5] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *CoRR*, abs/1403.6382, 2014.

[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *Proc. ICLR*, 2014.

[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[1] http://www.robots.ox.ac.uk/~vgg/research/deep_eval/

Transductive Multi-label Zero-shot Learning

Yanwei Fu, Yongxin Yang
{y.fu,yongxin.yang}@qmul.ac.uk
Timothy Hospedales, Tao Xiang
{t.hospedales,t.xiang}@qmul.ac.uk
Shaogang Gong
{s.gong}@qmul.ac.uk

School of EECS
Queen Mary University of London
London, E1 4NS, UK

Zero-shot learning has received increasing interest as a means to alleviate the prohibitive expense of annotating training data for large scale recognition problems. These methods have achieved great success via learning intermediate semantic representations in the form of attributes and more recently, semantic word vectors. However, many real-world data are intrinsically multi-label. For example, an image on Flickr often contains multiple objects with cluttered background, thus requiring more than one label to describe its content. And different labels are often correlated (e.g. cows often appear on grass). In order to better predict these labels given an image, the label correlation must be modelled: for n labels, there are 2^n possible multi-label combinations and to collect sufficient training samples for each combination to learn the correlations of labels is infeasible.

It is thus surprising to note that there is little if any existing work for general multi-label zero-shot learning. Is it because there is a trivial extension of existing single label ZSL approaches to this new problem? By assuming each label is independent from one another, it is indeed possible to decompose a multi-label ZSL problem into multiple single label ZSL problems and solve them using existing single label ZSL methods. However this does not exploit label correlation, and we demonstrate in this work that this naive extension leads to very poor label prediction for unseen classes. Any attempt to model this correlation, in particular for the unseen classes with zero examples, is extremely challenging.

Multi-Label Zero-Shot Framework In this paper, we propose a novel framework for multi-label zero-shot learning. Given a training/auxiliary dataset containing labelled images, and a test/target dataset with a set of unseen labels/classes (i.e. none of the labels appear in the training set), we aim to learn a multi-label classification model from the training set and generalise/transfer it to the test set with unseen labels. This knowledge transfer is achieved using an intermediate semantic representation in the form of the skip-gram word vectors [3] which allows vector-oriented reasoning. Such a reasoning is critical for our zero-shot multi-label prediction to synthesise label combination prototypes in the semantic word space. For example, $Vec(`Moscow')$ should be much closer to $Vec(`Russia') + Vec(`capital')$ than $Vec(`Russia')$ or $Vec(`capital')$ only. For this purpose, we employ the skip-gram language model to learn the word space, which has shown to be able to capture such syntactic regularities. This representation is shared between the training and test classes, thus making the transfer possible.

Our framework has two main components: multi-output deep regression (Mul-DR) and zero-shot multi-label prediction (ZS-MLP). Mul-DR is a 9 layer neural network that exploits convolutional neural network (CNN) layers, and includes two multi-output regression layers as the final layers. It learns from auxiliary data the mapping from raw image pixels to a linguistic representation defined by the skip-gram language model [3]. With Mul-DR, each test image is now projected into the semantic word space where the unseen labels and their combinations can be represented as data points without the need to collect any visual data. ZS-MLP addresses the multi-label ZSL problem in this semantic word space by exploiting the property that label combinations can be synthesised. We exhaustively synthesise the power set of all possible prototypes (i.e., combinations of multi-labels) to be treated as if they were a set of labelled instances in the space. With this synthetic dataset, we are able to propose two new multi-label algorithms – direct multi-label zero-shot prediction (DMP) and transductive multi-label zero-shot prediction (TraMP). However, since Mul-DR is learned using the auxiliary classes/labels, it may not generalise well to the unseen classes/labels. To overcome this problem, we further exploit self-training to adapt Mul-DR to the test classes to improve its generalisation capability.

Experiments We evaluate our framework with the widely used Natural Scene and IAPRTC-12 multi-label datasets. **Natural Scene** consists of 2000 natural scene images where each is labelled as any combinations of desert, mountains, sea, sunset and trees. We use a multi-class single label dataset – Scene dataset (2688 images) as the auxiliary dataset which are labelled with a non-overlapping set of labels such as street, coast and highway. **IAPRTC-12** consists of 20000 images and a total of 275 different labels. Our experiments consider the subset of landscape-nature branch (around 9500 images) and use the top 8 most frequent labels from this branch with over 30% of multi-label test images. For this dataset, we employ both Scene and Natural Scene as the auxiliary dataset.

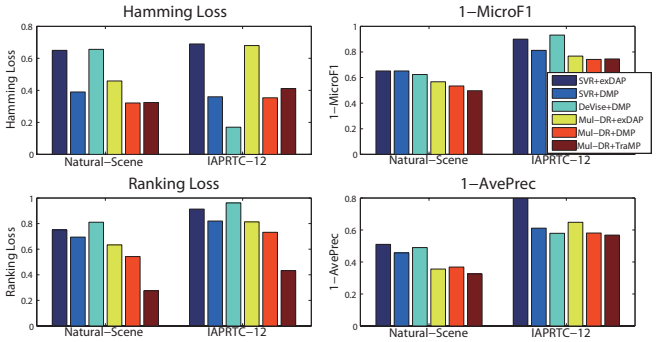


Figure 1: Comparing different zero-shot multi-label classification methods on Natural Scene and IAPRTC-12.So smaller values for all metrics are preferred.

Groundtruth	sand-beach, mountain,sky	landscape-nature, mountain, sky	grass	sand-beach, sky
Mul-DR+DMP	sand-beach, sky	landscape-nature, mountain, sky	grass	sand-beach, sky
Mul-DR+TraMP	sand-beach, mountain, sky	landscape-nature, mountain, sky	grass, ground, landscape-nature	ground, sky, sand-beach
DeViSE+DMP	sky	–	–	sky

Table 1: Examples of multi-label zero-shot predictions on IAPRTC-12. Top 8 most frequent labels of landscape-nature branch are considered.

The results in Fig 1 and Tab 1 show the efficacy of our framework for multi-label ZSL over a variety of baselines: (1) Comparing regression models: Our Mul-DR significantly improves the results compared to both conventional SVR [2] regression (Mul-DR+DMP>SVR+DMP, Mul-DR+exDAP>SVR+exDAP as well as DeViSE [1] (Mul-DR+DMP vs. DeViSE+DMP). (2) Comparing multi-label annotation strategy with the same regression model: Our transductive multi label approach outperforms the generalisation of the conventional DAP [2] to the multi-label setting (Mul-DR+DMP>Mul- DR+exDAP). For more detailed discussion, please read our paper. All the data/codes can be downloaded from
`http://www.eecs.qmul.ac.uk/~yf300/multilabelZSL/`.

[1] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[2] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

Optimal Representation of Multi-View Video

Marco Volino
m.volino@surrey.ac.uk
Dan Casas
d.casas@surrey.ac.uk
John Collomosse
j.collomosse@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk

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

Introduction: Multi-view video acquisition is widely used for reconstruction and free-viewpoint rendering (FVR) of dynamic scenes. Current approaches to FVR resample directly from the captured multi-view images at each time frame, achieving a high level of photo-realism but requiring storage and transmission of multi-video sequences. This is prohibitively expensive in both storage and bandwidth required for multiple video streams limiting applications to local rendering on high-performance hardware. This paper addresses the problem of optimally resampling and representing multi-view video to obtain a compact representation without loss of the view-dependent dynamic surface appearance.

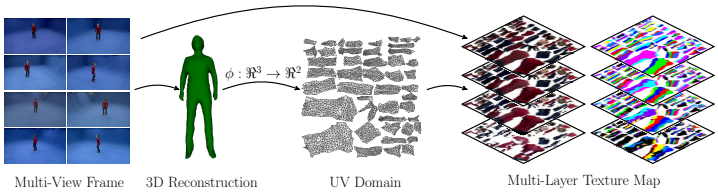


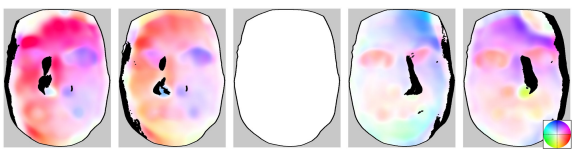
Figure 1: Overview of the resampling of multi-view video to a multi-layer texture video

Representation and Optimisation: Fig. 1 shows an overview of the proposed approach taking as input a set of camera images and an aligned mesh sequence. Texture coordinates, a 3D-2D mapping, are defined and the multi-view images are resampled into a hierarchy of texture maps with the views of each facet ordered by visibility. Optimal resampling from multiple views requires spatial and temporal coherence of the representation. The problem can be cast as a labelling problem where we seek the mapping $L : F \rightarrow C$ from the set of mesh facets F to the set of cameras $C = \{1...N_C\}$ which assigns a camera label $l_f \in C$ to each facet $f \in F$. We formulate the computation of the optimal labelling $L(t)$ as an energy minimisation of cost:

$$E(L(t)) = \sum_{\forall t}(E_v(L(t)) + \lambda_s E_s(L(t)) + \lambda_t E_t(L(t), L(t+1))). \quad (1)$$

where $E_v(L(t))$ is the unary visibility cost for all faces F to be assigned camera labels $L(t)$ at time t, $E_s()$ is the spatial coherence cost which enforces consistent camera labelling between adjacent mesh facets, $E_t()$ is the temporal coherence cost which enforces temporal coherence of the camera labelling, finally λ_s and λ_t are weighting terms for the spatial and temporal smoothness functions.

(a) Textured geometry C_i $i = 1...5$ (b) Rendered Image I_3^j for $j = 1...5$)



(c) Optical flow $O_{i\rightarrow j}$ between rendered image I_3 and I_3^j for $j = 1..5$. Optical flow colour mapped to direction (bottom right, maximum flow 5 pixels)

Figure 2: Results of surface-based optical flow alignment of appearance from multiple views

Multi-View Alignment: Simple projection and blending of camera views using the approximate reconstructed mesh geometry leads to blurring and ghosting artefacts. These artefacts are caused by misalignment between overlapping camera images projected onto to mesh surface from inaccurate geometry and camera calibration. In order to minimise these artefacts, we use optical flow based image warping to correct misalignments before sampling into the texture domain. To establish optical flow between camera views, we first render the geometry from the viewpoint of camera C_i and projectively texture the geometry using the image of camera C_j for all N_C cameras. This results in N_C^2 rendered images, R_i^j, which denotes the image rendered from the i^{th} camera viewpoint using the j^{th} camera image, Fig. 2(a) and (b). An optical flow correspondence field, $O_{i\rightarrow j}$, is computed between the rendered image $R_i = R_i^i$ and R_i^j where $i \neq j$. A binary confidence score is assigned to each flow vector, black indicates areas where occlusion or depth discontinuities occur these are assigned a zero confidence scores, Fig. 2(c). The magnitude of the correction vector is given by the weighted average of all visible and high-confidence flow vectors on the surface.

Results: Optimal resampling of the captured multi-view images as a layered texture map representation is achieved by combining the optical flow alignment of the captured images on the reconstructed surface with the spatio-temporal optimisation of camera label assignments for each mesh facet. Fig. 3 shows two examples of the multi-view alignment: (a) a texture map layer from dataset Dan. (b) First three layers from Cloth dataset blended together. This demonstrates that the approach corrects misalignment which reduces ghosting and blur artefacts during rendering. The representation is evaluated in terms of rendering quality and required storage when varying the size of the texture map and number used. We show that only 3 texture layers are required to maintain view dependence during rendering and no significant increase in quality occurs when using a texture size above 1024. This results in a >90% reduction in the required storage when compared to the captured data.

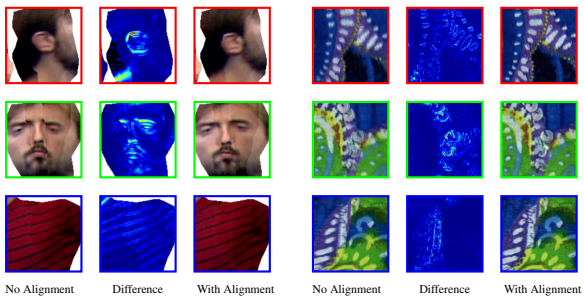


Figure 3: Results of multi-view optical flow based alignment

Conclusion: A method is presented for optimisation of the resampling from multi-view video sequences of a reconstructed surface into a multi-layer 2D texture map representation to obtain a compact, spatially and temporal coherent representation that minimises the loss of information from the captured data to maintain FVR quality. Spatio-temporal optimisation is combined with a surface-based optical flow alignment to significantly reduce the storage footprint and minimise artefacts due to errors in geometry and camera calibration. This demonstrates that the proposed approach results in an efficient representation that preserves the visual quality of the captured multiple view video for FVR whilst achieving approximately >90% reduction in size.

Unsupervised Spatio-Temporal Segmentation with Sparse Spectral Clustering

Mahsa Ghafarianzadeh[1]
masa@gwu.edu
Matthew B. Blaschko[2]
matthew.blaschko@inria.fr
Gabe Sibley[1]
gsibley@gwu.edu

[1] Computer Science Department
The George Washington University
Washington DC, USA

[2] École Centrale Paris
INRIA Saclay
Châtenay-Malabry, France

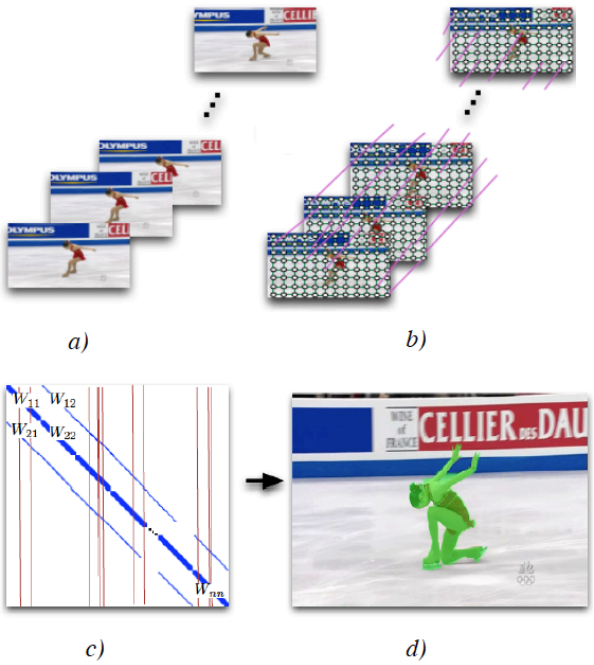


Figure 1: Approach overview: given a sequence of images a) we construct a graph b) connecting pixels in a neighborhood and also to their temporal correspondences. This is represented as a very large sparse matrix, from which we select a random subset of columns (which correspond to pixels) – only the randomly selected pixels are used for graph and matrix construction. d) we employ spectral clustering segmentation based on efficient and accurate low rank factorization based on the Nyström method to approximate the graph Laplacian.

Spatio-temporal cues are powerful sources of information for segmentation in videos. In this work we present an efficient and simple technique for spatio-temporal segmentation that is based on a low-rank spectral clustering algorithm. The complexity of graph-based spatio-temporal segmentation is dominated by the size of the graph, which is proportional to the number of pixels in a video sequence. In contrast to other works, we avoid oversegmenting the images into super-pixels and instead generalize a simple graph based image segmentation. Our graph construction encodes appearance and motion information with temporal links based on optical flow. For large scale data sets naïve graph construction is computationally and memory intensive, and has only been achieved previously using a high power compute cluster. We make feasible for the first time large scale graph-based spatio-temporal segmentation on a *single core* by exploiting the sparsity structure of the problem and a low rank factorization that has strong approximation guarantees.

The central contribution of this paper is to introduce a set of strategies that enable us to compute a dense graph based segmentation using a single processor and fitting in core memory. This is done primarily through two innovations: (i) we exploit the sparsity structure of the spatio-temporal graph, and (ii) we make use of an efficient and accurate low rank factorization based on the Nyström method to approximate the graph Laplacian in a spectral clustering approach. Our results show that not all of the pixels contain meaningful information about images, and just a subset of pixels can be a good representation of the entire scene.

Given an image I, we create a graph $G = (V, E, W)$, where the graph nodes V are the pixels in the image and are connected by edge E if they are within distance r from each other. W measures the similarity of pixels connected by an edge. We define W as the following: $W_{ij} = \exp \frac{-d^2(s_i,s_j)}{\sigma_i \sigma_j}$ where $W_{ij} = 0$ for $i =$j, and s_i denotes pixel color and $d(s_i,s_j)$ is the Euclidean distance. σ is a local scaling parameter [3] which takes into account the local statistics of the neighborhood around pixels i and j. Local scaling parameter is defined by: $\sigma_i = d(s_i,s_k)$ where s_k is the K'th neighbor of pixel i. In order to extend this to video, we make use of optical flow and add temporal motion information to the graph. We use optical flow to compute the motion vectors between frames. Then we connect pixel (x,y) in frame t to its 9 neighbors along the backward flow (u,v) in frame $t-1$, *e.g.* $(x+u(x,y)+\delta_x, y+v(x,y)+\delta_y)$ for $\delta_x, \delta_y \in \{-1,0,1\}$.

The similarity matrix for the video is a sparse symmetric block diagonal matrix of the size $n = number\ of\ frames \times number\ of\ pixels\ in\ one\ frame$.

Next, we use a time and space efficient spectral clustering via column sampling [1], that is similar to Nyström method, but with a further rank-k approximation of the normalized Laplacian using the sampled sub-matrix of the similarity matrix. This algorithm has shown promising results, since it reduces the time and space complexity of Nyström method and also it is able to recover all the degree information of the selected points. The time complexity of the algorithm is $O(nmk)$ and there is no need to store large similarity matrix W or its sampled columns in the memory. Also we are using the proposed inexpensive algorithm by [1] to orthogonalized estimated eigenvectors.

After performing spectral clustering on the similarity graph and obtaining the clusters, we first quantize each cluster into 256 bins (16 bins for each channel) and compute the *RGB* histogram. Then we merge adjacent clusters repeatedly if their similarity is more than a threshold τ to achieve the final segmentation.

We compared our method against other dense and sparse methods and achieved comparable performance while using just a subset of pixels (30%-50%) to label all of the pixels. In conclusion, we have demonstrated a novel method for spatio-temporal segmentation of dense pixel trajectories based on spectral clustering. We found that fully connecting pixels to their spatial neighbors within a given radius is an effective strategy for improving segmentation accuracy. Additionally, we use optical flow to more accurately compute temporal connectivity than a simple method based on an interpretation of the video sequence as a 3D volume. In contrast to previous work, we do not resort to super-pixel segmentation to achieve computational tractability and memory efficiency. Instead, we exploit the natural sparsity structure of the graph, and employ a low rank approximation of the Laplacian closely related to the Nyström method. We have found that sampling 30-50% of the pixels to index columns of the low rank approximation leads to comparable performance with a method that uses 100% of the columns. This strategy results in a spectral clustering method that can run on a single processor with the graph representation fitting in core memory. We have demonstrated the effectiveness of the approach on the Hopkins 155 data set, where we have achieved the best reported results for dense segmentation using an order of magnitude less computation than [2].

[1] Mu Li, Xiao-Chen Lian, James T Kwok, and Bao-Liang Lu. Time and space efficient spectral clustering via column sampling. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2297–2304. IEEE, 2011.

[2] Narayanan Sundaram and Kurt Keutzer. Long term video segmentation through pixel level spectral clustering on gpus. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 475–482. IEEE, 2011.

[3] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NIPS*, volume 17, page 16, 2004.

Depth Extraction from Videos Using Geometric Context and Occlusion Boundaries

S. Hussain Raza[1]
hussain.raza@gatech.edu
Omar Javed[2]
omar.javed@sri.com
Aveek Das[2]
aveek.das@sri.com
Harpreet Sawhney[2]
harpreet.sawhney@sri.com
Hui Cheng[2]
hui.cheng@sri.com
Irfan Essa[3]
irfan@cc.gatech.edu

[1] School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA

[2] SRI International
Princeton, New Jersey, USA

[3] School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA

http://www.cc.gatech.edu/cpl/projects/videodepth

Abstract

We present an algorithm to estimate depth in dynamic video scenes. We propose to learn and infer depth in videos from appearance, motion, occlusion boundaries, and geometric context of the scene. Using our method, depth can be estimated from unconstrained videos with no requirement of camera pose estimation, and with significant background/foreground motions. We start by decomposing a video into spatio-temporal regions. For each spatio-temporal region, we learn the relationship of depth to visual appearance, motion, and geometric classes. Then we infer the depth information of new scenes using piecewise planar parametrization estimated within a Markov random field (MRF) framework by combining appearance to depth learned mappings and occlusion boundary guided smoothness constraints. Subsequently, we perform temporal smoothing to obtain temporally consistent depth maps. We present a thorough evaluation of our algorithm on our new dataset and the publicly available Make3d static image dataset.

1 Introduction and Approach

Methods exploiting visual and contextual cues for depth can be used to provide an additional source of depth information to the structure from motion or multi-view stereo based depth estimation systems. In this paper, we focus on texture features, geometric context, motion boundary based monocular cues along with co-planarity, connectivity and spatio-temporal consistency constraints to predict depth in videos. We assume that a scene can be decomposed into planes, each with its own planar parameters. We over-segment a video into spatio-temporal regions and compute depth cues from each region along with scene structure from geometric contexts. These depth cues are used to train and predict depth from features. However, such appearance to depth mappings are typically noisy and ambiguous. We incorporate the independent features to depth mapping of each spatio-temporal region within in a MRF framework that encodes constraints from scene layout properties of co-planarity, connectivity and occlusions. To model the connectivity and co-planarity in a scene, we explicitly learn occlusion boundaries in videos. To further remove the inconsistencies from temporal depth prediction, we apply a sliding window to smooth the depth prediction. Our approach doesn't require camera translation or large rigid scene for depth estimation. Moreover, it provides a source of depth information that is largely complementary to triangulation based depth estimation methods [5]. **The primary contributions of our method to extract depth from videos are:**

- Adoption of a learning and inference approach that explicitly models appearance to geometry mappings and piecewise scene smoothness;
- Learning and estimating occlusion boundaries in videos and utilizing these to constrain smoothness across the scene;
- There is no requirement of a translating camera or a wide-baseline for depth estimation;
- An algorithm for video depth estimation that is complementary to traditional structure from motion approaches, and that can incorporate these approaches to compute depth estimates for natural scenes;

2 Experiments and Results

We perform extensive experiments on video depth data to evaluate our algorithm. We perform 5-fold cross-validation over 36 videos ($\sim$ 6400 frames). We compute average log-error $|\log d - \log \hat{d}|$ and average relative error $|\frac{d-\hat{d}}{d}|$ to report the accuracy of our method. We achieve an accuracy of 0.153 log-error and 0.44 on relative error (Table 1). Figure 1 shows some example scenes from our dataset with ground truth and predicted depth. Our approach for depth estimation can also be applied to images. We applied our algorithm over a publicly available Make3d depth image dataset [6]. Table 2 gives the comparison of the single image variant of our approach with the state of the art and we achieve competitive results. It should be noted that our algorithm depends on occlusion boundary detection and geometric context (for which motion based features are important and is not optimized to extract depth from single images.

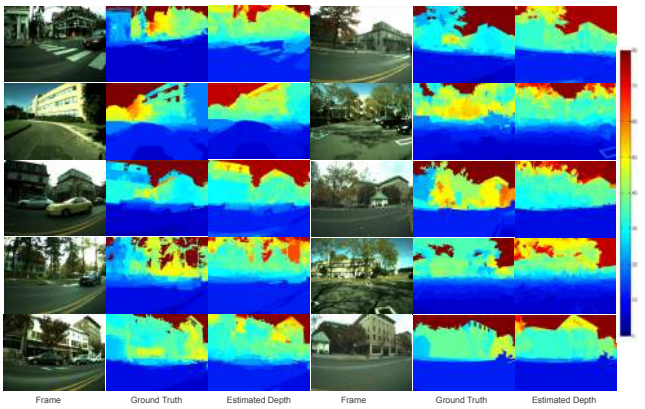


Figure 1: Examples of videos scenes, ground truth, and predicted depth by our method. Legend shows depth range from 0m (blue) to 80m (red).

Features	$log10$	rel-depth
ALL	0.153	0.44
App.+Flow	0.176	0.533
Appearance	0.175	0.512

Table 1: Performance of our algorithm on video dataset, combining appearance, flow, and surface layout features give best accuracy.

Algorithm	$log10$	rel-log
SCN [4]	0.198	0.530
HEH [1]	0.320	1.423
Baseline [6]	0.334	0.516
PP-MRF [6]	0.187	0.370
Depth Transfer [2]	0.148	0.362
Sematic Labels [3]	0.148	0.379
***Geom. Context Occl. Bound.**	**0.159**	**0.386**

Table 2: Our approach can also be applied to images. We apply it to Make3d depth image dataset [6].

[1] D. Hoiem, A.A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007.

[2] Kevin Karsch, Ce Liu, and Sing Kang. Depth extraction from video using non-parametric sampling. In *ECCV 2012*.

[3] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010.

[4] Ashutosh Saxena, Sung H Chung, and Andrew Ng. Learning depth from single monocular images. In *NIPS*, 2005.

[5] Ashutosh Saxena, Jamie Schulte, and Andrew Y Ng. Depth estimation using monocular and stereo cues. In *IJCAI*, 2007.

[6] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009.

Non-Rigid Shape-from-Motion for Isometric Surfaces using Infinitesimal Planarity

Ajad Chhatkuli

Daniel Pizarro

Adrien Bartoli
http://isit.u-clermont1.fr/~ab

ALCoV-ISIT, UMR 6284 CNRS / Université d'Auvergne, Clermont-Ferrand, France

Introduction. Non-Rigid Shape-from-Motion (NRSfM) is the general solution to the 3D reconstruction from multiple monocular images of deforming objects. Most previous attempts in NRSfM have been on learning a low dimensional shape basis from a set of contiguous images. NRSfM is very much related to the Shape-from-Template (SfT) problem, where shape is computed from a known 3D template and a single input image after deformation. Most SfT methods have been based on isometric deformations [1, 2]. Thus applying NRSfM in isometrically constrained deformations is a natural way forward. However, there has been a gap in the literature regarding the theory behind isometric NRSfM. Many of the isometric NRSfM solutions also have practical problems. Apart from that, most of the recent works in NRSfM are based on orthographic camera models. [3] uses the orthographic camera to recover the shape's normal locally; they suffer from local two-fold ambiguities and significantly degrade for shorter focal lengths. [5] recently solved the same problem for an orthographic and perspective camera. [4] specifically addresses the case of piecewise planar surfaces; it uses the perspective camera but still has patch-wise two-fold unresolved ambiguities induced by the processing of image pairs.

In the paper, we present a general framework to solve Non-Rigid Shape-from-Motion (NRSfM) with the perspective camera for isometric deformations. Isometry allows solving for complex shape deformations from a sparse set of images. First, we formulate isometric NRSfM as a system of first-order Partial Differential Equations (PDE) involving the shape's depth and normal field and an unknown template. Second, we show the system cannot be locally resolved as such. Third, we introduce the concept of infinitesimal planarity and show that it makes the system locally solvable for three or more views. Finally, we derive an analytical solution which involves convex, linear least-squares optimization only, outperforming existing work on challenging datasets.

Modeling. We present our parametrization as in figure 2. Here, given a collection of surfaces related by isometric deformations we concentrate on two: S_i and S_j. The embedding functions going from the unknown flat template $\mathcal{T}$ to the 3D surfaces are φ_i and φ_j respectively. ξ_i and ξ_j are the normal fields on the surfaces S_i and S_j. η_i and η_j are the image warps from $\mathcal{T}$ and $\eta_{i,j}$ is the image to image warp from the image of the surface S_i to that of S_j. Finally we represent the Jacobian of any function θ as $\mathbf{J}_\theta$.

NRSfM as a system of PDEs and its solvability. In the paper we describe the SfT problem and how we can go from SfT to NRSfM by recognizing that the 3D template and its flattening are unknown in NRSfM. This leads to the following, which we refer to as the NRSfM problem:

$$\text{Find} \left| \begin{array}{l} \mathcal{T} \subset \mathbb{R}^2 \\ \varphi_i \in C^2(\mathcal{T};\mathbb{R}^3) \\ i = 1,\dots,n \end{array} \right. \quad \text{st} \quad \left\{ \begin{array}{ll} \eta_{i,j} = \eta_j \circ \eta_i^{-1} & j = 1,\dots,n \quad j \neq i \\ \eta_i = \Pi \circ \varphi_i & \\ (\mathbf{J}_{\varphi_i}\, \lambda\xi_i)(\mathbf{J}_{\varphi_i}\, \lambda\xi_i)^\top = \lambda^2 \mathbf{I}_{3\times 3} & \end{array} \right. \tag{1}$$

System (1) has three constraints on the right. The first is the *consistency constraint* and describes how the inter-image warp is related to the image warps (from the unknown template to the images). The second is the *reprojection constraint*, which simply says that the image is obtained from a perspective projection of the surface embedding φ_i. Finally the third describes the *deformation constraint* for isometry. Here, we suppose the unknown flat template $\mathcal{T}$ is obtained using a conformal flattening with some unknown scaling function λ.

System (1) is analyzed in the paper to reveal that it has effectively 6 equations with 8 unknowns for 2 views. Similarly for n views, we obtain $3n$ equations with $3n+2$ unknowns. This summarizes our second result that one cannot solve isometric NRSfM by relaxing the relationship between the depth and the normal.

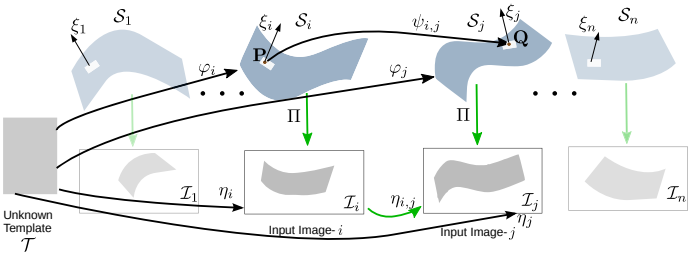


Figure 1: Geometric modeling of NRSfM.

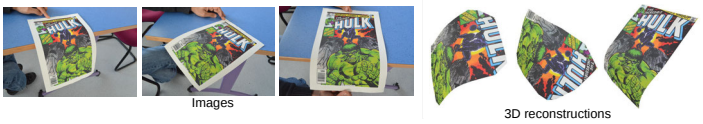


Figure 2: Images and their corresponding 3D reconstructions using our method.

Infinitesimal planarity and the solution to NRSfM. To obtain a solution to the NRSfM problem described by system (1), we assume that the surface is locally planar with zero second and higher-order derivatives. By doing this we show how isometry implies that the tangent planes on the surfaces are actually related by rigid transforms. Thus their images in turn are locally related by homographies. This is a very important result as it allows us to instead compute the aforementioned local homography to solve NRSfM. Given a known global inter-image warp we demonstrate a way to compute exactly such local homographies.

Homographies are well-understood in the literature and they can be decomposed to get the normals on the respective planes in 3D. However, there exists a two-fold ambiguity in computing the normal of a plane from its related homography. In the paper we present an algorithm to obtain the correct surface normal using at least 3 views that also robustly handles higher number of views. We tested our method with synthetic and real datasets. The results of these experiments show that our method outperforms the state of the art and is able to give a good 3D reconstruction under wide baseline viewpoints with significant deformations. We conclude that an algebraic solution of isometric NRSfM is not possible without further assumptions. However the use of infinitesimal planarity gives useful 3D reconstructions and is thus a valid assumption.

[1] A. Bartoli, Y. Gerard, 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 *CVPR*, 2012.

[2] M. Salzmann and P. Fua. Linear local models for monocular reconstruction of deformable surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(5):931–944, 2011.

[3] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *CVPR*, 2010.

[4] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-free monocular reconstruction of deformable surfaces. In *CVPR*, 2009.

[5] S. Vicente and L. Agapito. Soft inextensibility constraints for template-free non-rigid reconstruction. In *ECCV*, 2012.

Virtual Insertion: Robust Bundle Adjustment over Long Video Sequences

Ziyan Wu*[1]
ziyan.wu@siemens.com
Zhiwei Zhu[2]
zhiwei.zhu@sri.com
Han-Pang Chiu[2]
han-pang.chiu@sri.com

[1] Siemens Corporate Technology
Princeton, NJ, USA
[2] SRI International
Princeton, NJ, USA

Bundle Adjustment is a key process to enhance the global accuracy of the 3D camera pose and structure estimation in the framework of structure from motion over long video sequences. However, most bundle adjustment algorithms require sufficient visual feature correspondences from each camera frame to its neighboring frames in video sequences, which are hard to collect in real environments, especially for indoor real-time navigation applications. A camera may not observe enough common scene points over a long period of time due to occlusions or non-texture background such as the white walls etc.. With the use of video images as the only input, bundle adjustment will easily fail due to the constant link outage of visual landmarks in the scene. We call it the effect of "visual breaks", and the issue of "visual breaks" has hindered the usage of bundle adjustment. It is particularly critical for sequential Structure from Motion (sSfM) applications where motion estimation is from "chaining" neighboring key frames.

On the other hand, to deal with this issue of "visual breaks", vision-based navigation systems, such as Simultaneous Localization and Mapping (SLAM), typically do not rely on the video cameras only for robustness. Different techniques have been proposed to reduce the drift caused by"visual breaks" and other sources (e.g. inaccurate calibration) by fusing non-vision sensors, such as Inertial Measurement Unit (IMU) [1], LiDAR [3] or GPS [2]. As a result, good motion measurements from non-vision sensors or motion assumptions can be obtained easily at these "visual breaks" locations. However the bundle adjustment is still not able to use the motion estimates from these techniques directly due to a missing approach to incorporate them inside the cost function during optimization.

In this paper, in order to overcome the above issue, we propose a "Virtual Insertion" scheme to construct elastic virtual links on these "visual breaks" positions to fill visual landmark link outage with the measurements provided by other sensors or motion assumptions, so that all camera positions can be linked in the long video by the real or virtual scene landmarks before bundle adjustment. This way enables the traditional bundle adjustment algorithms to achieve robust large-area structure from motion over long video sequences. Specifically, with the measurements from non-vision sensors at the "visual break" positions, we actually convert them into a set of virtual landmark links that will serve as 3D-2D projection constraints in the cost function of bundle adjustment optimization. As a result, measurements from other sensors can be integrated into existing bundle adjustment framework. Experiments on real-world long video sequences show that the virtual insertion scheme can significantly enhance both robustness and global accuracy of bundle adjustment over long video sequences in challenging real-world environments.

A "visual break" is critical especially for sequential structure from motion, where usually a camera position has feature correspondences only with neighboring positions. With the help of IMU and Kalman filter [3], a visual odometry system is able to output reasonable and continuous poses using measurements from IMU especially at the "visual breaks". However the measurements from IMU cannot be integrated into the framework of bundle adjustment directly, resulting large jumps and drifts. This is because "visual breaks" can severely affect the bundle adjustment, in the sense that the global-minima of the whole sequence becomes the combination of local-minimas in each of the two segments of the sequence because the transition between the two sets of locations is unconstrained. This is the reason why large jumps can be found in the output trajectory from bundle adjustment when "visual breaks" exist in the sequence.

Figure 1 shows the illustration of a typical motion estimation over a video sequence with a "visual break" annotated with a red link arrow. Due to drift in the initial poses estimation, the loop does not close although the person travels back to the origin. It can be seen from Figure 1 that the initial estimated trajectory from is continuous and smooth. After feature matching, frames at the end are matched with frames at the beginning, and for the other locations, frames are only matched with their neighboring frames. During bundle adjustment, with the constraints provided the loop closure, the drifts at the end can be reduced. However a large jump can be observed at the "visual break" location, as showed in Figure 1, since the constraints cannot be propagated to the other locations because of the "visual break". It is straightforward to consider this "visual break" as a "broken joint".

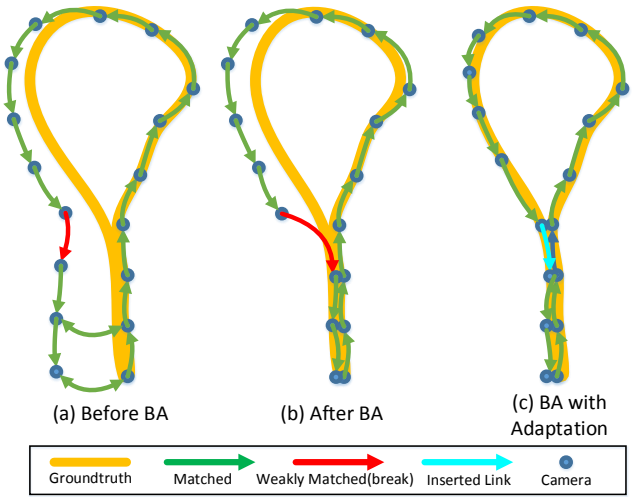


Figure 1: Linking the "visual break".

It is natural for us to think about adapting the "broken joints" with artificial links. From initial estimation of the camera poses fused with IMU, we can set up artificial links on the "visual breaks". Although drift will accumulate over long period in general, within a small period of time, the estimation fused with IMU can be considered as reliable and trustworthy. As shown in Figure 1, a virtual link estimated by IMU motion estimation can be inserted to the break location so that the constraints from loop closure can be propagated to the whole sequence. Hence, as it can be seen that the whole trajectory can reach global optima with drifts reduced on every location. In other words, this method is transferring the motion measurements from non-vision sensors into 3D-2D visual projection constraints, which are integrated into the cost function of bundle adjustment for a joint global optimization. This forms the base of proposed virtual insertion techniques.

[1] H. Chiu, S. Williams, F. Dellaert, S. Samarasekera, and R. Kumar. Robust vision-aided navigation using Sliding-Window Factor Graphs. *ICRA*, 2013.

[2] M. Lhuillier. Incremental fusion of Structure-from-Motion and GPS using constrained bundle adjustments. *IEEE PAMI*, 34(12):2489–95, December 2012.

[3] Z. Zhu, H. Chiu, T. Oskiper, S. Ali, R. Hadsell, S. Samarasekera, and R. Kumar. High-precision localization using visual landmarks fused with range data. *CVPR*, 2011.

*This work was done while the author was a student associate at SRI International, Princeton, NJ, USA.

Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization

Adria Ruiz[1]
adria.ruiz@upf.edu
Joost Van de Weijer[2]
joost@cvc.uab.es
Xavier Binefa[1]
xavier.binefa@upf.edu

[1] Universitat Pompeu Fabra (DTIC)
Barcelona, Spain
[2] Centre de Visió per Computador
Barcelona, Spain

Introduction: Most efforts in facial behavior analysis have focused on proposing supervised methods to detect a set of predefined gestures such as the Action Units. However, supervised AU detection is a difficult task which requires a huge labelling effort to annotate spontaneous behavior databases. In contrast, we focus on a different problem which we call facial behavior categorization. The goal is to estimate high-level semantic labels for videos of recorded people by means of analysing their facial expressions. As an example, consider a set of videos of people recorded while watching an advertisement. The videos are labelled with the subject's appreciation of the advertisement, revealing whether or not he liked it. The task of facial behavior categorization is to analyse the set of subject facial expressions during the whole recording and estimate the "Like/Not Like" label. This problem can be considered a weakly-supervised learning problem because we do not have access to frame-by-frame facial gesture annotations but only weak-labels at the video level are available. From this weak-annotations, we aim to learn a set of discriminative expressions and how they determine the high-level labels. Similar to [5], we pose facial behavior categorization as a Multiple Instance Learning problem. In MIL, the training set $\mathcal{T} = \{(X_1, y_1), (X_i, y_i), \ldots, (X_N, y_N)\}$ is formed by N pairs of bags $X_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$. Every $X_i = \{\mathbf{x}_{i1}, \mathbf{x}_{ij}, ..., \mathbf{x}_{iM}\}$ is a set of M instances $\mathbf{x}_{ij} \in R^D$. The labels $y_i \in \{0,1\}$ are binary variables indicating whether the class of the bag is positive or negative. The goal is to lean a classifier $F(X_*) = y_*$ able to predict a label y_* from a new test bag X_*. In facial behavior categorization, we consider a video as a bag X_i, its instances x_{ij} correspond to facial-descriptors extracted at each video-frame and y_i refers to the video weak-label.

Contributions: We propose a novel MIL method called Regularized Multi-Concept MIL for facial behavior categorization. In contrast to previous MIL methods applied to facial behavior analysis which use a Single-Concept approach, RMC-MIL follows a Multi-Concept assumption which allows different facial expressions (concepts) to contribute differently to the video-label. Moreover, to handle with the potential large number of non-informative features present in the high-dimensional facial-descriptors, RMC-MIL uses a discriminative approach to model the concepts and structured sparsity regularization. As a consequence, the concepts use only a common subset of features expected to be related with facial expression changes.

Rregularized Multi-Concept MIL: An overview of RMC-MIL is illustrated in Fig. 1. Our model learns a set of K hyperplanes $\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \ldots \ \mathbf{z}_K]$ in the instance space which classify instances depending when they belong or not to the k-th concept. This concepts are expected to represent different types of discriminative facial expressions. A bag

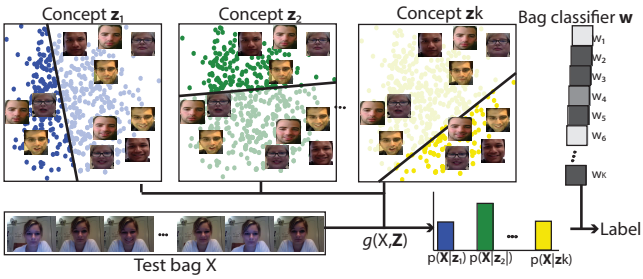


Figure 1: Overview of RMC-MIL. Concepts are modelled as a set of K classifiers $\mathbf{z_k}$ in instance space. A bag is represented using the probability of its instances given each concept. The bag-classifier $\mathbf{w}$ maps this bag-representation into high-level labels. Both $\mathbf{Z}$ and $\mathbf{w}$ parameters are jointly optimized during training.

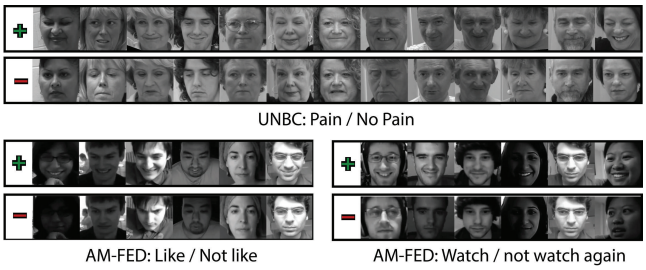


Figure 2: Most positive and negative instances estimated by RMC-MIL in a set of randomly selected videos for different facial behavior categorization problems in the AM-FED and UNBC datasets

(video) is represented as a K dimensional vector:

$$g(X_i, \mathbf{Z}) = < p(X_i|\mathbf{z}_1), p(X_i|\mathbf{z}_2), \ldots, p(X_i|\mathbf{z}_K) >, \quad (1)$$

where the value in the k-th dimension is the probability of that a concept k appears in the bag X_i. This probability is defined as the maximum probability $p(X_i|\mathbf{z_k}) = \max_j p(\mathbf{x}_{ij}|\mathbf{z}_k)$ among all the bag-instances. Finally, the bag-classifier is defined as $F(X) = \text{sgn}(\mathbf{w}^T g(X, \mathbf{Z}))$, where $\mathbf{w} = [w_1, w_2, \ldots, w_K]$ are the parameters of a linear classifier separating positive and negative bags embedded in the K dimensional space. In the training stage, RMC-MIL jointly optimize the bag-classifier $\mathbf{w}$ and concept-classifiers $\mathbf{Z}$ by using a logistic-loss function ℓ and solving:

$$\min_{\mathbf{w},\mathbf{Z}} \quad \mathcal{L}(\mathbf{w}, \mathbf{Z}) = \sum_{i=1}^{N} \ell(\mathbf{w}^T g(X_i, \mathbf{Z}), y_i) \quad s.t. \ ||\mathbf{Z}||_{2,1} \leq \tau_Z \quad (2)$$

The use of $L_{2,1}$ regularization is motivated by previous work [1] in Multi-Task Learning for supervised facial expression recognition. Concretely, it has been shown that the use of $L_{2,1}$ regularization to force joint sparsity between independent facial expressions classifiers increase their performance. Similarly, in the case of RMC-MIL, this regularization encourages the concept hyperplanes to use a common subset of features expected to be related with facial expression changes. Eq. 2 is a convex-constrained optimization problem and we use the Projected-Quasi-Newton [3] method to efficiently solve it.

Experiments: In our experiments, we evaluate the proposed approach in two different facial behavior categorization problems. Using the AM-FED [4] and UNBC [2] public datasets, we attempt to categorize viewer's responses to advertisements and detect pain from patients from weakly-labelled videos. We demonstrate the advantages of using multiple concepts in facial behavior categorization and the effectiveness of structured sparsity regularization in this context. Moreover, the results show the improvement of RMC-MIL over existing Single-Concept and Multi-Concept MIL methods and its ability to learn discriminative facial gestures from weakly-labeled data (Fig. 2).

[1] Lin Zhong et al. Learning active facial patches for expression analysis. In *CVPR*, June 2012.

[2] Lucey et al. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*, 2011.

[3] M. Schmidt et al. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *AISTATS*, 2009.

[4] McDuff et. al. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected in-the-wild. In *CVPR Workshops*, 2013.

[5] Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *FG*, 2013.

Expression-Invariant Age Estimation

Fares Alnajar*,[1]
F.alnajar@uva.nl
Zhongyu Lou*,[1]
z.lou@uva.nl
Jose Alvarez[2]
jose.alvarez@nicta.com.au
Theo Gevers[1]
th.gevers@uva.nl

[1] ISLA Lab, Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
[2] NICTA
Canberra ACT 2601
Australia

We investigate and exploit the influence of facial expressions on automatic age estimation. Different from existing approaches, our method jointly learns the age and the expression by introducing a new graphical model with a latent layer between the age/expression labels and the features. This layer aims to learn the relationship between the age and the expression and captures the face changes which induce the aging and the expression appearance, and thus obtaining expression-invariant age estimation.

External factors like facial expressions cause changes in facial muscles which distort the aging cues. A problem in age estimation is that expression-related muscles overlap with aging-induced facial changes. For example, smiling involves the activation of some facial muscles leading to raising the cheeks and pulling the lip corners. This influences the aging wrinkles around the mouth and near the eyes. Consequently, the aging cues changes caused by expressions show the necessity of separating the influence of expression when estimating the age.

We jointly learn the age and expression and model their relationship. More specifically, we introduce a new graphical model which contains a latent layer between the age/expression labels and the facial features. This layer captures the relationship between the age and expression. To predict the age, the age and expression are inferred jointly, and hence prior-knowledge of the expression of the test face is not required. The contributions of our work are: 1) we show how age-expression joint learning improves the age prediction compared to learning independently from expression. 2) As opposed to existing methods [2, 5], the proposed method predicts the age across different facial expressions without prior-knowledge of the expression labels of the test faces. 3) Finally, our results outperform the best reported results on age-expression datasets (FACES [1] and Lifespan [3]).

The proposed graphical model has four sets of connections: First, connections between the face subregions and the latent variables. These connections are designed to capture the changes of face appearance related to age and expression. Second, connections between the face subregions and the age/expression labels are formed. The aim here is to directly infer the age/expression from the features. Third, connections between the latent variable modeling the relationship between the face subregions. Finally, connections are established between the latent variables, the age, and the expression. The last type of connections is designed to relate the age with the expression which allows the joint learning between them.

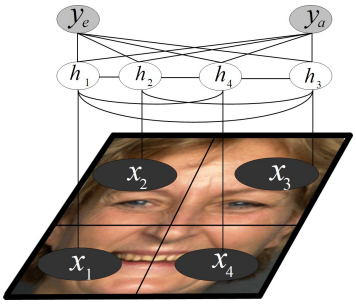


Figure 1: Our graphical model to jointly learn the age and the expression. $\mathbf{x} = [x_1, x_2, x_3, x_4]$ represents the feature vector, $\mathbf{h} = [h_1, h_2, h_3, h_4]$ denotes the latent variables, y_a and y_e are the corresponding age and expression respectively. Note that, while all x_i are connected with y_a and y_e, we do not show these connections in this figure for the sake of clarity.

Our model maximizes the conditional probability of the joint assignment of $\mathbf{y}$ given observation $\mathbf{x}$:

$$\mathbf{y}^* = argmax_{\mathbf{y}} P(\mathbf{y}|\mathbf{x};\theta). \tag{1}$$

$$P(\mathbf{y}|\mathbf{x};\theta) = \sum_{\mathbf{h}\in\mathcal{H}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\theta) = \frac{\sum_{\mathbf{h}\in\mathcal{H}} exp(\psi(\mathbf{y},\mathbf{h},\mathbf{x};\theta))}{\sum_{\mathbf{y}'\in\mathcal{Y},\mathbf{h}\in\mathcal{H}} exp(\psi(\mathbf{y}',\mathbf{h},\mathbf{x};\theta))}.$$

Where $\psi(.)$ is the potential function which measures the compatibility between the (observed) features, the joint assignment of the latent variables, and the output labels. The potential function is decomposed into four potential functions corresponding to the connections of the model (Figure 1).

$$\psi(\mathbf{y},\mathbf{h},\mathbf{x};\theta) = \sum_{i=1}^{4} \psi_1(y_a,x_i;\theta_i^1) + \sum_{i=1}^{4} \psi_2(y_e,x_i;\theta_i^2) + \sum_{i=1}^{4} \psi_3(h_i,x_i;\theta_i^3) + \psi_4(\mathbf{h},y_a,y_e;\theta^4). \tag{2}$$

To learn the parameters θ, we exploit the max margin approach [4]. The inference involves a combinatorial search of the joint assignment of $\mathbf{h}$, y_e and y_a which results in the maximum conditional probability:

$$(\hat{\mathbf{y}},\hat{\mathbf{h}}) = argmax_{\mathbf{y}\in\mathcal{Y},\mathbf{h}\in\mathcal{H}} \psi(\mathbf{x},\mathbf{y},\mathbf{h};\theta). \tag{3}$$

In the paper, we evaluate our model on FACES [1] (6 expressions) and Lifespan [3] (2 expressions) datasets. The experiments show the improvement in performance when the age is jointly learnt with the expression in comparison to expression-independent age estimation. The age estimation error is reduced by 14.43% and 37.75% for FACES and Lifespan datasets respectively. We show (Figure 2) the face regions corresponding to each hidden state (3).

Figure 2: Average face regions corresponding to different hidden states (from left to right) for the bottom and top face regions. For the bottom regions, the first hidden state corresponds to the face appearance where the mouth is open, the third hidden state represents a depressed lip corner, and the second hidden state corresponds to a normal face appearance. For the top regions, the second hidden state represents the face appearance where the eye is slightly closed while the first and the third states correspond to open eye appearances.

[1] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. Faces: database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1):351–362, 2010.

[2] Guodong Guo and Xiaolong Wang. A study on human age estimation under facial expression changes. *Computer Vision and Pattern Recognition*, pages 2547–2553, 2012.

[3] Meredith Minear and Denise C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, and Computers*, 36(4):630–633, 2004.

[4] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, pages 1453–1484, 2005.

[5] Chao Zhang and Guodong Guo. Age estimation with expression changes using multiple aging subspaces. *Biometrics: Theory, Applications and Systems*, pages 1–6, 2013.

* Both authors contributed equally.

A Stochastic Cost Function for Stereo Vision

Christian Unger
http://campar.in.tum.de/Main/ChristianUnger
BMW Group
Munich, Germany

Slobodan Ilic
http://campar.in.tum.de/Main/SlobodanIlic
Siemens AG
Research & Technology Center
Munich, Germany

The goal of this paper is to present a novel stochastic cost function for binocular stereo vision that delivers statistics about the most probable disparities on the pixel level. We drive these statistics by many independent stochastic processes so that robustness to outliers can be achieved. Each of these stochastic processes may be understood as an individual who is requested to deliver his opinion about the depth. Finally, the idea is to fuse all these individual measurements into one global disparity map. In this paper, we use random walks for this.

Figure 1: Examples for random walks.

Random walks, like the ones shown in Fig. 1, randomly traverse the image where at each step of the walk, an adjacent pixel location is chosen based on color similarity. In this sense, a random walk can be viewed as a local segmentation which is assumed to be robust along discontinuities. The set of pixels which is covered by the random walk is used to infer information about the disparity. In this paper, we demonstrate that random walks are useful for gathering important statistics about disparities. One strong property of our method is that our cost function is statistically motivated and we show that our proposed statistical consistency is a powerful and very useful confidence measure with which occlusions may be filtered out effectively.

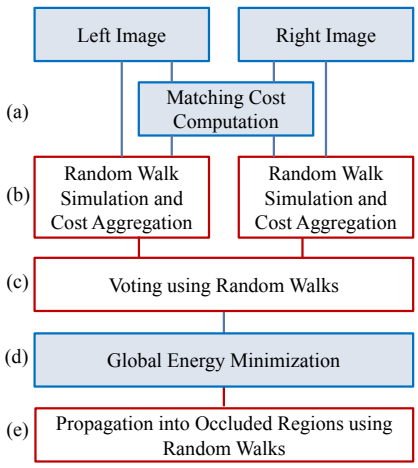


Figure 2: The processing steps of our method.

In Fig. 2 we depict the processing steps of our method: (a) the computation of pixel-wise matching costs. (b) In our cost aggregation stage, we simulate random walks for every pixel of the image. The costs are then summed for all pixels of a random walk. Since random walks rarely cross large image gradients, this can be understood as a pre-segmentation step to increase robustness at discontinuities. We also explicitly consider slanted surfaces by evaluating different surface orientations and by simulating random walks in both left and right images, we address occluded regions. (c) Once we computed a cost function for every random walk, we introduce a novel voting technique that fuses information of all random walks into one global voting space. The collected votes contain statistical information about the likelihood of every disparity at every pixel location and also reveal inconsistencies in places where matching is ambiguous. (d) The votes may be used in a global optimization or for direct disparity selection. (e) After this step, random walks may be used to propagate reliable matches into inconsistent regions.

In our paper, we provide a statistical consistency measure that serves as a confidence for every disparity value. The idea is that the confidence is high if many random walks confirm to the same disparity. Given that $\mathcal{V}(\mathbf{x},d)$ is the number of votes for disparity d at pixel $\mathbf{x}$, the *consistency* is defined as: $\frac{\mathcal{V}(\mathbf{x},\hat{d})}{1+\sum_i \mathcal{V}(\mathbf{x},i)}$, where $\hat{d}$ is the disparity with most votes at pixel $\mathbf{x}$. We analyze the reliability of this confidence value, also by discussing its ROC curve, which we present as an example in Fig. 3. Finally, we show disparity maps of challenging stereo images and we compare to other related methods.

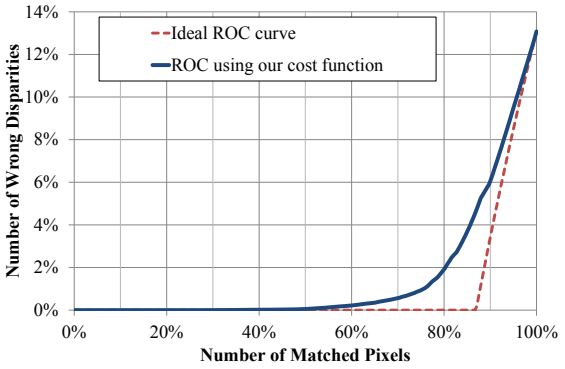


Figure 3: The ROC curve of our method.

To summarize, this paper proposes a novel stochastic cost function based on random walks which enables statistical reasoning on the discovered depth measurements. In particular, we introduce (1) a cost aggregation technique based on random walks which is orientation- and occlusion-robust, (2) a novel voting technique based on random walks to obtain statistical information about the disparity likelihood and (3) a strong novel statistical consistency measure. In our experiments we show impressive results on challenging stereo images. Given the obtained results we believe that our cost function together with the confidence is useful for other stereo methods and is valuable in practical applications.

Fusing Multiple Features for Shape-based 3D Model Retrieval

Takahiko Furuya
g13dm003@yamanashi.ac.jp
Ryutarou Ohbuchi
ohbuchi@yamanashi.ac.jp

Graduate School of Medicine and Engineering,
University of Yamanashi,
Yamanashi, Japan

Abstract

Fusing multiple features is a promising approach for accurate shape-based 3D Model Retrieval (3DMR). Most of the previous algorithms either simply concatenate feature vectors or sum similarities derived from features. However, ranking results due to these methods may not be optimal as they don't exploit distributions, i.e., manifold structures, of multiple features. This paper proposes a novel 3DMR algorithm that effectively and efficiently fuses multiple features. The proposed algorithm employs a Multi-Feature Anchor Manifold (MFAM) that approximates multiple manifolds of heterogeneous features with small number of "anchor" features. Given a query, ranks of 3D models are computed efficiently by diffusing relevance on the MFAM. Distance metrics of heterogeneous features are fused during the diffusion for better ranking. Experiments show that our proposed algorithm is more accurate and much faster than 3DMR algorithms we have compared against.

Proposed algorithm

For accurate and efficient 3DMR, we propose *3D model retrieval by Visual Feature Fusion* (*3DVFF*) algorithm that fuses multiple visual features of 3D models via an unsupervised distance metric fusion algorithm called *Multi-Feature Anchor Manifold Ranking* (*MFAMR*). Figure 1 shows an overview of the proposed algorithm. The 3DVFF algorithm first extracts two visual features SV-DSIFT and LL-MO1SIFT from each 3D model in a database. The SV-DSIFT aggregates local visual features extracted from multiple viewpoints by Super Vector (SV) coding [1], and shows high accuracy for models having global deformation and/or articulation. The LL-MO1SIFT aggregates per-view global image features by using Locality-constrained Linear (LL) coding [2], and shows high accuracy for rigid models. For each feature, to reduce computational cost, a manifold of all the features is approximated by a manifold of anchors. Then, the two anchor manifold graphs are fused into a MFAM graph [3]. Ranking of the 3D models in the database for a given query is efficiently computed by relevance diffusion from the query to the 3D models over the MFAM. The two heterogeneous manifolds, one for the SV-DSIFT and the other for the LL-MO1SIFT, are fused during the relevance diffusion over the MFAM to yield a fused distance metric.

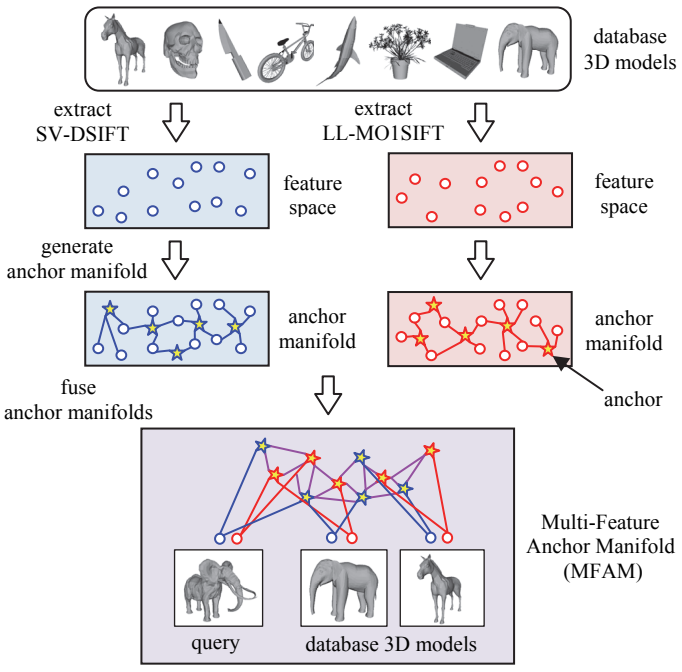


Figure 1: Overview of the proposed algorithm.

Experiments and results

To evaluate accuracy and efficiency of the proposed 3DVFF algorithm, we use two benchmarks; the *Princeton Shape Benchmark* (*PSB*) [4] and the SH14LC [5] (Results for other benchmarks are presented in a full paper). Figure 2 shows examples of 3D models for the benchmarks. For the PSB, we use 400 anchors for SV-DSIFT and 500 anchors for LL-MO1SIFT to approximate structure of multi-feature manifold. For the SH14LC, we use 2,000 anchors for SV-DSIFT and 2,500 anchors for LL-MO1SIFT. We use *Mean Average Precision* (*MAP*) [%] for quantitative evaluation of retrieval accuracy.

Figure 2: Examples of 3D models for the benchmarks.

Efficiency: Table 1 shows computation time per query. Feature extraction is accelerated by using a GPU and multi-core CPUs. MFAMR is computed on a single thread. Our 3DVFF using MFAMR is much faster than previous feature fusion method MR-early, which sums multiple feature similarities to generate a manifold graph and performs naive Manifold Ranking [6] at every query. The 3DVFF takes less than 3 seconds per query for the SH14LC having 8,987 3D models.

algorithms	Feature extraction	Ranking	Total
MR-early	2.575	34.779	37.354
3DVFF (proposed)	2.575	0.031	**2.606**

Table 1: Computation time per query for the SH14LC benchmark.

Accuracy: Table 2 compares retrieval accuracy of the proposed 3DVFF algorithm with other state-of-the-art 3D model retrieval algorithms. For both the PSB and the SH14LC, our proposed 3DVFF algorithm achieved the highest retrieval accuracy among the algorithms listed in Table 2. For the SH14LC benchmark, MAP score of our 3DVFF is 3% higher than LCDR-DBSVC, the best performing algorithm among the SH14LC track entries [5].

algorithms	PSB	SH14LC
MR-D1SIFT [6]	62.9	46.4
LCDR-DBSVC [5]		54.1
SV-DSIFT	63.4	46.4
LL-MO1SIFT	55.3	39.9
3DVFF (proposed)	**72.6**	**57.2**

Table 2: Comparison of retrieval accuracy (MAP [%]).

[1] X. Zhou, K. Yu, T. Zhang, and T.S. Huang. Image Classification using Super-Vector Coding of Local Image Descriptors, *Proc. ECCV 2010*:141–154, 2010.

[2] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained Linear Coding for Image Classification, *Proc. CVPR 2010*:3360–3367, 2010.

[3] S. Kim and S. Choi. Multi-view anchor graph hashing, *Proc. ICASSP 2013*:3123–3127, 2013.

[4] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The Princeton Shape Benchmark, *Proc. SMI 2004*:167–178, 2004.

[5] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, Q. Chen, N. K. Chowdhury, B. Fang, T. Furuya, H. Johan, R. Kosaka, H. Koyanagi, R. Ohbuchi, and A. Tatsuma. Large Scale Comprehensive 3D Shape Retrieval, *Proc. EG 3DOR 2014*:131–140, 2014.

[6] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on Data Manifolds, *Proc. NIPS 2004*, 2004.

[7] R. Ohbuchi and T. Furuya. Distance metric learning and feature combination for shape-based 3D model retrieval, *Proc. 3DOR 2010*:63–68, 2010.

Unsupervised RGB-D image segmentation using joint clustering and region merging

Md. Abul Hasnat
mdabul.hasnat@univ-st-etienne.fr
Olivier Alata
olivier.alata@univ-st-etienne.fr
Alain Trémeau
alain.tremeau@univ-st-etienne.fr

Hubert Curien Lab., UMR CNRS 5516,
Jean Monnet University, Saint Etienne, France.

Recent advances in imaging sensors, such as Kinect, provide access to the synchronized depth with color, called RGB-D image. Numerous researches [2, 4] have shown that the use of depth as an additional feature improves accuracy of scene segmentation. However, it remains an important issue - what is the best way to fuse color and geometry in an unsupervised manner? We focus on this issue and propose a solution.

In this paper, we propose an unsupervised method for indoor RGB-D image segmentation and analysis. The proposed method combines a clustering method with a region merging method. First, it identifies the possible image regions using clustering w.r.t. a statistical image generation model. Then, it merges regions based on planar statistics.

We consider a statistical image generation model in order to fuse color and shape (3D and surface normal) features. The model assumes that the features are independently (*naïve Bayes* assumption) issued from a finite mixture of multivariate Gaussian (for color and 3D) and a multivariate Watson distribution [6] (for surface normal). Mathematically, such a model with k components has the following form:

$$g(\mathbf{x}_i|\Theta_k) = \sum_{j=1}^{k} \pi_{j,k} f_g(\mathbf{x}_i^C|\mu_{j,k}^C, \Sigma_{j,k}^C) f_g(\mathbf{x}_i^P|\mu_{j,k}^P, \Sigma_{j,k}^P) f_w\left(\mathbf{x}_i^N|\mu_{j,k}^N, \kappa_{j,k}^N\right)$$

Here $\mathbf{x}_i = \{\mathbf{x}_i^C, \mathbf{x}_i^P, \mathbf{x}_i^N\}$ is the feature vector of the ith pixel with $i = 1,...,M$. Superscripts denote: C - color, P - 3D position and N - normal. $\Theta_k = \{\pi_{j,k}, \mu_{j,k}^C, \Sigma_{j,k}^C, \mu_{j,k}^P, \Sigma_{j,k}^P, \mu_{j,k}^N, \kappa_{j,k}^N\}_{j=1...k}$ denotes the set of model parameters where $\pi_{j,k}$ is the prior probability, $\mu_{j,k}$ is the mean, $\Sigma_{j,k}$ is the variance-covariance matrix and $\kappa_{j,k}$ is the concentration of the jth component. $f_g(.)$ and $f_w(.)$ are the density functions of the multivariate Gaussian distribution and the multivariate Watson [6] distribution respectively.

Fig. 1 illustrates the work flow of our RGB-D segmentation method that consists of two tasks: (1) cluster features and (2) merge regions. The first task performs a joint color-spatial-axial clustering and generates a set of regions. The second task performs a refinement on the set with the aim to merge regions which are susceptible to be over-segmented.

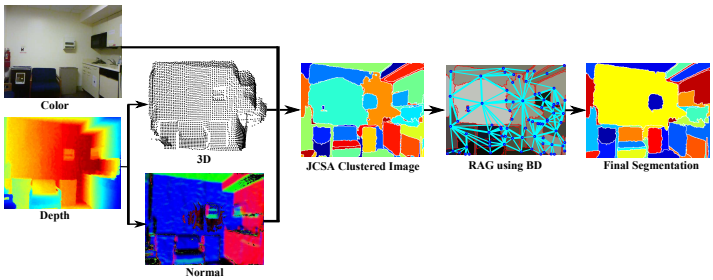


Figure 1: Work flow of the proposed RGB-D segmentation method.

We develop a Joint Color-Spatial-Axial (JCSA) clustering method to cluster pixels w.r.t. our image model. We exploit *Bregman Soft Clustering (BSC)* [1] method which has been effectively employed for mixture models based on exponential family of distributions. Compared to the traditional Expectation Maximization algorithms, BSC provides additional benefits: (a) it considers *Bregman Divergence* that generalizes a large number of distortion functions [1]; (b) simplifies computationally expensive Maximization step and (c) is applicable to mixed data type. Details of the JCSA clustering method is presented in the paper.

In an unsupervised setting the true number of segments are unknown. Therefore using JCSA, we cluster image features with an assumption of maximum number of clusters ($k = k_{max}$). Such an assumption often causes an over-segmentation of the image.

In order to tackle the over-segmentation issue mentioned above, we develop a statistical region merging method. It exploits planar property, which is related to the parameters (μ and κ) of the Watson distribution associated with each region. Our method first builds a region adjacency graph $G = (V, E)$. Each node $v_i \in V$ consists of concentration κ_i of the surface normals of its corresponding region. Each edge e_{ij} consists of two weights: w_d, based on statistical dissimilarity and w_b, based on boundary strength between adjacent nodes v_i and v_j. Then, following the standard region merging methods [3], we define a *region merging predicate* as:

$$P_{ij} = \begin{cases} true, & \text{if } (a)\ \kappa_j > \kappa_p \text{ and} \\ & (b)\ w_d(v_i, v_j) < th_d \text{ and } w_b(v_i, v_j) < th_b \text{ and} \\ & (c)\ planar\,outlier\,ratio > th_r; \\ false, & \text{otherwise.} \end{cases}$$

where κ_p is the threshold to define the planar property of a region. th_d and th_b are the thresholds associated with the distance weight w_d and boundary weight w_b. th_r is the threshold associated with the plane outlier ratio. The details of these thresholds are discussed in the paper. The *region merging order* sorts the adjacent regions that should be evaluated and merged sequentially.

Our proposed method is called JCSA-RM (joint color-spatial-axial clustering and region merging). We evaluate JCSA-RM on the benchmark image database NYUD2 [5] which consists of 1449 indoor RGB-D images with ground-truth segmentation. We evaluate its performance using five standard benchmarks: (1) Probability Rand Index (*PRI*); (2) Variation of Information (*VoI*); (3) Boundary Displacement Error (*BDE*); (4) Ground Truth Region Covering (*GTRC*) and (5) Boundary based F-Measure (*BFM*).

First, we study the sensitivity of JCSA-RM w.r.t. the parameters (k, κ_p, th_b, th_d). Then, we compare JCSA-RM with several unsupervised RGB-D segmentation methods. Among them, RGB-D extension of OWT-UCM [4] (UCM-RGBD) method is the most competitive method. Results (presented in the paper) show that JCSA-RM performs best in PRI, VoI and GTRC and comparable in BDE and BFM. We compared these two competitive methods based on computation time and observe that JCSA-RM (MATLAB) is ≈3 times faster than UCM-RGBD (C++).

JCSA-RM is an unsupervised RGB-D image segmentation method. It is comparable with the state of the art methods and it needs less computation time. It opens interesting perspectives to fuse color and geometry in an unsupervised manner. We foresee several possible extensions, such as: more complex image model and clustering with additional features, region merging with additional hypothesis based on color.

[1] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] Carlo Dal Mutto, Pietro Zanuttigh, and Guido M Cortelazzo. Fusion of geometry and color information for scene segmentation. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):505–521, 2012.

[3] Richard Nock and Frank Nielsen. Statistical region merging. *IEEE TPAMI*, 26(11):1452–1458, 2004.

[4] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, pages 2759–2766. IEEE, 2012.

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012*, pages 746–760. Springer, 2012.

[6] Suvrit Sra and Dmitrii Karp. The multivariate watson distribution: Maximum-likelihood estimation and other aspects. *J Multivar Anal*, 114:256 – 269, 2013.

CP-Census: A Novel Model for Dense Variational Scene Flow from *RGB-D* Data

David Ferstl
ferstl@icg.tugraz.at
Gernot Riegler
riegler@icg.tugraz.at
Matthias Ruether
ruether@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

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

Dynamic scene understanding is an essential topic in computer vision. It tries to combine information from tracking, 3D reconstruction, segmentation, motion estimation to infer information about an ever changing 3D environment. While structure from motion for measuring movements in space is well understood on static scenes, the motion estimation of non-static scenes, known as Scene Flow (*SF*), still pose a challenging problem. This gets even harder if the moving objects are non-rigid. A popular way to estimate *SF* is to use a calibrated and synchronized multi-view setup and combine traditional Optical Flow (*OF*) estimation with simultaneous 3D reconstruction [1, 4]. With the recent range sensor developments, such as the Microsoft Kinect or the Intel Gesture Camera, the *SF* estimation solely from RGB-D data became a popular alternative [2, 3, 5].

In this paper we show a novel method for accurate and robust *SF* estimation of non-rigid scenes from RGB-D data. This estimation is solved in an dense variational energy minimization framework

$$\min_{\mathbf{u}} G_I(I_1, I_2, \mathbf{u}) + G_D(D_1, D_2, \mathbf{u}) + R(\mathbf{u}) \tag{1}$$

based on a multi-scale Ternary Census Transform (*TCT*) for the intensity data term G_I in combination with a depth data term G_D based on the patch-wise Closest Point (*CP*) distance, as shown in Figure 1. The motion in our estimation is modeled as *direct* projection and image warping W in 3D.

In particular, we propose an intensity data term G_I to estimate the scene correspondences given by the *TCT* on a local neighborhood $\mathcal{N}$:

$$G_I(\mathbf{x}, \mathbf{u}) = \frac{1}{|\mathcal{N}|-1} \sum_{i=1}^{|\mathcal{N}|-1} 1 - [C_i(I_2, W(\mathbf{x}, \mathbf{u})) = C_i(I_1, \mathbf{x})], \tag{2}$$

Where C is the ternary census signature of each patch. This *TCT* term calculates the intensity difference by an encoding of the illumination invariant local structure. The similarity is calculated by the Hamming distance between the signature patches. The a depth data term G_D is calculated as the patch-wise distance to the *CP* in 3D, which makes it more robust in low structured regions and in case of acquisition noise:

$$G_D(\mathbf{x}, \mathbf{u}) = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} \|\mathbf{X}_2(\mathbf{y}) - \mathbf{u}(\mathbf{y}) - \mathbf{X}_1(\mathbf{y}^*)\|_2. \tag{3}$$

Compared to traditional pointwise constancy terms our method is invariant to most illumination changes, more robust to acquisition noise and delivers better guidance in regions with low structure or low texture. The *SF* constraints are combined with a higher order regularization term R, namely Total Generalized Variation (*TGV*). The regularizer is weighted and directed by an anisotropic diffusion tensor based on the input data. Because both the intensity as well as the depth data are highly non-convex a simple linearization as in traditional methods is not longer sufficient. We therefore perform a direct second-order Taylor expansion of the pointwise data terms, similar to [6]. The proposed whole variational energy model is efficiently solved based on the primal-dual formulation and is efficiently parallelized to run at high frame rates.

In an extensive evaluation we show the different properties and contributions of the different terms in our model. The applicability of our method to different kinds of camera modalities is shown in Figure 2. Beyond that, we show that the accuracy of our method is superior compared to current *SF* approaches based on the Middlebury Benchmark, as shown in Table 1. Our method better handles scenes with low texture or low structure and is robust to illumination changes. It can cope with smooth flow transitions, which occur at rotations or non-rigid movements, while sharp boundaries of the flow field are preserved.

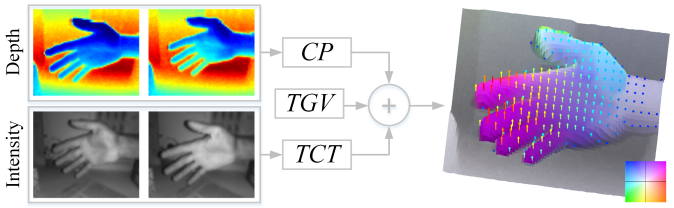
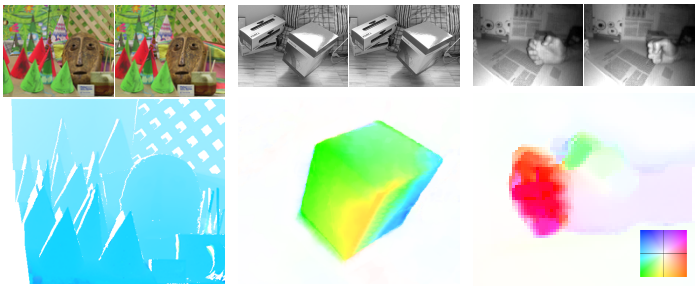


Figure 1: The *scene flow* is estimated from two consecutive depth and intensity acquisitions. The depth data term is calculated as patch-wise Closest Point (*CP*) search and the intensity data term is calculated as Ternary Census Transform (*TCT*). For regularization we propose an anisotropic Total Generalized Variation (*TGV*). The flow is visualized as a color coded X, Y map (motion key in the bottom right). The Z component is shown as arrows colored according to their magnitude.



(a) Middlebury (b) K4Wv2 (c) PMD Nano

Figure 2: *CP-Census SF* results on real image sequences. In the first column the results of the Middlebury *Cones* sequence, in the second column the flow of a rotated box with the K4Wv2 and in the third column a hand closing sequence (non-rigid movement) acquired with the PMD Nano are shown.

[1] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. In *CVPR*, 2010.

[2] Simon Hadfield and Richard Bowden. Scene particles: Unregularized particle-based scene flow estimation. *TPAMI*, 36(3):564–576, 2014.

[3] Michael Hornáček, Andrew Fitzgibbon, and Carsten Rother. Sphereflow: 6dof scene flow from rgb-d pairs. In *CVPR*, 2014.

[4] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *ICCV*, 2007.

[5] Julian Quiroga, Frédéric Devernay, and James L. Crowley. Local/global scene flow estimation. In *ICIP*, 2013.

[6] Manuel Werlberger, Thomas Pock, and Horst Bischof. Motion estimation with non-local total variation regularization. In *CVPR*, 2010.

	Cones EPE / RMS_{V_Z} / AAE			*Teddy* EPE / RMS_{V_Z} / AAE			*Venus* EPE / RMS_{V_Z} / AAE		
Basha et al. [1](2 views) (st)	0.58	N/A	0.39	0.57	N/A	1.01	0.16	N/A	1.58
Huguet and Devernay [4] (st)	1.10	N/A	0.69	1.25	N/A	0.51	0.31	N/A	0.98
Hadfield and Bowden [2]	1.24	0.06	1.01	0.83	0.03	0.83	0.36	0.02	1.03
Quiroga et al. [5]	0.57	0.05	0.52	0.69	0.04	0.71	0.31	**0.00**	1.26
Hornáček et al. [3]	0.54	**0.02**	0.52	0.35	**0.01**	0.16	0.26	0.02	0.64
CP-Census	**0.40**	0.03	**0.04**	**0.31**	0.02	**0.05**	**0.15**	**0.00**	**0.41**

Table 1: Quantitative comparison of *SF* methods on the Middlebury dataset. The error is measured by *EPE*/*AAE* in 2D, and *RMS* in Z direction. The best result for each dataset is highlighted and the second best is underlined. Methods that calculate *SF* from stereo are marked with (st).

Is 2D Information Enough For Viewpoint Estimation?

Amir Ghodrati
amir.ghodrati@esat.kuleuven.be
Marco Pedersoli
marco.pedersoli@esat.kuleuven.be
Tinne Tuytelaars
tinne.tuytelaars@esat.kuleuven.be

KU Leuven, ESAT - PSI, iMinds
Leuven, Belgium

Context. Estimating the pose of objects is a classical problem in vision. It aims at predicting a discrete or continuous viewpoint. Recent top performing methods for viewpoint estimation use 3D information. These 3D annotations are expensive and not really available for many classes.

What does this paper demonstrate. We show that a very simple 2D architecture (in the sense that it does not make any assumption or reasoning about the 3D information of the object) generally used for object classification, if properly adapted to the specific task, can provide top performance also for pose estimation. More specifically, we demonstrate how a 1-vs-all classification framework based on a Fisher Vector (FV) [1] pyramid or convolutional neural network (CNN) based features [2] can be used for pose estimation. In addition, suppressing neighboring viewpoints during training seems key to get good results.

The pipeline. Our method takes as input a detection bounding box, extracts features and assigns to the bounding box a pose. The estimation of the pose is done with a one-vs-all classifier of a discrete set of viewpoints.

- Detection: we use the deformable part models (DPM). We train our viewpoint estimation on the detected objects.
- Feature Extraction: we extract dense SIFT descriptors from the output of the detector. They are enriched by augmenting the location of the patch centre with respect to the upper-left corner of the bounding box, normalized by its size.
- Pose representation: We compare two representations commonly used in visual classification: Fisher Vector [1] + spatial pyramid matching and convolutional neural network based features [2].
- Learning: we consider each viewpoint as a different class. In this scenario an important difference with a standard 1-vs-all multi-class problem is that nearby viewpoints are generally visually very correlated. In the experimental results we show that eliminating nearby poses from negative samples always improves the viewpoint estimation. We call this procedure neighboring viewpoint suppression or briefly `nv-suppression`.

Experimental Evaluation. We evaluated our method on four datasets: Annotated faces-in-the-wild (AFW), EPFL multi-view car dataset, PASCAL3D+ and 3DObject dataset.

In table 1, we evaluate the performance of different features and encodings. we clearly notice that Bag-of-Words (`BoW`) representation is the poorest method for pose representation. The best representation on both datasets is `fisher` with spatial pyramid `spm`. Also embedding spatial information in the low-level (`sift+loc`) is still advantageous. Finally, CNN-based features, `decaf`, performs quite good as well, especially considering their much lower dimensionality.

		EPFL (8 poses)	**AFW (13 poses)**
Feature Type	Encoding	MPPE	FVP±15
`sift`	`BoW`	54.8%	49.4%
`sift`	`fisher`	68.2%	54.3%
`sift`	`fisher+spm`	80.1%	69.7%
`sift+loc`	`fisher+spm`	**81.8%**	**70.3%**
`decaf`	-	72.0%	67.9%

Table 1: An evaluation with training and testing data from output of detector on the EPFL car dataset and AFW faces dataset. MPPE is computed as the average of the diagonal of the confusion matrix. FVP±15 is the fraction of faces that are within ±15 degrees error interval, counting missed detections as infinite error.

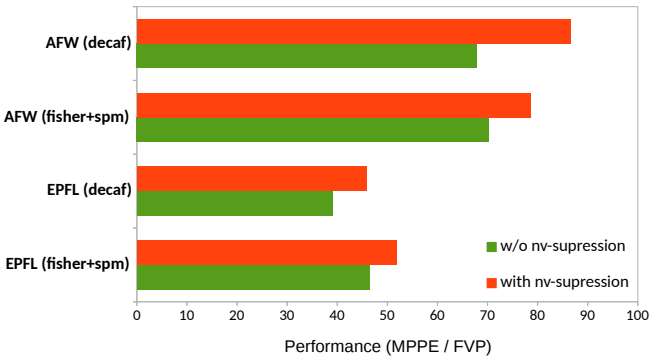


Figure 1: The effect of `nv-supression` using 36 poses for EPFL and 13 poses for AFW dataset.

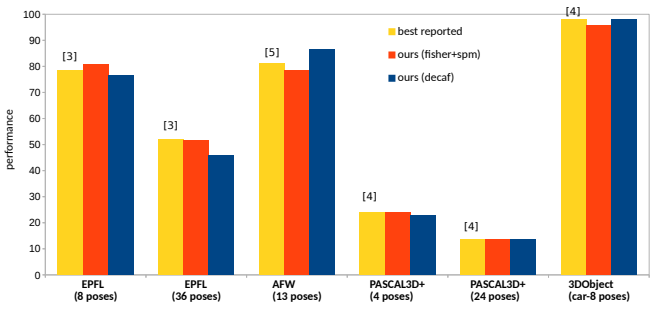


Figure 2: Viewpoint estimation in terms of MPPE, FVP(±15), mean AVP (Average Viewpoint Precision) and MPPE for EPFL, AFW, PASCAL3D+ and 3DObjects datasets respectively.

Figure 1 shows the effect of the neighboring viewpoints supression (`nv-suppression`). Its advantage is quite evident for the finer binning pose estimation for both types of features.

Comparison with state-of-the-art. Figure 2 shows the results of our methods and the current state-of-the-art on four datasets.

Conclusion. Through an extensive evaluation we can clearly see that for the fine-grained task of pose estimation, in contrast to common believe, the very simple framework based on the extraction of modern features (`decaf`) or in combination with modern encodings (`fisher+spm`) can in most of the cases get similar results as the 3D methods previously proposed and designed specifically for the problem of pose estimation.

References

[1] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *NIPS*, 1999.

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

[3] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In *ECCV*, 2012.

[4] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[5] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

Discrete Multi Atlas Segmentation using Agreement Constraints

Stavros Alchatzidis[134]
stavros.alchatzidis@ecp.fr
Aristeidis Sotiras[2]
aristeidis.sotiras@uphs.upenn.edu
Nikos Paragios[134]
nikos.paragios@ecp.fr

[1] Equipe GALEN, INRIA Saclay, Île-de-France, Orsay, France
[2] Section of Biomedical Image Analysis,Department of Radiology, University of Pennsylvania, Pennsylvania, USA
[3] Ecole des Ponts Paristech, Champs-sur-Marne, Île-de-France, France
[4] Ecole Centrale de Paris, Châtenay-Malabry, Île-de-France, France

Atlas-based segmentation describes a class of methods based on the registration of an annotated representative volume to a target one. When a single atlas is used a segmentation of the target image is obtained by warping the annotations using the deformation field found by the registration process. Recently, it has been shown that performing multiple such registrations allows for much improved results comparing to the single atlas case. In Multi Atlas segmentation the target image annotations are produced by fusing the multiple hypotheses either in a local [2, 4] or a global [1] fashion. In most methods, the segmentation problem is solved in two discrete steps and registration is merely seen as a fixed preprocessing step.

In this paper, we aim to couple the registration and segmentation problem through a unified formulation for multi-atlas segmentation. Registration terms seek optimal visual correspondences between atlases and target volumes while imposing smoothness. Segmentation terms seek voxel-wise consensus on the labeling of the target with respect to the deformed segmentation maps. Prior per voxel probabilities, produced by learning of local features, are taken into account in a seamless manner. In order to mathematically formulate these components, we adopt a pairwise Markov Random Field (MRF) graphical model where each atlas is associated with a deformation field, while the target image is associated with a segmentation map.

MRF Energy The dicrete pairwise energy function takes the form of:

$$E_{MRF}(\mathbf{l}) = \sum_{p \in \mathcal{V}} g_p(l_p) + \sum_{(p,q) \in \mathcal{E}} f_{pq}(l_p, l_q) \tag{1}$$

Graph Structure. Graph $\mathcal{G}$ is made of a set of N isomorphic grid graphs $\mathcal{G}_D = \{\mathcal{G}_{D_0}, \ldots, \mathcal{G}_{D_{N-1}}\}$ that represent deformation fields and a set of nodes $\mathcal{V}_S$ that represent segmentation. Nodes $p \in \mathcal{V}_{D_i}$ encode control points. The solution space around a control point is quantized and indexed by a discrete set of variables $\mathcal{L}_D$. This set represents possible control point displacements. We refer to a potential control point displacement attributed to a deformation node by l^d. The edge system of each grid in $\mathcal{G}_D$, $\mathcal{E}_{D_i}$ is created by a regular connectivity scheme.Each node $p \in \mathcal{V}_S$ corresponds to a random variable. The set of possible solutions for nodes of $\mathcal{V}_S$, $\mathcal{L}_S$ represents the set of anatomical regions augmented by the background label.

The energy in Eq.1 can be seen as broken into 4 terms:

Matching. The matching term, quantifies how well an atlas matches the target image.

$$g^M_{p_i}(l^d_{p_i}) = \int_\Omega \hat{\omega}_{p_i}(x) \rho(A_i \circ D_i^{l^d_{p_i}}, I(x)) dx. \tag{2}$$

$D_i^{l^d_{p_i}}$ is the transformation induced by the movement of the control point p in the ith deformation grid by the displacement $l^d_{p_i}$. The weighting function $\hat{\omega}_{p_i}$ determines the contribution of the point x to the unary potential of the control point p.

Deformation smoothness. [3] shows that deformation regularization can be efficiently modeled by pairwise potentials by:

$$f^R_{p_i q_i}(l^d_{p_i}, l^d_{q_i}) = \|\mathbf{d}^{l^d_{p_i}} - \mathbf{d}^{l^d_{q_i}}\|, \tag{3}$$

where $\mathbf{d}^{l^d_{p_i}}$ is the displacement applied to node p in the i-th deformation grid, indexed by l^d_{pi}.

Segmentation. This term takes into account a per-voxel probability distribution $\pi_x(l)$ and is incorporated in the MRF model by setting the unary potentials of the segmentation grid for every label to the negative log-probability of the respective class:

$$g^{SP}_{q_S}(l^s_{q_S}) = \int_\Omega \hat{\omega}_{q_S}(x)(-log(\pi_x(l^s_{q_S}))). \tag{4}$$

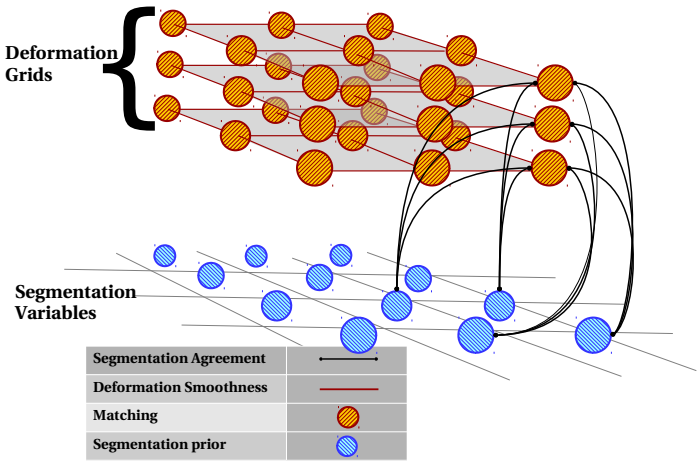


Figure 1: Graph structure of the proposed pairwise model.

where $\hat{\omega}_{q_S}(x)$ denotes the support of the segmentation node. For our experiments we have corresponded each segmentation node to a voxel, but a coarser or softer assignment could be envisaged.

Integrated Segmentation and Multi-Atlas Registration. To encourage the agreement between the estimated segmentation and the warped segmentation we penalize control point displacements of grid $\mathcal{G}_{D_i}$ that result in the warped segmentation mask corresponding to atlas i not agreeing with our final segmentation:

$$f^C_{p_i q_S}(l^d_{p_i}, l^s_{q_S}) = \int_\Omega \hat{\omega}_{q_S}(x) \hat{\omega}_{p_i}(x) \hat{\rho}(A_i \circ D_i^{l^d_{p_i}}, I(x)) Ind(S_i \circ D_i^{l^d_{p_i}}(x), l^s_{q_S}) dx \tag{5}$$

where p_i belongs to the grid $\mathcal{G}_{D_i}$ and q_S belongs to $\mathcal{G}_S$. $Ind(x,y) = 1$ except from $Ind(x,x) = 0$.

Validation. Comparing to independent pairwise registrations, our method is shown to increase registration quality in terms of overlap and harmonic energy. In addition, concensus between hypotheses is enforced leading to more concordant pairwise registrations, rejecting agressive deformation fields produced merely by good matchings.In addition, classical label fusion methods are outperformed by the annotations produced by our method.

[1] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738, 2009.

[2] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano. Combination Strategies in Multi-Atlas Image Segmentation: Application to Brain MR Data. *Medical Imaging, IEEE Transactions on*, 28(8):1266–1277, August 2009.

[3] Ben Glocker, Aristeidis Sotiras, Nikos Komodakis, and Nikos Paragios. Deformable Medical Image Registration: Setting the State of the Art with Discrete Methods*. *Annual Review of Biomedical Engineering*, 13(1):219–244, 2011.

[4] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken. Multi-Atlas-Based Segmentation With Local Decision Fusion. Application to Cardiac and Aortic Segmentation in CT Scans. *Medical Imaging, IEEE Transactions on*, 28(7):1000–1010, July 2009.

Video Segmentation by Non-Local Consensus Voting

Alon Faktor
http://www.wisdom.weizmann.ac.il/~alonf/
Michal Irani
http://www.wisdom.weizmann.ac.il/~irani/

Dept. of Computer Science and Applied Math
The Weizmann Institute of Science
ISRAEL

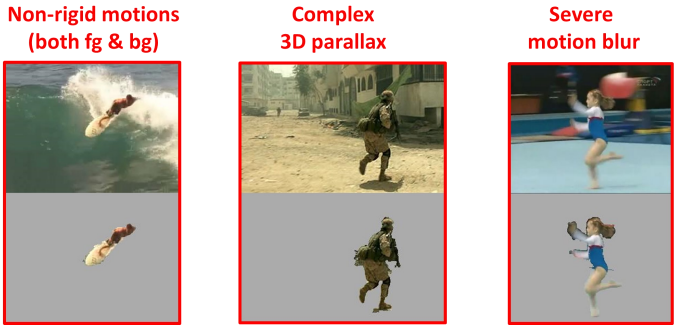


Figure 1: **A unified approach to foreground/background video segmentation in *unconstrained* videos.** Our algorithm can handle in a single framework video sequences which contain highly non-rigid foreground and background motions, complex 3D parallax as well as simple 2D motions and severe motion blur.

We address the problem of Foreground/Background (fg/bg) segmentation of "unconstrained" video. By "unconstrained" we mean that the moving objects and the background scene may be highly non-rigid (e.g., waves in the sea); the camera may undergo a complex motion with 3D parallax; moving objects may suffer from motion blur, large scale and illumination changes, etc. Fig. 1 shows a few such examples. Most existing segmentation methods fail on such unconstrained videos, especially in the presence of highly non-rigid motion and low resolution. Unconstrained video has thus become the focus of most recent video segmentation methods [5, 6, 9, 13].

In this paper, we suggest a simple yet general algorithm for performing fg/bg video segmentation, which handles complex unconstrained videos. We cast the video segmentation problem as a voting scheme on the graph of similar ("re-occurring") regions in the video sequence. 'Re-occurring' regions can be quite far both in space and in time, but are constrained to be close in the appearance feature space. We start from crude saliency votes at each pixel, and iteratively correct those votes by "consensus voting" of re-occurring regions across the video sequence. **The power of our consensus voting comes from the *non-locality* of the region re-occurrence, both in space and in time – enabling fast propagation of diverse and rich information across the entire video sequence.** This enables the correction of large errors in the initial fg/bg votes.

In contrast to trajectory-based methods [1, 2, 3, 4, 7, 8, 10, 11], we do not try to explicitly estimate long-term correspondences via flow estimation or tracking, but rather obtain long-term "probabilistic" correspondences using re-occurring regions across distant frames. This avoids the inherent uncertainties of explicit optical flow estimation, whose errors tend to accumulate over time. Similarly, MRF-based video segmentation methods [5, 6, 9, 13] tend to propagate information only *locally* in space-time. Their temporal links are based on optical-flow, whose rapidly accumulated errors induce weak (often zero) weights between related parts in faraway frames. The segmentation performance of video-MRF methods thus strongly depends on the quality of their initial fg/bg data term. However, fg/bg initializations tend to be very noisy, whether based on mining moving object proposals [5, 6, 13], or based on motion saliency maps [9] (especially in unconstrained low-quality videos). Therefore, current video segmentation methods encounter difficulties in such challenging videos. In contrast, our *non-local* consensus voting allows us to start with very 'noisy' fg/bg votes, and clean them rapidly according to 'consensus voting' of distant re-occurring regions.

Qualitative and quantitative experiments indicate that our approach outperforms current state-of-the-art methods. Some visual examples can be found in Fig. 2. **Full videos can be found on our project website www.wisdom.weizmann.ac.il/~vision/NonLocalVideoSegmentation.html.** Empirical comparisons on the SegTrack Dataset [12] can be found in the paper.

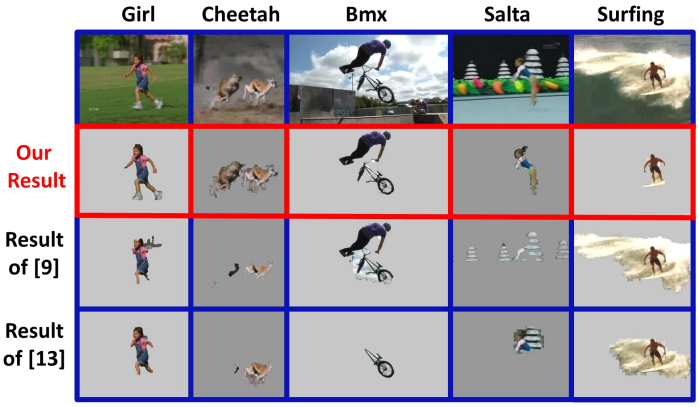


Figure 2: **Visual comparison of results.** Visual comparisons to [9, 13] using their publicly available code. The 3 first sequences are from the SegTrack dataset and the rest are new challenging sequences. For 'Bmx' and 'Salta', we show results of [13] using object selection without Grab-Cut (whereas for all other sequences with Grab-Cut), since these settings gave best results for [13]. See full videos on our Project Website (link in the text).

[1] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[2] J. Costeira and T. Kanande. A multi-body factorization method for motion analysis. In *ICCV*, 1995.

[3] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR*, 2009.

[4] K. Fragkiadaki and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012.

[5] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[6] T. Ma and L. J. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012.

[7] P. Ochs and T. Brox. Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011.

[8] P. Ochs and T. Brox. Higher order motion models and spectral clustering. In *CVPR*, 2012.

[9] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[10] S. R. Rao, R. Tron, R.Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *CVPR*, 2008.

[11] J. Shi and J.Malik. Motion segmentation and tracking using normalized cuts. In *ICCV*, 1998.

[12] D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC*, 2010.

[13] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013.

Embedding Geometry in Generative Models for Pose Estimation of Object Categories

Michele Fenzi
http://www.tnt.uni-hannover.de/staff/fenzi
Jörn Ostermann
http://www.tnt.uni-hannover.de/staff/ostermann

Institut für Informationsverarbeitung (TNT)
Leibniz Universität Hannover
Hannover, Germany

Pose estimation for object classes is central in many Computer Vision tasks. Many approaches have been proposed to estimate the pose of an unknown object from a given category, and those based on local features have shown to be very effective. While some use 3D information obtained through CAD models [4] or 3D reconstructions [2], others have shown that coupling feature regression and view labeling efficiently solves this task [1, 5]. However, they rely solely on the discriminative power of local features, and this is problematic if objects have similar appearance in different views, as Figure 1 shows. To handle these situations they need to resort to external coarse-grained pose estimators for disambiguation.

We propose a method that solves this problem by integrating feature regression and graph matching in a unified probabilistic framework. The former predicts the descriptor of each patch in a query pose, while the latter evaluates the geometrical consistency between pairs of matches. As a consequence, our approach does not resort to external pose pre-processing and in addition experimentally shows to be more accurate in comparison. This permits to avoid any initial hard decision, postponing it to a later stage when more data is available.

Feature regression allows to treat pose estimation as a continuous problem, unlike most methods that provide only discrete values for the pose [3, 4]. Graph matching permits to *softly* align the unknown object to the class model, bringing additional consistency and precision to the solution. In a nutshell, our method retains the benefits of regression-based methods, like continuity and generality, while favoring geometrically consistent results through graph matching.

Our feature regression method leverages [1]. Regression functions model feature descriptors as a function of the pose. Given a patch i, $t^i = \{(f_1^i, \alpha_1^i), (f_2^i, \alpha_2^i), \ldots, (f_n^i, \alpha_n^i)\}$, *i.e.*, t^i is a set of feature descriptors f_j^i labelled by their corresponding viewing angle α_j^i. For each t^i, a generative feature model F^i is defined as a linear combination of Gaussian kernels centered at the training poses,

$$F^i(\alpha) = \sum_{j=1}^{n} G(\alpha, \alpha_j^i)\mathbf{w}_j^i,$$

where $\mathbf{w}_j^i$ are estimated from t^i, and G measures the distance between two viewing angles. The class model is built by grouping all tracks from all class instances on the basis of their similarity in descriptor and pose space through spectral clustering.

At run time, query features are matched against a set of model representatives, which are the cluster centers in descriptor space. The nearest neighbor matching in [1] is prone to ambiguities occurring with similar views. Graph matching permits to favor geometrically consistent poses by exploiting the inherent spatial ordering of the features.

According to the graph matching paradigm, each feature set is interpreted as an attributed graph defined by $G = (V, E, A)$, where V is the set of vertices, E is the set of edges and A is an attribute matrix. We consider all test features as nodes of the test graph G and a subset of the model features as nodes of the model graph G'. Each entry A_{mn} represents some *relationship* between vertices $m, n \in V$. We defined $A_{mm} = f_m$, where f_m is the feature descriptor, and $A_{mn} = (\alpha_{mn}, r_{mn})$, where α_{mn} is the angle between the x-axis and the directed segment P_{mn} connecting the locations of features f_m and f_n, r_{mn} is the length of P_{mn}. $A'_{m'n'}$ is similarly defined.

We search for a mapping $M = \{(m, m') | m \in V, m' \in V'\}$ of the vertices that best respects the original attributes by maximizing the score

$$S = \sum_{(m,m')\in M,(n,n')\in M} g(A_{mn}, A'_{m'n'}),$$

where g evaluates the attribute similarity. If M is expressed as a binary vector $\mathbf{x}$, such that $x_{mm'} = 1$ if $(m, m') \in M$, the problem solution is

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} S = \arg\max_{\mathbf{x}} \mathbf{x}^T \mathbf{W} \mathbf{x}, \qquad \text{s.t. } x_{mn} \in \{0, 1\} \text{ and } \mathbf{C}\mathbf{x} = \mathbf{b},$$

Figure 1: Query features (left) are matched to the class model represented by the two rightmost images. If matching is based on descriptor distance and absolute spatial distance between features, ambiguity still remains. If oriented distances are considered, the correct configuration is favored.

where $\mathbf{W}$ is a matrix such that $W_{mm',nn'} = g(A_{mn}, A'_{m'n'})$. $\mathbf{C}\mathbf{x} = \mathbf{b}$ is a set of linear constraints that may be imposed on the solution. Diagonal entries in $\mathbf{W}$ are defined in terms of the descriptor distance, so that a high entry is assigned to feature pairs close in descriptor space; off-diagonal entries are defined in terms of the absolute angular distance and the Euclidean distance ratio of the corresponding segments. Therefore, a high entry is assigned to feature pairs whose locations are geometrically consistent in orientation and distance.

By relaxing the integer quadratic problem, the solution $\mathbf{x}^*$ is the principal eigenvector of $\mathbf{W}$. As $\mathbf{W}$ has only non-negative entries, all entries in $\mathbf{x}^*$ are in $[0, 1]$, and the solution can be interpreted in probabilistic terms.

If $p(\alpha, c|f) = p(\alpha|f, c)p(c|f)$ expresses the likelihood of observing feature f from viewpoint α and c being the correct match ($f \sim c$), then the best pose and matching for the query feature set $\mathcal{F} = \{f\}_{q=1}^{Q}$ and model set $\mathcal{C} = \{c\}_{r=1}^{N}$ is

$$(\alpha^*, c^*) = \arg\max_{(\alpha,c)} \sum_{(q,r):f_q\sim c_r} p(\alpha|f_q, c_r)p(c_r|f_q),$$

where $p(\alpha|f, c)$ is expressed in terms of the generative feature model and $p(c|f)$ in terms of the graph matching results. As $\|\mathbf{x}\| = 1$ and $x_{fc}^* \in [0, 1]$, the square of each score can be interpreted as a probability.

Method	MAE [°] (90^{th} percentile)	MAE [°]
Ozuysal et al. [3]	-	46.48
Torki et al. [5]	19.40	33.98
Fenzi et al. [1]	14.51	31.27
Ours	**12.67**	**23.38**

Table 1: EPFL dataset [3].

Experiments on two car datasets show that our approach outperforms state-of-the-art algorithms by 25%, as Table 1 shows. Even when the pose classifier is almost perfect, our method not only recovers the correct orientation over the whole pose range, instead of the smaller correct interval given by the classifier, but it is also more accurate.

The increase in performance is due to the higher capability of our algorithm to solve view-problematic situations as well as to an overall additional accuracy given by the introduction of geometric context in the process.

[1] M. Fenzi, L. Leal-Taixé, B. Rosenhahn, and J. Ostermann. Class Generative Models based on Feature Regression for Pose Estimation of Object Categories. In *CVPR*, 2013.

[2] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-Aware Object Detection and Pose Estimation. In *ICCV*, 2011.

[3] M. Özuysal, V. Lepetit, and P. Fua. Pose Estimation for Category Specific Multiview Object Localization. In *CVPR*, 2009.

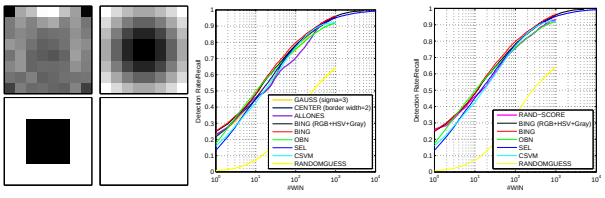
[4] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3D^{2}PM - 3D Deformable Part Models. In *ECCV*, 2012.

[5] M. Torki and A. M. Elgammal. Regression from Local Features for Viewpoint and Pose Estimation. In *ICCV*, 2011.

Cracking BING and Beyond

Qiyang Zhao
zhaoqy@buaa.edu.cn
Zhibin Liu
liuzhibin@nlsde.buaa.edu.cn
Baolin Yin
yin@nlsde.buaa.edu.cn

State Key Laboratory of
Software Development Environment,
Beihang University



(a) (b)

Figure 1: Experiments on templates in BING: (a) four learned/hand-tuned templates and their performances. (b) performance of RAND-SCORE.

The problem, *generic objectness proposal*, aims to reduce the candidate windows for object detection tasks. The popular evaluation criterion for related methods is detection-rate/windows-amount(*DR-#WIN*), where *DR* is the percentage of groundtruth objects covered by proposal windows. An object is considered "covered" by a window only if the strict PASCAL-overall criterion [3] is satisfied (the intersection of a proposal window and the object rectangle is not smaller than half of their union, so we call it "0.5-criterion" for short). Under the *DR-#WIN* evaluation framework, BING [2] in CVPR2014, obtains the best performance on the VOC2007 test set. It recalls 96.2% objects with only 1,000 proposal windows. The more surprising is the method is totally a realtime one.

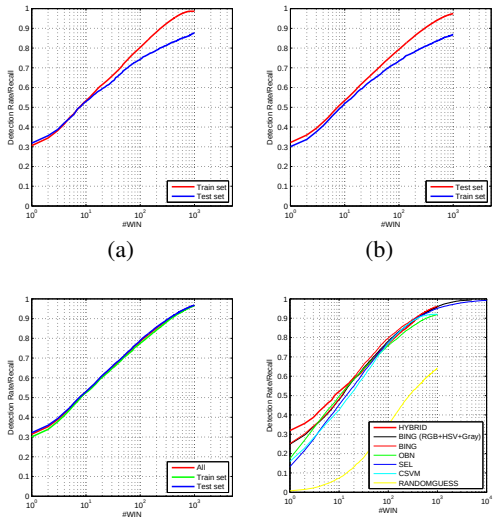
The authors of BING suggest that, after being resized to a fixed size (8×8), almost all annotated rectangle regions share a common characteristics in gradients [2]. This commonness is captured by a template W learned from training images with a linear SVM. Besides this, the subtle differences between diverse width/height configurations are captured in a re-weighting model. Therefore BING consists of two stages: calculating W in stage I, and learning the re-weighting model in stage II. Furthermore, BING uses smart bitwise operations to calculate the inner product of W and candidate windows, so to improve the efficiency.

We designed several templates by hands to substitute W, to verify whether templates play a key role in BING. These templates become less correlated to W in turn, but their performances on VOC 2007 test set are very close, see Fig.1.a. Next we discarded any templates and directly assigned the scores of stage I with uniformly random values (we call this method RAND-SCORE). Surprisingly, the performance of RAND-SCORE is even very close to BING, as shown in Fig.1.b. It is clear that these templates do not have as strong significance as suggested in [2]. Then what on earth makes BING performing so well?

To get the deep insight, we finished a theoretical analysis from the view of combinatorial geometry. We try to construct a small set of windows to "cover" all legal rectangles (we call it a *full cover set*). This is an atypical covering problem in combinatorial geometry [1]. We proposed four lemmas to solve it in the full paper. In conclusion, for an image of the width M and height N, we can use $s(i,j)$ windows of the width $2^i \cdot \sqrt{2}$ and height $2^j \cdot \sqrt{2}$ to cover all $2^i \le w \le 2^{i+1}, 2^j \le h \le 2^{j+1}$ rectangle regions, where $s(i,j) = \lceil \frac{M-2^i}{\left(1-\frac{\sqrt{2}}{2}\right)\cdot 2^i \cdot \sqrt{2}} \rceil \cdot \lceil \frac{N-2^j}{\left(1-\frac{\sqrt{2}}{2}\right)\cdot 2^j \cdot \sqrt{2}} \rceil$. Suppose the image size is $M = 2^m, N = 2^n$, and the object rectangles' widths and heights start from 2^k, then the amount of all windows in our *full cover set* is

$$\sum_{i=k}^{m-1}\sum_{j=k}^{n-1} s(i,j) = \sum_{i=1}^{m-k}\sum_{j=1}^{n-k} \lceil \frac{2^i-1}{\sqrt{2}-1} \rceil \cdot \lceil \frac{2^j-1}{\sqrt{2}-1} \rceil = O(2^{-2k} \cdot MN) \qquad (1)$$

Particularly, when the widths/heights of all object rectangles are at least 16, the amount is 19,600. While on the restriction of 32, we need only 4,225 windows. These amounts are far less than what people imagined before. We call it the Achilles' heel of the *DR-#WIN* evaluation framework. Recall that in BING, the widths/heights of proposal windows are doubled each time, in the same way as in Lemma 3.1-3.4. Its non-max suppression step, 0.25 relative to the normalized size 8, is very close to the step $(1-\frac{\sqrt{2}}{2}) \approx 0.29$ in Lemma 3.2. These two settings meet our analysis well and bring the success to BING.



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

Figure 2: Performance of greedy scheme and hybrid scheme: (a) cover set of training images, and its performance on test images; (b) cover set of test images, and its performance on training images; (c) cover set of all images, and its performance on two sets respectively; (d) comparison of hybrid scheme with other methods.

In real applications, we should pay more attention to those "hot" locations/sizes instead of all possibilities. We designed a greedy scheme to pick the "hottest" window in each round to construct an identical cover set for all images. We also proposed a hybrid scheme to address the huge difference between low-probability sample spaces of the training and test sets. With the increase of the number of proposal windows, we replace the windows in the greedy set with those of RAND-SCORE with increasing probabilities.

In our experiments, our greedy scheme performs considerably well: all *full cover sets* are reduced to about 1,000 windows, and the first windows have 0.3+ *DR*'s in all experiments. In most time, the *DR*'s of our hybrid scheme are higher than OBN and CSVM, and close to SEL and BING. It recalls 95.68% objects with 1000 proposal windows. Especially, its *DR*'s are 13.99% $\sim$ 40.29% (relatively) higher than all other methods in average on the first ten windows. At last, the time consumptions are all nearly zero because the major computations are to resize proposal windows for specific images.

To sum up, what can we benefit from the two schemes for object detection researches? We argue it needs a bigger picture to answer this question because it depends on whether the 0.5-criterion is effective and objective. If the 0.5-criterion is still adopted in future, the baseline should be RAND-SCORE or our hybrid scheme instead of random guesses. Both of them bring more challenges to future researches.

[1] Pach J. and Agarwal P. *Combinatorial Geometry*. John wiley & Sons, ISBN: 9780471588900, 1995.

[2] Cheng M. M., Zhang Z. M., Lin W. Y., and Torr P. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proc. CVPR*, 2014.

[3] Everingham M., Van Gool L., Williams C. K. I., Winn J., and Zisserman A. The pascal visual object classes (voc) challenge. *IJCV*, 88(2): 303-338, 2010.

How good are detection proposals, really?

Jan Hosang
http://mpi-inf.mpg.de/~jhosang
Rodrigo Benenson
http://mpi-inf.mpg.de/~benenson
Bernt Schiele
http://mpi-inf.mpg.de/~schiele

MPI Informatics
Saarbrücken, Germany

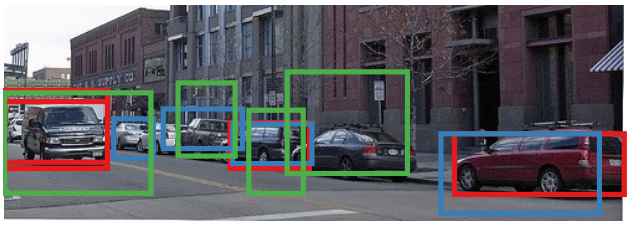


Figure 1: Each colour indicates a different set of detection proposals. How to evaluate which method provides the best proposals?

Object detection is traditionally instantiated in the well known "sliding window" paradigm where a classifier is evaluated over an exhaustive list of positions, scales, and aspect ratios. This approach evaluates the classifier at about $\sim 10^6$ different locations. To alleviate computation pressure by avoiding exhaustive search, recent detectors delegate the selection of candidate detections to a pre-processing step. If these class-agnostic candidate detectors can achieve high recall with $\sim 10^4$ or less windows, significant speed-ups can be achieved, enabling the use of more sophisticated classifiers.

Current top performing Pascal VOC object detectors employ detection proposals to guide the search for objects, thereby avoiding exhaustive sliding window search across images. Despite the popularity of detection proposals, it is unclear which trade-offs are made when using them during object detection. We provide an in depth analysis of ten object proposal methods (from 2009 to 2014) along with four baselines regarding: a) ground truth annotation recall (on Pascal VOC 2007 and ImageNet 2013), b) repeatability, and c) impact on DPM detector performance. See table 1. Our findings show common weaknesses of existing methods, and provide insights for practitioners seeking to choose the most adequate method for their application.

Repeatability We introduce the notion of repeatability which captures how much a detection proposal method is affected by different image perturbations. For this analysis we compute how well a method repeats the selection of candidates after applying an image transformation (see figure 2 for perturbation examples). We argue that repeatability is important when a detector uses candidate detection for negative mining, as it requires the distribution of negative windows to be very similar between training and test set. Our results indicate that repeatability seems to be an issue for most methods. Even very small changes cause most methods to have a strong drop in repeatability.

Recall Different papers evaluate based on recall at different operating points. We give a full picture evaluation regarding recall of ground-truth bounding boxes, and establish common ground for a proper comparison between different methods. To this end we analyse the recall as a function of both the number of candidates and the localisation quality. Recall is important because objects lost by the proposal method will not be recovered by the detector. Our results show that a handful of methods dominate quality in multiple settings (see figure 3). The ImageNet experiments show that, despite being tuned on Pascal VOC, current proposal methods have excellent generalization towards the larger (10×) set of ImageNet classes, indicating that they are true "objectness" measures.

Detection Finally, we do experiments regarding the effect of selective search over detection quality. As an initial approach, we filter the detections of a pre-trained DPM detector method using different proposal methods, to emulate having done the detections directly from these windows. Results show that detection quality is directly related to the accuracy and recall level of the underlying detection proposals method.

Our paper provide detailed result curves and tables, summarising more than 500 experiments over different data sets, totalling to more than 2.5 months of CPU computation.

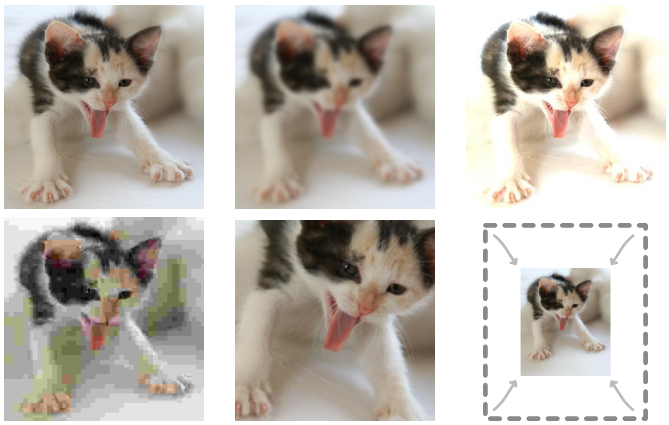


Figure 2: Example of the image perturbations considered for the repeatability experiments. Top to bottom, left to right: original, then blur, illumination, JPEG artefact, rotation, and scale perturbations.
How stable are detection proposals to slight changes in the input image?

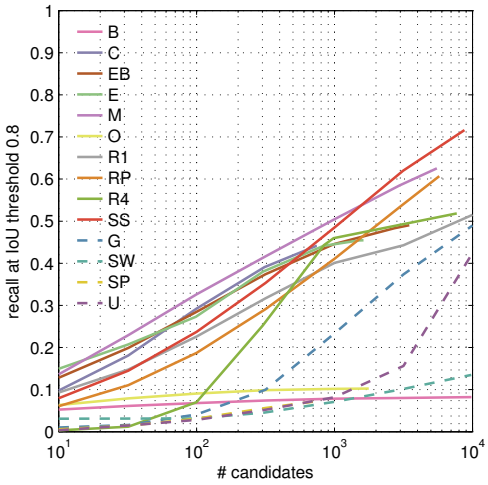


Figure 3: Recall versus number of proposals on the Pascal VOC 2007 test set for IoU above 0.8. At varying number of windows the best method to pick changes, what about when considering computational cost too?

Method		Time	Repeatability	Recall	Detection
Objectness	O	3	·	⋆	·
CPMC	C	250	-	⋆⋆	⋆
Endres2010	E	100	-	⋆⋆	⋆⋆
Sel.Search	SS	10	⋆⋆	⋆⋆⋆	⋆⋆
Rahtu2011	R1	3	·	·	⋆
Rand.Prim	RP	1	⋆	⋆	⋆
Bing	B	0.2	⋆⋆⋆	⋆	·
MCG	M	30	⋆	⋆⋆⋆	⋆⋆
Ranta.2014	R4	10	⋆⋆	·	⋆
EdgeBoxes	EB	0.3	⋆⋆	⋆⋆⋆	⋆⋆
Uniform	U	0	·	·	·
Gaussian	G	0	·	·	⋆
SlidingWindow	SW	0	⋆⋆⋆	·	·
Superpixels	SP	1	⋆	·	·

Table 1: Overview of detection proposal methods.
Time is in seconds. Repeatability, quality, and detection rankings are provided as rough qualitative overview; "-" indicates no data, "·", "⋆", "⋆⋆", "⋆⋆⋆" indicate progressively better results. See paper's text for details and quantitative evaluations.

Part Context Learning for Visual Tracking

Guibo Zhu
gbzhu@nlpr.ia.ac.cn
Jinqiao Wang
jqwang@nlpr.ia.ac.cn
Chaoyang Zhao
chaoyang.zhao@nlpr.ia.ac.cn
Hanqing Lu
luhq@nlpr.ia.ac.cn

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

Context information is widely used in computer vision for tracking arbitrary objects[1, 3, 4]. Global context cannot deal with the object deformation problem, while the local part context interactions are relatively stable. When the target appearance changes gradually, the intrinsic property of internal interaction between the parts inside object and context interaction between object and background are relatively stable in spatio-temporal 3D space of tracking.

To explore the structure property and stable relationship for overcoming complex environments, we propose a novel part context tracker. The Part Context Tracker (PCT) consists of an appearance model, an internal relation model and an context relation model. The internal relation model formulates the temporal relations of the object itself or the in-object parts themselves and the spatio-temporal relations between the object and in-object parts. The context relation model constructs the spatio-temporal relations between the in-object parts and the context parts and the temporal relations of the context parts themselves. Hence the physical properties and the appearance information are considered in the optimization process through parts and relations. The contributions are as follows:

(1) We first propose a unified context framework which formulates the single object tracking as a part context learning problem.

(2) The in-object parts and context parts are selected so that we not only pay attention to the appearance of object, but also focus on the relations among the object, the in-object parts and the context parts.

(3) A simple yet robust update strategy using median filter is utilized, thereby enabling the tracker to deal with appearance change effectively and alleviate the drift problem.

Our framework not only models the object with in-object parts, but also incorporates the interaction between the object and background with context parts. The deformable configuration [2, 5] together with the temporal structure of these parts are also considered in.

In Fig. 1, with the object bounding box as the root R, the in-object parts I are defined as the parts selected inside R, which covers part of the object appearance. The context parts C are selected from the overlapping area between the object and the background. For a target with K in-object parts and M context parts, the configuration is denoted as $B = (B_0, B_1 ... B_K, B_{K+1}, ..., B_{K+M})$. Where B_0 stands for the target bounding box R, $(B_1, .., B_K) \in I$ are the K in-object part boxes, and $(B_{K+1}, ..., B_{K+M}) \in C$ are the M context part boxes. The corresponding features of the root and parts are represented as $X = (x_0, ..., x_K, x_{K+1}, ..., x_{K+M})$. In a word, our framework models the object with three components:

$$M = M_A + M_I + M_C, \tag{1}$$

where M_A, M_I and M_C are the appearance model, the internal relation model and the context relation model respectively.

For online tracking, an appearance model is essential. It represents the intrinsic property of one object or the discriminative information between the object and background. To better mine the information, we factorize the appearance model M_A as Eq. (1):

$$M_A = A_R + A_I + A_C \tag{2}$$

where A_R, A_I and A_C are the global root appearance model, in-object parts appearance model and context parts model separately.

In addition to the appearance model, all spatio-temporal relative stable relations between the object and its corresponding parts frame-to-frame should be utilized in tracking. Therefore we design the internal relation model to formulate the interactions between root and the in-object parts, which includes the spatial constrains and the temporal constrains between them, we define M_I as:

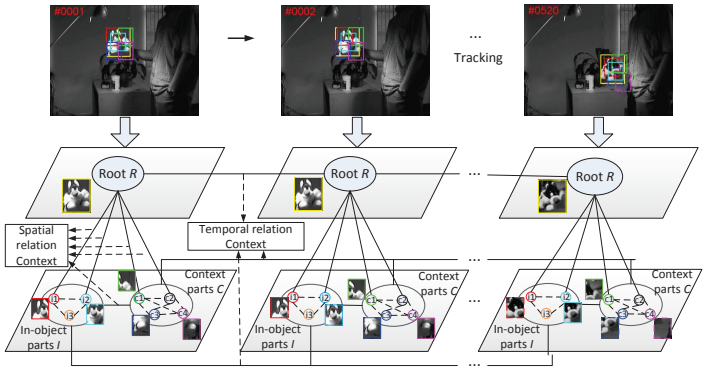


Figure 1: Illustration of our Part Context tracking framework using the "sylvester" video.

$$M_I = S_I + E_R + E_I \tag{3}$$

where S_I, E_R, and E_I are spatial relation between root and in-object parts, temporal relation between root and their historical root, and temporal relation between in-object parts and their historical information respectively.

Except internal relations inside the object, some information in latent intersection area between the object and background is neglected by previous works, such as the partial contour and the object are consensus in motion. To make full use of the information, we formulate the context relation model to express the interactions between root and the context parts, which also includes the spatial and temporal constrains between them. Similar to Eq. (3), we describe the context relation model mathematically as:

$$M_C = S_C + S_{C,I} + E_C \tag{4}$$

where S_C, $S_{C,I}$ and E_C denote spatial relation between root and context parts, spatial relation between in-object parts and context parts, and temporal relation between context parts and their historical information.

Implementation of this method by model definition is described in the paper, as are the details of the model optimization in inference and learning. Our conclusion is that one tracker consists of an appearance model, an internal relation model and an context relation model in a maximum margin structured learning framework, which is robust to certain conditions of occlusion, illumination and out-of-view.

[1] T.B. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, pages 1177–1184. IEEE, 2011.

[2] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[3] L. Wen, Z. Cai, Z. Lei, D. Yi, and S.Z. Li. Online spatio-temporal structural context learning for visual tracking. In *ECCV*, pages 716–729. Springer, 2012.

[4] K. Zhang, L. Zhang, M. H. Yang, and D. Zhang. Fast tracking via spatio-temporal context learning. *arXiv preprint arXiv:1311.1939*, 2013.

[5] L. Zhang and L. van der Maaten. Preserving structure in model-free tracking. *IEEE-TPAMI*, 36(4):756–769, 2014.

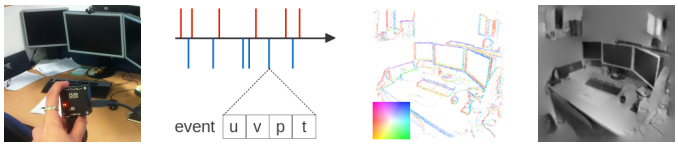
Simultaneous Mosaicing and Tracking with an Event Camera

Hanme Kim[1]
hanme.kim@imperial.ac.uk
Ankur Handa[2]
ah781@cam.ac.uk
Ryad Benosman[3]
ryad.benosman@upmc.fr
Sio-Hoi Ieng[3]
sio-hoi.ieng@upmc.fr
Andrew J. Davison[1]
a.davison@imperial.ac.uk

[1] Department of Computing, Imperial College London, London, UK
[2] Department of Engineering, University of Cambridge, Cambridge, UK
[3] INSERM, U968, Paris, F-75012, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 968, Institut de la Vision, Paris, F-75012, France; CNRS, UMR_7210, Paris, F-75012, France

An event camera is a silicon retina which outputs not a sequence of video frames like a standard camera, but a stream of asynchronous spikes, each with pixel location, sign and precise timing, indicating when individual pixels record a threshold log intensity change (positive or negative). By encoding only image change, it offers the potential to transmit the information in a standard video but at vastly reduced bitrate, and with huge added advantages of very high dynamic range and temporal resolution.

In this paper, we show for the first time that an event stream from an event camera (e.g. Figure 1(b)), with no additional sensing, can be used to track accurate camera rotation while building a persistent and high quality mosaic of a scene (e.g. Figure 1(d)) which is super-resolution accurate and has high dynamic range; we use the first commercial event camera [1] (Figure 1(a)). Our method involves parallel camera rotation tracking and template reconstruction from estimated gradients (e.g. Figure 1(c)), both operating on an event-by-event basis and based on probabilistic filtering.



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

Figure 1: Proposed algorithm: (a) scene and DVS camera; (b) event stream; (c) estimated gradient map; (d) reconstructed intensity map.

In our particle filter based tracking, the posterior density function at time t is represented by N particles, each of which is a set consisting of a hypothesis of the current state $\mathrm{R}_i^{(t)} \in \mathbf{SO}(3)$ and a normalised weight $w_i^{(t)}$. As a new event is received, all particles are perturbed based on a constant position motion model; we perturb the current $\mathfrak{so}(3)$ vector on the tangent plane with Gaussian noise independently in all three axes and reproject it onto the $\mathbf{SO}(3)$ unit sphere to obtain the corresponding predicted rotation. The noise is the predicted change the current rotation might have undergone since the previous event was generated. The weights of these perturbed particles are then updated through the measurement update step which applies Bayes rule to each particle and normalised subsequently. A measurement given an event, the current state $\mathrm{R}_i^{(t)}$ and the previous state $\mathrm{R}_i^{(t-\tau_c)}$, where τ_c is the time elapsed since the previous event at a specific pixel, is a log intensity difference between the corresponding intensity map positions which is to be used to calculate the likelihood for each particle, essentially asking 'how likely was this event relative to our mosaic given a particular hypothesis of camera pose?'. For the next measurement update and the reconstruction step, a particle mean pose is saved for each pixel.

We now turn to incrementally improving an estimate of the intensity mosaic. This takes two steps; pixel-wise incremental Extended Kalman Filter (EKF) estimation of the log gradient at each template pixel, and interleaved Poisson reconstruction to recover absolute log intensity. Each pixel of the gradient map has an independent gradient estimate and covariance matrix. Now, we want to improve a gradient estimate based on a new incoming event and a tracking result using the pixel-wise EKF. Assuming, based on the rapidity of events, that the gradient $\mathbf{g}$ in the template and the camera velocity $\mathbf{v}$ can be considered locally constant, we now say $(\mathbf{g}\cdot\mathbf{v})\tau_c$ is the amount of log grey level change that has happened since the last event. Therefore, if we have an event camera where log intensity change C should trigger an event, the brightness constancy tells us $\left(\mathbf{g}^{(t)}\cdot\mathbf{v}^{(t)}\right)\tau_c = \pm C$ which leads to define a measurement $z^{(t)} = \frac{1}{\tau_c}$ and its measurement model $h^{(t)} = \frac{\mathbf{g}^{(t)}\cdot\mathbf{v}^{(t)}}{C}$. The gradient estimate and the uncertainty covariance matrix are then updated using the standard EKF equations. Essentially, each new event which lines up with a particular template pixel improves our gradient estimate in the direction parallel to the camera motion over the scene at that pixel while we learn nothing about the gradient in the direction perpendicular to the motion. Finally, we reconstruct the log intensity of the image whose gradients across the whole image domain are close to the estimated gradients in a least squares sense inspired by [2].

We conducted the spherical mosaicing reconstruction in both indoor and outdoor scenes as shown in Figure 2. Also, we show the potential for reconstructing high resolution and dynamic range scenes from very small camera motion as shown in Figure 3.

Figure 2: Spherical mosaicing for indoor and outdoor scenes. The overlaid boxes represent the field of view of the event camera.

(a) (b)

Figure 3: (a) Comparison of a reconstructed high resolution image and a down sampled normal camera image; (b) comparison of a reconstructed high dynamic range image and a normal CCD camera image.

We believe these are breakthrough results, showing how joint sequential and global estimation permits the great benefits of an event camera to be applied to a real problem of mosaicing, and hopefully opening the door to similar approaches in dense 3D reconstruction and many other vision problems.

[1] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128×128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 43(2):566–576, 2008.

[2] J. Tumblin, A. Agrawal, and R. Raskar. Why I want a Gradient Camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

Deformable Template Tracking in 1ms

David Joseph Tan[1]
tanda@in.tum.de
Stefan Holzer[1]
holzers@in.tum.de
Nassir Navab[1]
navab@in.tum.de
Slobodan Ilic[2]
Slobodan.Ilic@siemens.com

[1] CAMP
Technische Universität München
Munich, Germany

[2] Siemens AG
Munich, Germany

The goal of template tracking is to estimate the transformation parameters that define the motion of the planar template. Typically, the transformation parameters encode linear transformation based on homographies with 8 degrees of freedom, which allows them to track rigid motions of the template. However, when it comes to tracking surfaces that undergo non-rigid deformations, deformable transformations with higher degrees of freedom must be used.

Most works on deformable template tracking rely on feature points [1, 5, 6, 7, 8]. It follows the traditional algorithm where it detects feature points on the current frame; finds feature point correspondences between the current frame and the template; remove outliers from these correspondences; and, estimate the deformation. However, since deformable models have a higher degrees of freedom, it becomes more difficult to detect outliers. As a result, it requires a longer runtime. Therefore, this paper aims to address the problem of real-time deformable template tracking.

Instead of using tracking-by-detection approaches with feature points, our work focuses on a frame-to-frame tracking approach with a dense pixel arrangement on the template. Hence, to track the template, we use the linear predictor $\mathbf{A}$ which establishes a linear relation between the vector of image intensity differences $\delta\mathbf{i}$ of a template and the corresponding template transformation parameters $\delta\mu$, which is written as [3]:

$$\delta\mu = \mathbf{A}\delta\mathbf{i}. \tag{1}$$

The main benefit of using dense pixel intensities is that a lack of a large number of feature points is compensated by the dense pixel information and, thus, allows tracking of less textured surfaces such as faces.

Up to this work, linear predictors have only been used to handle linear transformations such as homographies to track planar surfaces. In this paper, we introduce a method to learn non-linear template transformations that allows us to track surfaces that undergo non-rigid deformations. These deformations are mathematically modelled using 2D Free Form Deformations (FFD) with cubic B-Splines [2, 4]. It uses control points that are uniformly arranged around the template such that the deformation of the template is modelled by the displacement of the control points. In this way, the transformation parameters in $\delta\mu$ is defined by the displacements of the control points.

Linear predictors are learned using a dataset of n_ω images. Each image is a deformed version of the template, where random movements are induced on its control points. These movements correspond to the change in the parameter vector $\delta\mu$. Using FFD, the location of the sample points are deformed that creates the image intensity differences $\delta\mathbf{i}$. Therefore, we can concatenate the vectors from $\{(\delta\mathbf{i}_\omega, \delta\mu_\omega)\}_{\omega=1}^{n_\omega}$ to construct the matrices $\mathbf{Y} = \left[\delta\mu_1, \delta\mu_2, \cdots, \delta\mu_{n_\omega}\right]$ and $\mathbf{H} = \left[\delta\mathbf{i}_1, \delta\mathbf{i}_2, \cdots, \delta\mathbf{i}_{n_\omega}\right]$ with the relation $\mathbf{Y} = \mathbf{A}\mathbf{H}$, and learn the linear predictor $\mathbf{A}$ using [3]:

$$\mathbf{A} = \mathbf{Y}\mathbf{H}^\top \left(\mathbf{H}\mathbf{H}^\top\right)^{-1}. \tag{2}$$

The simplicity of our approach allows us to track deformable surfaces at extremely high speed of approximately 1 ms per frame with a single core of the CPU, which has never been shown before.

To evaluate our algorithm, we perform an extensive analysis of our method's performance on synthetic and real sequences with different types of surface deformations. In addition, we compare our results from the real sequences to the feature-based tracking-by-detection method [5], and show that the tracking precisions are similar but our method performs 100 times faster.

Our *Supplementary Material* includes a video that shows quantitative and qualitative results to demonstrate our tracking performance under different deformations and in low-lighting condition as illustrated in Fig. 1.

Figure 1: These images are exemplary examples of tracking a template.

[1] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 2003.

[2] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through mrfs and efficient linear programming. *Medical Image Analysis*, 2008.

[3] F. Jurie and M. Dhome. Hyperplane approximation for template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[4] S. Lee, G. Wolberg, and S.Y. Shin. Scattered data interpolation with multilevel b-splines. *IEEE Transactions on Visualization and Computer Graphics*, 1997.

[5] J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration and realistic augmentation. *International Journal of Computer Vision*, 2008.

[6] D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. *International Journal of Computer Vision*, 2012.

[7] J. Zhu and M. Lyu. Progressive finite newton approach to real-time nonrigid surface detection. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[8] J. Zhu, S. Hoi, and M. Lyu. Nonrigid shape recovery by gaussian process regression. In *Conference on Computer Vision and Pattern Recognition*, 2009.

Learn++ for Robust Object Tracking

Feng Zheng[1]
cip12fz@sheffield.ac.uk

Ling Shao[1]
ling.shao@ieee.org

James Brownjohn[2]
J.Brownjohn@exeter.ac.uk

Vitomir Racic[3]
v.racic@sheffield.ac.uk

[1] Department of Electronic and Electrical Engineering, The University of Sheffield, UK

[2] College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK

[3] Department of Civil and Structural Engineering, The University of Sheffield, UK

Figure 1: Assuming that the best classifiers for the previous frames are available, which classifiers should be used in the current frame (bottom right)? f_2, f_5 or their combination? Also, when the target moves out of view then comes back, which classifiers are the best to be used? This paper tries to solve these problems in object tracking.

Motivation. Most machine learning algorithms can learn from data that are assumed to be drawn from a fixed but unknown distribution. However, this assumption cannot be valid in case of the tracking problem. Traditional machine learning methods applied to the tracking problem, such as tracking-by-detection approaches [1, 2], will fail when there is a "concept drift" in the non-stationary environment, because the function learnt on a fixed sample set previously collected may not reflect the current state of nature due to a change in the underlying environment [3]. In object tracking, the distribution of samples changes a lot due to the deformation of the object and the change of the background. Especially during the transition between different difficulties (sub-problems), such as from occlusion to varying viewpoints, the samples in the two different situations differ significantly. Thus, the separability of features and classifiers used in previous frames will decrease in the new situation as shown in Fig. 1.

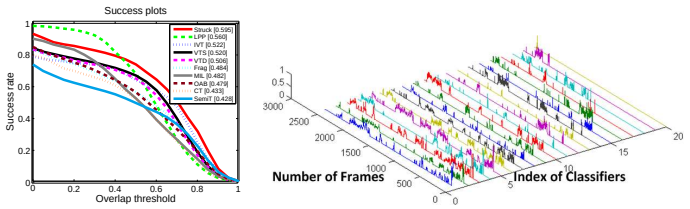
Contributions. Our idea is to build a basic classifier for each sub-problem and these basic classifiers learnt fromdifferent sample sets are independent from each other. In this paper, by enabling and designing these critical and flexible functions, we propose a new Learn++ method for robust and long-term object tracking, named as LPP tracker. LPP tracker dynamically maintains a set of basic classifiers $f_i \in \Omega_e^t$ which are trained sequentially without accessing original data but preserving the previously acquired knowledge. The "concept drift" problems can be solved by adaptively selecting the most suitable classifiers (called the active subset $\Omega_a^t \subset \Omega_e^t$) which are corresponding to the non-zero weights w_i^t. Thus, given the samples x_l^t and their labels y_l^t, the objective function is defined as:

$$\mathbf{w}^t = \arg\min_{w^t} \sum_l L\Big(\sum_{f_i \in \Omega_e^t} w_i^t f_i(x_l^t), y_l^t\Big) + \lambda \left|\left|w^t\right|\right|_0 \qquad (1)$$

where L and λ are the loss function and regularization parameter, respectively.

By using the classifiers that have yielded good performance in recent n frames or in the same situations, the optimal classifier $\mathbf{f}^t$ in the present environment can be fast approximated in a function space linearly spanned by these basic classifiers in the active subset. For each frame, the democratic mechanism can be adopted, where all classifiers should compete with each other to be added into an active subset to suit the present environment. To achieve this goal, four steps are adopted, including reviving old classifiers, training a new one, resampling and evaluating all of them. Therefore, we obtain the hypothesis:

$$\mathbf{f}^t = \sum_{f_i \in \Omega_e^t} \mathbf{w}_i^t f_i \qquad (2)$$



(a) (b)

Figure 2: (a) The success plot and AUC rankings of 10 tracking methods on challenging sequences. (b) The weights for the optimal classifier $\mathbf{f}^t$.

	LPP	Struck	VTS	IVT	VTD	MIL	CT	SemiT
IV	0.932	0.860	**0.957**	0.900	0.888	0.569	0.704	0.463
OPR	**0.858**	0.775	0.754	0.718	0.766	0.670	0.625	0.544
SV	**0.928**	0.816	0.763	0.779	0.771	0.769	0.805	0.449
OCC	**0.772**	0.659	0.723	0.749	0.733	0.583	0.608	0.519
DEF	**0.871**	0.682	0.595	0.674	0.6000	0.618	0.643	0.677
MB	**0.919**	0.776	0.726	0.513	0.710	0.847	0.614	0.293
FM	**0.875**	0.856	0.546	0.459	0.547	0.767	0.571	0.448
IPR	0.861	0.867	0.869	0.819	**0.885**	0.778	0.659	0.489
BC	0.882	**0.912**	0.725	0.769	0.709	0.714	0.594	0.609
Overall	**0.844**	0.817	0.736	0.734	0.664	0.658	0.605	0.552

Table 1: The precision rankings of 10 tracking methods on challenging sequences. Bold numbers denote the best precision scores.

Results. Our experiments follow the setting in [4] and compare with the results of 9 state-of-the-art methods from this report as well. From Fig. 2(a) and Table 1, we can see that, in total, LPP tracker gains six firsts, two seconds and one fourth by the precision ranking, and it gains three firsts, three seconds and two fourths by the AUC ranking. Further investigations are given in Fig. 3, Struck fails when the target starts to move out of view but LPP tracker tackles all the problems. From Fig. 2(b), we can see that the weights are very sparse and just a few members will be selected for each frame.

[1] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Robust object tracking with online multiple instance learning. *IEEE Tansactions on PAMI*, 33(8):1619–1632, 2011.

[2] Sam Hare, Amir Saffari, and Philip H. S. Torr. Struck: Structured output tracking with kernels. In *Proc. ICCV*, 2011.

[3] Matthew Karnick, Metin Ahiskali, Michael D. Muhlbaier, and Robi Polikar. Learning concept drift in nonstationary environments using an ensemble of classifiers based approach. In *Proc. IJCNN*, 2008.

[4] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proc. CVPR*, 2013.

Figure 3: Results on three more challenging sequences between LPP tracker (Red) and Struck (Green).

L_0-Regularized Object Representation for Visual Tracking

Jinshan Pan[1]
sdluran@gmail.com
Jongwoo Lim[2]
jlim@hanyang.ac.kr
Zhixun Su[1]
zxsu@dlut.edu.cn
Ming-Hsuan Yang[3]
mhyang@ucmerced.edu

[1] School of Mathematical Sciences
Dalian University of Technology Dalian, China
[2] Division of Computer Science & Engineering
Hanyang University
Seoul, Korea
[3] Electrical Engineering and Computer Science
University of California at Merced
California, USA

Thursday

Introduction & Motivation: Visual tracking is a highly researched topic in the computer vision community since it has been widely applied in visual surveillance, driver assistant system, and many others. Although much progress has been made in the past decades, designing a practical visual tracking system is still a challenging problem due to numerous challenges in real world.

Very recent efforts have been made to improve this method in terms of both speed and accuracy by using APG algorithm [1] or modeling the similarity between different candidates [6]. The works in [4, 5] point out that the aforementioned methods do not exploit rich and redundant image properties which can be captured compactly with subspace representations. Thus, they propose combining the strength of subspace learning [3] and sparse representation for modeling object appearance. In their work the raw pixel templates used in in [1, 2] are replaced with the orthogonal basis vectors (*e.g.*, PCA basis), and the coefficients for an image are obtained by a least square (LS) method. However, we empirically find that such linear combination of the orthogonal basis vectors sometimes include redundant parts (*e.g.*, background portions), which will interfere with the accuracy of object representation.

We in this paper address this problem by proposing a tracking method based on an L_0 regularized object representation. The L_0 regularized object representation is able to reduce the redundant features while keeping the most important part. Furthermore, the estimation of the L_0 regularized parameters can be efficiently conducted by the proposed APG algorithm.

L_0 Regularized Object Representation: We assume that the target region $\mathbf{y} \in \mathbb{R}^{d\times 1}$ can be represented by an image subspace with corruption,

$$\mathbf{y} = D\boldsymbol{\alpha} + \mathbf{e}, \tag{1}$$

where the columns of $D \in \mathbb{R}^{d\times n}$ are orthogonal basis vectors of the subspace, $\boldsymbol{\alpha}$ is the sparse coefficient vector, and $\mathbf{e}$ represents additive errors modeled by a Laplacian noise.

We propose an L_0 regularized prior to select useful features, which is defined as

$$\min_{\boldsymbol{\alpha},\mathbf{e}} \frac{1}{2}\|\mathbf{y} - D\boldsymbol{\alpha} - \mathbf{e}\|_2^2 + \lambda\|\mathbf{e}\|_1 + \gamma\|\boldsymbol{\alpha}\|_0, \tag{2}$$

where $D^\top D = I$, $\|\cdot\|_0$ denotes the L_0 norm which counts the number of non-zero elements, $\|\cdot\|_2$ and $\|\cdot\|_1$ denote L_2 and L_1 norms, respectively, γ and λ are regularization parameters, and I is an identity matrix. The term $\|\mathbf{e}\|_1$ is used to reject outliers (e.g., occlusions), while $\|\alpha\|_0$ is used to select the useful features. We note that if we set $\gamma = 0$, (2) is reduced to [4].

Analysis on the Effectiveness of L_0 Representation: The benefit of the L_0 norm regularized prior is that it is able to reduce the redundant features while keeping the most important part, thereby facilitating the tracking result.

When there are no errors (*e.g.*, occlusion) in the observation $\mathbf{y}$, i.e., $\mathbf{e} \approx 0$, we can think of L_p regularized error metric in general,

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - D\boldsymbol{\alpha}\|_2^2 + \gamma\|\boldsymbol{\alpha}\|_p^p, \quad \text{where } D^\top D = I, \tag{3}$$

and the solutions for different p are given in the following theorem.

Theorem 1 *Assume that $D \in \mathbb{R}^{d\times d}$ and $D^\top D = I$. The solution of (3) when p is 0 is given by*

$$\boldsymbol{\alpha} = H_{2\gamma}(D^\top \mathbf{y}), \tag{4}$$

when p is 1, the solution is

$$\boldsymbol{\alpha} = \mathcal{S}_\gamma(D^\top \mathbf{y}), \tag{5}$$

and when p is 2, the solution becomes

$$\boldsymbol{\alpha} = \frac{D^\top y}{1+2\gamma}. \tag{6}$$

Here $\mathcal{S}_\theta(x) = \text{sign}(x)\max(|x| - \theta, 0)$, and $H_\theta(x)$ is a hard thresholding operator, which is defined as $H_\theta(x) = x$; if $x^2 > \theta$ and 0 otherwise.

Based on Theorem 1, we have the following corollary.

Corollary 1 *We assume D is redundant and contains all possible basis. Let $\mathbf{u}^*$ denote the non-zero elements of $D^\top y$. If we set $\gamma = \frac{1}{2}\min_i\{|\mathbf{u}_i^*|^2\}$, the solution of L_0 regularized error metric (*i.e.*, (3) when $p = 0$) can exactly recover the data $\mathbf{y}$.*

Figure 1 shows the tracking results by using LS method [4] (i.e., $\gamma = 0$ in (2)), L_0 and L_2 norm under the same dictionary D, respectively. We note that using L_0 regularized method is able to find the good candidate when there exists occlusion, then facilitating the tracking results.

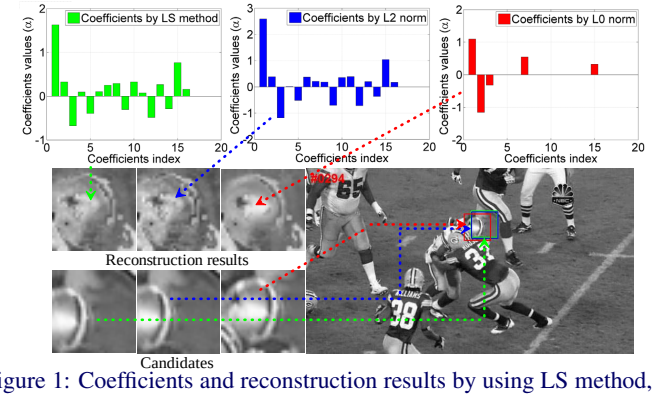


Figure 1: Coefficients and reconstruction results by using LS method, L_0 and L_2 norm under the same dictionary D, respectively. The coefficients by using L_0 norm are more sparse than those by L_2 norm and LS method, and the reconstruction result and the best candidate are also better. The rectangles in the last image represent MAP states for particle filers.

[1] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji. Real time robust ℓ_1 tracker using accelerated proximal gradient approach. In *CVPR*, pages 1830–1837, 2012.

[2] Xue Mei and Haibin Ling. Robust visual tracking using ℓ_1 minimization. In *ICCV*, pages 1436–1443, 2009.

[3] David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3): 125–141, 2008.

[4] Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Least soft-thresold squares tracking. In *CVPR*, pages 2371–2378, 2013.

[5] Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Online object tracking with sparse prototypes. *IEEE TIP*, 22(1):314–325, 2013.

[6] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust visual tracking via structured multi-task sparse learning. *IJCV*, 101(2):367–383, 2013.

You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video

Dima Damen
Dima.Damen@bristol.ac.uk
Teesid Leelasawassuk
Csztl@bristol.ac.uk
Osian Haines
Osian.Haines@bristol.ac.uk
Andrew Calway
Andrew.Calway@bristol.ac.uk
Walterio Mayol-Cuevas
Walterio.Mayol-Cuevas@bristol.ac.uk

Computer Science Department
University of Bristol
Bristol, UK

We present a **fully unsupervised** approach for the discovery of i) task relevant objects and ii) how these objects have been used. A **Task Relevant Object (TRO)** is an object, or part of an object, with which a person interacts during task performance. Given egocentric video from multiple operators, the approach can discover objects with which the users interact, both static objects such as a coffee machine as well as movable ones such as a cup. Importantly, we also introduce the term **Mode of Interaction (MOI)** to refer to the different ways in which TROs are used. Say, a cup can be lifted, washed, or poured into. When harvesting interactions with the same object from multiple operators, common MOIs can be found.

Setup and Dataset: Using a wearable camera and gaze tracker (Mobile Eye-XG from ASL), egocentric video is collected of users performing tasks, along with their gaze in pixel coordinates. Six locations were chosen: kitchen, workspace, laser printer, corridor with a locked door, cardiac gym and weight-lifting machine. The Bristol Egocentric Object Interactions Dataset is publically available [1].

Figure 1: An overview of the locations in the dataset.

Discovering TROs: Given a sequence of images $\{I_1,..,I_T\}$ collected from multiple operators around a common environment, we aim to extract K TROs, where each object TRO_k is represented by the images from the sequence that feature the object of interest . We investigate using appearance, position and attention, and present results using each and a combination of relevant features. For attention, we exploit the high quality and predictive nature of eye gaze fixations.

Results compare k-means clustering to spectral clustering, and propose estimating the optimal number of clusters using the standard Davies-Bouldin (DB) index. Figure 2 shows the best performance for discovering TROs by combining position (relative to a map of the scene) and appearance (HOG features within BoW) over a sliding window $w = 25$, using gaze fixations for attention, spectral clustering and estimating the number of clusters using the Davies-Bouldin (DB) index.

Finding MOIs: Given consecutive images $(I_t, I_{t+1}, I_{t+\rho})$ clustered into the same TRO, a video snippet u_i^k for TRO k is defined as

$$u_i^k = \{\Psi(I_j, \Delta(j), \omega); \quad I_j \in TRO_k; \quad j = t..t+\rho; \quad \rho \geq \xi\} \qquad (1)$$

where Ψ crops a window of size ω from image I_j around $\Delta(j)$, and $\Delta(j)$ is the interpolated gaze at frame j as gaze information is missing in some frames. The collection of all video snippets $U_k = \{u_i^k\}$ shows different ways in which TRO_k was used.

On average, 16.6 video snippets are extracted for each TRO ($\sigma = 7.4$). We cluster U_j, and represent each cluster by the video snippet $\hat{u_j}$ closest to the centre of the cluster μ_j (i.e. mean snippet), as well as the percentage of snippets within that cluster $p(MOI_j)$. We vary the threshold λ to accept $p(MOI_j)$ to produce recall-precision curves. Figure 3 shows an example of the method successfully discovering two MOIs for the 'socket'.

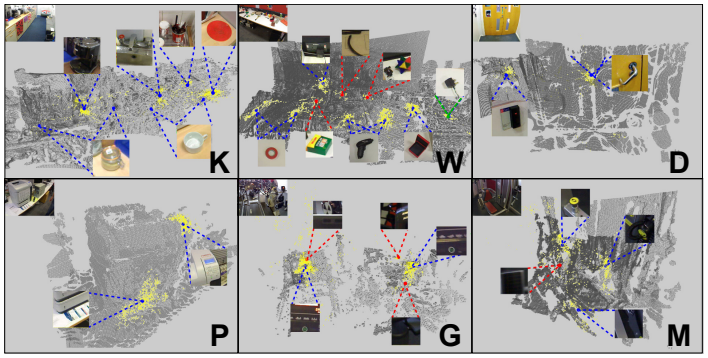


Figure 2: Discovered TROs. An overview of the locations is shown at the top. Blue dots represent true-positive (19 objs), red dots represent false positive (7 objs) and green dots represent false negative (1 obj).

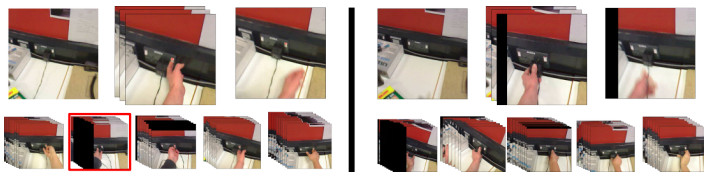


Figure 3: For the 'socket', the two common MOIs ('switching', 'plugging') are found (left & right). The representative *video snippet* is shown (up) with the other snippets in the same cluster (below) - only one snippet is incorrectly clustered (shown in red).

Video Guides: In addition, the approach enables the automatic generation of *help snippets* on how objects have been used before. We showcase video help guides using inserts on a pre-recorded video. A suitable video insert (i.e. MOI snippet) is chosen every time a gazed-at object is first recognised. In this assistive mode, we use the real-time texture-minimal scalable detector [2] due to its light-weight computational load that makes it amendable to wearable systems. Figure 4 shows frames from the help videos and a full sequence is available [3]. Recall that these inserts are *extracted, selected and displayed* fully automatically.

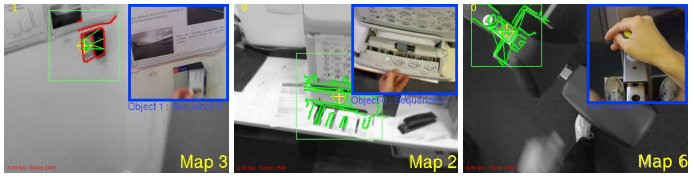


Figure 4: In the assistive mode, when a TRO is detected, video snippet is inserted showing the most relevant common MOI based on the object's current appearance.

[1] http://www.cs.bris.ac.uk/~damen/BEOID/

[2] http://www.cs.bris.ac.uk/~damen/MultiObjDetector.htm

[3] http://www.cs.bris.ac.uk/~damen/You-Do-I-Learn

Hierarchical Cascade of Classifiers for Efficient Poselet Evaluation

Bo Chen[1]
bchen3@caltech.edu
Pietro Perona[1]
perona@caltech.edu
Lubomir Bourdev[2]
lubomir@fb.com

[1] Computation and Neural Systems
California Institute of Technology
California, USA
[2] Facebook AI Research,
Menlo Park, California, USA

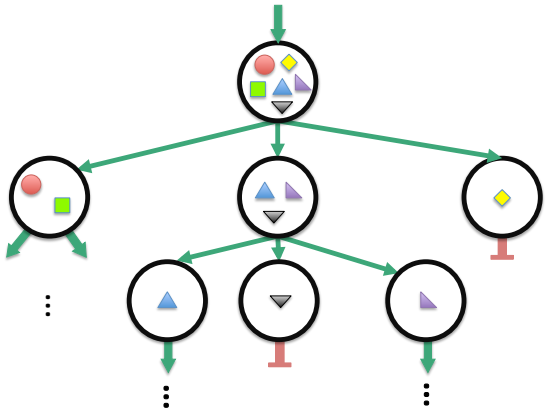


Figure 1: Cascade hierarchy. Each node is a classifier trained to let through examples of a set of parts (represented as shapes within the node) while filtering out the background class. The set of parts at the node is partitioned and each child is responsible for a subset of them. When an example passes a node classifier, it is evaluated by all of its children.

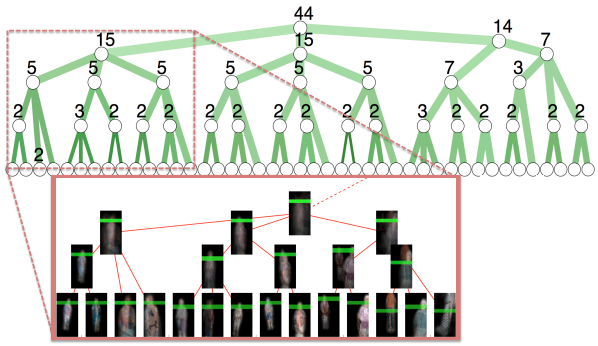


Figure 2: A classification tree generated by our algorithm for classifying 44 poselets at 90% target detection rate. The thickness of the edges denotes the retention rate of the classifiers. The number of poselet types classified by each node is indicated. **Left corner:** A zoom on part of the tree. At each node we show the average mask over all classifiers captured by the node, along with the horizontal stripe that was used to classify the node.

Poselets [1] have been used in a variety of computer vision tasks, such as detection, segmentation, action classification, pose estimation and action recognition, often achieving state-of-the-art performance. Poselets are part classifiers trained to detect part of a human pose under a given viewpoint. Examples of poselet classifiers are a frontal face, a part of a face and left shoulder, or a hand next to a hip in a side view. Poselet evaluation, however, is computationally intensive as it involves running thousands of scanning window classifiers to detect hundreds of poselet types. We present an algorithm for training a hierarchical cascade of part-based detectors and apply it to speed up poselet evaluation. Our cascade hierarchy leverages common components shared across poselets. We generate a family of cascade hierarchies, including trees that grow logarithmically on the number of poselet classifiers.

Example of our cascade and evaluation algorithm is shown on Figure 1. At each node we train a classifier designed to distinguish between a subset of the parts and the background class. We compute two values at the node: the detection rate (the fraction of positive examples that the node classifier passes) and the retention rate (the fraction of examples the node classifier passes). The detection rate of the cascade is the product of detection rates of the chain of nodes from the root to the leaves, and the computational cost is inversely proportional to the retention rate. Our algorithm finds the cascade structure that minimizes the computational cost while preserving a given target detection rate. Since the space of all possible trees is intractably large, we restrict it using a few simplifying assumptions: (1) the detection rate tradeoff between a node and its children is the same throughout the tree, (2) the number of children is no larger than 4, and (3) the partitioning of a set of parts into K subsets (one for each child) is fixed using a clustering algorithm.

While our algorithm is generic, in the case of HOG-based poselets, our node classifiers are linear SVMs over the HOG features in a horizontal stripe of the input image patch. We pick the stripe that best separates the node parts from the background class. Our design choices allow for efficient and memory-cache friendly classifier.

We use a dynamic programming approach to find the optimal cascade structure in this restricted state space. An example of the classification cascade is shown on Figure 2. We test our system on the PASCAL dataset [2] and show an order of magnitude speedup at less than 1% loss in AP (Figure 3-left). We also show that our algorithm evaluation cost scales logarithmically with the number of poselet classifiers (Figure 3-right).

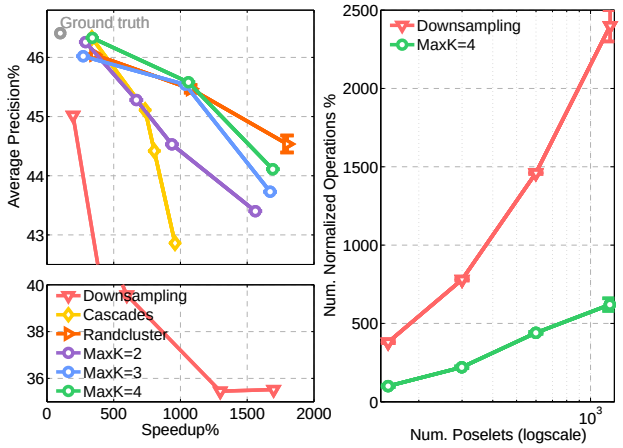


Figure 3: **Left:** Average precision of our classifier (*MaxK*=4) on the PASCAL 2007 set for the Person category as a function of evaluation speed. We compare against the AP of *Downsampling*: standard poselet detector with coarser sampling in space and scale; *Cascades*: independent cascades for each individual poselet; *Randcluster*: cascade hierarchy with random partition and *MaxK=k*: cascade hierarchy where each node can have at most k children. **Right:** Computation time for the same detection rate as a function of the number of poselets.

[1] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[2] Mark Everingham, Luc Van Gool, Chris K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

Regularized Max Pooling for Image Categorization

Minh Hoai
http://www.robots.ox.ac.uk/~minhhoai/

Visual Geometry Group, Department of
Engineering Science, Oxford University

We propose Regularized Max Pooling (RMP) for image classification. RMP classifies an image (or image region) by extracting feature vectors at multiple subwindows at multiple locations and scales. Unlike Spatial Pyramid Matching where the subwindows are defined purely based on geometric correspondence, RMP accounts for the deformation of discriminative parts. The amount of deformation and the discriminative ability for multiple parts are jointly learned during training.

An RMP model is a collection filters. Each filter is anchored to a specific image subwindow and associated with a set of deformation coefficients. The anchoring subwindows are predetermined at various locations and scales, while the filters and deformation coefficients are learnable parameters of the model. Fig. 1 shows a possible way to define subwindows. To classify a test image, RMP extracts feature vectors for all anchoring subwindows. The classification score of an image is the weighted sum of all filter responses. Each filter yields a set of filter responses, one for each level of deformation. The deformation coefficients are the weights for these filter responses.

Given a set of images $\{\mathbf{I}_i\}_{i=1}^n$ and labels $\{y_i | y_i \in \{1,-1\}\}_{i=1}^n$, consider a particular set of geometrically defined subwindows which can encode semantic content of an image at different locations and scales (e.g., Fig 1). Let $\{\mathbf{I}^j\}_{j=1}^m$ denote the set of subwindows for image $\mathbf{I}$. Let ϕ be the feature function of which the input is an image region and the output is a column vector. Let $\mathbf{D}^j$ be the feature matrix computed at location j for all images and $\mathbf{K}^j$ the corresponding kernel, i.e., $\mathbf{D}^j = [\phi(\mathbf{I}_1^j)\cdots\phi(\mathbf{I}_n^j)]$ and $\mathbf{K}^j = (\mathbf{D}^j)^T\mathbf{D}^j$. The joint kernel for all subwindows is the sum of all kernels: $\mathbf{K} = \sum_{j=1}^m \mathbf{K}^j$; this corresponds to concatenating all feature vectors computed at all subwindows. Given the kernel $\mathbf{K}$, we train an Least-Squares SVM and obtain a coefficient vector and bias term α, b. The filter for subwindow j can be computed as $\mathbf{w}^j = \mathbf{D}^j\alpha$.

For a particular subwindow j and an image $\mathbf{I}$, the regularized maximum score is defined:

$$f^j(\gamma) = \max_{k\in\{1,\cdots,m\}} \left\{(\mathbf{w}^j)^T\phi(\mathbf{I}^k) - \gamma.dist(\mathbf{I}^k,\mathbf{I}^j)\right\}. \quad (1)$$

Here γ is a non-negative regularization parameter and $dist(\cdot,\cdot)$ is the square geometric distance between two regions. The square geometric distance from a region R' to a reference region R is defined as:

$$dist(R',R) = \left(\frac{x'-x}{w}\right)^2 + \left(\frac{y'-y}{h}\right)^2 + \log_2^2\left(\frac{w'}{w}\right) + \log_2^2\left(\frac{h'}{h}\right), \quad (2)$$

where (x,y,w,h) and (x',y',w',h') are the center locations, the widths, and the heights of regions R and R' respectively. This distance function is asymmetric. It is invariant to the scale of the coordinate system. The last two terms of Eq. (2) measure the scale distance between R' and R. We use $\log_2(\cdot)$ to ensure that the scale distance from R' to R is the same for the following two cases: (i) R' is k times bigger than R; (ii) R' is k times smaller than R.

The value of $f^j(\gamma)$ is the regularized maximum response; it seeks a location with high filter response and low deformation cost w.r.t. to the anchor region $\mathbf{I}^j$. If γ is 0, $f^j(\gamma)$ is the maximum filter response. If γ is big, $\gamma.dist(\mathbf{I}^k,\mathbf{I}^j)$ will be big except for $k=j$ where $dist(\mathbf{I}^j,\mathbf{I}^j)=0$. Thus, for a big γ, $f^j(\gamma) = (\mathbf{w}^j)^T\phi(\mathbf{I}^j)$, which is the filter response of the anchor region.

The right setting for γ depends on the level of deformation of region j of the semantic class in consideration. Since the deformation level of a region is unknown, we start with an over-complete set of γ's and learn the tradeoff between deformation and discrimination. For each region j of an image $\mathbf{I}$, we construct a feature vector by varying the value of $\gamma \in \{\gamma_1,\cdots,\gamma_k\}$ and compute the regularized maximum response. Let $\mathbf{f}^j$ be the vector of obtained values, i.e., $\mathbf{f}^j = [f^j(\gamma_1),\cdots,f^j(\gamma_k)]^T$. For each image, we obtain a feature matrix by accumulating the filter responses for all regions $\mathbf{F} = [\mathbf{f}^1\cdots\mathbf{f}^m]$. Let $\mathbf{F}_i$ be the feature matrix for image $\mathbf{I}_i$. We jointly learn the deformation and discriminative ability of all regions by

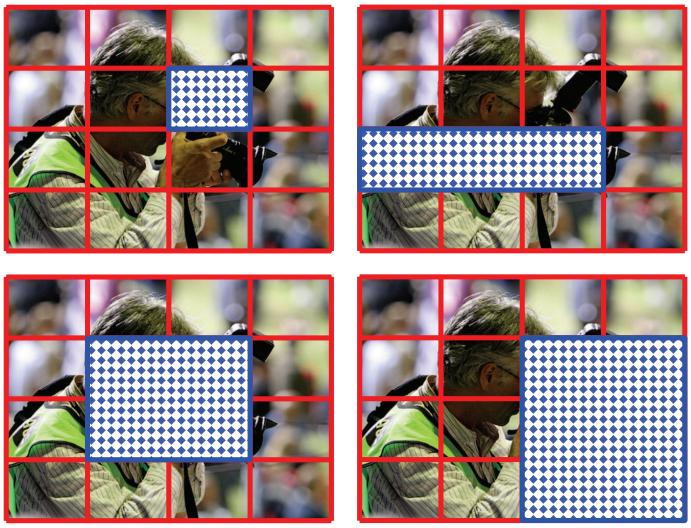


Figure 1: **From grid division to subwindows**. An image is divided into 4×4 blocks. We consider rectangular subwindows that can be formed by a contiguous chunk of blocks. There are 100 such subwindows, and this figure shows four examples.

solving the following optimization problem:

$$\underset{\mathbf{S},\overline{b}}{\text{minimize}} \sum_{i=1}^n (trace(\mathbf{S}^T\mathbf{F}_i) + \overline{b} - y_i)^2 \quad (3)$$

$$\text{s.t. } s_{lj} \geq 0 \;\forall l = 1,\cdots,k,\; \forall j = 1,\cdots,m. \quad (4)$$

The above optimizes over a weight matrix $\mathbf{S} \in \Re^{k\times m}$ and a bias term $\overline{b}$. Each column of $\mathbf{S}$ is a weight vector for a particular region; it learns weights for the regularized maximum responses for different values of γ's. The weights should be non-negative to emphasize the relative importance of higher filter responses. The objective of the above formulation minimizes the sum of L_2 losses.

We start with an over-complete set of γ's and let the algorithm determines the right level of allowable deformation. In our experiments, we use $\gamma_1 = 0$, $\gamma_k = \infty$, $\gamma_l = 2^l/10^4$ for $l = 2,\cdots,k-1$, with $k = 15$. The feasible set of $\mathbf{S}$ is suitable for different levels of deformation, including the following two extreme cases:

1. Well-aligned semantic concept. For an image categorization task where the semantic concepts are well aligned, rigid geometric alignment is the right model. In this case, the weight matrix $\mathbf{S}$ could be all zeros except for the last row of all ones (the last row corresponds to $\gamma = \infty$).
2. Highly deformed semantic concept. For categorization tasks where the semantic concepts have high level of deformation, geometric correspondence should be ignored. In this case, the weight matrix $\mathbf{S}$ could be all zeros except for the first row of all ones (the first row corresponds to $\gamma = 0$).

This formulation corresponds to a linear program, which can be optimized efficiently using a linear programming solver such as Cplex.

We demonstrate the benefits of RMP in recognizing human actions in still images. RMP outperforms Deformable Part Models and Spatial Pyramid Matching, especially for action classes with high level of deformation. Furthermore, the simplicity and flexibility of RMP allow it to be used with any type of features, including Convolutional Neural Network (CNN) features. Together with CNN features, RMP establishes the new state-of-the-art performance for human action recognition in still images, evaluated on the challenging dataset of PASCAL VOC 2012.

Discriminative Embedding via Image-to-Class Distances

Xiantong Zhen
zhenxt@gmail.com

Ling Shao
ling.shao@ieee.org

Feng Zheng
cip12fz@sheffield.ac.uk

Department of Medical Biophysics
The University of Western Ontario
London, ON, Canada

Department of Electronic and Electrical Engineering
The University of Sheffield

Department of Electronic and Electrical Engineering
The University of Sheffield

Thursday

Image-to-Class (I2C) distance firstly proposed in the naive Bayes nearest neighbour (NBNN) classifier [1, 5, 6] has shown its effectiveness in image classification. However, due to the large number of nearest-neighbour search, I2C-based methods are extremely time-consuming, especially with high-dimensional local features. In this paper, with the aim to improve and speed up I2C-based methods, we propose a novel discriminative embedding method based on I2C for local feature dimensionality reduction. Our method **1)** greatly reduces the computational burden and improves the performance of I2C-based methods after reduction; **2)** can well preserve the discriminative ability of local features, thanks to the use of I2C distances; and **3)** provides an efficient closed-form solution by formulating the objective function as an eigenvector decomposition problem. We apply the proposed method to action recognition showing that it can significantly improve I2C-based classifiers.

We incorporate the I2C distance to propose a novel dimensionality reduction method to embed high-dimensional local features into a discriminative low-dimensional space. The use of the I2C distance benefits in two aspects. On the one hand, local features from one image are treated as a whole and class labels can be directly used for supervised learning. This increases the discriminative capacity of local features. On the other hand, it provides an intuitive and effective venue to couple local feature reduction with classification, which can improve the performance of classification. In the low-dimensional space, local features from each image are aligned according to the I2C distances and the I2C distance to its own class is minimized and the I2C distances to other classes are maximized.

Our work contributes in the following aspects: **1)** a novel discriminative subspace learning algorithm based on the I2C distances is proposed for the dimensionality reduction of local features; **2)** after embedding, I2C-based methods are remarkably speeded up and scale well with a large number of local features and therefore become more attractive in real-world applications; and **3)** we formulate the method as an eigenvector decomposition problem, which is efficient with a closed-form solution.

The image-to-class (I2C) distance was first defined in the naive Bayes nearest neighbour (NBNN) classifier. NBNN is an approximation of the optimal MAP naive-Bayes classifier under some assumptions. Given an image Q represented as a set of local features, $\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots,\mathbf{x}_N$, where $\mathbf{x}_i \in R^D$ and D is the dimensionality of local features. The summation of all the distances from the local features of an image to their corresponding nearest neighbours in each class is defined as the Image-To-Class (I2C) distance, which can be calculated by:

$$D_X^c = \sum_{\mathbf{x}\in X} ||\mathbf{x} - NN^c(\mathbf{x})||^2, \tag{1}$$

where NN^c is the nearest neighbour of $\mathbf{x}$ in class c. The resulting classifier takes the form as:

$$\bar{c} = arg\min_c D_X^c, \tag{2}$$

Our task is to classify a collection of videos $\{X_i\}$, each of which is represented by a set of local features: $\{\mathbf{x}_{i1},\ldots,\mathbf{x}_{ij},\ldots,\mathbf{x}_{im_i}\}$, e.g., HOG3D [4], where m_i is the number of local features from image X_i. Given an image/video X_i, its I2C distance to class c is computed according to Eq. 1 as:

$$D_{X_i}^c = \sum_{j=1}^{m_i} ||\mathbf{x}_{ij} - \mathbf{x}_{ij}^c||^2, \tag{3}$$

where $\mathbf{x}_{ij}^c$ is the nearest neighbour in class c.

We aim to find a linear projection $\mathbf{W} \in R^{D\times d}$ to embed the local features into a lower-dimensional space R^d. Unlike the methods in [3], [2], our aim in the embedded space is to minimize the I2C distances from images to the classes they belong to while simultaneously maximizing the I2C distances to the classes they do not belong to. The objective function we used takes the form as:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\sum_{n=1}^{N_i}\sum_i \Delta X_{in}\Delta X_{in}^{\mathrm{T}})\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\sum_i \Delta X_{iP}\Delta X_{iP}^{\mathrm{T}})\mathbf{W})}, \tag{4}$$

where ΔX_{iP} is the auxiliary matrix associated with the class (positive class) that image X_i belongs to and ΔX_{in} is with the class (negative class) that image X_i does not belong to. Note that, given a dataset, the number of negative classes N_i is the same for all images in the dataset.

We can now seek the embedding $\mathbf{W}^*$ to maximize the ratio in Eq. 4. The above equation can be rewritten in terms of covariance matrices as:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_N\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_P\mathbf{W})}, \tag{5}$$

where $\mathbf{C}_N = \sum_{n=1}^{N_i}\sum_i \Delta X_{in}\Delta X_{in}^{\mathrm{T}}$, and $\mathbf{C}_P = \sum_i \Delta X_{iP}\Delta X_{iP}^{\mathrm{T}}$.

It can be seen that maximizing the objective function in Eq. 5 is a well-known eigensystem problem [2]:

$$\mathbf{C}_N\mathbf{W} = \lambda\mathbf{C}_P\mathbf{W} \tag{6}$$

The linear projection is composed of d eigenvectors corresponding to the d largest eigenvalues $\lambda_1,\ldots,\lambda_d$. The whole procedure of the embedding is illustrated in Fig 1.

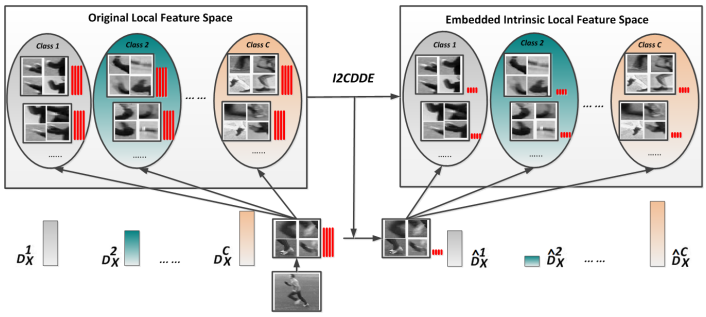


Figure 1: Illustration of the discriminative embedding based on the I2C distance. Action classes are represented by the ellipses in which the rectangles denote local patches from frames (Classes 1, 2 and c represent '*Boxing*', '*Handwaving*' and '*Running*' from the KTH dataset, respectively). The length of the red bars indicates the dimensionality of the local features. The color bars are the I2C distances. D_X^c is the I2C distance from the action X to class c. $\hat{D}_X^c$ is the I2C distance in the embedded space.

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

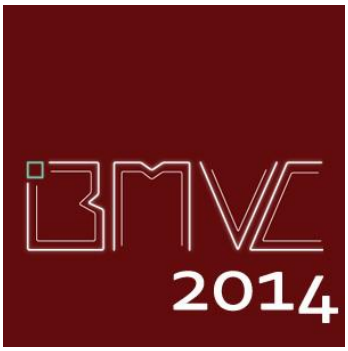
[2] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE TPAMI*, 33(2):338–352, 2011.

[3] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. In *ICCV*, 2007.

[4] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[5] S. McCann and D.G. Lowe. Local naive bayes nearest neighbor for image classification. In *CVPR*, 2012.

[6] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The nbnn kernel. In *ICCV*, 2011.



BMVC 2014

Poster abstracts

Optimized Transform Coding for Approximate KNN Search

Minwoo Park
mpark@objectvideo.com
Kiran Gunda
kgunda@objectvideo.com
Himaanshu, Gupta
hgupta@objectvideo.com
Khurram, Shafique
kshafique@objectvideo.com

Research and Development Services
ObjectVideo
11600 Sunrise Valley Dr, Ste 210
Reston, USA
http://www.objectvideo.com

Transform coding (TC) is an efficient and effective vector quantization approach where the resulting compact representation can be the basis for a more elaborate hierarchical framework for sub-linear approximate search. However, as compared to the state-of-the-art product quantization methods, there is a significant performance gap in terms of matching accuracy. One of the main shortcomings of TC is that the solution for bit allocation relies on an assumption that probability density of each component of the vector can be made identical after normalization. Motivated by this, we propose an optimized transform coding (OTC) such that bit allocation is optimized directly on the binned kernel estimator of each component of the vector. Experiments on public datasets show that our optimized transform coding approach achieves performance comparable to the state-of-the-art product quantization methods, while maintaining learning speed comparable to TC.

Introduction: In the context of general vector quantization, a quantizer encoder $e(\mathbf{x})$ is a real-valued function $\mathcal{E}:\mathbb{R}^n \rightarrow \mathcal{I}$ characterized by the region it induces on the input space, $\mathcal{R}^n_{\mathbf{x}} = \{\mathbf{x} \in R^n(i) : e(\mathbf{x}) = i\}$ and $\cup_{i=1}^{L} R^n(i) = \mathbb{R}^n$ where $\mathcal{I} = \{1,\cdots,L\}$ and $\mathbf{x}$ is an input vector. The decoder $d(i)$ is a real-valued function $\mathcal{D}:\mathcal{I} \rightarrow \mathbb{R}^n$ characterized by the codebook $\mathcal{C} = \{i \in \mathcal{I} : d(i) = y_i\} \subset \mathbb{R}^n$. The mean distortion error of the given quantization level L (MDE) of the quantization is given as:

$$MDE(L) = \sum_{i=1}^{L} \int_{R^n(i)} f(\mathbf{x}) Dist(\mathbf{x}, d(e(\mathbf{x}))) d\mathbf{x} \quad (1)$$

where f is an estimated probability density function of multi-dimensional vector $\mathbf{x}$ and $Dist(\mathbf{x},\mathbf{x}')$ is a distortion error between $\mathbf{x}$ and $\mathbf{x}'$.

In general, to find the optimal set of region $\mathcal{R}_{\mathbf{x}}$, the codebook $\mathcal{C}$, and the given quantization level L, minimum-distortion quantizer aims to minimize mean distortion error (MDE) as follows:

$$\left(\mathcal{R}^{opt}_{\mathbf{x}}, \mathcal{C}^{opt}\right) = arg \min_{\mathcal{R}_{\mathbf{x}},\mathcal{C}} MDE(L) \quad (2)$$

Although design of such a scalar quantizer to satisfy the minimum distortion criterion is well understood, vector quantization is still an open problem. For instance, it can be challenging to obtain sufficient sample data to characterize $f(\mathbf{x})$. Moreover, solving Eq. (2) is computationally expensive in high dimensions.

However, if $p(\mathbf{x})$ is independent in its components (dimensions), and the metric is of the form given as:

$$Dist(\mathbf{x},\mathbf{x}') = \sum_{k=1}^{D} dist(x_k, x'_k), \quad (3)$$

where D is a dimension of $\mathbf{x}$, $\mathbf{x}_k$ are the k^{th} component of $\mathbf{x}$, and $dist(x_k, x'_k)$ is a distance metric between $\mathbf{x}_k$ and $\mathbf{x'}_k$, we can obtain a minimum distortion quantizer by forming the Cartesian product of the independently quantized components. That is, the vector quantization encoder can be of a form, $e(\mathbf{x}) = [e_1(x_1),\cdots,e_D(x_D)]^T$. In the original PQ [4, 5], D dimensional space is divided into M sub-spaces (typical M is 8) to form given as:

$$e(\mathbf{x}) = [e_{1\sim K}(x_{1\sim K}),\cdots,e_{7K+1\sim 8K}(x_{7K+1\sim 8K})]^T \text{ where } K = D/M. \quad (4)$$

However, each component is not independent in practice. Therefore, TC [1] and OPQ [2] aim to minimize inter-component dependencies using the principal component analysis (PCA) and show great success over the original PQ [4, 5]. After minimizing the inter-component statistical dependencies using PCA, the quantizer design problem reduces to a set of M number of independent K dimensional problems. In TC, $K = 1$ and $M = D$. The major difference between OPQ and TC lies in the bit-allocation approach used in each method. The key difference is that OPQ assigns the same number of bits per sub-space, while TC assigns a different number of bits per sub-space. Therefore OPQ finds the best combination of components for each sub-space while maintaining the same number of bits for each sub-space while TC finds the number of bits suitable for each sub-space.

In the context of TC, each quantizing encoder e_k at the k^{th} dimension is designed independently for every $1 \leq k \leq D$ to minimize the expected distortion given as:

$$MDE_k(L_k) = \sum_{i=1}^{L_k} \int_{R_k(i)} f_k(c_k) dist_k(c_k, d_k(e_k(c_k))) dc_k. \quad (5)$$

where c_k is PCA coefficient after projection of $\mathbf{x}$ to PCA subspace k.

Therefore, a vector quantization using B-bits code is summarized as follows:

$$(\mathcal{L},\mathcal{R}_{\mathbf{c}},\mathcal{C})^{opt} = arg \min_{\mathcal{L},\mathcal{R}_{\mathbf{c}},\mathcal{C}} \sum_{k=1}^{D} MDE_k(L_k) \text{ subject to } \sum_{k=1}^{D} log_2(L_k) = B. \quad (6)$$

If the number of distinct quantization levels per k^{th} component L_k is known for a total target bit B, a product quantizer can be obtained by using the minimum distortion criterion. Optimal bit allocation is achieved by minimizing the expected distortion due to quantization. However, solution to this optimization problem for general distributions and distortion functions requires computationally prohibitive numerical search [1].

Instead, Brandt [1] adopted greedy integer-constrained allocation algorithm [3] to assign bits. Number of the quantization level set to be proportional to the variance of the data under the two assumptions that 1) probability density of each component can be made identical after the normalization and 2) per-component distortion functions are identical. However, the first assumption can be easily violated in many cases (e.g., non-Gaussian probability density function). Motivated by this problem, we propose to solve Eq. (6) directly in our proposed optimized transform coding (OTC). Details can be found in the paper.

[1] Jonathan Brandt. Transform coding for fast approximate nearest neighbor search in high dimensions. volume 0, pages 1815–1822, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[2] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2946–2953, Washington, DC, USA, 2013. IEEE Computer Society.

[3] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991. ISBN 0-7923-9181-0.

[4] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[5] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

Acknowledgments: Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

Associating locations from wearable cameras

Jose Rivera-Rubio
http://www.bicv.org
Ioannis Alexiou
i.alexiou09@imperial.ac.uk
Anil Bharath
a.bharath@imperial.ac.uk
Riccardo Secoli
r.secoli@imperial.ac.uk
Luke Dickens
luke.dickens@imperial.ac.uk
Emil Lupu
e.c.lupu@imperial.ac.uk

Imperial College London
South Kensington Campus, UK

1 Motivation and contributions

In this paper, we address a specific use-case of wearable or hand-held camera technology: indoor navigation. We explore the possibility of crowdsourcing navigational data in the form of video sequences that are captured from wearable or hand-held cameras. Without using geometric inference techniques (such as SLAM), we test video data for navigational content, and algorithms for extracting that content. We do not include tracking in this evaluation: our purpose is to explore the hypothesis that visual content, on its own, contains cues that can be mined to infer a person's location. We test this hypothesis through estimating positional error distributions inferred during one journey with respect to other journeys along the same approximate path.

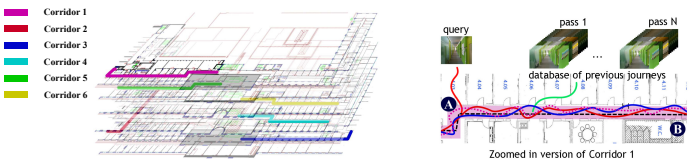


Figure 1: Maps of the recording locations (left). A sample path (Corridor 1, C1) with the multiple passes overlaid (right). Each of these passes represents a database sequence.

The contributions of this work are threefold. First, we propose alternative methods for video feature extraction that identify candidate matches between query sequences and a database of sequences from journeys made at different times. Secondly, we suggest an evaluation methodology that estimates the error distributions in position inference with respect to a ground truth. We assess and compare standard approaches in the retrieval context, such as SIFT [2] and HOG3D [1], to establish positional estimates. The final contribution is a publicly available database comprising over 90,000 frames of video-sequences with positional ground-truth. The data was acquired along more than 3 km worth of indoor journeys with a hand-held device (Nexus 4) and a wearable device (Google Glass).

2 The RSM dataset

The dataset contains 3.05 km of journey data. For each corridor, ten passes (i.e. 10 separate visual paths) were obtained. Five of these videos were acquired with the hand-held Nexus, and the remainder with Glass. The dataset is publicly available at [3].

	Photo	Length (m)			No. of frames		
		Avg	Min	Max	Avg	Min	Max
C1		57.9	57.7	58.7	2157	1860	2338
C2		31.0	30.6	31.5	909	687	1168
C3		52.7	51.4	53.3	1427	1070	1777
C4		49.3	46.4	56.2	1583	1090	2154
C5		54.3	49.3	58.4	1782	1326	1900
C6		55.9	55.4	56.4	1471	1180	1817
Total		3.042 km			90,302 frames		

Table 1: A summary of the dataset with thumbnails.

3 Methods

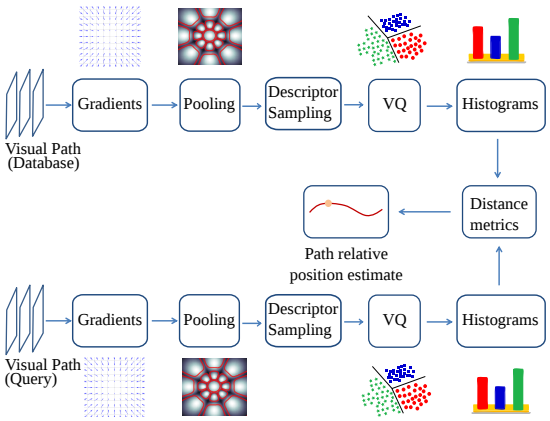
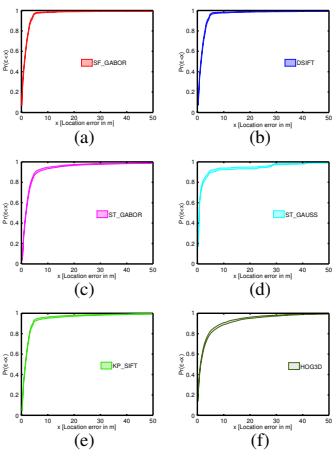
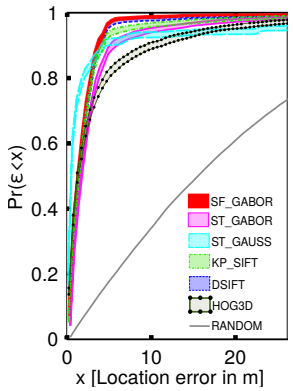


Figure 2: The stages in processing image sequences from database and query visual paths are illustrated above.

4 Evaluation





(g) Comparison between the error distribution obtained with the different methods. The results for a random test (RANDOM) were introduced as a "sanity check".

Figure 3: Cumulative Distribution Function of the methods under study.

[1] A Kläser, M Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, pages 995–1004, 2008.

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

[3] Jose Rivera-Rubio, Ioannis Alexiou, and Anil A. Bharath. RSM dataset, 2014. URL `http://rsm.bicv.org`.

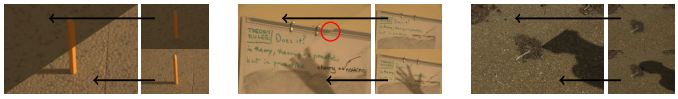
Interactive Shadow Removal and Ground Truth for Variable Scene Categories

Han Gong
http://www.cs.bath.ac.uk/~hg299
Darren Cosker
http://www.cs.bath.ac.uk/~dpc

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

Shadows are ubiquitous in image and video data, and their removal is of interest in both Computer Vision and Graphics. We present an interactive, robust and high quality method for fast shadow removal. To perform detection we use an on-the-fly learning approach guided by two rough user inputs for the pixels of the shadow and the lit area. From this we derive a fusion image that magnifies shadow boundary intensity change due to illumination variation. After detection, we perform shadow removal by registering the penumbra to a normalised frame which allows us to efficiently estimate non-uniform shadow illumination changes, resulting in accurate and robust removal. We also present a reliable, validated and multi-scene category ground truth for shadow removal algorithms which overcomes issues such as inconsistencies between shadow and shadow-free images and limited variations in shadows. Using our data, we perform the most thorough comparison of state of the art shadow removal methods to date. Our algorithm outperforms the state of the art, and we supply our code and evaluation data and scripts to encourage future open comparisons.

Shadow removal ground truth The first public data set was supplied in [2]. In our work, we propose a new data set that introduces multiple shadow categories, and overcomes potential environmental illumination and registration errors between the shadow and ground truth images. An example of comparison is shown in Fig. 1. Our new data set avoids these issues using a careful capture setup and a quantitative test for rejecting unavoidable capture failures due to environmental effects. Our images are also categorised according to 4 different attributes.



(a) mismatched illuminaiton (b) unregistered pixels (c) our data (no artefacts)

Figure 1: For each image: top left segment – shadow-free image; bottom right segment – shadow image. (a) and (b) are taken from [2]. An example from our data without these properties is shown in (c).

Our algorithm consists of 3 steps (see Fig. 2):

1) Pre-processing We detect an initial shadow mask (Fig. 2(b)) using a KNN classifier trained from data from two rough user inputs (e.g. Fig. 2(a)). We generate a *fusion image*, which magnifies illumination discontinuities around shadow boundaries, by fusing channels of YCrCb colour space and suppressing texture (Fig. 2(c)).

2) Penumbra unwrapping Based on the detected shadow mask and fusion image, we sample the pixel intensities of sampling lines perpendicular to the shadow boundary (Fig. 2(d)), remove noisy ones and store the remaining as columns for the initial penumbra strip (Fig. 2(e)). We align the initial columns' illumination changes using its intensity conversion image (Fig. 2(f)). This results in an aligned penumbra strip (Fig. 2(g)) whose conversion image (Fig. 2(h)) exhibits a stabler profile.

3) Estimation of shadow scale and relighting Unlike previous work [1, 2], we do not assume a constrained model of illumination change. The columns of penumbra strip are first clustered into a few small groups. A unified sample can be synthesised by averaging the samples of each group (e.g. Fig. 2(i)). Our shadow scale is adaptively and quickly derived from the unified samples which cancel texture noise. The derived sparse scales for all sampled sites (Fig. 2(j)) are then propagated to form a dense scale field (Fig. 2(k)). We remove shadows by inverse scaling using this non-uniform field (Fig. 2(l)).

Evaluation Directly using the per-pixel error [2, 3] between the shadow removal result and shadow-free ground truth does not take into account the size of the shadow, or the fact that some shadows are darker than others. We therefore compute the error ratio $\mathbf{E}_r = \mathbf{E}_n/\mathbf{E}_o$ as our quality measurement where $\mathbf{E}_n$ is the RMSE between the ground truth and shadow removal result, and $\mathbf{E}_o$ is the RMSE between the ground truth and the original shadow image. This normalised measure better reflects removal

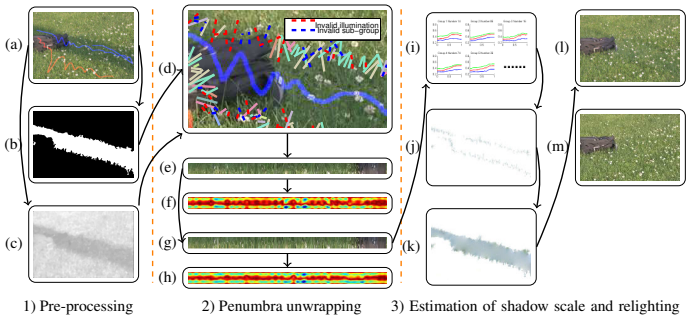


Figure 2: Our shadow removal pipeline. (a) input: a shadow image and user strokes (blue for lit pixels and red for shadowed pixels); (b) detected shadow mask; (c) fusion image; (d) initial penumbra sampling (solid lines in different colours indicate valid samples of different sub-groups. Dashed lines are invalid samples); (e) initial penumbra regularisation; (f) initial penumbra conversion image; (g) final penumbra regularisation; (h) final penumbra conversion image; (i) penumbra illumination estimation; (j) sparse shadow scale; (k) dense shadow scale; (l) output; (m) GT.

improvements towards the ground truth independent of original shadow intensity and size. Our removal test is based on our data set of 186 cases, which contains shadows in variable scenarios as well as simpler shadows, plus 28 example cases from [2] – resulting in 214 test cases in total. Each case is rated according to 4 attributes, which are *texture*, *brokenness*, *colourfulness* and *softness*, in 3 perceptual degrees from weak to strong. Our method is compared with three state-of-the-art methods [1, 2, 4] and shows leading performance across all scores. Tab. 1 shows some typical visual results of shadow removal on various scenarios[1].

	Original	Yang [4]	Guo [2]	Gong [1]	Ours	GT
Tex.						
Sof.						
Bro.						
Col.						
Other						

Table 1: Comparisons using images in different categories.

Application Our method is exclusively suitable for real-time interactive shadow editing which offers free controls for shape, darkness and smoothness of either new or original shadows (see our supplementary material).

Conclusions We have presented an interactive method for fast shadow removal together with a state of the art ground truth. Our method balances the complexity of user input with robust shadow removal performance. Our quantitatively-verified ground truth data set overcomes issues of mismatched illumination and registration. We have evaluated our method against several state of the art methods using a thorough quantitative test and shown leading state of the art performance.

[1] H. Gong, D. Cosker, C. Li, and M. Brown. User-aided single image shadow removal. In *ICME*, pages 1–6, 2013.

[2] R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. *PAMI*, PP(99):1–1, 2012.

[3] Y. Shor and D. Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *CGF*, 27(2):577–586, 2008.

[4] Q. Yang, K.-H. Tan, and N. Ahuja. Shadow removal using bilateral filtering. *IEEE Trans. on Image Proc.*, 21(10):4361–4368, 2012.

[1] Our supplementary material shows a wide range of other removal results with higher resolution images.

Board #4

Segmentation of Dynamic Scenes with Distributions of Spatiotemporally Oriented Energies

Damien Teney
d.teney@bath.ac.uk
Matthew Brown
m.brown@bath.ac.uk

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

Overview

In video segmentation, disambiguating appearance cues by grouping similar motions or dynamics is potentially powerful, though non-trivial. Dynamic changes of appearance can occur from rigid or non-rigid motion, as well as complex dynamic textures. While the former are easily captured by optical flow, phenomena such as a dissipating cloud of smoke, or flickering reflections on water, do not satisfy the assumption of brightness constancy, or cannot be modelled with rigid displacements in the image. To tackle this problem, we propose a robust representation of image dynamics as histograms of motion energy (*HoME*) obtained from convolutions of the video with spatiotemporal filters. They capture a wide range of dynamics and handle problems previously studied separately (motion and dynamic texture segmentation). They thus offer a potential solution for a new class of problems that contain these effects in the same scene. Our representation of image dynamics is integrated in a graph-based segmentation framework [3] and combined with colour histograms to represent the appearance of regions. In the case of translating and occluding segments, the proposed features additionally serve to characterize the motion of the boundary between pairs of segments, to identify the occluder and inferring a local depth ordering. The resulting segmentation method is completely model-free and unsupervised, and achieves state-of-the-art results on the SynthDB dataset for dynamic texture segmentation, on the MIT dataset for motion segmentation, and reasonable performance on the CMU dataset for occlusion boundaries.

Proposed approach

Our approach to identify motion is based on existing work on steerable spatiotemporal filters [1, 2]. Similarly to 2D filters used to identify 2D structure in images (*e.g.* edges), these 3D filters can reveal structure in the spatiotemporal video volume. We employ Gaussian second derivative filters $G2_{\hat{\theta}}$ and their Hilbert transforms $H2_{\hat{\theta}}$. They are both steered to a spatiotemporal orientation parameterized by the unit vector $\hat{\theta}$ (the symmetry axis of the $G2$ filter). They are convolved with the video volume $\mathcal{V}$ of stacked frames, and give an energy response

$$E_{\hat{\theta}}(x,y,t) \;=\; (G2_{\hat{\theta}} * \mathcal{V})^2 + (H2_{\hat{\theta}} * \mathcal{V})^2 \quad . \tag{1}$$

In the frequency domain, a pattern moving in the video with a certain direction and velocity correspond to a plane passing through the origin. We obtain a representation of image dynamics by measuring the energy along a number of those planes, obtained by summing responses of filters consistent with the orientation of each plane. The resulting **motion energy** ME along the plane of unit normal $\hat{n}$ is given by

$$ME_{\hat{n}}(x,y,t) \;=\; \sum_{i=0}^{N} E_{\hat{\theta}_i}(x,y,t) \quad , \tag{2}$$

where N=2 is the order of the derivative of the filter, and $\hat{\theta}_i$ are filter orientations whose response lie in the plane specified by $\hat{n}$ (see [1] for details). This provides a representation of *dynamics* only, marginalizing the filter responses over appearance. The measurements $ME_{\hat{n}_i}$ can be compared to the extraction of optical flow, since each $\hat{n}_i$ specifies a particular orientation and velocity (*e.g.* patterns moving rightwards at 2 pixels per frame). The complete set of measurements $ME_{\hat{n}_i}$ is potentially capable of representing multiple, superimposed motions at a single location, offering definitive advantages over optical flow. Using the observation that motion- and color-based segmentation are two intrinsically similar problems, we adapt the segmentation algorithm of [3] to use our representation of motion. In addition to the original color histograms that represent the appearance of regions, we similarly accumulate our features into motion histograms (as in [3]). These motion histograms have 2 dimensions, corresponding to the (spatial) orientations and (spatiotemporal) velocities of the different $\hat{n}_i$ considered. The agglomerative segmentation iteratively produces results at decreasing levels of granularity.

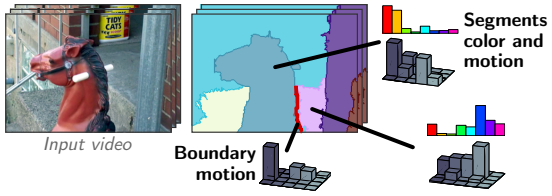


Figure 1: We represent dynamics in regions of the video with histograms of motion energies (HoME) measured at various space-time orientations. They are combined with colour histograms in a graph-based segmentation framework [3]. Post segmentation, HoMEs are additionally used to compare the motion of boundaries with their adjacent segments'. We thereby identify the occluders and infer a local depth ordering.

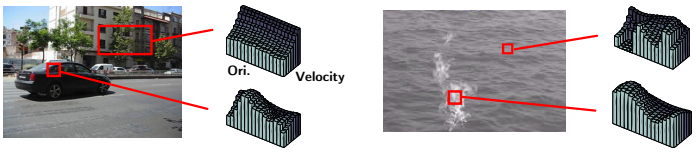


Figure 2: Actual HoMEs of real sequences, visualized as 2D histograms, of image (spatial) orientations and (spatiotemporal) velocities (lighter colours represent higher velocities; a limited set of velocities is represented for compactness). **(Left)** The background is mostly static with a uniform range orientations, whereas the moving car produces a single mode in the histogram. **(Right)** The sea waves exhibit multiple motion modes; the upwards motion of the flame is more simply defined.

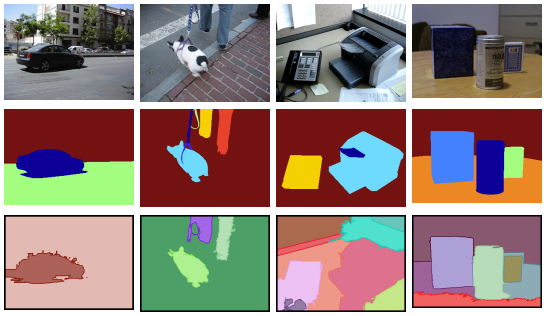


Figure 3: Motion segmentation (MIT dataset); input frame, ground truth, and segmentation. Different objects are correctly segmented, whether from their intrinsic motion (first two examples) or different relative motion induced by parallax and a translating camera (last two examples).

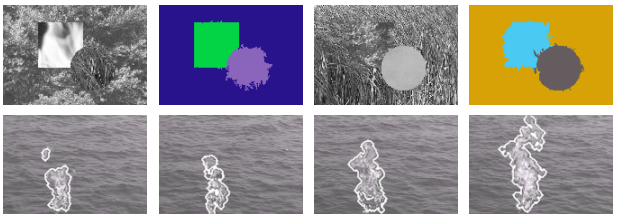


Figure 4: Segmentation of dynamic textures (SynthDB dataset). Static appearance of different textures may be very similar, and image dynamics are then crucial to distinguish them.

[1] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6):1193–1205, 2012.

[2] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

[3] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, pages 2141–2148, 2010.

Board #5

The State of the Art: Object Retrieval in Paintings using Discriminative Regions

Elliot J. Crowley
elliot@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

The objective of this work is to recognize object categories (such as animals and vehicles) in paintings, whilst learning these categories from natural images. This is a challenging problem given the substantial differences between paintings and natural images, and variations in depiction of objects in paintings [5] – see figure 1.

Contributions. (i) We show that object category classifiers learnt using Fisher Vectors [4] extracted from natural images can retrieve paintings containing that category with some success; (ii) we then introduce a method of re-ranking these retrieved paintings based on spatial consistency of Mid-Level Discriminative Patch (MLDP) correspondences with the original training images and show that the precision of the top ranked paintings (i.e. the ones that would appear on the first webpage in an image search) can be significantly improved using this method.

Motivation. Obtaining paintings with a particular object is of interest to Art Historians who currently find paintings manually or from memory. They can then study the change in the depiction style over time or determine when an object first appeared in paintings.

Summary of method. Object category classifiers are learnt from training sets of natural images (e.g. PASCAL VOC) and applied to paintings. The top ranked paintings for each category are re-ranked based on their spatial consistency with the natural images as follows: (i) discriminative regions are extracted from the natural images using the method of Aubry *et al.* [2] (figure 2); (ii) these regions are used to learn LDA [3] classifiers which are applied as sliding window detectors to the top ranked paintings to find matching regions and a RANSAC style algorithm is used to remove outlying matches (figure 3); (iii) each painting is scored by the maximum number of inlying matches shared with a natural image and are re-ranked accordingly (figure 4).

Figure 1: **Example Paintings** from top to bottom row, those containing: dog, horse, train. Objects have a variety of sizes, poses and depictive styles, and can be partially occluded or truncated. The paintings have been obtained from [1].

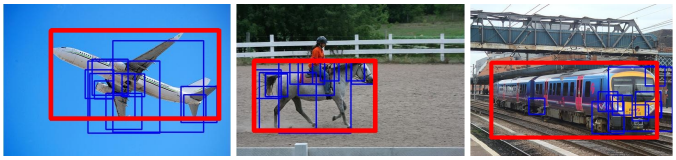


Figure 2: **Discriminative regions are extracted from natural images.** For a given object category square regions are sampled from each object ROI at a variety of scales; each region is represented by a HOG descriptor and assigned a score based on how much that HOG descriptor differs from the mean HOG descriptor of many natural images. Regions that score the highest this way are retained and are considered to be MLDPs for the object. The figure above shows a subset of discriminative regions (blue) overlapping with PASCAL VOC ROIs (red) for several images. Notice that informative areas of the objects are picked out such as a horse's head, and even within the ROI no indiscriminate background patches are selected.

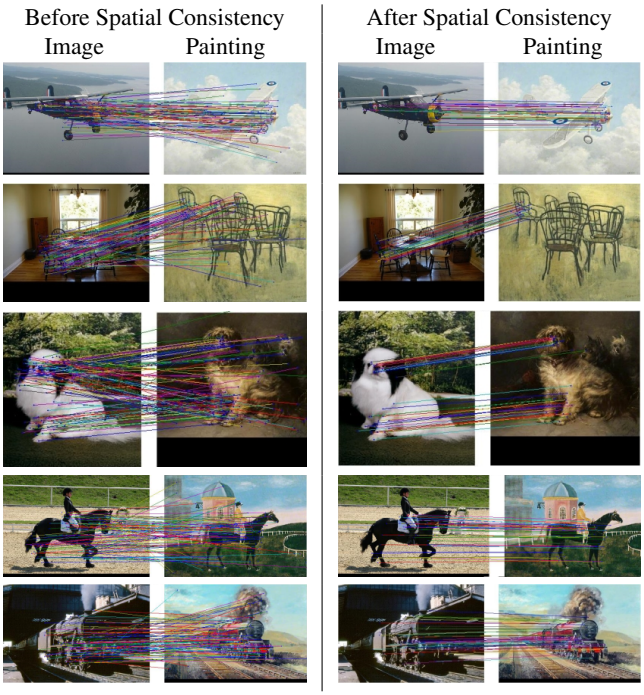


Figure 3: **Obtaining correspondences.** Each sliding window detector obtained from a mid-level discriminative patch (MLDP) on the natural image, defines a possible correspondence at the highest scoring detection window on each painting. This gives a set of provisional correspondences between each image-painting pair for an object. For each pair a RANSAC style algorithm is used to select a subset of these correspondences that are spatially consistent, and the image-painting pair is scored based on the size of this subset. Note, that the MLDP correspondences are able to generalize slightly over viewpoint, intra-class differences, and between natural images and paintings.

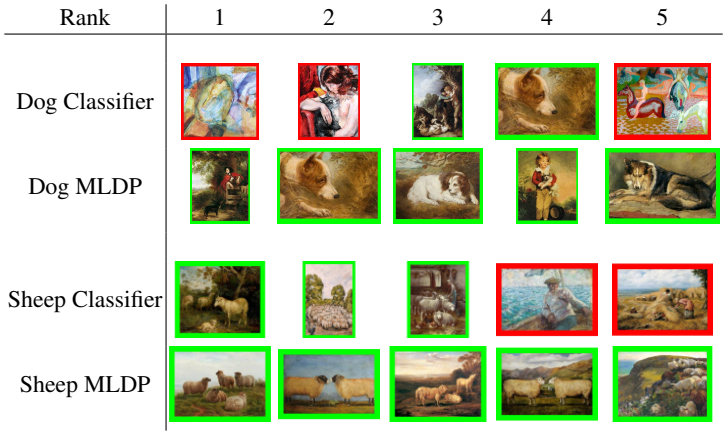


Figure 4: Top 5 ranked paintings before and after re-ranking using MLDPs for the dog and sheep category. A green border indicates a correct classification and a red border an incorrect one. Classification results are improved using our method.

[1] BBC – Your Paintings. http://www.bbc.co.uk/arts/yourpaintings/.

[2] M. Aubry, B. Russell, and J. Sivic. Painting-to-3D model alignment via discriminative visual elements. In *ACM Transactions of Graphics*, 2013.

[3] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Proc. ECCV*, 2012.

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

[5] Q. Wu and P. Hall. Modelling visual objects invariant to depictive style. In *Proc. BMVC.*, 2013.

Variational Level Set Segmentation in Riemannian Sobolev Spaces

Maximilian Baust[1]
campar.in.tum.de/Main/MaximilianBaust
Darko Zikic[2]
research.microsoft.com/en-us/people/darko
Nassir Navab[1]
campar.in.tum.de/Main/NassirNavab

[1] Computer Aided Medical Procedures & Augmented Reality, Technische Universität München, Munich, Germany
[2] Machine Learning and Perception, Microsoft Research, Cambridge, GB

The variational level set method [9] is still one of the most widely used methods in computer vision – especially for image segmentation [1]. This popularity might seem surprising, because variational level set segmentation is known to be non-convex, e.g., [3]. All the more, because since the seminal work of Chan et al. [3] a lot of research has been carried out in order to develop efficient methods for solving convex models for image segmentation, cf. [1, 2, 5].

The non-convexity of the variational level set approach is caused by the usage of continuous but non-convex approximations of the Heaviside and Dirac distribution for defining area and boundary integrals. This non-convexity is, however, not always a bane, because variational level set formulations for localized active contours models [6] or image segmentation in the presence of intensity inhomogeneities [7] make extensive usage of smeared-out Heaviside and Dirac distributions. As a consequence, it is still of interest to develop efficient methods for the non-convex variational level set method for image segmentation, which is the goal of this paper. Thereby, we will consider so-called Sobolev gradient flows, which have recently been shown to be superior to classical L^2-based gradient flows [4, 8]. Inspired by [10], we extend these approaches by changing the notion of distance in H^1. The main observation which leads to the proposed approach is that standard gradient for variational level set segmentation take the form

$$\nabla\mathcal{E}(\phi(x)) = F(x, I(x), \phi(x), \nabla\phi(x)), \tag{1}$$

where $\mathcal{E}$ is the energy to be minimized, I denotes the image to be segmented, and ϕ is the level set function. As a consequence, the gradient does not only inherit the very local behavior of the image, making the resulting level set evolution prone to get stuck in local minima, but also varies significantly w.r.t. to the individual problem dimensions, i.e., pixels. Both of the issues can be cure with the proposed approach which essentially projects this gradient into a Sobolev space endowed with a carefully chosen inner product. Thus, the minimizing gradient flow in the Riemannian Sobolev space exhibits a significantly improved convergence, compared to gradient flow in H^1. This advantage in convergence translates directly to an improvement of the overall runtime, cf. Fig. 1.

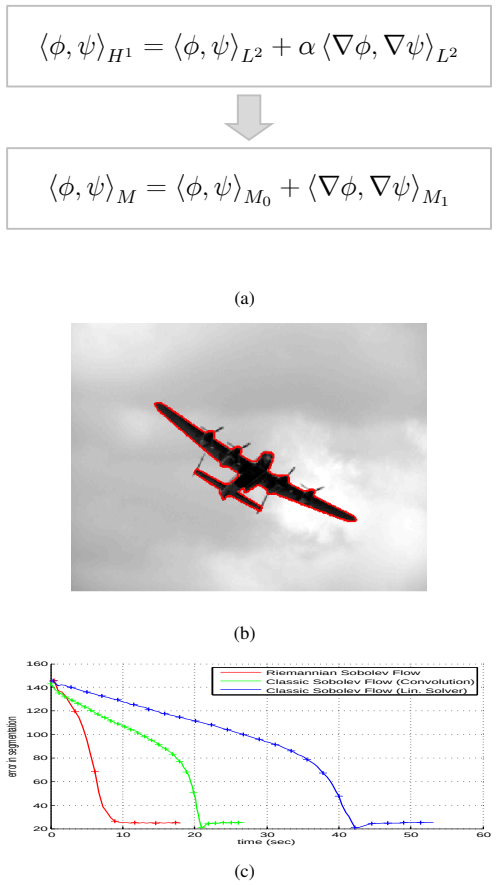


Figure 1: **The proposed generalization** (a) results in efficient Riemannian Sobolev flows, which provide accurate results (b), however with significantly improved convergence and overall runtime (c). Every 5th iteration is marked with a +.

[1] X. Bresson, S. Esedoglu, P. Vandergheynst, J. P. Thiran, and S. Osher. Fast global minimization of the active contour/snake model. *Journal of Mathematical Imaging and Vision*, 28(2):151–167, 2007.

[2] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[3] Tony F. Chan, Selim Esedo Glu, and Mila Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. Technical report, SIAM Journal on Applied Mathematics, 2004.

[4] G. Charpiat, P. Maurel, J. P. Pons, R. Keriven, and O. Faugerás. Generalized gradients: Priors on minimization flows. *International Journal on Computer Vision*, 73(3):325–344, 2007.

[5] T. Goldstein and S. Osher. The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences*, 2(2): 323–343, 2009.

[6] S. Lankton and A. Tannenbaum. Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17(11):2029 – 2039, 2008.

[7] C. Li, R. Huang, Z. Ding, C. Gatenby, D. N. Metaxas, and J. C. Gore. A level set method for image segmentation in the presence of intensity inhomogeneities with application to mri. *IEEE Transactions on Image Processing*, 20(7):2007–2016, 2011.

[8] G. Sundaramoorthi, A. J. Yezzi, and A. Mennucci. Sobolev active contours. *International Journal of Computer Vision*, 73(3):345–366, 2007.

[9] H. K. Zhao, T. Chan, B. Merriman, and S. Osher. A variational level set approach to multiphase motion. *Journal of Computational Physics*, 127(1):179–195, 1996.

[10] Darko Zikic, Maximilian Baust, Ali Kamen, and Nassir Navab. A general preconditioning scheme for difference measures in deformable registration. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2011.

[1] This fact is proven by 579 hits since 2012 obtained by a Google scholar search for the query "'variational level set' + 'segmentation'" performed on 9th of May 2014.

Robust segment-based Stereo using Cost Aggregation

Veldandi Muninder[1]
veldandi.muninder@nokia.com
Ukil Soumik[2]
soumik.ukil@nokia.com
Govindarao Krishna[2]
krishna.govindarao@nokia.com

[1] Nokia Technologies,
Sunnyvale
California, USA

[2] Nokia Technologies
Bangalore
India

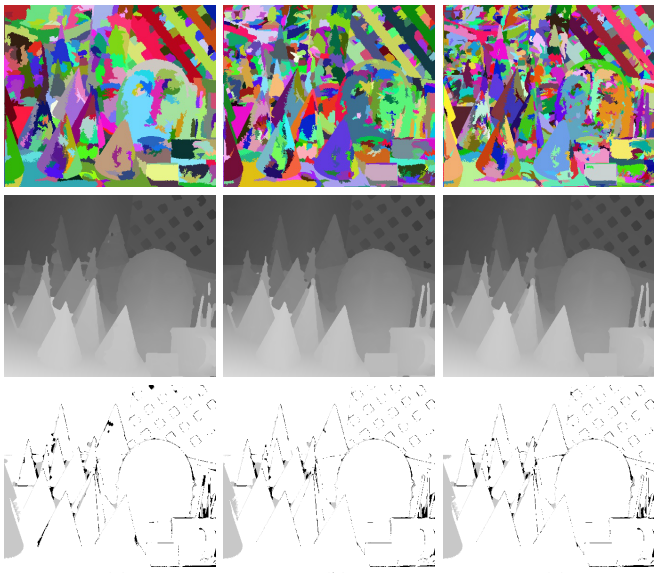
Introduction Most segment based stereo methods estimate disparity by modeling color segments as 3-D planes [2]. Inherently, such methods are sensitive to segmentation parameters and intolerant to segmentation errors. Two main dependencies of these methods on the underlying segmentation algorithm are: size of segments used for estimating planes, and assignment of a single plane to the whole segment. Specifically, in the case of under-segmentation, there is a higher chance of merging multiple objects (with multiple plane surfaces) into a single segment. Consequently, planes estimated using these segments are erroneous. The effect propagates to the disparity map, wherein a larger segment encompassing multiple objects is incorrectly represented by a single disparity plane. In the over-segmentation case, which gives smaller color segments, the estimated planes may be unreliable, leading to an inaccurate disparity map. Popular segment based methods try to solve this problem by re-fitting the planes on grouped segments, in an iterative manner [2]. We propose a novel algorithm for generating sub-pixel accurate disparities on a per-pixel basis, thus alleviating the problems arising from methods that estimate disparities on a per-segment basis. The proposed method computes sub-pixel precision disparity maps using the recent minimum spanning tree (MST) [4] based cost aggregation framework. Since the disparity at every pixel is modeled by a plane equation, the goal is to ensure that all pixels belonging to a planar surface are labeled with the same plane equation. We show that using a reduced and refined set of planes as candidate labels in the aggregation framework ensures homogeneous labeling within a color segment.

Proposed Method Our method computes an initial set of plane equations (label set) by fitting planes inside a color segment using the consistent disparities from an initial disparity map. The initial disparity map may be generated using any local or global algorithm. These plane equations form the initial label set and a matching cost volume is computed over this set for every pixel. This cost volume is aggregated using MST based cost aggregation framework and a WTA over the aggregated cost volume gives the initial labeling. The number of labels in the initial set is of the order of the number of segments, with a plane estimate for every segment. The initial labeling is used along with the color segmentation to filter and generate a reduced set of planes. This framework of plane filtering followed by re-labeling leads to a more accurate disparity map. In addition, segment analysis is also used to modify the plane matching cost. We weigh the pixel matching cost by a support factor, where the support factor is derived from the distribution of plane labels within the color segment, as follows:

$$D(p,l) = \rho(p,q)e^{\frac{-n_{l,s}}{\tau n_s}} \quad (1)$$

where $\rho(p,q)$ computes the pixel dissimilarity between the pixels p and q, n_s is the number of pixels in the segment s that contains p, $n_{l,s}$ is the number of pixels in the segment s that are assigned plane label l, and τ is a constant. This cost update adds a bias towards locally dominant labels, whilst suppressing labels with smaller support. The labeling derived from modified cost volume with reduced set of labels is more locally homogenous than previous labelings. The above matching cost modification is also used in occlusion filling step, which encourages labeling the occlusion region with a dominant plane label in the color segment the occlusion belongs to. The core algorithm block of plane labeling can be iterated on, in a feedback loop. The sub-pixel precision disparity map generated from the final plane labeling is used along with the initial color segments to re-estimate the set of planes. While a convergence criteria based on change in absolute disparities between iterations can be used, we have empirically found that convergence is reached in 3 iterations.

Results We report the experimental results using the proposed method on the Middlebury set [3] and also on natural scenes. We demonstrate the robustness of proposed method to the quality of the initial disparity map by considering two different methods for creating input fronto-parallel disparity maps. First, we initialize our method with a disparity map generated using simple WTA, without cost aggregation. The overall Middlebury rank [3] with this initialization is 21 after three iterations of our algorithm. Next, we initialize our method with the disparity generated by [4]. Three iterations of our algorithm using this initialization leads to an improvement in overall Middlebury rank from 43 to 11. Additionally, we report the lowest average percentage of bad pixels (3.58), of all methods in the Middlebury evaluation. The results indicate that our method adds a refinement step that is robust and can be added to any local or global algorithm generating fronto-parallel disparities. The recent method of Bleyer et al. [1] which also estimates a plane assignment per pixel takes 1 minute on an average to compute a disparity map on the Middlebury. The average run-time of our method on the Middlebury set is 25 seconds on a 2.67 GHz Intel Core i7 CPU with 8 GB memory.



(a) (b) (c)

Figure 1: Effect of segmentation variance on disparity (Cones): (a) 266 segments, error = 2.58, rank = 23; (b) 507 segments, error = 2.10, rank = 3; (c) 836 segments, error = 2.16, rank = 6;

Next, we demonstrate the robustness to segmentation parameter variation. The minimum segment size parameter in mean-shift segmentation is varied to generate varying segmentation maps. Our method is robust to these variations, resulting in accurate disparity maps in all instances as shown as shown in Fig. 1 of Middlebury Cones image. Observing the bottom right corner of the disparity maps in Fig. 1(a), 1(b), the pencils belong to a segment that spans multiple objects. Despite this leakage our algorithm is able to recover and assign correct disparities. The methods of [2] inherently generate labels on a per-segment basis, leading to a lower tolerance for such variations.

[1] M. Bleyer, C. Rhemann. C, and C. Rother. Patchmatch stereo - stereo matching with slanted support windows. *BMVC*, pages 1–11, 2011.

[2] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. *ICPR*, pages 15–18, 2006.

[3] D. Scharstein and R. Szeliski. Middlebury stereo evaluation. `http://vision.middlebury.edu/stereo/eval/`.

[4] Q. Yang. A non-local cost aggregation method for stereo matching. *CVPR*, pages 1402–1409, 2012.

Coloured signed distance fields for full 3D object reconstruction

Wadim Kehl[1]
kehl@in.tum.de
Nassir Navab[1]
navab@in.tum.de
Slobodan Ilic[2]
slobodan.ilic@siemens.com

[1] CAMP Chair
Computer Science Department
TU Munich, Germany

[2] Siemens AG
Research & Technology Center
Munich, Germany

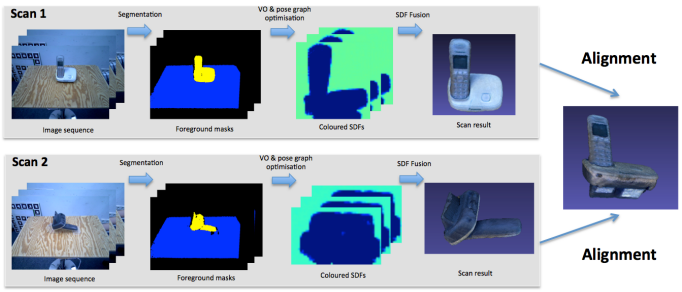


Figure 1: We take foreground-masked sequences, pose-optimise and fuse each of them and eventually align them to one coherent 3D model.

We propose a full 3D object reconstruction framework with a RGB-D sensor, requiring no marker boards and allowing for objects to be displaced during scanning. The proposed framework consists of three stages and provides a novel fusion and registration procedure for coloured signed distance fields (CSDFs) resulting in complete 3D models with high fidelity. It is suitable for a large variety of objects and outperforms the state-of-the-art both in terms of visual quality and geometrical accuracy.

The first step in our pipeline is the camera trajectory estimation via RGB-D visual odometry [5] similar to [2]. The goal is to compute the rigid-body movement $\Xi \in SE(3)$ of the camera between two consecutive foreground-masked sensor pairs $[I_0, D_0]$,$[I_1, D_1]$ by minimising

$$E(\Xi) = \int_{\Omega_2} [I_1(w_\Xi(x)) - I_0(x)]^2 dx$$

with a warp function $w_\Xi : \Omega_2 \to \Omega_2$ defined via the depth maps as $w_\Xi(x) = \pi_{D_1}(\Xi \cdot \pi_{D_0}^{-1}(x))$. We move the support surface while collecting keyframes along the way and eventually refine the trajectory globally with a pose-graph optimisation after loop closure detection.

After one full scan and pose refinement, we refer to our final result as a hemisphere $\mathcal{H} = \{(I_i, D_i, P_i)_i\}$ consisting of masked sensor pairs and poses. We create a 3D model ϕ by fusing the data, analogously to [1, 3], into a CSDF in a variational fashion with an approximate L^1 minimisation. We cast our data into volumetric geometry fields $f_i : \Omega_3 \subset \mathbb{R}^3 \to \mathbb{R}$ and colour fields $c_i : \Omega_3 \to [0,1]^3$ and we seek the minimisers of the functional

$$\mathcal{E}(u,v) = \int_{\Omega_3} [\mathcal{D}(\mathbf{f}, \mathbf{w}, \mathbf{c}, u, v) + \alpha \mathcal{S}(\nabla u) + \beta \mathcal{S}(\nabla v)] \; d\mathbf{x}$$

with a data term $\mathcal{D}$ that strives to uphold the solution's fidelity to all the observations $\mathbf{f} = \{f_1, ..., f_n\}, \mathbf{c} = \{c_1, ..., c_n\}$ and two weighted regularisers $\mathcal{S}(\nabla u)$ and $\mathcal{S}(\nabla v)$. In contrast to the original work [4], which only fuses the geometrical fields, we also include colour information into the formulation and solve simultaneously for both.

A suitable data term for many vision problems usually involves an outlier-robust L^1-norm whereas for regularisation purposes the total variation (TV) of the function is often employed:

$$\mathcal{D}(\mathbf{f}, \mathbf{w}, \mathbf{c}, u, v) = \frac{1}{\varepsilon + \sum_i w_i} \sum_i w_i \cdot (|u - f_i| + |v - c_i|) \;\; , \;\; \mathcal{S}(\nabla u) = |\nabla u|.$$

Due to the problematic aspect of solving such energies, specific minimisation schemes are employed (e.g. a ROF-variant or (iterated) primal-dual solutions). An alternative has been proposed in [4] where the problematic terms have been replaced with a smooth $epsL^1$ approximation $\Gamma(x) := \sqrt{x^2 + \varepsilon}$. We define it similarly as

$$\mathcal{D}(\mathbf{f}, \mathbf{w}, \mathbf{c}, u, v) = \Gamma(\sum_i w_i)^{-1} \sum_i w_i \cdot (\Gamma(u - f_i) + \Gamma(v - c_i)) \;\; , \;\; \mathcal{S}(\nabla u) = \Gamma(|\nabla u|)$$

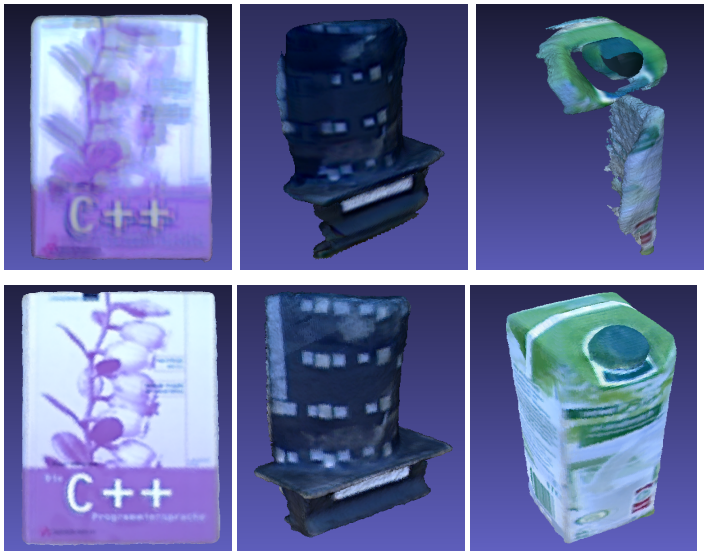


Figure 2: KinectFusion (top) vs. our approach (bottom). We recover richer texture as well as geometry, even for visually poor objects.

where we regard the weighted approximate absolute differences together with an additional normalisation factor and an approximate TV-regulariser.

Usually, one such scan does not expose the full geometry of the object. To this end, we propose to create multiple scans of the same object but placed differently in order to reveal hitherto unseen parts, thus acquiring multiple hemispheres $\mathcal{H}_j$. Then the transformations Ξ_j that map the models from all hemispheres to the first one $\mathcal{H}_0$ need to be determined. In order to retrieve those Ξ_j, we use the reconstructed models ϕ_j and align them automatically using a dense approximate-L^1 registration framework:

$$\mathcal{E}(\Xi)_{L^1} = \int_{\Omega_3} \Gamma(\phi_0(\mathbf{x}) - \phi_j(\Xi(\mathbf{x}))) \; d\mathbf{x}.$$

We compared our method to a commercial state-of-the-art KinectFusion implementation on eight real-life objects. Even though KinectFusion performed well, it failed for some of the objects due to poor geometry leading to tracking failure and supplied only mediocre results in terms of texture. For two models ground-truth data was available and was used to measure the geometrical error of the reconstructions. We show that we tremendously boost the geometrical and textural fidelity for all scanned objects due to the pose graph optimisation and the L^1 sensor fusion.

[1] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996.

[2] Maria Dimashova, Ilya Lysenkov, Vincent Rabaud, and Victor Eruhimov. Tabletop Object Scanning with an RGB-D Sensor. *ICRA Workshop*, 2013.

[3] Richard A. Newcombe, Andrew J. Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[4] Christopher Schroers, Henning Zimmer, Levi Valgaerts, Oliver Demetz, and Joachim Weickert. Anisotropic Range Image Integration. In *Pattern Recognition, LNCS*, 2012.

[5] Frank Steinbrucker, Jurgen Sturm, and Daniel Cremers. Real-time visual odometry from dense RGB-D images. In *ICCV Workshop*, 2011.

Board #9

Automatic Camera Calibration for Traffic Understanding

Markéta Dubská[1]
idubska@fit.vutbr.cz
Jakub Sochor[1]
isochor@fit.vutbr.cz
Adam Herout[12]
herout@fit.vutbr.cz

[1] Graph@FIT
Brno University of Technology
Czech Republic

[2] click2stream, Inc.

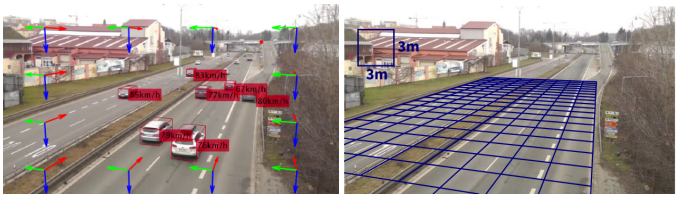


Figure 1: We automatically determine 3 orthogonal vanishing points, construct vehicle bounding boxes (*left*), and automatically determine the camera scale by knowing the statistics of vehicle dimensions. This allows us to measure dimensions and speed (*right*) and analyze the traffic scene.

This paper proposes a method for fully automatic calibration of traffic surveillance cameras. Our method allows for calibration of the camera – including scale – without any user input, only from several minutes of input surveillance video. The targeted applications include speed measurement, measurement of vehicle dimensions, vehicle classification, etc.

The first step of our approach is camera calibration by determining three vanishing points defining the stream of vehicles (Fig. 2, [3]). The second step is construction of 3D bounding boxes of vehicles (Fig. 3) and their measurement up to scale. In the third step, we use the dimensions of the 3D bounding boxes for calibration of the scene scale (Fig. 4).

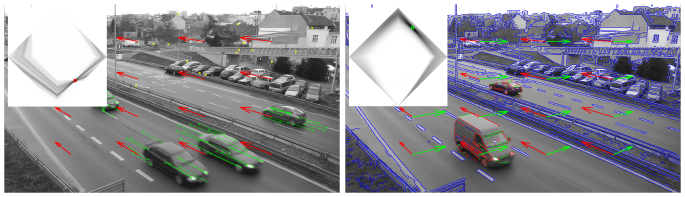


Figure 2: (*left*) Tracked points used for estimation of the 1st VP. Points exhibiting a significant movement (green) are accumulated. (*right*) Accumulation of the 2nd vanishing point. Only edges excluding the vertical ones and those with their direction towards the first VP (green) are accumulated to the diamond space.

Our method for VP detection uses Hough transform based on parallel coordinates [2], which maps the projective plane into a finite space referred to as the *diamond space* by a piecewise linear mapping of lines.

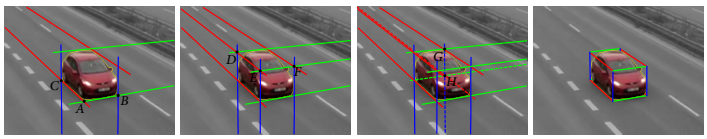


Figure 3: Construction of vehicle's 3D bounding box. From left to right: tangent lines and their relevant intersections A,B,C; derived lines and their intersections E,D,F; derived lines and intersection H; constructed bounding box.

The next step of our approach is construction of 3D bounding boxes of the observed vehicles (Fig. 3). We assume that the vehicle silhouettes can be extracted by background modeling and foreground detection and that the vehicles of interest are moving from/towards the first vanishing point. The 3D bounding box is constructed using tangent lines from vanishing points to the blob's boundary.

Having the bounding box projection, it is directly possible to calculate the 3D bounding box dimensions (and position in the scene) up to precise scale. By fitting the statistics of known dimensions and the measured data from the traffic, we obtain the scale of the scene (Fig. 4).

Camera orientation together with a know distance enables for measuring of vehicle speed/size or distances in the scene. We measured several

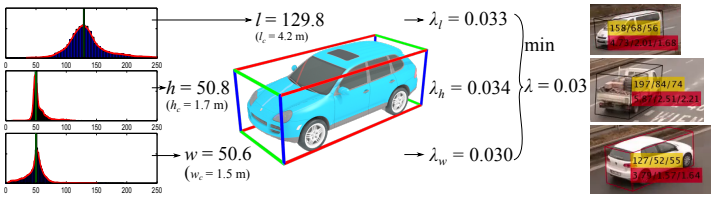


Figure 4: Calculation of scene scale. *(left)* Median (green bar) for each dimension is found in the measured data. *(middle)* Scales are derived separately based on known median car size and the final scale is derived as the minimum from these three scales. *(right)* Examples of relative size of the vehicles (yellow) and real dimensions in meters after scaling.

distances on the road plane and evaluated the error in measurements by our approach. Similar evaluation was provided by Zhang [5], who report average error of measurement "less then 10%". Our average error is 1.9% with worst case 5.6%, (Tab. 1).

dist	1.5 m	3 m	3.5 m	5.3 m	6 m	all
#	85	32	15	16	15	163
mean (%)	1.8	1.7	2.0	2.8	1.5	**1.9**
worst (%)	3.6	3.9	5.5	5.6	3.3	**5.6**

Table 1: Percentage error of absolute distance measurements. The error is evaluated as $|l_m - l_{gt}|/l_{gt} * 100\%$, where l_{gt} is the ground truth value and l_m is the distance measured by the presented algorithm. For each distance we evaluate the average and worst error. The numbers in the row labeled '#' are the number of measurements of the given length (from 5 videos).

When measuring the vehicle speed (Tab. 2), we take into account one corner of the bounding box which lies directly on the road). Vehicles in the video are tracked and their velocity is evaluated over the whole straight part of the track. The average speed of the vehicles was $75\frac{km}{h}$ and therefore 2% error causes $\pm 1.5\frac{km}{h}$ deviation. A similar evaluation was provided by Dailey [1] who used distribution of car lengths for scale calculation and reached average deviation $6.4\frac{km}{h}$ or by Grammatikopoulos [4] whose algorithm has accuracy $\pm 3\frac{km}{h}$ but requires manual distance measurements to obtain the scale.

	a (5)	b (3)	c (5)	d (5)	e (5)	f (5)	all (28)
mean (%)	2.39	2.90	1.49	1.65	1.31	2.58	**1.99**
worst (%)	3.47	3.63	3.18	3.77	2.40	4.26	**4.26**

Table 2: Percentage error in speed measurement. For obtaining the ground truth values, we drove cars with cruise control and get the speed from GPS. The error is evaluated as $|s_m - s_{gt}|/s_{gt} * 100\%$, where s_{gt} is speed from GPS and s_m is speed calculated by presented algorithm. The number in parentheses stands for the number of evaluated measurements for given video.

[1] D.J. Dailey, F.W. Cathey, and S. Pumrin. An algorithm to estimate mean traffic speed using uncalibrated cameras. *IEEE T-ITS*, 2000.

[2] M. Dubská and A. Herout. Real projective plane mapping for detection of orthogonal vanishing points. In *BMVC*, 2013.

[3] M. Dubská, A. Herout, J. Sochor, and R. Juránek. Fully automatic roadside camera calib. for traffic surv. *To appear in IEEE T-ITS*, 2014.

[4] L. Grammatikopoulos, G. Karras, and E. Petsa. Autom. estimation of vehicle speed from uncalibrated video seq. In *ISMTEPPGRF*, 2005.

[5] Z. Zhang, T. Tan, K. Huang, and Y. Wang. Practical camera calib. from moving objects for traffic scene surveill. *IEEE T-CSVT*, 2013.

Learning to Rank Bag-of-Word Histograms for Large-scale Object Retrieval

Danfeng Qin
http://www.vision.ee.ethz.ch/~qind/
Yuhua Chen
yuhchen@ee.ethz.ch
Matthieu Guillaumin
http://www.vision.ee.ethz.ch/~mguillau/
Luc Van Gool
http://www.vision.ee.ethz.ch/~vangool/

Computer Vision Laboratory
ETH Zurich
Switzerland

Retrieving images of a particular query object in a large database of images is an important problem for computer vision with applications in object discovery, 3D reconstruction, location recognition and mobile visual search. Most recent state-of-the-art large-scale image retrieval systems rely on local features, in particular the SIFT descriptor and its variants. Typically, those local descriptors are aggregated into a histogram-based representation of the image referred to as the Bag-of-Words model (BoW) [4]. BoW models considerably reduce the computational burden and the memory footprint of the systems, because local descriptors are quantised into *visual words*.

For BoW histograms, it is common to use simple similarity functions such as the inner product or cosine similarity. However, such functions are not optimal for modelling the visual similarity between BoW features and thus lead to sub-optimal performance for retrieval [2, 3, 6]. The potential problems are the following: a) The evidence coming from co-missing visual features is under-estimated [2]; b) The similarity between a query image and a database image should not be symmetric [6]; c) Statistical properties of visual words are not taken into account [1, 3, 5].

Even though different methods have been proposed to address each of these problems individually, none provides a satisfying solution to properly account for all of them. Moreover, most authors propose ad-hoc solutions by means of functions controlled by very few parameters. These parameters are then hand-tuned or exhaustively searched on validation/test data to adapt them to each dataset. In this work, our goal is to replace those ad-hoc similarities in measuring histograms with ones that are specifically trained to maximize the retrieval accuracy. We propose to use a simple and very general linear model whose weights directly represent the similarity values. We devise a variant of rank-SVM to learn those weights automatically from training data with fast convergence and we propose techniques to limit the weights to a tractable number to avoid overfitting. Importantly, the flexibility of our model allows us to seamlessly incorporate well-known image retrieval schemes such as burstiness, negative evidence and idf weighting, and still exploit inverted files for efficiency in the large-scale setting. In our experiments, as shown in Table. 1, our approach consistently and significantly outperforms the similarities used in several state-of-the-art systems on 4 standard benchmark datasets.

Most of existing similarity measures [2, 3, 6] can be written in a very general form as:

$$s(q,d) = \tau(q)\tau(d)\sum_{i=1}^{K} s_i(q_i,d_i). \tag{1}$$

Rather than trying to design τ and s_i manually, we propose to resort to learning and discover the patterns of a good similarity function for image search, automatically from training data. Looking at Eq. (1), we aim at learning the values $s_i(q_i,d_i)$ directly. This is notably impractical, as each q_i and d_i can be arbitrarily large. However, state-of-the-art methods use very large visual codebook ($K \approx 10^6$) leading to sparse of BoW representations, with few occurrences of any visual word in any given image. As a result, using a *truncated histogram* $\hat{q}_i = \min(q_i, n)$ with $n \in \mathbb{N}^+$ will provide an excellent approximation of the original histogram while limiting the number of possible values of $s_i(\hat{q}_i,\hat{d}_i)$ to $(n+1)^2$. Additionally, because we learn the values of $s_i(\hat{q}_i,\hat{d}_i)$ directly, these terms can be learned to incorporate a *contribution to the normalisation functions*. This leads to a modified similarity $\hat{s}_i$ and our approximated model becomes additive and writes as:

$$s(q,d) = \tau(q)\tau(d)\sum_{i=1}^{K} s_i(q_i,d_i) \approx \sum_{i=1}^{K} \hat{s}_i(\hat{q}_i,\hat{d}_i), \tag{2}$$

where $\hat{s}_i(j,l)$ for $j,l \in [0,n]$ are the $K\cdot(n+1)^2$ parameters to learn. Notably, this additive approximation allows to rewrite Eq. (2) as a linear combination of indicator functions:

$$\hat{s}_i(\hat{q}_i,\hat{d}_i) = w_{i\hat{q}_i\hat{d}_i} = \sum_{j=0}^{n}\sum_{l=0}^{n} w_{ijl}\mathbb{I}(\hat{q}_i = j)\mathbb{I}(\hat{d}_i = l), \tag{3}$$

where $w_{ijl} = \hat{s}_i(j,l)$. In other words, if we define $\Psi(q,d)$ as the binary vector indexed by (i,j,l) such that $\Psi_{ijl}(q,d) = \mathbb{I}(\hat{q}_i = j)\mathbb{I}(\hat{d}_i = l)$ and define $\mathbf{w} = [w_{ijl}]_{i,j,l}$, then:

$$s(q,d) \approx \mathbf{w}^\top \Psi(q,d). \tag{4}$$

Importantly, Eq. (4) highlights that Ψ acts as a feature encoding for the query-document pair (q,d) in a linear prediction model. Despite its simplicity, this model is very general and flexible, and is able to incorporate many of the properties discussed in [2, 3, 6], and potentially others, without having to explicitly model them. To illustrate this, let us first consider the simple case of $n = 1$. In such case, the truncated histogram $\hat{q}$ simply encodes the absence or presence of visual words (an encoding often referred to as *binary bag-of-words*), and there are only 4 weights to learn per visual word: co-absence $\hat{s}_i(0,0)$, co-occurrence $\hat{s}_i(1,1)$ and either case of mutual exclusion $\hat{s}_i(0,1)$ and $\hat{s}_i(1,0)$. If we learn that $\hat{s}_i(0,0) > \hat{s}_i(0,1)$, then not only have we implicitly learned that co-absence of the visual word i contribute more to the similarity than mutual exclusion (as argued by [2]) but also exactly by which amount. If we learn that $\hat{s}_i(0,1) \neq \hat{s}_i(1,0)$, then this implies that the ideal similarity is indeed asymmetric [6]. Finally, learning all the weights together allows to identify which visual words are more important than others, as indicated by the relative weight of $\hat{s}_i(1,1)$ and $\hat{s}_j(1,1)$. Hence, it automatically models re-weighting schemes such as IDF. Finally, when $n > 1$, phenomena such as burstiness [3] are also learnt.

	Oxford5k^s	Oxford105k^s	Holidays^s	UKbench^s
Cosine Similarity	0.819 (0)	0.725 (0)	0.862 (0)	3.51 (0)
Burstiness Weighting [3]	0.826 (0)	0.748 (0)	0.858 (0)	3.54 (0)
Negative Evidence [2]	0.830 (0)	0.684 (0)	0.848 (0)	3.44 (0)
Adaptive Asymmetric [6]	0.839 (1)	0.758 (0)	0.795 (0)	3.38 (0)
This paper	**0.870 (9)**	**0.816 (10)**	**0.871 (10)**	**3.70 (10)**

Table 1: Comparison to alternative similarities. We report the average performance over the 10 splits of the data (mAP or top-4 score depending on the dataset) and in parenthesis the number of runs where the method is the best. In bold is the best result for each dataset.

[1] Ondrej Chum, James Philbin, and Andrew Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *BMVC*, 2008.

[2] Hervé Jégou and Ondřej Chum. Negative evidences and cooccurences in image retrieval: The benefit of pca and whitening. In *Computer Vision–ECCV 2012*, pages 774–787. Springer, 2012.

[3] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1169–1176. IEEE, 2009.

[4] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[5] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. Lp-norm idf for large scale image search. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1626–1633. IEEE, 2013.

[6] Cai-Zhi Zhu, Hervé Jégou, and Shin'ichi Satoh. Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *ICCV-International Conference on Computer Vision*, 2013.

Board #11

Optimal Intrinsic Descriptors for Non-Rigid Shape Analysis

Thomas Windheuser
Matthias Vestner
Emanuele Rodolà
Rudolph Triebel
Daniel Cremers

Computer Vision Group,
Department of Computer Science,
Technische Universität München

We propose novel point descriptors for 3D shapes with the potential to match two shapes representing the same object undergoing natural deformations. These deformations are more general than the often assumed isometries, and we use labeled training data to learn optimal descriptors for such cases. Furthermore, instead of explicitly defining the descriptor, we introduce new Mercer kernels, for which we formally show that their corresponding feature space mapping is a generalization of either the Heat Kernel Signature (HKS) [3] or the Wave Kernel Signature (WKS) [1]. I.e. the proposed descriptors are guaranteed to be at least as precise as any Heat Kernel Signature or Wave Kernel Signature of any parameterisation.

A point descriptor $\phi : \mathcal{P} \to \mathbb{R}^T$ takes points from a set of shapes $\mathcal{P} := \cup_i \mathcal{M}_i$ and maps them to a space $\mathbb{R}^T$. Ideally, the descriptors of points that are at corresponding locations on the shapes should have a small distance in the descriptor space. Points at distinct locations on the shapes should be mapped to distinct locations in the descriptor space (see Figure 1).

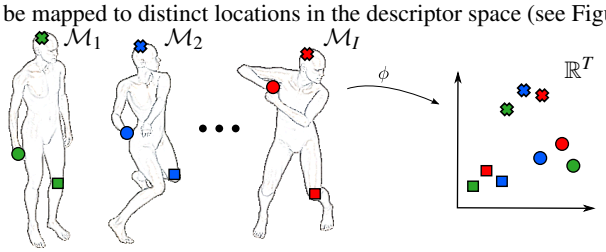


Figure 1: **Comparing points with a point descriptor:** Ideally the point descriptor $\phi : \mathcal{P} \to \mathbb{R}^T$ should map corresponding points to nearby locations and non-corresponding points to distinct locations.

In general one cannot assume that a given descriptor ϕ groups similar points as well as depicted in Fig. 1. The proposed method optimizes for the positive semi-definite matrix $M = L^T L$ inducing a pseudo distance in the descriptor space $\mathbb{R}^T$ via $d_M^2(x,y) = \langle x-y, x-y \rangle_M$ such that the point descriptors are grouped as good as possible. Optimizing for M is equivalent to looking for the best linear transformation L of the descriptor space with respect to the Euclidean distance, since $d_M(x,y) = \|L(x-y)\|$. In Figure 2 we see that L projects the images of ϕ onto the dotted line resulting in the much better descriptor $L \circ \phi$. As an optimization criterion for L we use LMNN [4] (see Figure 3).

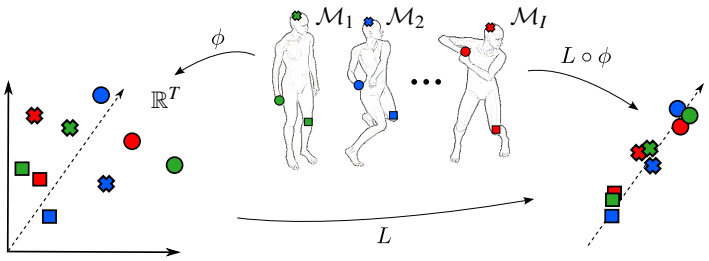


Figure 2: **Optimal distance in the descriptor space:** We are optimizing for a linear transformation L of the descriptor space $\mathbb{R}^T$. This is equivalent to a new distance $d_M^{\phi}(x,y) := \langle x-y, x-y \rangle_M$, where $M = L^\top L$, in the original descriptor space.

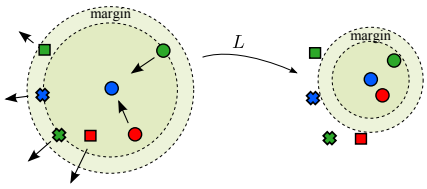
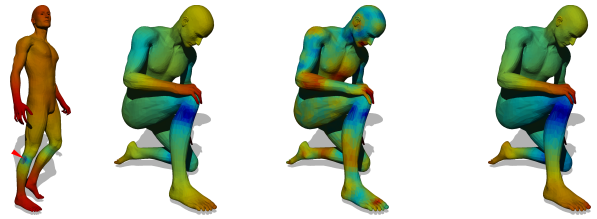


Figure 3: **Optimal LMNN distance [4]:** The neighbourhood of an input sample (blue circle) changes as a result of the training process. In this example, the learned distance is such that the nearest intra-class neighbours lie within a smaller radius after application of the linear mapping L. Similarly, the extra-class neighbours are left outside this optimized neighbourhood by a fixed margin.



reference shape | wave kernel signature [1] | learned cubic spline descriptor [2] | proposed descriptor with kernel $\kappa^{\mathcal{L}}$

Figure 4: **Qualitative comparison of descriptors:** Distance maps between the descriptor at a reference point (indicated by a red arrow) and the descriptors computed on the shape after deformation. Colours range from blue (small distance) to red (large distance). Qualitatively, WKS and the proposed method do very well at indicating the right location while the cubic spline descriptor exhibits several local minima across the shape. Both test shapes are from the class *michael* (TOSCA), whereas the proposed descriptor and the spline descriptor were trained on the class *david*. The distances on the reference shape a generated by the proposed method.

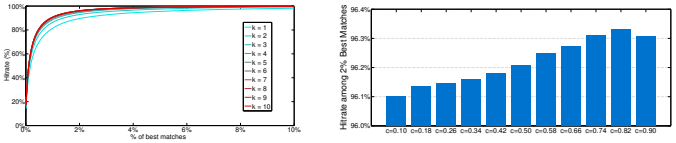


Figure 5: **Parameter sensitivity analysis Left:** Precision of the learned descriptor on the test set *michael* with kernel $\kappa^{\mathcal{L}}$, fixed $c = 0.5$ depending on different values for learning parameter k. Precision directly increases with higher values of k. **Right:** Same experimental setup as on the left, with fixed $k = 7$ and different values of c. The plot shows the hitrate when looking at the 2% best matches. The difference in precision among different values of c is only visible in this close-up.

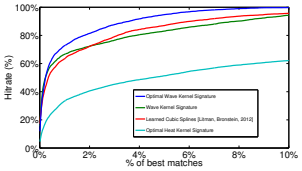


Figure 6: **Descriptor comparison** CMC curves of different descriptors on the TOSCA dataset; the method proposed in this paper with kernel $\kappa^{\mathcal{L}}$ is plotted as a blue curve.

The contributions of our paper can be summarized as follows:

- The method eliminates the need of tuning descriptor parameters. Neither does it have time parameters such as the HKS, nor do we need to choose the dimensionality of the descriptor as in [2]. In contrast, the adjustment of the descriptor is completely driven by the data, i. e. the shapes' deformations fed to the training process. The only two parameters of the objective function are directly related to the descriptor precision. Experiments suggest they can be fixed to constant values across applications, making the framework virtually parameter free.
- The method is a true generalization of the WKS and HKS and can potentially generalize other descriptors as well. Most importantly, we formally show that the proposed descriptors are guaranteed to be at least as accurate as WKS and HKS under any parameterisation with respect to the given shapes. Applications using WKS or HKS can avoid the parameter tuning problem by plugging in the proposed descriptor and are guaranteed to get optimal precision.

[1] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops)*, 2011.

[2] Roee Litman and Alexander M Bronstein. Learning spectral descriptors for deformable shape correspondence. *Transactions on Pattern Analysis and Machine Intelligence*, 1(1), 2013.

[3] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer Graphics Forum*, volume 28, pages 1383–1392, 2009.

[4] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

Fully Associative Ensemble Learning for Hierarchical Multi-Label Classification

Lingfeng Zhang
lzhang34@uh.edu
Shishir K. Shah
sshah@central.uh.edu
Ioannis A. Kakadiaris
ioannisk@uh.edu

Computational Biomedicine Lab
Department of Computer Science
University of Houston
Houston, TX, USA

In Hierarchical Multi-label Classification (HMC), rich hierarchical information is used to improve classification performance. Global approaches learn a single model for the whole class hierarchy [3, 6]. Local approaches introduce hierarchical information to the local prediction results of all the local classifiers to obtain the global prediction results for all the nodes [2, 5].

In this paper, we propose a novel local HMC framework, Fully Associative Ensemble Learning (FAEL). Specifically, a multi-variable regression model is built to minimize the empirical loss between the global predictions of all the training samples and their corresponding true label observations. Let X and Y represent local prediction matrix and label observation matrix, respectively. We define $W = \{w_{ij}\}$ as a weight matrix, where w_{ij} represents the weight of the i^{th} label's local prediction to the j^{th} label's global prediction. In the basic model, the objective function is:

$$\min_W \|Y - XW\|_F^2 + \lambda_1 \|W\|_F^2, \tag{1}$$

where the first term measures the empirical loss of the training set, the second term controls the generalization error, and λ_1 is a regularization parameter. The above function is known as ridge regression. We have:

$$W = \left(X^T X + \lambda_1 I_l\right)^{-1} X^T Y, \tag{2}$$

where I_l represents the $l \times l$ identity matrix.

To capture the complex correlation between global and local prediction, we can generalize the above basic model using the kernel trick. Let Φ represent the map applied to each example's local prediction vector $\mathbf{x}_i$. A kernel function is induced by $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$. By replacing the term X in (1), we obtain:

$$\min_{W_k} \|Y - \Phi W_k\|_F^2 + \lambda_1 \|W_k\|_F^2. \tag{3}$$

After several matrix manipulations [1], the solution of W_k becomes:

$$W_k = \left(\Phi^T \Phi + \lambda_1 I_l\right)^{-1} \Phi^T Y = \Phi^T \left(\Phi \Phi^T + \lambda_1 I_n\right)^{-1} Y, \tag{4}$$

where I_n represents the $n \times n$ identity matrix. For a given testing example s^t and its local prediction $\mathbf{x}^t$, the global prediction $\widehat{\mathbf{y}}^t$ is obtained by $\widehat{\mathbf{y}}^t = \mathbf{x}^t W$. For a kernel version, we obtain:

$$\widehat{\mathbf{y}}_k^t = K(\mathbf{x}^t, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \lambda_1 I_n\right)^{-1} Y. \tag{5}$$

To make full use of the hierarchical relationships between different nodes, we introduce a regularization term to the optimization function in (1). Let $\mathcal{R} = \{r_i(c_p, c_q)\}$ denote the binary constraint set of hierarchy $\mathcal{H}$. Each member $r_i(c_p, c_q)$ meets either $c_p =\uparrow c_q$ or $c_p =\Uparrow c_q$, where "$\uparrow$" and "$\Uparrow$" represent the "parent-child" constraint and the "ancestor-descendent" constraint, respectively. We introduce a weight restriction to each pair of nodes in $\mathcal{R}$. Define coefficient $m_{pq} \in \mathbb{R}^+$ for the i^{th} pair $r_i(c_p, c_q)$, so that:

$$w_{pk} = m_{pq} * w_{qk}. \tag{6}$$

For the global prediction of node k, the weight of node p is m_{pq} times the weight of node q. The value of m_{pq} is set by:

$$m_{pq} = \begin{cases} \mu & c_p =\uparrow c_q \\ \mu * (e_{pq} + 1) & c_p =\Uparrow c_q \end{cases}, \tag{7}$$

where μ is a positive constant and e_{pq} represents the number of nodes between p and q. All the restrictions over the hierarchy are summarized as:

$$\sum_{r_i(c_p,c_q)}^{|\mathcal{R}|} \sum_{k=1}^{l} \left(w_{pk} - m_{pq} * w_{qk}\right)^2. \tag{8}$$

Figure 1: (Top) The "human" sub-hierarchy. (Bottom) The weight matrix W^* learned from B-FAEL.

We introduce a sparse matrix $M = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_{|R|}]^T$, in which the i^{th} row $\mathbf{m}_i$ corresponds to the i^{th} pair in $\mathcal{R}$. Each row in M has only two non-zero entries. The p^{th} entry is 1 and the q^{th} entry is $-m_{pq}$, all the other entries are zero. Thus, we obtain the regularization term of the binary constraint model:

$$\sum_{r_i(c_p,c_q)}^{|\mathcal{R}|} \sum_{k=1}^{l} \left(w_{pk} - m_{pq} * w_{qk}\right)^2 = \|MW_b\|_F^2. \tag{9}$$

Adding this term to (1), the optimization function becomes:

$$\min_W \|Y - XW_b\|_F^2 + \lambda_1 \|W_b\|_F^2 + \lambda_2 \|MW_b\|_F^2. \tag{10}$$

The analytical solution of the binary constraint model is given by:

$$W_b = \left(X^T X + \lambda_1 I_l + \lambda_2 M^T M\right)^{-1} X^T Y. \tag{11}$$

Take the "human" sub-hierarchy from the extended IAPR TC-12 image dataset [4] for example, Figure 1 depicts the merits of our model and shows the contribution of hierarchical and sibling nodes on each local prediction. The weight matrix computed indicates that each local node influences its own decision positively while nodes not directly connected in the hierarchy provide a negative influence.

[1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, Minneapolis, MN, USA, 2007.

[2] Z. Barutcuoglu and C. DeCoro. Hierarchical shape classification using bayesian aggregation. In *Proc. IEEE International Conference on Shape Modeling and Applications*, Matsushima, Japan, 2006.

[3] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10): 2436–2449, 2011.

[4] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428, 2010.

[5] G. Valentini. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):832–847, 2011.

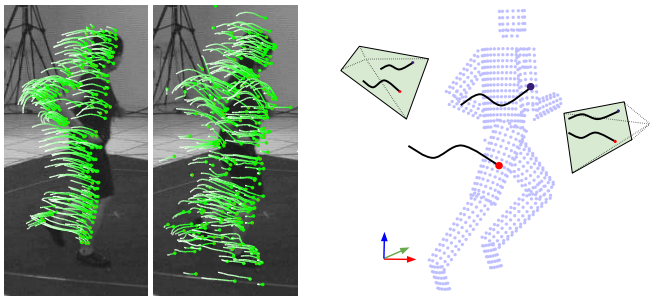
[6] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

Board #13

Unlabelled 3D Motion Examples Improve Cross-View Action Recognition

Ankur Gupta
ankgupta@cs.ubc.ca
Alireza Shafaei
shafaei@cs.ubc.ca
James J. Little
little@cs.ubc.ca
Robert J. Woodham
woodham@cs.ubc.ca

Department of Computer Science
University of British Columbia
Vancouver, Canada



(a) (b)

Figure 1: (a) We exploit the visual similarity between mocap-generated trajectories (left) and dense trajectories (right) to improve cross-view action recognition. (b) For mocap-trajectories, we can easily obtain corresponding features (i.e., descriptors for trajectories that originate from the same 3D point) in two views. We use these pairs of features to learn the transformation function for viewpoint change.

1 Overview

A view-invariant representation of human motion is crucial for effective action recognition. However, most view-invariant representations require either tracking of body parts or multi-view video data for learning which may not be a practical approach in many real-life scenarios. We describe a view-independent model for human action which is flexible, action-independent, and requires no multi-view video data or additional labelling effort.

We present a novel method for cross-view action recognition. Using a large collection of motion capture data we synthesize mocap-trajectory features from multiple viewpoints. Features originating from the same 3D point on the surface correspond, and this allows us to learn a feature transformation function for viewpoint change. Given this function, we can "hallucinate" the action descriptors of a video for different viewing angles. We use these hallucinated examples as additional training data to make our model view-invariant. We demonstrate the effectiveness of our approach on the unsupervised scenario of the INRIA IXMAS dataset.

2 Methodology

The approach has three steps:

Generating training data We adapt the mocap trajectory generation pipeline of Gupta *et al*. [1], which uses a human model with cylindrical primitives (see Figure 1(b)). Each limb consists of a collection of points that are placed on a 3D surface. Given a camera viewpoint, these points are projected under orthographic projection and tracked for L(=15) consecutive frames to generate trajectory descriptors similar to the dense-trajectories of Wang *et al*. [3]. The resulting displacement vectors are then used to generate trajectory features. Given two arbitrary viewpoints, we can find a correspondence between features that originate from the same point on the surface (see Figure 1(b)).

Learning the transformation function We quantize the mocap trajectory features using a fixed codebook $\mathcal{C}$ of size n. Given a source camera elevation angle θ and relative change in viewpoint given by $\Delta = (\delta\theta, \delta\phi)$, we define the training set $\mathcal{D}_\theta^\Delta = \{(f_i, g_i)\}_1^m$ to be the set of m pairs $(f,g) \in \mathcal{C} \times \mathcal{C}$, where f_i and g_i are the codewords for two corresponding trajectory features.

Given the training data $\mathcal{D}_\theta^\Delta$, we can learn a joint probability mass function $\mathrm{P}(F,G)$ which captures the probability of having feature pairs (f_i, g_i) in $\mathcal{D}_\theta^\Delta$. We calculate the empirical probability by counting the co-occurrences of (f_i, g_i) in $\mathcal{D}_\theta^\Delta$ followed by normalization. After observing an instance of codeword f_i in the source view, $\mathrm{P}(G|F = f_i)$ allows us to infer the possible outcomes in the target view.

Synthesizing cross-view descriptors Given a BoW descriptor of an action, we wish to synthesize another descriptor for a viewpoint $\Delta = (\delta\theta, \delta\phi)$ away from the original view. Let $\mathbf{x} = [x_1, \ldots, x_n]^\mathrm{T}$ be the BoW descriptor in the source view, and $\mathbf{y} = [y_1, \ldots, y_n]^\mathrm{T}$ be the descriptor we want to estimate. Using the probabilistic mapping between the codewords across views, we return an expected transformed descriptor

$$\bar{\mathbf{y}} = [\mathbb{E}[y_1], \ldots, \mathbb{E}[y_n]]^\mathrm{T} \text{ and } \mathbb{E}[y_j] = \sum_{i=1}^{n} x_i \cdot \mathrm{P}(G = f_j | F = f_i)$$

By organizing $\mathrm{P}(G|F)$ in the form of a matrix (say N) where the i-th row is the categorical distribution $\mathrm{P}(G|F = f_i)$, we can rewrite the above formulation as a matrix multiplication $\bar{\mathbf{y}} = \mathrm{N}^\mathrm{T}\mathbf{x}$. We further l_2 normalize $\bar{\mathbf{y}}$ to make it consistent with the original descriptor.

Method	Average accuracy
Ours	**71.7%**
nCTE based matching [1]	67.4%
w/o aug.	62.1%
Hankelets [2]	56.4%

Table 1: Average accuracy for action recognition over all view pairs of the INRIA IXMAS dataset. Given the training data from one viewing angle, the task is to recognize actions from a previously unseen viewpoint. We compare with other state-of-the-art methods. *w/o aug.* is our baseline without any data augmentation.

3 Experiments

To test our method we use the INRIA IXMAS dataset which has short view clips of 10 actors performing 11 activities (3 trials each) captured from 5 diverse angles. To learn the mapping between codewords, we generate mocap trajectories from multiple viewpoints and quantize them using the same codebook $\mathcal{C}$. We also quantize the viewpoints into 18 bins.

We synthesize multiple descriptors per training examples (one per viewpoint change), as described above, to augment our original training data. We train an SVM with χ^2 kernel using one-vs-all strategy. The main results are summarized in Table 1. Our code is publicly available: `http://cs.ubc.ca/research/motion-view-translation/`.

[1] Ankur Gupta, Julieta Martinez, James J. Little, and Robert J. Woodham. 3D Pose from Motion for Cross-view Action Recognition via Non-linear Circulant Temporal Encoding. In *CVPR*, 2014.

[2] Binlong Li, Octavia I. Camps, and Mario Sznaier. Cross-view Activity Recognition using Hankelets. In *CVPR*, 2012.

[3] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

Board #14

Location Constrained Pixel Classifiers for Image Parsing with Regular Spatial Layout

Kang Dang
kangdang@gmail.com
Junsong Yuan
jsyuan@ntu.edu.sg
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Location is useful for a variety of image parsing problems with regular spatial layout, such as pedestrian parsing after detection, street view scene parsing and medical image segmentation. This paper proposes a novel way to leverage both location and appearance information for pixel labeling (Fig. 1).

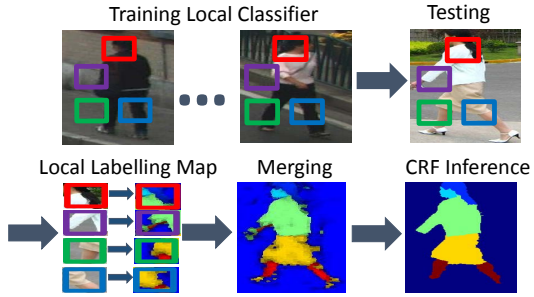


Figure 1: Overview of our method. (1) At each location we train a position dependent local pixel classifier with training pixel samples from its neighborhood region represented by a patch. (2) Assume the training and testing images have similar layout, the trained classifier is used for the same region in the testing images. (3) To ensure smoothed labeling results, we allow the local classifiers to overlap with each other, such that each pixel will be voted by multiple local classifiers. The final score is the average score of all engaged local classifiers. (4) The final result is obtained after a proper discretization of the labeling map with a conditional random field (CRF).

Existing approaches solve the pixel labeling problem with a single global model. In other words, they learn a single global pixel classifier for the entire image space, and all pixels of the image are used to train the classifier. In contrast, at each image location we learn a location constrained classifier, i.e. local classifier. Since each local pixel classifier is learned by only using the pixels in a local neighborhood, it is expected to better fit the local pixel distribution and capture local discriminative information. To prevent local classifiers overly depending on the image location and to improve the generalization, the neighborhood scale of local learning is important. We justify the significance of the neighborhood scale via the following theoretical studies.

Probabilistic Analysis. Given a pixel's central position (x,y) and its associated feature vector $\mathbf{f}$, our goal is to predict the class label L of that pixel. We are interested in learning a number of local classifiers $p_{\mathcal{N}}(L \mid \mathbf{f})$ at different spatial locations. $\mathcal{N}(x,y,s)$ stands for a local image neighborhood, which is a patch centered at (x,y) and of width $s \times \mathcal{W}$ and height $s \times \mathcal{H}$, where s is the neighborhood scale and $\mathcal{W}$ and $\mathcal{H}$ is the width and height of the image. In other words, the training set for each local classifier is $\{(L_i,\mathbf{f}_i) \mid \forall (x_i,y_i) \in \mathcal{N}(x,y,s)\}$. We show the local classifier approximates the following conditional distribution:

$$p_{\mathcal{N}(x,y,s)}(L \mid \mathbf{f}) \propto \sum_{(x,y)\in\mathcal{N}(x,y,s)} p(L \mid \mathbf{f},x,y)p(\mathbf{f} \mid x,y). \quad (1)$$

We see the proposed local classifier $p_{\mathcal{N}}(L \mid \mathbf{f})$ is a spatially smoothed version of the global classifier $p(L \mid \mathbf{f},x,y)$ in a local neighborhood, where the weight $p(\mathbf{f} \mid x,y)$ characterizes the dependency of the observed feature $\mathbf{f}$ at the pixel location (x,y). The neighborhood scale s plays an important role in building the local classifier. On one hand, when the local neighborhood contains only a single pixel, *i.e.*, $s=0$, our local classifier degenerates into: $p_{\mathcal{N}(x,y,0)}(L \mid \mathbf{f}) = p(L \mid \mathbf{f},x,y)$. On the other hand, when the local neighborhood expands to the entire image, *i.e.*, $s=1$, it becomes $p_{\mathcal{N}(x,y,1)}(L \mid \mathbf{f}) = p(L \mid \mathbf{f})$, which indicates position information (x,y) is not utilized at all. Our proposed classifier is a compromise between these two ends.

Bias-Variance Trade-Off. We discuss the implication of choosing an appropriate neighborhood scale s from the perspective of bias-variance analysis. Our main conclusion is a theorem stating that under certain assumptions, testing error variance monotonically decreases with the neighborhood scale s. In addition, our simulation shows that the bias increases with the neighborhood scale. Thus, an appropriate neighborhood scale is essential for balancing the bias and variance and minimizing the testing error.

	Penn-Fudan	PPSS
Feature Only	45.1	31.8
$(\mathbf{f},x,y)$ + SVM	54.3	39.7
$(\mathbf{f},x,y)$ + Boosting	60.3	45.1
Product of Expert	52.6	45.1
Ours	**63.1**	**53.5**

Penn-Fudan	
SBP[1]	57.3
P&S[4]	55.0
DL[3]	59.9
Ours	**63.1**
PPSS	
DDN[3]	47.2
Ours	**53.5**

Table 1: Benchmark results for Penn-Fudan and PPSS dataset. The performance metric is the average intersection over union(IOU) score over all labels. We compare our approach with three common methods of feature fusion and the state of arts. (1) $(\mathbf{f},x,y)$ + SVM: we concatenate feature and position information together to form $(\mathbf{f},x,y)$, and put it into a SVM classifier. (2) $(\mathbf{f},x,y)$ + Boosting: we put the concatenated feature vector $(\mathbf{f},x,y)$ into a joint boosting classifier. (3) Product of Experts: the merge is done by multiplying the two posterior probability map with weighting: $\frac{p(L|x,y)^k p(L|\mathbf{f})^{(1-k)}}{Z}$, where k is between 0 and 1, and Z is a normalization constant.

Figure 2: Image results from Penn-Fudan dataset. Visual quality is generally better than SBP[1].

Experiments. Our experimental evaluation is performed on two pedestrian parsing datasets Penn-Fudan [1] and PPSS dataset [3] as well as Weizmann horse segmentation[2]. Albeit simple, our proposed local learning works surprisingly well in these challenging image parsing problems. Some quantitative and qualitative results for pedestrian parsing datasets are shown in Table. 1 and Fig. 2. It confirms the advantages of our local classifiers which are better adapted to the local image characteristics than a global classifier.

[1] Yihang Bo and Charless C Fowlkes. Shape-based pedestrian parsing. In *CVPR*. IEEE, 2011.

[2] Eran Borenstein and Shimon Ullman. Class-specific, top-down segmentation. In *ECCV*. Springer, 2002.

[3] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *ICCV*. IEEE, 2013.

[4] Ingmar Rauschert and Robert T Collins. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In *ECCV*. Springer, 2012.

Unsupervised Learning of Generative Topic Saliency for Person Re-identification

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

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

Existing approaches to person re-identification (re-id) are dominated by supervised learning based methods, which requires a large number of manually labelled pairs of person images across every pair of camera views. This thus limits their ability to scale to large camera networks. To overcome this problem, a novel unsupervised re-id model, Generative Topic Saliency (GTS), is proposed in this paper for localised human appearance saliency selection in re-id by exploiting unsupervised generative topic modelling. It yields state-of-the-art re-id performance against existing unsupervised learning based re-id methods. For supervised methods, it also retains comparable re-id accuracy but without any need for pairwise labelled training data.

We are motivated by a very intuitive principle – humans often identify people by their salient appearances and ignore the more common traits in people's appearance. Compared to the pioneering work of [2] which is also based on learning appearance saliency for re-id, our model has two advantages: (1) Interpretability - our work explicitly models human appearances and backgrounds through learning a set of latent topics corresponding to localised human appearance components and also image backgrounds, so that the background cannot be mistaken as distractions to true foreground local salient region discovery. In addition, through associating saliency with *atypical* human appearances, the learned saliency is also more interpretable by human sense. (2) Complexity - only a *single* model is needed for computing saliency for all the images in a camera view, instead of learning a different discriminative saliency model (k-NN or one-class SVM) for every patches of every image.

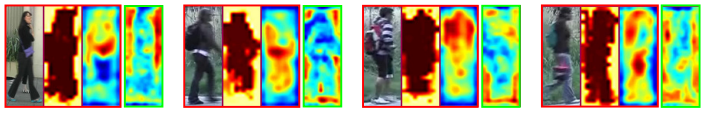


Figure 1: Saliency maps comparison (left to right): A person image in detected bounding box, GTS-computed background map, GTS-computed saliency map, saliency map computed by the model of [2] (green bounding box).

Our model is a generalisation of the Latent Dirichlet Allocation (LDA) model [1] with an added spatial variable to make the learned topics spatially coherent. Given a dataset of M images, each image will be factorised (clustered) into a unique combination of K shared topics, with each topic generating its own proportion of words on that image. Conceptually, one topic encodes a certain distribution of visual words (patches), whose vocabulary and spatial location revealing certain patterns, in our case the visual characteristics of human appearances and backgrounds. We thus learn two types of latent topics in our model corresponding to foreground and background respectively. Since foreground appearance are in general more 'compact' than background, we choose a Gaussian distribution to encode foreground human appearance topics and a Uniform distribution to encode more spread-out background topics.

A key objective of our model is to discover salient local foreground patches in a person's image that make the person stand out from other people, i.e. the model seeks not only visually distinctive but also *atypical* localised appearance characteristics of a person. In specific, we define a patch P_A's saliency according to three factors: The first one is how *unlikely* this patch will appear in a training set $\mathcal{I}^R$ of J images at the proximity of a particular spatial location in the images (i.e. its prevalence level). The less likely P_A repeatedly appears, the higher saliency score it should possess. Second, a patch with high probability of belonging to background topics should have low saliency scores. Third, even if a patch belongs to a human appearance topic, but if this topic is very dominant/popular in the training dataset (e.g. many people wearing jeans), the patch also should have low saliency score. With $Prevalence(P_A)$ measuring the prevalence level of P_A, Z_A denoting P_A's topic, T^{cb} the set of camera background topics, T^{pop} the set of *popular* human appearance topics, L and H the learned latent variables set and hyper-parameter set, patch P_A's saliency score is computed by:

$$\begin{aligned} Saliency(P_A) = h(Prevalence(P_A)) - \eta_1 \cdot \sum_{t_k \in T^{cb}} Pr(z_A = t_k | L, H) \\ - \eta_2 \cdot \sum_{t_k \in T^{pop}} Pr(z_A = t_k | L, H), \qquad 0 < \eta_1, \eta_2 < 1 \end{aligned} \tag{1}$$

where $h(x)$ is a inverse function defined as taking the additive inverse and normalising the result into the $[0,1]$ interval. The prevalence of P_A and the probability for P_A's topic Z_A falling into background topics and dominant/popular human appearance topics can all be computed from our model parameters inferred from training set. η_1, η_2 are the latter two factors' weights to affect the saliency score, determined by cross-validation. If one considers that $Prevalence(P_A)$ simply measures how likely the exact same patch appears repeatedly across images, its topic's *popularity* (the third component) takes much larger amounts of patches into consideration. These patches may even be visually different from P_A, but they are inherently related by the same topic. This model avoids the topic being simply data-driven; it also considers more inherent structure of the large-scaled data. The comparison between computed saliency are shown in Fig. 1.

Given the patch level saliency score, we adopt the same patch-based image matching scheme in [2]. In this patch-matching scheme, patches with higher saliency scores will contribute more to the distance between a pair of probe/gallery images. We conduct 10-trail experiments on both VIPeR and iLIDS dataset, compared with existing unsupervised learning methods, the GTS model improves re-id accuracy significantly, especially on Rank-1. The GTS model is also competitive against the state-of-the-art supervised learning based methods, but without requiring manual labelling of data, resulting in greater scalability to large scale re-id problems in many practical applications.

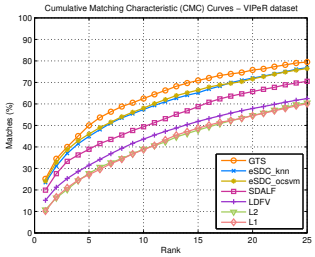


Figure 2: VIPeR test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
ELF	12.00	31.50	44.00	61.00
PRDC	15.66	38.42	53.86	70.09
PCCA	19.27	48.89	64.91	80.28
LMNN-R	20.00	49.00	66.00	79.00
KISSME	19.46	48.10	62.50	78.32
RPLM	27.00	-	69.00	83.00
LF	24.18	-	67.12	-
GTS	25.15	50.03	62.50	75.76

Table 1: VIPeR test: Comparing the GTS model to supervised learning based models.

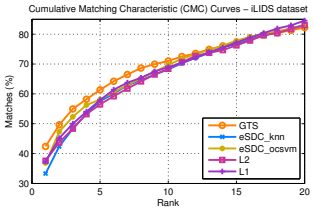


Figure 3: iLIDS test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
SDC_knn	33.31	57.55	68.22	83.13
SDC_ocsvm	36.81	58.10	69.69	82.94
PRDC	37.83	63.70	75.09	88.35
LMNN	27.97	53.75	66.14	82.33
PLS	22.10	46.04	59.95	78.68
ITM	28.96	53.99	70.50	86.67
GTS	42.39	61.35	71.04	82.21

Table 2: iLIDS test: Comparing the GTS model against other unsupervised (top) and supervised (bottom) learning based models.

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022, March 2003.

[2] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

Regularized ℓ^1-Graph for Data Clustering

Yingzhen Yang[1]
yyang58@ifp.uiuc.edu
Zhangyang Wang[1]
zwang119@ifp.uiuc.edu
Jianchao Yang[2]
jiayang@adobe.com
Jiawei Han[1]
hanj@cs.uiuc.edu
Thomas S. Huang[1]
huang@ifp.uiuc.edu

[1] University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
[2] Adobe Research
San Jose, CA 95110, USA

ℓ^1-Graph has been proven to be effective in data clustering, which partitions the data space by using the sparse representation of the data as the similarity measure. However, the sparse representation is performed for each datum independently without taking into account the geometric structure of the data. Motivated by ℓ^1-Graph and manifold leaning, we propose Regularized ℓ^1-Graph (Rℓ^1-Graph) for data clustering. Compared to ℓ^1-Graph, the sparse representations of Rℓ^1-Graph are regularized by the geometric information of the data. In accordance with the manifold assumption, the sparse representations vary smoothly along the geodesics of the data manifold through the graph Laplacian constructed by the sparse codes. Experimental results on various data sets demonstrate the superiority of our algorithm compared to ℓ^1-Graph and other competing clustering methods.

ℓ^1-graph [2, 3], which builds the graph by reconstructing each datum with all the other data, has been shown to be robust to noise and capable of producing superior results for high dimensional data, compared to K-means and spectral clustering. Compared to k-nearest-neighbor graph and ε-ball graph, ℓ^1-graph adaptively determines the neighborhood of each datum by solving sparse representation problem locally. Given the data $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, ℓ^1-graph seeks for the robust sparse representation for the entire data by solving the ℓ_1-norm optimization problem for each data point:

$$\min_{\boldsymbol{\alpha}^i} \|\boldsymbol{\alpha}^i\|_1 \quad s.t.\ \mathbf{x}_i = \boldsymbol{X}\boldsymbol{\alpha}^i \quad i = 1, \ldots, n \tag{1}$$

where $\boldsymbol{\alpha}^i \in \mathbb{R}^{n\times 1}$, and we denote by $\boldsymbol{\alpha}$ the coefficient matrix $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n] \in \mathbb{R}^{n\times n}$ with the element $\boldsymbol{\alpha}_{ij} = \boldsymbol{\alpha}_i^j$. Let $G = (\boldsymbol{X}, \mathbf{W})$ be the ℓ^1-graph where $\boldsymbol{X}$ is the set of vertices, $\mathbf{W}$ is the graph weight matrix and $\mathbf{W}_{ij}$ indicates the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. ℓ^1-graph sets the $n \times n$ matrix $\mathbf{W}$ as

$$\mathbf{W} = (|\boldsymbol{\alpha}| + |\boldsymbol{\alpha}^T|)/2 \tag{2}$$

where $|\boldsymbol{\alpha}|$ is the matrix whose elements are the absolute values of $\boldsymbol{\alpha}$, and then feed $\mathbf{W}$ as the pairwise similarity matrix into the spectral clustering algorithm to get the clustering result.

While ℓ^1-graph demonstrates better performance than many traditional similarity-based clustering methods, it performs sparse representation for each datum independently without considering the geometric information and manifold structure of the entire data. In order to obtain the sparse representations that account for the geometric information and manifold structure of the data, we employ the manifold assumption [1] and propose a novel Regularized ℓ^1-Graph (Rℓ^1-Graph). The manifold assumption in this case requires that if two points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the intrinsic geometry of the submanifold, their corresponding sparse codes $\boldsymbol{\alpha}^i$ and $\boldsymbol{\alpha}^j$ are also expected to be similar to each other. The following objective function for Rℓ^1-Graph is given below:

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\mathbf{W}} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_1 + \gamma \mathrm{Tr}(\boldsymbol{\alpha}\mathbf{L}_{\mathbf{W}}\boldsymbol{\alpha}^T) \\ &s.t.\ \mathbf{W} = (\mathbf{A}\circ|\boldsymbol{\alpha}| + \mathbf{A}^T \circ |\boldsymbol{\alpha}^T|)/2 \quad \boldsymbol{\alpha} \in S \end{aligned} \tag{3}$$

where $S = \{\boldsymbol{\alpha} \in \mathbb{R}^{n\times n} | \boldsymbol{\alpha}_{ii} = 0, 1 \le i \le n\}$, $\lambda > 0$ is the weight controlling the sparsity of the coefficients, and $\gamma > 0$ is the weight of the regularization term, $\mathbf{L}_{\mathbf{W}}$ is the graph Laplacian matrix constructed by the pairwise similarity matrix $\mathbf{W}$, $\mathbf{A}$ is a KNN adjacency matrix.

We simplified the optimization problem (3), and employ Alternating Direction Method of Multipliers (ADMM) to solve the nonconvex optimization problem. ADMM decomposes the original problem into a sequence of tractable subproblems which can be solved efficiently.

We demonstrate the performance of Rℓ^1-Graph with comparison to other competing methods, i.e. K-means (KM), Spectral Clustering (SC), ℓ^1-Graph and Laplacian regularized ℓ^1-Graph. There are two parameters that influence the regularization term in Rℓ^1-Graph, namely the weight of the regularization γ and the number of nearest neighbors K of the KNN adjacency matrix. The regularization term imposes stronger smoothness constraint on the sparse codes with larger γ and K, and vice versa. We investigate how the clustering accuracy on ORL face database changes when varying these two parameters, and illustrate the result in Figure 1. We observe that the performance of Rℓ^1-Graph is much better than other algorithms over a large range of both γ and K, revealing the robustness of our algorithm. Please refer to the paper for detailed description of our algorithm and more experimental results.

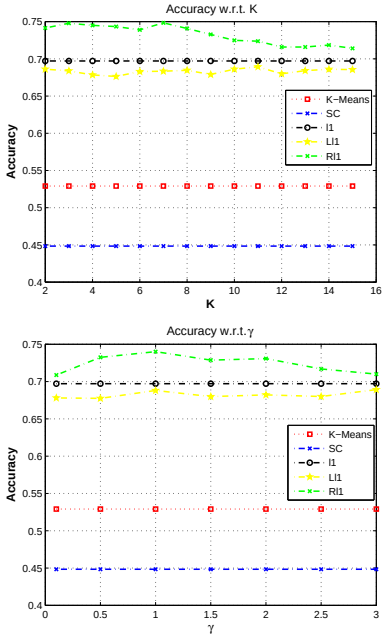


Figure 1: Clustering accuracy with different values of K and γ on ORL face database. Upper: K; Down: γ

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] Bin Cheng, Jianchao Yang, Shuicheng Yan, Yun Fu, and Thomas S. Huang. Learning with l1-graph for image analysis. *IEEE Transactions on Image Processing*, 19(4):858–866, 2010.

[3] Shuicheng Yan and Huan Wang. Semi-supervised learning by sparse representation. In *SDM*, pages 792–801, 2009.

Essential Matrix Estimation Using Adaptive Penalty Formulations

Mohammed E. Fathy
mefathy@cs.umd.edu
Michael C. Rotkowitz
mcrotk@umd.edu
University of Maryland, College Park
Maryland, USA

The Problem Given six or more pairs of corresponding points on two calibrated images, the accurate estimation of the essential matrix (EsM), which is a 3×3 matrix capturing the relative translation $\boldsymbol{t}$ and rotation $\mathbf{R}$ separating the two pinhole cameras, requires solving a nonlinear optimization problem subject to a set of constraints that guarantee the resulting 3×3 matrix has the structure of a valid EsM (i.e. $\mathbf{E}=[\boldsymbol{t}]_{\times}\mathbf{R}$, or equivalently $\mathrm{svd}(\mathbf{E})=\mathbf{U}\,\mathrm{diag}(1,1,0)\mathbf{V}'$, or equivalently $\mathbf{E}'\mathbf{E}\mathbf{E}'=0.5\,\mathrm{tr}(\mathbf{E}'\mathbf{E})\mathbf{E}'$). To the best of our knowledge, all existing schemes enforce the EsM constraints by performing the optimization on the manifold $\mathcal{E}$ of EsMs using either global [2] or local parametrizations [3]. No attempts were made to use the more straightforward approach of integrating the EsM constraint $\mathbf{E}'\mathbf{E}\mathbf{E}'=0.5\,\mathrm{tr}(\mathbf{E}'\mathbf{E})\mathbf{E}'$ directly into the optimization possibly because this 3×3 matrix equation as well as the homogeneity property of the EsM (i.e. $\mathbf{E}$ and $c\mathbf{E}$ represent the same EsM for all $c\neq0$) give a total of ten (nonlinearly dependent) constraints while the number of variables in a 3×3 matrix is only nine.

Idea To avoid this problem, we propose to use adaptive penalty methods [1] to incorporate the matrix constraint into the optimization. Penalty methods relax the constraints (and so do not suffer from the too-many-constraints problem) while making violating them expensive. Assuming that $f(\boldsymbol{e})$ is the cost function measuring the (robust) algebraic or geometric fitting error of the 9-vector $\boldsymbol{e}$ corresponding to $\mathbf{E}$ and $\boldsymbol{h}_2(\boldsymbol{e})=\mathrm{vec}\{\mathbf{E}'\mathbf{E}\mathbf{E}'-0.5\,\mathrm{tr}(\mathbf{E}'\mathbf{E})\mathbf{E}'\}$ is the EsM constraint function, we define the penalty-augmented cost function $f_c(\boldsymbol{e})=f(\boldsymbol{e})+0.5c||\boldsymbol{h}_2(\boldsymbol{e})||^2$ where $c>0$ is called the penalty parameter. The two functions $f(\boldsymbol{e})$ and $f_c(\boldsymbol{e})$ are equal iff $\boldsymbol{e}\in\mathcal{E}$. Otherwise, $f_c(\boldsymbol{e})>f(\boldsymbol{e})$. Ideally, one would set c to a very high number or ∞ so that the minimizers of the original and penalty-augmented problems coincide. Such a strategy would fail to locate the (local) minimum precisely due to finite machine precision. Instead, we repeatedly compute the minimum of f_c for a gradually increasing sequence $\{c_k\}$ and we use the minimizer of f_{c_k} as an initial guess for the minimizer of $f_{c_{k+1}}$. If at iteration k the current estimate of the EsM is $\boldsymbol{e}^k$, we compute the update $\boldsymbol{\delta}^k\in\mathbb{R}^9$ on $\boldsymbol{e}^k$ by solving the following optimization problem:

$$\underset{\boldsymbol{\delta}^k\in\mathbb{R}^9}{\text{argmin}} \quad f_{c_k}(\boldsymbol{e}^k+\boldsymbol{\delta}^k)=f(\boldsymbol{e}^k+\boldsymbol{\delta}^k)+0.5c_k||\boldsymbol{h}_2(\boldsymbol{e}^k+\boldsymbol{\delta}^k)||^2, \quad (1)$$

$$\text{subject to} \quad \boldsymbol{e}^{k\prime}\boldsymbol{\delta}^k=0 \text{ (to ensure } \boldsymbol{e}^{k+1} \text{ stays away from zero).} \quad (2)$$

Solution Procedure Here we use the popular Gauss-Newton iteration to solve the above problem. In particular, we build a convex quadratic program (QP) approximation to the above problem by (a) replacing f with a convex second-order Taylor approximation $0.5\boldsymbol{\delta}^{k\prime}\mathbf{H}_f(\boldsymbol{e}^k)\boldsymbol{\delta}^k+\nabla'f(\boldsymbol{e}^k)\boldsymbol{\delta}^k+f(\boldsymbol{e}^k)$ and (b) replacing $\boldsymbol{h}_2(\boldsymbol{e}^k+\boldsymbol{\delta}^k)$ with a linear Taylor approximation $\boldsymbol{h}_2^k+\mathbf{J}_2^k\boldsymbol{\delta}^k$ where $\boldsymbol{h}_2^k=\boldsymbol{h}_2(\boldsymbol{e}^k)$. The resulting QP is given by:

$$\underset{\boldsymbol{\delta}^k\in\mathbb{R}^9}{\text{argmin}} \quad \frac{1}{2}\boldsymbol{\delta}^{k\prime}(\mathbf{H}_f^k+c_k\mathbf{J}_2^{k\prime}\mathbf{J}_2^k)\boldsymbol{\delta}^k+(\nabla f^k+c_k\mathbf{J}_2^{k\prime}\boldsymbol{h}_2^k)'\boldsymbol{\delta}^k+\text{const}, \quad (3)$$

$$\text{subject to} \quad \boldsymbol{e}^{k\prime}\boldsymbol{\delta}^k=0. \quad (4)$$

where $\mathbf{H}_f^k=\mathbf{H}_f(\boldsymbol{e}^k)$ and $\nabla f^k=\nabla f(\boldsymbol{e}^k)$. Introducing a scalar Lagrange multiplier v allows us to write the corresponding Lagrangian as:

$$L(\boldsymbol{\delta}^k,v)=\frac{1}{2}\boldsymbol{\delta}^{k\prime}(\mathbf{H}_f^k+c_k\mathbf{J}_2^{k\prime}\mathbf{J}_2^k)\boldsymbol{\delta}^k+(\nabla f^k+c_k\mathbf{J}_2^{k\prime}\boldsymbol{h}_2^k)'\boldsymbol{\delta}^k+v\boldsymbol{e}^{k\prime}\boldsymbol{\delta}^k+\text{const}. \quad (5)$$

The partial derivatives $\nabla_{\boldsymbol{\delta}^k}L$ and ∇_vL must be zero at the optimal $(\boldsymbol{\delta}^k,v)$ [1]. This gives rise to the following 10×10 symmetric linear system of equations:

$$\begin{bmatrix}\mathbf{H}_f^k+c_k\mathbf{J}_2^{k\prime}\mathbf{J}_2^k & \boldsymbol{e}^k\\ \boldsymbol{e}^{k\prime} & 0\end{bmatrix}\begin{pmatrix}\boldsymbol{\delta}^k\\ v\end{pmatrix}=\begin{pmatrix}-(\nabla f^k+c_k\mathbf{J}_2^{k\prime}\boldsymbol{h}_2^k)\\ 0\end{pmatrix}. \quad (6)$$

$$\text{or more compactly as } \mathbf{B}^k\boldsymbol{x}^k=\boldsymbol{b}^k. \quad (7)$$

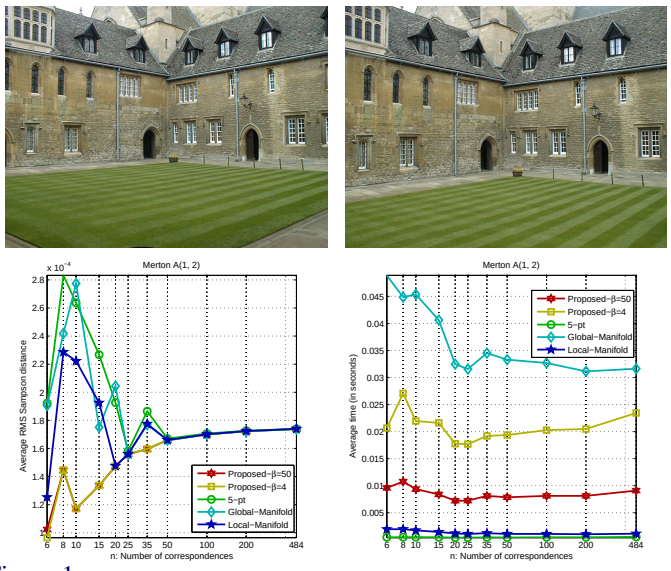


Figure 1: Row 1: (a) Merton A image 1. (b) Merton A image 2. Row 2: (a) RMS Sampson error for each point count with average taken over 75 different random subsets. (b) The average of the corresponding running times.

Rather than using the LDL or LU factorizations, we use the SVD factorization of $\mathbf{B}^k=\mathbf{U}\mathbf{S}\mathbf{V}'$ to solve for $\boldsymbol{x}^k$ as it is more numerically stable. We then use $\boldsymbol{\delta}^k$ to compute the new estimate $\boldsymbol{e}^{k+1}=\boldsymbol{e}^k+\boldsymbol{\delta}^k$.

Controlling The Penalty Parameter Finding an effective strategy for adapting the penalty parameter c_k is the most critical ingredient for the success of a penalty-based algorithm [1]. We consider updating c_k only if (a) we have done enough iterations (at least 3) with the current value of c_k to ensure the solution $\boldsymbol{e}^k$ has achieved some progress with the current value of c_k, and (b) the drop in the value of $||\boldsymbol{h}_2(\boldsymbol{e}^{k+1})||^2$ is found to be not adequate, i.e. $||\boldsymbol{h}_2(\boldsymbol{e}^{k+1})||^2>\gamma||\boldsymbol{h}_2(\boldsymbol{e}^k)||^2$ where we set $\gamma=0.5$. If any of the two conditions is not met, we keep $c_{k+1}=c_k$. Otherwise, we use the update rule $c_{k+1}=\min(\beta c_k,c_{\max})$ where the *penalty multiplier* $\beta>1$ controls the speed and the robustness of the convergence. We set $c_0=10^{-5}$ and $c_{\max}=10^9$.

Experimental Evaluation We compared the performance of the proposed scheme and existing schemes for EsM estimation using synthetic and real data. We included in the comparison two instances of the proposed penalty-based algorithm: one with the penalty multiplier $\beta=50$ (labeled as Proposed-$\beta=50$) and another with $\beta=4$ (Proposed-$\beta=4$) to demonstrate the effect of the penalty multiplier β on robustness and speed. The other schemes included in the comparison were (a) the overdetermined five-point scheme (5-pt), (b) a manifold-based scheme using a global over-parametrization $\boldsymbol{e}:\mathbb{R}^3\times\mathbb{R}^4\to\mathcal{E}$ with the 7-D parameter vector $\boldsymbol{\theta}$ consisting of a 3-vector representing translation and a 4-D quaternion encoding rotation (Global-Manifold (GM)) [2], and (c) Helmke's intrinsic manifold scheme using the local Cayley parametrization (Local-Manifold (LM)) [3]. All schemes were set to minimize the Sampson cost function. Results for one real image pair are shown in Fig. 1. The graphs indicate that the proposed scheme (especially when $\beta=4$) achieves generally lower error curves than the rest of the schemes. GM remains the slowest scheme and LM remains the fastest iterative scheme with the proposed scheme coming in between.

[1] Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[2] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[3] Uwe Helmke, Knut Hüper, Pei Yean Lee, and John Moore. Essential matrix estimation using gauss-newton iterations on a manifold. *Int'l J. Comput. Vision*, 74(2):117–136, 2007.

Board #18

Non-rectangular Part Discovery for Object Detection

Chunluan Zhou
czhou002@e.ntu.edu.sg
Junsong Yuan
jsyuan@ntu.edu.sg

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

The deformable part-based model (DPM) is commonly used for object detection and many efforts have been made to improve the model. However, much less work has been done to discover parts for DPM. Most DPM-based methods adopt the greedy search approach proposed in [2] to initialize a predefined number of parts of rectangular shapes, which may not be optimal for some object categories. Moreover, object structures are not well exploited by the approach. In [4], a three-layer spatial pyramid structure is used to simplify the initialization of parts. An And-Or tree model [3] is proposed to select discriminative part configurations by a dynamic programming algorithm. Although the method can determine part sizes automatically, part shapes are still restricted to rectangles. To address the limitations of these methods, we propose a novel data-driven approach to discover non-rectangular parts by exploiting object structures. Figure 1 shows rectangular and non-rectangular parts obtained by the greedy search approach and our approach, respectively. Generally, the parts obtained by our approach can better cover object regions.

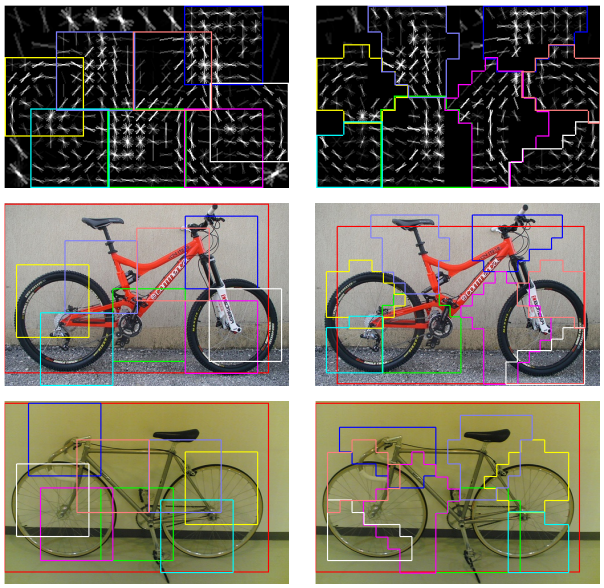


Figure 1: Rectangular parts vs. non-rectangular parts.

The DPM of an object category has several components representing different poses or orientations. Each component consists of a root which represents a whole object and a set of part filters which can move relatively to the root to capture structural deformations. As the training data only have bounding-box annotations specifying the image regions of training examples, the model is trained by first initializing the components and then learning model parameters in a latent structural SVM framework (See [2] for details). As the objective function used in the framework is not convex and as pointed out in [2] the training process is susceptible to local minima, it is necessary to select a good initialization of the components. In this paper, we focus on how to better initialize each component, especially its part filters.

Let M_c be the c-th component which has N_c part filters. The component M_c is defined by a $(2N_c+2)$-tuple $\beta_c = (\boldsymbol{F}_0, \boldsymbol{F}_1, ..., \boldsymbol{F}_{N_c}, \boldsymbol{d}_1, ..., \boldsymbol{d}_{N_c}, b)$, where $\boldsymbol{F}_0$ is the root filter, $\boldsymbol{d}_i \in \mathbb{R}^4$ is the deformation parameters of the part filter $\boldsymbol{F}_i$, and b is the bias term. Each filter $\boldsymbol{F}_i$ is an $H_i \times W_i$ array of n-dimensional weight vectors, where H_i and W_i are the height and width of $\boldsymbol{F}_i$, respectively. To initialize M_c, we first obtain the root filter $\boldsymbol{F}_0$ and then derive part filters from the root filter. The training examples are clustered into several groups each of which corresponds to one component. Let D_c be the set of object examples belonging to the c-th sub-category. $\boldsymbol{F}_0$ is obtained by training a linear SVM on the object examples in D_c and randomly sampled negative examples with each training example represented by histogram of oriented gradients (HOG) [1].

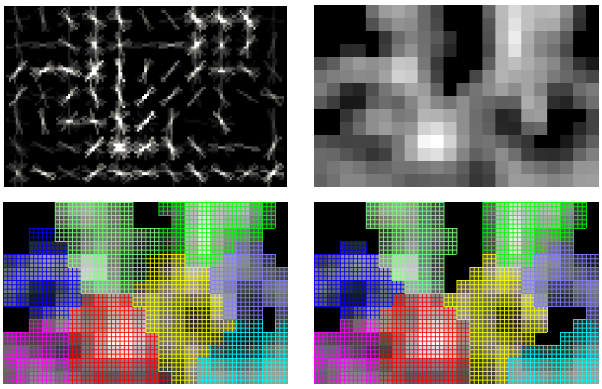


Figure 2: The process of our part discovery approach.

After F_0 is obtained, we find N_c part filters that have good matching regions on object examples in D_c and are consistent with these examples in terms of object structure. First, we double the size of the root filter $\boldsymbol{F}_0$ by interpolation, as in [2], to capture finer details. The enlarged root filter, denoted by $\boldsymbol{F}'_0$, is represented by a $2H_0 \times 2W_0$ array of cells C_k for $1 \le k \le 2H_0 \times 2W_0$, where each cell C_k corresponds to a n-dimensional weight vector in $\boldsymbol{F}'_0$. Then, from $\boldsymbol{F}'_0$, we obtain a configuration of N_c connected part filters, $\Lambda = \{\boldsymbol{F}_i | 1 \le i \le N_c\}$, which satisfies the following overlapping constraint:

$$O(\boldsymbol{F}_i, \boldsymbol{F}_j) = \frac{Area(\boldsymbol{F}_i \cap \boldsymbol{F}_j)}{Area(\boldsymbol{F}_i \cup \boldsymbol{F}_j)} < \tau \quad \text{for } i \neq j, \tag{1}$$

where τ is an overlapping threshold. This constraint prevents any two part filters from overlapping largely. We measure the fitness of the part filter configuration Λ to object examples in D_c by

$$F(\Lambda) = S_R(\Lambda)^{\lambda} \times S_C(\Lambda), \tag{2}$$

where $S_R(\Lambda)$ is the average matching response of Λ over object examples in D_c, $S_C(\Lambda)$ reflects the structural consistency of Λ with these examples, and λ is a parameter used to balance $S_R(\Lambda)$ and $S_C(\Lambda)$. Our goal is to find a feasible part-filter configuration Λ that maximizes $F(\Lambda)$. We refer readers to the paper for details on how $S_R(\Lambda)$ and $S_C(\Lambda)$ are defined and how the objective function is optimized. Figure 2 illustrates the process of our part discovery approach.

We test our approach on Pascal VOC2007 and VOC2010 datasets. Overall, our approach outperforms DPM for 19 and 17 out of 20 object categories in these two datsets respectively, which demonstrates the advantage of the discovered non-rectangular parts over the rectangular parts used in DPM. Implementation details and more experimental results are given in the paper.

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 2010.

[3] X. Song, T. Wu, Y. Jia, and S. Zhu. Discriminatively trained and-or models for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[4] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *International Conference on Computer Vision (ICCV)*, 2010.

Board #19

Weakly Supervised Object Detection with Posterior Regularization

Hakan Bilen
hakan.bilen@esat.kuleuven.be
Marco Pedersoli
marco.pedersoli@esat.kuleuven.be
Tinne Tuytelaars
tinne.tuytelaars@esat.kuleuven.be

KU Leuven, ESAT-PSI, iMinds
Leuven, Belgium

Motivation: In weakly supervised object detection where only the presence or absence of an object category as a binary label is available for training, the common practice is to model the object location with latent variables and jointly learn them with the object appearance model [1, 5]. An ideal weakly supervised learning method for object detection is expected to guide the latent variables to a solution that disentangles object instances from noisy and cluttered background. The learning algorithm should lead the appearance model and the latent variables to best explain the correlation between the training images and their binary labels. However, without complete supervision, maximizing the likelihood of observed data or minimizing the data-dependent cost function during training may result in latent variables that do not capture the expected regularities.

Contributions: In this paper, (i) we show that in a weakly-supervised setting, regulating the latent distribution and properly driving the latent variables are crucial for good performance and lead to state-of-the-art results in both classification and detection, (ii) we show how to introduce in the weakly supervised detection specific prior knowledge that helps to drive the latent variables by means of posterior regularization, and (iii) we better model the weakly-supervised object detection problem via the soft-max where multiple objects in the same image are considered and at the same time the optimization is smoother.

We focus on domain specific prior knowledge for object detection. In particular we exploit the fact that (i) each horizontal mirror of an object is still a valid object (*object symmetry*) and (ii) the same spatial region (in our case a bounding box) cannot represent more than one object class (*mutual exclusion*). We incorporate this prior knowledge via posterior regularization as proposed in [4].

Results: We evaluate our method and compare its performance to previous work [2, 6, 7] in the Pascal VOC 2007 dataset [3]. We first illustrate hard-max and soft-max outputs in Fig. 1, the posterior regularization on symmetry and mutual exclusion in Fig. 2 and Fig. 3 resp. We also report quantitative results in detection and classification tasks in Table 1 and 2 resp. We show the contribution of each added component and compare the final result to the state-of-the-art methods in both detection and classification.

[1] H. Bilen, V.P. Namboodiri, and L. Van Gool. Object and action classification with latent window parameters. *IJCV*, pages 1–15, 2013.

[2] R.G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

[3] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[4] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

[5] M.H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009.

[6] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[7] H.O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. One-bit object detection: On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

(a) images (b) hard-max output (c) soft-max output

Figure 1: Visual comparison of max-margin and soft-max margin learning on representative "cow" and "chair" images. While max outputs a single window, soft-max marginalizes over all windows and better represents multiple instances.

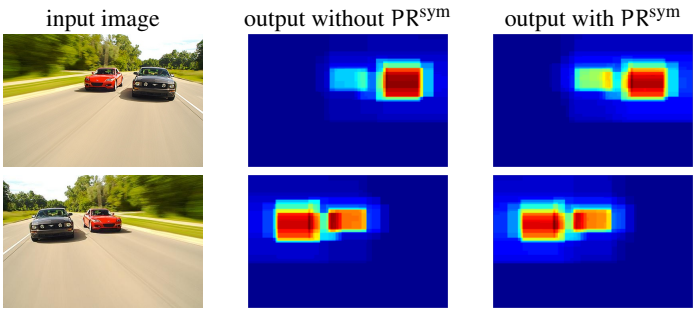


Figure 2: Output maps of "car" detectors on test and flipped images without and with posterior regularization for symmetry. Learning with the symmetrical constraints increase the scores of less confident detections.

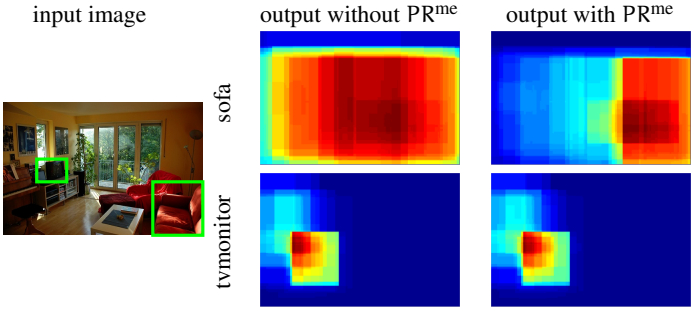


Figure 3: Output maps of "sofa" and "tvmonitor" detectors for input images. Adding the mutual exclusion constraint helps to separate two distributions by penalizing the bounding boxes with high probability for both detectors.

Ours					Others	
hard-max	soft-max	+flip	+PR^{sym}	+PR^{me}	[2]	[7]
22.7	24.0	24.8	26.0	**26.4**	22.4	22.7

Table 1: Weakly supervised detection results on the Pascal VOC 2007 in mean average precision (mAP). +flip indicates of adding horizontally mirrored training images to the training. PR^{sym} and PR^{me} denote the posterior regularization for symmetry and mutual exclusion. The components starting from +flip are consecutively added on the soft-max. Our method outperforms the state-of-the-art weakly supervised detectors [2, 7].

Ours			Others	
SVM	hard-max	Full	[2]	[6]
74.1	77.1	**80.9**	65.6	77.7

Table 2: Classification results on the Pascal VOC 2007 in mAP. SVM denotes training linear SVMs without any localization. hard-max and Full denote the latent SVM formulation and our full model. Our method outperforms the state-of-the-art classifiers [2, 6].

3D Pose-by-Detection of Vehicles via Discriminatively Reduced Ensembles of Correlation Filters

Yair Movshovitz-Attias[1]
www.cs.cmu.edu/~ymovshov
Vishnu Naresh Boddeti[2]
vishnu.boddeti.net
Zijun Wei[2]
hzwzijun@gmail.com
Yaser Sheikh[2]
www.cs.cmu.edu/~yaser/

[1] Computer Science Department
Carnegie Mellon University
Pennsylvania, USA

[2] Robotics Institute
Carnegie Mellon University
Pennsylvania, USA

Accurate estimation of the pose of a 3D model in an image is a fundamental operation in many computer vision and graphics applications, such as 3D scene understanding, inserting new objects into images, and manipulating current ones. One class of approaches to pose estimation is correspondence-based: individual parts of the object are detected, and a pose estimation algorithm (e.g., perspective-*N*-point) can be used to find the pose of the 3D object in the image. When the parts are visible, these methods produce accurate continuous estimates of pose. However, if the size of the object in the image is small or if the individual parts are not detectable (e.g., due to occlusion, specularities, or other imaging artifacts), the performance of such methods degrades precipitously. In contrast to correspondence-based approaches, pose-by-detection methods use a set of view-specific detectors to classify the correct pose; these methods have appeared in various forms such as filter banks, visual sub-categories, and exemplar classifier ensembles. While such approaches have been shown to be robust to many of the short-comings of correspondence-based methods, their primary limitation is that they provide discrete estimates of pose and as finer estimates of pose are required, larger and larger sets of detectors are needed.

Reduced representations are attractive because of their statistical and computational efficiency. Most approaches reduce the set of classifiers via the classic notion of minimizing the reconstruction error of the original filter set. Such a reduction does not directly guarantee optimal preservation of *detection* performance. This is particularly problematic in the case of viewpoint discrimination, as filters of proximal pose angles are similar. Reduction designed to minimize reconstruction error often results in a loss of view-point precision as the distinctive differences in proximal detectors are averaged out by the reduction.

In this paper, we present a pose-by-detection approach that uses an ensemble of correlation filters for precise viewpoint discrimination, by using a 3D CAD model of the vehicle to generate renders from viewpoints at the desired precision. A key contribution of this paper is a training framework that generates a discriminatively reduced ensemble of exemplar correlation filters by explicitly optimizing the detection objective. As the ensemble is estimated jointly, this approach intrinsically calibrates the ensemble of exemplar classifiers during construction, precluding the need for an after-the-fact calibration of the ensemble. The result is a scalable approach for pose-by-detection at the desired level of pose precision.

While our method can be applied to any object, we focus on 3D pose estimation of vehicles since cheap, high quality, 3D CAD models are readily available. We demonstrate results that outperform the state-of-the-art on the Weizmann Car View Point (WCVP) dataset, the EPFL car multi-view car dataset, and the VOC2007 car viewpoint dataset. We also report results on a new data-set based on the CMU-car dataset [1]. for precise viewpoint estimation and detection of cars. Fiigure 1 shows example results of our system on the WCVP dataset. Each row shows input images (top) and overlaid pose estimation results (bottom). Fiigure 2 These results demonstrate that pose-by-detection based on ensemble of exemplar correlation filters can achieve and exceed the level of precision of correspondence based methods in real datasets; and that discriminative reduction of an ensemble of exemplar classifiers allows scalable performance at higher precision levels.

[1] V. N. Boddeti, T. Kanade, and B. V. K. Vijaya Kumar. Correlation filters for object alignment. In *Computer Vision and Pattern Recognition*. IEEE, 2013.

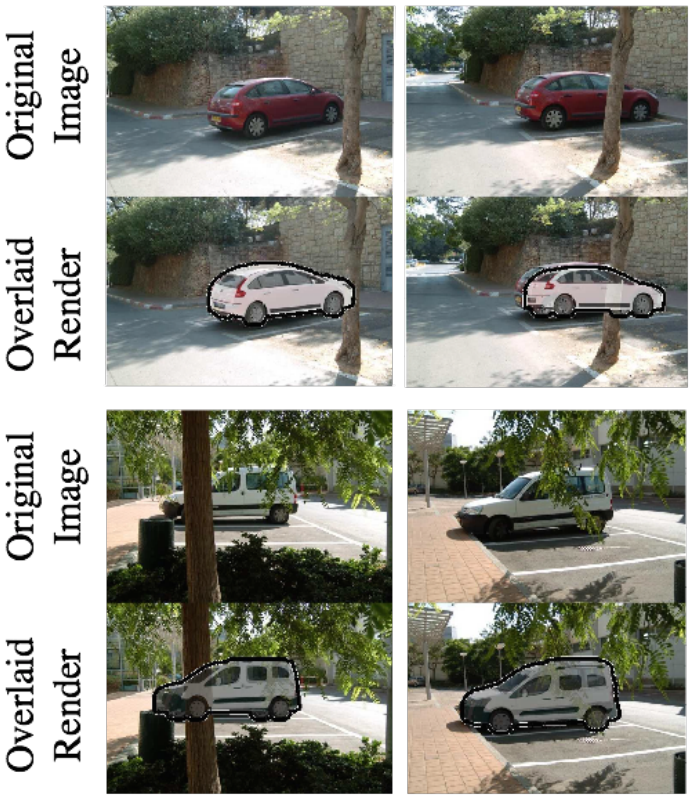


Figure 1: Example results. Each row shows input images (top) and overlaid pose estimation results (bottom).

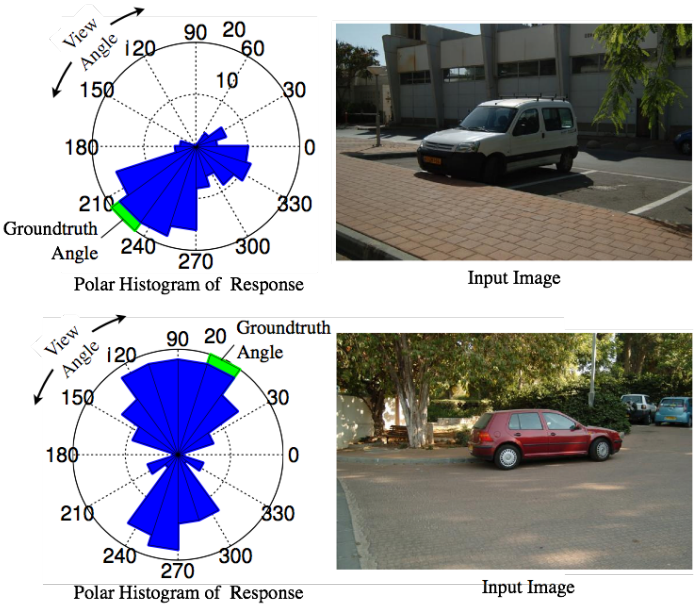


Figure 2: Polar histogram of scores. The left example shows a van at an oblique angle, with little ambiguity in the distribution of responses. The right example shows a side view with a distinctive symmetric ambiguity.

Board #21

Upper Body Pose Estimation with Temporal Sequential Forests

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 goal of this work is to recover the 2D layout of human upper body pose over long video sequences. The focus is on producing reliable and accurate pose estimates for use in gesture analysis and recognition.

We build on the recent successful applications of random forests (RF) classifiers and regressors [1], and develop a pose estimation model with the following novelties: (i) the joints are estimated sequentially, taking account of the human kinematic chain. This means that we don't have to make the simplifying assumption of most previous RF methods – that the joints are estimated independently; (ii) by combining both classifiers (as a mixture of experts) and regressors, we show that the learning problem is tractable and that more context can be taken into account; and (iii) dense optical flow is used to align multiple expert joint position proposals from nearby frames, and thereby improve the robustness of the estimates. The processing steps are divided into two stages.

Stage 1 – Sequential body joint detection

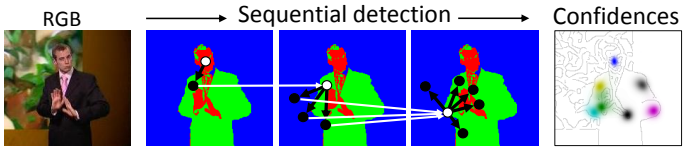
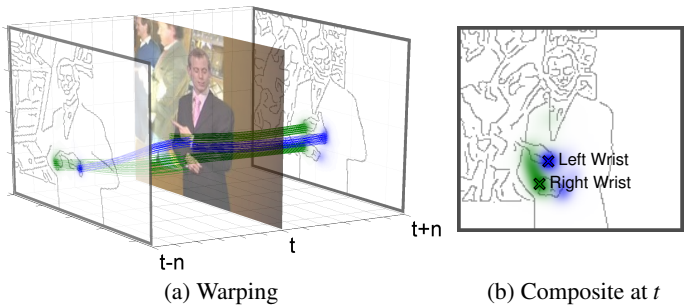


Figure 1: **Stage 1 – sequential upper body pose estimation.** (a) RGB input. (b) Sequential detection with random forest experts: the head is detected first, then shoulders, elbows and finally wrists. (c) Confidence map of body joints, with different colour for each joint (higher colour intensity indicates stronger confidence).

In Stage 1, body joints are detected sequentially in a single video frame. Each joint in the sequence depends on the location of the previous joint: the head is detected first, followed by shoulders, elbows, and wrists, separately for left and right arms. Figure 1(a-c) illustrates this sequential detection. Beginning with an RGB frame (a), the frame is first encoded into a feature representation, shown in Figure 1(b) as an image with pixels categorised as either skin (red), torso (green) or background (blue). For each joint, a separate mixture of experts (random forest) votes for the next joint location (votes shown as white lines in figure). Each expert (shown as black dots in figure) is responsible for a particular region of the image which depends upon the location of the previous joint in the sequence (positioned according to fixed learnt offset vectors, shown as black arrows). The output from this is a confidence map over pixels for each joint.

Stage 2 – Detection reinforcement with optical flow



(a) Warping (b) Composite at t

Figure 2: **Stage 2 – warping neighbouring confidence maps to improve wrist joint detections.** (a) Confidence maps from frames $(t-n)$ and $(t+n)$ warped to frame t using tracks from optical flow (green & blue lines). (b) Composite map with crosses indicating modes of confidence.

In Stage 2, confidences from Stage 1 produced at a frame t are reinforced with temporal context from nearby frames. Additional confidence maps are produced for neighbouring frames, and are aligned with frame t by warping them backwards or forwards using tracks from dense optical flow. This is illustrated in Figure 2(a) for wrist confidences produced at frame $(t-n)$ and $(t+n)$. Finally, body joint locations are estimated at frame t by selecting positions of maximum confidence from a composite map produced by combining warped confidences (see Figure 2(b)).

Figure 3: Pose estimates from our method on two different gesture datasets. Top: BBC TV dataset. Bottom: Chalearn gesture dataset.

Results

Our method takes advantage of the kinematic constraints of the human body and explicitly builds in spatial context which we know is of importance, such as elbow location when detecting the wrist. The locally trained RFs deal with less of the feature space compared to its sliding window counterparts, which makes learning easier and leads to improved accuracy over the state-of-the-art [1, 2].

Accuracy of the sequential forest at Stage 1 (SF) is shown to improve further when incorporating output from multiple expert opinions from neighbouring frames in Stage 2 (SF+flow) (see Figure 4). Example pose estimates on two different datasets are shown in Figure 3.

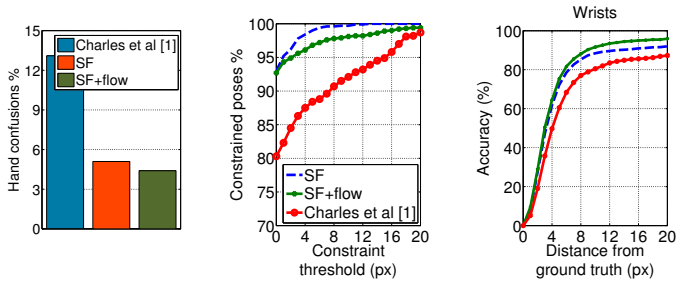


Figure 4: (a) SF+flow significantly reduces hand confusions. (b) SF and SF+flow achieve significantly better constrained pose estimates than state-of-the-art [1]. (c) Improvement in average wrist accuracy.

References

[1] J. Charles, T. Pfister, M. Everingham, and A. Zisserman. Automatic and efficient human pose estimation for sign language videos. *IJCV*, 2013.

[2] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. CVPR*, 2011.

Board #22

Cloud-scale Image Compression Through Content Deduplication

David Perra
perra@cs.unc.edu
Department of Computer Science,
University of North Carolina,
Chapel Hill, NC

Jan-Michael Frahm
jmf@cs.unc.edu
Department of Computer Science,
University of North Carolina,
Chapel Hill, NC

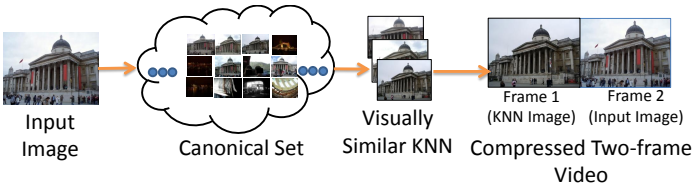


Figure 1: A high-level overview of our proposed technique. A canonical set, consisting of potentially millions of images of varying subjects, can be leveraged to find a set of images which are visually similar to a query image via k-nearest neighbors search. Those visually similar images are each used to compress the query image by use of a video codec such as H.265. The two-frame video which yields the best compression to quality ratio will be selected as the output of the pipeline, and the bits associated with the input image will be stored.

Modern large-scale photo services such as Google Drive, Microsoft OneDrive, Facebook, and Flickr are currently tasked with storing and serving unprecidented quantities of photo data. While most photo services still utilize jpeg compression to store photos, more elegant compression schemes will need to evolve to combat the storage costs associated with the exponential increase in data. To satisfy this need, two classes of solutions have been established: representative signal techniques [1, 6], and visual clustering techniques [3, 5, 8]. Representative signal techniques work by finding a common low-frequency signal within a set of images. The technique presented in this paper falls into the second class of techniques, which focuses upon sharing and reusing pixel data between multiple images by modelling the relationship between these images as a directional graph. Paths through this directional graph define image pseudosequences, or directionally related subsets of images which describe the visually shortest path between various images in an image set [5, 7, 8]. These pseudosequences can then be used for compression via image reconstruction or compression via traditional video codecs, such as H.265.

The primary shortcomings for state-of-the-art visual clustering techniques stem from a lack of scalability. Finding appropriate image pseudosequences becomes increasingly more difficult as an image set grows. This is because all-pairs comparisons must be performed between the images to find an optimal graph. Additionally, longer pseudosequences tend to result from larger image sets. Longer pseudosequences cause image compression and decompression to take longer, leading to a decrease in performance with an increase in dataset size.

In this paper, we present an efficient cloud-scale digital image compression scheme which overcomes the scalability issues found in the state-of-the-art techniques. Unlike current state-of-the-art systems, our image compression technique takes full advantage of redundant image data in the cloud by independently compressing each newly uploaded image with its GIST nearest neighbor taken from a canonical set of uncompressed images. This allows for fast identification of a size-restricted pseudosequence. We then leverage state-of-the-art video compression techniques, namely H.265, in order to efficiently reuse image content which is already stored server-side.

Previous state-of-the-art techniques used only the image data found within a particular image set to compress the entire set [5, 7, 8]. Our technique, on the other hand, avoids this through the use of the canonical set of images. Our key insight is that many photographs uploaded to the cloud are highly likely to have similar pixel patches, especially near landmarks and other commonly geotagged areas – even within the home of the user. Thus, we assume that the canonical set is a randomly selected, finite set of photos that is composed of tens or hundreds of millions of images depicting commonly photographed subjects and structures. Constructing such a set can be done, for example, by randomly sampling all photos currently stored in the cloud. Alternatively, techniques like Frahm *et al*. [2] and Raguram *et al*. [4] can be used to construct such a canonical set through iconic scene graphs. This process should naturally yield many views of popular subjects as more photos of those subjects are uploaded to the cloud. A sufficiently large canonical set contains enough photos to have a visually similar image for a large majority of photos uploaded in the future. Similarly, we foresee complementing the general canonical set with a user-specific canonical set if desired. Once an ideal canonical set is constructed, it can be used as a generic dataset for compressing any photo collection.

The implementation of our method is described in our paper, and extensive experiments are conducted. Experimental results demonstrate that our algorithm produces competitive image compression rates while reducing the computational effort by at least an order of magnitude in comparison to competing techniques, all while providing the necessary scalability for use in cloud-scale applications.

[1] Samy Ait-Aoudia and Abdelhalim Gabis. A comparison of set redundancy compression techniques. *EURASIP J. Appl. Signal Process.*, 2006:216–216, January 2006.

[2] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building rome on a cloudless day. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 368–381. Springer Berlin Heidelberg, 2010.

[3] Yang Lu, Tien-Tsin Wong, and Pheng-Ann Heng. Digital photo similarity analysis in frequency domain and photo album compression. In *Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia*, MUM '04, pages 237–244, New York, NY, USA, 2004. ACM.

[4] Rahul Raguram, Changchang Wu, Jan-Michael Frahm, and Svetlana Lazebnik. Modeling and recognition of landmark image collections using iconic scene graphs. *Int. J. Comput. Vision*, 95(3):213–239, December 2011.

[5] Zhongbo Shi, Xiaoyan Sun, and Feng Wu. Photo album compression for cloud storage using local features. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 4(1):17–28, March 2014.

[6] Chi-Ho Yeung, O.C. Au, Ketan Tang, Zhiding Yu, Enming Luo, Yannan Wu, and Shing-Fat Tu. Compressing similar image sets using low frequency template. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6, July 2011.

[7] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Cloud-based image coding for mobile devices 2014;toward thousands to one compression. *Multimedia, IEEE Transactions on*, 15(4):845–857, June 2013.

[8] Ruobing Zou, O.C. Au, Guyue Zhou, Wei Dai, Wei Hu, and Pengfei Wan. Personal photo album compression and management. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 1428–1431, May 2013.

DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking

Hanxi Li[12]
hanxi.li@nicta.com.au
Yi Li[1]
http://users.cecs.anu.edu.au/~yili/
Fatih Porikli[1]
http://www.porikli.com/

[1] NICTA and ANU, Canberra, Australia
[2] Jiangxi Normal University, Jiangxi, China

Defining hand-crafted feature representations needs expert knowledge, requires time-consuming manual adjustments, and besides, it is arguably one of the limiting factors of object tracking.

In this paper, we propose a novel solution to automatically relearn the most useful feature representations during the tracking process in order to accurately adapt appearance changes, pose and scale variations while preventing from drift and tracking failures. We employ a candidate pool of multiple Convolutional Neural Networks (CNNs) as a data-driven model of different instances of the target object. Individually, each CNN maintains a specific set of kernels that favourably discriminate object patches from their surrounding background using all available low-level cues (Fig. 1). These kernels are updated in an online manner at each frame after being trained with just one instance at the initialization of the corresponding CNN.

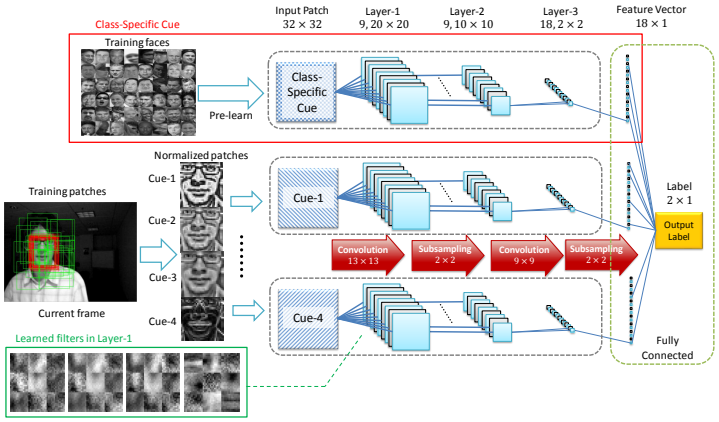


Figure 1: Overall architecture with (red box) and without (rest) the class-specific version.

Instead of learning one complicated and powerful CNN model for all the appearance observations in the past, we chose a relatively small number of filters in the CNN within a framework equipped with a temporal adaptation mechanism (Fig. 2). Given a frame, the most promising CNNs in the pool are selected to evaluate the hypothesises for the target object. The hypothesis with the highest score is assigned as the current detection window and the selected models are retrained using a warm-start back-propagation which optimizes a structural loss function.

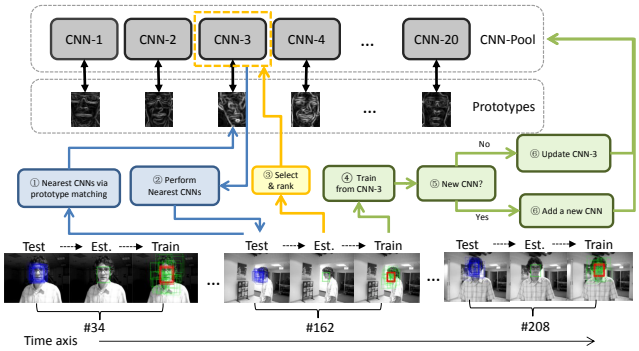


Figure 2: Illustration of the temporal adaptation mechanism.

Our experiments on a large selection of videos from the recent benchmarks demonstrate that our method outperforms the existing state-of-the-art algorithms and rarely loses the track of the target object. We evaluate our method on 16 benchmark video sequences that cover most challenging tracking scenarios such as scale changes, illumination changes, occlusions, cluttered backgrounds and fast motion (Fig. 3).

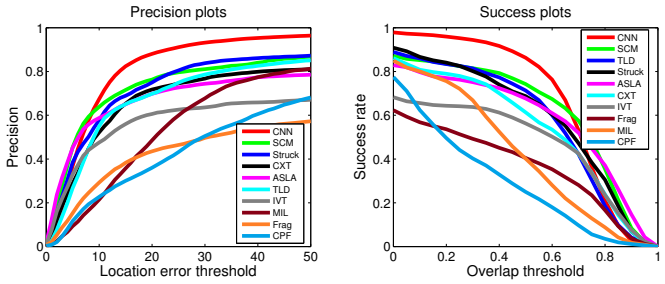


Figure 3: The Precision Plot (left) and the Success Plot (right) of the comparing visual tracking methods.

In certain applications, the target object is from a known class of objects such as human faces. Our method can use this prior information to leverage the performance of tracking by training a class-specific detector. In the tracking stage, given the particular instance information, one needs to combine the class-level detector and the instance-level tracker in a certain way, which usually leads to higher model complexity (Fig. 4).

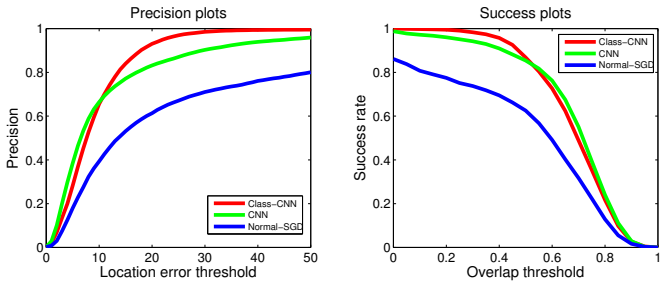


Figure 4: The Precision Plot (left) and the Success Plot (right) of the comparing visual tracking methods.

To conclude, we introduced DeepTrack, a CNN based online object tracker. We employed a CNN architecture and a structural loss function that handles multiple input cues and class-specific tracking. We also proposed an iterative procedure, which speeds up the training process significantly. Together with the CNN pool, our experiments demonstrate that DeepTrack performs very well on 16 sequences.

Board #24

Tri-Map Self-Validation Based on Least Gibbs Energy for Foreground Segmentation

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

NTT Communication Science Laboratories
3-1, Morinosato Wakamiya Atsugi-shi
Kanagawa, Japan 243-0198

Foreground segmentation plays an important role in high-level vision tasks. Of previously reported research, a large percentage is made up of Markov random field (MRF) based studies [2, 5, 6], in which optimal segmentation maximizes the posteriori probability given observations incorporated with a predefined tri-map. They are current to the state-of-the-art, but under the assumption that a sufficiently discriminative tri-map is given, *e.g.* specified by user interaction [6] or supervised by using class information [2, 5]. With a low-quality tri-map, although some attempts have been made to improve the MRF model, very little attention has been paid to enhancing the discernment of the tri-map itself. This constitutes the main problem that we tackle in this paper.

In contrast to the previous studies, which depended on strong assumptions, our aim is *unsupervised* foreground segmentation under only one weak (realistic) assumption. We assume that the location of a foreground is a normal deviate in the image space, whose expectation lies near the center of the image. We argue that the least Gibbs energy (LGE) can be formulated as a goal function of a tri-map optimization problem, and propose decomposing the complex problem into a series of tractable subproblems. A suboptimal optimization is gradually obtained by making decisions between pixel cluster-level set operations.

(a) $LGE = 2.61 \times 10^6$ (b) $LGE = 2.47 \times 10^6$ (c) $LGE = 2.42 \times 10^6$

Figure 1: Different tri-maps (left) exhibit differences in least Gibbs energies (LGE), incorporated in the segmentation (right) of the same image.

In terms of MRFs, the optimal segmentation $\hat{X}$ maximizes the a posteriori probability pertaining to an observed image Y and a tri-map T. It is equivalent to minimizing the Gibbs energy $E(X|Y,T)$:

$$E(X|Y,T) = \sum_p \sum_\alpha U_p^{(\alpha)}(y_p|T)\delta(\alpha,x_p) + \sum_{p,q} \frac{1-\delta(x_p,x_q)}{\|p-q\|} \exp(-\beta\|y_p - y_q\|) \quad (1)$$

where the right terms are known as the likelihood (first) and coherence (second) energies at the pixel level. We define the LGE as follows:

$$LGE(T|Y) = \min_X E(X|Y,T) \quad (2)$$

LGE is a function of T with a given observation Y, and is no longer dependent on the segmentation X. When the distributions of foreground and background pixels offer very low separability, as shown in Fig. 1(a), the likelihood term becomes non-contributory and the minimization over-fits the coherence term, resulting in a high LGE. When tri-maps lead to the same segmentation, i.e. to equivalent coherence energies, as shown in Fig. 1(b) and 1(c), the tri-map with the larger distribution overlap indicates a higher entropy. A desired tri-map $\hat{T}$ can be defined as one that minimizes $LGE(T|Y)$, more specifically

$$\hat{T} = \arg\min_T \min_X E(X|Y,T) \quad (3)$$

We propose a split-and-validate method for solving this problem. The splitting is determined by a non-parametric clustering method (see the paper). After splitting, the image is abstracted as a set of pixel clusters. Our tri-map validation is based on two types of cluster-level operations:

(Retaining) *Keeping a tri-map T unchanged, as denoted by $T \leftarrow T$.*

(Contracting) *For a tri-map $T = \{T_B, T_F\}$, in which T_B and T_F are background and foreground regions, and a pixel cluster c, subtracting c from T_F and adding c to T_B, as denoted by $T \leftarrow \{T_B \cup c, T_F \setminus c\}$.*

The self-validation of a tri-map is discretized to a tree-structured evolution process. $T^{(0)}$ is preliminarily treated as a rectangle in the center. Using Eq. 1, we can obtain $LGE(T^{(0)}|Y)$. All pixel clusters $\{c_1, c_2, \cdots\}$ are sorted in ascending order of image-space centrality. This is motivated by the assumption that a cluster of pixels is more likely to belong to the foreground if its location is closer to the center of the image. $T^{(0)}$ is then arguably refined by **Contracting** with the cluster at the top of the sorted queue, which leads to a tentative tri-map $T'^{(0)}$ and $LGE(T'^{(0)}|Y)$. An arbitrary T is contract-able if **Contracting** leads to a lower LGE than **Retaining**. If so, we update T to T' and continue this process iteratively until all clusters are incorporated in the validation. We obtain the segmentation by using an iterated graph cut [6] with the refined $\hat{T}$.

Figure 2: Example of tri-map optimization and segmentation. From left to right: initialized tri-map, segmentation of GC [6], optimized tri-map, and our segmentation.

Figure 2 compares the segmentations initialized by the same tri-map. Table 1 compares our method with advanced studies. More detail regarding the non-parametric clustering method determining the splitting and the experiments is described in the paper. Our conclusion is that the LGE can be a strong cue for capturing the discriminative power of a tri-map, and is useful when dealing with unsupervised foreground segmentation.

Table 1: Performance on Oxford Flower17 reported in the literature[1].

Method	MJI	MNHS
Nilsback and Zisserman [5]	93.0	–
Joulin *et al.* [3]	75.8	86.6
Chai *et al.* [2]	**94.7**	**98.3**
Najjar and Zagrouba [4]	84.0	–
Aydin and Ugur [1]	87.0	–
Suta *et al.* [7]	90.0	89.0
Our Method	**91.7**	**96.8**

[1] The definition of *MJI* and *MNHS* can be found in the paper.

[1] D. Aydin and A. Ugur. Extraction of flower regions in color images using ant colony optimization. *Procedia CS*, 3:530–536, 2011.

[2] Y. Chai, V. S. Lempitsky, and A. Zisserman. BiCoS: A Bi-level co-segmentation method for image classification. In *ICCV*, pages 2579–2586, 2011.

[3] A. Joulin, F. R. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[4] A. Najjar and E. Zagrouba. Flower image segmentation based on color analysis and a supervised evaluation. In *ICCIT*, pages 397–401, 2012.

[5] M.-E. Nilsback and A. Zisserman. Delving deeper into the whorl of flower segmentation. *Image Vision Comput.*, 28(6):1049–1062, 2010.

[6] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[7] L. Suta, F. Bessy, C. Veja, and M.-F. Vaida. Active contours: Application to plant recognition. In *ICCP*, pages 181–187, 2012.

Board #25

Surface Normal Integration for Convex Space-time Multi-view Reconstruction

Martin R. Oswald
martin.oswald@in.tum.de
Daniel Cremers
cremers@tum.de

Computer Vision Group,
Department of Computer Science,
Technische Universität München

We propose a convex variational approach to space-time reconstruction which estimates surface normal information and integrates it into the photoconsistency estimation as well as into an anisotropic spatio-temporal total variation regularization. As such the proposed method generalizes the works [4], [5]. Although [4] already studied anisotropic regularization they did not estimate normals but used the normals from [2]. The combination of these methods, [4] and [2], is more than 40 times slower than our method as [4] alone needs about 1h to compute a single frame. In contrast, our method only takes about 3 minutes per frame including normal estimation and temporal regularization due to the proposed efficient implementation. Moreover, the method by Kolev et al. [4] does not work well on the 4D data sets we consider, as shown in [5, Fig. 5]. With the estimated normals at hand, we further propose an improvement of the photoconsistency voting scheme by Hernández and Schmitt [1] resulting in superior accuracy especially for sparse camera setups.

We represent the space-time surface as a binary interior/exterior labeling function $u : V \times T \mapsto \{0,1\}$ and state the spatio-temporal 3D reconstruction task as a minimization problem of the following energy.

$$E(u) = \int_{V\times T} \Big[|\nabla_{\boldsymbol{x}} u|_{D_{\boldsymbol{x}}} + g_t |\nabla_t u| + \lambda\, f u \Big] \; d\boldsymbol{x}dt \tag{1}$$

The energy consists of three terms. An **anisotropic spatial regularization term**, defined be the norm $|\boldsymbol{y}|_{D_{\boldsymbol{x}}} = \langle \boldsymbol{y}, D_{\boldsymbol{x}}\boldsymbol{y}\rangle^{1/2}$ and the anisotropic diffusion matrix $D_{\boldsymbol{x}}(\boldsymbol{x},t) = \rho(\boldsymbol{x},t)^2 \boldsymbol{n}\boldsymbol{n}^T + \boldsymbol{n}_0\boldsymbol{n}_0^T + \boldsymbol{n}_1\boldsymbol{n}_1^T$ which lowers smoothing in the surface normal $\boldsymbol{n} \in \mathbb{R}^3$ direction and favors smoothness along the corresponding tangential directions $\boldsymbol{n}_0$ and $\boldsymbol{n}_1 = \boldsymbol{n} \times \boldsymbol{n}_0$. Further, the **temporal regularization term** weighted by function $g_t(\boldsymbol{x},t) = \exp\big(-a|\nabla f(\boldsymbol{x},t)|\big)$ accounts for a motion-dependent temporal smoothing. The purpose of the temporal regularization is to reduce surface jittering in scene parts with slow motion. Lastly, the **data term**, represented by function $f : V \times T \mapsto \mathbb{R}$ and smoothness weight λ, avoids trivial solutions of the energy and gives local preferences for an interior or exterior label. Both, the photoconsistency measure $\rho(\boldsymbol{x})$ in $D_{\boldsymbol{x}}$ and f depend on a voting scheme based on surface normal-dependent normalized cross-correlation (NCC) scores, represented by $C_i(\boldsymbol{x},d)$ for each point defined by the ray from camera i through point $\boldsymbol{x}$ at distance d.

$$\rho(\boldsymbol{x}) = \exp\Big[-\mu \underbrace{\sum_{i\in\mathcal{C}'} \delta\big(d_i^{\max} = \mathrm{depth}_i(\boldsymbol{x})\big) \cdot C_i(\boldsymbol{x}, d_i^{\max})}_{\mathrm{VOTE}_i(\boldsymbol{x})} \Big] \tag{2}$$

The original voting scheme [1] computes the best depth hypothesis per camera ray as $d_i^{\max} = \arg\max_d C_i(\boldsymbol{x},d)$ and does not enforce any spatial regularity of the votes, which we introduce by the following normal-dependent regularized voting scheme:

$$d_i^{\max} = \arg\max_d \int_{V_{\boldsymbol{x}}} C_i(\boldsymbol{x}-\boldsymbol{y},d)\; \mathcal{G}(\boldsymbol{y};\Sigma_{\boldsymbol{n}}) \; d\boldsymbol{y} \;, \tag{3}$$

where $\mathcal{G}(\boldsymbol{y};\S igma_{\boldsymbol{n}})$ is a normal-aligned anisotropic 3D Gaussian. We use surface normals at three places within our method: (a) NCC score, (b) voting scheme regularization and (c) anisotropic surface regularization. To estimate normals, we run our algorithm in two passes (see Fig. 1):

Pass 1: camera-to-point direction as normal for (a) and (b), isotropic surface regularization with high λ for (c)

Pass 2: normals from the previous pass for (a),(b) and (c) with lower λ for surface smoothness as desired

Finally, we propose an efficient GPU-accelerated primal-dual optimization of energy (1) which allows for comparatively low computation times. Our model yields significantly improved results over [5] which also compare well to other state-of-the-art reconstruction methods (see Fig. 2).

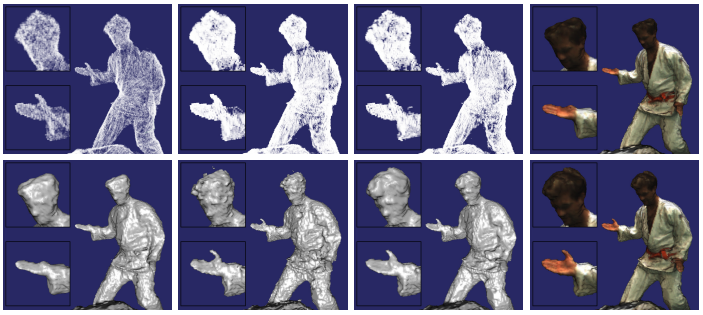


Figure 1: The photoconsistency measure $\rho(\boldsymbol{x})$ (top) and corresponding reconstructions (bottom) with and without the proposed normal integration. Pass 1 demonstrates the effect of the proposed normal-based voting regularization. Pass 2 adds anisotropic regularization and improved photoconsistency scores from pass 1.

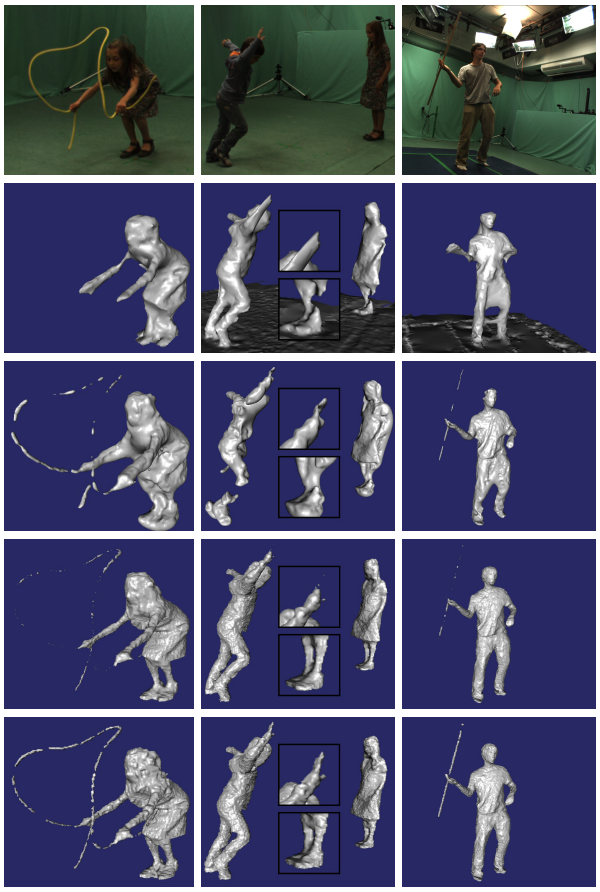


Figure 2: Reconstruction results of different methods. From top to bottom: input, Jancosek and Pajdla [3], PMVS+Poisson [2], Oswald and Cremers [5] and proposed. Our method recovers fine details and reduces temporal surface jittering.

[1] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *CVIU*, 96(3):367–392, Dec 2004.

[2] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE TPAMI*, 32(8):1362–1376, Aug 2010.

[3] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, pages 3121–3128, 2011.

[4] K. Kolev, T. Pock, and D. Cremers. Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. In *ECCV*, Heraklion, Greece, Sep 2010.

[5] Martin R. Oswald and Daniel Cremers. A convex relaxation approach to space time multi-view 3d reconstruction. In *ICCV - Workshop on Dynamic Shape Capture and Analysis (4DMOD)*, 2013.

Contextually Constrained Deep Networks for Scene Labeling

Taygun Kekeç[1]
taygunkekec@gmail.com
Rémi Emonet[1]
remi.emonet@univ-st-etienne.fr
Elisa Fromont[1]
elisa.fromont@univ-st-etienne.fr
Alain Trémeau[1]
alain.tremeau@univ-st-etienne.fr
Christian Wolf[2]
christian.wolf@liris.cnrs.fr

[1] Université de Lyon, CNRS UMR 5516, Laboratoire Hubert-Curien
Université de Saint-Etienne, F-42000, Saint-Etienne, France
[2] Université de Lyon, CNRS INSA-Lyon, LIRIS, UMR5205, F-69622, France

Deep learning approaches, such as multi-layer neural networks, leverage the amount of available data to learn representations: instead of hand-crafting intermediate features, they are learned directly from the data. This is particularly relevant since there is no universal feature detector performing best for any given problem and these learned features have been shown to outperform hand-crafted features on many perception tasks.

In this work we focus scene labeling task with deep learning strategies. We first learn a CNN (Convolutional Neural Network) to predict contextual information. By forcing this network to capture some context information of our choice, we aim to improve the interpretability of the CNN and obtain meaningful feature maps. In parallel, we learn a second model for the original task assuming that contextual information is obtainable from ground truth labels at training step. Finally, we combine these networks and perform a last training phase with weakened supervision.

In traditional feature learning, the input processing is separated in two parts as illustrated in Figure 1a. The input I is first processed with a function $f(.)$, which has parameters θ_f and produces a set of features F. A predictor $p(.)$ having parameters θ_p takes the features F as input and produces a prediction. To constrain the whole network, we propose to split the function f into two parts: f_d and f_c (Fig. 1b). Function f_c aims at predicting some context and it is learned with additional supervision.

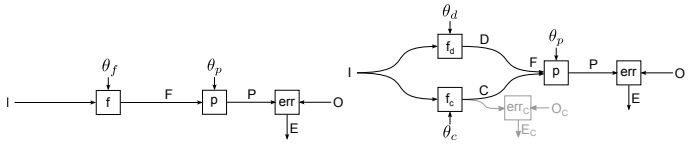


Figure 1: Functional representation of our feature learning approaches. (a) The target function is composed of a feature extraction function f and a prediction function p. (b) Our approach which distinguishes the learning of context features f_c and dependent features f_d.

Learning context – In this step, we start from a random initialization θ_c^0 and learn θ_c^1 where the superscript j in θ_c^j indicates the training stage. The context learning step minimizes the following error function: $\mathcal{L}_c = \sum_{k=1}^{K} \left\| p_{soft}^k(f_c(I,\theta_c) - O_c^k) \right\|^2$ where K is the number of context pixels for a patch I_i, p_{soft}^k is the softmax prediction output for $k'th$ pixel and O^k is the ground-truth label of $k'th$ context pixel.

The context learner is trained with a semantic label map containing the ground truth labels of the pixels to predict. At the end of this training step, the feature maps that correspond to the output of the *Context Learner* will be specialized in modeling the neighboring context of the target pixel.

As a standard CNN focuses only on learning the class of a given patch y_i, it is hard to infer what the last layers are actually learning. In contrast, our learner increases the interpretability of the whole network. In Fig. 2, we show the responses of our context learner maps for some input patches where feature maps learn to capture patch context.

Learning dependent features – The goal of this part of the augmented learner is to learn the parameters (θ_d^2, θ_p^2) from a random initialization of (θ_d^0, θ_p^0) and from parameters θ_c^1 learned in the previous step. We minimize $\mathcal{L}$ while keeping θ_c^1 fixed. Fixing θ_c prevents harming the parameters of the context learner while learning θ_d^2. We stochasticly replace context predictions with some true labels to regularize learning of $f_d(.)$.

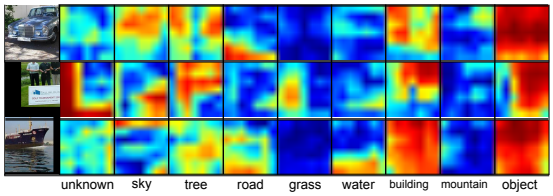


Figure 2: Feature maps of context learner for some input patches.

Fine tuning – In this step, we learn final parameters $\theta^3 = (\theta_c^3, \theta_d^3, \theta_p^3)$. We start from an initial value of $(\theta_c^1, \theta_d^2, \theta_p^2)$, and we minimize $\mathcal{L}$. This idea of this overall refinement step is to weaken the level of supervision and allow both θ_f and θ_d to adjust to this sudden lack of possible ground truth contextual information which is obviously not present during the test step.

Experiments Our approach has been tested on two scene labeling datasets: Stanford Background and SIFT Flow. The Stanford Background dataset contains 715 images of outdoor scenes having 9 classes. Our context learner transforms a 46×46 patch into a 7×7 context output. In the first layer, it has sixteen 7×7 filters and then 2×2 pooling operations for each feature map. Its second layer is composed of K filters (each of size 7×7) each encoding the context of a specific class followed by a 2×2 pooling operation. This layer has thus K output maps, where K corresponds to the number of classes.

(a) msConvnet (b) msAugLearner (c) GroundTruth

Figure 3: Raw image labeling of the multiscale ConvNet, our multiscale augmented learner and ground truth labels.

Both single scale and multiscale variants of the architecture has been analysed. While the accuracy gain varies between singlescale and multiscale implementations, we observe that our approach consistently improves both pixel and class accuracies. The gain on single-scale experiments are higher compared to multiscale implementations. This brings us to the empirical conclusion that contextual cues obtained implicitly through appearance cues of large support size provides valuable contextual information.

From a computational perspective, our approach increases the number of parameters by less than 1% compared to the ConvNet. Overall, we observe that our method provides better results for both the Stanford and the SIFT Flow datasets. Some labeling results from the Stanford dataset are shown in Figure 3. Our approach yields results that are more visually coherent than those obtained with the plain ConvNet architecture.

Adaptive Transductive Transfer Machines

Nazli Farajidavar
http://personal.ee.surrey.ac.uk/Personal/N.Farajidavar
Teofilo deCampos
http://personal.ee.surrey.ac.uk/Personal/T.Decampos
Josef Kittler
http://www.surrey.ac.uk/cvssp/people/josef_kittler

Center for vision, speech and signal processing
University of Surrey
Guildford, GU2 7XH
UK

Transductive transfer learning methods can potentially improve a very wide range of classification tasks, as it is often the case that a domain change happens between training and application of algorithms, and it is also very common that unlabelled samples are available in the target domain.

In this paper, we propose Adaptive Transductive Transfer Machine (ATTM) which combines methods that adapt the marginal and the conditional distribution of the samples, so that source and target datasets become more similar, facilitating classification (TTM). We further introduce two unsupervised dissimilarity measures which are the backbones of our classifier adaptation approach. ATTM uses these measures to select the best classifier and to further optimise its parameters for a new target domain. We show that our method obtains state-of-the-art results in cross-domain vision datasets using naïve features, with a significant gain in computational efficiency in comparison to related methods.

We propose the following TTM pipeline:

(a) A global linear transformation G^1 is applied to $\mathbf{X}^{src}$ and $\mathbf{X}^{trg}$ such that the marginal $P(G^1(\mathbf{X}^{src}))$ becomes more similar to $P(G^1(\mathbf{X}^{trg}))$. Following [2, 3, 4, 5] we adopt the Maximum Mean Discrepancy (MMD) for defining a projection matrix which aims to minimise the distance between the sample means of the source and target domains.

(b) With the same objective, a local transformation is applied to each transformed source domain sample $G_i^2(G^1(\mathrm{x}_{src}^i))$.

$$G_i^2(G^1(\mathrm{x}_{src}^i)) = G^1(\mathrm{x}_{src}^i) + \gamma \mathrm{b}^i \,, \tag{1}$$

Modeling the unlabelled target data, by a mixture of Gaussian probability density functions (GMM), we can formulate the problem of finding an optimal translation parameters b as one of maximising the likelihood of the translated source sample measured in the target domain.

$$\mathrm{b}^i = \frac{\sum_{k=1}^K P(\mathrm{x}^i + \mathrm{b}_0^i|\lambda_k)\Sigma_k^{-1}(\mathrm{x}^i - \mu_k)}{\sum_{k=1}^K P(\mathrm{x}^i + \mathrm{b}_0^i|\lambda_k)\Sigma_k^{-1}} \,, \tag{2}$$

where b_0^i is an initial value of b^i, which is set to a vector of zeros. In our experiments, we ran (2) only once, though one can iterate it further.

(c) Finally, aiming to reduce the difference between the conditional distributions in source and target spaces, a class-based transformation is applied to each of the transformed source samples $G_{y^i}^3(G_i^2(G^1(\mathrm{x}_{src}^i)))$ following the TST transformation of [1].

Figure 1 illustrates the effect of the three steps of the TTM pipeline.

In the Adaptive TTM we have an extra *classifier Selection and learning parameters adaptation* step where we introduce two unsupervised dissimilarity measures for selecting a proper classifier and for adapting its parameters. More specifically, when both dissimilarity measures indicate that the cross-domain datasets are very different, we suggest that it is better to use a non-parametric classifier, like Nearest Neighbour, so no optimisation is employed at training. When the two domains are similar at both global and cluster levels, it is sensible to use a classifier such as KDA, whose parameters optimised on the source domain have a better chance of working on the target space. And finally when two domains are similar at global levels but the clusters distribution in the two domains are different we propose to use the KDA but adapt the lengthscale σ of the RBF kernel using a linear function of the cluster dissimilarity measure.

Figure 2 demonstrates the full ATTM pipeline.

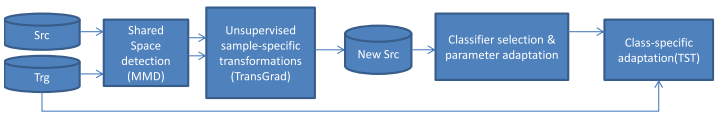
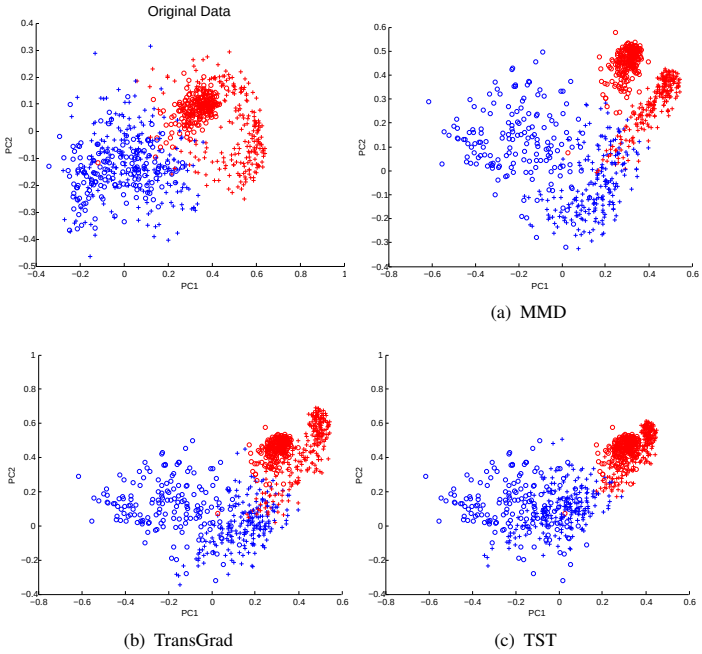


Figure 2: Adaptive Transductive Transfer Machine (ATTM).



(a) MMD (b) TransGrad (c) TST

Figure 1: The effect of different steps of our pipeline on digits 1 and 2 of the MNIST→USPS datasets, visualised in 2D through PCA. The source dataset (MNIST) is indicated by stars, the target dataset (USPS) is indicated by circles, red indicates samples of digit 1 and blue indicates digit 2 (better viewed on the screen).

Comprehensive experiments on MNIST, USPS, COIL20 and Caltech+ Office datasets show that our proposed TTM pipeline leverages the averaged performance by 1.32% compared to the best performing state-of-the-art of approach, JDA [3]. We have further tested our proposed *classifier Selection and learning parameters adaptation* on both JDA and TTM algorithms as AJDA and ATTM. The AJDA performance shows that the model adaptation drastically enhances the final classifier. The performance gains of **4.59** and **4.29** in ATTM and AJDA respectively validates the proposed dissimilarity measures for model selection and adaptation.

It is worth pointing out that ATTM is a general framework with applicability beyond image classification and could be easily applied to other domains, even outside Computer Vision. For future work, we suggest studying combinations of our method with instance reweighting methods and multi-source transfer learning.

[1] N. FarajiDavar, T. deCampos, J. Kittler, and F. Yang. Transductive transfer learning for action recognition in tennis games. In *ICCV,VECTaR workshop*, 2011.

[2] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel method for the two sample problem. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 19*, pages 513–520. MIT Press, 2007.

[3] M. Long, J. Wang, G. Ding, and P. Yu. Transfer learning with joint distribution adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[4] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1187–1192, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[5] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. A two-stage weighting framework for multi-source domain adaptation. In *NIPS*, pages 505–513, 2011. URL `http://dblp.uni-trier.de/db/conf/nips/nips2011.html#SunCPY11`.

Board #28

Randomized Support Vector Forest

Xutao Lv[1]
xutao.lv@sri.com
Tony X. Han[2]
hantx@missouri.edu
Zicheng Liu[3]
zliu@microsoft.com
Zhihai He[2]
hezhi@missouri.edu

[1] SRI International
Princeton, NJ, USA
[2] University of Missouri
Columbia, MO, USA
[3] Microsoft Research
Seatle, WA, USA

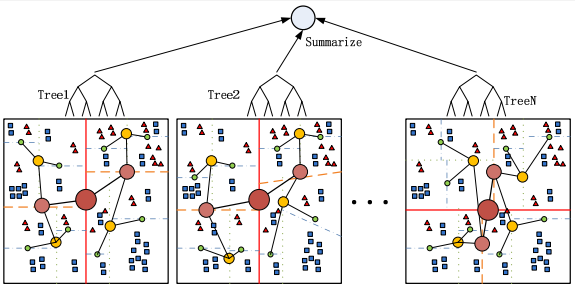


Figure 1: The structure of RSVF. This figure shows a RSVF with N trees. Each tree, with depth 5, is demonstrated in the last row of the figure. The small green dots are the LSVM classifiers; the other dots are the binary classifiers. Note, the binary classifiers mentioned in this paper represent decision nodes, which use a threshold to split the data into two child nodes.

Based on the structural risk minimization principle, the linear SVM aiming at finding the linear decision plane with the maximal margin in the input space has gained increasing popularity due to its generalizability, efficiency and acceptable performance. However, rarely training data are evenly distributed in the input space [1], which leads to a high global VC confidence [3], downgrading the performance of the linear SVM classifier. Partitioning the input space in tandem with local learning may alleviate the unevenly data distribution problem. However, the extra model complexity introduced by partitioning frequently leads to overfitting.

To solve this problem, we proposed a new supervised learning algorithm, Randomized Support Vector Forest (RSVF): Many partitions of the input space are constructed with partitioning regions amenable to the corresponding linear SVMs.

As illustrated in Figure 1, the RSVF consists of many Support Vector Trees (SVT). Each SVT represents a scheme of data partition and the corresponding local classifier. The final classification result of RSVF is a pooling from all the SVTs. After comparing various pooling methods including the majority voting, and max voting, i.e., taking the prediction from the SVT with the maximal confidence, we use majority voting from all of the trees in the forest for its simplicity and efficacy. We grow the RSVF through a procedure similar to growing the Classification And Regression Trees (CART) in random forest [7]. The steps of building RSVF is shown in Algorithm 1.

Input: Training dataset $\mathcal{X}$ and the number of trees N_{tree} in RSVF
Output: RSVF
for $t \leftarrow 1$ **to** N_{tree} **do**
  Randomly sample the bootstrap dataset $\mathcal{X}^*$ from $\mathcal{X}$; the Out-Of-Bag data will be $\mathcal{X} \backslash \mathcal{X}^*$;
  Train the SVTs $\mathcal{T}$ with both dataset $\mathcal{X}^*$ and $\mathcal{X} \backslash \mathcal{X}^*$;
end

Algorithm 1: Building RSVF

The generalization of the RSVF benefits from the randomness injected through random feature selection and bagging, which is also essential to the generalization of random forests [2].

The randomness of the partitions is injected through random feature selection and bagging. This partition randomness prevents the overfitting introduced by the over-complicated partitioning. With the injected randomness, the generalization error of RSVF can be proved to converge almost surely using the Law of Large Numbers when the number of SVTs increases. As the number of trees in RSVF increases, for almost surely all Θ, the generalization error e_g of RSVF converges to,

$$P_{\mathbf{X},Y}(P_\Theta(\mathcal{T}(\mathbf{X},\Theta)=Y) - \max_{j \neq Y} P_\Theta(\mathcal{T}(\mathbf{X},\Theta)=j<0) \qquad (1)$$

where $\mathcal{T}$ is an SVT; $\mathbf{X}$ is feature matrix; Y is the label of $\mathbf{X}$; and Θ is a set of parameters ϕ^* associated with the SVT $\mathcal{T}$.

We extensively evaluate the performance of the RSVF on several benchmark datasets, originated from various vision applications, including the four UCI datasets, the letter dataset, the KTH and the UCF sports dataset, and the Scene-15 dataset. The performance is shown in Table 1 and Table 2. The proposed RSVF outperforms linear SVM, kernel SVM, Random Forests (RF), and a local learning algorithm, SVM-KNN, on all of the evaluated datasets. The classification speed of the RSVF is comparable to linear SVM.

Method	LSVM	RF	RSVF	SVM-KNN	χ^2-KSVM	RBF-KSVM
KTH*	92.59%	91.67%	**93.98%**	87.04%	92.59%	92.13%
UCF	65.7 ± 5.8%	61.5 ± 7.3%	**72.2 ± 5.4%**	48.4 ± 5.6%	66.3 ± 6.6%	62.3 ± 6.7%
Scene15	75.1 ± 0.3%	63.3 ± 0.9%	**78.3** ± 0.4%	59.9 ± 0.9%	76.9 ± 0.4%	75.7 ± 0.6%

Table 1: Recognition accuracy on KTH, Scene-15 and UCF sports datasets. *Note: since the training and the testing sets are fixed in the KTH dataset, we just follow the standard setup so that our result can be compared with [4, 5, 6, 9].

Type	Best in [8]	Linear SVM	RBF-SVM	RSVF	RF
dna	0.059 ± 0.005	0.088 ± 0.017	0.054 ± 0.010	**0.052 ± 0.008**	0.056 ± 0.011
wine	0.030 ± 0.029	0.023 ± 0.024	0.016 ± 0.022	**0.002 ± 0.007**	0.014 ± 0.016
iris	0.057 ±0.022	0.038 ± 0.026	0.032 ± 0.025	**0.029 ±0.048**	0.041 ±0.029
glass	0.232 ±0.047	0.408 ± 0.091	0.300 ± 0.059	**0.223 ± 0.068**	0.234 ± 0.055

Table 2: Performance comparison on UCI datasets. The results in the first column is obtained from [8].

[1] Leon Bottou and Vladimir Vapnik. Local learning algorithms. *Neural Computation*, 4:888–900, 1992.

[2] Leo Breiman and E. Schapire. Random forests. In *Machine Learning*, volume 45, pages 5–32, 2001.

[3] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, 2005.

[5] Andrew Gilbert, John Illingworth, and Richard Bowden. Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):883–897, 2011.

[6] Nazli Ikizler-Cinbis and Stan Sclaroff. Object, scene and actions: combining multiple features for human action recognition. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 494–507, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15548-0, 978-3-642-15548-2.

[7] Daniel F. Schwarz, Inke R. König, and Andreas Ziegler. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 27(3):439, 2011.

[8] Chunhua Shen and Zhihui Hao. A direct formulation for totally-corrective multi-class boosting. In *CVPR*, pages 2585–2592, 2011.

[9] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

Board #29

Reverse Image Segmentation: A High-Level Solution to a Low-Level Task

Jiajun Wu
http://jiajunwu.com
Jun-Yan Zhu
http://www.eecs.berkeley.edu/~junyanz
Zhuowen Tu
http://pages.ucsd.edu/~ztu

CSAIL
Massachusetts Institute of Technology
Computer Science Division
University of California, Berkeley
Department of Cognitive Science
University of California, San Diego

Image segmentation is a fundamental and widely studied problem in computer vision [1, 2, 4]. Continuous efforts have been made to improve the performance of segmentation systems to match human capability [1]; however, it is generally acknowledged that solving the segmentation problem with low-level cues alone might not be possible. There has long been a discussion on solving this seemingly low-level task with high-level knowledge [3], but a clear and concrete solution is not yet available.

Two main issues (both due to the lack of semantic understanding) contribute to the main difficulty in image segmentation: (1) regions of different appearances might belong to the same segment, (2) and different image segments might have identical local appearances. In this paper, we propose to perform image segmentation in a reverse way. Our method takes a path of a high-level segmentation approach: at first per-pixel labeling of semantic categories is performed, followed by a procedure to obtain segmentations with per-pixel labels got discarded in the end. We are inspired from the observation that semantic labels give means of differentiating similar pixels and grouping dissimilar pixels. These labels can be viewed as a quantization of the solution space of segmentation, and the derived segmentations are mostly consistent even when the semantic level labels are not completely correct. For example, in Figure 1, a mammal is classified as a bird because of their similarity in color and texture, but the derived segmentation is mostly correct.

The LM+SUN dataset [5] can serve as a large-scale semantic knowledge base, which provides generic high-level information. To utilize this knowledge, we train a discriminative multi-class classifier on top of the superpixels of the outdoor images in the LM+SUN dataset, which we found to be sufficient for the task of general image segmentation.

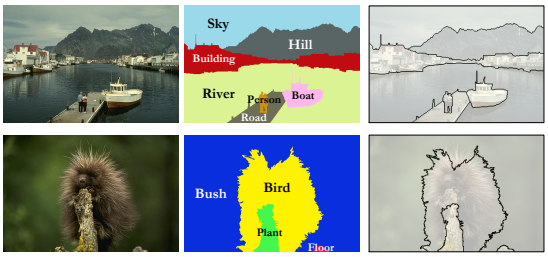
Specifically, we first assign each superpixel a semantic label. Following [5], a superpixel is associated with a semantic class if and only if at least half of the superpixel overlaps with a ground truth segment mask with that label. Then, according to the label frequencies on superpixels, 50 most frequent classes are picked out. For each class, 20,000 superpixels of the class are sampled as positive training examples, and another 20,000 superpixels unlabeled or with other class labels are randomly drawn as negative examples; a linear SVM is then trained on the data. These classifiers are generic and applicable to any images including those not in the dataset. For segmentation, each superpixel is tested by all learned classifiers to obtain a vector of confidence values.

We then formulate the problem under the framework of Conditional Random Fields (CRF). Constraints that allow us to reduce over/under segmentations near region boundaries are encoded as pairwise edge potentials. Denoting $S = \{s_i\}$ as a set of superpixels and $G(S,E)$ as an adjacency graph, the probability of class labels $\boldsymbol{c} = \{c_i\}$, given the set S and weights λ, μ, can be formulated as

$$-\log(\Pr(\boldsymbol{c}|G;\lambda,\mu)) = \sum_{s_i \in S} \Phi(c_i|s_i) + \sum_{(s_i,s_j)\in E} [\lambda \Psi(c_i,c_j) + \mu \Theta(c_i,c_j|s_i,s_j)]. \quad (1)$$

The unary potentials Φ are directly defined as the probability output of our multi-class classifier: $\Phi(c_i|s_i) = -\log(\Pr(c_i|s_i))$. Similar to [5], the first binary potentials Ψ are defined as probabilities of label co-occurrence: $\Psi(c_i,c_j) = -\log[(\Pr(c_i|c_j) + \Pr(c_j|c_i))/2] \cdot \delta[c_i \neq c_j]$, where $\Pr(c_i|c_j)$ is the conditional probability of one superpixel having label c_i given that its neighbor has label c', estimated from the training set, and $\delta[\cdot]$ is the indicator function. The second pairwise terms Θ are defined as $\Theta(c_i,c_j|s_i,s_j) = W(s_i,s_j)/(1+\|s_i - s_j\|) \cdot \delta[c_i \neq c_j]$, where $\|s_i - s_j\|$ is the L_2 difference between the feature vectors of superpixels s_i and s_j, and $W(s_i,s_j)$ is the normalized shared boundary length. W can be formulated as $W(s_i,s_j) = [L(s_i)^{-1} + L(s_j)^{-1}] \cdot L(s_i,s_j)$, where $L(s_i)$ is the length of boundary of superpixel s_i, and $L(s_i,s_j)$ is the shared boundary length between s_i and s_j.

There are two parameters λ and μ in our formulation, which represent the effects of high-level contextual information and low-level spatial regularization, respectively. Given λ and μ, we adopt MCMC methods for inference. Because the CRF is built on superpixels, the inference is highly efficient, taking approximately 0.1 second per image on average.



(a) (b) (c)

Figure 1: Example images and their semantic labeling and image segmentation results. Even if the semantic labels are not perfect, our pipeline could obtain satisfactory segmentation results.

	BSDS300					
	Covering ↑		PRI ↓		VoI ↑	
	ODS	OIS	ODS	OIS	ODS	OIS
Human	0.73	0.73	0.87	0.87	1.16	1.16
RIS+HL	**0.59**	**0.65**	**0.82**	**0.86**	1.71	1.53
RIS+H	0.55	0.60	0.80	0.84	1.82	1.63
RIS+L	0.57	0.63	0.79	0.82	1.80	1.60
RIS	0.52	—	0.77	—	1.99	—
SuperParsing	0.48	—	0.74	—	2.07	—
gPb-owt-ucm	**0.59**	**0.65**	0.81	0.85	**1.65**	1.47
fPb-owt-ucm	0.57	0.63	0.80	0.84	1.69	1.49
cPb-owt-ucm	**0.59**	**0.65**	0.81	0.85	1.66	**1.46**
MShift	0.54	0.58	0.78	0.80	1.83	1.63
FH	0.51	0.58	0.77	0.82	2.15	1.79
Canny	0.48	0.56	0.77	0.82	2.11	1.81
MNCuts	0.44	0.53	0.75	0.79	2.18	1.84
SWA	0.47	0.55	0.75	0.80	2.06	1.75
Quad-Tree	0.33	0.39	0.71	0.75	2.34	2.22

Table 1: Comparison on the test sets of BSDS300 and BSDS500 with both supervised and unsupervised methods. For each measure, the best algorithm is highlighted.

We finally discard the semantic labels produced by CRF to obtain segmentations. The proposed image segmentation framework is tested both with and without the high/low-level pairwise potentials, resulting in four variants (RIS, RIS+H, RIS+L, RIS+HL). For completeness, we also evaluate the segmentations derived from the outputs of a state-of-the-art nonparametric semantic labeling system (SuperParsing) [5].

As shown in Table 1, our solution yields highly competitive results on the famous Berkeley Segmentation Benchmark (BSDS300) [1]. When methods based purely on the ambiguous low-level features [1] tend to merge patches of similar appearances but different semantics, high-level semantic knowledge could help to figure out a correct segmentation. We also conduct experiments on multiple other datasets and obtain consistent results. Detailed illustrations and comparisons can be found in our paper and supplementary material.

[1] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2011.

[2] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE TPAMI*, 24(5):603–619, 2002.

[3] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *IJCV*, 72(2):195–215, 2007.

[4] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

[5] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 2010.

Board #30

All together now: Simultaneous Object Detection and Continuous Pose Estimation using a Hough Forest with Probabilistic Locally Enhanced Voting

Carolina Redondo-Cabrera[1]
carolina.redondoc@alu.uah.es
Roberto López-Sastre[1]
robertoj.lopez@uah.es
Tinne Tuytelaars[2]
Tinne.Tuytelaars@esat.kuleuven.be

[1] University of Alcalá
GRAM
Alcalá de Henares, ES

[2] K.U. Leuven,
ESAT-PSI, iMINDS
Leuven, BE

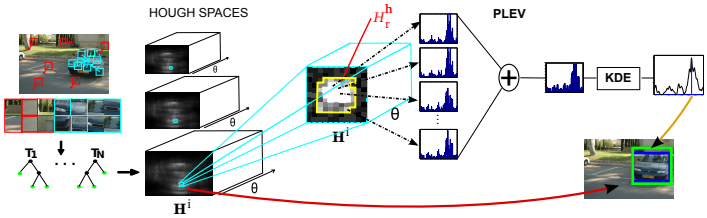


Figure 1: Our approach is able to jointly estimate the localization and the continuous pose of objects. To this end, we follow a HF regression voting in conjunction with our PLEV strategy, to integrate votes from a local region in the Hough space near the detected modes.

Object category detection has received a lot of attention over the last decades. Recently, several approaches have gone one step further proposing solutions for the problem of simultaneous object category detection and pose estimation [1, 3, 5]. In this paper, we tackle this problem using Hough Forests (HF) [2]. We propose a new approach (see Figure 1) which *jointly* solves both tasks, providing detection hypotheses and *probabilistic* estimates of their *continuous* pose.

We first introduce a new formulation for the regression to be performed with HF, incorporating an **uncertainty criterion for the continuous pose of the categories**. This uncertainty in pose is decoupled from the traditional localization uncertainty [2], which allows us to randomly choose between them during the HF learning. The resulting HF can effectively locate objects and estimate their pose.

For a set of patches S, we formulate this pose uncertainty as follows,

$$\mathcal{M}_p(S) = \sum_{child \in (left,right)} \sum_{j:c_j=1} \left(\frac{min\{(||\boldsymbol{\theta}_j - \boldsymbol{\theta}_A||), 360^\circ - (||\boldsymbol{\theta}_j - \boldsymbol{\theta}_A||)\}}{180^\circ} \right)^2, \quad (1)$$

where c_j is the class label of the j patch ($c_j = 1$ for foreground patches, and $c_j = 0$ for background patches), $\boldsymbol{\theta}_j$ encodes the continuous pose annotation for the patch j, and $\boldsymbol{\theta_A}$ is the viewpoint angle average over all foreground patches in the set of patches S^{child}. Randomly switching between this pose uncertainty and the localization uncertainty of [2] guarantees that the leaves of our decision trees gather image patches which vote not only for a similar object localization, but also for a similar pose.

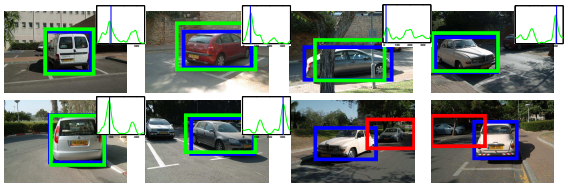
However, the extension of the Hough space to cover also the pose regression turns out to be suboptimal. The main reason is that the pose voting is very noisy, as we have experimentally observed, especially for views with shared appearance (*e.g.* think of a frontal vs. frontal-left views of a car). Instead, we propose to first localize the object, and then estimate its pose. For this second step, a novel regression strategy is introduced, named **Probabilistic Locally Enhanced Voting** (PLEV), which consists in modulating the regression with a kernel density estimation (KDE) to consolidate all the votes in a *local* Hough region near the maxima detected in the Hough space.

During testing, patches sampled from the test image traverse the trees and cast votes to the Hough space $\mathcal{H}$ based on the location and pose distributions stored in the leaves. The forest-based estimate is then computed by aggregating votes from different patches. The PLEV starts by collecting the votes in our multidimensional Hough space $\mathcal{H}$. We first project all votes on the (x,y) subspace of $\mathcal{H}$, and recover the object center hypothesis $\hat{\mathbf{h}}_\mathbf{d} = (\hat{x}, \hat{y})$ where the maximum is.

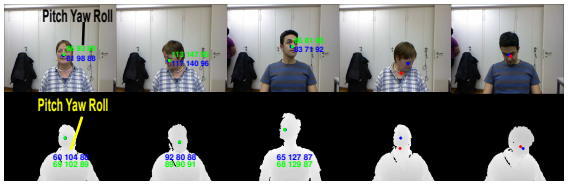
We then build a local Hough region $H_r^{\hat{\mathbf{h}}_\mathbf{d}} \subset \mathcal{H}$ for each detection hypothesis $\hat{\mathbf{h}}_\mathbf{d}$. We consider to be in the defined local region only those voting positions which receive at least one vote and are spatially close to the detected maximum. Then, PLEV aggregates all *pose* votes received within $H_r^{\hat{\mathbf{h}}_\mathbf{d}}$, obtaining the distribution of the poses in the Hough region (see Figure 1). Then, a Gaussian KDE is performed on that distribution in order to obtain a smooth probability density function (PDF) for the pose estimation. So, with the PLEV, our HF can cope with the uncertainty of the pose estimation votes.

To further improve the detections, we finally propose to integrate a novel **pose-based backprojection** (BP) strategy to boost the bounding box (BB) estimation using the pose cues. Essentially, we extend the traditional BP strategy [2]. When computing the BP mask, we want to penalize patches that vote not only for different object locations, as in [2], but also for different poses. For more details, see Section 2.3 in the paper.

As a conclusion, we have proposed a new object detection and continuous pose estimation solution using HF. It can successfully detect objects, while the pose is estimated with a probabilistic output using the PLEV. Our method reports state-of-the-art results on 4 different datasets [1, 3, 4, 5]. We show results on cars as well as faces, and using RGB as well as depth images as input. As a HF based approach with simple features, it is efficient. Being a voting-based scheme, it is intrinsically robust to occlusions. While many state-of-the-art approaches need 3D CAD models for the object class of interest during training, our approach is simple in the sense that we are able to learn the model directly from annotated images. Lastly, thanks to our PLEV strategy, we obtain a probabilistic output score, allowing easy integration as a building block in a larger probabilistic framework. Our extension to video-based pose estimation shows how to leverage the temporal continuity in video, even though poses may change from frame to frame. In Figure 2 we show qualitative results for different categories and for different modalities.



(a) Weizmann Cars Viewpoint dataset [3]



(b) Biwi Kinect Head Pose Database [1]

Figure 2: Qualitative results. Ground truth in blue, estimations in green and wrong detections in red.

[1] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *IJCV*, 101(3):437–458, 2013.

[2] J. Gall, N. Razavi, and L. Van Gool. *An Introduction to Random Forests for Multi-class Object Detection*, chapter 11, pages 243–263. Springer, 2012.

[3] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and continuous pose estimation. *IVC*, 30(12):923–933, 2012.

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

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

Board #31

Semi-Global 3D Line Modeling for Incremental Structure-from-Motion

Manuel Hofer
hofer@icg.tugraz.at
Michael Donoser
donoser@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

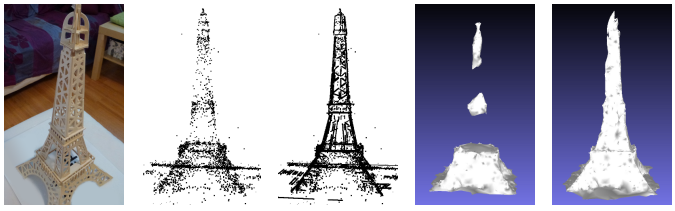
Institute for Computer Graphics and Vision
Graz University of Technology
Austria

1 Motivation

Recovering 3D information from a single moving camera is a widely studied field in the area of computer vision (e.g. [1]). Most of these Structure-from-Motion (SfM) approaches are based on so-called interest points (e.g. corners) in images, which can be accurately matched using powerful descriptors like SIFT [7]. Hence the output is usually a sparse 3D point cloud along with the camera poses for all successfully integrated images. While previous methods were only able to perform pose estimation and 3D reconstruction in an offline way, there are now more and more incremental SfM approaches available (e.g. [4]).

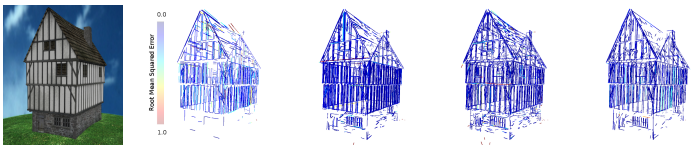
Since conventional SfM approaches are based on interest points, the distribution of the obtained 3D points is usually not uniform throughout the whole reconstruction. This is due to the fact that such interest points are usually located on highly textured areas, but not on homogeneous regions or along edges. Since the result of SfM pipelines is often used as basis to generate a more dense result or for localization and navigation tasks, it would be beneficial to generate additional complementary 3D information in an efficient way. From a SfM point of view, using line segments is especially interesting for urban and indoor environments, where linear structures frequently occur. While interest points are located mostly on richly textured image locations, line segments usually mark the boundaries of objects. Hence, incorporating such features in an online SfM pipeline to create 3D line segments naturally leads to a more complete 3D representation of the underlying scene, which is beneficial for all kinds of subsequent applications.

We propose a novel approach which generates 3D line models in a semi-global way directly on-the-fly, based solely on the output of a conventional incremental SfM pipeline. The goal of our method is to generate additional complementary 3D information to improve the sparse 3D representation of the scene. In this approach, we consider the SfM pipeline as a black box and do not interfere with the pose estimation procedure. We show that 3D line reconstructions can be obtained very efficiently by using purely geometric constraints, or by additionally incorporating appearance and collinearity information. Our approach enables accurate 3D reconstruction of texture-less as well as textured man-made objects, including complex structures such as wiry objects. Figure 1 shows a reconstruction result obtained by an incremental SfM system [4], followed by a surface generation method [5], with and without the usage of additional 3D line segments obtained by our proposed method. As we can see, additional 3D information significantly improves the completeness and overall appearance of the resulting reconstructions. For more technical details, we kindly refer to the full paper.



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

Figure 1: (a) An example image from the *EIFFEL* sequence. (b) The sparse 3D reconstruction result obtained by a conventional point-based SfM pipeline [4]. (c) The pointcloud combined with reconstructed 3D lines by our proposed method. On the right we can see an incrementally generated 3D mesh with (d) the 3D points only or (e) both points and lines. As we can see, the usage of complementary features significantly improves the completeness of the resulting 3D model in both cases.



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

Figure 2: Reconstruction results for the *Timberframe* sequence (240 images). (a) Example image. (b) The original result by [6] (offline, runtime of several hours), RMSE = 0.291. (c) The result by [2] (offline, runtime of 45 minutes), RMSE = 0.094. (d) The result by [3] (online, 5.7 minutes) RMSE = 0.196. (e) Our reconstruction with appearance and collinearity constraints enabled (online, 6.9 minutes), RMSE = 0.095.

2 Results

To demonstrate the capabilities of our proposed algorithm, we performed several quantitative and qualitative experimental evaluations. As a quantitative evaluation we used the synthetic *Timberframe* [1] dataset from [6], since there is a groundtruth CAD model available. Figure 2 shows our result in comparison to related state-of-the-art methods [2, 3, 6].

As can be seen, our proposed method achieves more accurate results than a previous incremental approach [3] (RMSE 0.095 vs. 0.196), while the runtime is not largely increased (6.9 vs. 5.7 min). That is off course due to the non-greedy nature of our approach and the incorporation of collinearity information. The accuracy with respect to the ground truth CAD model is almost as high as for the offline approach [2] (RMSE 0.095 vs. 0.094), which achieves the highest accuracy among the competitive algorithms, but with a significantly higher processing time (6.9 vs 45 min). For more results, please see the full paper.

Acknowledgements

This work has been supported by the Austrian Research Promotion Agency (FFG) project FreeLine (843459) and OMICRON electronics GmbH.

[1] S. Agarwal, N. Snavely, I. Simon, and S.M. Seitz. Building rome in a day, 2009. International Conference on Computer Vision (ICCV).

[2] M. Hofer, A. Wendel, and H. Bischof. Line-based 3D reconstruction of wiry objects, 2013. Computer Vision Winter Workshop (CVWW).

[3] M. Hofer, A. Wendel, and H. Bischof. Incremental line-based 3D reconstruction using geometric constraints, 2013. British Machine Vision Conference (BMVC).

[4] C. Hoppe, M. Klopschitz, M. Rumpler, A. Wendel, S. Kluckner, H. Bischof, and G. Reitmayr. Online feedback for structure-from-motion image acquisition, 2012. British Machine Vision Conference (BMVC).

[5] C. Hoppe, M. Klopschitz, M. Donoser, and H. Bischof. Incremental surface extraction from sparse structure-from-motion point clouds, 2013. British Machine Vision Conference (BMVC).

[6] A. Jain, C. Kurz, T. Thormaehlen, and H. Seidel. Exploiting global connectivity constraints for reconstruction of 3D line segments from images, 2010. International Conference on Computer Vision and Pattern Recognition (CVPR).

[7] D. Lowe. Distinctive image features from scale-invariant keypoints, 2004. International Journal of Computer Vision (IJCV).

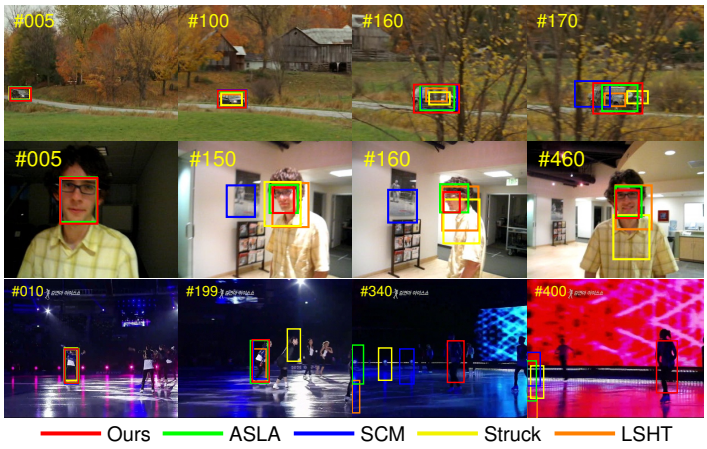
[1] http://www.mpi-inf.mpg.de/resources/LineReconstruction

Accurate Scale Estimation for Robust Visual Tracking

Martin Danelljan, Gustav Häger,
Fahad Shahbaz Khan, Michael Felsberg
martin.danelljan@liu.se, hager.gustav@gmail.com,
fahad.khan@liu.se, michael.felsberg@liu.se

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University
Linköping, Sweden

Robust scale estimation is a challenging problem in visual object tracking. Most existing methods fail to handle large scale variations in complex image sequences. This paper presents a novel approach for robust scale estimation in a tracking-by-detection framework. The proposed approach works by learning discriminative correlation filters based on a scale pyramid representation. We learn separate filters for translation and scale estimation, and show that this improves the performance compared to an exhaustive scale search while operating at real-time. Our scale estimation approach is generic as it can be incorporated into any tracking method with no inherent scale estimation.



Discriminative Correlation Filters. Our tracking approach is based on the discriminative correlation filters employed in the MOSSE tracker [1]. Similarly to [2], these filters are extended to multi-dimensional features for visual tracking. We use HOG features for the translation filter and concatenate it with image intensity features. In general, we consider a d-dimensional feature map representation of an image. Let f be a rectangular patch of the target, extracted from this feature map. We denote feature dimension number $l \in \{1,\dots,d\}$ of f by f^l. The objective is to find an optimal correlation filter h, consisting of one filter h^l per feature dimension. This is achieved by minimizing the cost function:

$$\varepsilon = \left\| \sum_{l=1}^{d} h^l \star f^l - g \right\|^2 + \lambda \sum_{l=1}^{d} \|h^l\|^2. \quad (1)$$

Here, g is the desired correlation output associated with the training example f and $\lambda \geq 0$ is a regularization parameter. The solution to (1) is:

$$H^l = \frac{\overline{G} F^l}{\sum_{k=1}^{d} \overline{F^k} F^k + \lambda}. \quad (2)$$

Capital letters denote the discrete Fourier transforms (DFTs) of the corresponding functions. We update the numerator A_t^l and denominator B_t of the correlation filter H_t^l in (2) separately using a learning rate η:

$$A_t^l = (1-\eta) A_{t-1}^l + \eta \overline{G}_t F_t^l \quad \text{and} \quad B_t = (1-\eta) B_{t-1} + \eta \sum_{k=1}^{d} \overline{F_t^k} F_t^k. \quad (3)$$

The correlation scores y at a patch z in the next frame are computed using (4). The new target state is found by maximizing the score y.

$$y = \mathscr{F}^{-1} \left\{ \frac{\sum_{l=1}^{d} \overline{A_t^l} Z^l}{B_t + \lambda} \right\}. \quad (4)$$

Our Scale Estimation Approach. Ideally, an accurate scale estimation approach should be robust while computationally efficient. To achieve this, we propose a fast scale estimation approach by learning separate filters for translation and scale. This helps by restricting the search area to smaller parts of the scale space. In addition, we gain the freedom of selecting the feature representation for each filter independently.

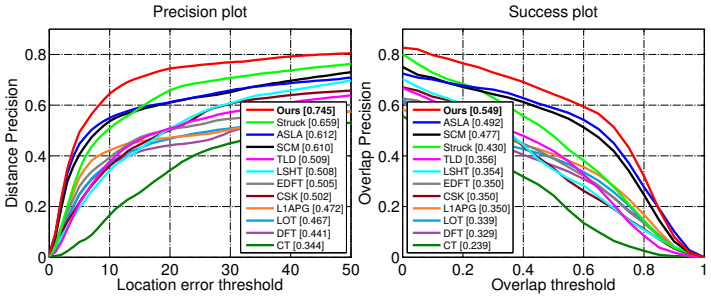


Figure 1: Precision and success plots illustrating the average distance and overlap precision respectively over all the 28 sequences. The average distance precision at 20 pixels for each method is reported in the legend of the precision plot. The legend of the success plot contains the *area-under-the-curve* (AUC) score for each tracker.

Method	median OP	median DP	median CLE	median FPS
Baseline (no scale)	37.8	74.5	15.9	**44.1**
Exhaustive Scale Search (this paper)	52.2	87.6	11.8	0.96
Fast Scale Search (this paper)	**75.5**	**93.3**	**10.9**	24.0

Table 1: Comparison of our fast scale estimation method with the baseline tracker and our exhaustive scale-space tracker.

We augment the baseline method by learning a separate 1-dimensional correlation filter to estimate the target scale in an image. The training example f for updating the scale filter is computed by extracting features using variable patch sizes centred around the target. Let $P \times R$ denote the target size in the current frame and S be the size of the scale filter. For each $n \in \left\{ \left\lfloor -\frac{S-1}{2} \right\rfloor, \dots, \left\lfloor \frac{S-1}{2} \right\rfloor \right\}$, we extract an image patch J_n of size $a^n P \times a^n R$ centred around the target. Here, a denotes the scale factor between feature layers. The value $f(n)$ of the training example f at scale level n is set to a HOG-based d-dimensional feature descriptor of J_n. Eq. 3 is then used to update the scale filter h_{scale} with the new sample f.

In visual tracking scenarios, the scale difference between two frames is typically smaller compared to the translation. Therefore, we first apply the translation filter h_{trans} given a new frame. Afterwards, the scale filter h_{scale} is applied at the new target location. An example z is extracted from this location using the same procedure as for f. By maximizing the correlation output (4) between h_{scale} and z, we obtain the scale difference.

Evaluation. We employ all the 28 sequences annotated with the scale variation attribute in the recent evaluation of tracking methods [3]. The sequences also pose challenging problems such as illumination variation, motion blur, background clutter and occlusion. The baseline HOG based tracker with no scale estimation capability is compared with our exhaustive scale space tracker and the fast scale estimation method in table 1.

We additionally compare our approach with 11 state-of-the-art trackers. Figure 1 contains the precision and success plots illustrating the *mean* distance and overlap precision over all the 28 sequences. In both precision and success plots, our approach significantly outperforms the compared methods. In summary, the precision plot demonstrates that our approach is superior in robustness compared to existing trackers. Similarly, the success plot shows that our method estimates the target scale more accurately on the benchmark sequences.

[1] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Yui M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010.

[2] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *CoRR*, abs/1404.7584, 2014.

[3] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

Board #33

Hough Networks for Head Pose Estimation and Facial Feature Localization

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

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

Head pose estimation and facial feature localization are keys to advanced human computer interaction systems and human behavior analysis. Due to their relevance, both tasks have gained a lot of attention in the computer vision community. Recent state-of-the-art methods like [1, 2, 3, 6] report impressive results and are real-time capable. However, those approaches rely on hand-crafted features. In contrast, we try to learn a feature representation from a set of training images. This is done by utilizing Convolutional Neural Networks (CNNs), which have shown to achieve outstanding results on various tasks such as image classification [5].

Instead of segmenting the head in a first step and then regressing the task-dependent parameters, we show in our paper a patch-based approach. Patches are densely extracted from the image along a regular grid and for each patch we perform a joint classification and regression. The classification segments the image patches into foreground and background, whereas the regression casts votes in a Hough space, but only for foreground patches. This is similar to the idea of Hough Forests (HFs) [4]. However, we replace the Random Forest (RF) with a CNN and call it therefore *Hough Network (HN)*.

Assuming that we have a training dataset $\{(x_s, \mathbf{t}_s)\}_{s=1}^{S}$ with S samples, where x_s denotes an image patch, and $\mathbf{t}_s$ encodes the foreground-background information as well as the regression targets, we want to train a CNN that minimizes the following error function

$$E_s(\theta) = \lambda_c E_{s,c} + \lambda_r E_{s,r}, \tag{1}$$

where $E_{s,c}$ and $E_{s,r}$ are the classification and regression error, respectively. The parameters λ_c and λ_r are weighting coefficients of the individual error functions and relate to increased or decreased delta values in the back-propagation algorithm. For classification, we utilize the cross-entropy error that is defined as follows

$$E_{s,c}(\theta) = -\left(t_{s,c} \ln(y_{s,c}) + (1 - t_{s,c}) \ln(1 - y_{s,c})\right). \tag{2}$$

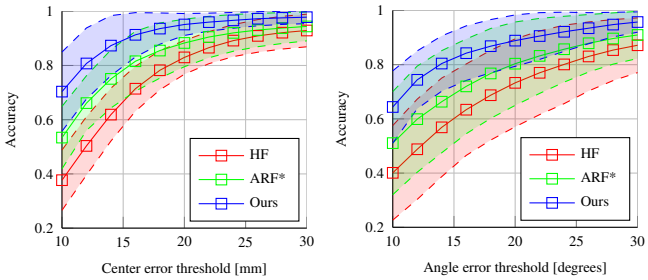
In contrast, for the regression targets we use the L_2 loss that minimizes the Euclidean distance between the target and predicted values:

$$E_{s,r}(\theta) = \frac{1}{2} \|\mathbf{y}_{s,r} - \mathbf{t}_{s,r}\|^2. \tag{3}$$

The objective function in Equation 1 allows that values in the single target vectors can be missing. In such cases we set the gradient values of the involved weights (which only effects connection to the output layer) to zero. We especially utilize this fact, if a patch does not belong to the foreground. In the case of a background patch, we back-propagate only the error values of the class information.

The straight-forward inference process in our HNs would be to densely extract overlapping patches from the image and evaluate the CNN for each patch independently. However, the structure of CNNs allows a more efficient method. We present the whole image as input to the CNN and if the patch stride (distance between two neighboring patch centers) is a multiply of the sum of the pooling widths, then the patches can be separated in the convolution and pooling layers. Only before the fully-connected layers we have to reshape the data to a matrix, where each patch corresponds to a single column. This allows us to perform classification and regression for all patches of an image in a single CNN evaluation.

We evaluated HNs on two challenging computer vision tasks. The first task deals with head pose estimation from consumer depth cameras. Given a depth image, we want to estimate the head center in 3D and its pose in Euler angles. We randomly split the sequences of the Biwi Kinect Headpose Database [3] into a train and a test set. A patch votes for a head center and a pose, if its foreground probability is > 0.99. Using a mean-shift variant [3], we find a single mode in the votes.



(a) (b)

Figure 1: Accuracy for the head center estimation error (a) and the angle error (b) of the HF [3], ARF* [6] and our approach. The curves visualize the fraction of correct estimates over an increasing success threshold. The solid lines represent the mean over five splits, whereas the shaded areas visualize the standard deviation.

Method	F. F. Error.	H. P. Error.
Conditional Random Forests (CRF) [2]	12.0	27.85
Robust Cascaded Pose Regression (RCPR) [1]	5.3	-
Ours	6.3	21.83
Human	4.5	-

Table 1: Performance of HNs compared to CRFs [2], RCPRs [1] and human performance on the LFW dataset as percentage of the inter-ocular distance.

The same approach can be utilized for facial feature localization. We evaluated our approach on the Labeled Faces in the Wild (LFW) dataset [2, 7], which also provides a discrete head pose. This information is incorporated into our HN by extending the error function:

$$E_s(\theta) = \lambda_c E_{s,c} + \lambda_r E_{s,r} - \lambda_h \sum_{i=1}^{5} \mathbf{t}_{s,h}^{(i)} \ln \mathbf{y}_{s,h}^{(i)}. \tag{4}$$

Further details and results can be found in our paper. Our conclusion is that HNs provide a powerful alternative to HFs, because it can learn a rich feature representation. Further, HNs could be adapted to various other tasks, such as human pose estimation and object detection.

[1] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, pages 1513–1520, 2013.

[2] Matthias Dantone, Juergen Gall, Gabriele Fanelli, and Luc J. Van Gool. Real-time facial feature detection using conditional regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.

[3] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc J. Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[4] Juergen Gall, Angela Yao, Nima Razavi, Luc J. Van Gool, and Victor S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (11):2188–2202, 2011.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[6] Samuel Schulter, Christian Leistner, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Alternating regression forests for object detection and pose estimation. In *International Conference on Computer Vision*, pages 417–424, 2013.

[7] Hai Wang, Bong-Nam Kang, and Daijin Kim. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

Board #34

Spherical Light Fields

Bernd Krolla[1]
http://av.dfki.de/~krolla
Maximilian Diebold[2]
http://hci.iwr.uni-heidelberg.de/Staff/mdiebold/
Bastian Goldlücke[3]
http://www.informatik.uni-konstanz.de/cvia/
Didier Stricker[1]
http://av.dfki.de/stricker

[1] German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[2] Heidelberg Collaboratory for Image Processing, Heidelberg, Germany
[3] Department for Computer and Information Science, University of Konstanz, Konstanz, Germany

A full view spherical camera exploits its extended field of view (FOV) to map its complete environment onto a 2D image plane. Thus, with a single shot, it delivers a lot more information about the surroundings than one can gather with a normal perspective or plenoptic camera, which are commonly used in light field imaging. However, in contrast to a light field camera, a spherical camera does not capture directional information about the incident light, and thus a single shot from a spherical camera is not sufficient to reconstruct 3D scene geometry.

In this paper, we introduce a method combining spherical imaging with the light field approach. To obtain 3D information with a spherical camera, we capture several independent spherical images by applying a constant vertical offset between the camera positions and combine the images in a *Spherical Light Field* (SLF).

Our approach differs from its related work in terms of expanded FOV and reduced acquisition time: Taguchi *et al*. [2] used an array of spherical mirrors to model catadioptric cameras for wide angle light field rendering, which implies decreasing tangential resolution close to the mirror borders and limits the FOV to $150^\circ \times 150^\circ$. Unger *et al*. [4] employed a fisheye-camera translated on a plane to capture hemispherical HDR images of a scene. The total acquisition time of up to 12 hours for a single scene restricts the application scenario to constantly illuminated indoor environments. Our proposed approach for SLF acquisition uses spherical cameras as shown in Figure 1(a) and allows to capture scenes within a few minutes, making it applicable to outdoor scenes.

A convenient description of this camera type is provided by Torii *et al*. [3], who consider a spherical camera to consist of a camera center C with a surrounding unit sphere acting as projection surface. This definition implies that no intrinsic parameters such as focal length or distortion values known from perspective imaging need to be considered (Figure 1(b)). By applying the Mercator projection [1], the spherical image is conformally mapped to an image on a cylinder surface Π (Figure 1(c)) allowing for epipolar plane image (EPI) reconstruction.

To describe a SLF, we define a new parametrization for the camera domain and the surrounding spherical 2D mapped image.We take the cylinder surface Π and denote the center line with Ω. The cylinder surface Π is parametrized by the image coordinates $(\phi,\theta) \in \Pi$. The line Ω contains the focal points $t \in \Omega$ of all possible camera positions in vertical direction.

A Spherical Light Field can then be described by a function

$$L : \Omega \times \Pi \to \mathbb{R} \qquad (t,\phi,\theta) \mapsto L(t,\phi,\theta), \tag{1}$$

where $L(t,\phi,\theta)$ defines the intensity of the incident light ray on the image plane (ϕ,θ) passing through the focal point t. To estimate the disparity, we address a 2D slice Σ_{ϕ^*} of the SLF by setting ϕ to a fixed value ϕ^*. The restriction of the light field to such a slice defines an EPI, being formally given as

$$S_{\phi^*} : \Sigma_{\phi^*} \to \mathbb{R} \tag{2}$$

$$(\theta,t) \mapsto S_{\phi^*}(\theta,t) := L(t,\phi^*,\theta). \tag{3}$$

Assuming a Lambertian scene, the EPI yields information about the disparity of a scene point in the form of orientated lines. To compute the disparity on the EPI, we can thus perform an orientation analysis on the given EPI S_{ϕ^*}, using a structure tensor. The orientation angle and thus the disparity map for the EPI S_{ϕ^*} can be computed directly from the components of the structure tensor.

A example for a resulting disparity map, is shown in Figure 1(d) and was computed by iterating over all EPIs from the SLF and storing the computed disparity at the corresponding azimuthal slice.

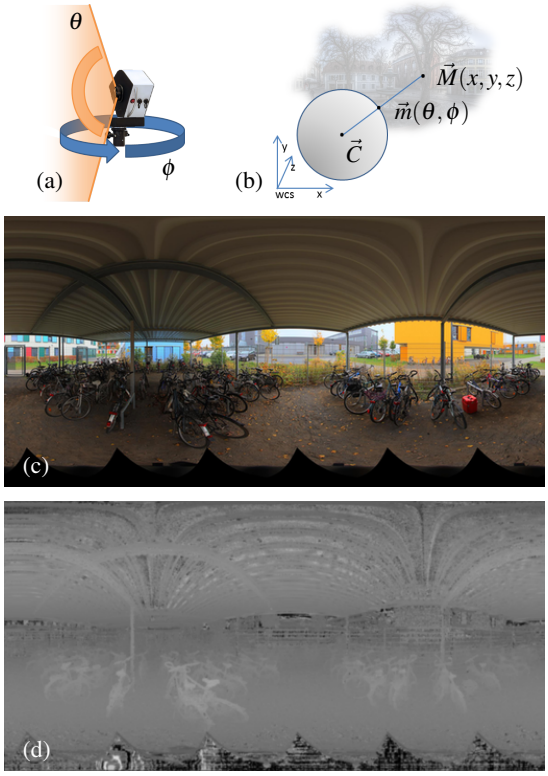


Figure 1: (a) Spherical image acquisition using a rotating tripod mounted camera equipped with a fish eye lens. (b) The spherical image results from the back projection of 3D points $M(x,y,z)$ to their corresponding image points $m(\theta,\phi)$ with $\phi[0,2\pi)$ and $\theta[0,\pi]$ assuming C to be the cameras center of projection. (c) In the current work, the resulting image is a *High Dynamic Range* (HDR) image with a resolution of 14000×7000 pixel. (d) shows the resulting disparity map of the captured scene.

The resulting full view spherical disparity map can then be employed for a 3D scene reconstruction of the camera's surroundings. Benchmarks on synthetic datasets demonstrate good accordance with the ground truth data. Finally simplifies the combination of spherical and HDR imaging approaches greatly the task of disparity estimation for real scenes, e.g. due to improved contrast, as shown in our work.

[1] Mats Bentsen, Geir Evensen, Helge Drange, and Alistair Jenkins. Coordinate transformation on a sphere using conformal mapping. *MW Review*, (127):2733–2740, 1999.

[2] Yuichi Taguchi, Amit Agrawal, Ashok Veeraraghavan, Srikumar Ramalingam, and Ramesh Raskar. Axial-cones: Modeling spherical catadioptric cameras for wide-angle light field rendering. *ACM Transactions on Graphics-TOG*, 29(6):172, 2010.

[3] Akihiko Torii, Atsushi Imiya, and Naoya Ohnishi. Two- and three-view geometry for spherical cameras. In *Proceedings of the sixth workshop on omnidirectional vision, camera networks and non-classical cameras*. Citeseer, 2005.

[4] Jonas Unger, Andreas Wenger, Tim Hawkins, Andrew Gardner, and Paul Debevec. Capturing and rendering with incident light fields. In *Proceedings of the 14th Eurographics workshop on Rendering*, pages 141–149. Eurographics Association, 2003.

Board #35

CoConut: Co-Classification with Output Space Regularization

Sameh Khamis
sameh@umiacs.umd.edu
University of Maryland
College Park, MD 20740

Christoph H. Lampert
chl@ist.ac.at
IST Austria
Am Campus 1, 3400 Klosterneuburg

Classification is one of the most fundamental and best understood machine learning problems. Different scenarios differ strongly in their training procedure, but agree fundamentally in their prediction step at test time: each test sample is assigned a label individually. However, in many real-world the samples to be classified occur in batches, such as words in a document, images in a photo collection, or stocks in a portfolio, and exploiting this fact should make it possible to achieve increased classification accuracy.

To motivate our framework, consider the situation of a linear classifier, which is efficiently trainable and exhibits good generalization capabilities but has a decision hypersurface that might not perfectly reflect the class boundaries in feature space. Given sufficiently many test samples it should be possible to modulate the classifier's decision boundary, for example, based on the cluster assumption, which states that class decision boundaries typically do not cross high density regions (see Figure 1 for an illustration).

Despite its potential, the task of co-classification, *i.e.* classifying a set of points jointly, has received little attention in the literature. In this work, we introduce CoConut, a method for co-classification based on the established principle of regularized risk minimization. It jointly labels all test points by minimizing a regularized risk functional that that incorporates additional information in the output (label) space. CoConut only requires the output of a set of classifiers as input, but makes no assumption on how they were trained. It is also efficient, as it requires no additional training step but only solves a regularized risk functional using efficient energy minimization techniques.

We formalize the co-classification scenario in the following way. We are given a set of (test) examples, $X = \{x_1, \dots, x_n\}$ from an input space $\mathcal{X}$, and we want to predict labels $Y = \{y_1, \dots, y_n\}$ from a label set $\mathcal{Y} = \{1, \dots, L\}$. For this task we have access to L fixed *base classifiers* with prediction functions, $f_1, \dots, f_L : \mathcal{X} \to \mathbb{R}$, where for any $x \in \mathcal{X}$ and $l \in \mathcal{Y}$ the value $f_l(x)$ reflects a confidence that the sample x belongs to class l. The straight-forward choice for labeling the test points is then to predict (greedily) the most confident label for each sample, $y_i = \operatorname{argmax}_{l=1,\dots,L} f_l(x_i)$.

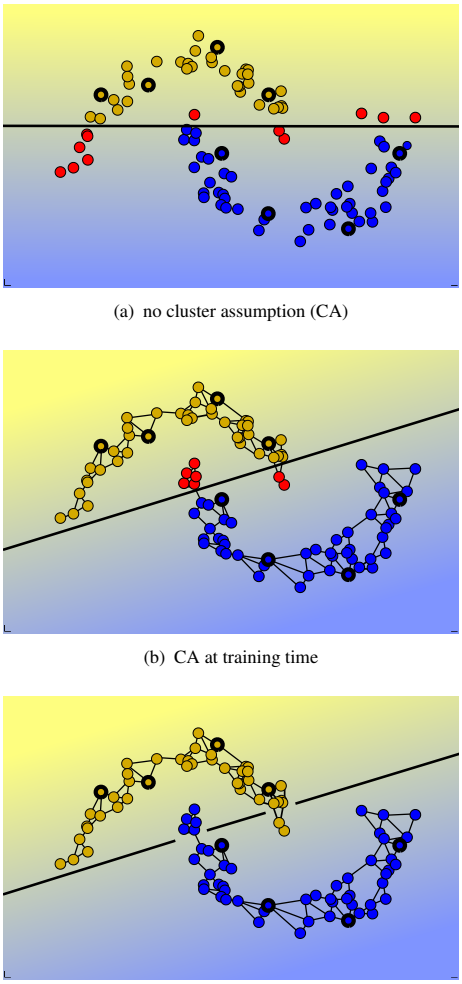
We propose to compute a joint labeling $y^* = (y_1^*, \dots, y_n^*) \in \mathcal{Y}^n$ of the test points by solving the following optimization problem:

$$y^* = \operatorname*{argmin}_{y \in \mathcal{Y}^n} \; -\sum_{l=1}^{L} [\![y_i = l]\!] f_l(x_i) + \lambda \Omega(y), \tag{1}$$

where Ω is a regularizer that penalizes undesirable label combinations and $\lambda \in \mathbb{R}^+$ is a constant that controls the regularization strength. Note that for $\lambda \to 0$ we recover independent per-sample predictions, showing that per-example label selection can be thought of as a special case of this framework. Equation (1) resembles the expressions occurring in the classical framework of *regularized risk minimization* [1]. The difference lies in the fact that we regularize in the output space (the space of all labelings), not in the space of classifier parameters. Therefore, we call the resulting approach *Co-Classification with output space regularization* (CoConut).

In our choice of regularizer we encode the *inductive bias* we have about the problem. Often this would be an assumption that the true labels vary smoothly with respect to the inputs. For any point x_i, let $N_i \subset X$ be the set of neighbors that are similar to x_i. Let w_{ij} denote the a measure of the similarity between two neighbors x_i and x_j. For any $x_j \in N_i$ the slope of g between x_i and x_j is $w_{ij}\delta_{ij}(g)$, where $\delta_{ij}(g) := [\![g(x_i) \neq g(x_j)]\!]$ indicates whether g changes value between x_i and x_j. Averaging this quantity across all neighbors and all points, we obtain a measure for the average discontinuity (lack of smoothness) of any labeling function $g \in \mathcal{G}$:

$$\Omega_S(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|N_i|} \sum_{x_j \in N_i} w_{ij} \delta_{ij}(g). \tag{2}$$



(a) no cluster assumption (CA)

(b) CA at training time

(c) CA at test time

Figure 1: Schematic illustration of the effect of the cluster assumption. Left: supervised training of a linear classifier with few training examples (bold circles): many mistakes occur at test time (red dots). Middle: cluster assumption during training reduces errors. Right: cluster assumption at test time reduces errors even further.

A regularizer can also encode a preference for a certain class label distribution at test time. This can counter the effect of the bias introduced by training with imbalanced class distributions. We assume that the target (expected) class label proportion for class l is Q_l, where $\sum_{l=1}^{L} Q_l = 1$. We define a measure for the disparity between the class label proportion $p_l(g)$ induced by labeling function g and the target proportion for each class l:

$$\Omega_D(g) = \sum_{l=1}^{L} |p_l(g) - Q_l| \tag{3}$$

where $p_l(g) = \frac{1}{n} \sum_{i=1}^{n} [\![g(x_i) = l]\!]$ are the label proportions the hypothesis g. The regularizer Ω_D penalizes the deviation from the target distribution, and this penalty is linear in the amount of deviation.

In the paper we discuss these regularizer choices, what information they can incorporate at test time, how they can be efficiently optimized, and their effect on the classification performance theoretically and empirically. We report our results using each regularizer on six different datasets, reporting consistent improvements over baselines.

[1] V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.

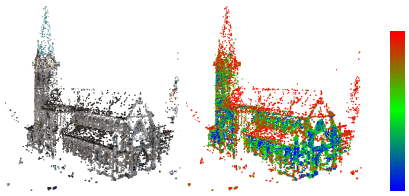
Board #36

A unified framework for content-aware view selection and planning through view importance

Massimo Mauro[1]
m.mauro001@unibs.it
Hayko Riemenschneider[2]
http://www.vision.ee.ethz.ch/~rhayko/
Alberto Signoroni[1]
http://www.ing.unibs.it/~signoron/
Riccardo Leonardi[1]
http://www.ing.unibs.it/~leon/
Luc Van Gool[2]
http://www.vision.ee.ethz.ch/~vangool/

[1] Department of Information Engineering
University of Brescia
Brescia, Italia

[2] Computer Vision Lab
Swiss Federal Institute of Technology
Zurich, Switzerland

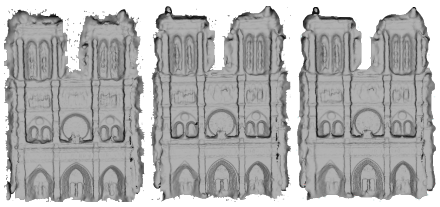


(a) SFM (b) E_{agg}

Figure 1: View importance as energy heatmap (the more red, the more salient and hence important) as example on *Fraumunster* SfM cloud.

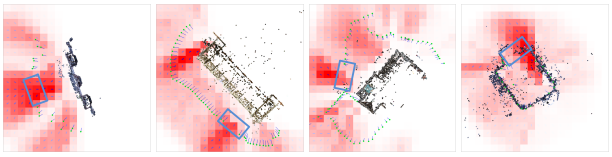
(a) Best1 (b) Best2 (c) Best3 (d) Worst1 (e) Worst2 (f) Worst3

Figure 2: Best and worst views on *Notre Dame* dataset.



(a) 715 img (all) (b) 167 (CMVS) (c) 106 (ours)

Figure 3: Similar 3D mesh results on *Notre Dame* for much smaller image sets. Our method effectively reduces yet keeps the salient 3D structures.



(a) Herz-Jesu-P25 (b) Hall (c) Fraumunster (d) Castle-P30

Figure 4: Next-Best-View grids. Importance is high in regions (blue rectangle) where cameras have been artificially removed.

Take home message: Reduction and selection of views through structure analysis of 3D point clouds. Our importance measure is much more effective without loosing salient structures.

Introduction: The great and unordered deal of images available on the Internet leads to two challenging problems for image-based 3D reconstructions: completeness and scalability. On one side, photographs are only taken from "popular" viewpoints, leading to incomplete 3D models. On the other side, the collected images are redundant. Next-Best-View (NBV) and Image Selection (IS) algorithms are thus needed to propose new and select from redundant viewpoints for efficient reconstruction.

In this work we propose two methods for IS and NBV, based on the idea of *view importance*: how important is a given viewpoint for a 3D reconstruction? Our answer is a unified framework for search of important viewsbased on a set of content-aware *quality features* extracted on the Structure-from-Motion (SfM) point cloud.

Quality Features. For every 3D point, we extract the following:

- *Density* is defined as the number of points contained in a sphere around the point.
- *Uncertainty* considers the maximum angle between the viewing directions of the evaluated point.
- *2D saliency* evaluates the meaningfulness of the 2D content around the point. It is estimated by reprojecting the point in the original images and measuring the gradient in the neighbourhood.
- *3D saliency* measures the geometric complexity around a point. It is estimated by the Difference of Normals (DoN) operator [2].

Feature aggregation. All the features have different ranges. We rescale them in the range [0,1] using a logistic function and we call *normalized energies* the obtained values. We note them as E_D, E_U, E_{2D} and E_{3D} respectively. The *aggregate energy* (example in Figure 1) is then defined as a linear combination

$$E_{agg} = w_D E_D + w_U E_U + w_{2D} E_{2D} + w_{3D} E_{3D} \quad (1)$$

View importance. The key concept behind both our IS and NBV algorithms is the *view importance*. Given a point cloud $\mathcal{P}$, the view importance I of a camera C is defined as the mean energy E_{agg} combined over all its visible points:

$$I(C,\mathcal{P}) = \frac{\sum_{p_i \in V_C} E_{agg}(p_i)}{|V_C|} \quad (2)$$

where V_C is the set of points in $\mathcal{P}$ visible from camera C. We use this basic definition in two variants I_{IS} and I_{NBV} (for IS and NBV respectively) to better adapt to the problem at hand. See paper for details.

View selection. The aim of image selection (IS) is to remove redundant images. We use an "importance-guided" approach: at every step our algorithm cuts out the *worst view* in terms of *view importance*, for an example see Figure 2. The worst view satisfies the relation:

$$C_{IS} = \arg\min_C I_{IS}(C,\mathcal{P}) \quad (3)$$

Next-Best-View planning. The goal of a Next-Best-View algorithm is to find the camera C_{NBV} with the largest view importance

$$C_{NBV} = \arg\max_C I_{NBV}(C,\mathcal{P}) \quad (4)$$

Since a great deal of images are collected by humans, we simplify the NBV search by fitting a plane primitive to the SfM camera centers. We then define a rectangular region around the point cloud and divide it in cells. We position a camera in every grid cell we evaluate the view importance for a given number of evenly spaced orientations, obtaining *view importance grids* as in Figure 4.

Experiments. The experiments show the effectiveness of the proposed content-aware methods. Our NBV planning effectively finds regions where viewpoints are missing. Our IS method reduces the number of images without losing salient regions of the scene, comparing favorably with the state-of-the-art image selection in CMVS [1]. E.g., For *Notre Dame*, we can remove more than 90% of the images (609/715), reducing the runtime of the reconstruction to 1/20th of the time without causing significant differences in reconstruction quality (Figure 3).

Acknowledgments. This work was supported by the European Research Council (ERC) under the project VarCity (#273940) and by the Italian Ministry of Education, University and Research under the PRIN project BHIMM (Built Heritage Information Modeling and Management).

[1] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010.

[2] Y. Ioanou, B. Taati, R. Harrap, and M. Greenspan. Difference of normals as a multi-scale operator in unorganized point clouds. In *3DPVT*, 2012.

Reproduction Angular Error: An Improved Performance Metric for Illuminant Estimation

Graham D. Finlayson
g.finlayson@uea.ac.uk
Roshanak Zakizadeh
r.zakizadeh@uea.ac.uk

School of Computing Sciences
The University of East Anglia
Norwich, UK

Illuminant Estimation which is the process of estimating the colour of the prevailing light and discounting it from the image is often done as the preprocessing step in computer vision, so that the image colour be used as a stable cue for indexing, recognition, tracking, etc. [4, 5].

Almost all illumination estimation research uses the angle between the RGB of the actual measured illuminant colour and that estimated one as the recovery error, which is defined as:

$$err_{recovery} = \cos^{-1}\left(\frac{(\underline{\rho}^{E} \cdot \underline{\rho}^{Est})}{\|\underline{\rho}^{E}\|\|\underline{\rho}^{Est}\|}\right) \quad (1)$$

where $\underline{\rho}^{E}$ denotes the RGB of the actual measured light, $\underline{\rho}^{Est}$ denotes the RGB estimated by an illuminant estimation algorithm and '.' denotes the vector dot product. Over a benchmark set, the average angular performance is calculated (including mean, median, and quantiles) and different algorithms are ranked according to these summary statistics [3].

This paper argues that recovery angular error despite its wide spread adoption has a fundamental weakness which casts doubt on its suitability. We observe that the same scene, viewed under two different coloured lights, leads to different recovery errors for the same illuminant estimation algorithm, despite the fact that when we remove the colour bias due to illuminant (we divide out by light) exactly the same reproduction is produced.

To illustrate this point we show at the top of Figure 1 four images of the same scene from the SFU Lab dataset [1] which are captured under different chromatic lights, from left to right: solux-4700K+blue filter; Sylvania warm white fluorescent; solux-4700K+3202+blue filter and Philips Ultralume fluorescent. Notice how much the colour (due to illumination) varies from left to right. Now, using the simple gray-world algorithm [2] for illuminant estimation we estimate the RGB of the light (the average image colour is the estimated colour of the light). Dividing the images by this estimate we produce the image outputs shown in the second row. In this case gray-world works reasonably well and the object colours look correct (though, of course this is not always the case). It is easy to show that dividing out by the gray-world estimate (or, indeed the estimates made by most algorithms) that the **same** output reproduction is made. In the 3rd row of Figure 1 we show the recovery angular errors (the plot with open bullets). Even though the same reproduction is produced the recovery angular error varies from 5.5° to 9° (an 80% difference).

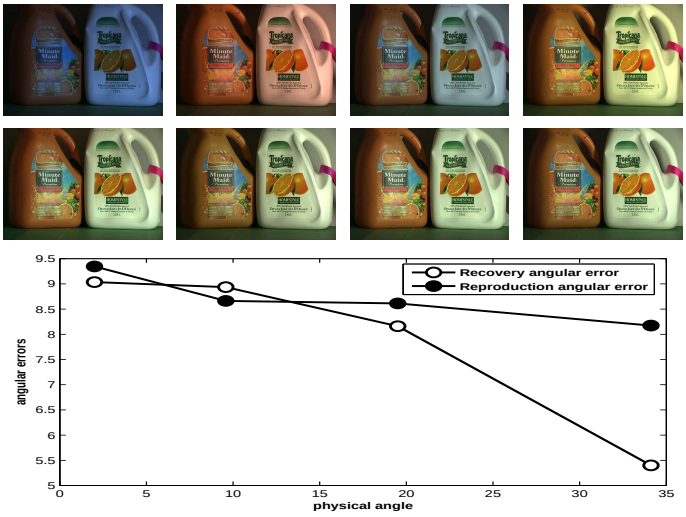


Figure 1: Row 1: four images captured under very chromatic illuminants. Row 2: corrected images using general gray-world [2] algorithm (Images are from [1]). Row 3: The Recovery angular error (conventional error measure) versus the Reproduction angular error (proposed error measure).

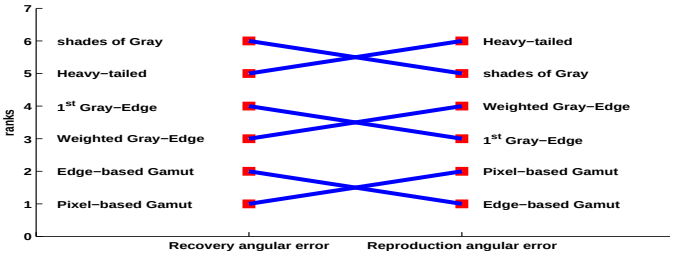


Figure 2: A pictorial scheme of the changed rank algorithms for SFU dataset [1] based on both median recovery and reproduction angular errors.

We begin this paper by quantifying the scale of this problem. For a given scene and algorithm, we solve for the range of recovery angular errors that can be observed given all colours of light. We define a theory which states that the lowest errors are for red, green and blue lights and the largest for cyans, magentas and yellows.

In the second part of the paper, we propose a new *reproduction angular error* which is defined as the angle between the RGB of a white surface when the ground-truth ($\underline{\rho}^{E,W}$ in Eq. (2)) and estimated illuminations ($\underline{\rho}^{Est}$ in Eq. (2)) are 'divided out' :

$$err_{reproduction} = \cos^{-1}\left(\frac{(\underline{\rho}^{E,W}/\underline{\rho}^{Est}).\underline{U}}{|(\underline{\rho}^{E,W}/\underline{\rho}^{Est})|\sqrt{(3)}}\right), \quad \underline{U} = \frac{\underline{\rho}^{E,W}}{\underline{\rho}^{E,W}} \quad (2)$$

We prove that this reproduction error metric, by construction, gives the same error for the same algorithm-scene pair. The reproduction angular errors for the reproduced images in Figure 1 are shown in the 3rd row of the same figure (the plot with the black bullets). Compared to the recovery angular error, the reproduction error is almost similar for the same scene captured under different colours of illuminants (almost since the process of image formation does not only depend on the color of the illuminant).

For many algorithms and many benchmark datasets we recompute the illuminant estimation performance of a range of algorithms for the new reproduction error and then compare against the algorithm rankings for the old recovery error. We find that using the new measure, the rankings of algorithms remains, while broadly unchanged can change and there can be local switches in rank (see Figure 2). Also the algorithm parameters which can be tuned to provide that best illuminant estimation performance can be chosen differently, depending on whether the reproduction angular error or the recovery angular error is used for evaluation.

[1] Kobus Barnard, Lindsay Martin, Brian Funt, and Adam Coath. A data set for color research. *Color Research & Application*, 27(3):147–151, 2002.

[2] Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

[3] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011. URL `http://colorconstancy.com/`.

[4] David Slater and Glenn Healey. The illumination-invariant recognition of 3d objects using local color invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):206–210, 1996.

[5] Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

Board #38

Texture Similarity Estimation Using Contours

Xinghui Dong
http://www.macs.hw.ac.uk/texturelab
Mike J. Chantler
http://www.macs.hw.ac.uk/~mjc/

The Texture Lab
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, UK

Although performances in the high nineties are typically obtained for tasks such as texture segmentation and classification the same cannot be said of judging texture similarity where a classifier has to estimate the degree to which pairs of textures appear similar to human observers. In an investigation of 51 computational feature sets Dong et *al*. [1] showed that none of these managed to estimate similarity data derived from a population of human observers better than an average agreement rate of 57.76%. Coincidently, none of these computed higher order statistics (HOS) over large regions (≥ 19×19 pixels).

We have discovered few methods that encode long-range, aperiodic characteristics of texture; however, it is well-known that such data are critical to human perception of imagery [2, 3]. For instance, scrambling phase spectra (while leaving the power spectra intact) will often render imagery unintelligible to the human observer [3]. It is also well-known that humans are extremely adept at exploiting the long-range visual interactions evident in contour information [2, 4]. Therefore, we designed an experiment with human observers in order to determine which of three different types of information (2nd-order statistics, local higher order statistics and contour information, see Figure 1) are more important for the perception of texture.

Ten human observers were used in a 2AFC (two-alternative forced choice) scheme with 334 texture images drawn from the *Pertex* database [5]. In each trial the observer was required to compare an original texture image quarter and one variant image quarter ("variant" being one of either contour, power spectrum or randomized block) and decide whether the variant represented the original texture or not (50% of the time they did not). Different quarters of the same texture sample were used in order to prevent observers from performing pixel-wise comparisons. It was found that contour data is more important than local image patches, or 2nd-order global data, to human observers.

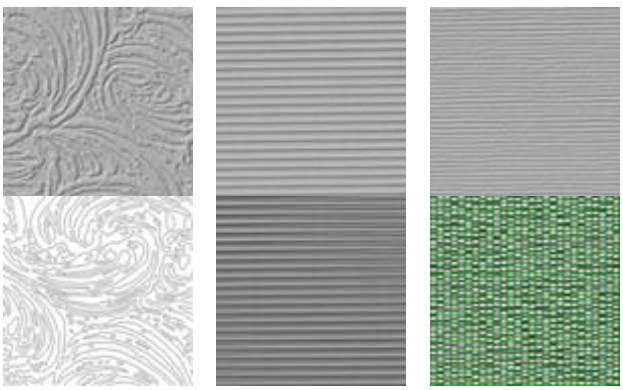
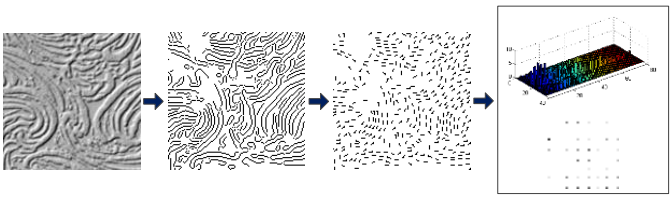


Figure 1: Each of the three columns shows two images derived from the same texture sample (although not the same physical texture area). The upper row shows unprocessed images. The lower row shows, from left to right, the corresponding contour map, power spectrum image and randomized, blocked image.

We therefore developed a contour-based feature set that exploits the long-range HOS encoded in the spatial distribution and orientation of contour segments. A contour is first fragmented into a set of equidistant segments and is then encoded using the spatial distribution and orientation of these segments. Note that images are first processed with the Canny edge detector [6] followed by a morphological erosion operator [7] in order to produce skeleton maps (see Figure 2 (b)). Connected component labelling [7] is performed on skeleton maps. Subsequently, the Moore-neighbour tracing algorithm with Jacob's stopping criteria [7] is applied to each contour and a sequence of points is obtained from each contour. Each contour is then divided into a series of equidistant segments. We represent segments by their mid-point position (on themselves) and chord orientation θ ($\theta \in (0^o, 180^o]$).

We use these data in two ways as outlined in Figure 2. In the first we encode the average shape of the contours in a segment joint orientation/distance histogram (see Figure 2 (d) upper). This provides data on the long-range higher-order visual interactions. In the second we used basic aura matrices [8] (see Figure 2 (d) lower) to encode the spatial distributions and orientations of the all of the segments within a local window without regard to which contour they belong. These data naturally provide relatively short-range (23×23 or less) HOS. The mean of all segment orientation/distance histograms and each basic aura matrix were concatenated into one feature vector which we refer to as "SDoCS" (spatial distribution of contour segments). We test it with two different segment angle quantization schemes (using A bins, $A \in \{18,36\}$), five different segment lengths ($SL \in \{3,5,7,9,11\}$) and one multi-scale case ($SL =$"MS") which concatenates all five feature vectors derived from the five different segment lengths.



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

Figure 2: A representation of the basic information flow: (a) original texture image; (b) skeleton map; (c) segment map. For display purposes, only a part of pixels are shown for each approximate segment; and (d) the joint histogram (upper) and basic aura matrix [8] (lower, only one is shown here).

SDoCS was compared against the 51 feature sets tested by Dong et *al*. [1, 9] and another contour model derived from shape recognition. A pair-of-pairs based evaluation method and a ranking-based evaluation method [1, 9] were applied. The results show that the proposed method outperforms all the other feature sets in the pairs-of-pairs task and all but two feature sets in the ranking task.

We feel that the key point, however, is that we have showed the usefulness of long-range HOS in computing texture similarity and hope that this will inspire other developments of texture features based on such information.

[1] X. Dong and M. J. Chantler. The Importance of Long-Range Interactions to Texture Similarity. Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns, 8047: 425-432, 2013.

[2] D. J. Field, A. Hayes and R. F. Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision Research*. 33: 173-193, 1993.

[3] A. V. Oppenheim and J. S. Lim. The Importance of Phase in Signals. *Proceedings of the IEEE*, 69 (5):529-541, 1991.

[4] L. Spillmann and J. S. Werner. Long-range interactions in visual perception. *Trends in Neurosciences*. 19: 428-434, 1996.

[5] A. D. F. Clarke, F. Halley, A. J. Newell, L. D. Griffin and M. J. Chantler. Perceptual Similarity: A Texture Challenge. *Proceedings of British Machine Vision Conference*, 120.1-120.10, 2011.

[6] J. Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6): 679-698, 1986.

[7] R. C. Gonzalez and R. E. Woods. *Digital Image processing*, Prentice Hall Upper Saddle River. NJ, 2002.

[8] X. Qin and Y. Yang. Basic Grey level aura matrices: theory and its application to texture synthesis. *Proceedings of the Tenth IEEE International Conference on Computer Vision*, 1: 128-135, 2005.

[9] X. Dong, T. Methven, and M. J. Chantler. How Well Do Computational Features Perceptually Rank Textures? A Comparative Evaluation. Proceedings of the ACM 2014 International Conference on Multimedia Retrieval, 281-288, 2014.

Board #39

Generic Object Detection with Dense Neural Patterns and Regionlets

Will Y. Zou[1]
http://ai.stanford.edu/~wzou
Xiaoyu Wang[2]
http://www.xiaoyumu.com
Miao Sun[3]
http://vision.ece.missouri.edu/~miao
Yuanqing Lin[2]
http://www.linyq.com

[1] Stanford University
Stanford, CA, 94305
[2] NEC Laboratories America
Cupertino, CA, 95014
[3] University of Missouri
Columbia, MO, 65201

This paper addresses the challenge of establishing a bridge between deep convolutional neural networks and conventional object detection frameworks for accurate and efficient generic object detection. We introduce Dense Neural Patterns, short for DNPs, which are dense local features derived from discriminatively trained deep convolutional neural networks. DNPs can be easily plugged into conventional detection frameworks in the same way as other dense local features(like HOG or LBP). The effectiveness of the proposed approach is demonstrated with the Regionlets object detection framework. It is the first approach efficiently applying deep convolutional features for conventional object detection models.

Detecting generic objects in high-resolution images is one of the most valuable pattern recognition tasks, useful for large-scale image labeling, scene understanding, action recognition, self-driving vehicles and robotics. At the same time, accurate detection is a highly challenging task due to cluttered backgrounds, occlusions, and perspective changes. Predominant approaches use deformable template matching with hand-designed features. However, these methods are not flexible when dealing with variable aspect ratios. Wang *et al*. recently proposed a radically different approach, named *Regionlets*, for generic object detection [4]. It extends classic cascaded boosting classifiers with a two-layer feature extraction hierarchy , and is dedicatedly designed for region based object detection. Despite the success of these sophisticated detection methods, the features employed in these frameworks are still traditional features based on low-level cues such as histogram of oriented gradients(HOG), local binary patterns(LBP) or covariance [3] built on image gradients.

With the success in large scale image classification [1], object detection using a deep convolutional neural network also shows promising performance [2]. The dramatic improvements from the application of deep neural networks are believed to be attributable to their capability to learn hierarchically more complex features from large data-sets. Despite their excellent performance, the application of deep CNNs has been centered around image classification, which is computationally expensive when transferred to perform object detection. Furthermore, their formulation does not take advantage of venerable and successful object detection frameworks such as DPM or *Regionlets* which are powerful designs for modeling object deformation, sub-categories and multiple aspect ratios.

These observations motivate us to propose an approach to efficiently incorporate a deep neural network into conventional object detection frameworks. To that end, we introduce the *Dense Neural Pattern* (DNP), a local feature densely extracted from an image with an arbitrary resolution using a deep convolutional neural network trained with image classification datasets. The DNPs not only encode high-level features learned from a large image data-set, but are also local and flexible like other dense local features (like HOG or LBP). It is easy to integrate DNPs into the conventional detection frameworks. More specifically, the receptive field location of a neuron in a deep CNN can be back-tracked to exact coordinates in the image. This implies that spatial information of neural activations is preserved. Activations from the same receptive field but different feature maps can be concatenated to form a feature vector for that receptive field. These feature vectors can be extracted from any convolutional layers before the fully connected layers. Because spatial locations of receptive fields are mixed in fully connected layers, neuron activations from fully connected layers do not encode spatial information. The convolutional layers naturally produce multiple feature vectors that are evenly distributed in the evaluated image crop (a 224×224 crop for example). To obtain dense features for the whole image which may be significantly larger than the network input, we resort to "network-convolution" which shifts the crop location and forward-propagate the neural network until

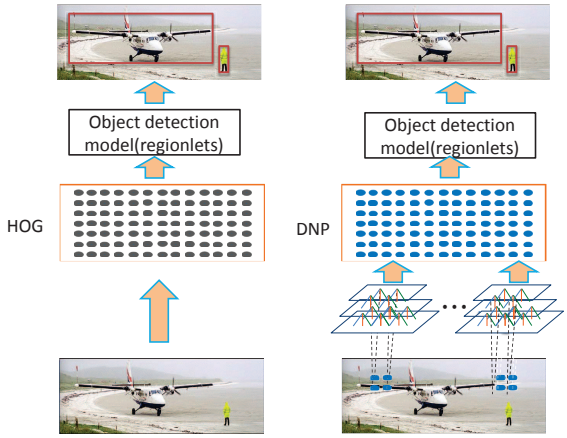


Figure 1: Deep Neural Patterns (DNP) for object detection

features at all desired locations in the image are extracted. As the result, for a typical PASCAL VOC image, we only need to run the neural network several times to produce DNPs for the whole image depending on the required feature stride, promising low computational cost for feature extraction. To adapt our features for the *Regionlets* framework, we build normalized histograms of DNPs inside each sub-region of arbitrary resolution within the detection window and add these histograms to the feature pool for the boosting learning process. DNPs can also be easily combined with traditional features in the *Regionlets* framework.

Our experiments show that the proposed DNPs from the top convolutional layers in deep CNN are very effective and also complementary to traditional features. It achieved 46.1% mean average precision on the PASCAL VOC 2007 dataset, and 44.1% on the PASCAL VOC 2010 dataset, which dramatically improves the original Regionlets approach without DNPs. Combing DNPs and hand-crafted low-level features produces compelling object detection performance. On the contrary, putting together lower layer features and higher layer features from the convolutional neural network does not improve the detection performance. It indicates that these features are correlated. While traditional hand-crafted features are not supervised learned which largely complement the neural network features.

The major contribution of the paper is two-fold: 1) We propose a method to incorporate a discriminatively-trained deep neural network into a generic object detection framework. This approach is very effective and efficient. 2) We apply the proposed method to the *Regionlets* object detection framework and achieved competitive and state-of-the-art performance on the PASCAL VOC datasets.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[2] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[3] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on riemannian manifolds. *T-PAMI*, 2008.

[4] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *ICCV*, 2013.

Top down saliency estimation via superpixel-based discriminative dictionaries

Aysun Kocak
aysunkocak@cs.hacettepe.edu.tr
Kemal Cizmeciler
kemalcizmeci@gmail.com
Aykut Erdem
aykut@cs.hacettepe.edu.tr
Erkut Erdem
erkut@cs.hacettepe.edu.tr

Computer Vision Lab
Department of Computer Engineering
Hacettepe University
Ankara, Turkey

We present a method for learning top-down visual saliency, which is well-suited to locate objects of interest in complex scenes. Our approach is inspired in part by the recent dictionary-based top-down saliency approaches [4, 9] and the new superpixel-based bottom-up salient object detection methods [5, 7, 8]. Specifically, we approach top-down saliency estimation as an image labeling problem in which higher saliency scores are assigned to the image locations corresponding to the target object.

Given a set of training images containing object level annotations, we first segment the images into superpixels. Additionally, we extract objectness maps of these images. For each object category, we then jointly learn a dictionary and a CRF, which leads to a discriminative model that better distinguishes target objects from the background. When given a test image and a search task, we compute sparse codes of superpixels with the corresponding dictionaries learned from data, estimate the objectness map and use the CRF model to infer saliency scores (see Figure 1).

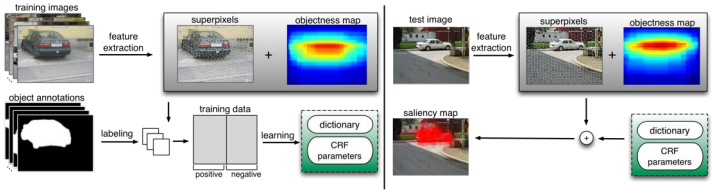


Figure 1: System overview.

Superpixel representation. We segment the images into superpixels and represent them by means of the the first and the second order statistics of simple visual features including color, edge orientation and spatial information. For this step, we employ the sigma points descriptor [3] which provides a compact and effective way of encoding statistical relationships among simple visual features.

CRF and dictionary learning for saliency estimation. We construct a CRF model with nodes $\mathcal{V}$ representing the superpixels and edges $\mathcal{E}$ describing the connections among them. The saliency map is determined by finding the maximum posterior $P(\mathbf{Y}|\mathbf{X})$ of labels $\mathbf{Y} = \{y_i\}_{i=1}^n$ given the set of superpixels $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$:

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \sum_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta) + \sum_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta) + \sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta) - \log Z(\theta,\mathbf{D}) \quad (1)$$

where $y_i \in \{1,-1\}$ denotes the binary label of node $i \in \mathcal{V}$ indicating the presence or absence of the target object, ψ_i are the dictionary potentials, γ_i are the objectness potentials, $\phi_{i,j}$ are the edge potentials, θ are the parameters of the CRF model, and $Z(\theta,\mathbf{D})$ is the partition function. The model parameters $\theta = \{\mathbf{w},\beta,\rho\}$ include the parameter of the dictionary potentials $\mathbf{w}$, the parameter of the objectness potentials β and the parameter of the edge potential ρ. The dictionary $\mathbf{D}$ used in ψ_i encodes the prior knowledge about the target object category.

We test the proposed model under three different settings. In setting 1, we ignore objectness potential and learn discriminative dictionaries and CRF model at superpixel level. In setting 2, we jointly learn dictionary and CRF model by including objectness prior. Setting 3 is extended version of the first one which determines the parameter of the objectness potential β later via cross-validation, while keeping the learned dictionary D and the other CRF parameters fixed.

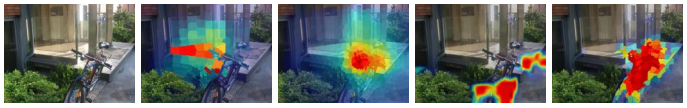
We demonstrate the effectiveness of our approach by comparing it with several bottom-up and top-down models and a generic objectness approach (see Table 1 and 2 for overall results and Figure 2 for a sample comparison). In general, bottom-up models and generic objectness approach do not capture the object of interest due to lack of prior knowledge about the object of interest, and the patch-based top-down saliency models either partly capture the target objects or provide very coarse localizations of the target objects. Our saliency model results in considerably better top-down saliency maps.

	Bike	Car	People
Margolin [5]	25.6	16.9	17.4
Perazzi [7]	11.4	13.8	14.3
Yang and Zhang [8]	14.8	13.7	14.9
Objectness [2]	53.5	48.3	43.5
Aldavert [1]	71.9	64.9	58.6
Khan and Tappen [4]	72.1	-	-
Marszalek and Schmid [6]	61.8	53.8	44.1
Yang and Yang [9]	62.4	60.0	62.0
Our approach (setting 1)	71.9	61.9	65.5
Our approach (setting 2)	71.7	62.0	64.9
Our approach (setting 3)	**73.9**	**68.4**	**68.2**

Table 1: EER results on the Graz-02 dataset.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Yang and Yang [9]	15.2	39.0	9.4	5.7	3.4	22.0	30.5	15.8	5.7	8
Our result	**49.4**	**46.6**	**33.7**	**60.9**	**26.1**	**51.8**	**35.1**	**64.9**	**21.1**	**34.8**
	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv-monitor
Yang and Yang [9]	11.1	12.8	10.9	23.7	**42.0**	2.0	20.2	10.4	24.7	10.5
Our result	**43.7**	**35.1**	**41.4**	**71.4**	32.6	**42**	**42.5**	**13.8**	**63.8**	**27.8**

Table 2: EER results on the PASCAL VOC 2007 dataset.



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

Figure 2: Results for the *look for a bike* task. (a) Input image, and the results of (b) a bottom-up saliency model [8], (c) the objectness map generated by [2], (d) the top-down saliency model of [9] and (e) our approach.

[1] D. Aldavert, A. Ramisa, R.L. de Mantaras, and R. Toledo. Fast and robust object segmentation with the integral linear classifier. In *CVPR*, pages 1046–1053, 2010.

[2] B. Alexe, T. Deselares, and V. Ferrari. What is an object? In *CVPR*, 2010.

[3] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: A small second order statistical region descriptor. In *CVPR*, pages 1802–1809, 2009.

[4] N. Khan and M.F. Tappen. Discriminative dictionary learning with spatial priors. In *ICIP*, pages 166–170, 2013.

[5] R. Margolin, A. Tal, and Zelnik-Manori L. What makes a patch distinct? In *CVPR*, 2009.

[6] M. Marszalek and C. Schmid. Accurate object recognition with shape masks. *Int. J. Comput. Vision*, 97(2):191–209, 2012.

[7] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012.

[8] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[9] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, pages 2296–2303, 2012.

Board #41

Sparse Codes as Alpha Matte

Jubin Johnson
JUBIN001@e.ntu.edu.sg
Deepu Rajan
ASDRAJAN@ntu.edu.sg
Hisham Cholakkal
HISHAM002@ntu.edu.sg

School of Computer Engineering
Nanyang Technological University
Singapore

Matting is a useful tool for image and video editing where foreground objects need to be extracted and pasted onto a different background. A matte is represented by α which defines the opacity of a pixel and is a value in [0, 1], with 0 for background (B) pixels and 1 for foreground (F) pixels. There are three main approaches for image matting: In sampling-based approaches, a foreground-background sample pair is picked from few candidate samples taken from F and B regions by optimizing an objective function. This (F,B) pair is then used to estimate α at a pixel with color I by

$$\alpha_z = \frac{(I-B)(F-B)}{\|(F-B)\|^2}. \tag{1}$$

α-propagation based methods assume correlation between the neighboring pixels under some image statistics and use their affinities to propagate alpha values from known regions to unknown ones. The third category is a combination of the two in which the matting problem is cast as an optimization problem.

The method proposed in this paper is based on sampling. However, there is one important difference between our method and other sampling-based approaches. Matting is cast as a sparse coding problem wherein the sparse codes directly give the estimate of the alpha matte. Hence, there is no need to use the matting equation that restricts the estimate of α from a single pair of foreground and background samples. This allows the matting framework to determine α based on more relevant F and B samples than with only one of each.

A dictionary of color values of F and B pixels is employed to determine the sparse codes for a pixel in an unknown region. The sum of the sparse codes for F pixels directly provides the α. Initially, the pixels in the trimap are classified into high-confidence and low-confidence based on probabilistic segmentation. Since the feature used for coding is color and the complexity of a region for matting is dependent on the overlap of foreground and background colors, we use probabilistic segmentation [2] as a cue to determine the confidence of a pixel as follows:

$$p(I_i) = \frac{p_f(I_i)}{p_f(I_i)+p_b(I_i)}, \tag{2}$$

where $p_f(I_i)$ is the foreground color probability value given by

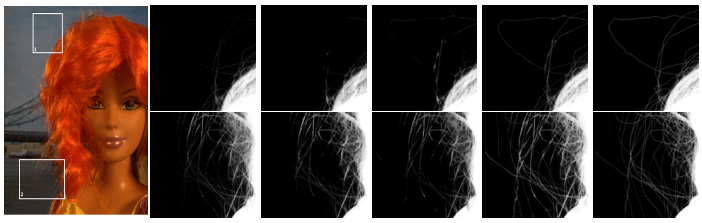
$$p_f(I_i) = exp\left(-\frac{\sum_{k=1}^{m}\|c(I_i)-c(f_k)\|^2}{m\cdot\delta}\right), \tag{3}$$

where $c(\cdot)$ is the RGB color value, m is the number of spatially close foreground samples. A similar formulation exists for background color probability $p_b(I_i)$.

(a) (b) (c)

Figure 1: Alpha matte extracted using our proposed sparse coding method. (a) Input image, (b) Extracted matte and (c) Ground truth.

The size of the dictionary for high-confidence pixels is smaller than that for low-confidence pixels. A universal sample set is generated using a superpixel-based sampling strategy, which is detailed in the paper. For a given unknown pixel of low-confidence, the final dictionary is a larger subset of the universal sample set than that of high-confidence pixels.



(a) (b) (c) (d) (e) (f)

Figure 2: Visual comparison of alpha matte generated by our proposed method using sparse coding with other state-of-the-art methods. Top and bottom rows show zoomed in regions of windows 1 and 2 respectively. (a) Input image, (b) Closed form, (c) Weighted color and texture, (d) Comprehensive sampling, (e) Proposed method and (f) Ground truth.

Given the final dictionary $\mathbf{D}$ for an unknown pixel i, its alpha matte is determined by sparse coding as

$$\begin{aligned} \beta &= argmin\,\|v_i - \mathbf{D}\beta_i\|_2^2 \quad s.t \;\; \|\beta_i\|_1 \le 1\,;\, \beta_i \ge 0, \\ \alpha &= \sum_{p\in F}\beta^{(p)}, \end{aligned} \tag{4}$$

where v_i is the feature vector at i composed of (R,G,B,L,a,b). The sparse codes β_i are generated using a modified version of the Lasso algorithm [3]. The sparse coding procedure is presented with an appropriate set of F and B samples and the sparse coefficients sum up to less than or equal to 1. In order to avoid negative sparse coefficients, the second constraint forces all coefficients to be positive. The sparse codes corresponding to atoms in the dictionary that belong to foreground are added to form the α for the unknown pixel.

The alpha matte obtained by sparse coding is further refined to obtain a smooth matte by considering the correlation between neighboring pixels' matte. We adopt the post-processing approach [4] where a cost function consisting of the data term and a confidence value together with a smoothness term is minimized with respect to α.

Implementation of cost function optimization is described in the paper. The contribution of each part of our proposed method is analyzed with quantitative and qualitative experiments conducted on a benchmark database [1] used universally for image matting evaluation. Our conclusion is that the simplicity of the sparse coding model, coupled with its ability to break away from the $F-B$ pair assumption in matting, makes it a useful tool for future insight into understanding the matting process.

[1] `http://www.alphamatting.com`.

[2] J. Ju, J. Wang, Y. Liu, H. Wang, and Q. Dai. A progressive tri-level segmentation approach for topology-change-aware video matting. In *Computer Graphics Forum*, volume 32, pages 245–253, 2013.

[3] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[4] E. Shahrian, D. Rajan, B. Price, and S. Cohen. Improving image matting using comprehensive sampling sets. In *CVPR*, 2013.

Board #42

Scene-driven Cues for Viewpoint Classification of Elongated Object Classes

José Oramas M.
http://homes.esat.kuleuven.be/~joramas
Tinne Tuytelaars
http://homes.esat.kuleuven.be/~tuytelaa

KU Leuven, ESAT-PSI, iMinds
Leuven, Belgium

Motivation

Object viewpoint classification, also referred to as object pose estimation, is a task of interest for several applications. However, since the early days of computer vision, it has been addressed from a very "local" perspective. This perspective focuses on learning from the features on the object itself, e.g. color, texture, or gradients [1, 2], to identify the different viewpoints in which an object may appear in an image. Lately, this trend has been extended from reasoning about local visual properties of the object in the image space to properties in the 3D scene [3, 4, 5]. Despite the effectiveness of the mentioned methods, they have the weakness of ignoring scene-related cues that can assist the classification process.

Contributions

We complement existing work by exploiting scene-driven cues for object viewpoint classification. The main contributions of this work are:

- We exploit the orientation of the elongation of the object as a cue to estimate its viewpoint. For example, in Fig. 1a. even when we have no direct access to the local features of the object, we are able to predict, up to some level, the orientation of the object (Fig. 1b).
- We enforce scene-consistency in the viewpoint classification process by exploring specific regions of the scene that are more likely to host certain objects with particular features such as class, orientation or size. For example, note how the orientation of the objects in Fig. 1c is closely related to the regions of the scene in which they occur.

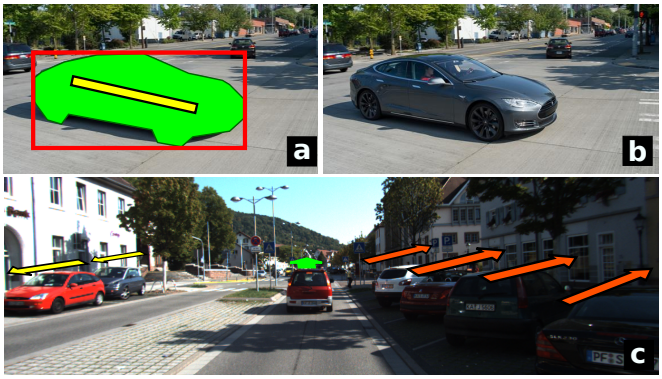


Figure 1: Note how the shape (a,b) and the location (c) of the bounding box of an object is related to its viewpoint.

Proposed method

Our method can be summarized in five steps (Fig. 2): First, we run a viewpoint-aware object detector to collect a set of hypotheses o_i. Then, we generate a set of scene-driven object proposals o'_i, Third, we estimate a correspondence descriptor d_i between each hypothesis o_i and its matching proposal o'_i. Then, we estimate the elongation orientation of the hypothesis o_i via multiclass classification of the descriptor d_i. Finally, the viewpoint of the objects is estimated by the fusion of the responses of the local detector and the elongation orientation classifier.

Figure 3: Viewpoint classification results encoded in jet scale. Continuous line, local detector prediction; Dashed line, scene-driven object proposals. Circle, ground-truth viewpoint (Best viewed in color).

Findings

Experiments on the KITTI object detection dataset show that:

- Considering scene-driven object elongation orientations brings improvements over purely appearance-based viewpoint-aware object detectors on the task of viewpoint classification (see Fig. 3).
- Our results based on 3D object proposals confirms the emerging consensus that coarse 3D scene-level reasoning, apart from context, is specially beneficial for these problems.
- This work complements very recent work, by sending the message that there are relatively simple cues in the scene that can bring improvements for the task of object viewpoint classification.

References

[1] R. J. Lopez-Sastre, T. Tuytelaars, and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *ICCV WS*, 2011.

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

[3] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[4] Y. Xiang and S. Savarese. Object detection by 3d aspectlets and occlusion reasoning. In *3ddr@ICCV*, 2013.

[5] Z. Zia, M. Stark, and K. Schindler. Are cars just 3d boxes? - jointly estimating the 3d shape of multiple objects. In *CVPR*, 2014.

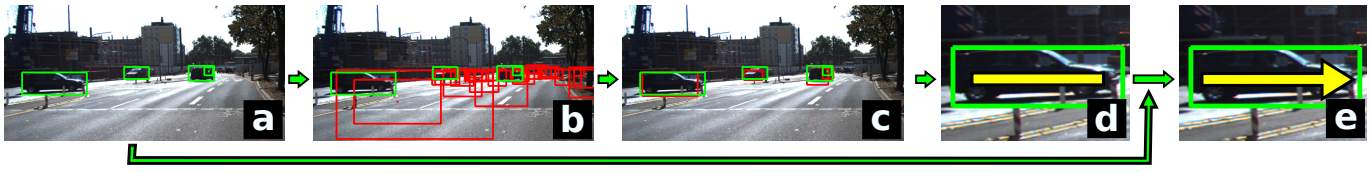


Figure 2: Algorithm Pipeline: a) Object Detection, b) Scene-driven Object Proposal Generation c) Object-hypotheses - Object-Proposal Matching, d) Elongation Classification, and e) Viewpoint Classification.

Board #43

Multi-View Depth Map Estimation With Cross-View Consistency

Jian Wei
jian.wei@graphics.uni-tuebingen.de
Benjamin Resch
benjamin.resch@uni-tuebingen.de
Hendrik P. A. Lensch
hendrik.lensch@uni-tuebingen.de

Computer Graphics
Tübingen University
72076 Tübingen
Germany

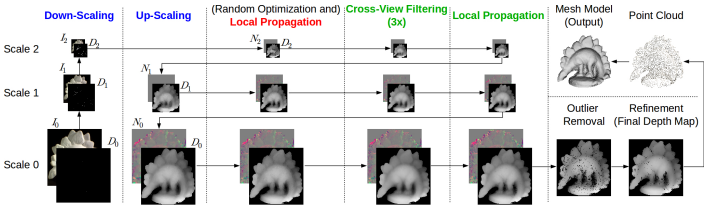


Figure 1: Our processing pipeline for one view of Dino dataset. Our key steps include: hierarchical framework (blue), local propagation (red), and cross-view filtering with an additional propagation pass (green).

Motivation. Multi View Stereo (MVS) aims to establish 3D models from multiple calibrated images. Some works use region growing to estimate depth map per view, and then merge the results. They either only deal with reliable regions, or have difficulty in parallelizing. More crucially, due to the view-independent estimation, inconsistent outliers may exist and grow during propagation, producing unstable estimates across views. This leads to a large amount of estimates removed in the merging stage after consistency checking, and diminishes the reconstruction quality.

To increase robustness of depth-map-based MVS methods, we combine several techniques: Depth estimates are propagated in parallel in the local neighborhood to efficiently spread reliable depth information into regions without prominent structures. A faster coarse-to-fine strategy fills in larger holes. Most importantly, a novel cross-view filtering stage based on free-space constraints and variance filtering, enforces consistency among the depth maps of different views. Our algorithm alternates between correlation and consistency optimization. This way, noisy patches and spikes are excluded so that the subsequent depth map fusion becomes easier.

Workflow. Figure 1 shows our workflow. I_k, D_k, and N_k are the image, depth map, and normal map of a reference view at scale k. I_0 is the input image. Each view selects at most 6 secondary images. Before the first propagation step at each scale, randomly shifted depths and random normals are assigned if smaller matching errors are obtained.

Initialization. For a pixel p, we initialize its depth $D_0(p)$ from bundler if p is feature point; otherwise $D_0(p) = 0$. Its normal $N_k(p)$ including the gradients of the tangent plane in x and y directions, is initialized fronto-parallel at the coarsest scale, *i.e.* $N_2(p) = \{0,0\}$. Before the estimation at each scale, E_k is initialized using the existing depth and normal estimates.

Local Propagation (LP). Good depth and normal estimates are dispersed into the neighborhoods by traversing all pixels if the propagated value improves the correlation measure. The depth hypothesis considers the normal of the tilted patch. Pixels are traversed along parallel scanlines on GPU. We shorten the traversal distance of the work [2] such that more GPU threads can be assigned. In every other iteration vertical and horizontal propagations are applied alternately.

Hierarchical Framework (HF). For textureless regions with few initializations, one propagation alone at the original scale is insufficient due to the locality of short scanlines. We down-scale the depth map and spread the sparse data into neighborhoods. This way, one propagation at the coarsest scale can fill most of the holes. Then the estimates are used for the consecutive finer scale by up-scaling. The overall time is also reduced since the scaling is negligible compared with the speed-up of propagation. We also down-scale the images and up-scale the normal maps.

Cross-View Filtering (CVF). Inspired by the temporally consistent optical flow estimation [3], after local propagation of all views, we perform a cross-view filtering for each reference view to improve the depth consistency. Then a second propagation spreads the optimized estimates.

The projection relationships of pixels between views are considered using the depth information. For each depth value, we find the corresponding pixels in the secondary views, and project them back into the reference view obtaining new depth candidates. These candidates are weighted by the depth difference between the reference and secondary views to get an optimized depth. In some cases, this depth projection from secondary views can even fill holes in the reference, spawning further, more consistent propagation. To avoid slight shifting for some inliers which were accurate before, we additionally check three randomly shifted depth values around the new depth.

Step		Bailer *et al.* [2]	Only LP	LP+HF	LP+CVF	LP+HF+CVF
Downscaling				8.4s		8.4s
1st	Propagation	174.3s	142.2s	13.6s	142.0s	13.6s
	Cross-View Filtering				151.2s	10.8s
	Propagation				226.9s	16.0s
	Upscaling			0.4s		0.4s
2nd	Propagation	1126.6s	880.3s	224.5s	1000.7s	250.0s
	Cross-View Filtering				193.3s	49.2s
	Propagation				951.9s	228.7s
	Upscaling			2.1s		2.1s
3rd	Propagation	418.0s	410.2s	417.4s	450.6s	458.2s
	Cross-View Filtering				204.1s	214.0s
	Propagation				279.9s	280.6s
Outlier removal		42.2s	41.2s	44.2s	48.1s	51.1s
Refinement			121.7s	144.1s	151.3s	189.5s
Overall		1984.4s	1866.8s	1079.1s	4020.3s	1844.1s

Table 1: Timings of each step using Bailer *et al.* [2] and different combinations of our processing steps, when reconstructing all views of Fountain-P11.

Measurement	Bailer *et al.* [2]	Only LP	LP+HF	LP+CVF	LP+HF+CVF	LP+HF+CVF[1]	LP+HF+CVF[2]
Mean Rel. Error ($\times 10^{-3}$) ↓	1.663	1.414	1.236	2.407	1.732	1.505	2.062
Completeness (%) ↑	64.0	63.9	66.9	74.6	79.6	75.9	80.5
Mean Consistency ↑	9.083	9.019	9.124	9.611	9.556	9.253	10.090
Mean Variance ($\times 10^{-6}$) ↓	1.790	1.722	1.626	1.602	1.092	1.179	1.099
Mean Rel. Error of LP+HF+CVF on Pixels of Other Methods ($\times 10^{-3}$) ↓	1.102	1.068	1.142	1.292		1.319	1.368

Table 2: Statistical comparisons for the center view of Fountain-P11 after outlier removal. LP+HF+CVF[1] uses cross-view filtering only for post-processing, and LP+HF+CVF[2] uses propagation-filtering at each scale without the second propagation. The arrows indicate preferred directions.

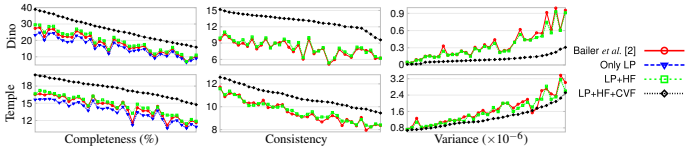


Figure 2: Completeness, mean consistency rating, and mean variance comparisons for some views of Dino and Temple datasets.

Bailer *et al.* [2] (no depth map refinement) | LP+HF (no depth map refinement) | LP+HF+CVF (no depth map refinement) | LP+HF+CVF (with depth map refinement)

Figure 3: Depth maps and 3D models of a region in Sofa dataset after outlier removal and our final results with depth map refinement.

Outlier Removal and Refinement. Inconsistent outliers are filtered out from the resulting depth maps. Results are finally refined by filling the holes and then filtering the noise.

Results. Some results are presented in Tables 1 and 2, as well as Figs. 2 and 3. The relative error evaluates depth accuracy between the estimates and ground truth. The completeness relates the number of recovered pixels to the image size. The consistency [2] and variance (see the paper) measure the multi-view coherence. Combining improved propagation, hierarchical estimation, and iterative multi-view consistency optimization, our method increases the estimation speed, generates dense depth maps with desirable global consistency, and yields convincing 3D reconstruction results. The benchmark results of our full pipeline using the Middlebury evaluation website [1] demonstrate that, our work is competitive with other methods and placed among the most efficient approaches.

[1] Multi-view stereo evaluation. `http://vision.middlebury.edu/mview/`.

[2] C. Bailer, M. Finckh, and H.P.A. Lensch. Scale robust multi view stereo. In *Proc. ECCV*, 2012.

[3] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 31(4), 2012.

Board #44

Improving Detection of Deformable Objects in Volumetric Data

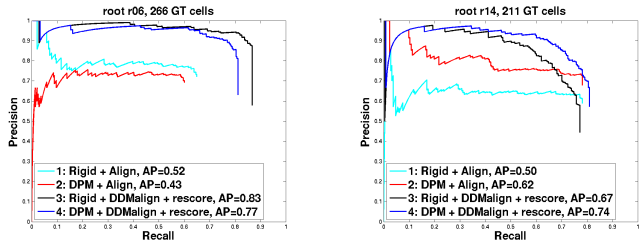
Dominic Mai [1,3]
maid@informatik.uni-freiburg.de
Jasmin Dürr [2]
jasmin.duerr@biologie.uni-freiburg.de
Klaus Palme [2,3]
klaus.palme@biologie.uni-freiburg.de
Olaf Ronneberger [1,3]
ronneber@informatik.uni-freiburg.de

[1] Computer Science Department
University of Freiburg
Germany
[2] Institute of Biology II - Botany
University of Freiburg
Germany
[3] BIOSS Centre for Biological Signalling Studies
University of Freiburg
Germany

Overview. We investigate class level object detection of deformable objects. To this end, we aim for cell detection in volumetric images of dense plant tissue (*Arabidopsis Thaliana*), obtained from a confocal laser scanning microscope. In 3D volumetric data, the detection model does not have to deal with scale, occlusion and viewpoint dependent changes of the appearance, however, our application needs high recall and precision. We implement Felsenszwalb's Deformable Part Model for volumetric data. Corresponding locations for part training are obtained via elastic registration. We identify limitations of its star shaped deformation model and show that a pairwise connected detection model can outperform the DPM in this setting.

Contribution. We combine the ideas of discriminative detection and elastic registration by using a discriminative similarity measure with a pairwise deformation model. To this end, we show that deformable detection approaches can be formulated in a general elastic registration framework. We propose the *Discriminative Deformable Model* (Fig. 2(d,h)): A set of pairwisely connected patch detectors. Each patch detector is realized as a *linear Support Vector Machine*. The optimization of the model is cast as a discrete labeling problem (*Markov Random Field*) and efficiently solved with iterated graph cuts (*FastPD*). The patch detectors are trained jointly and yield the *unary costs*, while the relative motion of neighboring patches gives the *binary costs* of the model: Only connected patches that move inconsistently have to pay displacement penalties.

Results. We show that we can improve the detection of deformable objects in volumetric image data substantially by using the more meaningful scores from the *Discriminative Deformable Model*. We obtain the fine grained localization of the elastic deformation model combined with the expressive scores that stem from the discriminative data term. The strategies based on the *DDM* based alignment with rescoring outperform the rigid and DPM based detection approaches by a margin of 0.23 percentage points with a mean average precision of 0.75. The average intersection over union of the valid detections with the ground truth data is 0.69 (Precision Recall Graphs in Fig. 1).



(a) (b)

Figure 1: Precision-recall Graphs of the different detection strategies for the two roots (a) r06 and (b) r14. The alignment and rescoring with the proposed *Discriminative Deformable Model* (DDMalign, **black** curve and **blue** curve) produces the best results, independent of the underlying detector.

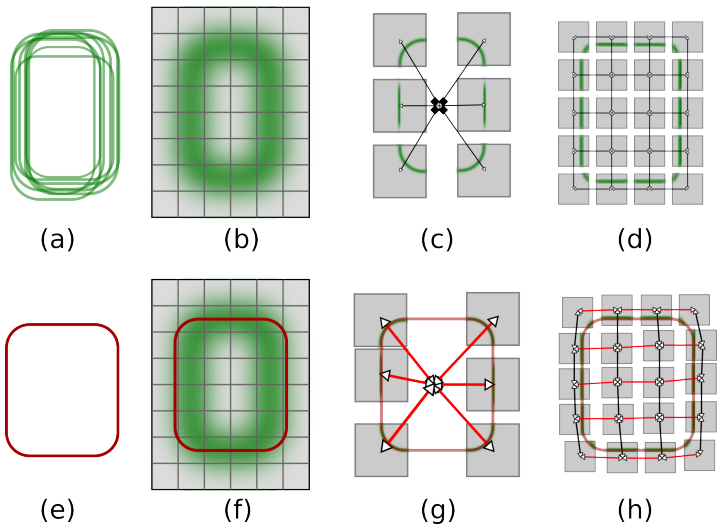


Figure 2: Illustration of different detection approaches and how they deal with deformation. (a) Overlay of the rigidly aligned positive training examples. (b) A rigid detection model allows for small local deformations due to the (soft-) binning of the gradients in the HOG cells. (c) The star shaped structure of the DPM allows parts to move independently. (d) *Proposed model*: The parts are connected pairwisely. (e) A Detection sample that is wider than most of the training examples. (f) The rigid filter barely detects the object. (g) Every part filter of the DPM has to pay a displacement penalty. (h) The parts of the proposed model only get penalties for horizontal displacements.

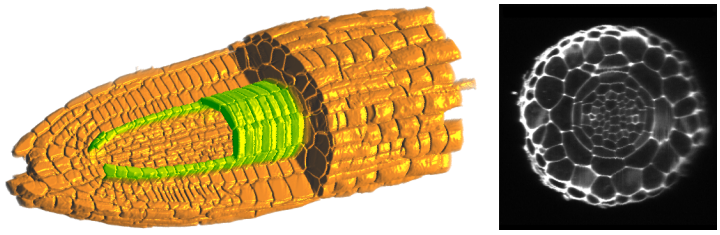


Figure 3: We work with 3D volumetric data of *Arabidopsis Thaliana* that was recorded with a confocal laser scanning microscope. Our goal is to detect and segment single cells of a specific layer. **(left)** A Volume rendering of the root *r06*, the layer used for training and detection is colored green. **(right)** A slice of the original raw data.

Board #45

Reasoning about Photo Collections using Models of Outdoor Illumination

Daniel Hauagge
hauagge@cs.cornell.edu
Scott Wehrwein
swehrwein@cs.cornell.edu
Paul Upchurch
paulup@cs.cornell.edu
Kavita Bala
kb@cs.cornell.edu
Noah Snavely
snavely@cs.cornell.edu

Cornell University
Ithaca, NY, USA

Natural illumination from the sun and sky plays a significant role in the appearance of outdoor scenes. We propose the use of sophisticated outdoor illumination models, developed in the computer graphics community, for estimating appearance and timestamps from a large set of uncalibrated images of an outdoor scene. We first present an analysis of the relationship between these illumination models and the geolocation, time, surface orientation, and local visibility at a scene point. We then use this relationship to devise a data-driven method for estimating per-point albedo and local visibility information from a set of Internet photos taken under varying, unknown illuminations. Our approach significantly extends prior work on appearance estimation to work with sun-sky models, and enables new applications, such as computing timestamps for individual photos using shading information.

1 Modeling illumination in outdoor scenes

The illumination arriving at a point in an outdoor scene depends on several key factors, including:

- geographic location
- time and date
- surface orientation
- local visibility

Our model describes the irradiance incident at an outdoor scene point on a clear day as a function $L(\phi,\lambda,t,\alpha,\vec{n})$ where ϕ,λ are latitude and longitude, t is the time and date, $\vec{n}$ is the normal, and α is the *local visibility angle*. This angle α is a parameterization of local visibility based on a model of ambient occlusion proposed by Hauagge et al. [1], which models local geometry around a point as a cylindrical hole with angle α from the normal to the opening. Figure 1 shows examples of L, in the form of spheres rendered under predicted outdoor illumination at various times and α angles, at a given location on Earth.

2 Method

A georegistered 3D point cloud built using SfM and MVS provides geographic location (ϕ,λ), surface normals $(\vec{n})$, and a set of observed pixel values for each point (I_x). We first estimate the albedo of each point, then use the albedo to estimate lighting and capture time for each photo.

Estimating Albedo. We adopt a simple Lambertian image formation model $I_x = \rho L_x$ where I_x is the observed color of a point x in a given image I, ρ_x is the (assumed constant) albedo at that point, and L_x is the irradiance as defined above. Given many observations of a point I_x, we derive the albedo ρ_x by dividing the average observed color $\mathcal{E}[I_x]$ by an estimate of the average illumination $\mathcal{E}[L_x]$.

Our key insight is that we can use a sun/sky model to predict illumination for a given condition, or indeed the *average* illumination for a given scene. For a given location, time, and visibility angle, we compute a physically-based environment map (we use the model of Hosek and Wilkie [2]) and, for each normal, integrate over the visible portion of the environment map to produce a database of spheres giving values for L at each normal direction, as illustrated in Figure 1(a-b). We then estimate expected illumination $\bar{L}(\vec{n},\alpha)$ as a function of normal and visibility angle by taking the average over a set of times sampled throughout the year.

For each point x, we have a surface normal estimate $\vec{n}_x$ from the 3D reconstruction; however we also need the visibility angle α_x to look up the appropriate expected illumination $\bar{L}(\vec{n}_x,\alpha_x)$. Under a simpler lighting model, Hauagge et al.showed that α can be determined analytically as a function of an albedo-invariant image statistic $\kappa_x = \mathcal{E}[I_x]^2/\mathcal{E}[I_x^2]$. Under our more complex illumination model, we instead relate κ to α numerically by computing $\kappa(\vec{n},\alpha)$ over the predicted illumination values provided by the sun/sky model, as shown in Figure 1(c). We let α_x be the alpha for which $\kappa(\vec{n}_x,\alpha)$ most closely matches the observed κ_x.

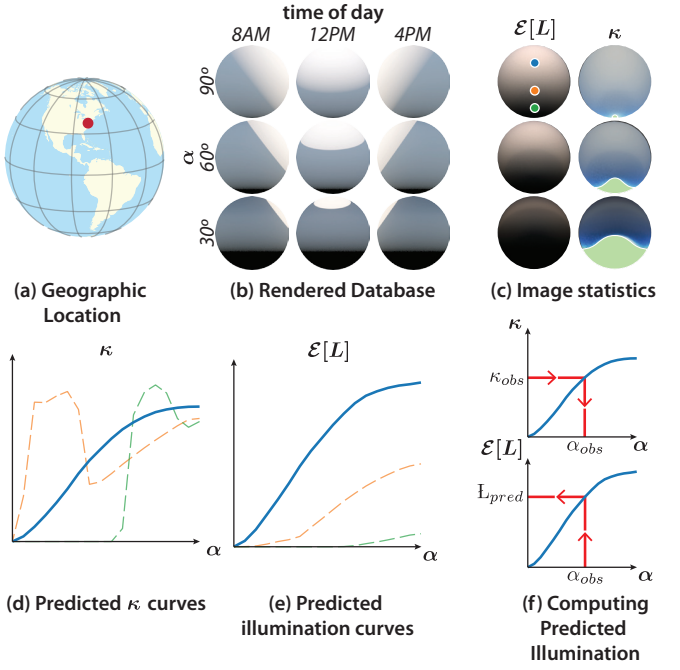


Figure 1: For a given geographic location (a), we render a database of spheres (b) covering all possible times (over a full year) and visibility angles. (c) We compute $\mathcal{E}[L]$ and κ for each α. Green regions correspond to combinations of normal direction and crevice for which we cannot reliably recover albedo.

Estimating Time of Day. With albedos in hand, we can estimate illumination for an image by dividing each visible point's observed color value by the estimated albedo $L_x = \frac{I_x}{\rho_x}$. To estimate the time for that image, we can compare this estimated per-point illumination to the illumination predicted by the sun/sky model at a set of times candidate times t (potentially sampled over the entire year). The predicted time t^* is then the time for which the observed and predicted illumination are most similar.

Results. Our technique recovers the albedo of outdoor scenes more accurately than Hauagge et al. [1] and successfully identifies and discards points whose albedo cannot be recovered. The timestamp estimates using our albedo have median error under one hour (about 15 degrees of sun position) on our test datasets, which significantly outperforms random chance and a state-of-the-art single image method. Please see the full paper for more details and complete results.

[1] Daniel Hauagge, Scott Wehrwein, Kavita Bala, and Noah Snavely. Photometric ambient occlusion. *CVPR*, 2013.

[2] Lukas Hosek and Alexander Wilkie. An analytic model for full spectral sky-dome radiance. *ACM Transactions on Graphics*, 2012.

Board #46

Online quality assessment of human movement from skeleton data

Adeline Paiement
csatmp@bristol.ac.uk
Lili Tao
lili.tao@bristol.ac.uk
Sion Hannuna
sh1670@bristol.ac.uk
Massimo Camplani
massimo.camplani@bristol.ac.uk
Dima Damen
dima.damen@bristol.ac.uk
Majid Mirmehdi
majid@cs.bris.ac.uk

Visual Information Laboratory
Department of Computer Science
University of Bristol
Bristol, UK

This work was performed under the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

This work addresses the challenge of analysing the quality of human movements from visual information which has use in a broad range of applications, from diagnosis and rehabilitation to movement optimisation in sports science. Traditionally, such assessment is performed as a binary classification between normal and abnormal by comparison against normal and abnormal movement models, e.g. [5]. Since a single model of abnormal movement cannot encompass the variety of abnormalities, another class of methods only compares against one model of normal movement, e.g. [4]. We adopt this latter strategy and propose a continuous assessment of movement quality, rather than a binary classification, by quantifying the deviation from a normal model. In addition, while most methods can only analyse a movement after its completion e.g. [6], this assessment is performed on a frame-by-frame basis in order to allow fast system response in case of an emergency, such as a fall.

Methods such as [4, 6] are specific to one type of movement, mostly due to the features used. In this work, we aim to represent a large variety of movements by exploiting full body information. We use a depth camera and a skeleton tracker [3] to obtain the position of the main joints of the body, as seen in Fig. 1. We normalise this skeleton for global position and orientation of the camera, and for the varying height of the subjects, e.g. using Procrustes analysis.

The normalised skeletons have high dimensionality and tend to contain outliers. Thus, the dimensionality is reduced using Diffusion Maps [1] which is modified by including the extension that Gerber et al. [2] presented to deal with outliers in Laplacian Eigenmaps. The resulting high level feature vector $\mathbf{Y}$, obtained from the normalised skeleton at one frame, represents an individual pose and is used to build a statistical model of normal movement.

Our statistical model is made up of two components that describe the normal poses and the normal dynamics of the movement. The pose model is in the form of the probability density function (pdf) $f_Y(y)$ of a random variable Y that takes as value $y = \mathbf{Y}$ our pose feature vector $\mathbf{Y}$. The pdf is learnt from all the frames of training sequences that contain normal instances of the movement, using a Parzen window estimator. The quality of a new pose y_t at frame t is then assessed as the log-likelihood of being described by the pose model, i.e.

$$llh_{pose} = \log f_Y(y_t) \ . \tag{1}$$

The dynamics model is represented as the pdf $f_{Y_t}(y_t|y_1,\ldots,y_{t-1})$ which describes the likelihood of a pose y_t at a new frame t given the poses at the previous frames. In order to compute it, we introduce X_t with value $x_t \in [0,1]$, which is the stage of the (periodic or non-periodic) movement at frame t. Note, in the case of periodic movements, this movement stage can also be seen as the phase of the movement's cycle. Based on Markovian assumptions, we find that

$$f_{Y_t}(y_t|y_1,\ldots,y_{t-1}) \approx f_{Y_t}(y_t|\hat{x}_t)\, f_{X_t}(\hat{x}_t|\hat{x}_{t-1}) \ , \tag{2}$$

with $\hat{x}_t$ an approximation of x_t that minimises $f_{\{X_0,\ldots,X_t\}}(x_0,\ldots,x_t|y_1,\ldots,y_t)$. $f_{Y_t}(y_t|x_t)$ is learnt from training sequences using Parzen window estimation, while $f_{X_t}(x_t|x_{t-1})$ is set analytically so that x_t evolves steadily during a movement. The dynamics quality is then assessed as the log-likelihood of the model describing a sequence of poses within a window of size ω:

$$llh_{seq} \approx \frac{1}{\omega} \sum_{i=t-\omega+1}^{t} \log\left(f_{Y_i}(y_i|x_i)\, f_{X_i}(x_i|x_{i-1})\right) \ . \tag{3}$$

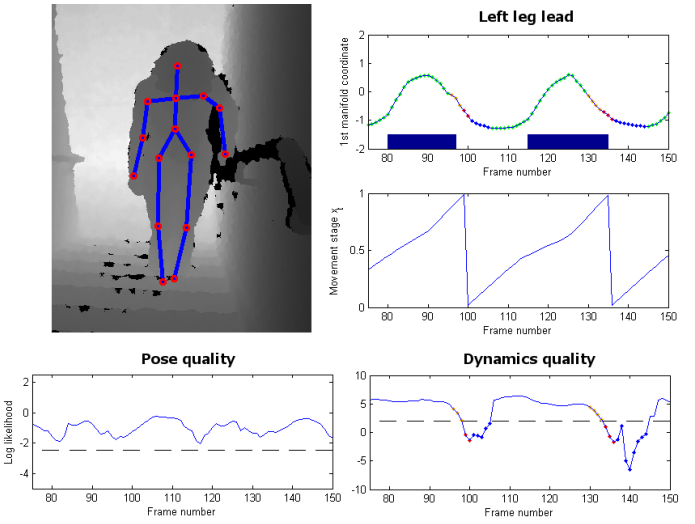


Figure 1: Analysis of gait on stairs. Clockwise from top left: raw skeleton data, high level description of the gait, dynamics and pose quality measures.

In our experiments, these two quality measures provided a continuous quality assessment of gait on stairs on a frame-by-frame basis, as illustrated at the bottom of Fig. 1. Two thresholds, set empirically, allowed deciding when gait becomes abnormal. More details and experiments can be found in the article and on our project's webpage[1].

[1] R. R. Coifman and S. Lafon. Diffusion maps. Applied and computational harmonic analysis, 21(1):5–30, 2006.

[2] S. Gerber, T. Tasdizen, and R. Whitaker. Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps. In Proc. of Int. Conf. on Machine Learning, pages 281–288. ACM, 2007.

[3] OpenNI User Guide. OpenNI organization, November 2010. URL http://www.openni.org/documentation.

[4] J. Snoek, J. Hoey, L. Stewart, R. S. Zemel, and A. Mihailidis. Automated detection of unusual events on stairs. Image and Vision Computing, 27(1):153–166, 2009.

[5] M. Z. Uddin, J. T. Kim, and T. S. Kim. Depth video-based gait recognition for smart home using local directional pattern features and hidden Markov model. Indoor and Built Environment, 23(1):133–140, 2014.

[6] R. Wang, G. Medioni, C. J. Winstein, and C. Blanco. Home monitoring musculo-skeletal disorders with a single 3D sensor. In CVPR Workshops, pages 521–528. IEEE, 2013.

[1] www.irc-sphere.ac.uk/work-package-2/movement-quality

Board #47

Depth Sweep Regression Forests for Estimating 3D Human Pose from Images

Ilya Kostrikov
ilya.kostrikov@rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Juergen Gall
gall@iai.uni-bonn.de
University of Bonn
Bonn, Germany

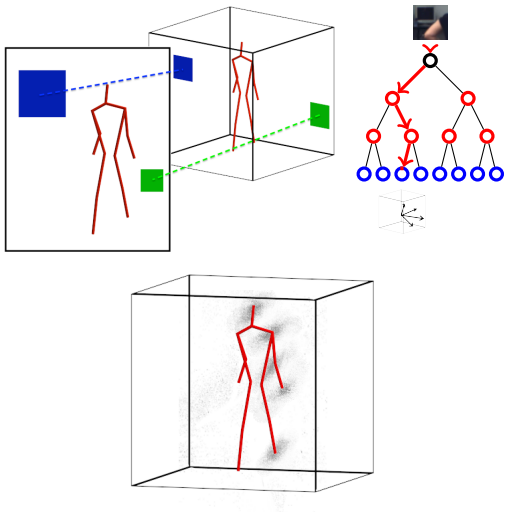


Figure 1: Illustration of a depth sweep regression forest for 3D pose estimation from a 2D image. **Top left.** Patches sampled from different depths project onto the image with different scale. **Top right.** The projected patches traverse the tree evaluating splitting functions in the intermediate nodes (black and red) until they reach a leaf node (blue). A leaf node contains 3D offsets that point to locations of a joint with associated weights. **Bottom.** Based on the offsets, the patches sampled in the 3D volume cast 3D votes for several joint locations.

Over decades estimating the human pose from still images has been an intensive research topic. In recent years, the majority of works has been focused on estimating the 2D pose, since this is already very challenging. However, many applications require the 3D pose. While some approaches estimate first the 2D pose and then reconstruct the 3D pose from the 2D pose estimate, estimating the 3D pose directly from the images is more practical since it directly solves the problem at hand. For this task, discriminative approaches that learn a mapping from image features to 3D pose, have been most successful. This is in contrast to state-of-the-art human pose estimation approaches that rely on discriminative parts and combine them within a pictorial structure model [2] that represents the human skeleton. A prominent example of these approaches is [4].

In this paper we address the problem of estimating the 3D pose from still images. However, instead of learning a regression from image features to the full pose, we regress the positions of the joints in 3D space and then infer the pose using a 3D pictorial structure framework. For regression, we rely on regression forests that have been shown to efficiently predict 2D pose from images [1]. These approaches, however, cannot be directly applied since each local image or depth feature estimates the relative positions of the joints from the feature location. While the relative position is well defined if feature and joint locations are given either in 2D or in 3D, it is not defined if the features are sampled from 2D images without depth information and the joint locations need to be predicted in a 3D world coordinate system.

Our approach consists of two parts: first, we independently estimate joint 3D location probabilities; second, we use the estimated probabilities together with the pictorial structure framework in order to infer the full skeleton. For the first part, we propose depth sweep regression forests which are regression forests that hypothesize the missing depth information of image features. For the second part, we extend the mixture of PSMs [3] for 3D inference.

In the context of pose estimation [1], a regression tree represents a mapping from the space of image patches and patch locations $\mathcal{P}\times\Omega$ to the space of probabilities over joint locations $\mathcal{X}$. In case of 2D pose estimation, we have $\Omega\subset\mathbb{R}^2$, $\mathcal{X}\subset\mathbb{R}^2$ and $\mathbf{d}(\mathbf{x},\mathbf{y})=\mathbf{x}-\mathbf{y}$. For localizing a joint j, the probabilities of all trees of a forest are averaged and summed over all patches sampled from locations $\mathbf{y}\in\Omega$:

$$\phi_j(\mathbf{x})=\sum_{\mathbf{y}\in\Omega}\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}}p(j|L_T(P,\mathbf{y}))p_j(\mathbf{d}(\mathbf{x},\mathbf{y})|L_T(P,\mathbf{y})). \quad (1)$$

where $p(j|L)$ denotes the class probability of joint j stored at leaf L and $p_j(\mathbf{d}|L)$ denotes the probability of relative locations of the joint j.

In order to predict 3D joint locations from 2D images, the approach briefly described above cannot be directly applied since $\Omega\subset\mathbb{R}^2$ and $\mathcal{X}\subset\mathbb{R}^3$. The relative location $\mathbf{d}$ of a 3D joint given the 2D location of a patch, and thus (1), are not defined. We therefore propose to perform the inference in $\Omega'\subset\mathbb{R}^3$ instead:

$$\phi_j^{ds}(\mathbf{x})=\sum_{\mathbf{y}'\in\Omega'}\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}}p\left(j|L_T(P,\mathbf{y}')\right)p_j\left(\mathbf{d}(\mathbf{x},\mathbf{y}')|L_T(P,\mathbf{y}')\right). \quad (2)$$

In this formulation $\mathbf{d}(\mathbf{x},\mathbf{y}')=\mathbf{x}-\mathbf{y}'$ is well defined, but the regression trees have to learn a mapping from $\mathcal{P}\times\Omega'$ to $\mathcal{X}$. This causes a problem since $\mathcal{P}\times\Omega'$ is not observed neither for training nor for testing. However, assuming that the camera projection π is known, which maps a point from Ω' to the image plane Ω, we can rephrase the problem as learning a mapping from $\mathcal{P}\times\Omega\times\mathcal{Z}$ to $\mathcal{X}$, where the appearance of a 2D patch P depends on the 2D image location and the depth z. Since we do not observe depth for training or testing, we hypothesize it by sweeping with a plane parallel to the image plane along the z-axis through a 3D volume. The patch P corresponding to the 3D point y' is then the patch centered at the projection $\pi(\mathbf{y}')\in\Omega$ and the leaf it ends depends on $z'\in\mathcal{Z}$:

$$\phi_j^{ds}(\mathbf{x})=\sum_{\mathbf{y}'\in\Omega'}\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}}p\left(j|L_T(P,\pi(\mathbf{y}'),z')\right)p_j\left(\mathbf{d}(\mathbf{x},\mathbf{y}')|L_T(P,\pi(\mathbf{y}'),z')\right). \quad (3)$$

Since the appearance of patches changes for different depth values, the maximum of (3) corresponds to a set of patches that are associated to the correct hypothesized depth values and agree on the 3D joint location.

Inferring 3D joint locations independently from 2D RGB images is prone to depth ambiguities. Many of the ambiguities, however, can be resolved by using a kinematic body model that provides information about constraints between joint locations. To this end, we use the well known pictorial structure framework [2] that provides accurate results while keeping the inference tractable:

$$p(X_J|I,\vartheta)\propto\prod_{j\in J}\left(\phi_j^{ds}(\mathbf{x}_j)\right)^{\alpha}\prod_{(i,j)\in E}\psi_{ij}(\mathbf{x}_i,\mathbf{x}_j|\vartheta_{ij}), \quad (4)$$

As proposed in [3], we use a mixture of PS models to overcome the limitations of a single tree model. Given a set of training poses M, we cluster the relative poses by k-means and estimate the parameters of a PS model for each cluster. Inference is first performed for each PS model independently and the solution of the model with highest confidence is taken. We weight the confidence of each model by the prior probability of the model.

We compare our approach with other methods on HumanEva I and Human3.6m where our approach achieves state-of-the-art performance.

[1] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[2] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 2005.

[3] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[4] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.

Board #48

AAAMS : Anisotropic Agglomerative Adaptive Mean-Shift

Rahul Sawhney[1]
rahul.sawhney@gatech.edu
Henrik I. Christensen[1]
hic@gatech.edu
Gary R. Bradski[2]
gbradski@magicleap.com

[1] Institute of Robotics & Intelligent Machines, Georgia Tech
[2] Magic Leap Inc.

Mean Shift today, is widely used for mode detection and clustering. The technique though, is challenged in practice due to assumptions of isotropicity and homoscedasticity. Isotropic/scalar bandwidths tend to smooth anisotropic patterns and affect partition boundaries, while homoscedastic / global bandwidths are inappropriate when clusters (or modes) at different scales need to be identified.

We present an adaptive Mean Shift methodology that allows for anisotropic clustering, through unsupervised local bandwidth selection. The bandwidth matrices evolve naturally, adapting locally through agglomeration, and in turn guiding further agglomeration. The online methodology is practical for low-dimensional feature spaces, preserving better detail and clustering salience. Additionally, conventional Mean Shift either critically depends on a per instance choice of bandwidth, or relies on offline methods which are inflexible and/or again data instance specific. The presented approach, due to its adaptive design, also alleviates this issue - with a default form performing generally well. The methodology though, allows for effective tuning of results.

In the proposed approach, clusters arise on the fly, as a consequence of agglomeration of extant clusters. Local bandwidths which evolve anisotropically every iteration, are associated with each cluster; by design, all members of a cluster converge to the same local mode. By evolving as function of a cluster's aggregated trajectory points, these bandwidths are able to adapt to the underlying mode structure (shape, scale, orientation) - and in turn, guide future cluster trajectory and agglomeration. This results in robust mode detection and with increased partition saliency (*Figs. 1, 2(a)*). The supplementary presents a convergence proof when anisotropic bandwidths vary between Mean shift iterations, as is the case here.

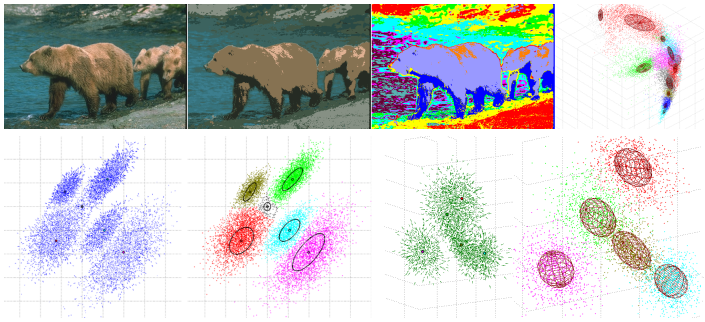


Figure 1: Single domain clustering examples over color data (top row with segment and label images are shown in middle, 11 clusters), and simulated gaussian mixtures (second row) in 2D & 3D respectively. $1-sigma$ final trajectory set bandwidths have been overlaid at converged mode positions. Post processing was disabled.

The approach involves running Mean Shift fixed point iterations at cluster levels, over a single data point per cluster. Starting out as trivial clusters (solitary data points), the clusters agglomerate between iterations. By algorithm design, clusters are merged only when they are tending towards the same mode. All member points of a cluster, u, which will eventually converge to a common local mode, share a common bandwidth, Σ_u - referred to as the local bandwidth. This bandwidth evolves every iteration, adapting to the structure of the local mode and to an extent, its basin.

The standard MS fixed point iteration, is reformulated through local bandwidth based decomposition, as a fixed point update over clusters :

$$u^{\tau+1} = f(u^{\tau}), where\, u^{\tau=0} \equiv x_{u,u} \tag{1a}$$

$$f(u^{\tau}) = \Big(\sum_{\forall g \in G} \frac{1}{c_g}\Sigma_g^{-1} \sum_{\forall i | x_{i,g} \in Ne_x(u^{\tau})} K'\left(\|u^{\tau} - x_i\|_{\Sigma_g}\right)\Big)^{-1} \ldots \times \Big(\sum_{\forall g \in G} \frac{1}{c_g}\Sigma_g^{-1} \sum_{\forall i | x_{i,g} \in Ne_x(u^{\tau})} K'\left(\|u^{\tau} - x_i\|_{\Sigma_g}\right) x_i\Big) \tag{1b}$$

For ascertaining cluster merges, the data points in the vicinity of a cluster u's trajectory, u^{τ}, are considered. If a data point, y, in vicinity of u^{τ}, is ascertained to be heading to the same mode as u^{τ}, then by transitivity - all the members of its parent cluster, $\Pi(y)$, are heading to that mode too - the clusters u and $\Pi(y)$, can then be merged. The cluster which is higher up the mode (higher density) assimilates the other cluster into itself, thus accelerating convergence. This also helps in avoiding spurious merges.

The bandwidth, Σ_u, of a cluster, u, is updated every iteration utlizing T_u - the set of trajectory points arising from its constituent members.

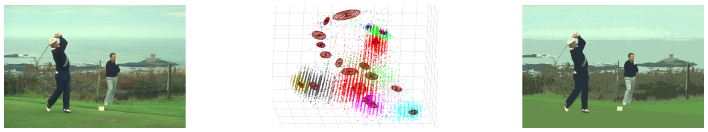
$$\Sigma_u = \frac{\Sigma_{\forall v \in T_u} \rho(v) v v^{\tau}}{\Sigma_{\forall v \in T_u} \rho(v)} - \eta_u \eta_u^T + \xi I,\ where\, \eta_u = \frac{\Sigma_{\forall v \in T_u} \rho(v) v}{\Sigma_{\forall v \in T_u} \rho(v)} \tag{2}$$

$\rho(v)$ is the data density in the immediate vicinity of a point $v \in T_u$. η_u and Σ_u are then the expectation, and variance of the localized distribution. *Eq.* 2 results in conservative but more localized and robust bandwidth estimates – more immune to long tails.

So starting with an initial base scalar, σ_{base}, the bandwidth matrices evolve by themselves. The nice part is that just a low base value suffices for reasonably dense data, with the bandwidths scaling data driven thereon and adapting to the local structure's scale, shape and orientation. σ_{base} thus becomes indicative of the minimum desired detail in the data space. This is opposed to traditional Mean Shift - where the bandwidth scalar is indicative of the scale at which data space has to be partitioned.

As *Figs. 1, 2(a)* indicate, reasonable local bandwidths arise, robustly identifying modes and salient clusters, by adapting according to local structure.

(*a*) Clustering (23 clusters) over color data (left) by the proposed approach. Segment image is shown on right.



(*b*) Comparitive results with standard MS (left) and variable-bandwidth isotropic MS, (VarMS, right), at similar clustering levels, 25 & 27 respectively, are shown. MS with correctly chosen bandwidth detected more coherent modes than VarMS, but looses partition saliency (bushes, water, sky in background). VarMS better adapts to scales but oversegments at places, and smooths over others (face). Both smoothed over details, failed to detect some modes at lower scales (trouser edges, maroon on shirt & shoes).

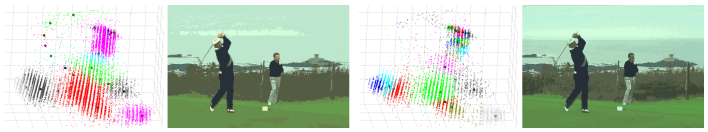


Figure 2: Exemplar illustrative result of our approach, AAAMS (a), is shown along with conventional MS results (b), at comparable clustering levels. As is indicated by the plots and segment images, AAAMS effectively adapts to local scale and preserves anisotropic details. This results in more salient yet parsimonius partitions.

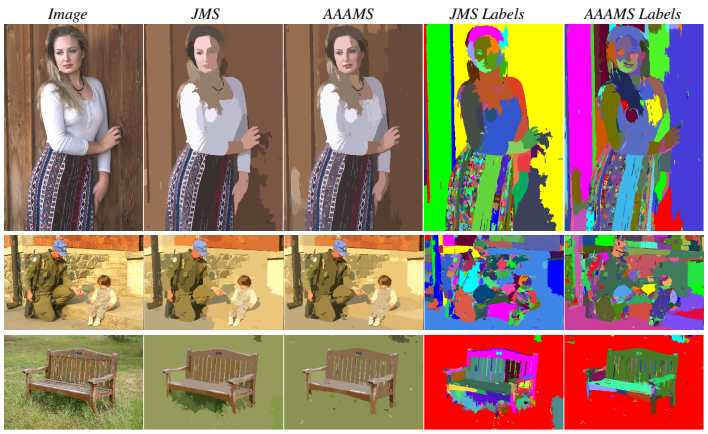


Figure 3: Example AAAMS results are shown, along with comparisons with standard joint domain Mean Shift (JMS). A single parameter set was used for AAAMS to show its adaptivity on varied images. At similar clustering levels, AAAMS preserved more details and affected more salient segmentations.

Promising qualitative and quantitative results were attained over image and point datasets - indicating the efficacy of the presented approach.

Future work would focus on experimenting with different merging schemes, and on more varied data spaces.

Board #49

From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains

Baochen Sun
http://www.cs.uml.edu/~bsun
Kate Saenko
http://www.cs.uml.edu/~saenko

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

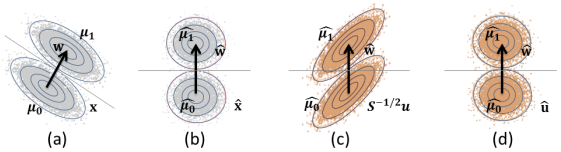


Figure 1: (a) Applying a linear classifier $\mathbf{w}$ learned by LDA to source data $\mathbf{x}$ is equivalent to (b) applying classifier $\hat{\mathbf{w}} = \mathbf{S}^{-1/2}\mathbf{w}$ to decorrelated points $\mathbf{S}^{-1/2}\mathbf{x}$. (c) However, target points $\mathbf{u}$ may still be correlated after $\mathbf{S}^{-1/2}\mathbf{u}$, hurting performance. (d) Our method uses target-specific covariance to obtain properly decorrelated $\hat{\mathbf{u}}$.

Abstract. The most successful 2D object detection methods require a large number of images annotated with object bounding boxes to be collected for training. We present an alternative approach that trains on virtual data rendered from 3D models, avoiding the need for manual labeling. Growing demand for virtual reality applications is quickly bringing about an abundance of available 3D models for a large variety of object categories. While mainstream use of 3D models in vision has focused on predicting the 3D pose of objects, we investigate the use of such freely available 3D models for multicategory 2D object detection. To address the issue of dataset bias that arises from training on virtual data and testing on real images, we propose a simple and fast adaptation approach based on decorrelated features.

Background. In recent years, use of the linear SVM with Histogram of Gradients (HOG) as the features has emerged as the predominant object detection paradigm. Yet, as observed by Hariharan *et al.* [3], training SVMs can be expensive, especially because it usually involves costly rounds of hard negative mining. Furthermore, the training must be repeated for each object category, which makes it scale poorly with the number of categories. Hariharan et al. proposed a much more efficient alternative using Linear Discriminant Analysis (LDA). LDA is a well-known linear classifier that models the training set of examples $\mathbf{x}$ with labels $y \in \{0,1\}$ as being generated by $p(\mathbf{x},y) = p(\mathbf{x}|y)p(y)$. $p(y)$ is the prior on class labels and the class-conditional densities are normal distributions $p(\mathbf{x}|y) = N(\mathbf{x};\mu^y,\mathbf{S})$, where the feature vector covariance $\mathbf{S}$ is assumed to be the same for both positive and negative (background) classes. In our case, the feature is represented by $\mathbf{x} = \phi(I,b)$. The resulting classifier is given by The innovation in [3] was to re-use $\mathbf{S}$ and $\mu_\mathbf{0}$, the background mean, for all categories, reducing the task of learning a new category model to computing the average positive feature, $\mu_\mathbf{1}$. This was accomplished by calculating $\mathbf{S}$ and $\mu_\mathbf{0}$ for the largest possible window and subsampling to estimate all other smaller window sizes. Also, $\mathbf{S}$ was shown to have a sparse local structure, with correlation falling off sharply beyond a few nearby image locations. LDA was shown in [3] to have competitive performance to SVM, and can be implemented both as an exemplar-based [4] or as deformable parts model (DPM) [1].

Approach. We observe that estimating global statistics $\mathbf{S}$ and $\mu_\mathbf{0}$ once and re-using them for all tasks may work when training and testing in the same domain, but in our case, the virtual training data is likely to have different statistics from the target real data. Figure 2 illustrates the effect of centering and decorrelating a positive mean using global statistics from the wrong domain. The effect is clear: important discriminative information is removed while irrelevant structures are not.

Based on this observation, we propose an adaptive decorrelation approach to detection. Assume that we are given labeled training data $\{\mathbf{x},y\}$ in the source domain (*e.g.* virtual images rendered from 3D models), and unlabeled examples $\mathbf{u}$ in the target domain (*e.g.* real images collected in an office environment). Evaluating the scoring function $f_\mathbf{w}(\mathbf{x})$ in the source domain is equivalent to first decorrelating the training features $\hat{\mathbf{x}} = \mathbf{S}^{-1/2}\mathbf{x}$, computing their positive and negative class means $\hat{\mu_\mathbf{1}} = \mathbf{S}^{-1/2}\mu_\mathbf{1}$ and $\hat{\mu_\mathbf{0}} = \mathbf{S}^{-1/2}\mu_\mathbf{0}$ and then projecting the decorrelated feature onto the decorrelated difference between means, $f_\mathbf{w}(\mathbf{x}) = \hat{\mathbf{w}}^T\hat{\mathbf{x}}$, where $\hat{\mathbf{w}} = (\hat{\mu_\mathbf{1}} - \hat{\mu_\mathbf{0}})$. This is illustrated in Figure 1(a-b). However, as we saw in Figure 2, the assumption that the input is properly decorrelated does not hold if the input comes from a target domain with a different covariance structure. Figure 1(c) illustrates this case, showing that $\mathbf{S}^{-1/2}\mathbf{u}$ does not have isotropic covariance. Therefore, $\mathbf{w}$ cannot be used directly.

We may be able to compute the covariance of the target domain on the unlabeled target points $\mathbf{u}$, but not the positive class mean. Therefore, we would like to re-use the decorrelated mean difference $\hat{\mathbf{w}}$, but adapt to the covariance of the target domain. In this paper, we make the assumption that the difference between positive and negative means is the same in the source and target.

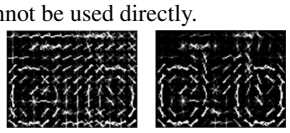


Figure 2: Mean bicycle decorrelated with mismatched-domain covariance (left) vs. with same-domain covariance (right).

Let the estimated target covariance be $\mathbf{T}$. We first decorrelate the target input feature with its inverse square root, and then apply $\hat{\mathbf{w}}$ directly, as shown in Figure 1(d). The resulting scoring function is $f_{\hat{\mathbf{w}}}(\mathbf{u}) = ((\mathbf{T}^{-1/2})^T\mathbf{S}^{-1/2}(\mu_\mathbf{1} - \mu_\mathbf{0}))^T\mathbf{u}$. This corresponds to a transformation of $(\mathbf{T}^{-1/2})^T(\mathbf{S}^{-1/2})$ instead of the original whitening $\mathbf{S}^{-1}$ being applied to the difference between means to compute $\mathbf{w}$. Note that if source and target domains are the same, then $(\mathbf{T}^{-1/2})^T(\mathbf{S}^{-1/2})$ equals to $\mathbf{S}^{-1}$ since $\mathbf{S}$ is positive definite.

In practice, either the source or the target component of the above transformation may also work, or even statitstics from similar domains. However, as shown by our experiments, dissimilar domain statistics can significantly hurt performance. Furthermore, if either source or target has only images of the positive category available, and cannot be used to properly compute background statistics, the other domain can still be used.

We also extend our approach to supervised adaptation when a few labeled examples are available in the target domain. Following [2], a simple adaptation method is used whereby the template learned on source positives is combined with a template learned on target positives, using a weighted linear combination. The key difference with our approach is that the target template uses target-specific statistics. In [2], the author uses the same background statistics as [3] which were estimated on 10,000 natural images from the PASCAL VOC 2010 dataset. Based on our analysis, even though these background statistics were estimated from a very large amount of real image data, they will not work for all domains. Our results confirms this claim.

We evaluate our technique by training on virtual labeled examples and testing on real images from a benchmark domain adaptation dataset. We compare two kinds of virtual data, one rendered with real-image textures and one without. The evaluation demonstrates that with our method, performance of classifiers trained on virtual data is comparable to that of classifiers trained on large-scale real image domains.

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

[2] Daniel Goehring, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Interactive adaptation of real-time object detectors. In *International Conference on Robotics and Automation (ICRA)*, 2014.

[3] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *Computer Vision–ECCV 2012*, pages 459–472. Springer, 2012.

[4] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 89–96. IEEE, 2011.

Leveraging Feature Uncertainty in the PnP Problem

Luis Ferraz[1]
luis.ferraz@upf.edu
Xavier Binefa[1]
xavier.binefa@upf.edu
Francesc Moreno-Noguer[2]
fmoreno@iri.upc.edu

[1] Department of Information and Communication Technologies
Universitat Pompeu Fabra
08018, Barcelona, Spain

[2] Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
08028, Barcelona, Spain

Introduction: The goal of the Perspective-n-Point (PnP) problem is to estimate the position and orientation of a calibrated camera from a set of n 3D-to-2D point matches. State-of-the-art PnP solutions assume that these correspondences may be corrupted by noise and show robustness against large amounts of it. Yet, none of these works considers that the particular structure of the uncertainty associated to each correspondence could indeed to be used to further improve the accuracy of the estimated pose. Specifically, existing solutions, as [3, 4], assume all 2D correspondences to be affected by the same model of noise, a zero mean Gaussian distribution, and consider all correspondences to equally contribute to the estimated pose, independently of the precision of their actual location.

Contributions: In this paper we propose a real-time and accurate PnP solution that exploits the fact that in practice the 2D position of not all 2D features is estimated with the same accuracy (see Fig.1(a,b)). Assuming a model of such feature uncertainties is known in advance, we reformulate the PnP problem as a Maximum Likelihood minimization approximated by an unconstrained Sampson error function, which naturally penalizes the most noisy correspondences. Pre-estimating feature uncertainty in real experiments is, though, not easy. In this paper we model it as 2D Gaussian distributions representing the sensitivity of the underlying 2D feature detectors to different camera viewpoints. When using these noise models with our PnP formulation we still obtain promising pose estimation results that outperform most recent approaches.

Method: Let $\mathbf{u}_i = [u_i, v_i]^\top$ be an observed 2D point obtained using a feature detector. This observed value can be regarded as the true 2D projection $\overline{\mathbf{u}}_i$ perturbed by a random variable $\Delta\mathbf{u}_i$,

$$\mathbf{u}_i = \overline{\mathbf{u}}_i + \Delta\mathbf{u}_i \tag{1}$$

We assume that $\Delta\mathbf{u}_i$ is small, independent and unbiased, and model it as a Gaussian distribution with expectation $E[\Delta\mathbf{u}_i] = \mathbf{0}$ and 2×2 covariance matrix $E[\Delta\mathbf{u}_i\Delta\mathbf{u}_i^\top] = \mathbf{C}_{\mathbf{u}_i}$, which is known in advance.

Taking into account these uncertainties the PnP problem can be solved as the following Maximum Likelihood for all n correspondences,

$$\arg\min_{\Delta\mathbf{u}_i,\mathbf{x}} \sum_{i=1}^{n} \|\Delta\mathbf{u}_i\|^2_{\mathbf{C}_{\mathbf{u}_i}^{-1}} \qquad \text{subject to} \quad \mathbf{M}_{\overline{\mathbf{u}}_i}\mathbf{x} = \mathbf{0} \tag{2}$$

where $\mathbf{M}_{\overline{\mathbf{u}}_i}\mathbf{x} = \mathbf{0}$ enforce the 3D-to-2D projective constraints of the noise-free correspondences and $\mathbf{x}$ represents a set of control points in camera coordinates. Since we assumed the uncertainty $\Delta\mathbf{u}_i = [\Delta u_i \;\; \Delta v_i]^\top$ to be small, the perspective constraint can be approximated using first order perturbation analysis

$$\mathbf{M}_{\overline{\mathbf{u}}_i}\mathbf{x} = \mathbf{M}_{\mathbf{u}_i}\mathbf{x} - \Delta u_i \nabla_u \mathbf{M}_{\mathbf{u}_i}\mathbf{x} - \Delta v_i \nabla_v \mathbf{M}_{\mathbf{u}_i}\mathbf{x} = \mathbf{0} \tag{3}$$

where $\nabla_u\mathbf{M}_{\mathbf{u}_i}$ and $\nabla_v\mathbf{M}_{\mathbf{u}_i}$ are the partial derivatives of $\mathbf{M}_{\mathbf{u}_i}$ with respect to u and v; and as in [2], $\mathbf{M}_{\mathbf{u}_i}$ encodes the perspective constraints.

Using Lagrange Multipliers Eq. 2 is rewritten as an unconstrained minimization of a Sampson Error function and solved using the Fundamental Numerical Scheme (FNS) approach [1].

Finally, once $\mathbf{x}$ is estimated, the PnP problem is solved following the Procrustes analysis proposed in [2].

Uncertainties estimation: Estimating 2D feature uncertainties $\mathbf{C}_{\mathbf{u}_i}$ in real images is still an open problem. Our approach starts by detecting features on a given reference view $\mathbf{V}_r$ of the object of interest. Then, we synthesize m novel views $\{\mathbf{I}_1, \ldots, \mathbf{I}_m\}$ of the object, which sample poses around $\mathbf{V}_r$. We then extract 2D features for each $\mathbf{I}_j$, and reproject them back to $\mathbf{V}_r$, creating feature point clouds (see Figure 1c).

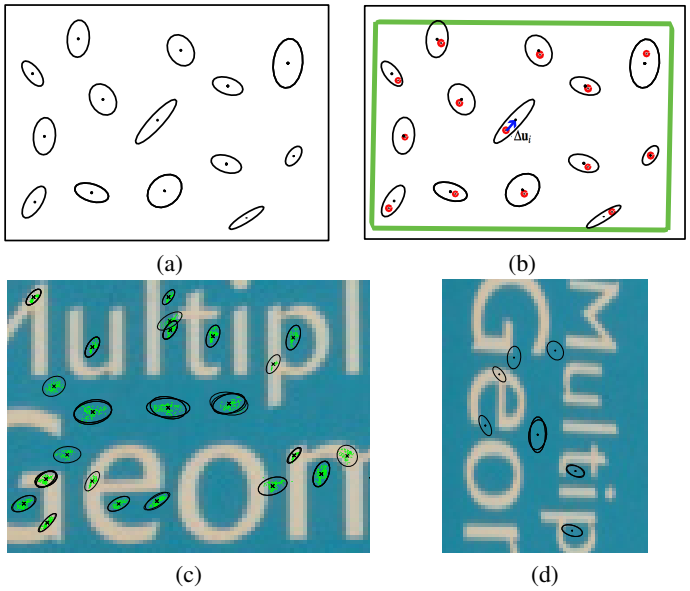


Figure 1: PnP problem with noisy correspondences. We assume 2D feature points are associated to particular noise models, as shown in (a). Our approach estimates a solution of the PnP problem that minimizes the Mahalanobis distances $\Delta\mathbf{u}_i$ shown in (b). The Green rectangle and red dots are the true projection of the 3D model and 3D points. Using our approach feature uncertainties $\mathbf{C}_{\mathbf{u}_i}$ (black ellipses) on real images are estimated for each reference view (c). In (d) uncertainties are aligned with a test image.

Once features are grouped we model each cluster i with a covariance matrix $\mathbf{C}_{\mathbf{u}_i}$. Note that this covariance tends to be anisotropic, thus it is not rotationally invariant. To achieve this invariance we use the main gradients as done by the SIFT detector. Fig.1(d) shows how each $\mathbf{C}_{\mathbf{u}_i}$ is rotated respect to the main gradients.

In practice, we found that $\mathbf{C}_{\mathbf{u}_i}$ accurately describes the uncertainties when the pose of $\mathbf{I}_j$ is close to the pose of the reference $\mathbf{V}_r$. This accuracy drops when camera moves away. In order to handle this, we defined a set of l reference images $\{\mathbf{V}_1, ..., \mathbf{V}_l\}$ under different poses and each one with its own uncertainty models. We experimentally found that a grid of reference images, taken all around the 3D object at every 20° in yaw and pitch angles, yielded precise uncertainty models.

Algorithm for real images is split into the following three main steps:

1. Estimate an initial camera pose without considering feature uncertainties using EPPnP. Let $[\mathbf{R}|\mathbf{t}]_{EPPnP}$ be this initial pose.
2. Pick the nearest reference view $\mathbf{V}_k$. Solving $\max_k \left(\frac{\mathbf{c}_k^\top}{\|\mathbf{c}_k\|} \cdot \frac{\mathbf{c}_{EPPnP}}{\|\mathbf{c}_{EPPnP}\|} \right)$, where $\mathbf{c}_*/\|\mathbf{c}_*\|$ are the normalized camera centers in world coordinates, being $\mathbf{c}_* = -\mathbf{R}_*^\top\mathbf{t}_*$.
3. Solve Eq. 2 using the covariances $\mathbf{C}_{\mathbf{u}_i}$ of the reference image $\mathbf{V}_k$, and $[\mathbf{R}|\mathbf{t}]_{EPPnP}$ for initializing the iterative process. The final pose $[\mathbf{R}|\mathbf{t}]_{CEPPnP}$ is obtained using Procrustes as in [2].

[1] W. Chojnacki, M. J. Brooks, A. Van Den Hengel, and D. Gawley. On the fitting of surfaces to data with covariances. *PAMI*, 2000.

[2] L. Ferraz, X. Binefa, and F. Moreno-Noguer. Very fast solution to the PnP problem with algebraic outlier rejection. In *CVPR*, 2014.

[3] C-P. Lu, G. D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *PAMI*, 22(6):610–622, 2000.

[4] Y. Zheng, Y. Kuang, S. Sugimoto, K. Aström, and M. Okutomi. Revisiting the PnP problem: A fast, general and optimal solution. In *ICCV*, 2013.

Board #51

Exploiting Colour Information for Better Scene Text Recognition

Muhammad Fraz[1]
M.Fraz@lboro.ac.uk
M. Saquib Sarfraz[2]
Muhammad.Sarfraz@kit.edu
Eran A. Edirisinghe[1]
E.A.Edirisinghe@lboro.ac.uk

[1] Department of Computer Science
Loughborough University
Loughborough, UK

[2] Computer Vision for Human Computer Interaction Lab
Karlsruhe Institute of Technology
Karlsruhe, Germany

The problem of scene text recognition has gained significant importance because of its numerous applications. A variety of methods has been recently proposed that explore various theoretical and practical aspects to solve this problem. In this work, we focus towards a framework to recognize the text present in outdoor scene images. The text information carries one important property, that is, its colour in comparison to its background. Text information is always placed in such a way that it stands out from its background. In the same way, most of the time the characters in a word possess similar colour that helps us to recognize the letters of a particular word. We exploit this characteristic of text regions to solve the problem of character recognition. The character recognition pipeline is further extended in to a word recognition framework where the estimated word combinations are matched against a lexicon.

The existing approaches for scene text recognition can be roughly divided in to two broad categories: Region grouping based methods and object recognition based methods. In this work, we have combined region grouping method with object recognition based strategy to achieve the advantages of both techniques. First, we binarize the image using colour information and perform foreground segmentation to separate characters from background. Next, we extract shape representation features on binary images and perform character classification using a pre-trained classifier. The recognized characters form words that are fed in to a string similarity matching stage where lexicon based search is performed to find the closest matching word.

Character Identification: We use the bilateral regression [2] for character identification. However, our approach is different than the original method in that we only use it to estimate the horizontal location of each character in word image. The bilateral regression models the foreground pixels by using a weighted regression that assigns weight to each pixel according to its location with respect to foreground in feature space. The pixels that belong to the foreground get high weights in comparison to the pixels belonging to background. In this case, the regression model in equation 1 represents the quadratic surface that best models the image as a function of pixel locations.

$$z = ax^2 + by^2 + cxy + dx + ey + f \quad (1)$$

We enhance the operation of bilateral regression by a pre-processing step where the foreground colour is estimated a priori. We apply n-level colour quantization to achieve binary image for each quantization level. We use Minimum Variance Quantization (MVQ) originally proposed by Heckbert [3]. We quantize each word image in to three colours and analyse the respective binary maps for three quantization levels to estimate the foreground. The characters are cropped from the actual word images using the estimated horizontal location and width from bilateral regression while the height is kept same as the height of the actual word image. the segmented masks are used to crop the characters from original (coloured) image and fed into the character recognition pipeline explained next.

Character Classification: Similar to the character identification stage, we use colour quantization to enhance the character. We found on the basis of extensive experimentation that for a character image 2-level quantization is good enough to recover the full character pixels from background. We therefore generate two binary images corresponding to the two colour levels by assigning the pixels for each colour cluster a value '1' (white similar to the previous stage), we categorize the two binary images as foreground character map by simply analysing the white pixels density along the borders of each binary map. The binary map that possesses the higher total number of corner white pixels is considered as background and the other binary map is classified as the character map. We compute HOG-SVM for character binary map representation and classification.

Word Recognition: The errors in character recognition are inevitable because of high interclass similarity between various characters. In order to find the correct word from various character combinations, the predicted words are aligned with the words available in the lexicon using a string similarity measure. The closest matched word in the lexicon is declared as the word in the image. We adopted a simple strategy where the alignment is performed using Lavenshtien distance.

Figure 1: Improved character identification. Row 1 shows the original images. Row 2 shows the results of character segmentation using Bilateral Regression. Row 3 shows the results of character segmentation using the combination of proposed pre-processing and Bilateral Regression.

Results: The proposed characters recognition pipeline outperforms the current state-of-the-art on Chars74k [1] ICDAR03-CH [5] dataset. Further to that, the proposed word recognition pipeline outperforms the state-of-the-art on challenging ICDAR03-Word [5] and ICDAR11-Word [4] benchmark datasets.

Computational Performance: The proposed framework is implemented in MATLAB. The average execution time for the proposed word recognition pipeline on a standard PC is 1.7 seconds. The separate average execution time for three stages: Character Identification, Character Recognition and Word Recognition is 1.2 sec., 0.4 sec. and 0.1 sec. respectively. Note that the code is unoptimized. The execution time can be further reduced near real-time with the inclusion of code optimization and parallel processing techniques.

Conclusively, the proposed recognition method combines region grouping method with object recognition based strategy to achieve state-of-the-art performance on benchmark datasets. The proposed modification for bilateral regression based segmentation drastically improved character identification performance. The binary maps of the segmented characters have been directly used to extract shape features and fed in to the trained SVM classifier. Finally, a basic string similarity measure has been used to align the estimated words with the lexicon to remove inaccuracies. The experimental results show that proposed framework is accurate, fast, simple and exploitable for practical applications.

[1] T. de. Campos, B. Babu, and M. Verma. Character recognition in natural images. In *VISAPP*, 2009.

[2] Jacqueline L. Feild and Erik G. Learned-Miller. Improving open-vocabulary scene text recognition. In *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, ICDAR '13, pages 604–608, Washington, DC, USA, 2013. IEEE Computer Society. ISBN 978-0-7695-4999-6. doi: 10.1109/ICDAR.2013.125.

[3] Paul Heckbert. Color image quantization for frame buffer display. *SIGGRAPH Comput. Graph.*, 16(3):297–307, July 1982. ISSN 0097-8930. doi: 10.1145/965145.801294.

[4] A Shahab, F. Shafait, and A Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1491–1496, Sept 2011. doi: 10.1109/ICDAR.2011.296.

[5] L. P. Sosa, S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions. In *In Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 682–687. IEEE Press, 2003.

Structured Semi-supervised Forest for Facial Landmarks Localization with Face Mask Reasoning

Xuhui Jia[1]
xhjia@cs.hku.hk
Heng Yang[2]
heng.yang@eecs.qmul.ac.uk
Angran Lin[1]
arlin@cs.hku.hk
Kwok-Ping Chan[1]
kpchan@cs.hku.hk
Ioannis Patras[2]
i.patras@eecs.qmul.ac.uk

[1] Department of Computer Science
The Univ. of Hong Kong, HK
[2] School of EECS
Queen Mary Univ. of London, UK

Motivation. Despite the great success of recent facial landmarks localization approaches, the presence of occlusions significantly degrades the performance of the systems [2, 5]. Though occlusion occur frequently in realistic scenarios (e.g. the use of scarf or sunglasses, hands or hair on the face), very few works have addressed this problem explicitly due to the high diversity of occlusion in real world. While [4] tried to model a few synthetic occlusion patterns, the recent method of [1] dealt with the occlusion problem in more realistic sceneries. Both of them only focused on modelling the occlusion in an unstructured way, i.e. treating the visibility of each landmark independently. However in realistic conditions, the occlusion patterns (or called occluders) often occupy a continuous region instead of an individual pixel location, as depicted in Fig 2. Thereby the whole occluded region will consistently affect the landmarks localization.

Contribution. This work attempts to address the face mask reasoning and facial landmarks localization in an unified Structured Decision Forests framework. We first have built a rich face image dataset with face mask annotation. The dataset was built as an extension of the recent datasets: Caltech Occluded Faces in the Wild (COFW), Labeled Face Parts in the Wild (LFPW) and Labeled Face in the Wild (LFW). We manually annotate a portion of images in these datasets with face masks. The face mask indicates whether or not each pixel belongs to the face. Then we incorporate such additional information of dense pixel labelling into training the Structured Classification-Regression Decision Forest. The classification nodes aim at decreasing the variance of the pixel labels of the patches by using our proposed structured criterion while the regression nodes aim at decreasing the variance of the displacements between the patches and the facial landmarks. The proposed framework allows us to predict the face mask and facial landmarks locations jointly. The proposed framework with following properties. First, semi-supervised, it uses training images from the above described augmented dataset, only a portion of which are with face masks. Second, structured, it has a novel structured criterion for split function selection for the pixel labelling (face mask reasoning) problem. Third, joint classification-regression, it predicts face mask label for each pixel (classification) and the landmark locations (regression) at the same time, and more importantly it uses the face mask reasoning results

Figure 2: The images on the left side of the two pairs show the results from the standard Random Forests for facial landmarks localization [3], with failure cases under occlusion. The images on the right side of the two pairs show the results of our proposed method. It first explicitly predicts the face mask (the semi-transparent region), then use the face mask information to improve the localization and to predict the occlusion status of the landmarks.

to improve the accuracy of landmark localization . Experiments show our method 1) yields promising results in face mask reasoning; 2) improves the existing Decision Forests approaches in facial landmark localisation, aided by face mask reasoning.

[1] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.
[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.
[3] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012.
[4] Golnaz Ghiasi and Charless Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, 2014.
[5] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

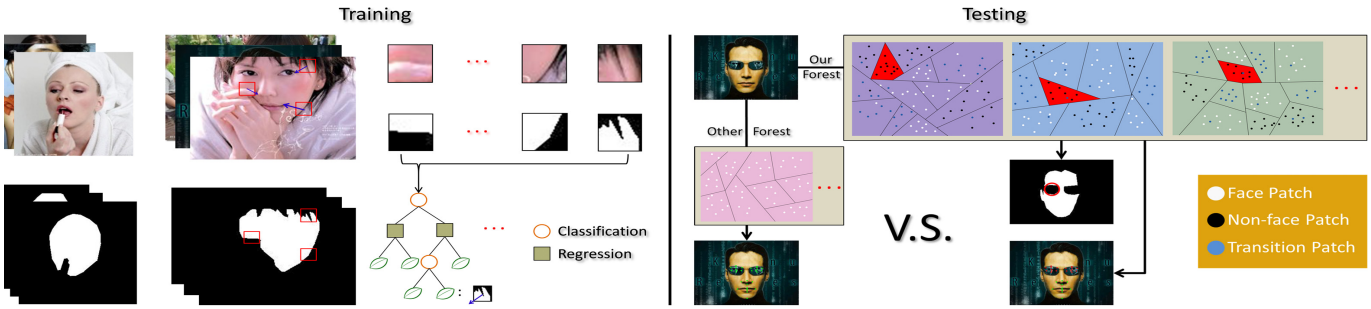


Figure 1: The framework of proposed method. We use face images with annotation of facial landmarks and face masks for training. By randomly switching the information gain function at the internal nodes, the decision trees are optimized with respect to both the offsets to landmarks (regression) and to the local structured label configuration (classification). The forest model is able to predict the face mask and landmark locations jointly. We exploit the face mask prediction to further improve the landmark localization.

Board #53

Action Recognition from Weak Alignment of Body Parts

Minh Hoai[1]
http://www.robots.ox.ac.uk/~minhhoai/
L'ubor Ladický[2]
http://www.inf.ethz.ch/personal/ladickyl/
Andrew Zisserman[1]
http://www.robots.ox.ac.uk/~az/

[1] Visual Geometry Group
Department of Engineering Science
University of Oxford,
Oxford, UK
[2] ETH Zürich
Zürich, Switzerland

The objective of this paper is to recognize human actions in still images. The contribution of this work is a novel framework for obtaining weak alignment of human body-parts to improve the recognition performance. Our framework implicitly exploits physical constraints of human body parts (e.g., heads are above necks, hands are attached to forearms). It uses the locations of some detected body parts to aid the alignment of some others. Specifically, we demonstrate the benefit of our framework for computing registered feature descriptors from automatically detected upper bodies and silhouettes. Fig. 1 illustrates the benefits of our approach over the grid-alignment approach.

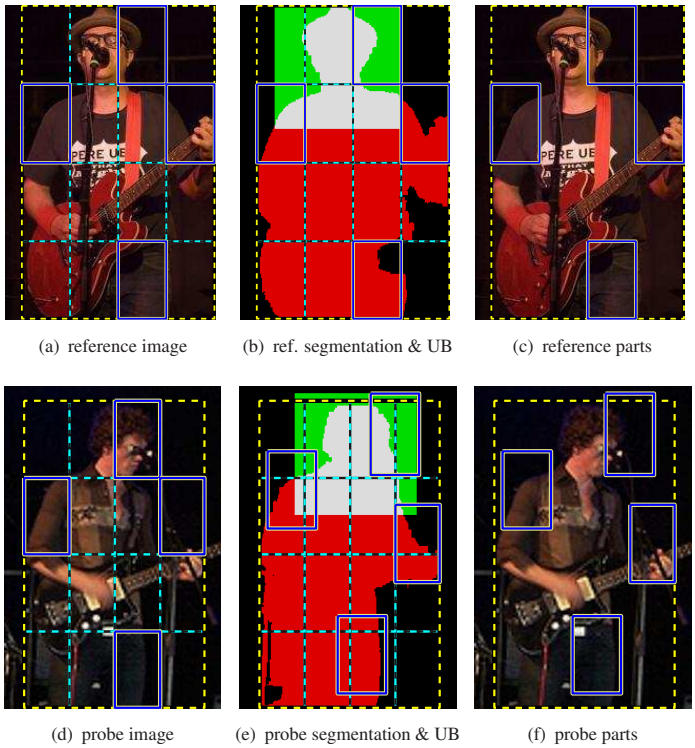
Given the bounding box of a human, we approximate the human body by a set of deformable rectangular parts, which is similar to a DPM [1]. The goal is to align these rectangular parts between two images, referred to as *reference* and *probe* images. We formulate the problem as a minimization of a deformation energy between the parts of the reference (which are fixed as a default grid formation) and those of the probe (which deform to best match those of the reference). The energy encourages the parts to overlap the silhouette and upper body in a consistent way (between reference and probe) whilst penalizing severe deformations. The energy is defined for a configuration of parts, and it is formulated as the sum of unary and pairwise terms. Consider aligning a human specified by a bounding box $\mathbf{b}$ in the probe image $\mathbf{I}$ to another human specified by the bounding $\mathbf{b}^{ref}$ in the reference image $\mathbf{I}^{ref}$. Let $\mathbf{p}_1^{ref},\cdots\mathbf{p}_k^{ref}$ be the default configuration of parts for the reference image at the bounding box $\mathbf{b}^{ref}$. We consider the following energy function for a configuration of parts $\mathbf{p}_1,\cdots,\mathbf{p}_k$ of a probe image $\mathbf{I}$:

$$E(\{\mathbf{p}_i\}) = \sum_{i=1}^{k} ||\phi(\mathbf{I},\mathbf{p}_i) - \phi(\mathbf{I}^{ref},\mathbf{p}_i^{ref})||^2 + \lambda \sum_{i=1}^{k} ||\psi(\mathbf{p}_i, par(\mathbf{p}_i)) - \psi_i^{def}||^2. \quad (1)$$

The above energy function factors into a sum of local and pairwise energies. $\phi(\mathbf{I},\mathbf{p}_i)$ is the feature vector computed at the location specified by part $\mathbf{p}_i$ of image $\mathbf{I}$. In this work, it is a vector of two components. The first component is the proportion of pixels inside $\mathbf{p}_i$ that belong to the detected upper body, and the second component is the proportion of pixels inside $\mathbf{p}_i$ that belong to the human segmentation. $par(\mathbf{p}_i)$ is the parent of $\mathbf{p}_i$; the parent of the root part is the provided bounding box $\mathbf{b}$. ψ is the function that computes the relative displacement of a part and its parent. ψ_i^{def} is the displacement computed for the default configuration of parts. The energy for a configuration of parts is given by the difference of each part at its respective location w.r.t. the corresponding part in the reference image (data term) plus a deformation cost that depends on the relative positions of each part w.r.t. the parent (spatial prior).

We align an image with a set of training (or reference) images as follows. We first divide the training images into three roughly equal subsets, based on the aspect ratios of the provided person bounding boxes. Given a probe image (either training or testing), we determine the subset that has similar aspect ratio, and compute the matching energy between the probe image and every training image in the subset. The matching energy is the difference (in the occupancy of silhouette and upper body) between the two default configurations of parts, as defined in Eq. 1. The m training images that yield the lowest matching energies, referred to as m nearest neighbors, are used as the references for aligning the probe image. This produces m configurations of parts for the probe image, defining its deformation space.

The alignment of a probe image w.r.t. its nearest neighbors can be used to compute an improved feature descriptor for any type of feature, including HOG and color. For example, consider a feature descriptor in which a HOG template is computed for each part. Using our approach, for each of the nearest neighbors, the HOG template can be computed at the deformed configuration of parts. We pool the HOGs for each corresponding part by averaging. The process can be thought as alignment-informed jittering.



(a) reference image (b) ref. segmentation & UB (c) reference parts

(d) probe image (e) probe segmentation & UB (f) probe parts

Figure 1: **Aligning body parts for action recognition**. (a) & (d): the alignment induced by a regular grid is not suited for registering body parts (c.f., solid blue boxes). The geometric constraints provided by the silhouettes and upper bodies ((b) & (e)) lead to a good alignment of parts. (c) & (f): alignment results – the translated parts are better aligned with the reference parts (e.g., the rightmost blue boxes both correspond to a hand holding the guitar fretboard).

Human silhouettes are obtained using a foreground/background segmentation algorithm. This algorithm is based on a joint energy minimization framework [2] that consists of energy potentials from a pose model, a color model, and texture classifiers. To localize the upper body, the Calvin upper-body detector is used.

We train a kernel SVM for each action class. The SVM kernel is a convex combination of base kernels, which capture different visual cues: HOG, SIFT, color, pose, object detection scores. Some of these cues are computed at various relative locations of the provided human bounding box, yielding a total of 20 kernels. We evaluated the descriptors on the default and on the deformed part configurations. We optimize the weights for kernel combination using randomized grid search.

Experiments on the challenging PASCAL VOC 2012 dataset show that our method outperforms the state-of-the-art on the majority of action classes.

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9):1627–1645, 2010.

[2] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *Proc. CVPR*, 2013.

Board #54

Bird Species Categorization Using Pose Normalized Deep Convolutional Nets

Steve Branson[1]
sbranson@caltech.edu
Grant Van Horn[2]
gvanhorn@ucsd.edu
Serge Belongie[3]
tech.cornell.edu
Pietro Perona[1]
vision.caltech.edu

[1] California Institute of Technology Pasadena, CA, USA
[2] University of California, San Diego
La Jolla, CA, USA
[3] Cornell Tech
New York, NY, USA

In this work we propose an architecture for fine-grained visual categorization that approaches expert human performance in the classification of bird species. We perform a detailed investigation of state-of-the-art deep convolutional feature implementations and fine-tuning feature learning for fine-grained classification. We observe that a model that integrates lower-level feature layers with pose-normalized extraction routines and higher-level feature layers with unaligned image features works best. Our experiments advance state-of-the-art performance on bird species recognition, with a large improvement of correct classification rates over previous methods (75% vs. 55-65%).

Our architecture can be organized into 4 components: keypoint detection, region alignment, feature extraction, and classification. We predict 2D locations and visibility of 13 semantic part keypoints of the birds using the DPM implementation from [1] . These keypoints are then used to warp the bird to a normalized, prototype representation. To determine the prototype representations, we propose a novel graph-based clustering algorithm for learning a compact pose normalization space. Features, including HOG, Fisher-encoded SIFT, and outputs of layers from a CNN [3], are extracted (and in some cases combined) from the warped region. The final feature vectors are then classified using an SVM.

Although we believe our methods will generalize to other fine-grained datasets, we forgo experiments on other datasets in favor of performing more extensive empirical studies and analysis of the most important factors to achieving good performance on CUB-200-2011. Specifically, we analyze the effect of different types of features, alignment models, and CNN learning methods. We believe that the results will be informative to researchers who work on object recognition in general.

Our fully automatic approach achieves a classification accuracy of 75.7%, a 30% reduction in error from the highest performing (to our knowledge) existing method [2]. We note that our method does not assume ground truth object bounding boxes are provided at test time (unlike many/most methods). If we assume ground truth part locations are provided at test time, accuracy is boosted to 85.4%. These results were obtained using prototype learning using a similarity warping function computed using 5 keypoints per region, CNN fine-tuning, and concatenating features from all layers of the CNN for each region. The major factors that explain performance trends and improvements are:

1. Choice of features caused the most significant jumps in performance. The earliest methods that used bag-of-words features achieved performance in the $10-30\%$ range. Recently methods that employed more modern features like POOF, Fisher-encoded SIFT and color descriptors, and Kernel Descriptors (KDES) significantly boosted performance into the $50-62\%$ range. CNN features have helped yield a second major jump in performance to $65-76\%$. See Figure 1.
2. Incorporating a stronger localization/alignment model is also important. Among alignment models, a similarity transformation model fairly significantly outperformed a simpler translation-based model. Using more keypoints to estimate warpings and learning pose regions yielded minor improvements in performance. See Figure 2.
3. When using CNN features, fine-tuning the weights of the network and extracting features from mid-level layers yielded substantial improvements in performance. See Figure 3.

[1] Steve Branson, Oscar Beijbom, and Serge Belongie. Efficient large-scale structured learning. In *CVPR*, 2013.

[2] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

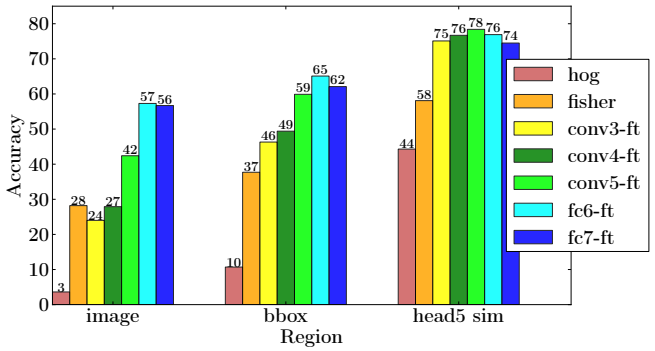


Figure 1: **Feature Performance Comparison:** CNN features significantly outperform HOG and Fisher features for all levels of alignment (image, bounding box, head).

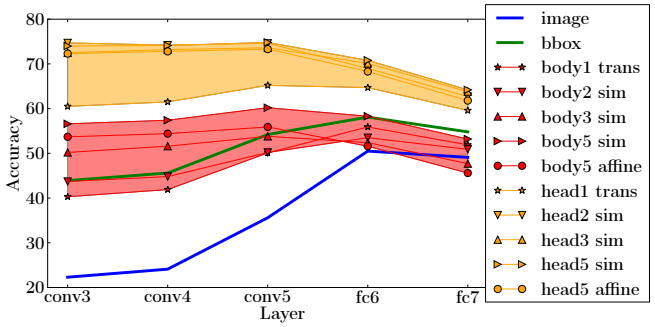


Figure 2: **Effect of CNN Layers For Different Regions:** The later fully connected layers (fc6 & fc7) significantly outperform earlier layers when a crude alignment model is used (image-level alignment), whereas convolutional layers (conv5) begin to dominate performance as we move to a stronger alignment model (from image $\rightarrow$ bbox $\rightarrow$ body $\rightarrow$ head).

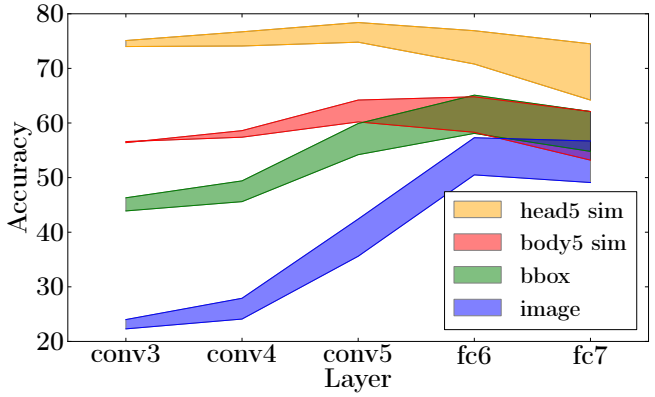


Figure 3: **Effect of Fine-Tuning with GT Parts:** Fine-tuning significantly improves performance for all alignment levels (width of each tube). Improvements occur for all CNN layers; however, the effect is largest for fully connected layers.

 Board #55

Speeding up Convolutional Neural Networks with Low Rank Expansions

Max Jaderberg
max@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
Oxford, UK

The focus of this paper is speeding up the application of convolutional neural networks (CNNs). While delivering impressive results across a range of computer vision and machine learning tasks, these networks are computationally demanding, limiting their deployability. Convolutional layers generally consume the bulk of the processing time, and so in this work we present two simple schemes for drastically speeding up these layers. This is achieved by exploiting cross-channel or filter redundancy to construct a low rank basis of filters that are rank-1 in the spatial domain. Our methods are architecture agnostic, and can be easily applied to existing CPU and GPU convolutional frameworks for tuneable speedup performance. We demonstrate this with a real world network designed for scene text character recognition [1], showing a possible 2.5× speedup with no loss in accuracy, and 4.5× speedup with less than 1% drop in accuracy, still achieving state-of-the-art on standard benchmarks.

Approximation Schemes. We provide the frameworks for two methods to approximate the N 3D filters W_n of a convolutional layer acting on the input z with C channels z_n^c such that $W_n * z = \sum_{c=1}^{C} W_n^c * z^c$. Both methods exploit the redundancy that exists between different feature channels and filters of convolutional layers.

Scheme 1 builds upon the work of Rigamonti *et al.* [2] and approximates the original set of full-rank filters as a linear combination of a smaller set of separable (rank-1) filters. The separability of these filters allows convolutions to be computed much more efficiently than the full-rank filters by splitting the full convolution in to horizontal convolution followed by vertical convolution. For a layer of N convolutional filters W_n^c, $n \in [1 \dots N]$, where each filter acts on a single channel c of the 3D input, $c \in [1 \dots C]$, we learn a basis of M separable filters s_m^c, $m \in [1 \dots M]$, where $M < N$, as well as the coefficients a_n^{cm} to linearly combine them, such that the original filter $W_n^c \simeq \sum_{m=1}^{M} a_n^{cm} s_m^c$, offering a speedup due to the separable convolution and smaller basis of filters required for convolution.

Scheme 2 also employs the idea of separable convolutions, but uses a separate basis of vertical filters and horizontal filters. The original convolutional layer is approximated by a vertical convolution layer with K vertical filters $\{v_k : k \in [1 \dots K]\}$ followed by a horizontal convolutional layer with horizontal filters $\{h_n : n \in [1 \dots N]\}$. This results in the original filters being approximated by the sequence of these two layers, *i.e.* $W_n^c \simeq \sum_{k=1}^{K} h_n^k * v_k^c$. The advantage of this method over Scheme 1 is that it can be plugged straight in to any CNN toolbox that supports rectangular filters.

For both schemes, the speedup can be tuned by varying the number of filters for each basis.

Optimization. The separable approximations can be optimized using two methods – *filter reconstruction optimization* and *data reconstruction optimization*. Filter reconstruction optimization aims to minimise the reconstruction error of the original filters by the approximation, whereas data reconstruction optimization aims to reconstruct the output of the original layer by the approximated layer for the training data inputs, minimising the reconstruction error using traditional back-propagation of errors.

Results. We provide the results of our approximation schemes on scene text character recognition using the state-of-the-art classifier of [1], under different scenarios and settings. A 4.5× speedup can be obtained with virtually no loss in classifier accuracy (Fig. 2), with data reconstruction optimization improving the accuracy of both schemes.

[1] M. Jaderberg, A Vedaldi, and A. Zisserman. Deep features for text spotting. In *European Conference on Computer Vision*, 2014.

[2] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua. Learning separable filters. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2754–2761. IEEE, 2013.

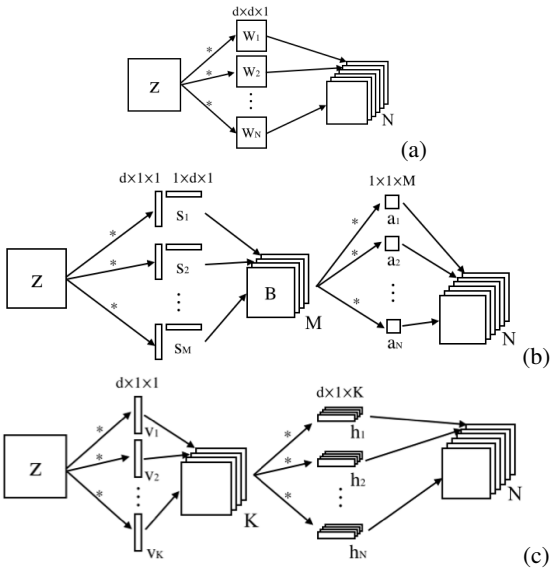
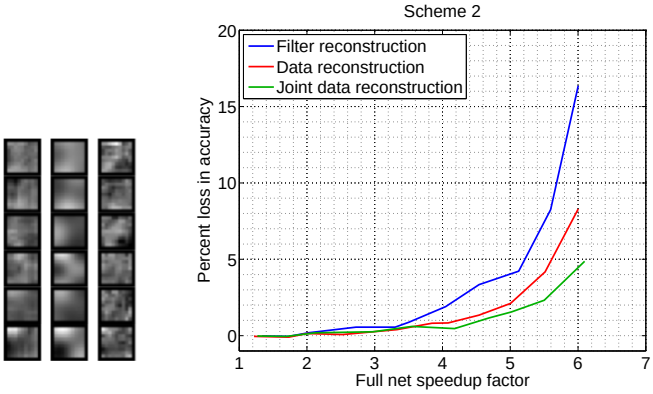


Figure 1: **Approximation Frameworks** (a) The original convolutional layer acting on a single-channel input *i.e.* C=1. (b) The approximation to that layer using the method of Scheme 1. (c) The approximation to that layer using the method of Scheme 2.



(a) (b)

Figure 2: **Approximation Results** (a) A selection of Conv2 filters from the original CNN (left), and the reconstructed versions under Scheme 1 (centre) and Scheme 2 (right), where both schemes have the same model capacity corresponding to 10x theoretical speedup. (b) The percent loss in performance as a result of the speedups attained with Scheme 2 (c). Joint data reconstruction optimizes the solution of multiple layers' approximations jointly, rather than optimizing each layer in isolation.

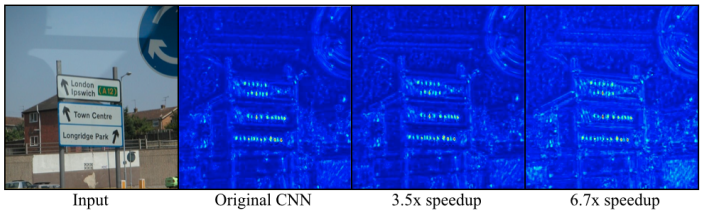


Figure 3: **Qualitative Result** Text spotting using the CNN character classifiers run in sliding window mode. The maximum response map over the character classes of the CNN output with Scheme 2 indicates the scene text positions. The approximations have sufficient quality to locate the text, even at 6.7× speedup.

Board #56

Real-time Hybrid Stereo Vision System for HD resolution disparity map

Jiho Chang
changjh@etri.re.kr
Jae-chan Jeong
channij80@etri.re.kr
Dae-hwan Hwang
hdh@etri.re.kr

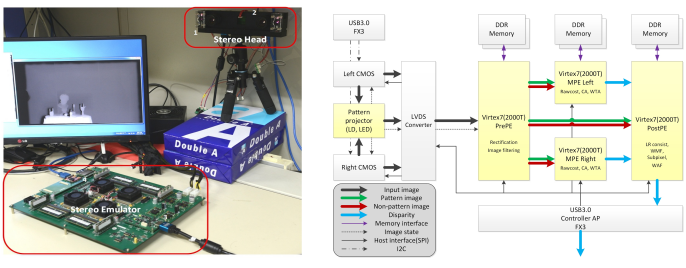
Intelligent Cognitive Technology Research Department
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Stereo matching is a traditional method used to obtain 3D depth information and has been studied for decades. However, it is still difficult to apply stereo matching algorithms to practical devices due to real-time issues as well as the technique's inability to adequately handle untextured regions. In this paper, we propose a hybrid stereo matching system to remedy the disadvantages of active and passive stereo vision.

Stereo matching algorithm Following Scharstein's taxonomy [5], the stereo matching algorithm divides into four steps: matching cost computation, cost aggregation, disparity computation and disparity refinement. First of all, we calculate raw cost volume using the AD-Census [2, 6]. At this time, we use cost combining with alpha-blending for the AD-Census, and the final raw cost that is the sum of the pattern cost (T_1) and the non-pattern cost (T_2). The information permeability filtering(PF) proposed by Cevahir Cigla *et al*. [2] is an ASW approach that has simple parameters and provides constant operational time for calculating cost aggregation. However, because there is no proximity weight term, PF can encounter problems with images containing large untextured regions. Modified information permeability(MPF), including a proximity weight term, is defined in Section 2.2. Our proposed system uses WTA as disparity computation, because it is very simple algorithm.

Stereo vision system design The proposed system is composed of the stereo head and the stereo emulator. The stereo head includes a LVDS module, an LD/LED projection module and two CMOS sensors. Input from the two CMOS sensors is received in the form of 10-bit monochrome images at a resolution of 1280 x 720 pixels at a rate of 60fps. Image streams from the left & right cameras are transferred to the FPGA board through the LVDS module, which includes control signals. The stereo emulator is based on FPGA modules to obtain disparity maps from stereo pairs. The stereo emulator contains a deserialized module, a USB 3.0 controller and four FPGAs. The USB 3.0 controller transfers the results to the computer and is received parameters for FPGA. Figure 1 (a) represents our entire system. The hybrid system captures a pair of pattern images and an alternating pair of non-pattern images to evaluate disparity. The CMOS sensors are synchronized with LD/LED projection module, so that the images are obtained in operating time with the LD (pattern) in one frame and with the LED (non-pattern) in the next frame.

Implemeting stereo algorithm Figure 1 (b) is a block diagram of the hybrid stereo vision system. The stereo matching algorithm consists of four processing elements implemented in each single FPGA. The PrePE is first composed through rectification(based on the Caltech method) and image filtering(based on bilateral filtering). The MPE Left generates a left-referenced disparity that is implemented using the algorithm described in Section 2. Certain modifications are taken into account while implementing the system on the FPGA, and usually appear in the cost aggregation. One of the issues is whether the operations can be processed in a limited clock cycle. In order to solve this problem, we allow as much parallel processing, pipeline insertion and using LUT. Also, we configure to a power of two to use the data shifter instead of the divider. Another issue is whether to use the PF 4way in cost aggregation. When implementing the FPGA in cost aggregation using the PF, data from the entire image must be saved, and hence involves cost volume. If horizontal cost volume data is assumed to be 16-bit, 4-direction processing is needed, which requires at least 570MB. For the above reasons, other research[1, 4] implements FPGA in only the horizontal direction. However, this causes streak noise and thus increases the error due to disparity. Our system is configured to run 3way(including top-to-bottom) while maintaining consistent performance and reducing memory consumption. In this case, we use memories for two lines of aggregated cost and two lines of non-pattern images. In PostPE, we use left-right consistency check and weighted median filtering to eliminate error pixels from the results of disparity calculation. Sub-pixel estimation is implemented as a parabola fitting the costs and is calculated to divide 4 bits more than 256 steps that are separated by a maximum disparity. So, steps of final disparity are 4096(256x16).



(a) (b)

Figure 1: (a) The entire system with stereo head and stereo emulator (b) Block diagram of proposed stereo vision system

Algorithm		error(%)
Not using pattern image (Passive stereo matching)	PF 4way	61.67%
	MPF 4way	54.54%
Using pattern image (Hybrid system)	PF 2way(Horizontal)	11.39%
	MPF 3way for FPGA	3.80%

Table 1: Error pixel rate on non-occlusion regions.

Performance Evaluation We evaluate performance in two ways: quantitative evaluation using the ground truth and qualitative evaluation using Microsoft Kinect. The ground truth is created using a space-time stereo [3]. The performance of the system is then analyzed by using different cost aggregations and comparison between passive stereo matching and hybrid system. When applying the hybrid approach, the MPF has lesser streak noises than the conventional real-time two-way PF. We also see that the MPF 3way for FPGA performs just as well as the MPF 4way, except for the fact that the sharpness of the object is different.

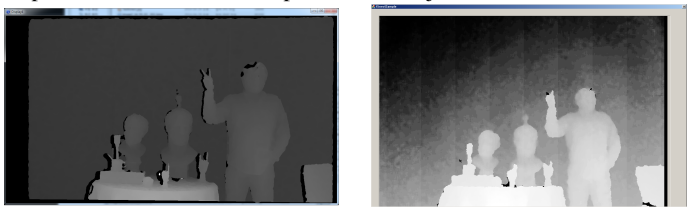


Figure 2: Comparison with Kinect, Left:proposed system, Right:Kinect

We proposed a hybrid stereo matching system that combines active and passive stereo vision. Using the active pattern, our system successfully detected disparity in untextured regions. Through comparison with other algorithms using the ground truth and with Microsoft Kinect, we found that our proposed system shows a significant improvement over current systems in processing untextured regions, and accurately calculates depth for a 1280 x 720 image at 60fps in indoor environments.

[1] A. Aysu, M. Sayinta, and C. Cigla. Low cost fpga design and implementation of a stereo matching system for 3d-tv applications. In *VLSI-SoC*, 2013.

[2] C. Cigla and A.A. Alatan. Efficient edge-preserving stereo matching. *ICCV Workshops*, 2011.

[3] J. Jeong, H. Shin, J. Chang, E. Lim, S. Choi, K. Yoon, and J. Cho. High-quality stereo depth map generation using infrared pattern projection. *ETRI Journal*, 2013.

[4] S. Mattoccia. Stereo vision algorithms for fpgas. In *CVPR Workshops*, 2013.

[5] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *SMBV*.

[6] Xun Sun, Xing Mei, shaohui Jiao, Mingcai Zhou, and Haitao Wang. Stereo matching with reliable disparity propagation. In *3DIMPVT*.

Board #57

Image Cosegmentation via Multi-task Learning

Qiang Zhang, Jiayu Zhou, Yilin Wang, Jieping Ye, Baoxin Li
qzhang53,jiayu.zhou,ywang370,jieping.ye,baoxin.li@asu.edu

Computer Science and Engineering
Arizona State Uni., Tempe, AZ, USA

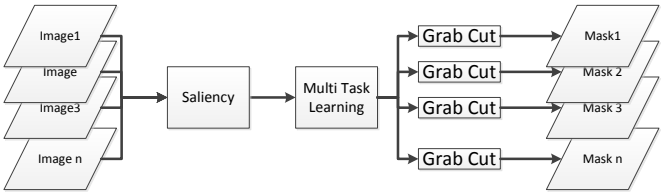


Figure 1: The overview of the proposed algorithm. We first extract the saliency map for the input images; according to the saliency map, we pick the seed regions to initialize the multi-task learning algorithm; and finally according to the output of multi-task learning algorithm, we use grab cut to obtain the final segmentation result.

1 Motivation

Image segmentation has been studied in computer vision for many years and yet it remains a challenging task. One major difficulty arises from the diversity of the foreground, which often results in ambiguity of background-foreground separation, especially when prior knowledge is missing. To overcome this difficulty, cosegmentation methods were proposed, where a set of images sharing some common foreground objects are segmented simultaneously. Different models have been employed for exploring such a prior of common foreground. In this paper, we propose to formulate the image cosegmentaion problem using a multi-task learning framework, where segmentation of each image is viewed as one task and the prior of shared foreground is modeled via the intrinsic relatedness among the tasks. Compared with other existing methods, the proposed approach is able to simultaneously segment more than two images with relatively low computational cost. The proposed formulation, with three different embodiments, is evaluated on two benchmark datasets, the CMU iCoseg dataset and the MSRC dataset, with comparison to leading existing methods. Experimental results demonstrate the effectiveness of the proposed method.

2 Proposed Method

An overview of the proposed method is illustrated in Fig 1. In experiments of this paper, we first over-segment the images into superpixels and then use them as basic units for subsequence processing. For obtaining the superpixels, we use SLIC and set the number of superpixels for each image to 200. For notations, we use $\mathbf{X}_i^j$ to represent the descriptor of the j_{th} superpixel in i_{th} image and y_i^j as its label.

Feature Extraction: for each superpixel, we extract the feature according to [15], which includes geometry measurements, color, texture and edges. The similarity measure of the superpixels is one of the most important component for image segmentation. For image cosegmentation, we need not only the similarities measure of the superpixels within each image, but also the similarities measure of superpixels cross different images. For the superpixels within each image, high similarity score is assigned to superpixels which are both spatially close and feature-wise similar. For the similarities of the superpixels cross images, we use nearest neighbor to find their correspondences.

$$A(i,j;p,q) = K(\mathbf{X}_i^j, \mathbf{X}_p^q) \times e^{\frac{-|loc(\mathbf{X}_i^j) - loc(\mathbf{X}_p^q)|_2^2}{2\sigma}} \text{ if } i = p \quad (1)$$

$$A(i,j;p,q) = K(\mathbf{X}_i^j, \mathbf{X}_p^q) \times \text{KNN}(i,j;p,q) \text{ if } i \neq p \quad (2)$$

Visual Cosaliency: recently visual cosaliency was proposed and utilized to initialize the image coegmentation algorithm. In the proposed cosaliency, a superpixel is cosalient, if it is not only salient in the corresponding image but also similar to salient superpixels of other images. After computing the cosaliency score s_i^j, we label the top 20% of the salient superpixel as the foreground and the bottom 70% ones as background to initialize the image cosegmentation algorithm.

$$s_i^j(t+1) = (1-\alpha)\hat{s}_i^j + \alpha \sum_{p,q:s_p^q(t)\geq\tau} s_p^q(t) \times A(i,j;p,q) \quad (3)$$

Mean	88.67%
$\ell_{2,1}$	87.81%
Low	88.33%

(a)

Mean	80.58%
$\ell_{2,1}$	80.87%
Low	81.20%

(b)

Table 1: (a) The result on iCoseg dataset. (b) The result on MSRC dataset.

Figure 2: Example of image segmentation on iCoseg dataset and MSRC dataset, where the green contour shows the segmentation results.

Multi-task learning: we formulate image cosegmentation as a multi-task learning problem, where the segmentation of each image is viewed as one task. Naturally, in this formulation the prior that common foreground objects are shared among the images is assumed to be captured by the intrinsic relatedness among the tasks. For developing a solution under this formulation, we focus on regularization-based modeling, due to its flexibility in incorporating existing computational models and supporting various assumptions of task relatedness. Especially we consider the following three multi-task learning assumptions:

1) the task parameters are drawn from the same distribution ("mean").

$$\{\mathbf{W}_i\} : \arg\min_{\{\mathbf{W}_i\}} f(\{\mathbf{W}_i\}) + \lambda \|\mathbf{W}_i - 1/N \sum_{j=1}^{N} \mathbf{W}_j\|_2^2. \quad (4)$$

2) the models of the tasks share a common low-rank subspace ("low").

$$\{\mathbf{W}_i\} : \arg\min_{\{\mathbf{W}_i\}} f(\{\mathbf{W}_i\}) + \lambda \|\mathbf{W}\|_* \quad (5)$$

3) the models of the tasks share the same subset of features ("$\ell_{2,1}$").

$$\{\mathbf{W}_i\} : \arg\min_{\{\mathbf{W}_i\}} f(\{\mathbf{W}_i\}) + \lambda \|\mathbf{W}\|_{2,1} \quad (6)$$

After we find the classifiers with the multi-task learning methods, we apply the classifiers to each superpixel of the images. The classifiers return a score within $[0,1]$ via logistic function. We then label a superpixel as background if its response is smaller than 80% of the mean response of all of the superpixels of all images, which will be used for initialization of Grabcut algorithm to obtain the final segmentation.

3 Experiment

We evaluate the proposed method on two widely-used datasets: CMU iCoseg (37 sets of images, 4 to 41 images per class) and MSRC (14 sets of images, around 30 images for each set). We compared with several existing methods, some of them being the current state-of-art. For performance metric, we compute the accuracy of the segmentation result over the manually labeled mask, which includes both foreground and background $p = \frac{\text{true postive+true negative}}{\text{area of image}}$.

Board #58

Geodesic Finite Mixture Models

Edgar Simo-Serra
esimo@iri.upc.edu
Carme Torras
torras@iri.upc.edu
Francesc Moreno-Noguer
fmoreno@iri.upc.edu

Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
08028, Barcelona, Spain

The use of Riemannian manifolds and their statistics has recently gained popularity in a wide range of applications involving non-linear data modeling. For instance, they have been used to model shape changes in the brain [1] and human motion [3]. In this work we tackle the problem of approximating the Probability Density Function (PDF) of a potentially large dataset that lies on a *known* Riemannian manifold. We address this by creating a completely data-driven algorithm consistent with the manifold, i.e., an algorithm that yields a PDF defined exclusively on the manifold.

In the proposed finite mixture model, we simultaneously consider multiple tangent spaces, distributed along the whole manifold as seen in Fig. 1. We draw inspiration on the unsupervised Expectation Maximization (EM) algorithm from [2], which given data lying in an Euclidean space, automatically computes the number of model components that Minimize a Message Length (MML) cost. By representing each component as a distribution on the tangent space at its corresponding mean on the manifold, we are then able to generalize the algorithm to Riemannian manifolds and at the same time mitigate the accuracy loss produced when using a single tangent space.

Given an input dataset, [2] starts by randomly initializing a large number of components. During the Maximization (M) step, the MML criterion is used to annihilate those components not well supported by the data. In addition, upon EM convergence, the least probable mixture component is also forcibly annihilated and the algorithm continues until a minimum number of components is reached.

In order to extend [2] to Riemannian manifolds, we define each mixture component as a normal distribution on its own tangent space $T_{\mu_k}\mathcal{M}$, with a mean μ_k and a concentration matrix $\Gamma_k = \Sigma_k^{-1}$:

$$p(x|\theta_k) \approx \mathcal{N}_{\mu_k}\left(0, \Sigma_k^{-1}\right)$$

where $\theta_k = (\mu_k, \Sigma_k)$. The mean μ_k is defined on the manifold $\mathcal{M}$, while the concentration matrix Γ_k is defined on the tangent space $T_{\mu_k}\mathcal{M}$ with the mean at the origin. Specifically, our algorithm proceeds as follows:

Let us assume we have K components after iteration $t-1$. Then, in the E-step we compute the *responsibility* that each component k takes for every sample x_i:

$$w_k^{(i)} = \frac{\alpha_k(t-1)p(x_i|\theta_k(t-1))}{\sum_{k=1}^{K}\alpha_k(t-1)p(x_i|\theta_k(t-1))},$$

for $k = 1,\ldots,K$ and $i = 1,\ldots,N$, and where $\alpha_k(t-1)$ are the relative weights of each component k.

In the M-step we update the weight α_k, the mean μ_k and covariance Σ_k for each of the components according to:

$$\alpha_k(t) = \frac{1}{N}\sum_i^N w_k^{(i)} = \frac{w_k}{N}, \qquad \mu_k(t) = \arg\min_p \sum_{i=1}^{N} d\left(\frac{N}{w_k} w_k^{(i)} x^{(i)},\ p\right)^2$$

$$\Sigma_k(t) = \frac{1}{w_k}\sum_{i=1}^{N}\left(\log_{\mu_k(t)}(x^{(i)})\right)\left(\log_{\mu_k(t)}(x^{(i)})\right)^\top w_k^{(i)}$$

where $d(\cdot,\cdot)$ is the geodesic distance between two points and $log_\mu(\cdot)$ is an operator that maps a point from the manifold $\mathcal{M}$ to the tangent space $T_p\mathcal{M}$ at point μ.

After each M-step, we eliminate the components whose accumulated responsibility w_k is below a threshold. A score for the remaining components based on MML is then computed. This EM process is repeated until convergence of the score or until reaching a minimum number of components K_{min}. If this number is not reached, the component with the least responsibility is eliminated and the EM process is repeated. Finally, the configuration with minimum score is retained, yielding a resulting distribution with the form

$$p(x|\theta) = \sum_{k=1}^{K}\alpha_k p(x|\theta_k)\ .$$

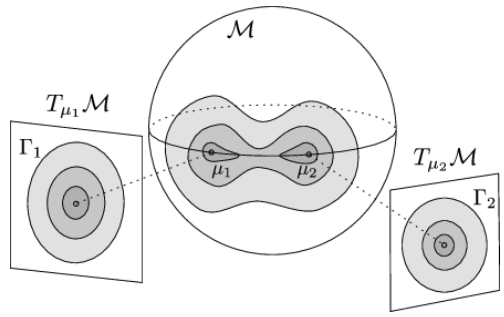
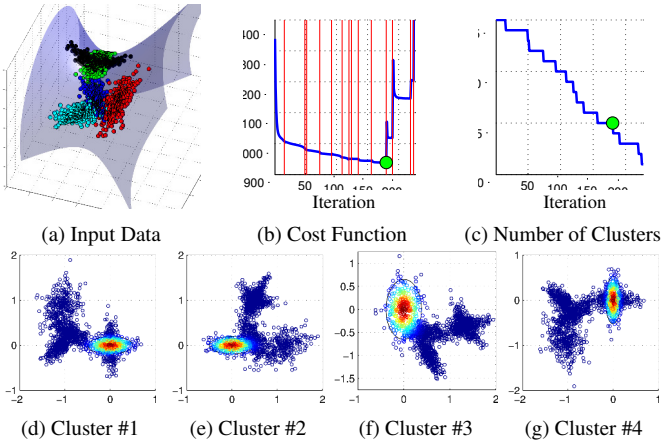


Figure 1: Illustration of the proposed mixture model approach. Each mixture component has its own tangent space, ensuring the consistency of the model while minimizing accuracy loss.



(a) Input Data (b) Cost Function (c) Number of Clusters

(d) Cluster #1 (e) Cluster #2 (f) Cluster #3 (g) Cluster #4

Figure 2: **Quadratic Surface Example.** (a) Section of the manifold with the input data. (b) Evolution of the cost function where vertical lines represent iterations in which a cluster is annihilated. The optimal mixture is marked with a green dot. (c) Evolution of the number of clusters. Some of the clusters from the solution are shown in (d) to (g).

We validate our method by providing extensive results on both synthetic and real examples. In particular, we show results on synthetic examples of a sphere and a quadric surface (see Fig. 2), and on a large and complex dataset of human poses, where the proposed model is used as a regression tool for hypothesizing the geometry of occluded parts of the body. We show that our approach outperforms the traditionally used Euclidean Gaussian Mixture Model, von Mises distributions and approaches using a single tangent space.

[1] B. C. Davis, E. Bullitt, P. T. Fletcher, and S. Joshi. Population Shape Regression from Random Design Data. In *International Conference on Computer Vision*, 2007.

[2] M. Figueiredo and A. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[3] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold Valued Statistics, Exact Principal Geodesic Analysis and the Effect of Linear Approximations. In *European Conference on Computer Vision*, 2010.

Board #59

Multiple Object Tracking Using Local Motion Patterns

Mehrsan Javan Roshtkhari
http://www.cim.mcgill.ca/~javan
Martin D. Levine
http://www.cim.mcgill.ca/~levine

Center For Intelligent Machines
Department of Electrical and Computer Engineering
McGill University
Montreal, QC, Canada

Object tracking is, perhaps, the most fundamental task for any high-level video content analysis system. Decades of research on this topic have produced a diverse set of approaches and a rich collection of tracking algorithms. Most of the reported algorithms are based on object detection followed by a data association algorithm. Thus a key assumption is that a reliable object detection algorithm exists [1, 5]. These methods use the detection response to construct an object trajectory. This is accomplished by using data association based on either the detection responses or a set of short tracks called tracklets that are associated with each detected object [1]. Subsequently, data association links these tracklets into multi-frame trajectories. On the other hand, there are other tracking algorithms, which are based on local spatio-temporal motion patterns in the scene. More closely related to our approach are those that construct motion models for the moving objects without performing any detection [2].

In this paper we concentrate on creating long-term trajectories for unknown moving objects by using a model-free tracking algorithm. As opposed to the tracking-by-detection algorithms [5], no object detection is involved. Each individual object is tracked only by modeling the temporal relationship between sequentially occurring local motion patterns. This is achieved by constructing two sets of initial tracks that code local and global motion patterns in videos. These local motion patterns are obtained by analyzing spatially and temporally varying structures in videos [3, 4].

Initially, the video is densely sampled, spatio-temporal video volumes (STVs) are constructed, and similar ones are grouped to reduce the dimension of the search space. This is called the low-level codebook, $\mathbf{C}^{\mathcal{L}}$. Then, a large contextual region containing many STVs (in space and time) around each pixel is examined and their compositional relationships are approximated using a probabilistic framework. They are then employed to form yet another codebook, called the high-level codebook, $\mathbf{C}^{\mathcal{H}}$. Therefore, two codewords are assigned to each pixel, one from the low level and the other from the high level codebook. By examining pairs of sequential video frames, the matching codewords for each video pixel are transitively linked into distinct tracks, whose total number is unknown a priori and which we will refer to as linklets. The linking process is separately performed for both codebooks. This is done under the hard constraint that no two linklets may share the same pixel at the same time, i.e. the assigned codewords. The end result at this step is two sets of independent linklets obtained from the low- and high-level codebooks.

Subsequently, a set of sparse tracks, referred to as tracklets in the literature, are produced by grouping the linklets that indicate similar motion patterns (see Figure 1). This produces two sets of independent tracklets, referred to as low- and high-level tracklets, $\mathbf{T}^{\mathcal{L}}$ and $\mathbf{T}^{\mathcal{H}}$, respectively. Given the resulting tracklets, high-level trajectories can be generated by linking them in space and time. We achieve this by formulating the data association required as a maximum a posteriori (MAP) problem and solve it with the Markov Chain Monte Carlo Data Association (MCMCDA) algorithm. The observations are taken to be the constructed tracklets, $\mathcal{O} = \left\{\mathbf{T}^{\mathcal{L}}, \mathbf{T}^{\mathcal{H}}\right\}$. Let $\mathbf{\Gamma}$ be a tracklet association result, which is a set of trajectories, $\Gamma_k \in \mathbf{\Gamma}$. Γ_k is defined as a set of the connected observations which is a subset of all observations, $\Gamma_k = \left\{T_i^{\mathcal{L}}, T_j^{\mathcal{H}}\right\} \subseteq \mathcal{O}$. The goal is to find the most probable set of object trajectories, $\mathbf{\Gamma}$, which is formulated as a MAP problem:

$$\mathbf{\Gamma}^* = \arg\max_{\mathbf{\Gamma}} P(\mathbf{\Gamma}|\mathcal{O}) = \arg\max_{\mathbf{\Gamma}} P(\mathcal{O}|\mathbf{\Gamma})\,P(\mathbf{\Gamma}) \tag{1}$$

The likelihood, $P(\mathcal{O}|\mathbf{\Gamma})$ indicates how well a set of trajectories matches the observations and the prior, $P(\mathbf{\Gamma})$ indicates how correct the data association is. By assuming that the likelihoods of the tracklets are conditionally independent, we can rewrite the likelihood, $P(\mathcal{O}|\mathbf{\Gamma})$, in (1) as follows:

$$P(\mathcal{O}|\mathbf{\Gamma}) = \prod_{\substack{T_i^{\mathcal{L}} \in \mathbf{T}^{\mathcal{L}} \\ T_j^{\mathcal{H}} \in \mathbf{T}^{\mathcal{H}}}} P\left(T_i^{\mathcal{L}}, T_j^{\mathcal{H}}|\mathbf{\Gamma}\right) \prod_{\Gamma_k \in \mathbf{\Gamma}} P(\Gamma_k) \tag{2}$$

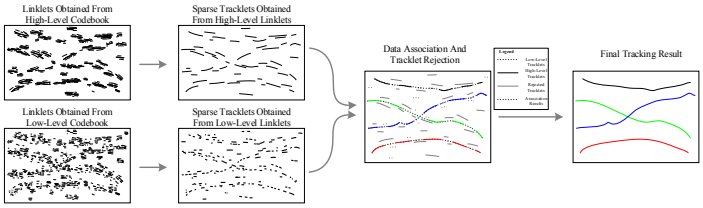


Figure 1: The goal is to estimate the trajectory of the moving objects in the video without invoking object detection. Initially two sets of linklets are constructed by chaining; the low-level considers small window fragments, while the high-level analyzes a larger region in order to impose a contextual influence. They are obtained by exploiting an activity understanding system [3, 4]. The resultant tracks (chains) are filtered and replaced by a set of sparse representative tracks, the so-called tracklets. Longer trajectories are then generated by using the Markov Chain Monte Carlo Data Association (MCMCDA) algorithm to solve the Maximum A Posteriori (MAP) problem using tracklet affinities. Thus this procedure uses low-level tracklets to connect high-level tracklets when there is a discontinuity in motion or time.

We adopt Markov Chain Monte Carlo Data Association (MCMCDA) to estimate an initially unspecified number of trajectories. To this end, we formulate the tracklet association problem as a Maximum A Posteriori (MAP) problem to produce a chain of tracklets. Data association is accomplished by considering temporal continuity and motion consistency of both the low- and high-level tracklets, with the additional option of rejecting irrelevant tracklets. The final output of the data association algorithm is a partition of the set of tracklets such that those belonging to each individual object have been grouped together. Implementation of this method is described in the paper, as are the details of the all other parts of this algorithm.

Although our algorithm possesses no information regarding either an object's color pattern or a human body model, it achieves promising results on challenging data sets. The results indicate that although the correct detections we obtain with our algorithm are comparable to the state of the art, they include more false positives. Perhaps one can expect this, since no object detection is employed in our algorithm. Recall that the scene observations that we use are motion descriptors and do not incorporate object appearance, as do object-centric trackers. As stated in the paper, the major drawback of our algorithm is the number of false positives and some problems in maintaining the trajectory identity when objects have similar shape and motion.

[1] Huang Chang, Li Yuan, and R. Nevatia. Multiple target tracking by learning-based hierarchical association of detection responses. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35 (4):898–910, 2013.

[2] L. Kratz and K. Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):987–1002, 2012.

[3] Mehrsan Javan Roshtkhari and Martin D. Levine. Online dominant and anomalous behavior detection in videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2609–2616, 2013.

[4] Mehrsan Javan Roshtkhari and Martin D. Levine. Human activity recognition in videos using a single example. *Image and Vision Computing*, 31(11):864–876, 2013.

[5] Bo Yang and Ramakant Nevatia. Multi-target tracking by online learning a crf model of appearance and motion patterns. *International Journal of Computer Vision*, 107(2):203–217, 2014.

Compact Video Code and Its Application to Robust Face Retrieval in TV-Series

Yan Li
yan.li@vipl.ict.ac.cn
Ruiping Wang
wangruiping@ict.ac.cn
Zhen Cui
zhen.cui@vipl.ict.ac.cn
Shiguang Shan
sgshan@ict.ac.cn
Xilin Chen
xlchen@ict.ac.cn

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

Problem: We address the problem of video face retrieval in TV-Series which searches video clips based on the presence of specific character, given one video clip of his/hers, see Figure 1. This is tremendously challenging because on one hand, faces in TV-Series are captured in largely uncontrolled conditions with complex appearance variations, and on the other hand retrieval task typically needs efficient representation with low time and space complexity.

Our Method: To solve this problem, we propose a compact and discriminative representation for the huge body of video data, named Compact Video Code (CVC). Our method first models the video clip by its sample (i.e., frame) covariance matrix to capture the video data variations in a statistical manner. Let $F = [f_1, f_2, ..., f_n]$ be the data matrix of a video clip with n frames, where $f_i \in \mathbb{R}^d$ denotes the i^{th} frame with d-dimensional feature. We represent the video clip with the $d \times d$ sample covariance matrix:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})(f_i - \bar{f})^T, \tag{1}$$

where $\bar{f}$ is the mean of all frames in the video clip. It is well known that the nonsingular covariance matrices do not lie in a Euclidean space but on a Riemannian manifold $\mathcal{M}$. However, it is not trivial to learn a binary code encoder on the manifold since typical code learning methods are devoted to operating in Euclidean space. So here we utilize the Log-Euclidean Distance (LED) to bridge the gap between Riemannian manifold and Euclidean space as in [2]:

$$d_{LED}(C_1, C_2) = \|\log(C_1) - \log(C_2)\|_F. \tag{2}$$

To incorporate discriminative information and obtain more compact video signature, the high-dimensional covariance matrix is further encoded as a much lower-dimensional binary vector, which finally yields the proposed CVC. Specifically, each bit of the code, i.e., each dimension of the binary vector, is produced via supervised learning in a max margin framework [1], which aims to make a balance between the ***discriminability*** and ***stability*** of the code.

Discriminability: We characterize the discriminability into two parts: within class compactness (S_W) and between class separability (S_B).

$$S_W = \sum_{c\in\{1:M\}}\sum_{m,n\in c} dis(b_m, b_n), \tag{3}$$

$$S_B = \sum_{\substack{c_1\in\{1:M\}\\ p\in c_1}}\sum_{\substack{c_2\in\{1:M\}\\ c_1\neq c_2, q\in c_2}} dis(b_p, b_q), \tag{4}$$

where M is the total number of training classes, $dis(\cdot)$ is the distance measurement of binary codes in Hamming space, $b \in \{-1,1\}^{N\times K}$ denotes the binary codes of training instances, and N and K denote the total number of training instances and the length of binary code respectively. Thus, to implement a strong discrimination, we should minimize the following energy function E_{disc}.

$$E_{disc} = S_W - \lambda_1 S_B. \tag{5}$$

Stability: To make better stability, we build the K hyperplanes by using SVM, and each generates one bit of the binary code. Concretely, we

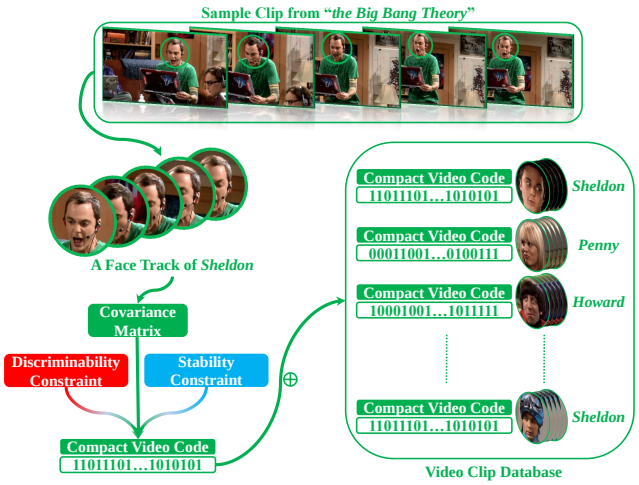


Figure 1: Illustration of the proposed method. Given a video clip of one character as query, we extract the proposed Compact Video Code (CVC) to represent it and use Hamming distance to retrieve video clips containing the specific character in database, which are also encoded in the form of CVC representations.

denote the k^{th} hyperplane by ω^k $(k = 1, \cdots, K)$, and the energy function can be formulated as follow.

$$E_{stab} = \frac{1}{2}\sum_{k\in\{1:K\}} {\omega^k}^T \omega^k + \lambda_2 \sum_{\substack{k\in\{1:K\}\\ i\in\{1:N\}}} \max(1 - b_i^k({\omega^k}^T x_i), 0), \tag{6}$$

where x_i denotes the input feature, b_i^k indicates in which side of the k^{th} hyperplane the i^{th} training instance lies, and λ_2 balances the empirical training error and the hyperplane margin.

After the above analysis, we can reach the final objective function by combining Eqn. (5) and Eqn. (6) to simultaneously consider the discriminability and stability of the target binary code:

$$\min_{b,\omega} \quad E_{disc} + E_{stab}. \tag{7}$$

Since the objective function is non-convex, in practice we independently optimize each individual component to iteratively update b and ω, where an efficient subgradient descend method proposed in [1] with compute-complexity $O(NK)$ was utilized to optimize b.

Experiment: Face retrieval experiments on two challenging TV-Series video databases have demonstrated the competitiveness of the proposed CVC over state-of-the-art retrieval methods. In addition, as a general video matching algorithm, our method is also evaluated in traditional video face recognition task on a standard Internet database, i.e., *YouTube Celebrities*, showing its quite promising performance by using an extremely compact code with only 128 bits.

[1] Mohammad Rastegari, Ali Farhadi, and David Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, pages 876–889. Springer, 2012.

[2] Ruiping Wang, Huimin Guo, Larry S. Davis, and Qionghai Dai. Covariance discriminative learning: a natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503. IEEE, 2012.

Board #61

Biologically Inspired Online Learning of Visual Autonomous Driving

Kristoffer Öfjäll
kristoffer.ofjall@liu.se
Michael Felsberg
michael.felsberg@liu.se

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University
Linköping, Sweden

While autonomously driving systems accumulate more and more sensors as well as highly specialized visual features and engineered solutions, the human visual system provides evidence that visual input and simple low level image features are sufficient for successful driving. In this paper we propose extensions (non-linear update and coherence weighting) to one of the simplest biologically inspired learning schemes (Hebbian learning). We show that this is sufficient for online learning of visual autonomous driving, where the system learns to directly map low level image features to control signals. After the initial training period, the system seamlessly continues autonomously. This extended Hebbian algorithm, qHebb, has constant bounds on time and memory complexity for training and evaluation, independent of the number of training samples presented to the system. Further, the proposed algorithm compares favorably to state of the art engineered batch learning algorithms in a visual head pose prediction challenge, where the algorithms can be more thoroughly evaluated.

The input layer of the system consists of a single gray-scale camera and a generic, holistic representation of the whole visual field, using visual Gist [4]. No information regarding what kind of track or what kind of visual features that define the track is provided in advance. The system is supposed to learn features of the track together with correct driving behavior from the visual Gist features and the manual control signals provided during the initial training phase, see Fig. 1. Table 1 summarizes previous and present approaches to this task.

The proposed approach is based on the channel representation [2] and associative learning [3]. The channel representation of a scalar entity is a coefficient vector similar to a soft histogram (Fig. 2). A corresponding representation in biological systems is the population coding of e.g. orientation in the visual field. An associative mapping $\mathbf{y} = \mathbf{C}\mathbf{x}$ is learned, relating the channel representation, $\mathbf{x}$, of each image from the camera to the channel representation, $\mathbf{y}$, of the steering signal. Although the mapping is linear in the channel domain, non-linear relations can be represented between the original domains. Our two main contributions are an online learning rule for $\mathbf{C}$ with decoupled learning and forgetting rates, and a weighted variant of $\mathbf{y} = \mathbf{C}\mathbf{x}$, where each coefficient in $\mathbf{x}$ is weighted with the specificity with which it predicts the control signal (encoded in $\mathbf{y}$).

Method	Online Driving	Training Data Proc. Speed
CN [5]	No	Days (batch)
RFR [1]	Yes	Hours (batch)
Assoc. Hebb	No	Video rate (online)
Proposed	Yes	Video rate (online)

Table 1: Summary of approaches for visual autonomous driving, including Random Forest Regression (RFR) and Convolutional Networks (CN).

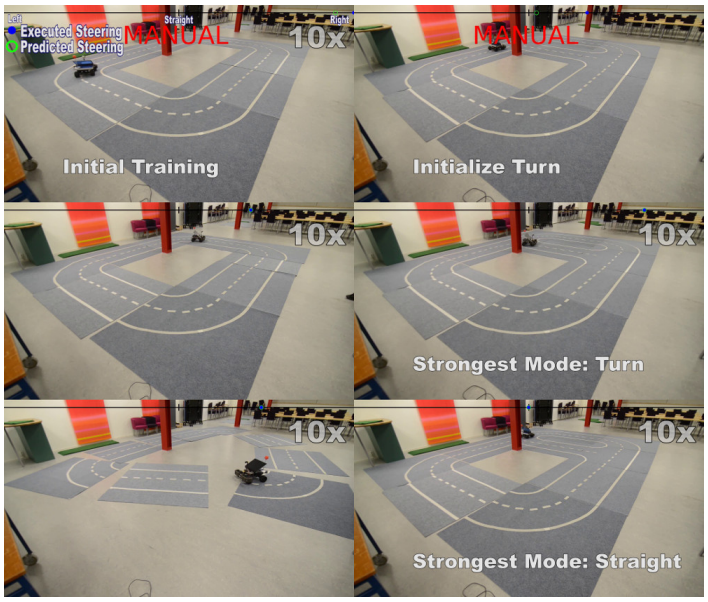


Figure 1: An autonomous vehicle learns a mapping from general low level image features to steering signal from demonstration by remotely operating the vehicle during parts of the first lap. When the remote override is released, the vehicle seamlessly continues autonomously. After the training period, the system showed robustness towards reconfiguration of the track (left column). After introducing an intersection, the vehicle can be forced to take the other way by a short application of manual control. After this new training data is acquired, both modes, *going straight* and *turning*, are present in the system. When the vehicle reaches the intersection again, it either goes straight or turns depending on which mode happen to get a stronger response (right column). However, the modes are never mixed, which would be noticed by the vehicle making half a turn at the intersection. The figure shows a selection of frames from the supplementary video available at *http://users.isy.liu.se/cvl/ofjall/bmvc2014.mp4*.

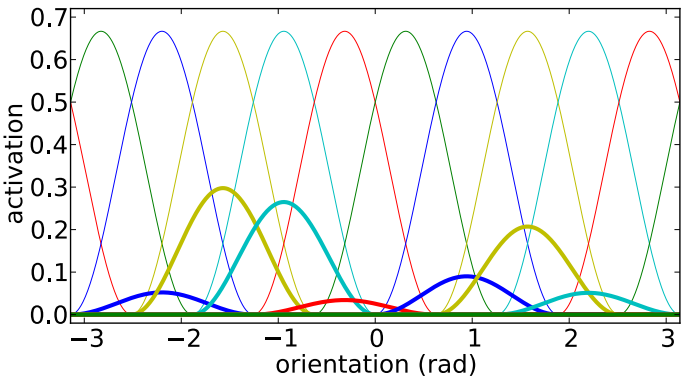


Figure 2: Illustration of a channel representation of a distribution of orientation data. The thin plots indicate the kernel functions (the channels) and the bold lines illustrate the corresponding channel coefficients as weighted kernel functions. The represented distribution has two modes.

[1] Liam Ellis, Nicolas Pugeault, Kristoffer Öfjäll, Johan Hedborg, Richard Bowden, and Michael Felsberg. Autonomous navigation and sign detector learning. In *Robot Vision (WORV), 2013 IEEE Workshop on*, pages 144–151. IEEE, 2013.

[2] G. H. Granlund. An Associative Perception-Action Structure Using a Localized Space Variant Information Representation. In *Proceedings of Algebraic Frames for the Perception-Action Cycle (AFPAC)*, Germany, September 2000.

[3] Björn Johansson. *Low Level Operations and Learning in Computer Vision*. PhD thesis, Linköping University, Computer Vision, The Institute of Technology, 2004.

[4] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[5] Maria Schmiterlöw. Autonomous path following using convolutional networks. Master's thesis, Linköping University, Computer Vision, The Institute of Technology, 2012.

Online segmentation and classification of modeled actions performed in the context of unmodeled ones

Dimitrios I. Kosmopoulos[1,3]
dkosmo@ics.forth.gr
Konstantinos Papoutsakis[1,2]
papoutsa@ics.forth.gr
Antonis A. Argyros[1,2]
argyros@ics.forth.gr

[1]Institute of Computer Science
FORTH, Greece
[2]Computer Science Department
University of Crete, Greece
[3]Dept. of Informatics Engineering
Technological Educational Institute
Crete, Greece

In this work we deal with the problem of online segmentation and classification of visually observable actions, i.e., we have to provide labels given the fact that the visual observations arrive stream-wise on a sequential fashion and we need to decide on the label shortly after they are received, without having available the full sequence.

The video segmentation has been traditionally treated separately from the classification step, however, these two problems are correlated and can be better handled considering simultaneously the low level cues and the high level models representing the candidate classes. Generative models have been used extensively given their ability to build probabilistic models of actions and provide the posterior of assigning labels to observations. Alternatively, discriminative models better predict the conditional probability of the states given the observed features. As a result, several researchers have investigated the use of discriminative models of actions such as CRFs, SVMs [2] or random forests [1]. However, the discriminative models are not without problems, since they cannot easily handle unknown actions, since they were not part of their optimisation process.

In this paper, we show how we seek to mitigate that limitation, by employing a discriminative framework for online simultaneous segmentation and classification of visual actions, which deals effectively with unknown sequences that may interrupt the known sequential patterns. Our framework comprises of two main components: (a) a Hough transform to vote in a 3D space for the begin and end points and the label of the segmented part of the input stream. An SVM is used to model each class and to suggest putative labeled segments on the timeline. (b) A dynamic programming algorithm to identify the most plausible segments among the putative ones, by maximising an objective function for label assignment in linear time.

Hypotheses generation via discriminative voting. In the proposed discriminative voting framework we seek to identify simultaneously (a) the instances of classes C of sub-sequences in time series data, (b) the location $\mathbf{x}$ of the class-specific subsequence, in other words the begin and the end time point in the data. It is inspired by the framework presented in [3], which dealt with Hough transform based object detection.

Let $\mathbf{f}_t$ denote the feature vector observed at time instance t, while $S(C,\mathbf{x})$ denotes the score of class C at a location $\mathbf{x}$. The implicit model framework obtains the overall score $S(C,\mathbf{x})$ by adding up the individual probabilities $p(C,\mathbf{x},\mathbf{f}_t,l_t)$ over all observations within a time window l_t.

We define M action primitives, which result from clustering of the visual observation vectors $\mathbf{f}_t$, using GMMs to represent the distributions of the observation vectors. Let P_i denote the i-th action primitive. Assuming a uniform prior over features and time locations and marginalizing over the primitive entries, we derive:

$$\begin{aligned} S(C,x) &= \sum_i p(C|P_i) \sum_t p(P_i|\mathbf{f}_t) p(\mathbf{x}|C,P_i,l_t) \\ &= \sum_i w_i \times a_i(\mathbf{x}) = W_c^T A(\mathbf{x}) \end{aligned} \quad (1)$$

We can use maximum margin optimisation, if we observe that the score $S(C,\mathbf{x})$ is a linear function of $p(C|P_i)$, where $A^T = [a_1 a_2 ... a_M]$, is noted as the activation vector and a_i is given by:

$$a_i(\mathbf{x}) = \sum_t p(\mathbf{x}|C,P_i,l_t) p(P_i|\mathbf{f}_t) \quad (2)$$

The weights W_c^T are class-specific and we notice that they can be optimised in a discriminative fashion to maximise the score for correct segmentations and labels. Given the labels $S(C,\mathbf{x}_i)$ and the respective $A(\mathbf{x}_i)$ we calculate the weights W_c using multiple one-versus-all binary SVM settings. In *testing* we vote in the 3D space using Eq.(1) and then we apply the SVMs in a sliding time window to get the putative segments,

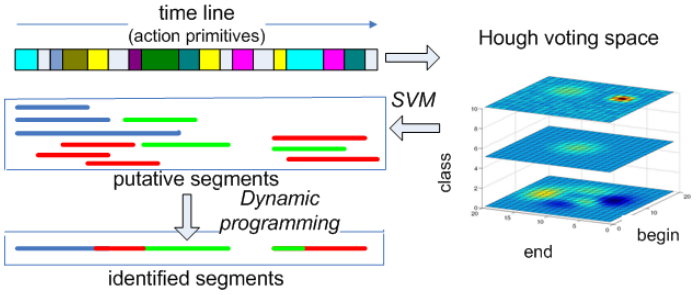


Figure 1: Overview of the proposed method: The action primitives span vote in a 3D Hough voting space (begin-end-class). The SVM receives the votes and suggests the putative segments.The final solution is computed by maximising an objective function via dynamic programming.

considering only the segments that collected enough votes. An additional evaluation step is normally applied to eliminate some false positives using a likelihood-based objective function. An illustrative example of the proposed hypotheses generation process is shown in Fig.1

Hypotheses evaluation via dynamic programming. We merge the proposed K putative segments that may overlap and have the same label. Assuming only one label for each time slot, we propose a variation of the Viterbi algorithm for linear-cost label assignment with regard to the number of input frames based on the likelihood δ_t, which is calculated after the optimal assignment of time instances to classes. The optimal sequence of classes for a time segment t=1..T, which contains overlapping candidate segments of different labels is given by the path $\psi_t = C_1, C_2, ..., C_t$, which is calculated based on dynamic programming.

Experimental Evaluation. The performance of the proposed method was evaluated on synthetic as well as on real data for action recognition (Weizmann and Berkeley MHAD). Actions were provided as segments. For the purpose of identifying actions in continuous data we concatenated those videos. We compare our method against two state of the art methods, [2] and [4], that do online segmentation like our method does. The proposed approach is of comparable accuracy to the state of the art for online stream segmentation and classification and performs considerably better in the presence of previously unseen actions.

Conclusions. Our work proposed a new framework for simultaneous segmentation and classification of sequential data interrupted by unknown actions and we have applied it on synthetic and visual action streams. Under a "closed world" assumption, our method performed similarly or better than the competing discriminative methods. When the actions of interest were interrupted by previously unseen actions our method was still able to classify them and detect the unknown ones. To knowledge, our discriminative method is the first one for online simultaneous segmentation and classification having this property.

[1] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Trans. on PAMI*, 33(11):2188–2202, November 2011.

[2] M. Hoai, Z.Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *IEEE CVPR*, 2011.

[3] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *IEEE CVPR*, 2009.

[4] Q. Shi, Li Wang, Li Cheng, and A. Smola. Discriminative human action segmentation and recognition using semi-markov model. In *IEEE CVPR*, 2008.

Board #63

Adaptive Multi-Level Region Merging for Salient Object Detection

Keren Fu[1,2]
fkrsuper@sjtu.edu.cn,keren@chalmers.se
Chen Gong[1]
goodgongchen@sjtu.edu.cn
Yixiao Yun[2]
yixiao@chalmers.se
Yijun Li[1]
leexiaoju@sjtu.edu.cn
Irene Yu-Hua Gu[2]
irenegu@chalmers.se
Jie Yang[1]
jieyang@sjtu.edu.cn
Jingyi Yu[3]
yu@eecis.udel.edu

[1] Institute of Image Processing and Pattern Recognition
Shanghai Jiao Tong University
Shanghai, P.R. China

[2] Department of Signals and Systems
Chalmers University of Technology
Gothenburg, Sweden

[3] University of Delaware
Newark, USA

Salient object detection is a long-standing problem in computer vision and plays a critical role in understanding the mechanism of human visual attention. In applications that require object-level prior (e.g. image retargeting), it is desirable that saliency detection highlights holistic objects. Lately over-segmentation techniques such as SLIC superpixel [6], Mean-shift [1], and graph-based [3] segmentations are popular among saliency detection due to their usefulness on eliminating background noise and reducing computation cost. However, individual small segments provide little information about global contents. Such schemes have limited capability on modeling global perceptual phenomena. Fig.1 shows a typical example. The entire flower tends to be perceived as a single entity by human visual system. It is easily imagined that saliency computation with the help of coarse segmentation is conducive to highlighting entire object while suppressing background. As it is important to control segmentation level to reflect proper image content, more recent approach benefits from multi-scale strategies to compute saliency on both coarse and fine scales with fusion [4]. [4] merges a region to its neighbor region if it is smaller than pre-defined sizes. The underlying problem may be that scale parameters in [4] are crucial to performance. A salient region may not appear in the proper level if it is smaller than the defined size. On the other hand, large background regions with close colors may not be merged together if they are larger than the defined size.

In this paper we propose an alternative solution, namely by quantifying *contour strength* to generate varied levels. Compared to [4], we use edge/contour strength and a globalization technique during merging. Our contributions include:

1. Develop an adaptive merging strategy for salient object detection rather than using several fixed "scales". Our method generates intrinsic optimal "scales" when the merging continues.

2. Incorporate additional global information by graph-based spectral decomposition to enhance salient contours. It is useful in salient object rendering.

3. Performance obtained is similar to other state-of-the-art methods even though simple region saliency measurements are adopted for each region.

As shown in Fig.2, our framework first performs over-segmentation on an input image by using SLIC superpixels [6], from which merging begins. To acquire holistic contour of salient objects as the merging process proceeds, we propose a modified graph-based merging scheme inspired by [3] which sets out to merge regions by quantifying a pre-defined region comparison criterion. Specifically before merging starts, a globalization

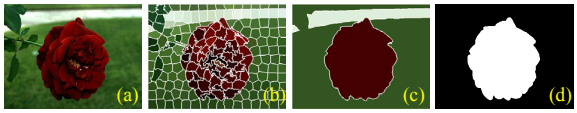


Figure 1: Multi-level segmentation for salient object detection. (a) shows a sample image from MSRA-1000 dataset [5]. (b) Over-segmentation using superpixels destroys the semantic content such as the flower. (c) A coarse segmentation derives from (b) maintains semantic holism. (d) object mask (ground truth).

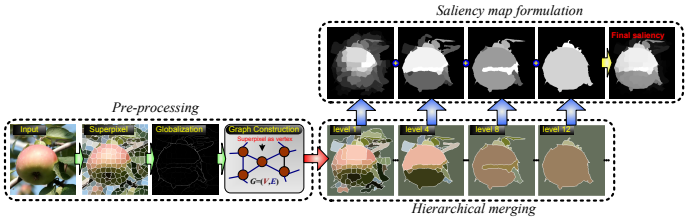


Figure 2: The processing pipeline of our approach.

procedure is proposed and conducted to pop out salient contours whereas suppress background clutter (Fig.2). At each level, we formulate an intermediate saliency map based on several simple region saliency measurements. Finally a salient object will be enhanced by summing across-level saliency maps (Fig.2).

Let initial SLIC superpixels be R_i^0, $i = 1,2,...,N$. A graph $G = (V,E)$ is defined where vertices V are superpixels, and E are graph edges. Let $R^l = \{R_1^l, R_2^l, ...\}$ be a partition of V in the lth level and $R_k^l \in R^l$ corresponds to its kth part (namely region). With the constructed edge E, a criterion D is defined to measure the pairwise difference of two regions R_i^l, R_j^l as:

$$D_{ij}^l = D(R_i^l, R_j^l) = \text{mean}_{v_k \in R_i^l, v_m \in R_j^l, e_{km} \in E}\{e_{km}\} \quad (1)$$

where "mean" is averaging operation over graph edges connecting R_i^l and R_j^l. In order to adapt merging to "large" differences (strong edges), we define a threshold Th to control the bandwidth of D_{ij}^l: at level l, we fuse two components R_i^l, R_j^l in R^l if their difference $D_{ij}^l \leq Th$. Suppose $R_i^l, R_j^l, R_k^l, ...$ are regions that have been merged into one larger region R_{new}^l at this level, we then update $R^l \leftarrow (R^l / \{R_i^l, R_j^l, R_k^l, ...\}) \bigcup R_{new}^l$ ("/" and "$\bigcup$" are set operation), where R_{new}^l is the newly generated region. At next level $l+1$, Th is increased as $Th \leftarrow Th + T_s$ where T_s is a step length. In graph edge construction (i.e. E), a globalization procedure is proposed inspired by a contour detector *gPb* [2]. The technique attempts to achieve area completion by solving the eigen-problem on the local affinity matrix. This operation also meets the Gestalt psychological laws properties [7, 8] i.e. *closure* and *connectivity* based on which human perceive figures.

To show the effectiveness of the proposed region merging and integration scheme, each merged region is just evaluated using several simple region saliency measurements. Even though like this, we show the proposed method already can achieve competitive results against the best methods among the state-of-the-art.

[1] D. Comaniciu et al. Mean shift: a robust approach toward feature space analysis. *TPAMI*, 24(5):603–619, 2002.

[2] P. Arbelaez et al. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5): 898–916, 2010.

[3] P. Felzenszwalb et al. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[4] Q. Yan et al. Hierarchical saliency detection. In *CVPR*, 2013.

[5] R. Achanta et al. Frequency-tuned salient region detection. In *CVPR*, 2009.

[6] R. Achanta et al. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012.

[7] K. Koffka. Principles of gestalt psychology. 1935.

[8] S. Palmer. Vision science: Photons to phenomenology. *The MIT press*, 1999.

Board #64

Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval

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

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

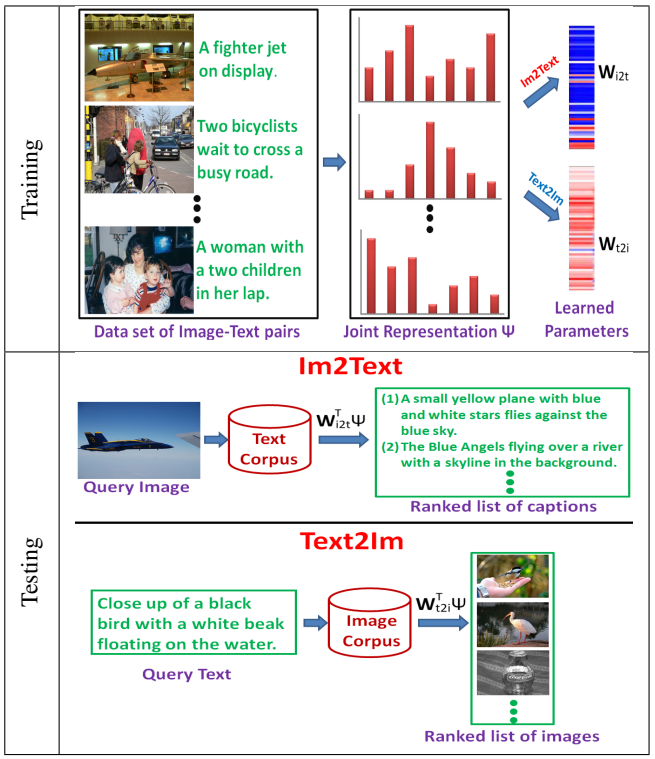


Figure 1: While training, given a dataset consisting of pairs of images and corresponding texts (here captions), we learn models for the two tasks (Im2Text and Text2Im) using a joint image-text representation. While testing for Im2Text, given a query image, we perform retrieval on a collection of only textual samples using the learned model. Similarly, for Text2Im, given a query text, retrieval is performed on a database consisting only of images.

To automatically describe image content using text is one of the challenging and interesting research problems in computer vision. A complementary problem to this is to automatically associate semantically relevant image(s) given a piece of text, and is commonly referred as the image retrieval task. In this work, we address the problem of learning bilateral associations between visual and textual data. We study two complementary tasks: (i) predicting text(s) given an image ("Im2Text"), and (ii) predicting image(s) given a piece of text ("Text2Im"). While several existing methods (*e.g.*, [1]) assume presence of data from both the modalities during the testing phase, the motivation of this work is similar to the few known works (*e.g.*, [2]) that do not make such assumption. This means that for Im2Text, given a query image, our method retrieves a ranked list of semantically relevant texts from a plain text-corpus that has no associated images. Similarly, for Text2Im, given a query text, it retrieves a ranked list of images from an independent image collection without any associated textual meta-data. The major contributions of this work are:
(1) We propose a novel Structural SVM based unified framework for both these tasks. We use vector representations for both visual (image) and textual data that are based on probability distributions over latent topics. From these, we form a joint feature vector using tensor product of input and output representations. Because the output data is represented in the form of a vector, we use Manhattan (M) and Euclidean (E) distances as our loss functions. As the proposed approach performs the two complementary tasks (Im2Text and Text2Im) under a single unified framework, we refer to it as Bilateral Image-Text Retrieval (or BITR). Figure 1 explains the gist of our framework.
(2) We examine generalization of different methods across datasets when textual data is in the form of captions. For this, we learn models from one dataset, and perform retrieval on other. To our best knowledge, ours is the first such study in this domain.

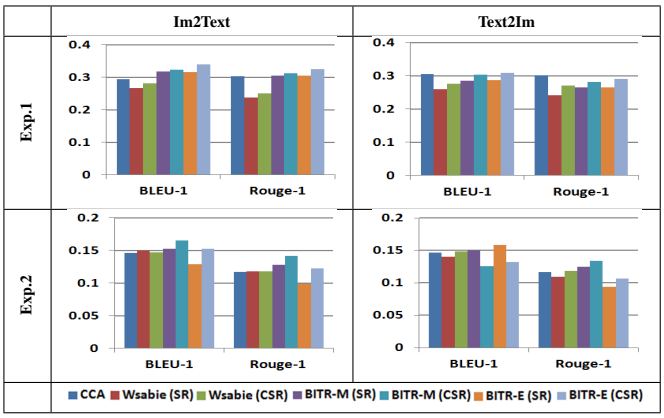


Figure 2: Results on IAPR TC-12 dataset for within-dataset (top) and cross-dataset (bottom) image-caption retrieval.

We conduct experiments on three datasets (UIUC Pascal Sentence dataset, IAPR TC-12 benchmark, and SBU-Captioned Photo dataset), and compare our approach with WSABIE [3] and CCA. These are two well-known methods that can scale to large datasets and have been shown to work well for learning cross-modal associations. While CCA based methods have been used previously under such settings [2], WSABIE was originally proposed for the task of label-ranking and hence can not be directly applied for captions. We do this by adapting it for captions, the details of which are provided in the supplementary file. We consider two types of representations for visual and textual data. The first representation captures high-level semantics of data in the form of unimodal topic distributions learned using latent Dirichlet allocation. We refer to this as semantic representation (or SR). The second representation combines SR with cross-modal correlations learned between input and output space. We refer to this as correlated semantic representation (or CSR).

We perform experiments under different settings when textual data is in the form of either captions, or phrases, or labels. Here we discuss the two experiments when textual data is in the form of captions. In the first experiment (Exp.1), we learn dataset-specific models separately for both the tasks (Im2Text and Text2Im). And in the second experiment (Exp.2), we analyze the generalization ability of different methods across datasets. For this, instead of learning models for each dataset individually, we use the models learned using SBU dataset in Exp.1 and evaluate the performance on the other two datasets, i.e. Pascal and IAPR TC-12. Precisely, for Im2Text, we consider query images from Pascal or IAPR TC-12 dataset, and perform retrieval on the captions of SBU dataset. Similarly, for Text2Im, we consider query caption from Pascal or IAPR TC-12 dataset, and perform retrieval on the image collection of SBU dataset. In both Exp.1 and Exp.2, we use BLEU and Rouge metrics for evaluation.

Figure 2 compares the performances of different methods on IAPR TC-12 dataset (please refer the paper for more results). Here, we can observe that: (a) For all the three methods, the performance usually improves by using CSR as compared to SR. This indicates the advantage of explicitly infusing cross-correlations into data representation. (b) In cross-dataset experiment (Exp.2), the performance of all the methods degrades significantly compared to that in Exp.1. This reflects the impact of dataset specific biases, and thus emphasizes the necessity of performing cross-dataset evaluations. (c) For most of the cases, the proposed method achieves promising results and mostly outperforms existing techniques.

[1] Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

[2] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[3] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

Open-World Person Re-Identification by Multi-Label Assignment Inference

Brais Cancela[1]
brais.cancela@udc.es
Timothy M. Hospedales[2]
t.hospedales@qmul.ac.uk
Shaogang Gong[2]
s.gong@qmul.ac.uk

[1] VARPA Group,
Universidade da Coruña,
A Coruña, 15071, Spain

[2] School of EECS,
Queen Mary University of London,
London, E1 4NS, U.K.

The task of re-identification (ReID) is defined as the recognition of the same individual at different times and locations. State-of-the-art techniques share two very strong assumptions: *the total number of people in the scene is known* a priori, and there exists a total *overlap of identity* between a camera pair, that is, every person appears in both camera views. This is unrealistic for real-world re-identification scenarios, when there is no prior information about the same people reappearing in the scene at different views. We refer to this unconstrained setting as the 'open world' ReID problem. The open-world problem is more challenging for two reasons: (i) the total number of unique people within each camera and the scene as a whole (cross-cameras) are both unknown, and (ii) each subject may appear in some unknown subset of the cameras.

In this paper we consider for the first time the most general open-world re-identification problem. To address this, we introduce a new Conditional Random Field (CRF) model, making three important contributions: (1) No label information is needed a priori, allowing the system to detect when a new person enters the camera network; (2) An 'open world' solver, that is, the model does not assume that a person will (re)appear in every camera; and (3) Producing a person count as a byproduct. Our approach provides generality that is lacking in existing state of the art closed world ReID solutions.

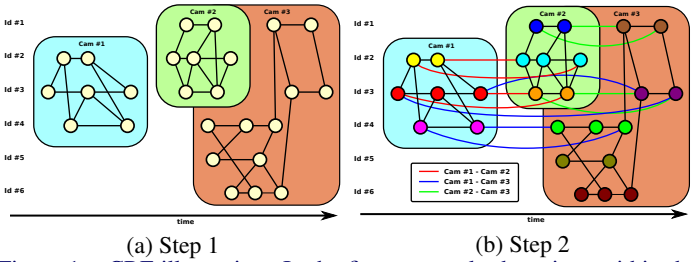
The objective of the CRF is to assign the most likely correct assignment of multiple id labels simultaneously to all the nodes in the CRF. We assume as input a set of N observations $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ across different camera views. Each observation $\mathbf{x}_i = \{c_i, t_i, \mathbf{p}_i, \mathbf{v}_i, \mathbf{a}_i\}$ consists of: A camera c_i making the detection; the time of detection t_i (we assume cameras are synchronized); the image position $\mathbf{p}_i$ and velocity $\mathbf{v}_i$ where the person was detected; and an appearance feature $\mathbf{a}_i$ from the detection bounding box. The re-identification task is to correctly assign identity labels $\mathcal{L} = \{l_i\}_{i=1}^{N}$, $l \in 1 \ldots L$ to all detections..

To address this task we propose a CRF $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where each node corresponds to a person detection (observation) $\mathcal{V} = \{v_i = x_i\}$. Each edge corresponds to a similarity between nodes/persons $\mathcal{E} = \{e_{ij} = (v_i, v_j)\}$, and the label of each node corresponds to the identity of that person/detection. Our aim is to find the set of labels $\mathcal{L}$ that best fits all the observations $\mathcal{X}$,

$$\mathcal{L}^* = \arg\min_{\mathcal{L}} \left(\sum_i U(l_i|\mathcal{X}) + \sum_{ij} B(l_i, l_j|\mathcal{X}) \right), \quad (1)$$

where $U(l_i|\mathcal{X})$ and $B(l_i, l_j|\mathcal{X})$ denote unary and pairwise energy functions, respectively. Our algorithm proceeds in two steps, as explained in Algorithm 1. First, we solve the CRF allowing connections only between detections within the same camera. Second, we use that solution as an initial condition to build the connections between different cameras, creating the final CRF model. The structure and parameterisation of CRF at each stage is the same. We only increase the information included.

To evaluate our contribution, we focus on the challenging SAIVT-Softbio database [1], that includes 150 people recorded using 8 different cameras. Different people appear in different subsets of the cameras.



(a) Step 1 (b) Step 2

Figure 1: CRF illustration. In the first step, only detections within the same camera are connected. In the second step, a restricted connection between cameras is allowed.

Input: Detections $\mathcal{X}$
Output: Associations between detections $\mathcal{L}$
begin
  Compute within camera weights W and U,
  Solve the CRF Eq (1) with Alpha-expansion
  Solve Initial Hungarian to obtain H,
  Compute across camera weights W and U
  Solve the CRF Eq (1) with Alpha-expansion
end

Algorithm 1: Overview of CRF algorithm for open-world ReID.

Table 1: Re-identification among three cameras from SAIVT (3, 5 and 8).

	F_1-Score	Precision	Recall
Naive RankSVM	26.2%	22,0%	42,1%
Naive KISS	29.5%	19.7%	**66.1%**
RankSVM+CRF	42.0%	**53.7%**	39.4%
KISS+CRF	**48.3%**	50.3%	49.8%

Our contribution is agnostic to the appearance feature, and the base pairwise matching model used. To test our methodology, we consider the ELF [4] feature along with RankSVM [2, 4] and KISS [3] pairwise models. Furthermore, spatial and temporal information are included as information between cameras. As we address the open world problem with no prior information about the number of people or their camera overlap, no existing models directly apply. For baselines, we therefore define a more conventional 'engineering' generalisation to open world based on thresholding pairwise RankSVM and KISS scores.

To evaluate the performance of open-world problems the conventional CMC metric is insufficient. We therefore apply statistical analysis techniques. Given the final and ground truth labels, $\mathcal{L}^*$ and $\mathcal{L}_{gt}$, we evaluate all pairs. If two nodes have the same label in $\mathcal{L}_{gt}$ and in $\mathcal{L}^*$, it is a true positive; if they have different labels a true negative, and so on.

According to the obtained results (Table 1), our CRF model is more robust, as evidenced by its maintenance of high precision values. Moreover, it improves both of the base methods it is paired with. Because of the dichotomy between obtaining high precision and high recall, we conclude that the F-Score is the best overall metric to validate an open-world ReID algorithm.

A byproduct of open-world inference is a person count. Table 2 shows the estimated number of unique people among the approximately 600 detections across all three cameras. The estimated number of people along with the standard deviation of the estimate over multiple runs are given. In each case our framework improves on the baseline result, with KISS+CRF obtaining the best and most stable estimate.

Table 2: Inferring the number of distinct people in the dataset.

GT	Naive RankSVM	Naive KISS	RankSVM+CRF	KISS+CRF
48	61 ± 17.6	57.8 ± 11.2	65 ± 13.2	$\mathbf{54.1} \pm \mathbf{7.9}$

[1] Alina Bialkowski, Simon Denman, Patrick Lucey, Sridha Sridharan, and Clinton B Fookes. A database for person re-identification in multi-camera surveillance networks. In *DICTA*, 2012.

[2] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD*, pages 217–226, 2006.

[3] M Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[4] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

Location recognition on lifelog images via a discriminative combination of generative models

Alessandro Perina
alessandro.perina@iit.it
Matteo Zanotto
matteo.zanotto@iit.it
Baochang Zhang
baochang.zhang@iit.it
Vittorio Murino
vittorio.murino@iit.it

Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia
Genova, Italy

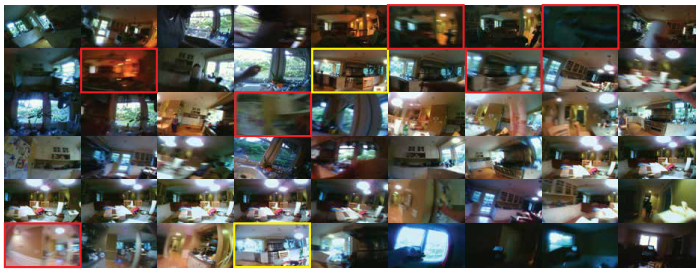


Figure 1: The first 54 images of a lifelog. Notice the high blur (red boxed images) and the dramatic changes in illumination (yellow box)

It is a common belief that in the near future, wearable technology will be the next computing revolution. Such wearable systems are intended to be used in a seamless way like a piece of clothing and they are at the basis of "lifelogging". Among all wearable sensors, the first lifelogging cameras are recently becoming available for a large number of people to use: all of them use a passive record-it-all approach, automatically shooting a photo every 10-30 seconds. However, the soon-to-be enormous amount of images must be organized in order to be useful, and simply using temporal arrangement of the shots is totally unsatisfactory. This paper represents a first step towards this goal: we focused on location recognition and we propose the use of a combination of heterogeneous generative models, each one able to capture the different aspects that characterize each location. Our approach of combining evidence outperforms each individual model as well as other advanced techniques.

Challenges. Lifelog images represent a serious challenge for computer vision researchers. Cameras are usually worn around the neck or attached to clothes and this causes non-linear and unpredictable motion which causes blur and rapid changes in the scene. Figure 1 shows 54 consecutive images spanning a period of ∼15 minutes over which the bearer changes location few times (kitchen, living room, garage). Notice how most of the frames are blurred, while few are highly blurred and difficult to understand even for a human. Moreover, the illumination exhibits dramatic changes over short time periods even when the bearer stays in the same location. Another intrinsic characteristics of lifelogs is that, in a real scenario, the labeled data available to accomplish a classification task are inherently scarce: most of the images, in fact, can only be labeled by the bearer of the camera and crowd-sourcing is difficult, if not impossible.

Motivations. This paper focuses on location recognition. It exploits several recent and classical generative models used for scene understanding to propose a framework able to learn a discriminative combination of weights dealing with the several complexities of multiple heterogeneous models for each location. This choice is motivated by an intuitive and a theoretical reason:

1. The locations one visits are so different that it does not exists a single model able to fit well everywhere. Our favorite grocery store, could nicely be modeled by a full bag-of-word approach like LDA, whereas locations like kitchen or living room are probably well recognized by looking at the objects that contain, and finally contained environments like our work cubicle or our car may well be modeled by an exemplar based-method or by a panoramic reconstruction method like the epitome.
2. When none of the models in an ensemble is the true data generator (TDG) model, there usually exists a combination that can replicate the behavior of the TDG more closely than any individual model on its own.

Overview of the proposed approach. Instead of searching for the best model, or for a combination that can more closely replicate the true data generator model behavior than any individual model on its own, we looked for a discriminative combination of weights. Furthermore, we computed it per-class as, in general, different combinations of models could be better suited for different classes.
Working in a one-vs-all setting, for each class l, we propose to compute the weights π^l which maximize the margin between the average conditional ensemble log-likelihood ratio **A-CLLR** of positive samples and that of negative samples (e.g., belonging to all the other classes). The average conditional log-likelihood of a set of bags of features $\mathbf{c}^t$, is defined as follows

$$A-CLL = \frac{1}{T}\sum_{t=1}^{T} \log p(l^t = l|\mathbf{c}^t) \quad (1)$$

where t indexes a sample, and l^t its class. The likelihood of the ensemble $\mathcal{E}$ is the likelihood of a mixture model whose components are the K individual models $\mathcal{M}_k$ themselves

$$p(l^t = l|\mathbf{c}^t, \mathcal{E}) \propto \sum_k \pi_k^l \cdot p(\mathbf{c}^t|\mathcal{M}_k^l) \quad (2)$$

Our technique allows to exploit all the data in both the generative and discriminative steps. This is crucial as lifelogs cannot have a lot of training data and standard methods could overtrain.

Results. We considered the SenseCam-32 dataset, a portion of lifelog where the dataset authors highlighted 32 recurrent classes visited by the camera bearer over a period of 21 days. We compared our approach with generative combination methods like Bayesian model averaging, discriminative fusion methods and kernel methods built from the log-likelihood of the individual models.
A snapshot of the results is reported in Fig.2. As visible, our combination method always outperforms each individual model in the ensemble, even with a very limited number of training images.
Further results are reported in the paper, where we also exploited the

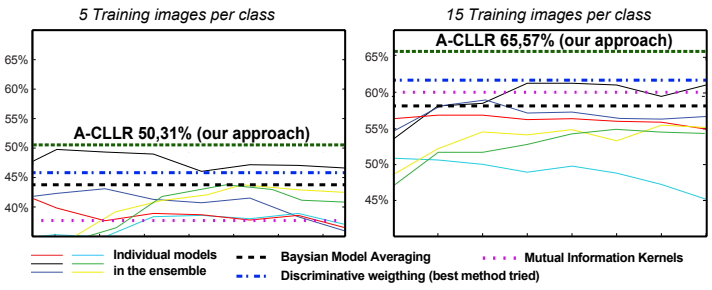


Figure 2: Model combination results on the SenseCam-32 dataset. On the x-axis the K complexities of each model $\mathcal{M}$; on the y-axis the classification accuracy over the 32 classes. See the paper for details.

weak temporal relationships between lifelog images and tested the framework on the 67-indoor scene dataset.

Board #67

Real-time Activity Recognition by Discerning Qualitative Relationships Between Randomly Chosen Visual Features

Ardhendu Behera
http://www.comp.leeds.ac.uk/behera/
Anthony G Cohn
http://www.comp.leeds.ac.uk/agc/
David C Hogg
http://www.comp.leeds.ac.uk/dch/

School of Computing
University of Leeds
Leeds, LS2 9JT, UK
Email: {A.Behera, A.G.Cohn, D.C.Hogg}@leeds.ac.uk

Motivation. Automatic recognition of human *activities* (or *events*) from video is important to many potential applications of computer vision. One of the most common approach is the *bag-of-visual-features*, which aggregate space-time features globally, from the entire video clip containing complete execution of a single activity. The *bag-of-visual-features* does not encode the spatio-temporal structure in the video. For this reason, there is a growing interest in modeling spatio-temporal structure between visual features in order to improve the results of activity recognition.

The proposed framework. We model the spatio-temporal structure by exploiting the qualitative relationships between a pair of visual features. The proposed approach is inspired by [3, 4]. The goal is to find a pair of visual features whose spatiotemporal relationships are discriminative enough, and temporally consistent for distinguishing various activities. The framework is applied to recognize activities from a continuous live video (egocentric view) of a person performing manipulative tasks in an industrial setup. In such environments, the purpose of activity recognition is to assist users by providing on-the-fly instructions from an automatic system that maintains an understanding of the on-going activities.

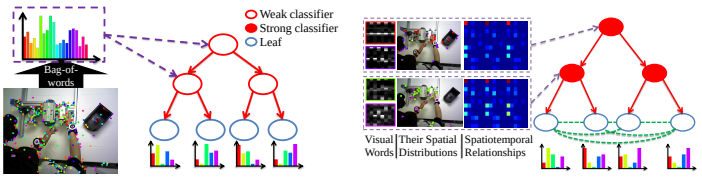
In order to recognize activities in real-time, we propose a *random forest with a discriminative Markov decision tree* algorithm that considers a random subset of relational features at a time and Markov temporal structure that provides temporally smoothed output (Fig. 1). Our algorithm is different from conventional decision trees [2] and uses a linear SVM as a classifier at each nonterminal node and effectively explores temporal dependency at terminal nodes of the trees. We explicitly model the spatial relationships of *left*, *right*, *top*, *bottom*, *very-near*, *near*, *far* and *very-far* as well as temporal relationships of *during*, *before* and *after* between a pair of visual features (Fig. 2), which are selected randomly at the non-terminal nodes of a given Markov decision tree. Our hypothesis is that the proposed relationships are particularly suitable for detecting complex non-periodic manipulative tasks and can easily be applied to the existing visual descriptors such as SIFT, STIP, CUBOID and SURF.

Growing discriminative Markov decision trees. Each tree is trained separately on a random subset of frames belonging to training videos. Learning proceeds recursively by splitting the training frames at internal nodes into the respective left and right subsets. This is done in the following four stages: randomly assign all frames from each activity class to a binary label; randomly sample a pair of visual words; compute the spatiotemporal relationships histogram $\mathbf{h}$ between them; and use a linear SVM to learn a binary split using the extracted $\mathbf{h}$. The binary SVM at each internal node sends the frame to the left child if $\mathbf{w}^T\mathbf{h} \leq 0$ otherwise to the right child, where $\mathbf{w}$ is the set of weights learned through the linear SVM. Using an information gain criteria, each binary split corresponds to a pair of visual words is evaluated on the training frames that falls in the current node. Finally, the split that maximizes the information gain is selected. The splitting process is repeated with the newly formed subsets until the current node is considered as a leaf node.

Inference. For real-time activity recognition, the proposed inference algorithm computes the posterior marginals $P(a_t|l_1^\tau \dots l_t^\tau)$ of all activities a_t over a frame I_t given a history of visited leaf nodes is $l_1^\tau \dots l_t^\tau$ (Fig.1b) for a particular tree τ. The smoothed output over the whole forest is achieved by averaging the posterior probabilities from all $\mathcal{T}$ trees:

$$a_t^* = \arg\max_{a_t} \sum_{\tau=1}^{\mathcal{T}} P(a_t|l_1^\tau \dots l_t^\tau)$$

Results. We evaluate our framework using an egocentric paradigm for recognizing complex manipulative tasks of assembling parts of a pump system in an industrial environment[1]. We compare our approach with our previous work in [1] which models the wrist-object and object-object interactions using qualitative and functional relationships. The accuracy of the proposed approach is 68.56% (using SIFT and STIP) and better than the method in [1], which is 52.09%. We also evaluated using *bag-of-visual-features* approach and the performance is 63.19%. This is achieved using a χ^2-SVM by concatenating STIP and SIFT *bag-of-visual-features*. Activity-wise performance comparison of live recognition is presented in Fig. 3.



(a) Conventional random DT (b) Discriminative Markov DT

Figure 1: (a) Conventional random Decision Trees (DT). The histogram below the leaf nodes represents the posterior probability distribution $P(a|l^\tau)$. (b) The proposed Markov DT sample a pair of visual words and the splitting criterion is based on the relationships between the sampled words. Green dotted lines illustrate the temporal dependencies.

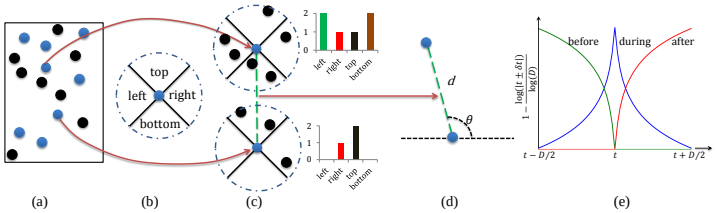


Figure 2: (a) A pair of visual word ('blue dots' and 'black dots') in an image. (b) *Local relationships* (c) Histogram representing *local relationships*. (d) *Global relationships* encode the oriented *very-near*, *near*, *far* and *very-far* relationships. (e) Temporal relationships of *before*, *during* and *after* over a sliding window of duration D.

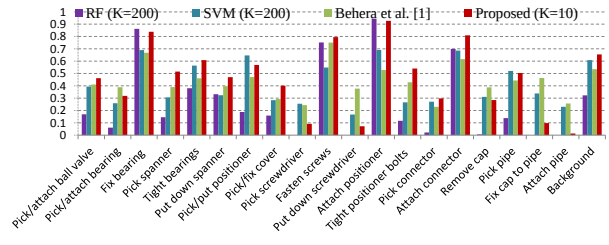


Figure 3: (a) Comparison of the performance of live activity recognition. SIFT bag-of-words ($K = 200$) results in accuracy of 53.21% using χ^2-SVM and 53.28% using conventional random forest. The method in [1] results in 52.09%. The proposed method is 66.20% ($K = 10$) significantly better than the baselines, where the random chance is 5%.

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

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[4] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

[1]Dataset and source code are available at `www.engineering.leeds.ac.uk/computing/research/ai/BallValve/index.htm`

Multi-target tracking in team-sports videos via multi-level context-conditioned latent behaviour models

Jingjing Xiao[1]
shine636363@sina.com
Rustam Stolkin[2]
r.stolkin@bham.ac.uk
Ales Leonardis[3]
a.leonardis@cs.bham.ac.uk

[1] School of Electronics, Electrical and Computer Engineering, University of Birmingham, Birmingham, UK
[2] School of Mechanical Engineering,University of Birmingham, Birmingham, UK
[3] School of Computer Science,University of Birmingham, Birmingham, UK

Sports team tracking poses challenges not present in conventional pedestrian tracking: motion is erratic and players wear similar uniforms with frequent inter-player occlusions. We propose a multi-level multitarget sports-team tracker, which overcomes these problems by modelling latent behaviours at both individual and player-pair levels, informed by team-level context dynamics Fig.1.

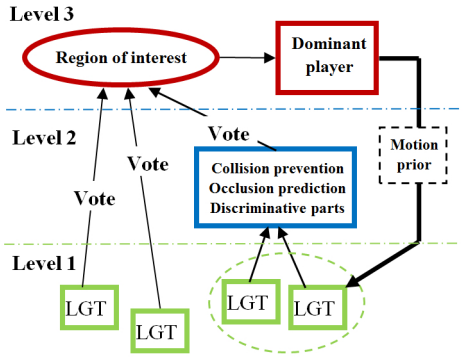


Figure 1: Multi-level tracking algorithm. Level 1: each player tracked by [1]. Level 2: player-player occlusions handled by player-pair behaviour model. Level 3: group or team-level context-dynamics gives dominant player trajectory prediction.

1 Individual player level (Level 1)

At the lowest level (Level 1), we track individual players using the state-of-the-art LGT "Local-Global" tracker [1].This, itself involves two "layers" of tracking: a parts-based set of "local" patches (based on intensity distributions), and a "global" target model (incorporating motion, shape and colour distributions). These local and global layers each provide constraints for re-learning the other, which enables stable adaptation, shown in Fig. 2.

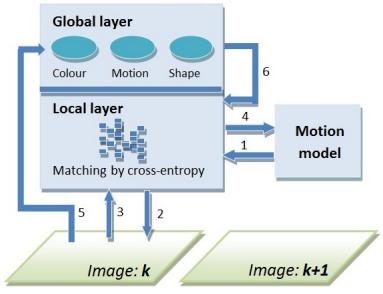


Figure 2: Single-target tracking steps at each frame. 1-spatiotemporal prediction, 2-match local layer, 3-update patches, 4-update motion model, 5-update global layer, 6- add new patches. Adapted from [1]

2 Local group-level (Level 2)

The LGT player models (Level 1) are next augmented by an additional model at the local group-level (Level 2), which encodes the motion preferences of two or more players in close proximity, in the form of a probability distribution representing their tendency to avoid collisions. The pair-wise collision-avoidance model is used to modify the local patch models and global target models of a target pair: the global motion model is modified by the collision avoidance model, providing a stronger motion prior; a prediction is made about which local patches will be occluded during the pair-wise player interaction; and remaining patches are weighted according to their predicted discriminative power during such interactions.

3 Global group-level (Level 3)

We next examine the motion of multiple players at the global group-level (Level 3). Based on player positions, provided by the lower tracking levels, we propose an adaptive approach to meshing the playing area in which the mesh resolution scales appropriately with player density. A player-voting method is then proposed which computes a region of interest (ROI), based on the distribution of player locations and their individual velocities Fig 3.

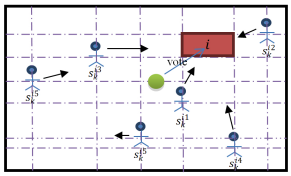


Figure 3: forming mesh according to players' distribution. Green circle: centre of players' distribution; Red region: potential region of interest.

The region of interest does not necessarily indicate the ball position, but may equally indicate the future ball position, or some other position of strategic importance, as predicted by the players. Using this information, it is possible to select one or more "dominant" players, who tend to move with a clearly identifiable trajectory towards the ROI, with a high degree of confidence.

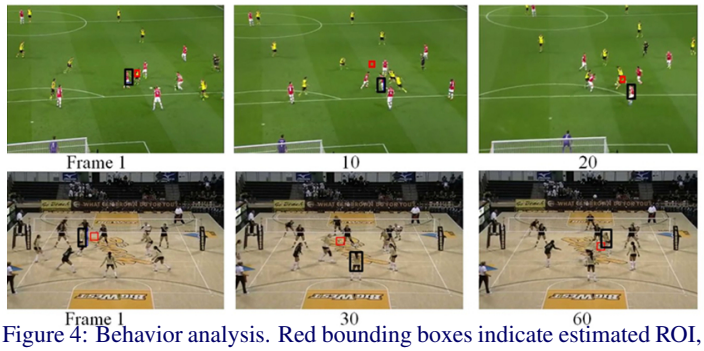


Figure 4: Behavior analysis. Red bounding boxes indicate estimated ROI, Black bounding boxes show a dominant player.

In Fig.5, the group-level models enable successful tracking of interacting/occluding player-pairs where LGT fails (see the right-most player-pair in the right-most image).

Figure 5: Frames 34, 81 of volleyball sequence: LGT (left pair) and our multi-level tracker (right pair). Green/red bounding boxes denote correct/erroneous tracking respectively.

[1] Luka Cehovin, Matej Kristan, and Ales Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(4):941–953, 2013.

Board #69

Parametric temporal alignment for the detection of facial action temporal segments

Bihan Jiang[1]
bi.jiang09@imperial.ac.uk
Brais Martinez[1]
b.martinez@imperial.ac.uk
Maja Pantic[12]
m.pantic@imperial.ac.uk

[1] Computing Department
Imperial College London, UK
[2] Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente,
Netherlands

We propose a new methodology for producing temporal alignment of facial behaviour, and apply it to the analysis of the facial action units (AU) temporal segments. Therefore, our contributions are twofold. In first place, we propose a new methodology for temporal alignment of two sequences of facial behaviour. Secondly, we propose a new way of segmenting the AU temporal segments that relies on the temporal alignment of an exemplar sequence (a template) with the test sequence.

Alignment methodology The temporal alignemnt strategy builds on the work of [4]. In this work, the authors managed to project a sequence into a parametric curve embedded into a lower-dimensional space by applying Laplacian eigenmaps. Furthermore, they were able to backproject from this curve into frame space by means of a simple linear transformation. Formally, if $X = \{\mathbf{x}_t\}_{t=1:n}$ is the original sequence, then this technique allows the construction of a continuous parametric approximation of the original sequence as:

$$\mathcal{X}(t) = A(X)\mathcal{Y}(t) + \bar{\mathbf{x}} \tag{1}$$

where $\mathcal{Y}(t)$ is the curve embedded in the lower dimensional space, and $A(X)$ is a matrix that depends on the original sequence. Crucially, $\mathcal{Y}(t)$ has an analytical form and can be derived analytically.

We then consider a family of parametric functions that represent the possible temporal transformations. For example, we can use a linear warp to account for constant differences on the speeds of actions, or a piecewise linear function. $W(-;\theta)$ represents such transformation parametrised by θ. If aligning the test function onto a template sequence, we define the loss function of the alignment between the template and the test sequence as:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i}^{n} \|\mathbf{x}_i^{\text{tmpl}} - \mathcal{X}\left(W(i;\theta)\right)\|_2^2 \tag{2}$$

Applying the chain rule and the fact that $\mathcal{Y}$ can be analytically differentiated, then we can compute:

$$\frac{\partial \mathcal{Y}(\mathcal{W}(i;\theta))}{\partial \theta_j} = \left.\frac{\partial \mathcal{Y}(t)}{\partial t}\right|_{\mathcal{W}(i;\theta)} \frac{\partial \mathcal{W}(t;\theta)}{\partial \theta_j} \tag{3}$$

It is then possible to minimise the loss function using a Gauss-Newton approach as:

$$\theta^{(it+1)} = \theta^{(it)} - \left(\mathbf{J}^t_{\theta^{(it)}}\mathbf{J}_{\theta^{(it)}}\right)^{-1}\mathbf{J}^t_{\theta^{(it)}}\mathbf{r}\left(\theta^{(it)}\right) \tag{4}$$

where $\mathbf{J}_{\theta^{(it)}}$ is the Jacobian of $\mathcal{X}$ respect to the warp parameters θ.

Application to AU temporal segment detection: The AU temporal segments are defined as neutral (no activation), onset (increase of intensity of the AU), apex (maintain) and offset (decay of intensity of the AU). The task is to label each frame of a sequence accordingly. This is typically done by training per-frame classifiers. However, we propose instead to align the test sequence with an exemplar sequence with known labels (a template). The template labels are then mapped through the alignment function to produce the test sequence labelling.

We define two different warp functions. The first one aligns a full activation episode to the test sequence by using a piecewise linear warping. This model adapts to linear differences in speed independently for each AU segment. This model is illustrated in Fig. 1. Specifically, the warp function is defined as:

$$W(i;\theta) = \begin{cases} \frac{\theta_2-\theta_1}{n_{\text{on}}}i + \theta_1 & : \theta_1 \le i < \theta_2 \\ \frac{\theta_3-\theta_2}{n_{\text{ap}}}i + \theta_2 & : \theta_2 \le i < \theta_3 \\ \frac{\theta_4-\theta_3}{n_{\text{off}}}i + \theta_3 & : \theta_3 \le i \le \theta_4 \end{cases} \tag{5}$$

However, this model does not account for different AU intensities. Smiles can be low intensity (closed mouth and low intensity of the mouth corner pulling) or broad smiles (with open stretched mouth). The second model accounts for this differences. In particular, the action exemplar and the test sequence do not need to be aligned in full. Therefore, the template should reach maximum intensity. This model is illustrated on the right hand side part of in Fig. 1.

$$W(i;\theta) = \begin{cases} 0 & : i < \theta_1 \text{ or } \theta_4 \le i \\ \frac{\theta_5}{\theta_2-\theta_1}(i-\theta_1) & : \theta_1 \le i < \theta_2 \\ \theta_5 & : \theta_2 \le i < \theta_3 \\ -\frac{\theta_5}{\theta_4-\theta_3}(i-\theta_3)+\theta_5 & : \theta_3 \le i < \theta_4 \end{cases} \tag{6}$$

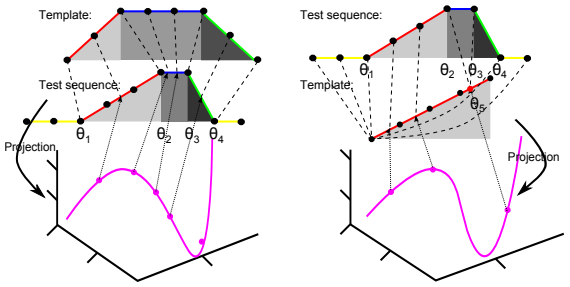


Figure 1: Depiction of the temporal alignment strategy for both of the models presented here (left: model1, right: model2).

The performance achieved by model 2 is the best. However, both models provide superior performance to other state of the art methods, as shown in Table 1.

Table 1: Comparison of AU temporal segment detection methods on the MMI database. $\mathbf{F1}_{\text{act}}$ is the F1-measure after converting into AU activation.

Systems	Neutral	Onset	Apex	Offset	$\mathbf{F1}_{\text{act}}$
Model1	83.42	54.15	78.86	57.87	77.83
Model2	**85.88**	**56.32**	**79.75**	**58.95**	**80.62**
Jiang et al. 2013[1]	78.50	53.38	72.12	48.73	67.53
Valstar et al. 2012[3]	76.60	56.75	69.38	48.87	-
Koelstra et al. 2010[2]	-	-	-	-	62.5

[1] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. Dynamic appearance descriptor approach to facial actions temporal modelling. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 44(2):161–174, 2014.

[2] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(11): 1940–1954, 2010.

[3] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 42(1):28–43, 2012.

[4] Z. Zhou, G. Zhao, and M. Pietikaïnen. Towards a practical lipreading system. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 137–144, 2011.

Modeling Sequential Domain Shift through Estimation of Optimal Sub-spaces for Categorization

Suranjana Samanta
ssamanta@cse.iitm.ac.in
Tirumarai A Selvan
tirumarai.selvan@gmail.com
Sukhendu Das
http://www.cse.iitm.ac.in/~sdas/

Visualization and Perception Lab
Dept. of CS&E
Indian Institute of Technology Madras
Chennai, India

Domain adaptation (DA) is the process in which labeled training samples available from one domain is used to improve the performance of statistical tasks performed on test samples drawn from a different domain. The domain from which the training samples are obtained is termed as the source domain, and the counterpart consisting of the test samples is termed as the target domain. Few unlabeled training samples are also taken from the target domain in order to approximate its distribution.

In this paper, we propose a new method of unsupervised DA, where a set of domain invariant sub-spaces are estimated using the geometrical and statistical properties of the source and target domains. This is a modification of the work done by Gopalan *et al*. [2], where the geodesic path from the principal components of the source to that of the target is considered in the Grassmann manifold, and the intermediary points are sampled to represent the incremental change in the geometric properties of the data in source and target domains. Instead of the geodesic path, we consider an alternate path of shortest length between the principal components of source and target, with the property that the intermediary sample points on the path form domain invariant sub-spaces using the concept of Maximum Mean Discrepancy (MMD) [3]. Thus we model the change in the geometric properties of data in both the domains sequentially, in a manner such that the distributions of projected data from both the domains always remain similar along the path. The entire formulation is done in the kernel space which makes it more robust to non-linear transformations.

Let X and Y be the source and target domains having n_X and n_Y number of instances respectively. If $\Phi(.)$ is a universal kernel function, then in kernel space the source and target domains are $\Phi(X) \in \mathbb{R}^{n_X \times d}$ and $\Phi(Y) \in \mathbb{R}^{n_Y \times d}$ respectively. Let K_{XX} and K_{YY} be the kernel gram matrices of $\Phi(X)$ and $\Phi(Y)$ respectively. Let $D = [X;Y]$ denote the combined source and target domain data, and the corresponding data in kernel space is given as $\Phi(D)$. The kernel gram matrix formed using D is given by $K = \begin{bmatrix} K_{XX} & K_{XY} \\ K_{XY}^T & K_{YY} \end{bmatrix}$, where $K_{XY} = \Phi(X)\Phi(Y)^T$.

Let $\Phi(\tilde{X})$ and $\Phi(\tilde{Y})$ represent the projections of $\Phi(X)$ and $\Phi(Y)$ respectively onto a subspace $W_i \in \mathbb{R}^{d \times p}$, which is a point on the Grassmann manifold $G_{d,p}$. Here, d is the dimension of both source and target domains in RKHS and p is the dimension of the optimal sub-spaces. Then, the square of the distance between the means of two domains is given as:

$$\delta_\mu^2 = tr\left(W_i^T \Phi(D)^T \begin{bmatrix} I_1 & -I_2 \\ -I_2 & I_3 \end{bmatrix} \Phi(D) W_i\right) = tr\left(Z_i^T \Gamma Z_i\right) \quad (1)$$

where, $W_i = \Phi(D)^T Z_i$, $Z_i \in \mathbb{R}^{(n_X+n_Y)\times p}$, $\Gamma = \left(K \begin{bmatrix} I_1 & -I_2 \\ -I_2 & I_3 \end{bmatrix} K\right)$ and $[I_1]_{n_X \times n_X}$, $[I_2]_{n_Y \times n_X}$ and $[I_3]_{n_Y \times n_Y}$ are matrices containing all elements as $1/n_X^2$, $1/n_X n_Y$ and $1/n_Y^2$ respectively and Z_i is the unknown variable to be estimated.

If U_X^Φ and U_Y^Φ are the principal components of $\Phi(X)$ and $\Phi(Y)$ respectively, it can be proved that the principal components of $\Phi(X)$ and $\Phi(Y)U_Y^\Phi U_X^{\Phi T}$ are the same. Hence, the starting point of the path P^W is the principal components of $\Phi(D_s) = [\Phi(X); \Phi(Y)U_Y^\Phi U_X^{\Phi T}]$ and the end point of P^W can be obtained by the principal components of $\Phi(D_t) = [\Phi(X)U_X^\Phi U_Y^{\Phi T}; \Phi(Y)]$. Let, U_s^Φ and U_t^Φ be the principal components of $\Phi(D_s)$ and $\Phi(D_t)$ respectively. Also, V_X^Φ and V_Y^Φ be the eigen-vectors of K_{XX} and K_{YY} respectively. Similarly, let V_s^Φ and V_t^Φ be the eigen-vectors of K_s and K_t respectively, where K_s and K_t are the kernel gram matrices built on $\Phi(D_s)$ and $\Phi(D_t)$ respectively.

Let, G_i denote the i^{th} sampled point on the geodesic path P^G and the i^{th} sample point on P^W represent the sub-space W_i. The start and the end points of P^W are given by $W_1 = V_s^\Phi$ and $W_{N'} = V_t^\Phi$ respectively, while the intermediate points are denoted by W_i, $i = 1, \ldots N'-1$. Now, P^W is the path of shortest length if the sampled points from P^W is closest to the corresponding sampled points from P^G, i.e. $d_{proj}(G_i, W_i)$ is minimum, $\forall i = 2, \ldots, (N'-1)$. The square of the distance between two sub-spaces, P_i^G and P_i^W in the kernel space, ia given as:

$$\delta_{proj}^2(W_i, G_i) = p - tr(Z_i^T \hat{K}_i V_i^\Phi V_i^{\Phi T} \hat{K}_i^T Z_i) = p - tr(Z_i^T \Pi_i Z_i) \quad (2)$$

where, $\Pi_i = \hat{K}_i V_i^\Phi V_i^{\Phi T} \hat{K}_i^T$. $\Phi(\hat{D}_i)$ is an appropriate projection of $\Phi(D)$. V_i^Φ is the i^{th} intermediary point sampled on the geodesic path from V_s^Φ to V_t^Φ and $\hat{K}_i$ is the kernel gram matrix (for i^{th} sub-space in the sequence) given as $KV_i^\Phi V_s^{\Phi T} K$.

For an optimal value of Z_i, δ_{mu}^2 and $\delta_{proj}^2(G_i, W_i)$ given in Eqns. 1 and 2 should be minimum. The optimization framework to estimate Z_i is:

$$\underset{Z_i}{\text{maximize}} \quad tr(Z_i^T \Pi_i \Gamma^{-1} Z_i) \quad (3)$$

$$\text{subject to} \quad Z_i^T Z_i = I \quad (4)$$

After obtaining the set of optimal Z_is, the projections of the data onto W_is are given as $\Phi(D)W_i = KZ_i, \quad \forall i = 2, \ldots, (N'-1)$. The projection of the data points onto the first and last (or initial and final) points of the path P^W i.e. on U_s^Φ and U_t^Φ are:

$$\Phi(D)U_s^\Phi = \Phi(D)\Phi(D_s)^T V_s^\Phi = \begin{bmatrix} K_{XX} & K_{XX}V_X^\Phi V_Y^{\Phi T} K_{YY} \\ K_{XY}^T & K_{XY}^T V_X^\Phi V_Y^{\Phi T} K_{YY} \end{bmatrix} V_s^\Phi \quad (5)$$

$$\Phi(D)U_t^\Phi = \Phi(D)\Phi(D_t)^T V_t^\Phi = \begin{bmatrix} K_{XY}V_Y^\Phi V_X^{\Phi T} K_{XX} & K_{XY} \\ K_{YY}V_Y^\Phi V_X^{\Phi T} K_{XX} & K_{YY} \end{bmatrix} V_t^\Phi \quad (6)$$

After obtaining the optimal sub-spaces, the projections of the source and target domains onto the intermediary sub-spaces are obtained and concatenated together, as done in [2], for training the KNN classifier.

We evaluate the performance of the proposed method of DA for improving the results of object categorization using Office + Caltech datasets [1]. The dataset contains four domains: Amazon (A), Caltech (C), Dslr (D) and Webcam (W), with 10 classes of objects in each of the domains. Table 1 shows the classification accuracies for 12 different pairs of source and target domains, using a 25-fold cross validation.

Table 1: Classification accuracies (in %-age) on Office+Caltech dataset [1], using different techniques of unsupervised domain adaptation.

Method	C→A	D→A	W→A	A→C	D→C	W→C
GFS [2]	36.9	32	27.5	35.3	29.4	21.7
GFK [1]	36.9	32.5	31.1	**35.6**	29.8	27.2
Proposed	**42.63**	**44.16**	**44.65**	34.40	**41.56**	**43.26**
Method	A→D	C→D	W→D	A→W	C→W	D→W
GFS [2]	30.7	32.6	54.3	31.0	30.6	66.0
GFK [1]	35.2	35.2	70.6	34.4	33.7	74.9
Proposed	**38.82**	**43.64**	**80.57**	**39.31**	**42.27**	**78.03**

The proposed method of unsupervised domain adaptation handles non-linear transformation of data as well as estimates intermediate domain invariant sub-spaces, making it more efficient.

[1] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

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

[3] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning*, 13:723–773, 2012.

Board #71

Geodesic pixel neighborhoods for multi-class image segmentation

Vladimir Haltakov[1]
http://campar.in.tum.de/Main/VladimirHaltakov
Christian Unger[1]
http://campar.in.tum.de/Main/ChristianUnger
Slobodan Ilic[2]
http://campar.in.tum.de/Main/SlobodanIlic

[1] BMW Group
Munich, Germany
[2] Siemens AG
Corporate Technology
Munich, Germany

Introduction

Multi-class image segmentation is a complex problem that poses several challenges: developing better classifiers, designing more discriminative features, finding efficient optimization techniques and modeling the relations between image pixels in different image regions. In this paper we focus on the last one. A common way to address the problem of structured prediction is to model it as a Conditional Random Field (CRF), but in this paper we take a different approach by using classification and integrating local and global semantic structure constraints directly in the features.

Our contribution is threefold. Firstly, we introduce a classification framework based on the concept of pixel neighborhoods, which captures structure constrains with a new histogram based neighborhood feature. Secondly, we propose a novel way to use the geodesic distance to compute the local pixel neighborhood. Thirdly, we introduce a new global rays based neighborhood, again using the geodesic distance, that can also capture global context.

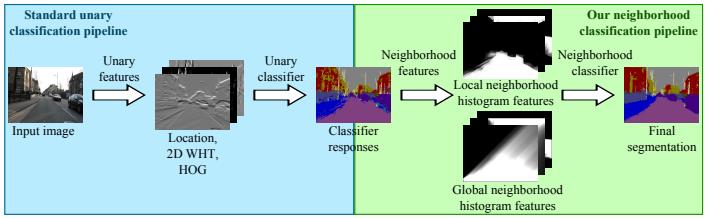


Figure 1: The blue part shows the standard unary classification process, while the green part shows the neighborhood classification pipeline.

Pixel neighborhoods

We introduce a classification framework based on the concept of pixel neighborhoods as visualized in Fig. 1. We define the neighborhood N_i of pixel i as a set of pixels that are related to the pixel i in some way. In the paper we explore several ways to define pixel neighborhoods by making use of the geodesic distance transform defined as:

$$d(i,j) = \inf_{\mathbf{G}\in\mathcal{P}_{i,j}} \int_0^l \sqrt{1+\gamma^2(\nabla I \cdot \mathbf{G}'(\mathbf{s}))^2}\,ds, \qquad (1)$$

where $\mathcal{P}_{i,j}$ is the set of all possible paths between pixels i and j, $\mathbf{G}$ is a path from this set with length l and $\mathbf{G}'$ is its spatial derivative. The parameter γ indicates the weight between the image gradient and the spatial distance between the two pixels. For $\gamma = 0$ the geodesic distance becomes equivalent to the euclidean distance, while for $\gamma = 1000$ it is dominated by the image gradients.

Figure 2: Visualization of the shapes of the presented neighborhoods for selected pixels (marked in black).

We introduce two types of neighborhoods: an adaptive local neighborhood and a rays based global neighborhood that are able to express local or global relations respectively (see Fig. 2). The local neighborhood consists of the closest N pixels to the pixel of interest according to the geodesic distance. This allows the neighborhood to cover a patch around the pixel that aligns well to strong image gradients which often correspond to object edges. For the global neighborhood we shoot 8 rays at 45° from the pixel of interest to the borders of the image. For each ray we define a separate neighborhood, which again makes use of the geodesic transform to accumulate the pixels along the ray. In this way our global neighborhood is able to capture long range context relations.

Neighborhood classification framework

We first classify each image pixel individually based on features computed from the image. Then, for each pixel we compute one or more pixel neighborhoods and summarize the responses of the classifier over each neighborhood by computing a new histogram based feature. This is done by letting each pixel in the neighborhood vote for its most probable label based on the responses of the unary classifier. We then create a normalized histogram over all votes from the neighborhood and use the values of the histogram as features. This new feature is a compact representation of the whole neighborhood which allows for very fast training.

We use the neighborhood features to train a second classifier which is again used to classify each pixel, but in contrast to the first one, it integrates local and global constraints from the neighborhood features and is therefore able to improve the results of the first classifier significantly.

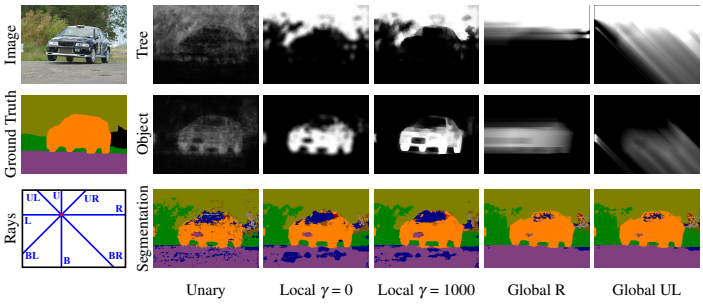


Figure 3: Histogram based neighborhood features. We show the bins corresponding to the classes TREE and OBJECT as a probability map, the segmentation from the neighborhood classifier for the local and global neighborhoods and the raw unary responses. The last two columns show two of the rays of the same global neighborhood. A high value for a pixel in the global neighborhood of a ray means that there is a region of this class in this direction.

Results

We evaluate our method on three widely used and very challenging datasets: CamVid, MSRC-21 and the Stanford background dataset. We analyze the performance of the different parts of our model and show how they contribute to increase the segmentation performance. Furthermore, we compare to two well known and strongly related methods: auto-context [2] and the robust P^n model of [1] and show an increase in performance especially around the object edges.

[1] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *IJCV*, 2009.

[2] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. In *PAMI*, 2010.

High Entropy Ensembles for Holistic Figure-ground Segmentation

Ignazio Gallo
ignazio.gallo@uninsubria.it
Alessandro Zamberletti
a.zamberletti@uninsubria.it
Simone Albertini
simone.albertini@uninsubria.it
Lucia Noce
lucia.noce@uninsubria.it

Applied Recognition Technology Laboratory
Department of Theoretical and Applied Science
University of Insubria
Varese, Italy

1 Overview and Results

In this paper we approach the task of figure-ground segmentation of natural images using a novel framework to generate highly collaborative tree-based structures, called High Entropy Ensembles (HEE).While many model combination frameworks adopt rejection rules to improve the classification time of the ensembles at the cost of restricting the interactions between the different elements in the structures, throughout our work we prove that, similarly to the Cascade Classification Model [3], when execution time is not critical, better results can be obtained when encouraging that kind of interaction by combining heterogeneous suboptimal classifiers into highly connected tree-based ensembles in which the different algorithms communicate with each other to let the strengths of one overcome the weaknesses of the others and vice versa. Inspired by random-based model combination approaches [2], we do not focus on looking for the optimal classifiers to be added to the HEE, instead we pick them from a pool of randomly configured segmentation algorithms. This randomness injection increases the effectiveness of HEE while also decreasing both the computational complexity of the model creation procedure and the risk of overfitting the training data, which is a common issue for most model combination frameworks.

2 Proposed Method

The proposed method consists in a *building phase* that creates a figure-ground segmentation ensemble by executing an initial *base* step followed by a recursive sequence of *bottom-up* and *top-down* steps. The building procedure is driven by the maximization of a *goodness* function that defines the quality of the HEE being built. The goal of the initial *base* step is to identify both the first suboptimal figure-ground segmentation algorithm a that needs to be added to the ensemble T and its set of input image features F_a, as shown in Fig. 1a. Once the first node has been identified and added to the structure, the ensemble T is progressively augmented by adding new suboptimal root and leaf nodes through a recursive sequence of *bottom-up* and *top-down steps*, as shown in Fig. 1b and Fig. 1c respectively. The *building phase* terminates once a *bottom-up* step followed by a *top-down step* do not increase the value of the goodness function computed for T. The resulting HEE T can be used to generate the soft figure-ground segmentation map $M_I^{t_r}$ for a given image I simply by providing I as input to every node in T, as summarized in Fig. 1d. The final binary segmentation map M_I for I is obtained by thresholding $M_I^{t_r}$.

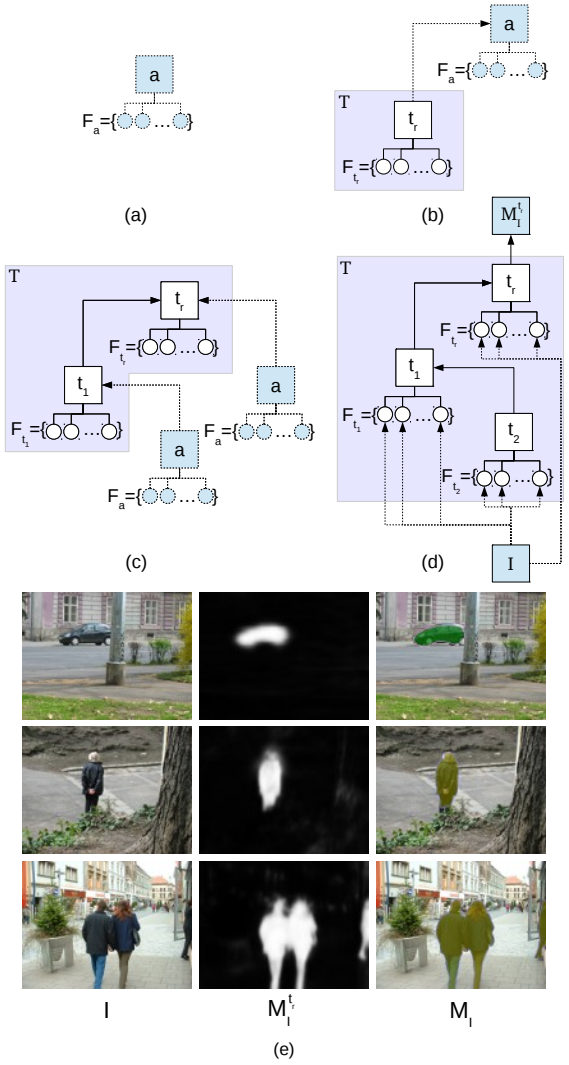


Figure 1: **HEE building procedure and segmentation examples.** (a) base step (b) bottom-up step (c) top-down step (d) segmentation of an image I (e) segmentation examples for images from INRIA Graz-02.

3 Results

An end-to-end experimental analysis is conducted in order to compare HEE against other state-of-the-art figure-ground segmentation algorithms and model combination methods on several challenging datasets: Weizmann Horses, Oxford Flowers 17, INRIA Graz-02 and the figure-ground variant of Pascal VOC 2010; despite the simplicity of our approach, HEE outperform all the evaluated competing methods.

4 Conclusion

The proposed method does not require any user input nor extensive tuning and constitutes a valid alternative to other frameworks when combining heterogeneous figure-ground segmentation algorithms. It is particularly interesting to observe that in many cases the set of image features automatically selected by the *building phase* as input features for the nodes in the HEE resembles the base set of Integral Channel Features [1] (LUV, gradient histogram and magnitude) widely used by state-of-the-art rigid object detection algorithms. This proves that, even tough the proposed model is heavily random-based, it tries to build optimal segmentation ensembles. It is an open question whether our method can pose a challenge to other similar approaches when applied to more challenging tasks, such as object classification or multi-class image segmentation.

[1] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *British Machine Vision Conference*, 2009.

[2] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[3] Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. Cascaded classification models: Combining models for holistic scene understanding. In *International Conference on Neural Information Processing Systems*, 2008.

Board #73

Frankenhorse: Automatic Completion of Articulating Objects from Image-based Reconstruction

Alex Mansfield[1]
mansfield@vision.ee.ethz.ch
Nikolay Kobyshev[1]
nk@vision.ee.ethz.ch
Hayko Riemenschneider[1]
hayko@vision.ee.ethz.ch
Will Chang[2]
wychang1@cs.ubc.ca
Luc Van Gool[1]
vangool@vision.ee.ethz.ch

[1] Computer Vision Lab
ETH Zürich
Switzerland
[2] Department of Computer Science
University of British Columbia
Canada

Reconstruction of scene geometry and semantics are important problems in vision, and increasingly brought together. The state of the art in Structure from Motion and Multi View Stereo (SfM+MVS) can already create accurate, dense reconstructions of scenes. Systems such as CMPMVS [2] are freely available and produce impressive results automatically. However, when assumptions break down or there is insufficient data, noise, extraneous geometry and holes appear in the reconstruction.

We propose to solve these problems by introducing prior knowledge. We focus on the difficult class of articulating objects, such as people and animals. Prior modelling of these classes is difficult due to the articulation and large intra-class variation. We propose an automatic method for completion which does not rely on a prior model of the deformation or training data captured under controlled conditions. Instead, given far from perfect reconstructions, we simultaneously complete each using the well-reconstructed parts of the others.

This is enabled by the data-driven piecewise-rigid 3D model alignment method of Chang and Zwicker [1]. This method estimates local coordinate frames on the meshes and proposes correspondences by matching local descriptors. Each correspondence determines a rigid alignment, which is used as a label in a graph labelling problem to determine a piecewise-rigid alignment which brings the meshes into correspondence while penalising stretching edges.

Our main contributions are as follows. We present a novel, fully automatic method for the completion of noisy real SfM+MVS reconstructions which (1) exploits a set of noisy reconstructions of objects of the class, rather than relying on a large clean training set which is expensive to collect, (2) handles the articulation structure in the class of objects, allowing larger holes to be filled and with greater accuracy than a generic smoothness prior and (3) is exemplar-based, allowing details to be maintained that may be smoothed out in related learning-based approaches.

Our method takes as its input sets of images of scenes each containing an object of a specific class. For each input image set, initially yielding an incomplete and cluttered reconstruction of the whole scene, the output is a completed model of the object, created using the other reconstructions. Our method consists of a pipeline of several stages, visualised in Figure 1.

In the first stage, each scene is reconstructed using a SfM+MVS pipeline [2]. We then segment the objects from the scene by combining object detections in the images. In the third stage, we align each of the segmented source models to the target model taking into account articulation using the method of Chang and Zwicker [1]. We exploit these aligned source models to remove clutter from the target model, and hence correctly identify the holes. Finally, we choose a completion for each hole from those proposed by the aligned source models, and reconstruct the final result filling small holes using screened Poisson reconstruction [3].

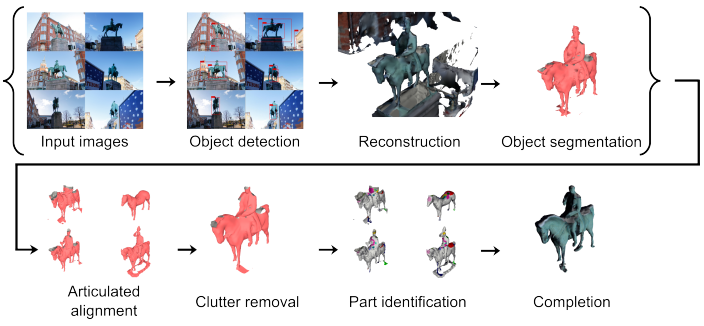


Figure 1: Our fully automatic pipeline takes at the input datasets of images, and processes each to obtain a segmented model of the object (upper row). Completion of a noisy target model from SfM+MVS reconstruction draws on the whole set of segmented models (lower row).

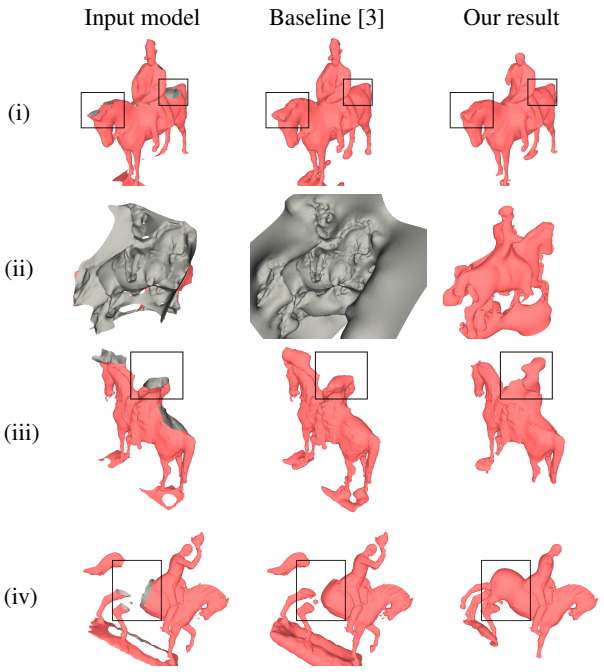


Figure 2: Results (i) – (iii) show the completion of real holes, in (i) the back, (ii) half of the horse, and (iii) the rider. Result (iv) shows the completion of synthetically created holes in the back. For small holes, the baseline also produces good results (i), but for larger holes, the smooth completion rounds off the hole, while our method can complete the part.

As our method performs completion as a post-process, we expect it to produce a plausible reconstruction of the real object. Given the large holes, there is a large variety of appropriate solutions. This is hard to model quantitatively and so visual inspection provides the best evaluation method. Using visual inspection, we perform a qualitative evaluation of our full set of our results, and show typical results in Figure 2. We also perform a quantitative analysis that proves effectiveness of our method.

We demonstrate that while small holes can be completed with local smoothness priors, completing large holes requires a global perspective. We successfully add missing parts like heads, legs and horse riders which are otherwise just smoothed out stumps. Our failure modes occur due to the registration of the models and confusing locally similar parts.

[1] Will Chang and Matthias Zwicker. Automatic registration for articulated shapes. *Computer Graphics Forum (Proc. SGP)*, 27(5):1459–1468, 2008.

[2] M. Jancosek and T Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Proc. CVPR*, 2011.

[3] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *Proc. SIGGRAPH*, 32(3), 2013.

Board #74

Online Dense Non-Rigid 3D Shape and Camera Motion Recovery

Antonio Agudo[1]
aagudo@unizar.es
J. M. M. Montiel[1]
josemari@unizar.es
Lourdes Agapito[2]
l.agapito@cs.ucl.ac.uk
Begoña Calvo[1,3]
bcalvo@unizar.es

[1] Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Zaragoza, Spain.
[2] Department of Computer Science, University College London, London, United Kingdom.
[3] Centro de Investigación en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Zaragoza, Spain.

Recovering 3D reconstruction from 2D images of a deforming object is an inherently ill-posed problem and it usually requires prior knowledge on the scene structure. Most approaches model the non-rigid shape using a low-rank shape constraint [5, 7, 12] combined with additional priors such as temporal smoothness [4, 6, 12], smooth-time trajectories [3, 8], spatial smoothness [7, 12] and inextensibility constraints [13]. Although accurate results have been obtained in recent years, these approaches process all the frames in the sequence in batch manner after video acquisition, preventing them from *online* and *real-time* applications. While sequential rigid real-time solutions exist for a sparse set of salient points [9] and even per-pixel dense reconstruction [10], online estimation of non-rigid objects from a single camera based only on the measurements up to that moment remains a challenging problem. Only recently, sequential formulations have emerged using either sparse [1, 11] or dense correspondences [2].

In this paper, we propose a *sequential solution* to simultaneously recover camera motion and the 3D reconstruction of non-rigid objects from 2D point tracks in a monocular image sequence as the data arrives. We employ a *probabilistic linear subspace* to encode the non-rigid 3D shape at each frame where the shape basis is computed by modal analysis. Our contribution is to propose a new mode shape computation algorithm that makes possible the full extension of the method to dense shapes, and a sequential expectation maximization based algorithm to solve the latent variable problem providing both efficient and more accurate solutions with respect to state-of-the-art sequential methods. Our approach works in two stages: shape basis computation and online estimation.

In stage one, we estimate a shape at rest using a few initial frames, and then the surface is discretized by means of a soup of triangular finite elements where applying the continuum mechanics. The mode shapes can be computed by modal analysis solving an eigenvalue problem [2] obtaining two non-rigid families: bending and stretching modes. The first one is affordable to compute even for dense cases, but only it is valid for out-of-plane stretching deformations. However, to code shapes undergoing stretching deformations, stretching modes have to be included in the shape basis. Unfortunately, computing these mode shapes may become prohibitive –sometimes unfeasible– for some dense cases in terms of computational and memory requirements. It is our first contribution to propose a *growth of modes* (Fig. 1) to easily compute all frequency spectrum and to obtain the stretching modes at quite affordable cost. We compute the mode shapes on a down-sampled shape at rest, and then the sparse shape basis is grown back to dense exploiting the shape functions within the finite element.

In stage two, equipped with this low-rank deformable shape basis, it is our second contribution to propose an online expectation maximization based algorithm over a sliding temporal window of frames to estimate the model parameters as the data arrives. Since the basis weights are modeled with hierarchical priors, these can be marginalized out and we only optimize a small number of parameters per frame obtaining a low cost system that potentially runs in real-time.

We show successful non-rigid 3D reconstruction results on several challenging sequences from highly extensible to inextensible deformations. We also show the advantages of our approach w.r.t. competing sequential methods. Our approach is also valid from sparse to dense data, do not require any training data and can deal with missing data.

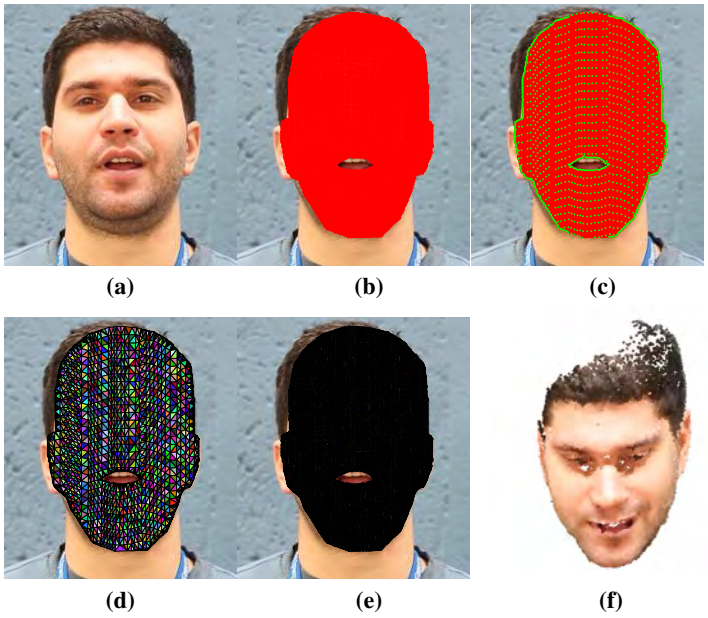


Figure 1: **Growth of modes for dense shapes. (a):** Reference image plane to compute optical flow. **(b):** Dense 2D tracking of p points. **(c):** Subsample of dense shape into q points (green points) with $q << p$. **(d):** Delaunay triangulation for sparse mesh. **(e):** Active search to match every point in the sparse mesh. **(f):** General viewpoint of the 3D reconstruction. Best viewed in color.

[1] A. Agudo, B. Calvo, and J. M. M. Montiel. Finite element based sequential bayesian non-rigid structure from motion. In *CVPR*, 2012.

[2] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *CVPR*, 2014.

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *TPAMI*, 33 (7):1442–1456, 2011.

[4] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *CVPR*, 2008.

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

[6] A. Del Bue, X. Llado, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *CVPR*, 2006.

[7] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, 2013.

[8] P. F. U. Gotardo and A. M. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *CVPR*, 2011.

[9] G. Klein and D. W. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007.

[10] R. Newcome, S. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *ICCV*, 2011.

[11] M. Paladini, A. Bartoli, and L. Agapito. Sequential non rigid structure from motion with the 3D implicit low rank shape model. In *ECCV*, 2010.

[12] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors. *TPAMI*, 30(5):878–892, 2008.

[13] S. Vicente and L. Agapito. Soft inextensibility constraints for template-free non-rigid reconstruction. In *ECCV*, 2012.

Board #75

Scene Flow Estimation using Intelligent Cost Functions

Simon Hadfield
S.Hadfield@surrey.ac.uk
Richard Bowden
R.Bowden@surrey.ac.uk

Centre for Vision Speech and Signal Processing
University of Surrey
Surrey, UK

Scene flow is the 3D counterpart to optical flow, describing the 3D motion field of a scene, independent of the cameras which view it. Motion estimation techniques (both scene flow and optical flow) is a fundamental tool in computer vision. It forms the basis or pre-processing step for many other algorithms, and is included in many vision libraries. These techniques are typically based upon the assumption of brightness constancy, or related assumptions such as gradient constancy and filter response constancy.

A lot of previous work has been dedicated to accurately modelling the behaviour of these consistency assumptions, in the motion fields of real scenes. This helps handling scene artifacts such as non lambertian surfaces, illumination changes and occlusions. In this paper we extend this analysis further, and examine the behaviour of visual consistency assumptions, in cases where the motion field has been incorrectly estimated.

Distinguishing truth from errors

Intuitively, the accurate modelling of visual consistency for ground truth motion fields helps ensure that correct motion fields are always recognised as such (reducing "False negatives"). However, it tells us nothing about the metrics ability to reject erroneous motion fields ("False positives").

For our analysis we examine a range of common visual constancy assumptions. These include the Optical Flow Constraint (OFC), L_2 brightness constancy (SQ), and gradient constancy counterparts OFC_g and SQ_g.

Ideally these metrics should provide a low cost for true motions and a high cost for incorrect motions, as illustrated by the PDF in fig. 1(a). In this case the true motion field registers no violation in the underlying assumption (the PDF contains all responses at 0), while the incorrect motions strongly violate the assumption (PDF is concentrated at 1).

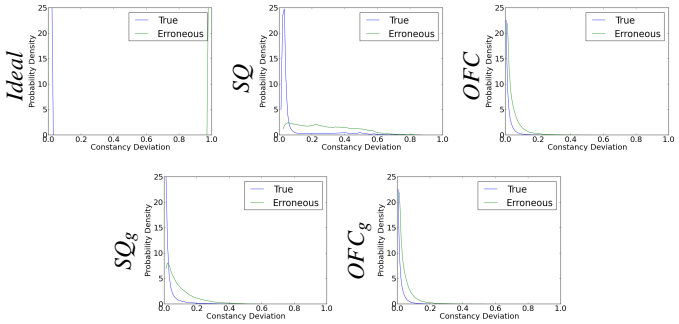


Figure 1: Responses for various motion estimation metrics (including an ideal example), applied to the ground truth and error motion fields of a real scene.

The actual response distributions found for the examined metrics tell an unfortunate story. Most ground truth motions are assigned to the lower 20% of the responses, with the occlusion and specularity effects seen previously being the minority. However, similar responses are produced, even for the significantly erroneous motions. Indeed the linearised brightness constancy metric OFC shows an 80% overlap between the two PDFs. Attempting to minimize these metric responses across the scene, will result in almost as many correct motions being discarded, as incorrect. In the full paper, further analysis is performed to examine how the metric response changes as the amount of error in the motion field varies (i.e. does the response smoothly decrease, as the error is reduced).

Intelligent Cost Functions

We propose a simple solution to this problem; Explicitly finding discriminatory metrics, using machine learning techniques. These "Intelligent cost functions" (ICFs) are able to embody more complex behaviours. As an example, it may be expected that in very light or dark parts of the scene, image contrast would be reduced. In this case, little variation may be expected naturally, and any appearance deviations may be more significant. Alternatively, specular effects may cause a large change in appearance across all colour channels, while a change in appearance for only one channel is more likely to relate to an erroneous motion.

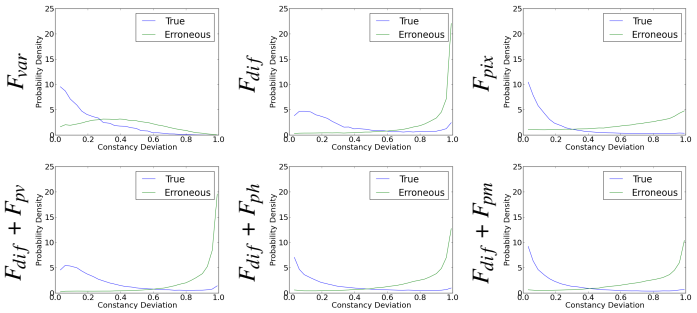


Figure 2: Distribution of responses for various ICFs, applied to ground truth and erroneous motions. Top row: pixel features. Bottom row: local context features.

To learn these Intelligent Cost Functions (ICFs) Gaussian Processes (GP) were trained to model the relationship between various input features and the level of motion error. The GP provides a non-parametric means for fitting these complex relationships, by estimating a distribution over the infinite set of possible cost functions.

Fig. 2 shows the performance of ICFs, based on various visual features (see supplementary material[1] for results on a range of additional sequences). The simplest (F_{var}) provides little additional separation. However, in the case of F_{dif} encoding, and the local context features, the ICF exploits richer infrmation to greatly improve separation.

Motion Estimation with ICFs

We've seen that standard motion estimation cost functions have some significant flaws, and that greater robustness may be obtained via ICFs. However, much work in motion estimation (particularly for optical flow where there are no problems with differing sensor responses) has looked at producing specialised subsystems to mitigate, rather than correct, these issues. As such, it is important to examine whether the use of ICFs does in fact translate to more accurate motion estimates.

Metric	ε_{of}	ε_{sf}	ε_{st}	ε_{ae}	Runtime (secs)
SQ	0.173	0.010	1.52	1.66	352
F_{var}	0.164	0.021	1.53	1.63	389
F_{dif}	0.111	0.009	**1.04**	1.41	363
F_{pix}	0.142	0.012	1.17	1.47	**340**
$F_{dif}+F_{pv}$	0.100	**0.005**	1.10	1.50	440
$F_{dif}+F_{ph}$	0.134	0.008	1.14	1.59	560
$F_{dif}+F_{pm}$	**0.098**	0.014	1.06	**1.23**	430

Table 1: Performance for scene flow estimation, based on the original SQ metric, and a range of ICFs.

To this end, a recent, publicly available, algorithm for scene flow estimation (based on the SQ cost function) is modified to exploit ICFs. Results in tab. 1 are averaged over all sequences from the Middlebury dataset. The results show an almost universal improvement in motion estimation accuracy, with $F_{dif}+F_{pm}$ providing improvements to magnitude, directional and structural accuracies of 44%, 20% and 30% respectively.

We also examine the behaviour of ICFs in optical flow scenarios. In this case we discover a more modest 20% improvement in accuracy.

OFC_g F_{pix} Ground truth

Figure 3: Example motion fields for one of the Middlebury sequences, comparing a standard metric and an ICF against the ground truth.

[1] personal.ee.surrey.ac.uk/Personal/S.Hadfield/icf.html

Board #76

DNN Flow: DNN Feature Pyramid based Image Matching

Wei Yu[1]
w.yu@hit.edu.cn
Kuiyuan Yang[2]
kuyang@microsoft.com
Yalong Bai[1]
ylbai@mtlab.hit.edu.cn
Hongxun Yao[1]
h.yao@hit.edu.cn
Yong Rui[2]
yongrui@microsoft.com

[1] Harbin Institute of Technology
Harbin, China
[2] Microsoft Research
Beijing, China

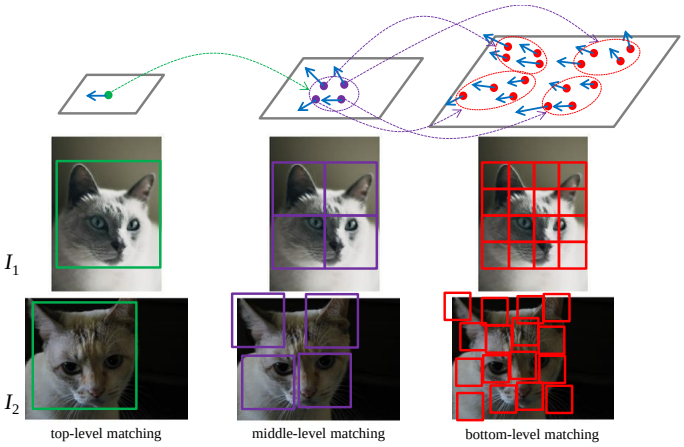


Figure 1: Matching I_1 and I_2 using DNN flow. Each column shows the matching of different levels. In first row, parallelogram denotes the DNN feature image of I_1, where dot represents the feature at that location. Line with arrow denotes the flow vector of the corresponding feature, while curve with arrow denotes guidance from high level to low level. In second row, the color rectangles show I_1's patches covered by the DNN features. Third row shows I_2's matching patches corresponding to the patches of second row.

As a fundamental problem in computer vision, image matching is the cornerstone for many vision problems, such as motion estimation [2], label propagation [3] and object modeling [1]. The goal of image matching is to find the corresponding pixels between two images. Based on the variations between the two images, we roughly divide image matching into two categories, i.e., instance-level matching and category-level matching. Compared to instance-level matching, category-level matching tries to match two images with more challenge variations, which belong to same category. Category-level matching aims to overcome the intra-class variability in shape and other visual properties, such as cars with various shapes and colors and cats with different poses and furs.

In this paper, we propose a DNN feature based image matching approach, which focuses on category-level matching. Recently, Deep Neural Network (DNN) has shown great ability in handling the variations under the same category. The ability comes from the gradual abstraction through several layers, where low layer detects simple patterns, such as edges and blobs, middle layer detects object parts and high layer detects objects. Considering the ability of DNN feature in handling semantic variations, we propose a novel image matching method based on DNN feature pyramid, named as DNN Flow. DNN Flow utilizes DNN features of different layers to achieve coarse to fine matching. As shown in Figure 1, top level matching attempts to achieve object level matching since top level features detect patterns at object level, middle level matching establishes correspondences at part level, finally bottom level matching achieves fine level matching through small patterns.

The main advantage of DNN flow is to utilize more targeted feature to achieve the matching goal of each level. The top level feature with semantic invariance helps to discriminate inter-class variance and stand intra-class variance. Therefore, top level feature is robust to fight against various visual variance. Even if two images of the same category are obvious similar at bottom level, high-level matching still produces helpful coarse flow field and guides low-level matching along with the reasonable direction.

For given two images I_1, I_2, and corresponding DNN feature pyramids F_1 and F_2, let $p = (x, y)$ be the grid coordinate in the feature pyramid, and $F_1(p,i)$ denotes the feature at p on the i^{th} level. Let w_i be the flow field on the i^{th} level, and $w_i(p) = (u_i(p), v_i(p))$ be the flow vector at p, where $u_i(p)$ and $v_i(p)$ are horizontal flow vector and vertical flow vector respectively.

Then, the DNN Flow's matching objective function can be formulated as:

$$E(w_i|w_{i-1},i) = \sum_p (E_D(p,w_i) + \alpha \sum_{q\in\varepsilon(p,i)} E_S(p,q,w_i) + \beta E_{SD}(p,w_i,w_{i-1})) \quad (1)$$

$$E_D(p,w_i) = |F_1(p,i) - F_2(p+w_i(p),i)| \quad (2)$$

$$E_S(p,q,w_i) = |u_i(p) - u_i(q)| + |v_i(p) - v_i(q)| \quad (3)$$

$$E_{SD}(p,w_i,w_{i-1}) = |u_i(p) - \widetilde{u}_{i-1}(p)| + |v_i(p) - \widetilde{v}_{i-1}(p)| \quad (4)$$

where E_D, E_S, E_{SD} are the data term, smoothness term and small displacement term respectively, $\varepsilon(p,i)$ is the neighborhoods of p on the i^{th} level, $(\widetilde{u}_{i-1}, \widetilde{v}_{i-1})$ is the w_{i-1} mapped to i^{th} level based on mapping of DNN. E_D measures the similarity between the correspondence features on the same level. E_S leverages the geometric prior that neighbors' flow vectors should be similar. E_{SD} uses the flow field of upper level to guide the optimization of low-level flow field.

We build a four-level pyramid to estimate dense correspondences. The DNN used for extracting DNN feature is learned by supervised back propagation on ILSVRC2012 training set, which contains eight layers with weights: five convolutional layers followed by three fully-connected layers. Three max-pooling layers are used following the first, seconde and fifth convolutional levels. The output of fifth convolutional layer is adopted as top-level feature, while the outputs of second and first convolutional layer are adopted as two mid-level features. In order to extract bottom-level feature for each pixel, the dense output of first convolutional layer is adopted as bottom-level feature through adjusting stride.

The performance of DNN Flow is demonstrated based on three experiments: rough image dense matching, fine object alignment and label transfer. The experiments are designed repectively on different datasets. Three image matching approaches, PatchMatch, SIFT FLow and DSP, are compared with DNN Flow in all experiments. The selected approaches are based on local feature or hierarchical local feature. In order to quantitatively evaluate image matching, two evaluation metrics are introduced into experiment: label transfer accuracy (**LT-ACC**) metric and intersection over union.

[1] Yan Li, Leon Gu, and Takeo Kanade. Robustly aligning a shape model and its application to car alignment of unknown pose. *PAMI*, 33 (9):1860–1876, 2011.

[2] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*. 2008.

[3] Michael Rubinstein, Ce Liu, and William T Freeman. Annotation propagation in large image databases via dense image correspondence. In *ECCV*. 2012.

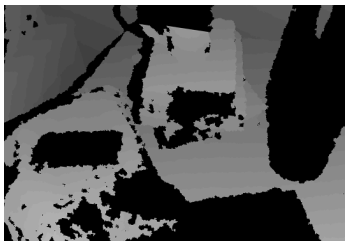
Board #77

Improved Depth Recovery In Consumer Depth Cameras via Disparity Space Fusion within Cross-spectral Stereo

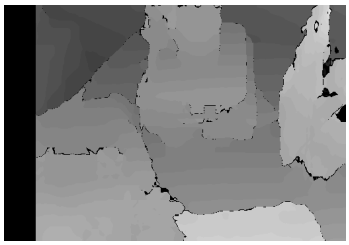
Gregoire Payen de La Garanderie
gregoire@hochet.info
Toby P. Breckon
toby.breckon@durham.ac.uk
School of Engineering,
Cranfield University, Bedfordshire, UK
School of Engineering and Computing Sciences,
Durham University, Durham, UK

(a) RGB Colour

(b) Infrared (normalised)



(c) SL disparity (from RGB-D camera)



(d) CS-DSI disparity (via proposed approach)

Figure 1: Fused disparity estimation:- $(SL, CS) \rightarrow CS-DSI$

Low-cost consumer depth cameras have seen the combined use of colour and 3D depth most commonly leverage the use of near infrared structured light projection (830nm) with regular visible-band colour sensing (400-700nm) to provide co-registered colour (RGB) and depth (D) as combined RGB-D image components. The common physical characteristics of such devices - comprising a colour camera (Fig. 1(a)), an infrared pattern projector and corresponding infrared camera (Fig. 1(b)) (e.g. Microsoft Kinect / PrimeSense Carmine) - pose an obvious, yet commonly under-utilized, cross-modal stereo configuration.

Depth coverage in consumer depth cameras can be considerably improved based on the combined use of such cross-spectral stereo (CS) and near infrared structured light sensing (SL). Our joint approach, leveraging disparity information from both structured light and cross-spectral stereo, facilitates the recovery of global scene depth comprising both texture-less object depth, where stereo sensing commonly fails, and highly reflective object depth, where structured light active sensing commonly fails. The proposed solution is illustrated using dense gradient feature matching and is shown to outperform prior approaches [1, 2] that use late-stage fused cross-spectral stereo depth as a facet of improved sensing for consumer depth cameras.

We propose the use of *"best in class"* dense gradient features from [3] to facilitate recovery of secondary cross-spectral stereo disparity directly from the depth camera $\{I_{ir}, I_{RGB}\}$ image pair. Our main contribution is the fusion of this secondary cross-spectral (CS) disparity information with *a priori* disparity information, obtained via conventional structured light (SL) sensing, within the disparity space image (DSI) constructed prior to conventional disparity optimization for scene depth recovery.

This is achieved by modifying the *disparity space image*, formed by $C(x,y,d)$, which constitutes the disparity cost space over which disparity optimization will be performed. We construct an alternative cost function, $C_{DSI}(x,y,d)$ such that the use of disparity from structured light sensing, $D_{SL}(x,y)$, is incorporated as follows:

$$C_{\text{DSI}}(x,y,d) = \begin{cases} C_{HOG}(x,y,d) & \text{if } D_{SL}(x,y) \text{ is unavailable at pixel } (x,y) \\ low_c & \text{if } d = D_{SL}(x,y) \\ high_c & \text{if } d \neq D_{SL}(x,y) \end{cases} \quad (1)$$

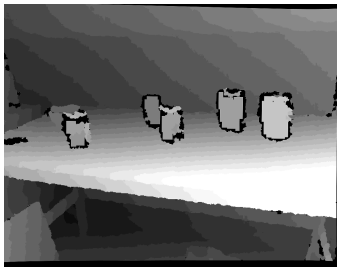
Figure 1 shows both the disparity recovered by conventional structured light (SL) within such a device (Fig. 1(c)) and that recovered by our proposed cross-spectral disparity space image (CS-DSI) approach (Fig. 1(d)) for a given $\{I_{ir}, I_{RGB}\}$ image pair (Fig. 1(a)) / 1(b)).

Figure 2 shows the disparity results obtained from the $\{I_{RGB}, I_{ir}, I_{depth}\}$ triplet shown in Fig. 2(a) - 2(c) for both the prior work of [1] (CS-union, Fig. 2(e)) and the proposed CS-DSI approach (Fig. 2(f)) against illustrative ground truth depth derived using manual depth labelling (Fig. 2(d)). The resulting CS-DSI disparity (Fig. 2(f)) presents a clearer disparity image with notably less missing disparity values and noise than CS-union (Fig. 2(e)) and the original SL disparity (Fig. 2(c)). This quality improvement resulting from CS-DSI is also present within Fig. 1.

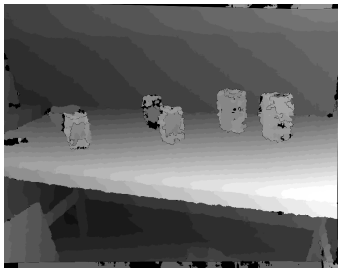
(a) RGB Colour

(b) Infrared (normalised)

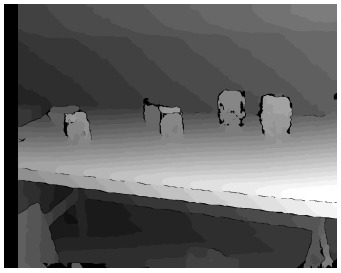
(c) SL disparity (from RGB-D camera)



(d) Ground truth disparity



(e) CS-Union disparity (prior work, [1])



(f) CS-DSI disparity (via proposed approach)

Figure 2: Disparity recovery on transparent and specular objects

Improved disparity can be recovered from a consumer depth camera based on the fusion of cross-spectral stereo and existing structured light sensing performed prior to conventional disparity space optimization. Missing depth information is recovered for transparent and specular objects in addition to that missing due to inter-object occlusions. This directly extends prior work [1, 2] which is shown to produce lesser depth recovery and requires computationally expensive scene dependant optimization.

[1] W. C. Chiu, U. Blanke, and M. Fritz. Improving the Kinect by Cross-Modal Stereo. In *Proceedings of the British Machine Vision Conference*, pages 1–10, 2011. doi: 10.5244/C.25.116.

[2] W. C. Chiu, U. Blanke, and M. Fritz. I spy with my little eye: Learning optimal filters for cross-modal stereo under projected patterns. *Proc. IEEE International Conference on Computer Vision Workshops*, pages 1209–1214, 2011. doi: 10.1109/ICCVW.2011.6130388.

[3] P. Pinggera, T.P. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. In *Proc. British Machine Vision Conference*, pages 526.1–526.12, September 2012. doi: 10.5244/C.26.103.

Action Recognition by Weakly-Supervised Discriminative Region Localization

Hakan Boyraz[12]
hakanb@amazon.com
Syed Zain Masood[13]
zainmasood@sighthound.com
Baoyuan Liu[1]
bliu@cs.ucf.edu
Marshall Tappen[12]
tappenm@amazon.com
Hassan Foroosh[1]
foroosh@cs.ucf.edu

[1] Department of EECS
University of Central Florida
Orlando, FL USA
[2] Amazon, Inc. *
Seattle, WA USA
[3] Sighthound, Inc.
Orlando, FL USA

In this paper, we present an action recognition system that *automatically* locates discriminative regions within a video and then uses information from these regions to classify the action being performed. The system is trained in a weakly supervised manner where the training data is annotated with only the action label i.e. no annotation of discriminative regions is provided.

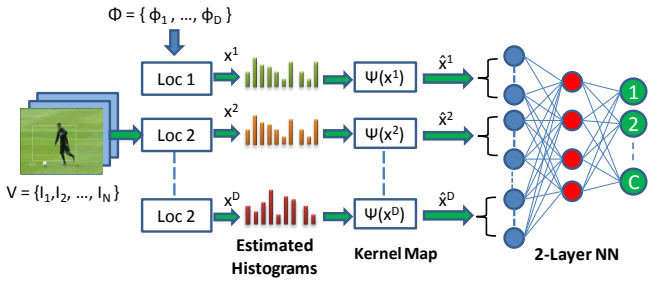


Figure 1: Our proposed framework for localizing discriminative regions and recognizing actions.

Figure 1 shows our proposed weakly-supervised framework for localizing discriminative regions and recognizing actions. The first step in recognizing the action is localizing discriminative sub-regions that best describe the action. These candidates are selected using a set of D discriminative sub-region localizers. A localizer ϕ_d, learned during training, is a vector of parameters describing the probability distribution of a latent location variable. Even though localizers are not associated with any action class explicitly, using multiple localizers allows the model to select different regions in each frame to capture variations in classes. For every sub-region in each frame of the video, localizers compute the probability of that sub-region being the most discriminative in that frame as follows:

$$p^f(r;\phi_d) = \frac{\exp\left(\phi_d^\top h_{f,r}\right)}{\sum_{r' \in R_f} \exp\left(\phi_d^\top h_{f,r'}\right)} \quad (1)$$

where $h_{f,r}$ denotes the histogram describing the frequency of visual words in the sub-region r contained in frame f and R_f is the set of all possible sub-regions in the frame. The final feature representation for localizer $d \in D$, denoted x_d, is obtained by aggregating the region histograms over all frames:

$$x_d(\phi_d) = \sum_{f \in F} \sum_{r \in R_f} h_{f,r} p^f(r;\phi_d), \quad (2)$$

The estimated histograms are then transformed to a high-dimensional feature space using Kernel Map and used as inputs to a two-layer neural network where the second layer is a C-way softmax classifier. Picking the class corresponding to the highest probability gives us our final classification.

While the focus of our approach is to find the most discriminative regions for action classification and not specifically the location of the actor in the video, our experiments on UCF Sports show that this method selects the actor location as the discriminative region with an accuracy comparable to systems trained explicitly for action localization on manually annotated data, as shown in Figures 2 and 3.

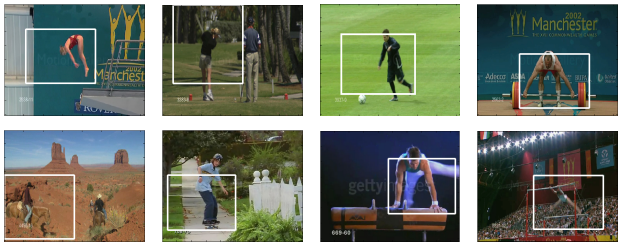


Figure 2: Localization results obtained using our method on the UCF Sports action dataset.

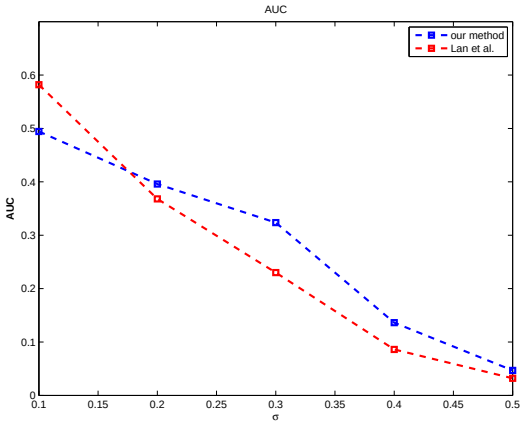


Figure 3: Comparison of action localization performance against Lan et al. [1].

Finally, Table 1 shows the comparison of our method with the global bag-of-words (BOW) model on HMDB and UCF101 datasets.

Method	HMDB	UCF101
Global BOW [HOG/HOF]	21.0%	43.94%
Global BOW [MBH]	36.6%	65.28%
Our Method [HOG/HOF]	29.56%	53.35%
Our Method [MBH]	45.29%	74.24%
Our Method [Combined]	47.24%	78.77%

Table 1: Comparison of our method with global BOW on HMDB and UCF101 datasets.

[1] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011.

* This work was performed while the authors were at the University of Central Florida.

Board #79

Adaptive Structured Pooling for Action Recognition

Svebor Karaman[1]
svebor.karaman@unifi.it
Lorenzo Seidenari[1]
lorenzo.seidenari@unifi.it
Shugao Ma[2]
shugaoma@bu.edu
Alberto Del Bimbo[1]
alberto.delbimbo@unifi.it
Stan Sclaroff[2]
sclaroff@bu.edu

[1] MICC (Media Integration and Communication Center)
University of Florence
Florence, Italy
[2] Boston University
Boston, USA

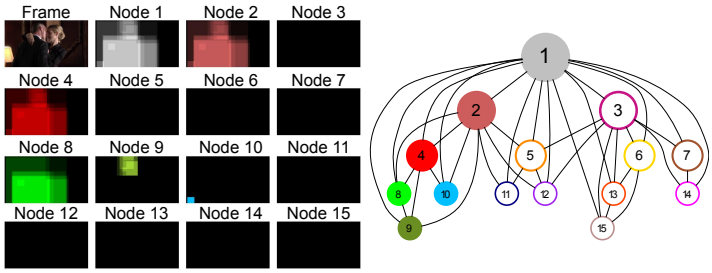


Figure 1: Overview of our method. Left: a frame of the "kiss" action of the HighFive dataset and pooling maps plots. Right: the video structure graph where nodes are spatio-temporal pooling regions at different granularities. Note how node 9 selects both actors faces.

We propose an adaptive structured pooling strategy to solve the action recognition problem in videos. Our method aims at individuating several spatio-temporal pooling regions each corresponding to a consistent spatial and temporal subset of the video. Each of them gives a pooling weight map and is represented as a Fisher vector computed from the soft weighted contributions of all dense trajectories evolving in it. We further represent each video through a graph structure, defined over multiple granularities of spatio-temporal subsets. The graph structures extracted from all videos are compared with an efficient graph matching kernel.

Soft pooling weigths. Given a set of Hierarchical Space-Time Segments (HSTS) [2] $\mathcal{S}_k$ we define a weighted pooling map M_k by accumulating how many segments of $\mathcal{S}_k$ are present in each frame at each position. For every pixel $p=(x,y)$ of frame t, we compute the pooling map value $M_k^t(p)$ as the count of segment enclosing this position $M_k^t = \sum_{s\in\mathcal{S}_k^t} \Psi_s$ where for each segment $s \in \mathcal{S}_k^t$ we define the function $\Psi_s(p)=1$ if $p \in s$ and $\Psi_s(p)=0$ otherwise.

The pooling map M_k^t is further normalized by the total number of segments in the frame and square-rooted. This pooling maps represent at any moment of the video, how much each pixel is relevant with respect to the set $\mathcal{S}_k$. The more segments overlap in one position the more likely this pixel is significant for the action taking place. Finally, for a video with T frames we define the spatio-temporal pooling map as:

$$M_k(x,y,t) = \left\{ M_k^1(x,y) \dots M_k^T(x,y) \right\} \tag{1}$$

For each local feature to be encoded, we estimate the weight with respect to to set $\mathcal{S}_k$ as a small local integral of the pooling map M_k around its centroid. That is for each $x_m \in X$ with the spatio-temporal coordinates of its centroid being $(x_{x_m}, y_{x_m}, t_{x_m})$, w_m^k is estimated as:

$$w_m^k = \int_{x_{x_m}-v_x}^{x_{x_m}+v_x} \int_{y_{x_m}-v_y}^{y_{x_m}+v_y} \int_{t_{x_m}-v_t}^{t_{x_m}+v_t} M_k(x,y,t)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}t \tag{2}$$

Finally, all weights of a pooling region are normalized to sum to one in order to have comparable representation no matter how many number of features are present in the region. We obtain soft-pooling by using the weight w_m^k of each feature $x_m \in X$ within the soft Fisher encoding formulation (see eq. 3 and 4).

Fisher encoding with soft pooling. Given the Gaussian Mixture Model (GMM) $u_\lambda = \sum_{n=1}^N \omega_n u_n(x;\mu_n,\sigma_n)$ and the M features of X, we compute for each component u_n the mean $\mathcal{G}_n^\mu(X)$ and covariance elements $\mathcal{G}_n^\sigma(X)$ of a Fisher vector as:

$$\mathcal{G}_n^\mu(X) = \frac{1}{\sqrt{\omega_n}} \sum_{m=1}^M w_m \gamma_n(x_m) \left(\frac{x_m-\mu_n}{\sigma_n} \right), \tag{3}$$

$$\mathcal{G}_n^\sigma(X) = \frac{1}{\sqrt{2\omega_n}} \sum_{m=1}^M w_m \gamma_n(x_m) \left(\frac{(x_m-\mu_n)^2}{\sigma_n^2} - 1 \right), \tag{4}$$

where $\gamma_n(x_m)$ is the posterior probability of the feature x_m for the component n of the GMM and w_m is the weight obtained from eq. 2.

Spatio-temporally structured pooling of a video. We want to build a structured representation of each video. We propose to find coherent subsets by grouping together segments according to their overlap. This will create a set of local (both spatially and temporally) pooling regions.

We first compute an affinity matrix A of all segments $\mathcal{S}$ of a video. The affinity of two segments s_i (alive from frame t_{is} to t_{ie}) and s_j (alive from frame t_{js} to t_{je}) is computed as:

$$A(s_i,s_j) = \frac{1}{\min(t_{ie}-t_{is}, t_{je}-t_{js})} \sum_{t\in[\max(t_{is},t_{js}),\min(t_{ie},t_{je})]} \frac{s_i^t \cap s_j^t}{s_i^t \cup s_j^t}. \tag{5}$$

Given this affinity matrix we run the normalized cuts algorithm to obtain the subsets of segments. Instead of choosing one fixed number of subsets, we use multiple increasing sizes that will each provide a set of finer local representations of the video. We represent each HSTS cluster as a node in the graph, and each node attribute is the soft pooling of dense trajectories features weighted by the map computed on all segments of this cluster. We link clusters based on their overlap, we create a link between all clusters that have at least a pair of overlapping segments (even partially). An illustration of one video graph is shown in Figure 1. To compare the video graphs we use the efficient GraphHopper kernel from [1].

Conclusions. Our structured representation is adaptive to the content of the video and does not rely on a fixed partition of neither space nor time. We exploit an unsupervised procedure to generate a structured representation of the video. Our representation jointly models the hierarchical and spatio-temporal relationship of videos without imposing a strict hierarchy.

Experiments conducted on two standard datasets for action recognition show a significant improvement over the state-of-the-art. We obtain **65.4%** mean AP on HighFive dataset and **90.4%** mean per class accuracy on UCF Sports dataset. In the future, we would like to see if our structured representation could also be used to solve the action localization problem by identifying the paths and/or nodes that are most relevant for the action.

[1] Aasa Feragen, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne, and Karsten Borgwardt. Scalable kernels for graphs with continuous attributes. In *Advances in Neural Information Processing Systems*, pages 216–224, 2013.

[2] Shugao Ma, Jianming Zhang, Nazli Ikizler-Cinbis, and Stan Sclaroff. Action recognition and localization by hierarchical space-time segments. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2013.

 Board #8o

Online Action Recognition via Nonparametric Incremental Learning

Rocco De Rosa
rocco.derosa@unimi.it

Nicolò Cesa-Bianchi
nicolo.cesa-bianchi@unimi.it

Ilaria Gori
ilaria.gori@iit.it

Fabio Cuzzolin
fabio.cuzzolin@brookes.ac.uk

Department of Mathematics "Federigo Enriques"
Università degli Studi di Milano, Milano, Italy

Dipartimento di Informatica
Università degli Studi di Milano, Milano, Italy

iCub Facility
Istituto Italiano di Tecnologia, Genova, Italy

Department of Computing and Communication Technologies
Oxford Brookes University, Oxford, UK

We introduce an *online action recognition system* that can be combined with any set of frame-by-frame feature descriptors. Our system covers the frame feature space with classifiers whose distribution adapts to the hardness of locally approximating the Bayes optimal classifier. An efficient nearest neighbour search is used to find and combine the local classifiers that are closest to the frames of a new video to be classified. The advantages of our approach are: *incremental training*, *frame by frame real-time prediction*, *nonparametric predictive modelling*, video segmentation for *continuous action recognition*, *no need to trim videos* to equal lengths and *only one tuning parameter* (which, for large datasets, can be safely set to the diameter of the feature space). Experiments on standard benchmarks (see Fig. 2 and Tab. 1) show that our system is competitive with state-of-the-art non-incremental and incremental baselines.

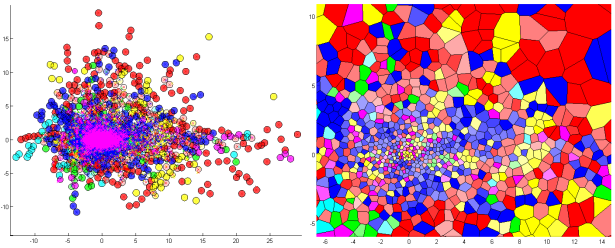


Figure 1: Left: the set of balls resulting from training on the first two principal components of local features extracted from the KTH dataset (colours denote labels, and color intensity expresses the 'purity' of the conditional class distribution within each ball). Right: a close-up of the central area represented as the Voronoi tessellation associated with the balls shows how the regions whose class statistics are more complex are covered by a finer set of balls.

Algorithm 1 ABACOC (Adaptive Ball Cover for Classification)

Input: Initial radius $R > 0$, metric ρ

Initialize set of ball centers $\mathcal{S} = \emptyset$ and set of labels $\mathcal{Y} = \emptyset$
for $i = 1, 2, \ldots$ **do**
  Receive labeled video (V_i, y_i)
  Create sequence of labeled frames $(x_1, y_i), \ldots, (x_{T_i-1}, y_i)$
  for $t = 1, \ldots, T_i - 1$ **do**
    if $\mathcal{S} \equiv \emptyset$ **then**
      $\mathcal{S} = \{x_t\}$, set $\varepsilon_t = R$, and use y_i to init. estimates p_t
    else
      Let $x_s \in \mathcal{S}$ be the nearest neighbour of x_t in $\mathcal{S}$
      if $\rho(x_s, x_t) \leq \varepsilon_s$ (x_t belongs to current ball centered on x_s) **then**
        if $y_i \neq \operatorname{argmax}_{c \in \mathcal{Y}} p_s(c)$ **then**
          Set $m_s = m_s + 1$ and update radius via $\varepsilon_s = R\, m_s^{-1/(2+d)}$
        end if
        Use y_i to update estimates p_s
      else
        $\mathcal{S} = \mathcal{S} \cup \{x_t\}$, set $\varepsilon_t = R$, and use y_i to init. estimates p_t
      end if
    end if
  end for
end for

The proposed method is a general framework for incremental *multivariate time series classification* (e.g. video frames) based on the following principles: (i) each video frame is a training example in a local feature space; (ii) incoming training examples are selected to cover the frame feature space with balls whose radius is adjusted according to the distribution of action classes within each ball; (iii) each ball is associated with an estimate of the conditional class probabilities, obtained by collecting statistics around its centre, which is used to make predictions on new unlabeled samples; (iv) the set of balls can be organized in a tree structure, allowing logarithmic queries in the number of balls. During training (see Alg. 1), a new ball is added whenever the input frame example does not belong to the ball whose center is the closest to the frame among the centers in the current set (Fig. 1, left). Otherwise, the ball statistics and its radius are updated. In the prediction phase, the conditional class probability estimates associated with the ball centre nearest to the input frames are used to select the action that maximises the sum of those scores (Fig. 1, right). The method allows us to work *incrementally* at frame level and *in real time*. Our learning method is also nonparametric. That is, the classifier structure is not pre-determined (as for linear classifiers), but it is inferred from the data (as for k-NN). As it handles videos on a frame-by-frame basis, the method is suitable to tackle the so-called "continuous action recognition" problem. To the best of our knowledge no other approach enjoys all these attractive features.

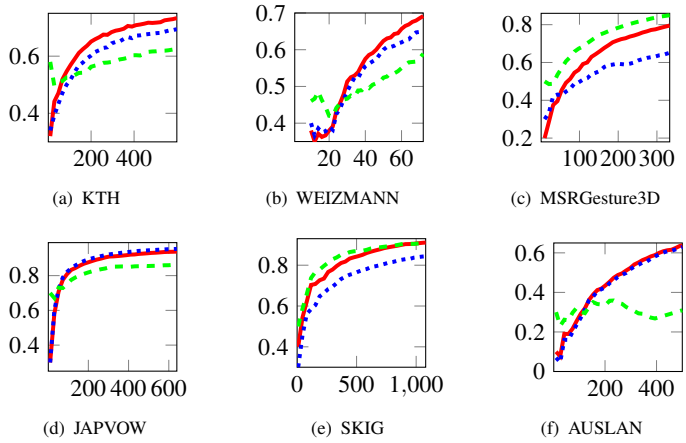


Figure 2: The plots show the online performance of ABACOC (red solid line) against SVM-b (green dashed line) and ALMA (blue dotted line). The x-axis is the number of videos fed to the algorithms and the y-axis is the average accuracy over the ten random permutations.

DATASET	HMM-1NN	DTW-d	SVM-b	ABACOC
KTH	68.28%	52.50%	69.83%	**83.20%**
Weizmann	87.50%	53.76%	97.22%	**98.61%**
SKIG	90.30%	95.74%	94.50%	**97.50%**
MSRGesture3D	78.20%	50.65%	**95.55%**	90.33%
JAPVOW	95.67%	69.72%	84.59%	**98.01%**
AUSLAN	67.07%	**83.81%**	44.78%	72.32%

Table 1: Multiclass accuracies of ABACOC compared against four baseline algorithms on the six benchmark datasets. All the methods share the same extracted features.

Single Image Dehazing Using Color Attenuation Prior

Qingsong Zhu[1]
qs.zhu@siat.ac.cn
Jiaming Mai[1, 2]
jiamingmai@163.com
Ling Shao[3]
ling.shao@ieee.org

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] South China Agricultural University, Guangzhou, China
[3] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, UK

We propose a simple but powerful prior, color attenuation prior, for haze removal from a single input hazy image. By creating a linear model for modelling the scene depth of the hazy image under this novel prior and learning the parameters of the model with a supervised learning method, the depth information can be well recovered. With the depth map of the hazy image, we can easily remove haze from a single image. Figure 1 shows an overview of the proposed dehazing method.

To describe the formation of a hazy image, the atmospheric scattering model is widely used and it can be expressed as follows:

$$\mathbf{I}(x) = \mathbf{J}(x)\mathrm{e}^{-\beta d(x)} + \mathbf{A}(1-\mathrm{e}^{-\beta d(x)}) \quad (1)$$

where $\mathbf{I}$ is the hazy image, $\mathbf{J}$ is the scene radiance representing the haze-free image, $\mathbf{A}$ is the atmospheric light, β is the scattering coefficient of the atmosphere and d is the depth of scene.

By doing a lot of experiments on the hazy images, we find the statistics that the density of the haze is positively correlated with the difference between the brightness and the saturation in a single hazy image. Since the haze density increases along with the change of scene depth in general, we can make an assumption that the depth of the scene is positively correlated with the density of the haze and we have:

$$d(x) \propto c(x) \propto v(x) - s(x) \quad (2)$$

As the difference between the brightness and the saturation can approximately represent the density of the haze, we boldly assume that the relationship among the scene depth d, the brightness v and the saturation s is linear. Based on this assumption, we can create a linear model as follows:

$$d(x) = \theta_0 + \theta_1 v(x) + \theta_2 s(x) \quad (3)$$

where d is the scene depth, v is the brightness, s is the saturation, and (θ_0, θ_1, θ_2) are the unknown linear coefficients.

In order to determine the coefficients (θ_0, θ_1, θ_2) accurately, a simple and efficient supervised learning method is used. The training data are necessary in the supervised learning method. A training sample consists of a hazy image and its corresponding ground truth depth map in our case. In order to obtain the accurate depth information as far as possible, we use the dehazing results of Kopf et al. [1] to make an inverse calculation to acquire the depth maps. In [1], Kopf used the city model from Bing to acquire the depths for the New York images and a plain 30-meter digital terrain model for the Yosemite images. To seek a solution that minimizes the difference between the scene depth $d(x)$ estimated by Equation (3) and the true depth, we minimize the following squared loss function:

$$L = \frac{1}{n|\omega|}\sum_{i=1}^{n}\sum_{j=1}^{\omega_i}(d_{ri}(x_j) - (\theta_0 + \theta_1 v_i(x_j) + \theta_2 s_i(x_j)))^2 \quad (4)$$

Here, n is the number of the training samples, ω_i is the size of the hazy image of the ith training sample, $|\omega|$ is the total number of the pixels of all the hazy images in the training set, d_{ri} is the depth map of the ith training sample, v_i and s_i are the brightness channel and the saturation channel of the hazy image of the ith training sample respectively. To facilitate the calculation, we first define the two matrices $\mathbf{X}$ and $\boldsymbol{\theta}$, and combine all the d_{ri} into a vector $\mathbf{D}$ as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ v_1 & v_2 & \cdots & v_n \\ s_1 & s_2 & \cdots & s_n \end{bmatrix}^{\mathrm{T}} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} d_{r1} & \cdots & d_{rn} \end{bmatrix}^{\mathrm{T}} \quad (5)$$

Now we can rewrite Equation (4) in a more concise way as below:

$$L = \frac{1}{n|\omega|}(\mathbf{D} - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{D} - \mathbf{X}\boldsymbol{\theta}) \quad (6)$$

The problem of estimating the linear coefficients (θ_0, θ_1, θ_2) can be converted into the problem of solving the following equation:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{2}{n|\omega|}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} - \frac{2}{n|\omega|}\mathbf{X}^{\mathrm{T}}\mathbf{D} = 0 \quad (7)$$

The solution of the equation above is given by:

$$\boldsymbol{\theta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D} \quad (8)$$

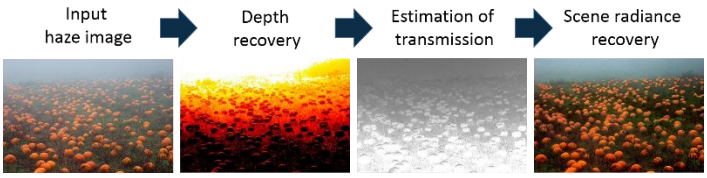


Figure 1: An overview of the proposed dehazing method.

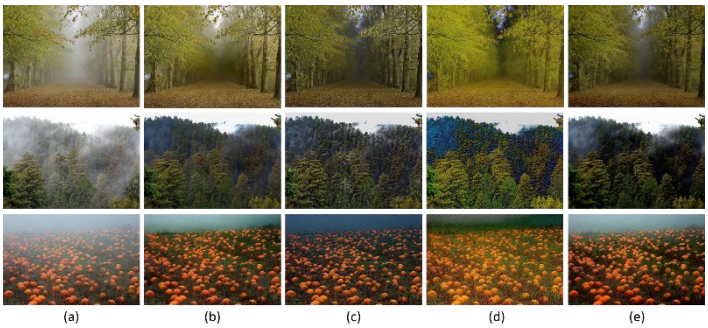


Figure 2: Comparison with other methods. (a-e) Hazy images, He et al.'s results [2], Tarel et al.'s results [3], Nishino et al.'s results [4] and ours.

We learn the linear coefficients according to Equation (8).

According to Equation (1), if $d(x) \to \infty$, then $\mathrm{e}^{-\beta d(x)} \to 0$ and $\mathbf{I}(x) = \mathbf{A}$. Based on this theory, we pick the top 0.1 percent brightest pixels in the depth map, and select the pixel with highest intensity in the corresponding hazy image $\mathbf{I}$ among these brightest pixels as the atmospheric light $\mathbf{A}$.

Now that the depth of the scene d and the atmospheric light $\mathbf{A}$ are known, we can recover the scene $\mathbf{J}$ in Equation (1). For convenience, we rewrite Equation (1) as follows:

$$\mathbf{J}(x) = \frac{\mathbf{I}(\mathrm{x}) - \mathbf{A}}{e^{-\beta d(x)}} + \mathbf{A} \quad (9)$$

where $\mathbf{J}$ is actually the haze-free image we want to obtain finally.

We implement the proposed method to test it on various hazy images and compare with the state-of-the-art methods. Figure 2 shows partial of the results. As can be seen, the dehazing effect of our method is outstanding. For an image of size $m \times n$, the complexity of the proposed dehazing algorithm is only O($m \times n$). In Table 1, we give the time consumption comparison with the state-of-the-art methods. As we can see, our approach is much faster than others and achieves the real-time requirement.

Image size	He [2]	Tarel [3]	Nishino [4]	Ours
600×450	12.2 s	8.2 s	104.7 s	**0.7** s
1024×768	36.9 s	69.3 s	317.4 s	**1.8** s
1536×1024	73.6 s	218.0 s	649.7 s	**3.0** s
1803×1080	90.7 s	351.1 s	861.4 s	**3.5** s

Table 1: Time consumption comparison.

All of these experimental results show that the proposed approach is highly efficient and it outperforms the state-of-the-art haze removal algorithms in terms of the dehazing effect as well.

[1] J. Kopt, B. Neubert, B. Chen, M. Cohen and D. Cohen-Or. Deep photo: Model-based photograph enhancement and viewing. *ACM Transactions on Graphics*, 27(5): 116, 2008.
[2] K. He, J. Sun and X. Tang. Guided image filtering. *IEEE TPAMI*, 35(6): 1397-1409, 2013.
[3] J. P. Tarel, and H. Nicolas. Fast visibility restoration from a single color or gray level image. In *Proc. ICCV*, 2009.
[4] K. Nishino, L. Kratz, and S. Lombardi. Bayesian defogging. *International journal of computer vision*, 98(3): 263-278, 2012.

Board #82

Fine-Grained Sketch-Based Image Retrieval by Matching Deformable Part Models

Yi Li
yi.li@qmul.ac.uk
Timothy Hospedales
t.hospedales@qmul.ac.uk
Yi-Zhe Song
y.song@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Queen Mary University of London
London, UK

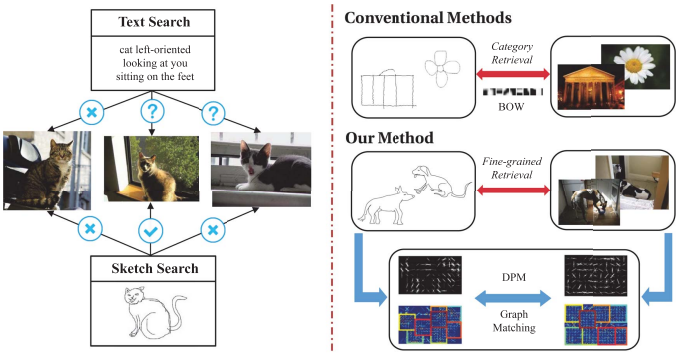


Figure 1: Comparison of traditional text-based image retrieval, conventional SBIR, and the proposed *fine-grained* SBIR framework.

Introduction Sketches are known to be able to capture object appearance and structure more intuitively and precisely than bare texts. However, to date the main focus of sketch-based image retrieval (SBIR) has been on retrieving photos of the same category, overlooking an important property of sketches — they can capture *fine-grained* variations of objects such as pose (standing vs. sitting) and iconic pattern (textures on a cow's body). By further leveraging this descriptive power of sketches, in this paper, for the first time we introduce *fine-grained* SBIR. That is to study how sketches can be used to differentiate *fine-grained* variations of objects for retrieval, specifically pose variations. Figure 1 contrasts text-based image retrieval and conventional SBIR with our proposed *fine-grained* SBIR.

Methodology Key to this problem is introducing a mid-level sketch representation that not only captures object pose, but also possesses the ability to traverse sketch and photo domains. Specifically, we learn deformable part-based model (DPM) [3] to discover and encode the various poses and parts in sketch and image domains independently, and employ graph matching [1] to establishing the correspondence between DPMs from different domains. The DPM is a two-layer structure, composed of root filter and part filters. We denote DPM as $M = (\mathbf{r}, G)$, where $\mathbf{r} = (w, h, f)$ specifies the width w, height h and global appearance feature of the root filter; and $G = (V, E, A)$ represents the star graph composed of the part filters. For the star graph G, V represents a set of nodes, E, edges, and A, attributes. Our matching objective for DPM accounts for both appearance and geometric information encoded in DPM, as well as both layers of representation, i.e., root filter $\mathbf{r}$ and part filter star graph G. Given two DPMs M^R and M^T, the similarity function is defined as:

$$S(M^R|M^T) = \gamma * S_{root}(M^R|M^T) + (1-\gamma) * S_{part}(M^R|M^T) \qquad (1)$$

where S_{root} is the root similarity and S_{part} is the part similarity; γ is a weighting factor balancing root and part similarities. The root filter similarity is generated considering appearance features, sizes and aspect ratios of the root filters, while the part similarity is solved as a graph matching problem on the part filter star graphs. The desired input of our proposed method is a sketch probe S with known category, and the output is a sequence of images from the same category ordered by their similarities with the probe S in terms of pose/appearance details. Achieving this *fine-grained* SBIR requires two major steps: (i) Training: DPM training and component alignment; (ii) Retrieval: *fine-grained* retrieval based on matching a probe sketch DPM detection with image DPM detections.

Experiment We propose an SBIR dataset by intersecting 14 common categories from the 20,000 sketch dataset [4] and PASCAL VOC dataset [2]. We divide the whole dataset into testing and training sets of the equal size. To enable quantitative evaluation, we manually annotate a subset of

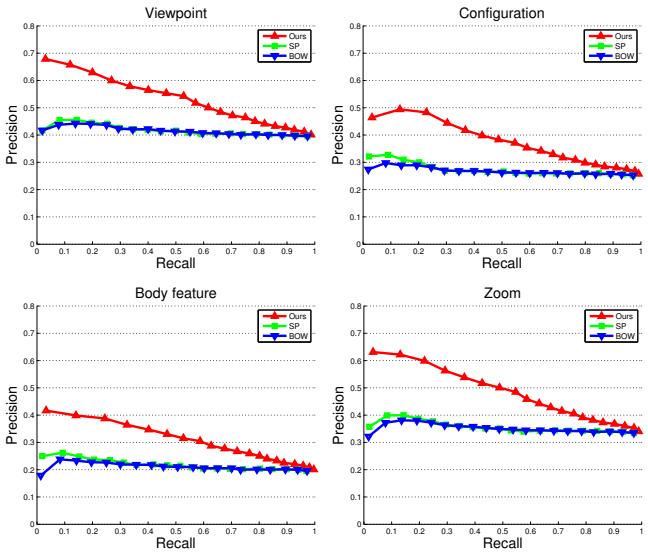


Figure 2: Precision-recall curves comparing bag-of-words (BOW), spatial pyramid (SP), and our method (Ours), using criterion: viewpoint, configuration, body feature, zoom separately.

Figure 3: Example retrievals of our method (Ours), spatial pyramid (SP) and bag-of-words (BOW). Ground truth similarity is also illustrated with the decomposition of viewpoint (V), configuration (C), body feature (B) and zoom (Z).

the testing set with exhaustive pairwise similarity ground-truth. For each sketch-image pair, we score their similarity in terms of four independent criteria: (i) viewpoint (V),(ii) zoom (Z),(iii) configuration (C), (iv) body feature (B). For each criterion, we annotate three levels of similarity: 0 for not similar, 1 for similar and 2 for very similar. The results in Figure 3 include some example annotations. We compare our method with conventional bag-of-words and spatial pyramid methods, both quantitative results (Figure 2) and qualitative results (Figure 3) have demonstrated our superior performance.

[1] M. Cho, J. Lee, and K. Lee. Reweighted random walks for graph matching. In *ECCV*, 2010.

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

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[4] E. Mathias, H. James, and A. Marc. How do humans sketch objects? *ACM TOG (Proceedings SIGGRAPH)*, 2012.

Board #83

Generalised Scalable Robust Principal Component Analysis

Georgios Papamakarios
georgios.papamakarios13@imperial.ac.uk
Yannis Panagakis
i.panagakis@imperial.ac.uk
Stefanos Zafeiriou
s.zafeiriou@imperial.ac.uk

Department of Computing
Imperial College London
London, UK

Real world visual data, while typically being very high-dimensional, often lie on a *low-dimensional* subspace. *Low-rank* is an attribute capturing the intrinsic low-dimensional structure of the data, when they are represented as column vectors of a matrix. Therefore, a natural approach in low-dimensional subspace recovery is to minimise the rank of the target matrix, subject to a constraint on the error in fitting the data.

By adopting the least squares error metric in fitting (i.e., assuming that the errors follow Gaussian distribution with small variance), the solution of the above mentioned rank minimisation problem is the classical Principal Component Analysis (PCA) [3]. However, visual data obeying postulated low-rank models may also contain gross errors and outliers to which the least squares metric is known to be sensitive.

To overcome the aforementioned drawbacks of the PCA, robust to gross but sparsely supported errors/outliers variants of the PCA have been proposed. With $\mathbf{X} \in \mathbb{R}^{F \times N}$ representing the data matrix, such methods aim to solve the following rank minimisation problem

$$\min_{\mathbf{E},\mathbf{A}} \operatorname{rank}(\mathbf{A}) + \lambda \|\mathbf{E}\|_0 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{A} + \mathbf{E}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{F \times N}$ is low-rank, $\mathbf{E} \in \mathbb{R}^{F \times N}$ is sparsely supported and accounts for gross errors/outliers and $\lambda > 0$ is a regularisation parameter.

Due to the discrete nature of the rank and the ℓ_0 quasi-norm, problem (1) is NP-hard and thus intractable. To overcome this, a convex relaxation is typically adopted, by surrogating the ℓ_0 quasi-norm of the fitting error matrix and the rank of the target matrix with their closest convex approximants, namely the ℓ_1-norm and the nuclear norm respectively. For instance, the RPCA [2] minimises $\|\mathbf{A}\|_* + \lambda \|\mathbf{E}\|_1$ subject to $\mathbf{X} = \mathbf{A} + \mathbf{E}$. The IRPCA [1] rewrites $\mathbf{A} = \mathbf{P}\mathbf{X}$ and minimises $\|\mathbf{P}\|_* + \lambda \|\mathbf{E}\|_1$ subject to $\mathbf{X} = \mathbf{P}\mathbf{X} + \mathbf{E}$. The active subspace RPCA [4] factorises $\mathbf{A} = \mathbf{U}\mathbf{V}$ with $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and minimises $\|\mathbf{V}\|_* + \lambda \|\mathbf{E}\|_1$ subject to $\mathbf{X} = \mathbf{U}\mathbf{V} + \mathbf{E}$.

Although the aforementioned nuclear/ℓ_1 norm-based methods mainly involve convex problems with global solutions, the relaxation may make the solutions seriously deviate from the original ones. Consequently, a better approximation to the original ℓ_0 quasi-norm-regularised rank minimisation problem (1) is necessary. In this paper, the *Generalised Scalable Robust PCA* (GSRPCA) is proposed, by reformulating the robust PCA problem using the Schatten p-norm $\|\cdot\|_{S_p}$ and the ℓ_q-norm $\|\cdot\|_q$ subject to orthonormality constraints. Let $\mathbf{U} \in \mathbb{R}^{F \times k}$ be column-orthogonal, such that $k \leq F$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and rewrite $\mathbf{A} = \mathbf{U}\mathbf{V}$. GSRPCA is formulated as the following non-convex optimisation problem

$$\min_{\mathbf{E},\mathbf{V},\mathbf{U}} \|\mathbf{V}\|_{S_p}^p + \lambda \|\mathbf{E}\|_q^q \quad \text{s.t.} \quad \begin{matrix} \mathbf{X} = \mathbf{U}\mathbf{V} + \mathbf{E} \\ \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{matrix} \tag{2}$$

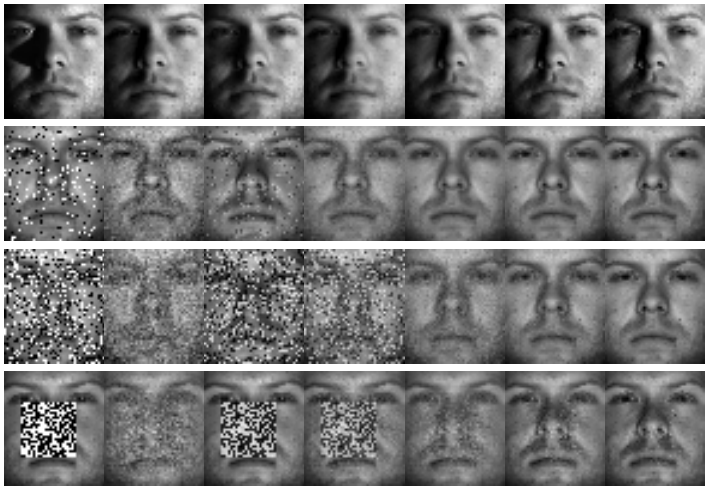
The column vectors of $\mathbf{U}$ can be interpreted as the principal components (base vectors) spanning the principal subspace and $\mathbf{V}$ as the projection of $\mathbf{X}$ onto the principal subspace. The state-of-the-art robust variants of the PCA in [1, 2, 4] are all special cases of the GSRPCA when $p = q = 1$ and by properly choosing the number k of principal components. The advantage of (2) is that, for $p \to 0$ and $q \to 0$, a closer approximation to the original rank minimisation problem in (1) can be achieved, by allowing the optimisation function to become non-convex, while retaining the scalability benefit introduced with the factorisation of $\mathbf{A}$.

An efficient alternating directions algorithm for GSRPCA is developed (Algorithm 1), based on the method of augmented Lagrange multipliers. The computational cost per iteration is dominated by 2 SVDs of size $k \times N$ and $F \times k$. Since for most applications typically $k \ll \min(F, N)$, the 2 SVDs can be computed in $\mathcal{O}\left(kN^2 + k^3\right)$ and $\mathcal{O}\left(kF^2 + k^3\right)$ respectively. In contrast, the RPCA [2] requires one SVD of size $F \times N$, which is $\mathcal{O}\left(NF^2 + N^3\right)$ per iteration (assuming $F \geq N$) and the IRPCA [1] requires one SVD of size $F \times F$, which is $\mathcal{O}\left(F^3\right)$ per iteration. Therefore, as long as k remains low, GSRPCA scales well to problems where F and/or N become large, contrary to RPCA and IRPCA.

Algorithm 1: Generalised Scalable Robust PCA

Input: Data matrix $\mathbf{X}$, number of components k, parameters p, q
Initialise: $\mathbf{U} = [k$ first singular vectors of $\mathbf{X}]$, $\mathbf{E} = \mathbf{Y} = \mathbf{0}$, $\mu = \frac{FN}{4\|\mathbf{X}\|_1}$

while *not converged* **do**
  Compute the SVD: $\mathbf{U}^T\left(\mathbf{X} - \mathbf{E} + \mu^{-1}\mathbf{Y}\right) = \mathbf{U}_S\mathbf{D}_S\mathbf{V}_S^T$.
  Update: $\mathbf{V} \leftarrow \mathbf{U}_S\mathcal{S}^p_{\mu^{-1}}\{\mathbf{D}_S\}\mathbf{V}_S^T$.
  Update: $\mathbf{E} \leftarrow \mathcal{S}^q_{\lambda\mu^{-1}}\left\{\mathbf{X} - \mathbf{U}\mathbf{V} + \mu^{-1}\mathbf{Y}\right\}$.
  Compute the SVD: $\left(\mathbf{X} - \mathbf{E} + \mu^{-1}\mathbf{Y}\right)\mathbf{V}^T = \mathbf{U}_S\mathbf{D}_S\mathbf{V}_S^T$.
  Update: $\mathbf{U} \leftarrow \mathbf{U}_S\mathbf{V}_S^T$.
  Update: $\mathbf{Y} \leftarrow \mathbf{Y} + \mu(\mathbf{X} - \mathbf{U}\mathbf{V} - \mathbf{E})$.
  Update: $\mu \leftarrow \min(\mu\xi, \mu_{max})$.
  Check convergence: $\|\mathbf{X} - \mathbf{U}\mathbf{V} - \mathbf{E}\|_F \leq \varepsilon\|\mathbf{X}\|_F$.
end while
Output: Principal components $\mathbf{U}$, projections $\mathbf{V}$, sparse errors $\mathbf{E}$



Original PCA RPCA IRPCA $p,q=1$ $p,q=0.5$ $p,q=0.1$

Figure 1: Denoising on the Extended Yale B database. 1st row: shadow removal; 2nd row: 10% salt & pepper noise; 3rd row: 30% salt & pepper noise; 4th row: random patch of maximum size 40×40. 1st column: original image; 2nd column: PCA; 3rd column: RPCA; 4th column: IRPCA; 5th–7th columns: GSRPCA with $p = q \in \{1,\ 0.5,\ 0.1\}$.

The performance of the GSRPCA is assessed by conducting experiments on both synthetic and real data (see for instance Fig. 1). The experimental results indicate that the GSRPCA outperforms the robust PCA methods [1, 2, 4] to which it is compared, without introducing much extra computational cost.

[1] B.-K. Bao, G. Liu, C. Xu, and S. Yan. Inductive robust principal component analysis. *IEEE Trans. Image Processing*, 21(8):3794–3800, 2012.

[2] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11–37, 2011.

[3] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educational Psychology*, 24:417–441,498–520, 1933.

[4] G. Liu and S. Yan. Active subspace: Toward scalable low-rank learning. *Neural Comput.*, 24(12):3371–3394, 2012.

Board #84

Solving Jigsaw Puzzles using Paths and Cycles

Lajanugen Logeswaran
lajanugenl.14@cse.mrt.ac.lk

Department of Electronic and Telecommunication Engineering
University of Moratuwa
Sri Lanka

There has been a growing interest in image jigsaw puzzles with square shaped pieces. A solver takes as input square shaped patches of the same size belonging to an image and attempts to reconstruct the image. The key components of a jigsaw solver are a compatibility metric and an assembly algorithm. A compatibility metric uses the color content of the image patches to identify which pairs of pieces are likely to be neighbors in the correct assembly. More specifically, given puzzle pieces x, y and a neighboring relationship $d \in \mathcal{D} = \{left, right, top, bottom\} = \{l, r, t, b\}$ a compatibility metric C assigns a numeric value $C(x, y, d)$ which represents how likely it is that piece y is the neighbor of piece x in the direction indicated by d. The assembly algorithm attempts to put the pieces together in the correct arrangement guided by these compatibility values. Prior work present several compatibility metrics and assembly algorithms.

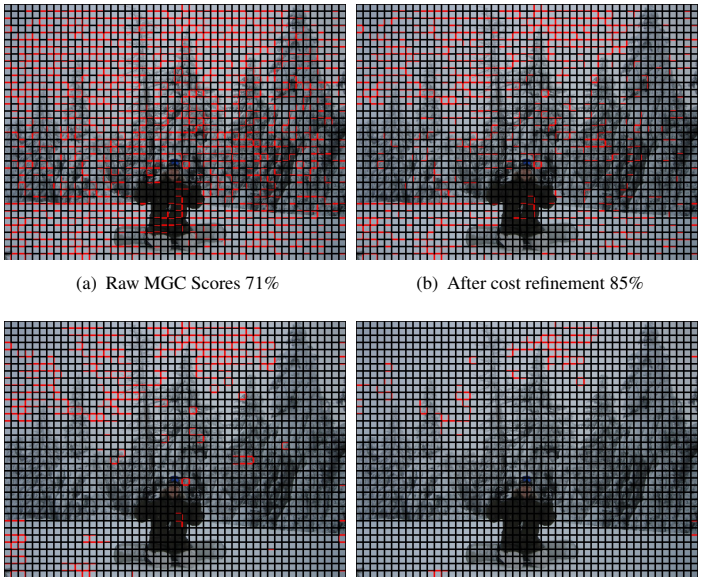
We propose techniques which attempt to exploit more contextual information provided by the compatibilty metric compared to previous work. We introduce the concept of paths and cycles in jigsaw puzzles and show that they provide a means of identifying correct and incorrect matches. Based on this concept we propose refinement techniques which incrementally modify the compatibility values suggested by a metric to improve its neighbor identification accuracy. We further propose a means of exploiting information provided by different compatibility metrics. We define a compatibility measure based on the idea of cycles and use it to guide a greedy solver. The solver beats state of the art performance and the improvements are significant in the more challenging situation of smaller piece size. We briefly discuss the proposed techniques below.

A neighbor matrix N represents point estimates $N(x, d)$ of the neighbor for each piece x and direction d. Based on the raw compatibility scores we may obtain these estimates as $N(x, d) = \text{argmin}_y C(x, y, d)$. The piece identified as the best candidate to be the top neighbor of x is $N(x, t)$. We may also observe that $N(N(N(x, l), t), r)$ is another estimate for the top neighbor of x. They may happen to be the same or different depending on the correctness of the entries in the neighbhor matrix (and whether x is located in the left or top borders in the correct assembly). In general, we may consider a sequence of directions $\mathbf{d} = (d_1, d_2, ..., d_n)$ to obtain beliefs about piece placement x_n at the location determined by $\mathbf{d}$, relative to a given piece x, where x_n is defined as following : $x_0 = x, x_i = N(x_{i-1}, d_i) \forall i \in \{1..n\}$. We define the sequence of pieces $(x_0, x_1, .., x_n)$ to be a ***path***, and say that the *links* (x_{i-1}, x_i, d_i) make up the path.

Consider the situation where the direction sequence $\mathbf{d}$ represents a closed curve (such as $(l, r), (l, t, r, b)$ etc.). For a path $(x_0, .., x_n)$ generated by such a direction sequence it has to be true that $x_n = x_0$ if all the links making up the path are correct. If not, we may conclude that atleast one of these links is incorrect. If the property does hold, intuitively this makes the constituent links likely to be correct. In this case we call the path a ***cycle***.

The idea of cycles motivated us to define an alternative measure of piece pair compatibility. We define the *strength* of a link (x, y, d) to be the number of cycles to which it belongs. This link strength measure guides our proposed techniques for improving the neighbor identification accuracy of a given compatibility metric. The proposed **cost refinement** technique iteratively modifies the scores suggested by a compatibilty metric in an attempt to use correctly and confidently identified piece neighbors to correct piece neighbors identified incorrectly. The proposed **neighbor refinement** procedure makes use of paths starting and ending at the same two pieces to repair incorrect entries in a given neighbor matrix.

Different compatibility metrics may use different image features and techniques to score piece pairs. There is no single metric which performs best for all types of pieces and puzzles. Although one may be dominant when considering the overall performance we found that different metrics taken together have more to offer than the individual metrics. We thus propose a means of **combining the strengths of multiple compatibility metrics** using the cycles idea. The incremental improvements in neighbor identification accuracy contributed by each of the aforementioned techniques are illustrated for a particular puzzle in Figure 1.



(a) Raw MGC Scores 71% (b) After cost refinement 85%

(c) After neighbor refinement 91% (d) After combining metrics 97%

Figure 1: Incorrect neighbor relationships (indicated by red markings on piece boundaries) after application of each improvement procedure - Image 11 of Cho et al.'s database [1] (piece size = 14, puzzle size = 1728)

(a) Pomeranz et al. [3] (b) Gallagher [2] (c) Ours

Figure 2: Visual comparison of puzzle assembly with two recently proposed solvers (piece size = 14, puzzle size = 1728)

Although high neighbor identification accuracies are favorable, the quality of puzzle assembly depends equally well on the assembly algorithm. In a greedy approach the order in which piece pairs are picked is important. Early mistakes may adversely affect assembly, depending on the robustness of the algorithm. While previous work have used the compatibility scores directly either to determine the order to pick piece pairs in a greedy approach or to define an energy function which is optimized, we use our link strength measure to guide a greedy solver. Significant improvements are observed in puzzle assembly compared to previous work, especially in the more challenging case of smaller piece size. Figure 2 compares our assembly procedure with two previously proposed algorithms on a puzzle instance.

We plan to explore further ways in which paths and cycles may be utilized to build robust solvers in future, complementing the limitations of compatibility metrics in identifying correct neighbor relationships.

[1] Taeg Sang Cho, Shai Avidan, and William T. Freeman. A probabilistic image jigsaw puzzle solver. In *CVPR*, pages 183–190. IEEE, 2010.

[2] Andrew C. Gallagher. Jigsaw puzzles with pieces of unknown orientation. In *CVPR*, pages 382–389. IEEE, 2012.

[3] Dolev Pomeranz, Michal Shemesh, and Ohad Ben-Shahar. A fully automated greedy square jigsaw puzzle solver. In *CVPR*, pages 9–16. IEEE, 2011.

Board #85

Parsing Semantic Parts of Cars Using Graphical Models and Segment Appearance Consistency

Wenhao Lu[1]
yourslewis@gmail.com
Xiaochen Lian[2]
lianxiaochen@gmail.com
Alan Yuille[2]
yuille@stat.ucla.edu

[1] Department of Electrical Engineering
Tsinghua University

[2] Department of Statistics
University of California, Los Angeles

Figure 1: The inputs (left) are images of a car taken from different viewpoints. The outputs (right) are the segmentation of car parts.

We attempt to parse cars into wheels, lights, windows, license plates and body, as illustrated in Figure 1. We formulate the problem as landmark identification. We first select representative locations on the boundaries of the parts to serve as landmarks. They are selected so that locating them yields the silhouette of the parts, and hence enables us to do object part segmentation (see Figure 2(a)). We use a mixture of graphical models to deal with different viewpoints so that we can take into account how the visibility and appearance of parts alter with viewpoint (see Figure 2(b)). We then use a mixture of graphical models to deal with different viewpoints so that we can take into account how the visibility and appearance of parts alter with viewpoint.

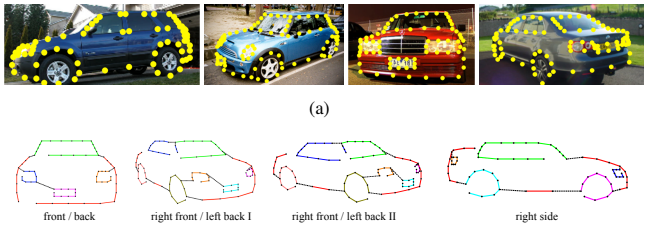


Figure 2: (a) The landmark annotations for cars of some viewpoints. (b) The proposed mixture-of-trees model. The landmarks connected by the solid lines of same colors belong to the same semantic parts. The black dashed lines show the links between different parts.

A novel aspect of our graphical model is that we couple the landmarks with the segmentation of the image to exploit the image contents when modeling the pairwise relation between neighboring landmarks. In the ideal case where part boundaries of the cars are all preserved by the segmentation, we can assume that the landmarks lie near the boundaries between different segments. Each landmark is then associated to the appearance of its two closest segments. This enables us to associate appearance information to the landmarks and to introduce pairwise coupling terms which enforce that the appearance is similar within parts and different between parts. We call this segmentation appearance consistency (SAC) between segments of neighboring landmarks. However, in practice, it is always impossible to capture all part boundaries using single level segmentation. Instead we couple the landmarks to a hierarchical segmentation of the image. We treat the level f the hierarchy for each part as a hidden variable, which is chosen *dynamically* during inference/parsing. By doing this, our model is able to automatically select the most suitable segmentation level for each part while parsing the image.

The model for each viewpoint is represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The nodes $\mathcal{V}$ correspond to landmark points. They are divided into subsets $\mathcal{V} = \bigcup_{p=1}^{N} \mathcal{V}_p$, where N is the number of parts and $\mathcal{V}_p$ consists of landmarks lying at the boundaries of semantic part p. The edge structures $\mathcal{E}$ are manually designed (see Figure 2(b)). Each node has pixel position of landmark $l_i = (x_i, y_i)$. The set of all positions is denoted by $\mathbf{L} = \{l_i\}_{i=1}^{|\mathcal{V}|}$. We denote by p_i the indicator specifying which part landmark i belongs to, and by $h(p)$ the segmentation level of part p. Then the segment pair of node i, $\mathbf{s_i}$, can be seen as the function of $h(p_i)$, which we denote by $\mathbf{s}_{i,h}$ for simplicity. Similar to the definitions of $\mathbf{L}$, we have $\mathbf{H} = \{h(p_i)\}_{i=1}^{N}$ and $\mathbf{S}(\mathbf{H}) = \{\mathbf{s}_{i,h}\}_{i=1}^{|\mathcal{V}|}$. The score function of the model for viewpoint v is

$$S(\mathbf{L}, \mathbf{H}, v \mid \mathbf{I}) = \phi(\mathbf{L}, \mathbf{H}, v \mid \mathbf{I}) + \psi(\mathbf{L}, \mathbf{H}, v \mid \mathbf{I}) + \beta_v \quad (1)$$

In the following we omit v for simplicity. The unary terms $\phi(\mathbf{L}, \mathbf{H} \mid \mathbf{I})$ is expressed as:

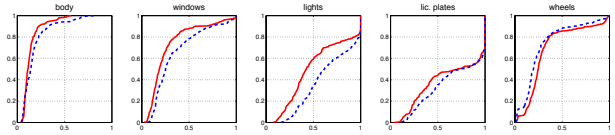
$$\phi(\mathbf{L}, \mathbf{H} \mid \mathbf{I}) = \sum_{i \in \mathcal{V}} \left[\mathbf{w}_i^f \cdot f(l_i \mid \mathbf{I}) + w_i^e e(h(p_i), l_i \mid \mathbf{I}) \right] \quad (2)$$

$\mathbf{w}_i^f \cdot f(l_i \mid \mathbf{I})$ measures the appearance evidence for landmark i at location l_i, where $f(l_i \mid \mathbf{I})$ is the HOG feature vector. The term $e(h(p_i), l_i \mid \mathbf{I})$ penalizes landmarks being far from edges. The binary term $\psi(\mathbf{L}, \mathbf{H} \mid \mathbf{I})$ is:

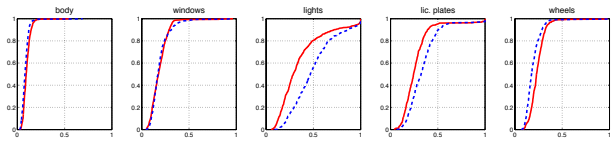
$$\psi(\mathbf{L}, \mathbf{H} \mid \mathbf{I}) = \sum_{(i,j) \in \mathcal{E}} \mathbf{w}_{i,j}^d \cdot d(l_i, l_j) + \sum_{\substack{(i,j) \in \mathcal{E} \\ p_i = p_j}} \mathbf{w}_{i,j}^A \cdot A(\mathbf{s}_{i,h}, \mathbf{s}_{j,h} \mid \mathbf{I}) \quad (3)$$

$d(l_i, l_j) = (-|x_i - x_j - \bar{x}_{ij}|, -|y_i - y_j - \bar{y}_{ij}|)$ measures the deformation cost for connected pairs of landmarks, where $\bar{x}_{ij}$ and $\bar{y}_{ij}$ are the anchor (mean) displacement of landmark i and j. We adopt L1 norm to enhance our model's robustness to deformation. In the second term of Equation 3, $A(\mathbf{s}_{i,h}, \mathbf{s}_{j,h} \mid \mathbf{I}) = (\alpha(s_{i,h}^1, s_{j,h}^1 \mid \mathbf{I}), \alpha(s_{i,h}^1, s_{j,h}^2 \mid \mathbf{I}), \alpha(s_{i,h}^2, s_{j,h}^1 \mid \mathbf{I}), \alpha(s_{i,h}^2, s_{j,h}^2 \mid \mathbf{I}))$ is a vector storing the pairwise similarity between segments of nodes i and j. This, together with the strength term $\mathbf{w}_{ij}^A$, models the SAC. Finally, β is a mixture-specific scalar bias. The parameters of the score function are $\mathcal{W} = \{\mathbf{w}_i^f\} \cup \{w_i^e\} \cup \{\mathbf{w}_{ij}^d\} \cup \{\mathbf{w}_{ij}^A\} \cup \{\beta\}$.

We validate our approach on a subset of PASCAL VOC2010 car images (VOC10) [1] and 3D car (CAR3D) [2]. The comparison with [3] are shown in Figure 3.



(a) PASCAL VOC 2010



(b) CAR3D

Figure 3: Cumulative segmentation error distribution for parts. X-axis is the average segmentation error normalized by image width, and Y-axis is the fraction of the number of testing images. The red solid lines are the performance using SAC and the blue dashed lines are from [3].

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[2] Silvio Savarese and Fei-Fei Li. 3d generic object categorization, localization and pose estimation. In *ICCV*, pages 1–8, 2007.

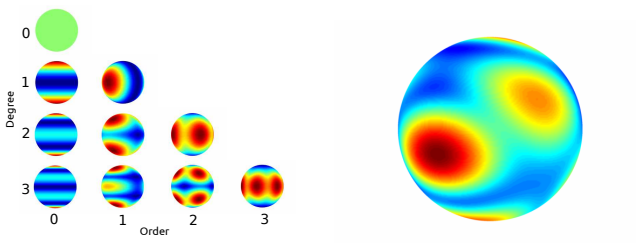
[3] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.

Board #86

An Image Based Approach to Recovering the Gravitational Field of Asteroids

Andrew Melim
Andrew.Melim@gatech.edu
Frank Dellaert
dellaert@cc.gatech.edu

College of Computing
Georgia Institute of Technology
Atlanta GA, 30332 USA



(a) (b)

Figure 1: **(a)** Gravitational field strength with harmonic coefficients of degree n and order m. **(b)** Vesta gravitational perturbations due to harmonics up to degree $n = 3$.

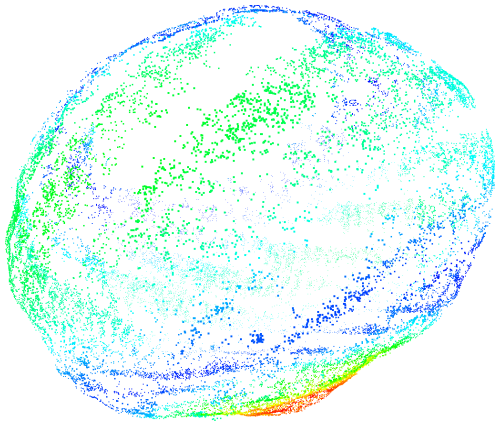
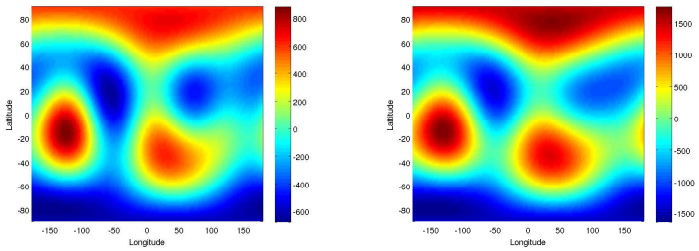


Figure 2: Vesta 3D Reconstruction (29143 landmarks) color mapped with our gravitational field results



(a) (b)

Figure 3: Gravity perturbation field results **(a)** VESTA20H solution with DSN tracking and optical landmarks **(b)** Our solution from a subset of HAMO-1 data using optical measurements only

This paper presents a pure vision based approach to solving for the gravitational field of extraterrestrial bodies with image data obtained by an orbiting spacecraft or satellite. Recovering a spacecraft's trajectory with modern day Structure from Motion approaches allows for further investigation for perturbations to accelerations due to variation in the strength of gravity. Understanding the variations of these forces, as well as developing a map, help to derive various models on the interior structure of the target planetary body or asteroid[1, 4].

Classical approaches for recovering the strength of a gravitational field study the motion of a satellite by tracking its position with Earth based telescopes. The basic principle behind this approach was developed in the field of satellite geodesy with the specific goal to define a highly accurate map of Earth's gravitational field. The same principle has not changed significantly, where the use of X-band Doppler and range measurements from a collection of Earth based radiometric tracking stations, known as the Deep Space Network, has been used to great effect.

In this paper, we introduce method to recover an estimate of the gravitational field without any need of radiometric tracking. We formulate constraints on a set of spherical harmonic coefficients, which defines a map of gravitational variations on a sphere, as shown in Figure 1, that integrate with graphical models used in modern Structure from Motion techniques [2, 3, 6]. Our approach is a complete image-based pipeline based around a two-step optimization that recovers 3D structure, spacecraft kinematics, and a gravitational model.

The basic process for gravity estimation is a two step iterative optimization. First, spacecraft pose and 3D landmark variables are estimated using batch bundle adjustment. The second step involves optimizing for the parameters of the gravitational field, in addition to camera pose velocities, using the local solutions found in step one. Here, tracking residuals are minimized with respect to global models.

Development of two key error terms for recovering the gravitational field are presented. First, a dynamics based gravitational potential function is used to compare the error of a point mass' orbiting trajectory with the recovered camera positions given a set of spherical harmonic coefficients. The spherical harmonics, commonly referred to as Stokes coefficients, define a basis for the gravitational model, similar to a Fourier series but instead map to the surface of a unit sphere. This is the key error term behind recovering the gravitational perturbations. A second error term based upon the Kaula power law constrains the magnitude of the harmonic coefficients as a function of their degree.

We evaluated our approach using camera data from the DAWN spacecraft's orbits around 4 Vesta, the second largest asteroid in the Solar System. Figure 2 shows our 3D reconstruction of Vesta color-mapped with the optimized gravitational perturbations.

Our approach, which only recovers up to degree three, develops an accurate representation of the accelerations when compared to the degree 20 NASA solution, referred to as VESTA20H [5], as seen in 3. Higher order terms governing the more complex structure are recovered more accurately than the lower degree coefficients. The contribution of higher degree terms $3 < N < 20$ are approximated in our solution by the lowest order terms, such as J_2, where we see the greatest difference with the VETSA20H solution.

[1] S.W. Asmar, A.S. Konopliv, R.S. Park, B.G. Bills, R. Gaskell, C.A. Raymond, C.T. Russell, D.E. Smith, M.J. Toplis, and M.T. Zuber. The gravity field of vesta and implications for interior structure. In *Lunar and Planetary Institute Science Conference Abstracts*, volume 43, page 2600, 2012.

[2] F. Dellaert and M. Kaess. Square Root SAM: Simultaneous localization and mapping via square root information smoothing. *Intl. J. of Robotics Research*, 25(12):1181–1203, Dec 2006.

[3] K. Konolige. Sparse sparse bundle adjustment. In *British Machine Vision Conf. (BMVC)*, September 2010.

[4] A.S. Konopliv, J.K. Miller, W.M. Owen, D.K. Yeomans, J.D. Giorgini, R. Garmier, and J-P. Barriot. A global solution for the gravity field, rotation, landmarks, and ephemeris of eros. *Icarus*, 160(2): 289–299, 2002.

[5] A.S. Konopliv, S.W. Asmar, R.S. Park, B.G. Bills, F. Centinello, A.B. Chamberlin, A. Ermakov, R.W. Gaskell, N. Rambaux, C.A. Raymond, et al. The vesta gravity field, spin pole and rotation period, landmark positions, and ephemeris from the dawn tracking and optical data. *Icarus*, 2013.

[6] M.I. A. Lourakis and A.A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36 (1):1–30, 2009. doi: http://doi.acm.org/10.1145/1486525.1486527.

Board #87

Incremental Domain Adaptation of Deformable Part-based Models

Jiaolong Xu[1,2]
jiaolong@cvc.uab.es
Sebastian Ramos[1,2]
sramosp@cvc.uab.es
David Vázquez[1]
dvazquez@cvc.uab.es
Antonio M. López[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

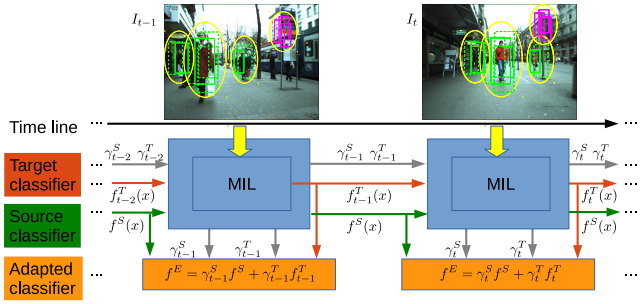


Figure 1: Incremental domain adaptation framework. $f_t^T(\mathbf{x})$ is the classifier trained by multiple instance learning (MIL) with current target image, while the final target-domain adapted classifier is f^E.

Algorithm 1 Incremental Domain Adaptation

Input:
source classifier f^S
target images $\{I_t, t \in [1,N]\}$
Output: $f^E = \gamma_N^S f^S + \gamma_N^T f_N^T$
0: $f_0^T \leftarrow f^S$, $\gamma_1^S = \gamma_1^T \leftarrow 0.5$
1: **for** t=1,2, ..., N, **do**
2: Receive image I_t, collect samples $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$.
3: Predict $\bar{y}_j^S$ by f^S, and $\bar{y}_j^T$ by f_{t-1}^T, $j \in \{1, N_t\}$.
4: Compute γ_t^S and γ_t^T by (2).
5: Generate training bags for MIL (see Figure 1).
6: Learn f_t^T with the collected bags (Eq. (3) and Eq. (4)).
7: **end for**

In this work we focus on performing an *incremental domain adaptation* of deformable part-based model (DPM) detectors [1]. The main benefit is to have an algorithm ready to improve existing source-oriented detectors as soon as a little amount of labeled target-domain training data is available, and keep improving as more of such data arrives in a continuous fashion.

We present our adaptation model as a weighted ensemble of source- and target-domain classifiers. This model is inspired in online transfer learning (OTL) [7]. Suppose we are given a set of training samples $(\mathbf{x}_1, y_1, \mathbf{h}_1), \ldots, (\mathbf{x}_N, y_N, \mathbf{h}_N) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$, where $\mathcal{X}$ is the input space, $\mathcal{Y} = \{+1, -1\}$ is the label space, and $\mathcal{H}$ is the hypothesis or output space. The DPM decision function can be written as $f(\mathbf{x}) = \max_{\mathbf{h} \in \mathcal{H}} \mathbf{w}' \Phi(\mathbf{x}, \mathbf{h})$, where $\Phi(\mathbf{x}, \mathbf{h})$ is a joint feature vector.

The basic idea is to learn an ensemble classifier $f^E(\mathbf{x})$ which is a weighted combination of the source domain classifier $f^S(\mathbf{x})$ and target domain classifier $f_t^T(\mathbf{x})$ at time t of the incremental learning task. We denote by γ_t^S and γ_t^T the combination coefficients. At the time t, given a sample $\mathbf{x}$, the ensemble decision function is written as follows:

$$f^E(\mathbf{x}) = \gamma_t^S f^S(\mathbf{x}) + \gamma_t^T f_t^T(\mathbf{x}), \tag{1}$$

where $f_t^T(\mathbf{x})$ is updated incrementally each time. Note that $f^S(\mathbf{x})$ and $f_t^T(\mathbf{x})$ are not independent as they maximize over the same $\mathbf{h}$ at training and testing time. In addition to updating $f_t^T(\mathbf{x})$, the two coefficients γ_t^S and γ_t^T are adjusted dynamically. The following updating scheme can be extended from OTL [7]:

$$\gamma_t^S = \frac{\gamma_{t-1}^S g_{t-1}(\bar{y}_i^S, y_i)}{\Gamma_{t-1}}, \quad \gamma_t^T = \frac{\gamma_{t-1}^T g_{t-1}(\bar{y}_i^T, y_i)}{\Gamma_{t-1}} \quad , \tag{2}$$

where $\Gamma_t = \gamma_t^S g_t(\bar{y}_i^S, y_i) + \gamma_t^T g_t(\bar{y}_i^T, y_i)$, $\bar{y}_i^S$ is the predicted label by f^S and $\bar{y}_i^T$ by f_{t-1}^T, $g_t(\bar{y}_i, y_i) = \frac{1}{N_t} \sum_{i=0}^{N_t} \exp\{-\frac{1}{2} l^*(\Pi(\bar{y}_i), \Pi(y_i))\}$, N_t is the number of target domain training samples at time t, $\Pi(s) = \max(0, \min(1, \frac{s+1}{2}))$ is a normalization function, and $l^*(\bar{y}, y) = (\bar{y} - y)^2$ is the square loss we use.

To train $f_t^T(\mathbf{x})$ in the target domain, we apply an incremental learning strategy similar to [2] under a frame-by-frame setting. Assume we receive an image I_t at time t and we learn f_t^T on that image by updating f_{t-1}^T learned at time $t-1$. Motivated by the online learning algorithms [3], we define f_t^T on instance $\mathbf{x}$ as follows:

$$f_t^T(\mathbf{x}) = \max_{\mathbf{h} \in \mathcal{H}} [\mathbf{w}_{t-1}' \Phi(\mathbf{x}, \mathbf{h}) + (\mathbf{w}_t' - \mathbf{w}_{t-1}') \Phi(\mathbf{x}, \mathbf{h})] = f_{t-1}^T(\mathbf{x}) + \Delta f_t^T(\mathbf{x}), \tag{3}$$

where $\Delta f_t^T(\mathbf{x})$ is the perturbation function. Given the training bags with instances $\mathbf{x}_i, \ldots, \mathbf{x}_{N_t}$, we learn the parameters $\mathbf{w}_t$ by minimizing the following objective function:

$$J(\mathbf{w}_t) = \frac{1}{2} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 + C \sum_{i=1}^{N_t} \mathcal{L}_{surr}(\mathbf{w}_t, \mathbf{x}_i, y_i, \mathbf{h}_i). \tag{4}$$

In some cases, the labels of target domain examples are weakly labeled, *e.g.*, pedestrian samples are collected by applying a pre-trained detector. We propose to handle weakly labeled examples by multiple instance learning (MIL) (see Figure 1), and the weakly labeled structured SVM (WL-SSVM) is used to train DPM by MIL. With the above learning strategy, f_t^T can be embedded into the OLT framework.The complete algorithm is presented in Alg. 1.

We evaluate the proposed method on several pedestrian datasets. We use a synthetic dataset [6] to train our source domain DPM detector, and adapt it to multiple real-world datasets [4, 5]. The incremental domain adaptation achieves comparable accuracy results to the batch learned model while being more flexible for learning with continuously coming target domain data. In the future, we plan to focus on improving the incremental domain adaptation with unlabeled target domain images.

[1] P. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *T-PAMI*, 32(9):1627–1645, 2010.

[2] W. Li, L. Duan, IW. Tsang, and D. Xu. Batch mode adaptive multiple instance learning for computer vision tasks. In *CVPR*, 2012.

[3] S. Shalev-shwartz, K. Crammer, O. Dekel, and Y. Singer. Online passive-aggressive algorithms. In *NIPS*, 2003.

[4] D. Vázquez, A.M. López, J. Marín, D. Ponsa, and D. Gerónimo. Virtual and real world adaptation for pedestrian detection. *T-PAMI*, 36 (4):797–809, 2014.

[5] J. Xu, S. Ramos, D. Vázquez, and A.M. López. Domain adaptation of deformable part-based models. *T-PAMI*, in press, 2014.

[6] J. Xu, D. Vazquez, A.M. Lopez, J. Marin, and D. Ponsa. Learning a part-based pedestrian detector in a virtual world. *T-ITS*, in press, 2014.

[7] P. Zhao and S. Hoi. OTL: A framework of online transfer learning. In *ICML*, 2010.

Contextual rescoring for Human Pose Estimation

Antonio Hernández-Vela[1]
ahernandez@cvc.uab.cat
Stan Sclaroff[2]
sclaroff@bu.edu
Sergio Escalera[1]
sergio@maia.ub.es

[1] Dept. of Applied Mathematics, Universitat de Barcelona, Spain
Computer Vision Center, UAB, Spain
[2] Dept. of Computer Science, Boston University, USA

Given an image of a person, the problem of human pose estimation can be briefly described as localizing the position and orientation of the body limbs. The complexity of the problem comes from issues like background clutter, changes in viewpoint, changes in appearance, self-occlusions of body parts, etc.

The pictorial structures framework [1] has been widely applied in human pose estimationn. Yang and Ramanan [7] proposed a simple yet efficient model that outperformed previous state of the art approaches. However, in addition to the difficulties of modelling small image patches for the body joints (see Fig. 1), the performance of their method is also compromised by the use of a tree-structured model. Although trees permit efficient and exact inference on graphical models, the restricted edge structure is insufficient for capturing all the important relations between parts.

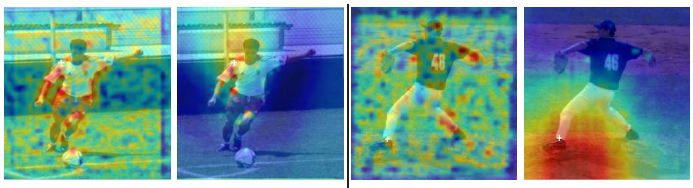


Figure 1: (a) Detection score map for the right shoulder using a classical sliding-window detection approach with a linear SVM trained on HOG features. (b) Rescored version of (a) produced by our context-based rescoring. The original map (a) has a strong score on the actual shoulder location, but also in other regions. Our proposed rescoring produces more spatially-consistent score maps, showing a high response near the correct location, and suppressing false positive locations. In addition, our rescoring method can hallucinate the location of a part, e.g. foot (d) even if there is not a high-scoring region in the original map (c).

In this work, we propose a new method for obtaining robust part detections in a pictorial structure formulation for human pose estimation. Motivated by the fact that small local HOG templates modelling the body joints ("basic parts" from now on) are sensitive to noise, we introduce information from a mid-level representation of the image in order to obtain more reliable basic part detections (see Fig. 2). More specifically, we make the following contributions:

- We introduce a method for the automatic discovery of a compact set of discriminative poselets [2] that offers both high detection precision and a covering of the different poses in a given validation dataset.
- Using this set of poselets as our mid-level image representation, we assign a new score to the detections of a certain basic part through a rescoring function that learns patterns of their contextual relationships.
- We extend the formulation from [7] in order to include the rescored detections.

Experimental evaluation is conducted on two benchmarks: UIUC Sports [6] and Leeds Sports [3]. In the experiments, pose estimation accuracy improves when our proposed rescoring functions are included in the unary potential of a pictorial structure model, using our mid-level part representation (see Fig. 3. In particular, among the different mid-level part representations in our comparative analysis, the automatic discovery of poselets with covering attains the best results in both datasets. In addition, we report a gain in the pose estimation performance comparable to the one in [4, 5], while reducing the size of the mid-level representation by an order of magnitude (40-50 poselets in our approach vs. more than 1000 in [4, 5]).

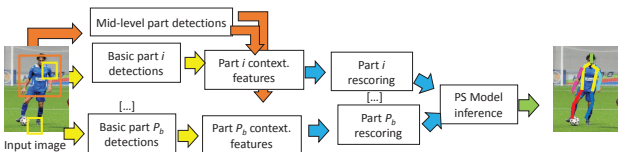


Figure 2: Proposed pipeline for human pose estimation. Given an input image, a set of basic and mid-level part detections is obtained. For each basic part *i* detection, a contextual representation is built based on mid-level part detections, which is used for rescoring the former. The original and rescored detections for all basic parts are then used in inference on a pictorial structure (PS) model to obtain the final pose estimate.

Figure 3: Qualitative results for the UIUC Sports dataset (row 1) and LSP dataset (row 2). Leftmost images show the results from [7] and rightmost images show our results.

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021. IEEE, 2009.

[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, volume 6316, pages 168–181, 2010.

[3] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, pages 12.1–11, 2010.

[4] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, pages 3487–3494, Dec 2013.

[5] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, pages 588–595, 2013.

[6] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, pages 1705–1712, 2011.

[7] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2878–2890, Dec 2013.

Board #89

Recognizing Image Style: Extended Abstract

Sergey Karayev[1]
Matthew Trentacoste[2]
Helen Han[1]
Aseem Agarwala[2]
Trevor Darrell[1]
Aaron Hertzmann[2]
Holger Winnemoeller[2]

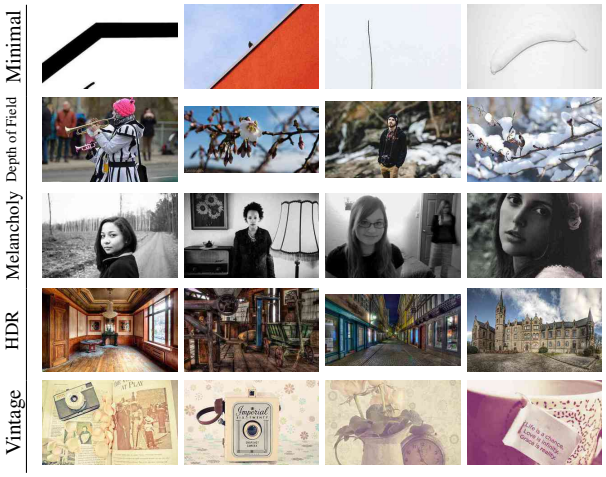
[1] University of California, Berkeley
[2] Adobe

Deliberately-created images convey meaning, and *visual style* is often a significant component of image meaning. For example, a political candidate portrait made in the lush colors of a Renoir painting tells a different story than if it were in the harsh, dark tones of a horror movie. While understanding style is crucial to image understanding, very little research in computer vision has explored visual style.

We present two novel datasets of image style, describe an approach to predicting style of images, and perform a thorough evaluation of different image features for these tasks. We find that features learned in a multi-layer network generally perform best – even when trained with object class (not style) labels. Our approach shows excellent classification performance on both datasets, and we use the learned classifiers to extend traditional tag-based image search to consider stylistic constraints.

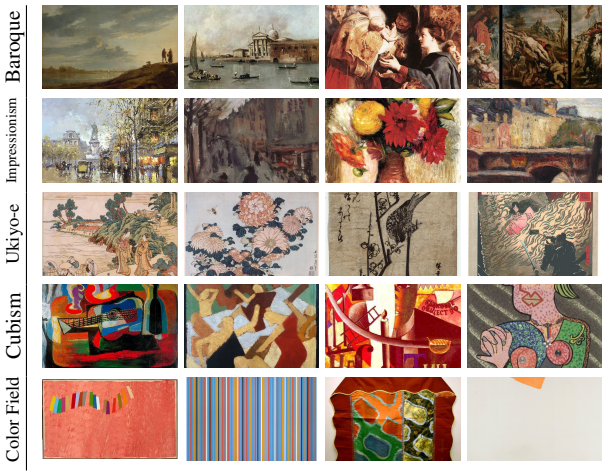
Flickr Style Using curated Flickr Groups, we gather 80K photographs annotated with 20 style labels, ranging from photographic techniques ("Macro," "HDR"), composition styles ("Minimal," "Geometric"), moods ("Serene," "Melancholy"), genres ("Vintage," "Romantic," "Horror"), to types of scenes ("Hazy," "Sunny").

Top five predictions on the test set for a selection of styles:



Wikipaintings Using community-annotated data, we gather 85K paintings annotated with 25 style/genre labels.

Top five predictions on the test set for a selection of styles:



Features and Learning We test the following features: **L*a*b color** histogram, **GIST** descriptor, Graph-based **visual saliency**, Meta-class binary (**MC-bit**) object features, and deep convolutional neural networks (CNN), using the Caffe implementation of Krizhevsky's ImageNet architeture (referred to as the **DeCAF** feature, with subscript denoting network layer). Notably, the last two of these are features designed and trained for object recognition.

As we hypothesize that style features may be content dependent, we also train **Content** classifiers using the CNN features and an aggregated version of the PASCAL VOC dataset, and use them in second-stage fusion with other features.

Evaluation Mean APs on three datasets for the considered single-channel features and their second-stage combination. Only the clearly superior features are evaluated on the Flickr and Wikipaintings datasets.

	Fusion x Content	$DeCAF_6$	MC-bit	L*a*b* Hist	GIST	random
AVA Style	**0.581**	0.579	0.539	0.288	0.220	0.132
Flickr	**0.368**	0.336	0.328	-	-	0.052
Wikipaintings	**0.473**	0.356	0.441	-	-	0.043

We compare our predictors to human observers, using Amazon Mechanical Turk experiments, and find that our classifiers predict Group membership at essentially the same level of accuracy as Turkers. We also test on the AVA aesthetic prediction task, and show that using the "deep" object recognition features improves over state-of-the-art results.

Applications Example of filtering image search results by style. Our Flickr Style classifiers are applied to images found on Pinterest. The images are searched by the text contents of their captions, then filtered by the response of the style classifiers. Here we show top five results for the query "Dress."

Code & Data All data, trained predictors, and code are available at `http://sergeykarayev.com/recognizing-image-style/`.

Multi-model fitting based on Minimum Spanning Tree

Radwa Fathalla
http://www.aast.edu/cv.php?ser=36825
George Vogiatzis
http://www.george-vogiatzis.org

College of Computing and Information Technology,
Arab Academy for Science and Technology
School of Engineering and Applied Science,
Aston University

Simultaneous parametric estimation of multiple primitive geometric models plays a key role in the overall interpretation of complex 3d scenes. This is characterized in the literature as a problem of irregular sites with discrete labels, on which techniques of unsupervised classification and optimization can be applied. This paper presents a novel approach to the computation of primitive geometrical structures, where no prior knowledge about the visual scene is available and a high level of noise is expected. We based our work on the grouping principles of proximity and similarity, of points and preliminary models. The former was realized using Minimum Spanning Trees (MST), on which we apply a *stable alignment* and a *goodness of fit criteria*. As for the latter, we used *spectral clustering* of preliminary models. The algorithm can be generalized to various model fitting settings in which the spatial coherence constraint applies, without fine tuning of run parameters.

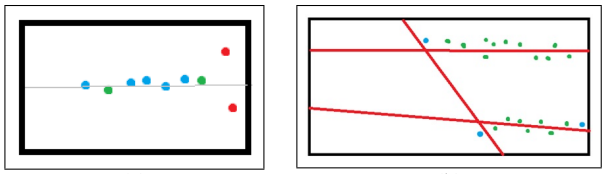
Stating our problem formally, let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ be a set of n data points. It is required to find $L = \{L_i\}_{i=1}^{M}$, such that L is a set of models that best describe$\mathbf{X}$. L_i is the parameter vector of model i which, together with the variable M are unknown a priori. In addition, the data points are contaminated by varying levels of outliers.

The literature of model fitting can be broadly categorized into the following subdivisions.

Energy-based formulation. Early attempts include the work in the RANSAC-adaptations to the multi-model case [5]. The initial randomly populated models compete according to some poor greedy heuristics which results in enhancing each model locally. Generally, the oversimplified single objective formulation overlooks cues that are inherent to the human visual system. These include, the compactness of points in areas that belong to the same model and the intuitive merging of adequately similar models. This gave rise to the need for regularized functions, as in PEARL [2]. It is inspired by the energy function of the incapacitated facility location problem (FLP) [4]. Mapped to our application, it incorporates a data cost and a cost for establishing a new label. They added a smoothness prior that ensures the spatial coherence in the search. The random initial set of hypotheses are verified using the α -expansion graph cut optimization. The main shortcomings of this paradigm are the determination of the trade off between the various energy terms and settling for approximate solutions to preserve computational feasibility causing it be susceptible to local minima. In addition, it is not difficult to find a counter example, as in figure 1 (a), for relying on absolute proximity to enforce spatial coherence (implying that possible inliers are closer to each other than to outliers).

Similarity-based formulation. This category exploits the fact that a structure can be detected by the presence of several entities sharing a certain property, defined upon a parameter, residual or conceptual space. The entities can be the given points, as the system proposed in [6]. An agglomerative algorithm clusters points based on the Jaccard distance and the final models are the best fits of these clusters. The points are expressed by their set of preferred models based on residuals. This can be very misleading in case of random generation of models that may result in cross structures (figure 1 (b)). Their assumption "Residuals for each data point have peaks corresponding to the true models" struggles in case of initial random sampling in high levels of outliers, because points can be equidistant to completely different models.

Our proposed algorithm provides a solution to the problems presented in figure 1 by relying on analysis of models layout in space, focusing on point arrangements rather than optimizing on residual values. It belongs to the category of model-based similarity formulation of the model fitting problem. To ensure the correct detection of clusters, the sampling of the hypothesized models should guarantee the repeated presence of optimal/sub-optimal models, in order to form agglomerated dense regions in some space. For this reason and because random sampling fails in this respect, we have resolved to the guided sampling paradigm. At each point we initiate a sample set. Gradually, this set is expanded by incorporating



(a) (b)

Figure 1: Snapshots of point arrangements, (a) showing 3 randomly formed models (lines) and some points scattered around them. The points in blue share very similar preference based on the cross structure despite the discrepancy in their true belonging; (b) showing 2 inliers in green with an in-between distance larger than the distances between one of them and the gross outliers in red.

more points that belong to the minimum spanning tree (MST) and we find the best fitting model. Due to the presence of geodesic paths between inliers of a model on its surface, we argue that MST-based sampling is more robust to varying noise levels than propagation-based ones. We introduce a novel *stable alignment* criterion to select the best size of the sample set that generates the model at the investigated point. It relies on the fact that at a certain phase, the model alignment is not drastically altered by the inclusion of gross outliers. Because, the growing size of the subtree enhances the spread of inlier points and the adherence of the generated model to the underlying structure. The subtree selection is further enhanced by the incorporation of the *margin of error* criterion, which geometrically indicates how well aligned and dense the points are in the consensus zone. For measuring the alteration in the model alignment, we introduce an arbitrary dissimilarity measure: *model deviation*. We have shown its superiority to the commonly utilized measure of Jaccard distance, with respect to being more linear and sensitive to small perturbations.

We construct a similarity matrix between populated models and then pass it to spectral clustering algorithm [3] to produce subsets. We handled the issue of the unknown number of models and subsequently clusters with the *Repeated 2-clustering method* in a top-down approach. The regularization function is the *Daives Bouldin* (DB) index [1]. Each cluster promotes its centroid model to the final set, defined as the model that is least dissimilar to the rest of models in the same cluster. Our algorithm was shown to outperform the state-of-the-art techniques, in some aspects.

[1] David L Davies and Donald W Bouldin. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227, February 1979. ISSN 0162-8828.

[2] Hossam Isack and Yuri Boykov. Energy-Based Geometric Multi-model Fitting. *Int. J. Comput. Vision*, 97(2):123–147, April 2012. ISSN 0920-5691.

[3] Francis R Bach Michael I Jordan. Learning spectral clustering. *Advances in Neural Information Processing Systems*, 16:305, 2004.

[4] Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. In *Proceedings of the 38th international conference on Automata, languages and programming*, ICALP'11, pages 77–88, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-22011-1.

[5] C.V. Stewart. Bias in Robust Estimation caused by discontinuities and multiple structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19 (8):331–338, 1997.

[6] Roberto Toldo and Andrea Fusiello. Robust Multiple Structures Estimation with J-Linkage. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 537–547, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-88681-5.

Board #91

Object Disambiguation for Augmented Reality Applications

Wei-Chen Chiu[1]
walon@mpi-inf.mpg.de
Gregory S. Johnson[2]
gregory.s.johnson@intel.com
Daniel Mcculley[2]
daniel.b.mcculley@intel.com
Oliver Grau[2]
oliver.grau@intel.com
Mario Fritz[1]
mfritz@mpi-inf.mpg.de

[1] Max Planck Institute for Informatics
Saarbrücken, Germany
[2] Intel Corporation

Abstract

The broad deployment of wearable camera technology in the foreseeable future offers new opportunities for augmented reality applications ranging from consumer (e.g. games) to professional (e.g. assistance). In order to span this wide scope of use cases, a markerless object detection and disambiguation technology is needed that is robust and can be easily adapted to new scenarios. Further, standardized benchmarking data and performance metrics are needed to establish the relative success rates of different detection and disambiguation methods designed for augmented reality applications.

Here, we propose a novel object recognition system that fuses state-of-the-art 2D detection with 3D context. We focus on assisting a maintenance worker by providing an augmented reality overlay that identifies and disambiguates potentially repetitive machine parts. In addition, we provide an annotated dataset that can be used to quantify the success rate of a variety of 2D and 3D systems for object detection and disambiguation. Finally, we evaluate several performance metrics for object disambiguation relative to the baseline success rate of a human.

Method

We seek a monocular system that operates markerless and exploits state-of-the-art object detectors in order to disambiguates objects as parts of a machine. Figure 1 shows an overview of our system.

We use the sparse 3D information generated by the SLAM system[1] in order to reproject the 2D object detections[2] to 3D. As all preceding frames are connected by SLAM track, we accumulate the reprojected 2D object detections over time. In addition to 3D detection clouds, we also require a 3D machine layout that specifies the relative locations of each object. Such description are often provided by the machine specifications, but it doesn't have to be metric or a complete model in our method, which provides easy deployment and adaptation to new scenarios.

In order to match the 3D layout with N objects g_n to the observed detections d, we define an energy function that is taking into account the object appearance ($E_{appearance}$), deformation of the layout ($E_{deformation}$), scale (E_{scale}), viewpoint ($E_{viewpoint}$) as well as amount of matched objects (optional part in the deformation energy). The energy on scale and viewpoint capture an expectation of typical viewpoints the machine is viewed in. We seek the best match by finding an assignment of detections $d_1,\ldots,d_N$ as well as a projection matrix M so that the following objective:

$$\underset{d_1,d_2,\ldots,d_N,M}{\arg\min} \; E_{deformation} + E_{appearance} + E_{scale} + E_{viewpoint} \tag{1}$$

where

$$E_{deformation} = \frac{\sum_{n=1}^{N}\delta_n}{N}\sum_{n=1}^{N}\delta_n \cdot log(\left\|\bar{M}(P_{g_n}) - P_{d_n}\right\|)$$
$$E_{appearance} = -\sum_{n=1}^{N}\delta_n \cdot A_{d_n} \tag{2}$$

P_{g_n} and P_{d_n} are the 3D coordinate of g_n and d_n, while A_{d_n} is the detection score of the match d_n. δ_n is for handling the non-matched machine parts, where $\delta_n = 1$ if $\left\|\bar{M}(P_{g_n}) - P_{d_n}\right\|$ smaller than a threshold and $\delta_n = 0$ otherwise. In both E_{scale} and $E_{viewpoint}$, the scale factor s and three view points included in the 3D transformation $M(\cdot)$ are hard-constrained according to their distributions learnt from the training videos.

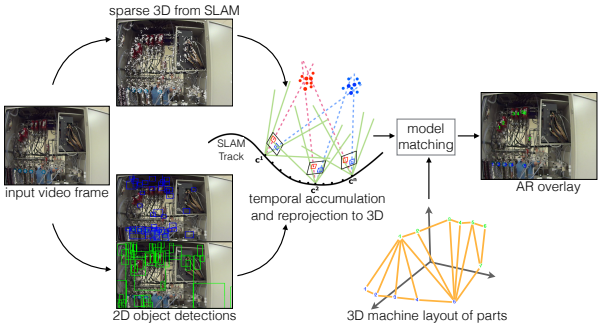


Figure 1: Overview of our system for object disambiguation.

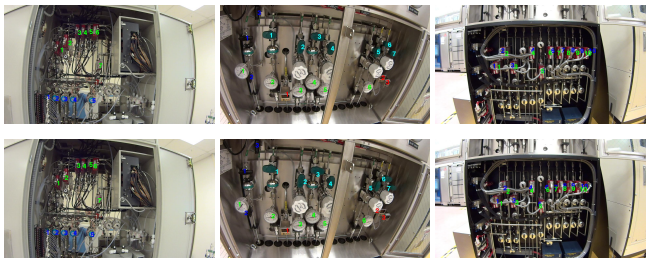


Figure 2: Example results. First row are for the groundtruh of each machine. Second row are the corresponding results from our method.

In order to minimize the objective, we follow a RANSAC pipeline by randomly selecting candidate alignments between the detections and the machine layout which results in an initial geometric transformation.

Experiments

In order to evaluate our approach, we propose the first benchmark for an object disambiguation task in maintenance work that is composed of an annotated dataset. Furthermore, instead of using traditional Pascal metric, we are interested in a metric that captures the object disambiguation performance of a human if provided with the produced overlay. Therefore we propose a set of candidate metrics and then evaluate which one is closest to actual human judgement on the task. Our proposed metric gives a more realistic estimate of the system performance than a traditional Pascal object detection metric that consistently underestimates the system performance. Figure 2 shows example results of our system in comparison to the groundtruth annotations.

References

[1] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007.

[2] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrew D Bagdanov, Maria Vanrell, and Antonio M Lopez. Color attributes for object detection. In *CVPR*, 2012.

Knowing Where I Am: Exploiting Multi-Task Learning for Multi-View Indoor Image-based Localization

Guoyu Lu[1]
luguoyu@udel.edu
Yan Yan[2]
yan@disi.unitn.it
Nicu Sebe[2]
sebe@disi.unitn.it
Chandra Kambhamettu[1]
chandrak@udel.edu

[1] Video/Image Modeling and Synthesis Lab
University of Delaware

[2] Department of Information Engineering and Computer Science
University of Trento

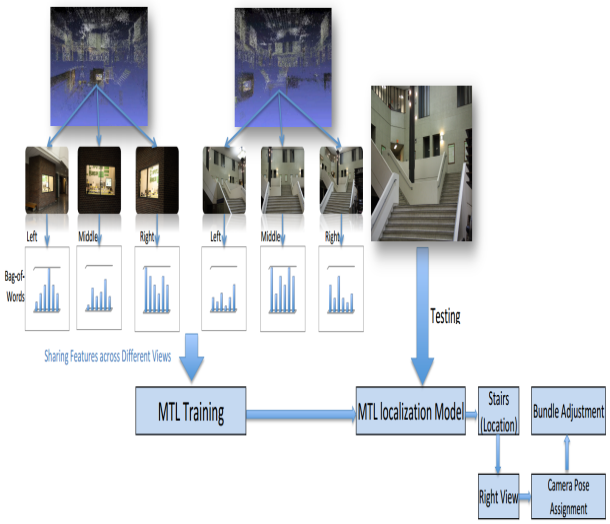


Figure 1: Multi-view image-based localization system

Indoor localization systems are applied to navigate people in large and complex indoor environment, such as shopping malls and museums where auxiliary information is necessary to help visitors localize themselves. In some urgent situation, like boarding an airplane and finding the emergency room in a hospital, providing accurate and timely location information is essential for travelers to catch planes and wounded people to get prompt medical assistance. The majority of the current indoor localization methods is based on WiFi and pre-deployed beacons. These methods usually require additional equipment to perform the localization task and the accuracy depends on the distribution of beacons and cellular stations in a large extent. Meanwhile, the WiFi and beacon based methods are lack of the orientation information, which is essential for navigation. GPS is quite successful in outdoor navigation. However, in indoor buildings with roofs and walls, weak GPS signals result in unreliable navigation information. Even in an outdoor large building area, GPS signals from satellite are attenuated by walls.

Image based localization has been mainly applied in outdoor environments in the past to overcome the weak GPS signal problem among large buildings. This method has been introduced to indoor environments in recent time. The main idea is to linearly search the image database consisting of indoor building images and find the best matched image. With the development of Structure-from-Motion (SfM) reconstruction techniques, 3D models are used for localization. Users can easily capture a 2D image with their mobile phone and register the 2D image with the 3D model to get the location information. In this process, features extracted from the 2D images are utilized to match against the features in the SfM 3D model; camera pose can be calculated based on the matching descriptors, providing users the location and orientation information. As the SfM technique does not require the cameras to be calibrated, the related images are easier to obtain, which makes the large scale reconstruction and 3D model based localization possible. Obtaining the location information is only half of the job. A map with the location information can help better perform the navigation task. With this purpose, a 3D model is suitable for localization purposes that facilitate users to understand the 3D building structure and schedule a visiting plan. However, a SfM model for localization usually contains millions of descriptors. Searching the correspondences within this scope is extremely time-consuming. Although k-d trees and visual word methods are applied to accelerate the corresponding search process, the reduced search scope may potentially add incorrect correspondences between 2D features and 3D points.

In this paper, we propose multi-view image based localization, which is a framework based on multi-task learning (MTL). MTL attempts to improve the performance of several specific tasks based on the shared common properties. Current research shows that it is beneficial to learn the tasks simultaneously instead of learning a single task separately when the tasks exhibit commonalities. During the learning process, the shared information across different tasks is extracted to simultaneously learn the multi-related tasks. With the purpose of guiding users with the location and orientation information, we divide the physical view direction into several regions. It is expected that images of the same object captured from different view directions contain similarities with regards to appearance, as well as differences due to the viewing perspectives.

Multi-view image based localization aims to learn the relationship of interior architecture appearance across different viewing directions. Ideally, the tasks within the same group should share the similar features while features extracted from tasks in different groups are expected to be different. Following this idea, images captured from the same direction are classified into one task, including same and different location images. The images captured from the same location across different camera angles are treated as the same group. We learn a multi-view regression model based on the correlated tasks scattering in different groups. During the testing phase, the query image retrieves the most relevant group for achieving the location information. Meanwhile, our MTL regression model assigns a direction to the query image based on multiple tasks for the orientation purpose. As we perform SfM reconstruction prior to the multiple view localization phases, every image used for SfM reconstruction is associated with a camera pose. The camera pose of the most corrected image within the same task, and the same group is assigned to the query image. We further apply bundle adjustment to the query image to refine the assigned camera pose. In this way, we can take benefits from localization methods both based on 2D image and 3D model. The whole multi-view image-based localization framework is illustrated in Figure 1.

To summarize, the contributions of this paper are the following: (i) To our knowledge, this work is the first to address the problem of indoor image-based localization from multi-view settings. (ii) We are the first to propose the multi-task learning approach for multi-view indoor image-based localization. (iii) Both the orientation of the image and the location information can be obtained by exploiting multi-task learning.

Making use of the multi-task learning method, we develop a multi-view image based localization system. By separating the view directions into 3 different partitions as tasks, we simultaneously learn the relationship among the tasks, which can improve the prediction accuracy of each view orientation. The learned multi-view regression model can accurately retrieve the location information. After learning the model, our multi-view system can retrieve the location and view orientation information by computing a dot product to assign a correlation score, avoiding large scale correspondences search. Leveraging the 3D localization system, we assign the camera pose of the nearest neighbor image of the same orientation and location used for SfM reconstruction to the query image, with further refinement using bundle adjustment. Embedding our multi-view method into the 3D localization system helps us better achieve the localization information in a 3D map.

Board #93

Duration Dependent Codebooks for Change Detection

Brandon A. Mayer
Brandon_Mayer@brown.edu
Joseph L. Mundy
mundy@lems.brown.edu

Brown University
School of Engineering
Rhode Island, USA

Change detection is a computer vision application that attempts to distinguish normal and abnormal scene activity in video sequences. However, natural scenes are composed of complex, dynamic events that make it difficult for a change detection system to distinguish between changes of interest and background. To further compound the problem, it is impossible to define what a system should consider as a relevant change without considering the context of the application. For example, are cars moving along a highway foreground or background? If the goal of the application is to count the number of cars entering and exiting a restricted area, it is necessary for the system to account for every car in the scene. However, if the requirement is to monitor a busy highway for irregular traffic activity such as a collision, then the system will need to consider common traffic patterns as normal and not declare routine traffic activity as change.

This paper describes a supervised system for pixel-level change detection for fixed, monocular surveillance cameras. Per-pixel intensity sequences are modeled by a class of Hidden Semi-Markov Models, Duration Dependent Hidden Markov Models (DDHMMs), to accurately account for stochastically periodic phenomena prevalent in real-world video. The per-pixel DDHMMs are used to assign discrete state labels to pixel intensity sequences that summarize the appearance and temporal statistics of the observations. State assignments are then used as a features for constructing per-pixel code books during a training phase to identify changes of interest in new video.

The per-pixel intensity model is validated by showing superior predictive performance to pixel representations commonly used in change detection applications. A new data set is presented which contain dynamic, periodic backgrounds with larger time scale variability than previous data sets and the proposed method is compared to state-of-the-art change detection methods using the new videos.

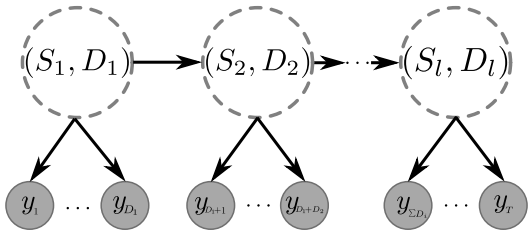


Figure 1: Graphical visualization of the DDHMM

A Duration Dependent Hidden Markov Model (DDHMM) models a sequence of observations, $Y = (y_1, y_2, \ldots, y_T)$, using a sequence of latent state pairs: $((S_1,D_1), (S_2,D_2), \ldots, (S_l,D_l))$ where S_i is a state label and D_i is a random variable that represents the time spent in state S_i. Note that capital letters denote random variables and lower case letters represent specific variable assignments. A graphical visualization of a DDHMM is shown in Figure 1 where dotted circles represent random variables and the shaded nodes represent observed quantities. The topology of the graphical model is variable since the number of state-duration tuples will change depending on the particular configuration of the duration random variables.

The observation and state sequences are related through three fundamental distributions: the duration $p(D_i = d_i \mid S_i = s_i)$, state transition $p(S_i = s_i \mid S_{i-1} = s_{i-1})$ and emission $p(y_t \mid S_i = s_i)$ distributions. The likelihood of an observation sequence given a particular latent state assignment is given by equation 1 where $r_i = \sum_{m=1}^{i} d_m$ and $p(s_1)$ is an initial distribution of state labels. The observation sequence is assumed to be left-censored, i.e., the last tuple (s_l, d_l) is distributed according to the state survival distribution $p(D_l \geq d_l \mid s_l)$, to mitigate the effect of the length of the observation sequence on the probability of a particular state sequence [2].

$$
\begin{aligned}
&p(y_1,\ldots,y_T|(s_1,d_1),\ldots,(s_l,d_l)) = \cdots \\
&p(s_1)p(d_1 \mid s_1)\prod_{m=1}^{d_1} p(y_m \mid s_1)\prod_{i=2}^{l-1} p(d_i \mid s_i)p(s_i \mid s_{i-1})\cdots \\
&\prod_{j=1}^{d_i} p(y_{r_i+j} \mid s_i)p(D_l \geq d_l)p(s_l \mid s_{(l-1)})\prod_{k=1}^{d_l} p(y_{r_{(l-1)}+k} \mid s_l) \qquad (1)
\end{aligned}
$$

A simple single-pass, greedy algorithm is used for learning the parameters and complexity of the per-pixel DDHMMs as well as computing the locally most likely state assignment under an AIC [1] based objective function. An unoptimized multithreaded C++ implementation, running on a 3.46 GHz Intel i7 processor, achieves real-time performance. Specifically, continuously updating a DDHMM at each pixel for a video sequence containing seventeen hundred frames with resolution 240×320 pixels takes an average of 31 milliseconds per frame.

The Swing video sequence shown in Figure 2 shows a mother pushing her daughter on a swing set and eventually, a previously unobserved pedestrian enters and exits the scene. This seemingly innocuous footage contains interesting periodic phenomena that modern change detection algorithms cannot model. The mother's motions are repetitive as she pushes the child with a periodic rhythm. The mother and daughter on the swing set are considered normal, they are using the swing set for the entirety of the video sequence, and the pedestrian is a change of interest. By modeling intensity persistence, the proposed method is able to explicitly model the dynamics of the swinging child and avoid false positive detections. Competing algorithms exhibit significantly higher false positive rates for this sequence.

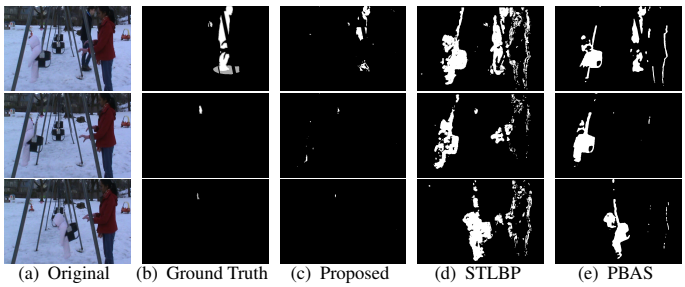


Figure 2: *Swing sequence*: Change detections are visualized as white pixels, normal scene activity as black. The proposed method is the only algorithm which can learn the swinging child is a normal part of the scene but still detect the previously unobserved pedestrian.

The paper discusses the implementation of the online DDHMM learning and inference algorithm as well as the construction of the DDHMM based code book and its application as a classifier for detecting changes in novel video segments. The proposed method is compared to current state of the art change detection algorithms and is shown to be superior, especially in environments containing complex periodic phenomena.

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

[2] Yann Guédon. Estimating hidden semi-markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3): pp. 604–639, 2003. ISSN 10618600. URL `http://www.jstor.org/stable/1391041`.

Real-time Dense Disparity Estimation based on Multi-Path Viterbi for Intelligent Vehicle Applications

Qian Long[1]
long@toyota-ti.ac.jp
Qiwei Xie[1]
qw_xie@toyota-ti.ac.jp
Seiichi Mita[1]
smita@toyota-ti.ac.jp
Hossein Tehrani[2]
hossein_tehrani@denso.co.jp
Kazuhisa Ishimaru[3]
kazuhisa_ishimaru@soken1.denso.co.jp
Chunzhao Guo[4]
czguo@mosk.tytlabs.co.jp

[1] Research Center for Smart Vehicles
Toyota Technological Institute
2-12-1 Hisakata, Tempaku, Nagoya, Aichi 468-8511, Japan
[2] DENSO CORPORATION
1-1, Showa cho, Kariya, Aichi, 448-8661 Japan
[3] NIPPON SOKEN Inc.
Nishio, Aichi, Japan
[4] Toyota Central R&D Labs., Inc.
Nagakute, Aichi, Japan

3D scene understanding plays an essential role for intelligent vehicle applications. In these applications, passive stereo vision systems offer some significant advantages to estimate depth information compared with active systems such as 3D LIDAR. To apply stereo vision in autonomous driving, a new real-time stereo matching algorithm paired with an online auto-rectification framework is proposed. This method uses a bi-directional Viterbi algorithm at 4 paths to decode the matching cost space and a hierarchical structure (as shown in Fig. 1) is proposed to merge the 4 paths to further decrease the decoding error. We introduce Total Variation [1] constraint into Viterbi path for approximately modeling 3D planes at different orientations to reach a similar effect as slanted-plane models. Structural similarity (SSIM)[3] is used to to measure the pixel difference between left and right images at epipolar lines to improve robustness to luminance variation. The equation for one Viterbi path is expressed by:

$$e(p,u) = \min_{u' \in L_u} \{e(p-1,u') + \lambda e^{-|G|}|u-u'| + SSIM(p,u)\} \quad (1)$$

where $e(p,u)$ is the energy of Viterbi node at pixel p and disparity u, G is the gradient information of image, λ is the parameter, L_u denotes connected Viterbi nodes to the Viterbi node at pixel p and disparity u.

Based on the output of Viterbi process, a convex optimization equation is derived to estimate epipolar line distortion. we summarize the properties of the epipolar line distortion caused by normal factors in intelligent vehicle applications. Based on these properties and inspired by the famous optical flow problem, we convert this distortion estimation problem to an optimization problem and employ the convex optimization theory to solve it. The Viterbi process and convex optimization are integrated into an online framework (as shown in Fig. 2) and two parts benefit each other without losing speed in this framework. It can automatically keep the epipolar line constraint to avoid the degradation of stereo matching results, which usually happens when other stereo matching methods being applied for driving vehicles.

Extensive experiments were conducted to compare proposed algorithm with other practical state-of-the-art methods for intelligent vehicle applications. According to evaluation results at the KITTI [2] training dataset which includes total 194 images, our method has 7.38% average error rate compared to SGBM's 12.88% and ELAS's 11.99%. We also test the proposed algorithm in our experimental autonomous vehicle at real driving environments. For any 640x480 images with maximum 40 disparities, the running time is about 196ms with GTX TITAN GPU and Xeon E5-2620 CPU. Real driving videos including featured cases and typical failure cases can be found in the supplementary material.

[1] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20 (1-2):89–97, 2004.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[3] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

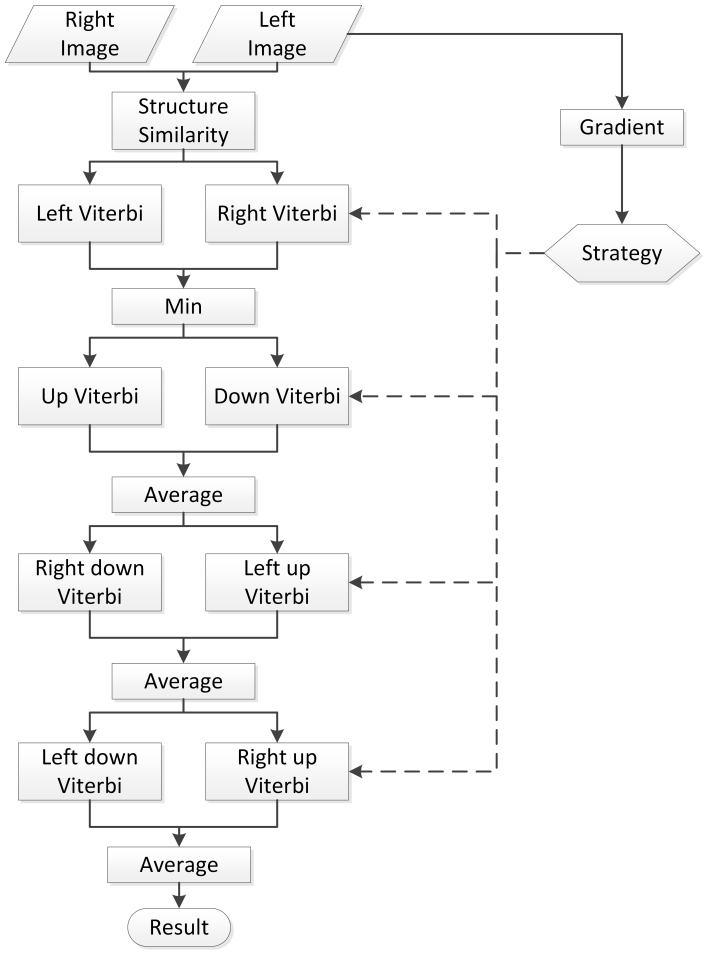


Figure 1: Hierarchical structure for the merging of multiple Viterbi paths.

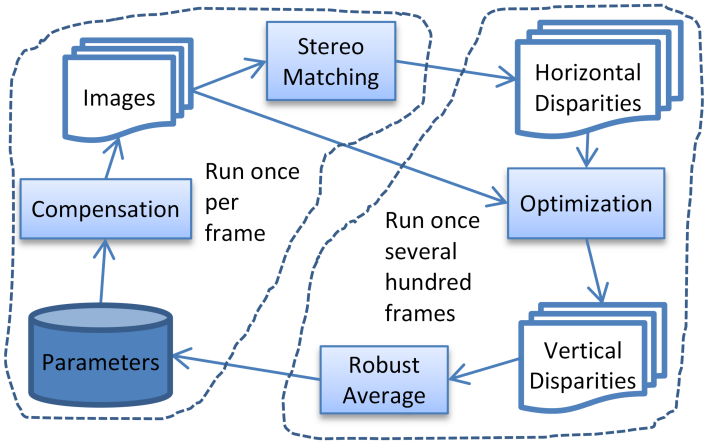


Figure 2: Framework of automatic online rectification.

Board #95

Uncalibrated Near-Light Photometric Stereo

Thoma Papadhimitri
http://www.cvg.unibe.ch
Paolo Favaro
http://www.cvg.unibe.ch

Universität Bern
Institut für Informatik und angewandte Mathematik
Bern, Switzerland

Photometric stereo (PS) [3] is a technique to accurately recover the normal map of a 3D scene from several pictures (at least three) taken from the same view point and under different illumination conditions. When the light directions and intensities are known, photometric stereo can be solved as a linear system. When the illumination is not known, one needs to solve a much harder problem: uncalibrated photometric stereo. Typical assumptions are the Lambertian reflectance, orthographic projection, absence of shadows and interreflections and that the light sources are far away from the object. In particular, the last assumption allows to consider parallel illumination and, consequently, a simpler image formation model.

The distant light assumption is a reasonable approximation as long as the dimensions of the scene are much smaller than the distance of the light sources. However, this may not be the case in many practical scenarios such as endoscopy, cultural heritage, reconstruction of big indoor objects, underground and underwater navigation, or full human body 3D reconstruction. Motivated by this fact, we introduce for the fist time an uncalibrated near-light photometric stereo method where no prior information about light position and intensities is needed. Only in [1] uncalibrated near-lights were considered. However, the method only recovers depth cues obtained from particular illumination configurations (lights moving on a line or plane), while in our algorithm we consider illuminants distributed arbitrarily in front of the object. We achieve this by first analyzing the reconstruction ambiguities and then by introducing an iterative technique to solve for the normals, reflectance and lights. We demonstrate the practical use and accuracy of our algorithm with real world experiments and compare it with the state-of-art in uncalibrated distant light photometric stereo.

The image formation model typically used for the near-light case under the Lambertian reflectance is

$$I_{pk} = \frac{\rho_p \mathbf{N}_p^T (\mathbf{L}_k - \mathbf{X}_p)}{\|\mathbf{L}_k - \mathbf{X}_p\|^q} e_k, \tag{1}$$

where $q = 3$, $\mathbf{N}_p$ is the normalized normal, $\mathbf{L}_k$ is the 3D position of the k-th light, e_k the corresponding intensity, $\mathbf{X}_p$ is the 3D position of a generic point of the surface and finally ρ_p is the albedo, where p denotes the pixel or spatial index. Notice that the intensity fall-off is inversely proportional to the square distance of the light source from the object. In [2] the attenuation term is considered to be inversely proportional to the distance instead of the square distance of the light from the surface point and in this case we have $q = 2$. In this work we investigate both cases ($q = 2$ and $q = 3$).

We solve the uncalibrated near-light photometric stereo via an alternating minimization procedure which consists of two steps: first we estimate the normals, the albedo and the depth and then we estimate the lights and their intensities given the normals, the depth and the albedo.

In Fig. 1 we show the experimental results in the case of the **Dwarf** and **Sphere** datasets. We captured images by randomly distributing 12 led lights in the upper hemisphere of the scene, which were positioned within a distance range of 40-60 cm. The light calibration was done manually (in order to have a ground truth reference) and the error is less than 0.5 cm. We have included additional profile photos in order to create a better perception of the 3D structure of the scene. It can be noticed that a choice of $q = 2$ in the image formation model yields to lower reconstruction errors compared to that for $q = 3$. The light estimation is more accurate as well: for the **Dwarf** dataset we obtain a mean error in the light coordinates estimation of 4.79 cm for $q = 2$ and 5.35 cm for $q = 3$, while for the **Sphere** dataset such error is 3.85 cm for $q = 2$ and 5.25 cm for $q = 3$. We also performed the reconstruction of a ground truth scene (planar scene made of paper) and obtained a mean angular error in the surface normals estimation of 4.05 angular degrees for $q = 2$ and 9.47 angular degrees for $q = 3$, while the distant light photometric stereo method gives a mean angular error of 24.85 angular degrees. These results seem to be in contradiction with the well established image formation model for near-light illumination which requires $q = 3$. This might be due to the light sources we chose for the illumination setup. However, for both cases the reconstruction results obtained with our method are very good. Indeed, notice the significant improvement of the reconstruction compared to the distant light photometric stereo. Conventional photometric stereo fails because the lights are close to the scene and the distant light assumption does not hold anymore and a strong distortion of the normal map can be noticed, especially towards the borders of the image.

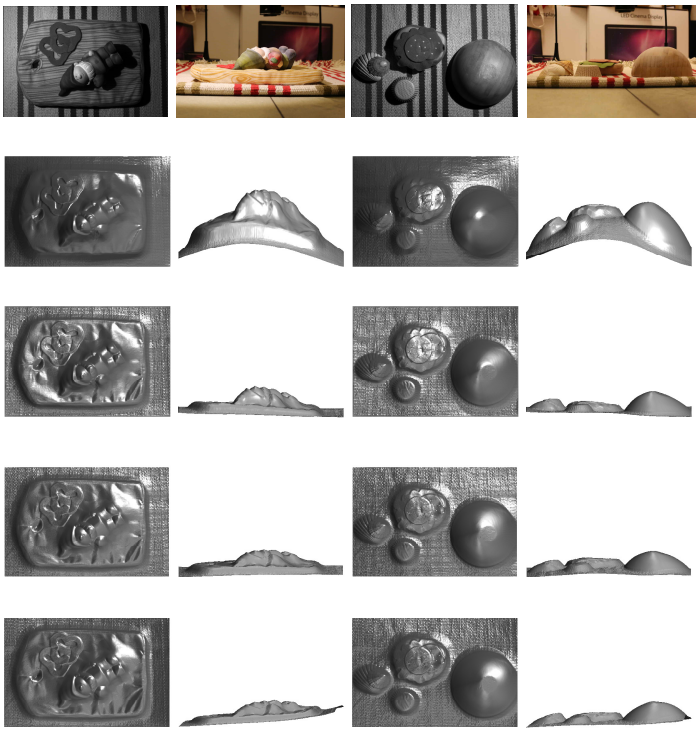


Figure 1: Reconstruction results for the **Dwarf** and **Sphere** scene obtained via our experimental setup. Rows from top to bottom: frontal (first and third column from left) and lateral (second and fourth column from left) view of the scene, reconstructed surfaces via calibrated distant light PS (second row from top), reconstructed surfaces via our calibrated near-light PS (third row from top), reconstructed surface via our uncalibrated near-light PS method with $q = 2$ (fourth row from top) and reconstructed surface via our uncalibrated near-light PS method with $q = 3$ (fifth row from top).

However, the surface is smoothed out at the borders of the object. This is because of the shadows which introduce non-negligible distortion to the imaging model, especially when the lights are closer to the scene, as in our experimental setup. Moreover, the effect of interreflections at these regions with strong concave edges is significant. Finally, the running time of our algorithm for the above datasets with resolution 0.2-0.3 megapixels varies between 3 and 4 minutes.

[1] Sanjeev Jagannatha Koppal and Srinivasa G. Narasimhan. Novel depth cues from uncalibrated near-field lighting. *ICCV, 2007*.

[2] Fumihiko Sakaue and Jun Sato. A new approach of photometric stereo from linear image representation under close lighting. In *ICCV Workshops*, pages 759–766, 2011.

[3] R.J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

Video-Based Face Recognition Using the Intra/Extra-Personal Difference Dictionary

Ming Du
mingdu@umd.edu
Rama Chellappa
rama@umiacs.umd.edu

Department of Electrical and Computer Engineering and Center for Automation Research, UMIACS
University of Maryland
College Park, USA

In recent years, with videos playing an increasingly important role in our everyday lives, video-based face recognition (VFR) has begun to attract considerable research interest. In this paper, we attempt to improve the performance of VFR based on the concept of intra-personal/extra-personal face variations. The concept was first proposed by Moghaddam et al. in [2] and has achieved great success in still-image based face recognition. Specifically, the intrapersonal subspace Ω_{In} is defined as the subspace constructed from within-class sample differences $\{\Delta_{In}\}$. It accounts for appearance variations of the same subject that arise from factors like pose, lighting, expression etc. Similarly, the extra-personal subspace Ω_{Ex}, which characterizes appearance variations caused by intrinsic identity differences, is constructed using the between-class sample differences $\{\Delta_{Ex}\}$. To apply this concept to the VFR problem, our solution is based on two aspects: To handle pose variations, we learn a Structural-SVM-based detector that simultaneously localizes the face fiducial points and estimates face pose. To model other face variations, we exploit the strengths of sparse codings by constructing intra-personal/extra-personal dictionaries. An overview of the proposed approach is shown in Figure 1.

For face normalization, we learn a mixture of fiducial point detectors which is used for geometric alignment. Each component of the mixture corresponds to a specific face pose. We localize the fiducial points L and estimate the face pose m jointly by maximizing the potential function: $\mathbf{z}^* = \{L^*, m^*\} = \underset{L,m}{\operatorname{argmax}}\, \mathbf{w}_m^T \phi_m(I, L)$. To learn the parameter $\mathbf{w}$, we solve the following margin re-scaling structure SVM problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \max_{\mathbf{z}\in\mathcal{Z}} [\Delta(\mathbf{z}, \mathbf{z}_n) + \mathbf{w}^T \mathbf{\Phi}(I_n, \mathbf{z})] - \mathbf{w}^T \mathbf{\Phi}(I_n, \mathbf{z}_n) \quad (1)$$

. In (1), $(I_n, \mathbf{z}_n)$ is an image-label pair in the training database and $\mathcal{Z}$ is the viable label configuration set. ξ_n is the slack variable . $\Delta(\mathbf{z}, \mathbf{z}_n)$ is the loss function of the output $\mathbf{z}$ when measured against the ground-truth label $\mathbf{z}_n$. Suppose there are S fiducial points in total and the subset of indexes of those fiducial points visible for the m-th pictorial model is $S(m)$. Compared with Zhu and Ramanan's recent Deformable Parts Model (DPM)-based face and feature detector [3], our objective function explicitly impose constraints on the margin between correct and wrong landmark predictions. Moreover, in our case the margin is re-scaled by a loss function $\Delta(\mathbf{z}, \mathbf{z}_n)$ which penalizes the negative training samples according to their misalignment errors. As a result, although our method is not designed to produce face detection output in addition to feature point locations, it has higher accuracy in localizing fiducial points.

Based on the estimated pose, the localized faces in a video are then aligned to pose-specific common reference coordinate frames. They are further clustered using a non-parametric Bayesian model to remove temporal redundancy. The resulting model has infinite number of Gaussian mixtures controlled by a Dirichlet process $DP(\beta, H)$ [1], where β is the concentration parameter and H is the base probability measure. The mixture weights are generated from the Griffiths-Engen-McClosky (GEM) process. By using the Dirichlet process mixture model, new clusters can be generated when more frames are observed, and there is no need to know the number of clusters a priori.

In recent years, sparse coding has gained popularity in the field of image classification. In general, a dictionary $\mathbf{D} = [D_1, D_2, ..., D_K]$, where $D_k \in R^d$, can be learned unsupervisedly from training samples $\mathbf{X} = \{\mathbf{x}_i, i = 1, 2, ..., N\} \in R^{d\times N}$ (In our case, the training samples are intra/extra-personal difference of feature vectors which are extracted from faces of the same pose.) by solving the following constrained optimization problem:

$$\min_{\mathbf{D},\alpha} \sum_{i=1}^{N} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \quad (2)$$

However, to serve the purpose of classification better, we follow the Label-Consistent K-SVD (LC-KSVD) algorithm to jointly learn a generative

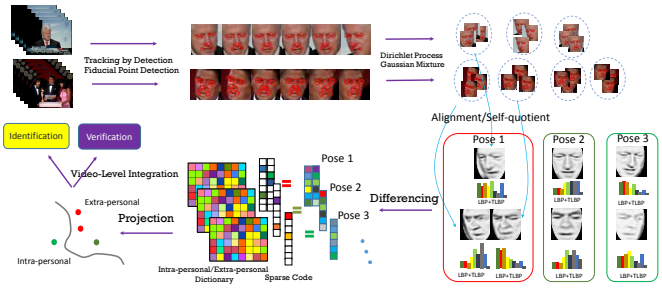


Figure 1: Processing pipeline of the proposed video-based face recognition algorithm.

shared dictionary and a discriminative projection matrix. Although the shared dictionary is composed of two sub-dictionaries corresponding to intrapersonal and extra-personal differences respectively, the sparse code of any input difference vector is computed by using the complete set of atoms in the dictionary. As a result, the final optimization problem has the following form:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_2^2 + \mu\|\mathbf{Q} - \mathbf{B}\mathbf{A}\|_2^2 + \sigma\|\mathbf{F} - \mathbf{W}\mathbf{A}\|_2^2 + \lambda\sum_i \|\alpha_i\|_1 \quad (3)$$

. In (3), the columns of $\mathbf{F} \in R^{2\times N}$ are labels of the training instances in $\mathbf{X}$, represented using the 1-of-K coding scheme. The matrix $\mathbf{W} \in R^{2\times d}$ encodes the discriminative information of the sparse codes $\mathbf{A}$ and is learned along with the shared dictionary. The linear transformation $\mathbf{B} \in R^{K\times d}$ encourages the samples from the same class to be reconstructed using similar atoms. This constraint can be written in the form: $\mathbf{B}\mathbf{X} = \mathbf{Q}$, where $\mathbf{Q} \in R^{K\times N}$ has a block diagonal form. At test time, for each probe video, we extract feature vectors from the centers of clusters formed using the non-parametric Bayesian method introduced above, and take differences between them and the feature vectors similarly extracted from clusters in the gallery videos. Recognition results are then obtained using the learned intra/extra-personal dictionaries and the discriminant matrix $\mathbf{W}$.

One advantage of the proposed algorithm is its scalability. Traditionally, it requires a large amount of training data to effectively characterize a subject. More often than not, we have insufficient training samples to account for all possible variations for each subject. As a result, decision boundaries of the classifiers are often highly dependent on the training data and are prone to change every time we add new subjects to the database. In contrast, because the intra/extra-personal face variations are generic, our algorithm is flexible enough to learn a dictionary using either the training set from the same database or that of an entirely different set of subjects (i.e. cross-database dictionary). We demonstrate through experiments that the proposed approach achieved state-of-arts performance in both modes. Moreover, the proposed framework naturally supports the face verification protocol in addition to the recognition one.

[1] K. Kurihara, M. Welling, and Teh Y. W. Collapsed variational Dirichlet process mixture models. In *International Joint Conference on Artificial Intelligence*, pages 2796–2801, January 2007.

[2] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(6):780–788, 2002.

[3] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, June 2012.

Board #97

An Efficient Online Hierarchical Supervoxel Segmentation Algorithm for Time-critical Applications

Yiliang Xu[1]
yiliang.xu@kitware.com
Dezhen Song[2]
dzsong@cse.tamu.edu
Anthony Hoogs[1]
anthony.hoogs@kitware.com

[1] Kitware Inc.
28 Corporate Drive
Clifton Park, New York, USA

[2] Department of Computer Science and Engineering
Texas A&M University
College Station, Texas, USA

Video segmentation has been an active research topic for the last decade. It is often used as a pre-processing procedure for subsequent vision algorithms. Despite its significant practical relevance, research on video segmentation does not catch up with its counterpart of image segmentation, due to multiple challenges including higher dimensional (3D) segmentation, temporal consistency, scalability and efficiency, and many more. Most existing algorithms require pre-loading all or part of the video and batch processing the frames, which introduces temporal latency and significantly increases memory and computational cost. Other algorithms rely on human specification for segmentation granularity control.

In this paper, we propose an efficient online hierarchical supervoxel segmentation algorithm for time-critical applications. Here by online, we mean the algorithm computes the supervoxel segmentation of the video stream up to the latest frame once it arrives. Therefore the algorithm requires no streaming buffer but the incoming frame and thus runs in the truly online manner. It also automatically segments the video with hierarchical granularity. The main contributions of the work include

1. an efficient, yet effective probabilistic segment label propagation across consecutive frames,
2. a new method for label initialization for the incoming frame, and
3. a temporally consistent hierarchical label merging scheme.

Figure 1 illustrates the processing flow of our algorithm. The algorithm starts with the over-segmentation and the corresponding hierarchical segmentations of the first frame using the hierarchical graph-based segmentation [3]. Then it propagates the over-segmentation labels onto the second frame based on both motion (dense optical flow) and appearance cues to form the "seed" segments and the corresponding new graph for the second frame. The seed segments grow in the second frame and new segments (if any) are naturally generated using the graph-based merging to complete the over-segmentation for the second frame. Finally, higher-level segmentations of the second frame are generated with a self-supervision merging scheme based on the segmentation at the same level in the previous frame. These steps are repeated when the new frame is coming to form the up-to-date video stream segmentation.

We test our algorithm on a public benchmark dataset [5], and use a wide range of performance metrics to thoroughly compare it with multiple state-of-the-art algorithms, namely, Segmentation by Weighted Aggregation (SWA) [1, 4], Graph-Based Hierarchical segmentation (GBH) [3], and Streaming Graph-Based Hierarchical segmentation (StreamGBH) [6]. In particular, SWA and GBH are offline algorithm which load the video at once. According to both [2] and [5], GBH is one of the top-performing algorithms. StreamGBH loads a buffer of K frames at a time. Here we test and compare two of its variations with $K = 10$ and $K = 1$, respectively.

Figure 2(a) shows the 3D boundary PR of all algorithms[1]. SWA appears to have the best PR tradeoff. Our algorithm is comparable to GBH and StreamGBH with $K = 10$, and outperforms StreamGBH with $K = 1$. Figure 2(b) shows the 3D volume precision-recall of all algorithms. Our algorithm is comparable to GBH and SWA and outperforms the two StreamGBH variations.

Table 1 shows that our algorithm is significantly faster than all the other algorithms including the StreamGBH with $K = 1$. This is because our graph-based segmentation is carried out only on individual 2D frames, while that in StreamGBH is carried out on a $(K+1)$-frame 3D volume. Even with $K = 1$, the number of edges that need to be cut (for each frame) in StreamGBH is at least a couple of times of that in our algorithm. Our algorithm is also memory-efficient. Offline algorithms or streaming algorithms requires the memory size proportional to the size of the 3D volume buffer, while our algorithm only requires memory size proportional to the 2D frame size.

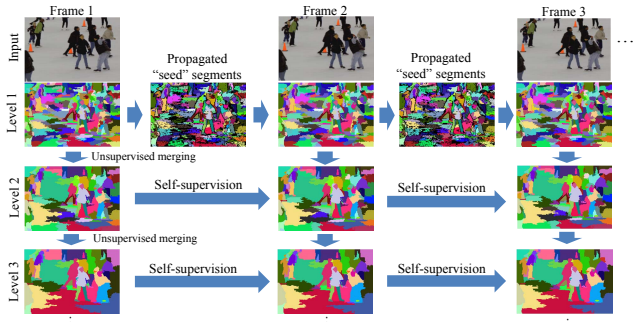
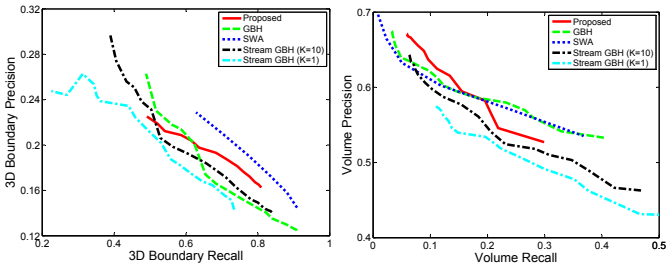


Figure 1: An illustration of the processing flow of the proposed algorithm. Each color corresponds to a supervoxel.



(a) 3D boundary (b) 3D volume

Figure 2: Comparison of Precision-Recall.

		Proposed	StreamGBH		GBH	SWA[2]
			$K = 1$	$K = 10$		
Time (sec.)	per frame[3]	0.72	4.27	8.23	12.96	5.88
	per segmentation	0.03	0.20	0.39	0.62	0.98
Memory (GB)		0.1	0.1	0.5	3.7	8.2

Table 1: Comparison on computation time and memory requirement.

[1] Jason J Corso, Eitan Sharon, Shishir Dube, Suzie El-Saden, Usha Sinha, and Alan Yuille. Efficient multilevel brain tumor segmentation with integrated bayesian model classification. *IEEE Transactions on Medical Imaging (T-MI)*, 27(5):629–640, 2008.

[2] Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jiménez Cárdenas, Thomas Brox, and Bernt Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*, 2013.

[3] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, pages 2141–2148. IEEE, 2010.

[4] Eitan Sharon, Meirav Galun, Dahlia Sharon, Ronen Basri, and Achi Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813, 2006.

[5] Chenliang Xu and Jason J Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

[6] Chenliang Xu, Caiming Xiong, and Jason J Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012.

[1] These are not typical PR curves. They are not generated by sweeping a threshold for recognition decision. Instead they are precision and recall pairs by different granularity configurations. The convex shape of the curve does not means the algorithm is worse than random guess.

[2] SWA is set to produce only 6 layers of segmentation to save time.

[3] For each input frame, we produce a number of layers of segmentations.

Board #98

Robust 3D Face Shape Reconstruction from Single Images via Two-Fold Coupled Structure Learning

Pengfei Dou
bensondou@gmail.com
Yuhang Wu
yuhang@cbl.uh.edu
Shishr K. Shah
sshah@central.uh.edu
Ioannis A. Kakadiaris
ioannisk@uh.edu

Computational Biomedicine Lab
Department of Computer Science
University of Houston
Houston, TX, USA

The problem of estimating the 3D shape of human faces from single images is of great interest and has attracted considerable research effort. Many approaches recently proposed to solve this problem could be considered extensions of Shape-from-Shading (SFS) methods, where a 3D shape is optimized to generate 2D renderings that match the input images [1, 5, 7]. Other methods in the literature propose to infer 3D face shape by fitting a set of feature points between the 2D image and the 3D model [3, 4, 6].

In this paper, we propose the Two-Fold Coupled Structure Learning (2FCSL) algorithm, which is capable of reconstructing 3D face models based on a sparse set of 2D landmarks that could be localized automatically by most of the recently proposed landmark detectors. By explicitly incorporating 3D-2D pose estimation and formulating the problem into a two-fold coupled structure learning problem, our method achieves better robustness to arbitrary pose variations and landmark localization noise.

Using a shape vector representation Y^i_{3D} of the dense 3D face, N 3D training faces are stacked together to construct the 3D dense landmark (3DDL) model $\gamma^d_{3D} = (Y^1_{3D}, \cdots, Y^N_{3D})$. Similarly, 3D sparse landmark (3DSL) model is represented by χ^s_{3D}=$(X^1_{3D}, \cdots, X^N_{3D})$, where X^i_{3D} is the vector representation of M 3D landmarks. Given a 2D image, a sparse set of landmarks X^I_{2D} is first detected with any off-the-shelf detector. Then, the 3D-2D projection matrix P is estimated using least squares minimization, such that $X^I_{2D} = P\bar{X}_{3D}$, where $\bar{X}_{3D}$ is the mean of 3DSLs in the training database. By projecting each 3DSL via P, the corresponding 2D sparse landmark (2DSL) model χ^s_{2D}=$(X^1_{2D}, \cdots, X^N_{2D})$, where X^i_{2D} is the vector representation of M 2D landmarks, is generated on-line.

By applying PCA to the 3DSL and the 2DSL models, we derive a compact representations of the corresponding shapes A_m and A_n, based on which a PLS regression P_{PLS} [2] is learned, $\hat{A}_m = A_n P_{PLS}$:

$$X^i_{3D} = \bar{X}_{3D} + \sum_{m=1}^{N-1} a^i_m U^s_{3D} \qquad A_m = [a^1_m, a^2_m, \cdots, a^{N-1}_m] \quad , \tag{1}$$

$$X^i_{2D} = \bar{X}_{2D} + \sum_{n=1}^{N-1} a^i_n U^s_{2D} \qquad A_n = [a^1_n, a^2_n, \cdots, a^{N-1}_n] \quad . \tag{2}$$

Following the same procedure, we compute the compact representation of X^I_{2D} by solving for $a^I_n = {U^s_{2D}}^{-1}(X^I_{2D} - \bar{X}_{2D})$. Then the a^I_m is recovered by $a^I_m = a^I_n P_{PLS}$ and the 3DSL is constructed through $X^R_{3D} = \bar{X}_{3D} + a^I_m U^s_{3D}$.

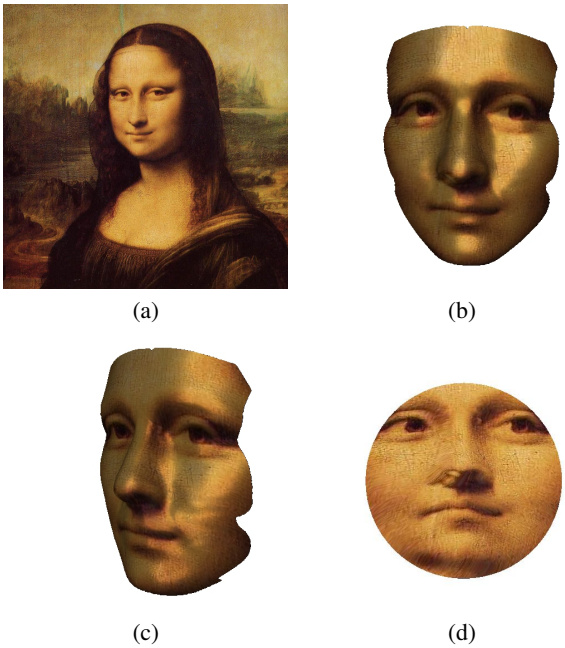
After we obtain the 3DSL X^R_{3D}, we aim to reconstruct the 3DDL Y^R_{3D}. In the training phase, the correlation between 3DSL and 3DDL is implicitly learned in a coupled manner.

$$\underset{\alpha, \Lambda^d_{3D}, \Lambda^s_{3D}}{\arg\min} \left\| \begin{bmatrix} \beta_0 \gamma^d_{3D} \\ \chi^s_{3D} \end{bmatrix} - \begin{bmatrix} \beta_0 \Lambda^d_{3D} \\ \Lambda^s_{3D} \end{bmatrix} \alpha \right\|^2_2 \quad s.t. \|\alpha\|_1 \leq \beta_1 \quad , \tag{3}$$

$$\underset{\alpha^*}{\arg\min} \left\| X^R_{3D} - \Lambda^s_{3D} \alpha^* \right\|^2_2 + \beta_2 \|\alpha^*\|_1 + \beta_3 \|\alpha^*\|_2 \quad . \tag{4}$$

By fitting X^R_{3D} to Λ^s_{3D}, the shared coefficient α^* could be recovered by solving Eq. 4. Then, the final Y^R_{3D} is reconstructed via Eq. 5:

$$Y^R_{3D} = \frac{\Lambda^d_{3D} \alpha^*}{\beta_0} \quad . \tag{5}$$



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

Figure 1: (a) The input 2D image; (b) frontal view of the face shape reconstruction; (c) profile view of the face shape reconstruction; and (d) lifted UV texture.

In the paper, we conducted several experiments using both synthetic data and real 2D face images from two face datasets. Compared with [6], our method demonstrates higher reconstruction accuracy and better robustness to face pose variations and landmark localization noise. Fig. 1 depicts the reconstructed 3D face of Mona Lisa using the famous painting by Leonardo da Vinci and the lifted texture in a pre-registered UV space.

[1] R. Dovgard and R. Basri. *Statistical symmetric shape from shading for 3D structure recovery of faces*. Springer, 2004.

[2] P. Geladi and B.R. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[3] T. Hassner. Viewing real-world faces in 3D. In *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 1-8 2013.

[4] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011.

[5] H. Rara, S. Elhabian, T. Starr, and A. Farag. Model-based shape recovery from single images of general and unknown lighting. In *Proc. IEEE International Conference on Image Processing*, pages 517–520, Cairo, November 2009.

[6] H.M. Rara, A.A. Farag, and T. Davis. Model-based 3D shape recovery from single images of unknown pose and illumination using a small number of feature points. In *Proc. International Joint Conference on Biometrics*, pages 1 –7, Washington, DC, October 2011.

[7] W.A.P. Smith and E.R. Hancock. Recovering facial shape using a statistical model of surface normal direction. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 28(12):1914–1930, December 2006.

BMVC 2014 Posters

There are two poster rooms:

Room 1 : Exchange building C3 (Poster board numbers **1-49**)
Room 2 : Exchange building C33 (Poster board numbers **50-98**)

Posters will be displayed throughout the conference. There are two formal poster sessions on Tuesday and Wednesday (13:30-14:45). Please do your best to ensure your poster is manned during both sessions.

Title	Authors	Board
Optimized Transform Coding for Approximate KNN Search	Minwoo Park, Kiran Gunda, Himaanshu Gupta, Khurram Shafique	1
Associating locations from wearable cameras	Jose Rivera-Rubio, Ioannis Alexiou, Anil Bharath, Luke Dickens, Riccardo Secoli, Emil Lupu	2
Interactive Shadow Removal and Ground Truth for Variable Scene Categories	Han Gong, Darren Cosker	3
Segmentation of Dynamic Scenes with Distributions of Spatiotemporally Oriented Energies	Damien Teney, Matthew Brown	4
The State of the Art: Object Retrieval in Paintings using Discriminative Regions	Elliot Crowley, Andrew Zisserman	5
Variational Level Set Segmentation in Riemannian Sobolev Spaces	Maximilian Baust, Darko Zikic, Nassir Navab	6
Robust segment-based Stereo using Cost Aggregation	Muninder Veldandi, Soumik Ukil, Krishna Govindarao	7
Coloured signed distance fields for full 3D object reconstruction	Wadim Kehl, Nassir Navab, Slobodan Ilic	8
Automatic Camera Calibration for Traffic Understanding	Marketa Dubska, Adam Herout, Jakub Sochor	9
Learning to Rank Bag-of-Word Histograms for Large-scale Object Retrieval	Danfeng Qin, Yuhua Chen, Matthieu Guillaumin, Luc Van Gool	10
Optimal Intrinsic Descriptors for Non-Rigid Shape Analysis	Thomas Windheuser, Matthias Vestner, Emanuele Rodola, Rudolph Triebel, Daniel Cremers	11
Fully Associative Ensemble Learning for Hierarchical Multi-Label Classification	Lingfeng Zhang, Shishir Shah, Ioannis Kakadiaris	12
Unlabelled 3D Motion Examples Improve Cross-View Action Recognition	Ankur Gupta, Alireza Shafaei, James Little, Robert Woodham	13
Location Constrained Pixel Classifiers for Image Parsing with Regular Spatial Layout	Kang Dang, Junsong Yuan	14
Unsupervised Learning of Generative Topic Saliency for Person Re-identification	Hanxiao Wang, Shaogang Gong, Tao Xiang	15
Regularized ℓ^1-Graph for Data Clustering	Yingzhen Yang, Zhangyang Wang, Jianchao Yang, Jiawei Han, Thomas Huang	16