
Bayesian Non-Parametrics for Multi-Modal Segmentation

Thesis for obtaining the title of
Doctor of Engineering Science
(Dr.-Ing.)
of the Faculty of Natural Science and Technology I
of Saarland University

by
Wei-Chen Chiu, M.Sc.

Saarbrücken
September 2016

Day of Colloquium 13th of September, 2016

Dean of the Faculty Univ.-Prof. Dr. Frank-Olaf Schreyer
Saarland University, Germany

Examination Committee

Chair Prof. Dr. Verena Wolf
Saarland University, Germany

Reviewer, Advisor Dr. Mario Fritz
Max Planck Institute for Informatics, Germany
Saarland University, Germany

Reviewer Prof. Dr. Vera Demberg
Saarland University, Germany

Reviewer Prof. Dr. Thomas Brox
University of Freiburg

Academic Assistant Dr. Shanshan Zhang
Max Planck Institute for Informatics, Germany

ABSTRACT

Segmentation is a fundamental and core problem in computer vision research which has applications in many tasks, such as object recognition, content-based image retrieval, and semantic labelling. To partition the data into groups coherent in one or more characteristics such as semantic classes, is often a first step towards understanding the content of data. As information in the real world is generally perceived in multiple modalities, segmentation performed on multi-modal data for extracting the latent structure usually encounters a challenge: how to combine features from multiple modalities and resolve accidental ambiguities. This thesis tackles three main axes of multi-modal segmentation problems: video segmentation and object discovery, activity segmentation and discovery, and segmentation in 3D data.

For the first two axes, we introduce non-parametric Bayesian approaches for segmenting multi-modal data collections, including groups of videos and context sensor streams. The proposed method shows benefits on: integrating multiple features and data dependencies in a probabilistic formulation, inferring the number of clusters from data and hierarchical semantic partitions, as well as resolving ambiguities by joint segmentation across videos or streams.

The third axis focuses on the robust use of 3D information for various applications, as 3D perception provides richer geometric structure and holistic observation of the visual scene. The studies covered in this thesis for utilizing various types of 3D data include: 3D object segmentation based on Kinect depth sensing improved by cross-modal stereo, matching 3D CAD models to objects on 2D image plane by exploiting the differentiability of the HOG descriptor, segmenting stereo videos based on adaptive ensemble models, and fusing 2D object detectors with 3D context information for an augmented reality application scenario.

ZUSAMMENFASSUNG

Segmentierung ist ein zentrales problem in der Computer Vision Forschung mit Anwendungen in vielen Bereichen wie der Objekterkennung, der inhaltsbasierten Bildsuche und dem semantischen Labelling. Daten in Gruppen zu partitionieren, die in einer oder mehreren Eigenschaften wie zum Beispiel der semantischen Klasse übereinstimmen, ist oft ein erster Schritt in Richtung Inhaltsanalyse. Da Informationen in der realen Welt im Allgemeinen multi-modal wahrgenommen werden, wird die Segmentierung auf multi-modale Daten angewendet und die latente Struktur dahinter extrahiert. Dies stellt in der Regel eine Herausforderung dar: Wie kombiniert man Merkmale aus mehreren Modalitäten und beseitigt zufällige Mehrdeutigkeiten? Diese Doktorarbeit befasst sich mit drei Hauptachsen multi-modaler Segmentierungsprobleme: Videosegmentierung und Objektentdeckung, Aktivitätssegmentierung und –entdeckung, sowie Segmentierung von 3D Daten.

Für die ersten beiden Achsen führen wir nichtparametrische Bayessche Ansätze ein um multi-modale Datensätze wie Videos und Kontextsensor-Ströme zu segmentieren. Die vorgeschlagene Methode zeigt Vorteile in folgenden Bereichen: Integration multipler Merkmale und Datenabhängigkeiten in probabilistischen Formulierungen, Bestimmung der Anzahl der Cluster und hierarchische, semantische Partitionen, sowie die Beseitigung von Mehrdeutigkeiten in gemeinsamen Segmentierungen in Videos und Sensor-Strömen.

Die dritte Achse konzentriert sich auf die robuste Nutzung von 3D Informationen für verschiedene Anwendungen. So bietet die 3D-Wahrnehmung zum Beispiel reichere geometrische Strukturen und eine holistische Betrachtung der sichtbaren Szene. Die Untersuchungen, die in dieser Arbeit zur Nutzung verschiedener Arten von 3D-Daten vorgestellt werden, umfassen: die 3D-Objektsegmentierung auf Basis der Kinect Tiefenmessung, verbessert durch cross-modale Stereoverfahren, die Anpassung von 3D-CAD-Modellen auf Objekte in der 2D-Bildebene durch Ausnutzung der Differenzierbarkeit des HOG-Descriptors, die Segmentierung von Stereo-Videos, basierend auf adaptiven Ensemble-Modellen, sowie der Verschmelzung von 2D-Objektdetektoren mit 3D-Kontextinformationen für ein Augmented-Reality Anwendungsszenario.

ACKNOWLEDGEMENTS

First of all, I would like to thank my doctoral supervisor, Dr. Mario Fritz, for all his support during my PhD career. His constructive advice and suggestions make me grow up in the research field of computer vision. I am grateful for his patience and encouragement when I faced the dilemma at work and life. He is more than a supervisor but also a mentor to me. Without his help, I couldn't go so far. Special thanks to Prof. Bernt Schiele, for his lead and also the efforts of creating a great research environment for the whole D2 group. I am grateful to Prof. Thomas Brox and Prof. Vera Demberg for serving as reviewers on my PhD thesis, and Prof. Verena Wolf for being the Chair of my defense.

I would like to thank people who I had chances to collaborate with: Dr. Ulf Blanke, Dr. Fabio Galasso, and Dr. Margret Keuper. Their kindness to share the experiences in research and also fruitful discussions help me improving the quality of my works. I am truly grateful for all my former or current colleagues of D2 group: Sandra, Marcus, Leonid, Bojan, Siyu, Maksim, Jan, Wenbin, Anja, Xucong, Mohamed, Anna, Yongqin, Yang, Abhishek, Joon, Hosna, Sabrina, Julian, Philipp, Qianru, Eldar, Evgeny, and Andrea. Those days of working and sharing life moments together will be always in my memory. Thanks for senior researchers and postdocs: Dr. Peter Gehler, Dr. Michael Stark, Dr. Andreas Bulling, Dr. Bjoern Andres, Dr. Mykhaylo Andriluka, Dr. Rodrigo Benenson, Dr. Zeynep Akata, Dr. Gaurav Sharma, Dr. Shanshan Zhang, and Dr. Yusuke Sugano, for showing me how a excellent researcher will be. Thanks for our secretary Connie Balzert for her helps of all the administrative stuffs. Especially, I would like to express my sincere thanks to my best office mate, Mateusz Malinowski, who is just like my brother to accompany me going through all the positive and negative periods, no matter in work or life.

I would like to thank my Taiwanese friends in Saarbrücken: Ting-Yen, Chia-Hui, Chia-Jui, Chi-Hsien, Jennifer, Yu-hsuan, Peggy, and all the others, for being my family in Germany. Especially thanks for I-Fang, her encouragement and company means a lots to me. Thanks my friends: Hung-Chun, Roxy, and Chun-Hao, for the mental support from the other parts of Germany. And thanks for my Kendo friends in the club Kendo ATV Dudweiler for making me able to continue my favorite sport.

Lastly and most importantly, my sincerest thanks to my dearest parents and families for unconditional love and support, and, everything.

CONTENTS

1	Introduction	1
1.1	Contributions	2
1.2	Thesis Outline	4
2	Related Work	7
2.1	Image Segmentation	7
2.2	Video Segmentation	8
2.3	Video Co-Segmentation	8
2.4	Segmentation on 3D Data	10
3	Video Co-Segmentation	12
3.1	Introduction	13
3.2	Overview of Approach	15
3.3	Chinese Restaurant Processes (CRP) and Dirichlet Process Mixture (DPM)	16
3.4	CRP-Based Video Co-Segmentation	17
3.4.1	Generative Procedure	17
3.4.2	Inference	18
3.4.3	Illustration on Synthetic Sequence	19
3.4.4	Implementation Details	20
3.4.5	Experimental Results of CRP-Based Video Co-Segmentation	20
3.4.5.1	Video Segmentation	21
3.4.5.2	Sketch-Based Video Retrieval	23
3.5	ddCRP-Based Video Co-Segmentation	25
3.5.1	Video Representation	25
3.5.2	Distance Dependent Chinese Restaurant Processes (ddCRP)	26
3.5.3	ddCRP Video Segmentation Prior	26
3.5.4	Generative Multi-Video Model	28
3.5.5	Posterior Inference via Gibbs Sampling	29
3.5.6	Implementation Details	30
3.5.7	Experimental Results of ddCRP-Based Video Co-Segmentation	31
3.5.7.1	Dataset	31
3.5.7.2	Follow-Up Datasets	32
3.5.7.3	Evaluation Metric	33
3.5.7.4	Comparison to Video Segmentation	34
3.5.7.5	Results	34
3.5.7.6	Discussion	35
3.5.7.7	Analysis with Over-Segmentation	37
3.5.7.8	Analysis on Different Granularities of Superpixels	38
3.5.7.9	Variants for Video Segmentation Prior	39

3.5.7.10	Runtime Comparison	40
3.6	Conclusion	42
4	Joint Segmentation and Activity Discovery	44
4.1	Introduction	45
4.2	Previous Works on Activity Discovery	46
4.3	Joint Segmentation and Discovery Approach	47
4.4	Discovery Framework	50
4.4.1	Context Word Extraction	50
4.4.2	Segmenting Context Words into Supersamples	50
4.4.3	Segmentation Priors for Activity Discovery	50
4.4.4	Joint Segmentation and Activity Discovery (ddCRP+CRP)	52
4.5	Evaluation Methodology	52
4.5.1	Dataset	53
4.5.2	Framework Implementation	54
4.5.3	Performance Estimation	54
4.6	Results	55
4.6.1	Semantic Relationships within and between Activities	55
4.6.2	Activity Discovery from Context Word Labels	55
4.6.2.1	ddCRP+CRP versus LDA	55
4.6.2.2	ddCRP+CRP versus CRF	56
4.6.2.3	Temporal and Semantic Priors	56
4.6.3	Sensitivity to Context Word Noise	57
4.6.4	Activity Discovery from Sensor Data	57
4.7	Discussion	58
4.8	Conclusion	60
5	3D Model Fitting by Differentiating HOG	61
5.1	Introduction	61
5.2	Related Work	63
5.3	∇ HOG	65
5.3.1	Gradients Computation	65
5.3.2	Weighted Vote into Spatial and Orientation Cells	66
5.3.3	Contrast Normalization	67
5.3.4	Implementation	67
5.4	Experimental Results	68
5.4.1	Reconstruction from HOG Descriptors	68
5.4.2	Pose Estimation	70
5.5	Conclusions	75
6	Adaptive Stereo Segmentation	77
6.1	Introduction	78
6.2	Consumer Stereo Video Segmentation Challenge (CSVSC)	80
6.3	Efficient Adaptive Segmentation of Stereo Videos	81
6.4	Efficient Segmentation Ensemble Model	82

6.4.1	Unifying Graph	82
6.4.2	Improved Efficiency with Graph Reduction	83
6.4.3	Details to Derive the Reduced Graph \mathcal{G}^Q	84
6.4.4	Implementation Details	86
6.5	Performance-Driven Adaptive Combination	87
6.5.1	Adaptive Combination by Regression	87
6.5.2	Performance-Driven Regressor Learning by Differentiable Proxies	87
6.5.2.1	Metric Specific Performance Proxy	88
6.5.2.2	Joint Learning of Regressor and Latent Parameter Combinations	90
6.5.3	Implementation Details	90
6.6	Experimental Results	90
6.6.1	Video Segmentations and Their (Static) Ensemble	91
6.6.2	EASVS and the State-of-the-art	91
6.6.3	Deeper Analysis of EASVS	92
6.7	Conclusions	95
7	Multi-Modal Stereo for 3D Object Segmentation	96
7.1	Introduction	97
7.2	Related Work	98
7.3	Cross-Modal Stereo	99
7.3.1	Stereo and Alignment to Kinect Depth	100
7.3.2	Cross-Modal Adaptation for IR-RGB-Stereo	102
7.3.3	Point Cloud Based Object Segmentation	102
7.3.4	Experiments	104
7.3.4.1	Results	104
7.3.4.2	Discussion	105
7.4	Learning Optimal Filters to Improve Cross-Modal Stereo	107
7.4.1	Capturing and Analyzing Sensor Characteristics of the Kinect	107
7.4.2	Experiments	111
7.4.2.1	Learning Filters	111
7.4.2.2	Evaluation	112
7.4.2.3	Discussion	113
7.5	Conclusions	114
8	Multi-Modal Multi-Part Object Detector	116
8.1	Introduction	116
8.2	Related Work	118
8.3	Object Disambiguation	118
8.4	Experiments	121
8.4.1	Object Disambiguation DataSet (ObDiDaS)	122
8.4.2	Object Disambiguation Metrics	122
8.4.3	Evaluation	125
8.5	Conclusion	127

9	Conclusions and Future Perspectives	129
9.1	Future Directions	130
	List of Figures	133
	List of Tables	135
	Bibliography	136
	Curriculum Vitae	152
	Publications	155

Contents

1.1	Contributions	2
1.2	Thesis Outline	4

MULTI-MODAL data, which contains the observation from multiple modalities such as image, motion, depth, and wearable sensors, is ubiquitous nowadays. Typically, the environment around us is too complex to be well described by a single modality. Therefore the multi-modal data with distinct characteristics from different sources provides an opportunity for holistic understanding of real-world scenarios. In computer vision research, the basic modality is the image, that captures appearance of the visual world into a 2D plane. When moving from static images to a sequence of images or videos, the temporal association between frames and the motion information exhibit another modality to outline the objects and understand how they move spatio-temporally. Furthermore, with extension from monocular system to stereo-vision or depth-perception sensors, the depth estimate provides again an additional informative modality capturing the 3D structure of the scene.

Learning feature presentations or proposing methods to incorporate the information from multiple modalities that are descriptive enough for discovering semantic knowledge is still one of the most important and challenging problems in computer vision. This thesis focuses especially on the segmentation task for multi-modal data, in order to partition the data into semantic groups as the latent structure. In general, the data distributions of different modalities represent diverse and distinct configurations. Accordingly, the data from multiple sources may not only contain additional observations, but also potentially lend to complementary information cross modalities. For instance, the motion information can be helpful to outline an animal with camouflage from grassland, whereas the appearance information is more discriminative to separate animals of different species under similar motion. Therefore the segmentation approaches should be able to combine and make the best use of the information from multiple modalities, while at the same time resolve ambiguities caused by the partial observation based on a specific modality.

This thesis tackles several segmentation problems upon multi-modal data along three main axes: video segmentation and object discovery, activity segmentation and discovery, and segmentation in 3D data. In the following the corresponding main contributions of this thesis with respect to the three axes are summarized.

1.1 CONTRIBUTIONS

Multi-Modal Video Segmentation and Object Discovery (Chiu and Fritz, 2013) Due to the popularity of smart phones with cameras and on-line platforms that share videos worldwide, video data is growing fast these years. Based on the rich information in appearance, motion, and spatio-temporal cues, segmenting video sequences into semantic regions representing the potential foreground, background, or objects of interest, provides an initial but important step for computers to understand video data. Furthermore, given a set of videos with shared object classes, the hierarchical structure of object classes across videos and local object instances within each video would lead to a much richer representation to benefit the segmentation. Therefore, we propose to address the *video co-segmentation* tasks jointly segmenting multiple videos where the discovered regions should correspond to objects and the regions belonging to the same object class are linked across videos.

- We propose non-parametric Bayesian approaches that incorporate appearance, motion and spatial-temporal features of visual data, define generative procedure of multiple video sequences, and infer the latent clusters which represent the local object instances and global object classes. Beyond the investigation of explicitly modelling motion and spatio-temporal distributions of local object instances, we further formulate the dependencies of motion and spatio-temporal distance between data points into a video segmentation prior in order to propose spatially contiguous segments of similar motion.
- We propose the first multi-object video co-segmentation dataset, which exposes challenges encountered in consumer or online video collections.
- The experimental results demonstrate superior performance of the proposed approach with respect to the state-of-the-art video segmentation and image co-segmentation methods. In addition, we improve co-segmentation results in comparison to independent video segmentation. This shows that the enriched feature representation across videos helps to resolve the ambiguities of incidental similarities in appearance or motion patterns within single videos.

Multi-Modal Activity Segmentation and Discovery (Seiter *et al.*, 2015) We analyze context word patterns of multi-modal sensor streams produced by multiple body worn and ambient sensors, which detect the mode of locomotion (e.g. walk, stand) as well as usages of objects (e.g. cup, spoon, and lazychair), in order to discover daily activities and routines for getting insights into human behaviour.

- We introduce a novel hierarchical topic model approach for joint segmentation and activity discovery that does not depend on manually selecting

parameters for segment size and number of topics. The method overcomes the limitation of time-invariant sliding windows by using a data-driven segmentation method based on state changes in context words to obtain supersamples, which are basic data units used in the topic model approach. We propose a segmentation prior to model the temporal distances and the semantic distances between supersamples based on word2vec representations of context words, and use the non-parametric distance dependent Chinese Restaurant Process to cluster supersamples into groups as potential activities.

- We present results on a multi-modal activity dataset and improve over both parametric LDA and non-parametric CRF approaches. Also, we provide an analysis of discovery performance based on context word labels, context word detections from sensor data, and synthetic context word noise.

Multi-Modal Segmentation in 3D Data We investigate multi-modal segmentation and recognition based on 3D data sensed in different manners, including consumer stereo cameras, Kinect depth sensors, monocular SLAM system, and 3D CAD models. In comparison to static images that represent a partial observation of the world, the 3D data provides richer geometric information for understanding the visual content.

- 3D Model Fitting (Chiu and Fritz, 2015)
 - We revisit the feature computation of the Histogram of Oriented Gradient (HOG) descriptor and exploit its piecewise differentiability. The experiments on pre-image visualization given HOG features present a proof-of-concept for the differentiable HOG.
 - We propose a CAD model-to-image alignment approach that re-implements the exemplar LDA pipeline by integrating differentiable HOG with an approximate renderer. By parameterizing the vertice locations of CAD model with pose parameters, our approach enables end-to-end optimization for continuous 3D pose estimation. The matching between the CAD model and target image not only can lift the 2D appearance information into 3D space but also has potential to help segment the object based on the 2D projection of 3D model.
 - We experimentally show that our proposed method improves over the state-of-the-art which relies on pre-rendering views exhaustively.
- Adaptive Stereo Segmentation (Chiu *et al.*, 2016)
 - We address the stereo video segmentation task by proposing an ensemble method which combines a pool of image and video segmentations. The model is represented by a graph and parameterized by the importances of pooled segmentations as well as multiple feature distances. At training time, the optimal parameters for training videos with

- respect to the stereo segmentation performance are found directly via a differentiable proxy. At test time, we regress the combination parameters based on color, depth and motion statistics of the target stereo video, and therefore achieve an adaptive stereo segmentation.
- We propose a consumer stereo video segmentation challenge which contains videos, annotations, and metrics to measure the segmentation performance.
 - Experimentally, we show that the adaptive scheme further improves over the static combination, the initial segmentations, video co-segmentation, as well as a most recent RGB-D segmentation technique.
- Cross-Modal Stereo for 3D Object Segmentation (Chiu *et al.*, 2011b,a)
 - The Kinect depth sensor fails to estimate the depth on some common materials in daily life, such as glasses and polishing metal. We tackle this problem by utilizing different modalities provided by computer vision techniques in order to reduce these artifacts. In particular, we propose a cross-modal stereo composed of Kinect’s RGB and IR cameras and further learn optimal spatial filters to improve the stereo matching performance.
 - We evaluate an object segmentation task on the 3D data estimated by the improved Kinect. The experimental results show better segmentation performance, due to improved depth sensing on reflective and transparent objects.
 - Multi-Part Object Detector and Disambiguation (Chiu *et al.*, 2014)
 - We consider an application scenario of recognizing a factory machine that consists of potentially repetitive machine parts. We propose a multi-part object detection system that utilizes the depth information extracted by SLAM approach to lift the 2D object-part detector outputs into 3D space. Our approach uses the spatial context in order to identify different machine parts and resolve the ambiguities between parts of the same category.
 - We propose the first benchmark for this task, that is composed of an annotated dataset as well as a metric that approximates human judgement. The experiments show the ability of the propose system to predict the identities of individual parts and localize them in the factory scene.

1.2 THESIS OUTLINE

The following overview states the relations of the chapters to corresponding publications.

Chapter 2: Related Work. In this chapter we show a systematic overview for the

prior works of *segmentation on multi-modal data*, with particular focus on several data types: video sequences, video sets, and visual data with 3D information.

Chapter 3: Video Co-Segmentation. We tackle the task of multi-class video co-segmentation. The proposed framework models the instance-class hierarchy of multiple videos as well as joint probability distribution over appearance, motion, and spatio-temporal observations, in order to discover global object classes shared across videos and outline object instances in the video. In addition to segmenting out objects, we also show an application of sketch-based video retrieval based on the latent structure discovered from videos.

The content of this chapter corresponds to the CVPR 2013 publication *Multi-Class Video Co-Segmentation with a Generative Multi-Video Model* (Chiu and Fritz, 2013) in Section 3.5 and a few unpublished materials in Section 3.4. Wei-Chen Chiu is the lead author of this paper.

Chapter 4: Joint Segmentation and Activity Discovery. While the previous chapter applies a non-parametric Bayesian framework in video co-segmentation task, its generative, probabilistic formulation for data clustering makes it a perfect fit to be generalized for different applications. In particular, in this chapter we show the application on the joint segmentation and activity discovery from context sensor data streams, with modelling the temporal and semantic distances between context words by distance-dependent Chinese Restaurant Process as prior.

The content of this chapter corresponds to the Percom 2015 publication *Joint Segmentation and Activity Discovery using Semantic and Temporal Priors* (Seiter *et al.*, 2015). The paper is based on a collaboration with the Wearable Computing Lab, ETH Zurich, Switzerland. Wei-Chen Chiu contributes to adapt the core non-parametric Bayesian approach from video co-segmentation scenario for the framework of inferring activity routines from context sensor signals.

Chapter 5: 3D Model Fitting by Differentiating HOG. In this chapter we show the differentiability of HOG feature representation and use it in the exemplar LDA pipeline combined with a differentiable renderer OpenDR, which enables the end-to-end optimization for continuous 3D pose estimation. In addition to using ∇ HOG for 3D model fitting, we also show its application on pre-image reconstruction given HOG features, as a proof-of-concept and another example of exploiting the differentiable HOG-based pipelines.

The content of this chapter corresponds to the ICCV 2015 publication *See the Difference: Direct Pre-Image Reconstruction and Pose Estimation by Differentiating HOG* (Chiu and Fritz, 2015). Wei-Chen Chiu is the lead author of this paper.

Chapter 6: Adaptive Stereo Segmentation. We tackle the task of segmenting consumer stereo video data in which there are many feature cues (e.g. motion, depth, color) as well as diverse set of segmentation approaches in image and video scenarios proposing different partitions of the visual data. We learn to predict the importance of the features cues and segmentation proposals depending on statistical properties of the data, and aggregate those information in a parametrized similarity graph which can be utilized by spectral clustering technique to produce the final stereo video segmentation.

The content of this chapter corresponds to an ACCV 2016 publication *Towards Segmenting Consumer Stereo Videos: Benchmark, Baselines and Ensembles* (Chiu *et al.*, 2016). Wei-Chen Chiu is the lead author of this paper.

Chapter 7: Multi-Modal Stereo for 3D Object Segmentation. We study the problem of object segmentation in multi-modal 3D data sensed by Kinect. As Kinect's active depth estimate has difficulties on specular or transparent surfaces, we propose cross-modal stereo built upon RGB and IR cameras of Kinect to reduce missing regions on Kinect's depth map. The improved depth estimate improves the 3D object segmentation performance.

The content of this chapter corresponds to the BMVC 2011 publication *Improving the kinect by cross-modal stereo* (Chiu *et al.*, 2011b) and its extension at the *Consumer Depth Cameras for Computer Vision* workshop held at ICCV 2011: *I spy with my little eye: Learning optimal filters for cross-modal stereo under projected patterns* (Chiu *et al.*, 2011a). Wei-Chen Chiu is the lead authors of both papers.

Chapter 8: Multi-Modal Multi-Part Object Detector. We present a system that recognizes objects composed of repetitive parts by utilizing 2D object detections, depth information from SLAM, and the 3D context of object layout. In particular, we focus on an application scenario of assisting a maintenance worker by providing an augmented reality overlay that identifies and disambiguates machine parts.

The content of this chapter corresponds to the BMVC 2014 publication *Object Disambiguation for Augmented Reality Applications* (Chiu *et al.*, 2014). The paper is based on a collaboration with Intel Visual Computing Institute, Saarbrücken, Germany. Wei-Chen Chiu is the lead author of this paper.

Contents

2.1	Image Segmentation	7
2.2	Video Segmentation	8
2.3	Video Co-Segmentation	8
2.4	Segmentation on 3D Data	10

IN computer vision, image segmentation means to partition the image into spatially contiguous regions that are similar regarding appearance features (e.g. bandpass filter responses, color, texture), respect image boundaries, or represent potential objects. Large body of works has been proposed to address this problem and it is still one of the major research areas to date. While generalizing the target data of segmentation task from images to multi-modal visual data such as videos and stereo sequences, it brings another big challenge of analysing the information from multiple sources simultaneously to extract the semantic groups and infer the latent structure, as different features can be extracted from different modalities. In this chapter we provide an overview over the related approaches on segmentation over different multi-modal data types, ranging from video sequences, video sets, to 3D data such as stereo videos.

2.1 IMAGE SEGMENTATION

Without loss of generality, image segmentation methods attempt to divide an image into groups of pixels such that the pixels of a group are similar, and pixels from different groups are dissimilar. These criteria correspond to two main categories of image segmentation approaches: clustering-based and graph-based methods. Clustering-based methods cluster the pixels with high similarity into a group based on certain feature representations, typical examples include K-means and Mean-shift algorithms (Comaniciu and Meer, 2002). And graph-based methods, e.g. (Felzenszwalb and Huttenlocher, 2004; Shi and Malik, 2000), normally define an affinity graph where the pixels are treated as vertices and the edges measure the distances between nodes regarding various feature cues, such that the edges between pixels in a segment are expected to have relatively low weights while edges between pixels in different segments are with higher weights. While generalizing image segmentation approaches to video data, the additional modality from motion information leads to a need of having algorithms that are able to segment multi-modal data.

2.2 VIDEO SEGMENTATION

Learning a better representation and inferring the latent structure of video sequences has been a concern of the computer vision community and has recently received much attention. By segmentation of videos into group of objects or regions which are coherent in appearance and motion, it delivers the first step to interpret the video content and thus become the base of many higher-level computer vision tasks, such as object tracking, scene labelling, and activity recognition. As image segmentation, many widely-used video segmentation approaches can be categorized into clustering or graph-cut based methods, under various choices for basic data units: pixels/voxels, superpixels/supervoxels, and tracking trajectories of previous ones. For instances, in (Ochs and Brox, 2011) long term point trajectories based on dense optical flow are used to cluster the feature points into temporally consistent segmentations of moving objects in the video. Similarly, in (Galasso *et al.*, 2011) with introduction of probabilistic region trajectories, they proposed to use spatial-temporal clustering on trajectories based on motion. The graph-based approaches (Grundmann *et al.*, 2010) and (Xu and Corso, 2012) define affinities between supervoxels and the edges are weighted by different cues such as color, motion or texture. The final segmentation is derived by grouping the supervoxels. Although their methods provide plausible solutions on video segmentation tasks, a single video only contain the partial observation of the objects thus can suffer from ambiguities due to the indistinguishable motion or appearance between nearby objects. In addition, some video segmentation works (Fragkiadaki *et al.*, 2012; Di *et al.*, 2013) face the problem of making an explicit choice on the number of clusters, which is undesirable in unsupervised settings.

2.3 VIDEO CO-SEGMENTATION

To ease the difficulties met by viewing a single video, the video co-segmentation task is studied where the objects belonging to common classes across videos can provide additional information and benefit the segmentation. The co-segmentation idea was originated from the works on image co-segmentation (Rother *et al.*, 2006), which denotes segmenting the common parts of an image pair. And with introducing an explicit object notion, it was extended to object co-segmentation and object discovery. This idea has seen several refinements (Vicente *et al.*, 2010; Kim *et al.*, 2012; Collins *et al.*, 2012; Mukherjee *et al.*, 2011; Kim *et al.*, 2011; Chang *et al.*, 2011) and today's state-of-the-art in co-segmentation can handle multiple objects (Joulin *et al.*, 2012; Kim and Xing, 2012). While image co-segmentation only looks at single frames, the video co-segmentation task starts to consider spatio-temporal structure and motion information within the video collection.

Initial attempts (Rubio *et al.*, 2012; Chen *et al.*, 2012) have been made to approach video co-segmentation task with a binary foreground/background segmentation formulation. But this setting makes quite strong assumptions and eliminates the

problem of associating segments of multiple classes across frames and videos. The segmentation is performed foreground vs. background and has only been evaluate on a set of visually very similar videos. Furthermore, (Chen *et al.*, 2012) utilize a set of videos with exact the same object instance and (Rubio *et al.*, 2012) restricts the condition for common objects to be of similar motion across videos. These simplifying assumptions limit the applicability of these methods.

In contrast, our method in this thesis is the first to address less structured videos containing multiple objects, denoted as *multi-class video co-segmentation* task. Our approach is based on a non-parametric Bayesian framework where the basic data units within each video are grouped into object instances subject to the appearance, motion, and spatio-temporal features. The appearance models of object instances are further clustered across videos again into global object classes. The proposed framework relates to the application of topic models in the image domain (Sivic *et al.*, 2005) in terms of discovering objects and appearances. While the work (Sivic *et al.*, 2005) has been extended to handle also spatial information (Wang and Grimson, 2007) as well as part notions in infinite mixture models (Sudderth *et al.*, 2008) and motion Kuettel *et al.* (2010), our model represents an extension of these idea to video sets where their hierarchical structures of object classes and instances are inferred. The details of how we build up the model from the basic Dirichlet Process Mixture to handle the video collection are described in the chapter 3. There are several benefits of the proposed non-parametric Bayesian approach: The employment of non-parametric prior overcomes the issue of choosing a particular number of classes or instances. In addition, the global appearance models in the proposed method relax the assumption of (Joulin *et al.*, 2012; Rubio *et al.*, 2012; Chen *et al.*, 2012) that object reappears to a weaker one which assumes that the objects are shared between videos, therefore co-segmentation can be encouraged but not enforced. Although this makes the co-segmentation assumption weaker, we believe that it is a more realistic one.

Moreover, it is worthwhile to mention that upon the publication of our method on multi-class video co-segmentation task (Chiu and Fritz, 2013), several research works have been following up by proposing different approaches (Wang *et al.*, 2014a; Fu *et al.*, 2014; Zhang *et al.*, 2014a; Wang *et al.*, 2014b; Guo *et al.*, 2013, 2014; Lou and Gevers, 2014; Barhoumi, 2015; Joulin *et al.*, 2014). For instance, the recent works such as (Fu *et al.*, 2014; Zhang *et al.*, 2014a) utilize object proposals (Endres and Hoiem, 2010) to first extract the potential object of interests, then construct a graph to model the inter-frame consistency and across-video similarity of candidate proposals, and finally optimize to discover the common object classes for co-segmentation. However, along with the studying of the video co-segmentation task, its problem definition is still under discussion with various assumptions across different works. For instance, as originated from the image co-segmentation task, some works impose hard constraints that all the object classes of interest must be contained in all images in the image set, in order to provide a weak supervision for the co-segmentation algorithms to identify the primary objects. And other works such as (Guo *et al.*, 2014; Fu *et al.*, 2014; Zhang *et al.*, 2014a) propose to handle the general cases of multiple

objects, temporary occlusions, and possible moving in/out for objects. While in our work, we follow a stricter unsupervised setting which covers almost all the general difficulties indicated in the related works to have no prior knowledge of the number of object classes, nor the assumption of that the common object classes should appear in each video sequence in the video set, as well as support the varying number of object instances and potential appearance changes of object classes in each video.

2.4 SEGMENTATION ON 3D DATA

In addition to the video sequences which exploit the appearance and motion characteristics, the use of depth is now becoming ubiquitous in many cases, e.g. gaming interface and robotics platform, to provide the additional information for the structural and semantic properties of the visual scenes. For instance, in a simple tabletop scheme shown as Figure 7.1 which is commonly seen in daily life, the depth information can represent the scene as point clouds in 3D space, then the naive clustering-based segmentation is able to group the 3D points into potential objects.

The utilization of depth in segmentation framework naturally relates to the research works on depth/stereo segmentation. There is a long tradition of work on 3D reconstruction which estimates 3D coordinates, thus depth, from pair or multiple views (Kanade and Okutomi, 1994; Scharstein and Szeliski, 2002). These efforts have been recently combined with reasoning on the object appearance and the physical constraints of the 3D scene in the work of (Bleyer *et al.*, 2012), whereby segmentation proposals are produced for semantic objects. These methods generally require high quality images. In other words, most of the stereo algorithms are based on the well-defined stereo-rigs which have fixed calibration parameters and controlled environments. However, in consumer stereo content or real-world cases, it is very easy to see the difficulties for current stereo algorithms from the problems such as wide variety of capture devices, zooming in/out effects, camera moving and synchronization. Therefore, making the segmentation algorithms capable of handling noisy 3D data is one of the challenging issues. In addition, the underlying assumption of a static scene for these methods does not allow direct applicability on video data with 3D information, such as stereo videos.

To move from segmentation on static 3D data to video types, the techniques used in video segmentation to model the spatio-temporal continuity provide a rich source of references, as mentioned above in the section 2.2. Although these video-based methods produce temporally consistent segmentation, we note that none of those techniques may seamlessly generalize to videos with 3D data. Another line of related works come from recent developments in scene flow estimation for stereo videos, which tackle the problem of jointly estimating optical flow and the depth, assuming calibrated (Huguet and Devernay, 2007; Basha *et al.*, 2013) or uncalibrated cameras (Vogel *et al.*, 2011). Although the scene flow provides the information of 3D motion fields which is able to give more insights into the geometric layout of the visual scene, these works do not address segmentation. Elsewhere, video

segmentation is addressed by considering RGB-D information (den Bergh and Gool, 2012; Weikersdorfer *et al.*, 2013; Hickson *et al.*, 2014), Kinect colour images with depth. Nevertheless, the same problem of potentially noisy or uninformative depth estimation is still unanswered, in particular for consumer stereo videos which are not assumed calibrated nor from the same camera.

Recent research on image and video co-segmentation tasks provide a way to jointly extract common objects across multiple images or videos (Kim and Xing, 2012; Joulin *et al.*, 2012; Fu *et al.*, 2014; Chiu and Fritz, 2013). The general assumption of co-segmentation problem on commonality of objects in a video set makes it a plausible fit to stereo video segmentation task when we consider left and right videos of stereo pairs as two separate sequences in the same set. However, from this perspective the depth cue in the stereo videos is not explicitly explored which can provide rich information to outline objects when other features are with ambiguities.

Another issue generally observed in segmenting multi-modal data is: the discriminative power of feature cues shown in different modalities changes along various conditions, due to the diverse properties and wide variations in the real-world environment. In addition, since the heterogeneity of the segmentation algorithms designed for various domains and application targets, almost no single method with fixed integration over multiple features can guarantee to perform best in all situations. Accordingly, there is a desire to dynamically select or combine suitable algorithms according to the characteristics of the data. There has been previous work which attempted to select the best algorithm from a candidate pool depending on the specific task. (Mac Aodha *et al.*, 2010) presented a supervised learning approach to predict the most suitable optical flow algorithm, based on the confidence measures of the optical flow estimates. By contrast, as going to be shown in chapter 6, for the segmentation problem on consumer stereo videos, we extend the work of (Li *et al.*, 2012) to *combine* the available image and video segmentation algorithms, adaptively weighting their contributions based on the statistical properties of the target stereo video.

Overall, one task covered in our research is aiming at developing segmentation methods on 3D data which are able to cope with noisy 3D input as well as combine the information from multiple modalities, e.g. appearance, motion, depth, to resolve the ambiguities happened occasionally in any single modality.

Contents

3.1	Introduction	13
3.2	Overview of Approach	15
3.3	Chinese Restaurant Processes (CRP) and Dirichlet Process Mixture (DPM)	16
3.4	CRP-Based Video Co-Segmentation	17
3.4.1	Generative Procedure	17
3.4.2	Inference	18
3.4.3	Illustration on Synthetic Sequence	19
3.4.4	Implementation Details	20
3.4.5	Experimental Results of CRP-Based Video Co-Segmentation	20
3.5	ddCRP-Based Video Co-Segmentation	25
3.5.1	Video Representation	25
3.5.2	Distance Dependent Chinese Restaurant Processes (ddCRP)	26
3.5.3	ddCRP Video Segmentation Prior	26
3.5.4	Generative Multi-Video Model	28
3.5.5	Posterior Inference via Gibbs Sampling	29
3.5.6	Implementation Details	30
3.5.7	Experimental Results of ddCRP-Based Video Co-Segmentation	31
3.6	Conclusion	42

IN this chapter we present the video co-segmentation task for discovering the latent structure and object instance-class hierarchies from a set of videos. Based on feature points or superpixels as basic observations, the proposed approaches infer the semantic information of videos in an unsupervised manner based on non-parametric Bayesian models which provide the generative, probabilistic formulation for cluster analysis and have the property of allowing the data to determine the complexity of the model. We investigate different modelling aspects to utilize the appearance, spatial-temporal and motion features of the video content. The clusters of the data result in segmentations that not only outline object instances within each video but also establish their class correspondences across different video sequences. The experiments demonstrate the flexibility and capabilities of the proposed methods in various applications, including video segmentation, video indexing and multi-class video co-segmentation.

3.1 INTRODUCTION

Video data is one of the fastest growing resource of publicly available data on the web. Leveraging such resources for learning and making it accessible and searchable in a easy way is a big opportunity – but equally a big challenge. It is desirable to provide an initial analysis of the video sequence in order to support object extraction, recognition or indexing. In order to achieve this goal, algorithm must be able to deal with the unstructured nature of such videos which is still an open challenge nowadays.

Processing video data also plays an important role in applications such as surveillance, digital asset management and robotics. Key tasks are indexing and retrieval of video content, summarization and recognition of activities. All these tasks refer to an underlying semantic structure in the video. We might see this as an underlying plot which the observed sequence seems to obey. For humans, this underlying structure is a natural and also compact way to communicate about videos.

Video segmentation and tracking-based approaches have been proposed in order to approach this problem (Darrell and Pentland, 1991; Wang and Adelson, 1993, 1994). As motion and spatio-temporal structures in videos provide rich cues about potential object boundaries and independently moving objects, good progress has been made by first forming low level feature tracks which are later aggregated to potential object segments in a clustering scheme (Brox and Malik, 2010; Ochs and Brox, 2011; Galasso *et al.*, 2011; Lezama *et al.*, 2011).

However, this approach has multiple inherent limitations. The main reoccurring problem in these approaches is the difficulty of choosing the right number of segments or objects in a sequence which requires different heuristics. Also, as a single video might only expose a partial view, accidental similarities in appearance and motion patterns might lead to an ambiguous or even wrong segmentation. In addition, performing video segmentation independently on each video of a video collection does not reveal any object class structure between the segments that would lead to a much richer representation.

We draw two conclusions. First, segmentations should be treated in a probabilistic framework in order to account for uncertainty. Second, a richer problem set should be investigated where the approach is enabled to reason across multiple video sequences in order to collect additional evidence that is able to link segments across videos, which corresponds to the so-called co-segmentation task.

Our approach is based on a non-parametric Bayesian model which addresses the first problem in a principled way by forming prior over different object groupings over time, and also builds global appearance models shared across videos to refer object class notions in the process of inferring the latent structure of the video. Therefore the obtained representation can indeed be seen as a summary of how different entities move in a video sequence as it relates segments across video sequences by global appearance classes. In addition to learning of global appearance classes, to better utilize the important information from motion and spatio-temporal cues in a video for discovering object instances, in this chapter we sequentially study

two variants of our non-parametric Bayesian model.

For the first variant (Section 3.4), we explicitly model the motion and spatio-temporal distributions of object instances within each video sequence which shows how objects move in the video volume over time. In experiments, We show that the algorithm – despite it’s minimal set of modelling assumption performs competitively with state of the art video segmentation algorithms. Beyond video segmentation, we also show the flexibility of our approach by exemplifying its use case on the application scenarios of video indexing, where we make use of the low dimensional structure of the video in order to compare video sequences and find the most relevant one. We illustrate how a sketch of a simple action taken place in a video can be related to the latent structure and therefore be used to retrieve relevant videos.

For the second variant (Section 3.5), we base on distance-dependent Chinese Restaurant Processes to formulate a video segmentation prior in order to propose contiguous segments of coherent motion for video data. For the purpose of quantitatively evaluate the performance of outlining local object instances as well as discovering global object classes across videos, we establishes a multi-class video co-segmentation challenge which is performed on realistic video sequences. We give further insights by analysing the different model components of video segmentation prior in isolation and providing an analysis of runtime.

The main contribution of this chapter is to propose a non-parametric Bayesian based model that is able to describe the generative procedure of video sequences including the object instances built upon basic feature points or superpixels, and also the property of global appearance sharing across object instances belonging to the same global class. The proposed method does not rely on any low level tracks and rather solve feature association as well as global appearance classes jointly, while neither the global number of appearance classes nor the number of instances in each video is known.

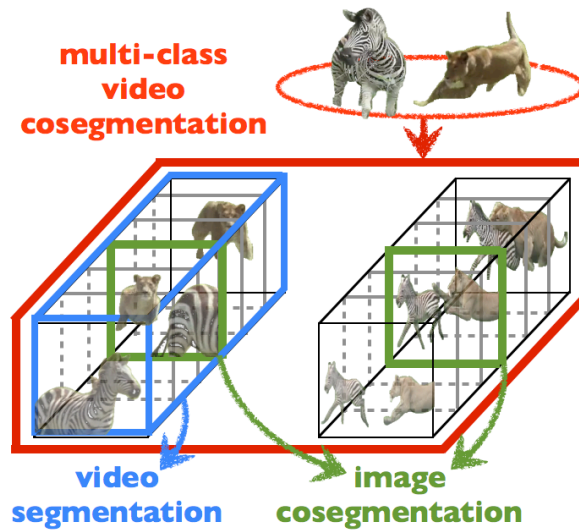


Figure 3.1: Our proposed multi-class video co-segmentation model addresses segmentation of multiple object classes across multiple videos. The segments are linked within and across videos via the global object classes.

3.2 OVERVIEW OF APPROACH

The goal of this chapter is to perform segmentation across multiple videos where the segments should correspond to the objects and segments of the same object class are linked together within and across videos.

In order to provide a clear problem statement of the multi-class video co-segmentation task, we extend the definition of image co-segmentation from the work of (Vicente *et al.*, 2010): given a set of video sequences, the task of *multi-class video co-segmentation* is to segment the common foreground objects from the potentially arbitrary backgrounds given that the distribution of appearance features of the common foreground objects are shared in subsets of videos. The property of shared distribution of appearance for common foreground objects supports the hierarchy of object instances and classes. The common object classes can appear in some videos with similar appearance but are able to have different motions, deformations for its instances across different sequences.

As motivated above, video segmentation on each video independently can lead to ambiguities that only can be resolved by reasoning across sequences. In order to deal with this problem, we approach video cosegmentation by a generative model where videos are linked by a global appearance model. In order to be able to deal with an unknown number of object classes and object instances in each video, we make use of non-parametric Bayesian modelling based on Dirichlet Processes.

In the following sections, we first give an overview of Chinese Restaurant Processes (CRP) (Pitman, 2006) which gives an effective way to represent a Dirichlet Process (Section 3.3). Then two different formulations of the our proposed probabilis-

tic model and their corresponding experimental results are described step by step. First, in Section 3.4 we present a CRP-based video cosegmentation approach which treats all appearance, spatial-temporal and motion features as observations, and uses typical CRP as the prior over data partition. The hierarchical structure with local mixture models for spatial-temporal and motion features in each video as well as an infinite mixture model for the global appearance classes across videos addresses the object instances with different position preferences and moving patterns together with their belonging object classes. We demonstrate the applications on video segmentation and also the video retrieval which benefits from the explicit modelling of motion distributions. However, a significant challenge is that single Gaussian models for local motion and spatial-temporal features, although good in some instances (e.g. cars), does not work well on articulated or irregular-shaped objects. Thus we present another formulation in Section 3.5 (Chiu and Fritz, 2013) which instead bases on distance dependent Chinese Restaurant Process (ddCRP) to encode the local dependencies coming from spatial-temporal and motion distances between data points, noted as video segmentation prior, in order to propose contiguous segments of coherent motion. Similar to the previous approach, this formulation of video co-segmentation also contains a global appearance model for representing object classes shared across multiple videos.

3.3 CHINESE RESTAURANT PROCESSES (CRP) AND DIRICHLET PROCESS MIXTURE (DPM)

We briefly introduce the basic idea of Chinese Restaurant Processes (CRP). CRP is an alternative representation of Dirichlet process model and it can be understood as the following procedure. Imagine a Chinese restaurant with an infinite number of tables. A sequence of customers come enter the restaurant and sit at randomly to any of the occupied tables or to the first available empty tables. The i -th customer sits down at a table with a probability that is proportional to how many customers are already sitting at that table or opens up a new table with a probability proportional to a hyperparameter, which is usually named *concentration parameter*. Their seating configuration represents a random partition also called *table assignments*. Thus CRP provides a flexible prior distribution over table assignments where the number of tables is potentially infinite. Since the table assignment of each customer just depends on the number of people sitting at each table and is independent of the other ones, the ordering of customers does not affect the distribution over partitions and therefore exchangeability holds.

Dirichlet Process Mixture Model (DPM) (Escobar and West, 1995; Neal, 2000; Teh *et al.*, 2006) constructs a single mixture model in order to perform data clustering in which the number of mixture components can be infinite therefore we do not need to manually assign the number of clusters. The generative procedure of DPM can be

written as:

$$\begin{aligned} G|H_a &\sim DP(\gamma, H) \\ \theta_i|G &\sim G \\ x_i|\theta_i &\sim F(\theta_i) \end{aligned} \tag{3.1}$$

where G is the set of cluster parameters and it is distributed according to a Dirichlet Process $DP(\gamma, H)$. For the observation x_i belongs to the class i it is generated conditionally depending on the generative distribution $F(\theta_i)$ where the cluster parameter θ_i is sampled from G . In practice, the base distribution H of Dirichlet Process is usually chosen to be conjugate prior to the generative distribution F in order to have more efficient computation for the posterior inference to get the clustering results. In the inference procedure, it is required to estimate the cluster assignment for each observation x_i , as shown in the generative procedure above. The most popular tool used for inference on the DPM models is the Gibbs sampling scheme to iteratively update the assignments of observations, which utilizes the Chinese Restaurant Process to consider the conditional distribution of one cluster assignment given all others.

3.4 CRP-BASED VIDEO CO-SEGMENTATION

3.4.1 Generative Procedure

For the representation of video sequences, we consider a generative procedure: Videos consist of different global object classes with different appearances, and for every video there are arbitrary number of instances which are located at different locations and possibly move over time. Those object instances emit appearance, spatial-temporal and motion features as observed variables here in our model. Our probabilistic, generative model aims to infer the latent structure and object instance-class hierarchies of a video set. Our model is based on non-parametric Bayesian approach and its graphical model is shown in Fig. 3.2, we describe the details of our proposed method in following section.

First the Dirichlet Process Mixture Model is utilized to model the appearance classes as multinomial distributions over the codewords, where only appearance information is considered. Suppose G is the global set of appearance classes which is drawn from a Dirichlet Process $DP(\gamma, H_a)$. The multinomial distribution $F(\theta_i)$ is configured by object-specific parameter θ_i and is used to generate a_i observations. The visualization of this model is illustrated in Fig 3.2(a).

In order to also take spatial-temporal and motion information into consideration, the position of each image patch located in the video volume and the motion vector of central pixel in the patch are used as features. Given the observation that for the object instances belonging to the same category will be with the similar appearance, but they can have different local spatial-temporal distributions and different motions, we would like to extend our graphical model to capture this characteristic, hence Hierarchical Dirichlet Process (HDP) is used (Teh *et al.*, 2006). Instead of DPM

which models observations as a single set of object classes, the HDP can share multiple parameters among several object instances. Therefore, we extend the HDP to put a Dirichlet prior H_a on feature appearance distribution in higher layer of the hierarchical structure to flexibly model its globally sharing property, then put the normal-inverse-Wishart priors H_s and H_m on 3-dimensional Gaussian of spatial-temporal distribution and on 2-dimensional Gaussian of motion distribution in lower layer to locally model different position preferences and motion patterns of object instances as shown in Fig 3.2(b). In addition, because now the spatial-temporal and motion distributions are located on different layer from the appearance model, the position and motion of the object instance is independent of its belonging object class. We use absolute image position.

To interpret the generative procedure more clearly, for video sequences, we first sample object classes with distinctive appearance distributions G_0 from Dirichlet Process prior $DP(\gamma, H_a)$, then for every object instance vj in the video v , we will combine the class-specific appearance distribution with its spatial-temporal and motion distribution sampled according to the prior H_s and H_m . Finally we can get generate feature points $(a_{vji}, s_{vji}, m_{vji})$ using the distribution of parameter θ_{vji} :

$$\begin{aligned}
 G_0 | H_a &\sim DP(\gamma, H_a) \\
 G_{vj} | G_0, H_s &\sim DP(\alpha, G_0 \times H_s \times H_m) \\
 \theta_{vji} | G_{vj} &\sim G_{vj} \\
 a_{vji}, s_{vji}, m_{vji} | \theta_{vji} &\sim F(\theta_{vji})
 \end{aligned} \tag{3.2}$$

Note here if the number of videos is more than one, then the proposed model can learn the global classes across videos which provides potential for further applications, for instance, the similar object retrieval on a group of videos.

3.4.2 Inference

To do the inference for our proposed model, we need to infer the correspondences between extracted image patches and object instances, similar to the one between object instances and classes. Here we introduce two indicator variables to present those correspondences: t_{vji} for the i^{th} image patch belonging to j^{th} object instance in video v ; and k_{lvj} for mapping from object instance vj to object class l . For the i^{th} image patch in object instance vj is described by $(a_{vji}, s_{vji}, m_{vji})$ where a_{vji}, s_{vji} and m_{vji} denote the appearance, spatial-temporal position and motion features. With each object instance vj associated with parameters $\theta_{vj} = (\eta_{vj}, \{\mu_{vj}^s, \Lambda_{vj}^s\}, \{\mu_{vj}^m, \Lambda_{vj}^m\})$ where η_{vj} is the parameter of multinomial function for appearance information and $\{\mu_{vj}^s, \Lambda_{vj}^s\}, \{\mu_{vj}^m, \Lambda_{vj}^m\}$ are Gaussian parameters for spatial-temporal distribution and the motion distribution, respectively. In addition, for every object class l it only contains parameter η_l related to appearance information.

Given the assignment t_{vji} appearance, spatial-temporal and motion features become independent and we get the likelihood of image patches given the object

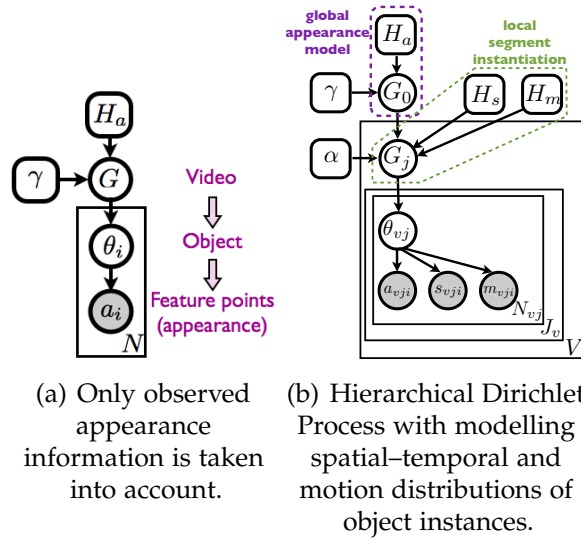


Figure 3.2: (a)The graphical model for “video-objects-feature points” generative procedure, only observed appearance information is taken into account. (b)Graphical model of our extended hierarchical Dirichlet Process. The prior of appearance distribution is on the highest layer to handle the appearance sharing among object instances belonging to the same object category. Meanwhile, the priors of spatial-temporal and motion distributions are located on the lower layer to flexibly model the local position preference and motion pattern of every object instance. J_v denotes the number of object instances in video v .

instance:

$$\begin{aligned}
 p(a_{vji}, s_{vji}, m_{vji} | t_{vji}) &= p(a_{vji}, s_{vji}, m_{vji} | \eta_{vj}, \mu_{vj}^s, \Lambda_{vj}^s, \mu_{vj}^m, \Lambda_{vj}^m) \\
 &= p(a_{vji} | \eta_{vj}) p(s_{vji} | \mu_{vj}^s, \Lambda_{vj}^s) p(m_{vji} | \mu_{vj}^m, \Lambda_{vj}^m)
 \end{aligned} \tag{3.3}$$

For the likelihood of object instance vj given the object class l with assignment k_{lvj} , we can simply accumulate all the appearance likelihood of image patches belonging to the object instance given the parameter of the object class.

Combining with conjugate Dirichlet Process prior H_a to multinomial function of appearance information and also the conjugate inverse-Wishart prior H_s, H_m to the Gaussian distribution of spatial-temporal and motion features, we can utilize the posterior sampling scheme where Gibbs sampling is used in our method to iteratively sample the assignment of t_{vji} and k_{lvj} .

3.4.3 Illustration on Synthetic Sequence

We give an example of a toy video sequence in Fig 3.3 in order to illustrate the properties of our approach. There are four moving objects in the video and two of them have the similar appearance (rectangle) but with different spatial-temporal and

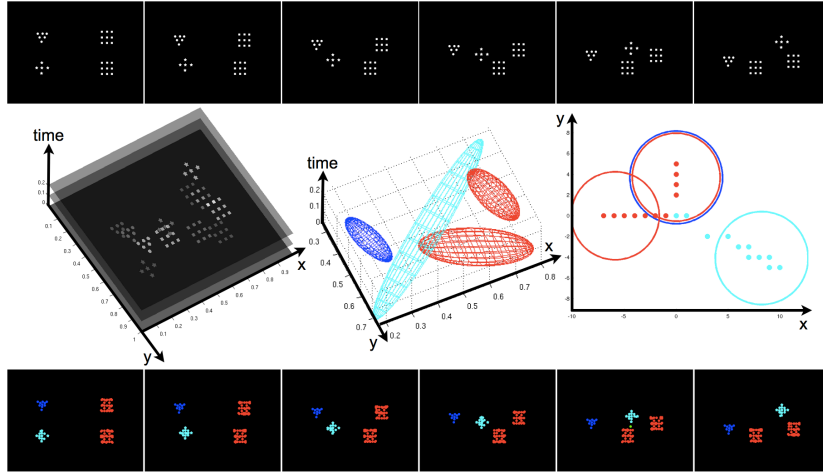


Figure 3.3: **First Row:** Toy video sequence with four moving objects. **Second Row, from left to right:** Illustration for (a)the video cube with inferred (b)spatial-temporal and (c)motion distributions of four moving objects inside. **Third Row:** Video sequence with labelled feature points. Note here different colors stand for different global classes.

motion distributions. Our method clusters the feature points into multiple groups of distinct motion and spatio-temporal distributions shown in the middle of the figure, as well as assigns these groups into different global appearance classes, which are shown in the figure with different color codes.

3.4.4 Implementation Details

Our implementation are based on modification from the non-parametric Bayesian toolbox published in (Teh, 2004). For the concentration parameters we simply follow the default setting in the toolbox with weakly informative priors $\alpha \sim \text{Gamma}(1, 1)$ and $\gamma \sim \text{Gamma}(1, 1)$. The hyperparameter of multinomial distribution for appearance information is assigned symmetric Dirichlet prior $H_a = \text{Dir}(W/10)$, where $W = 1000$ is the dictionary size; for the hyper-parameters for spatial-temporal and motion distribution we chose H_s and H_m to have 7 and 6 degrees of freedom for covariances while the means have non-informative priors. In our experiments the approach turns out to be not very sensitive to different settings of these parameters wherefore we kept them fixed.

3.4.5 Experimental Results of CRP-Based Video Co-Segmentation

We apply the proposed algorithm for inferring latent structure in videos to the task of video segmentation and video indexing. In particular, we demonstrate the benefits of our model that infers a global appearance model and local segment

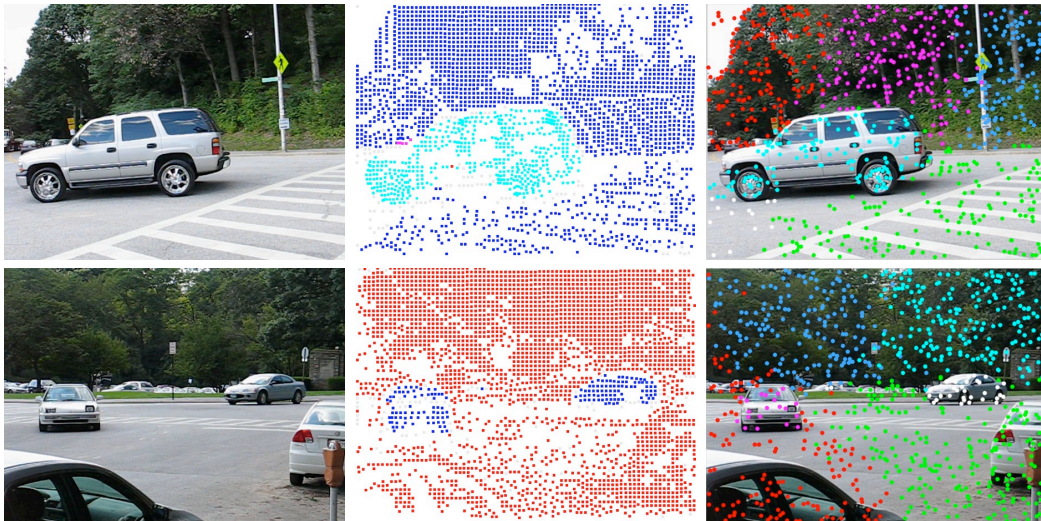


Figure 3.4: **First column:** Original frames. **Second column:** Object segmentation results from (Brox and Malik, 2010). **Third column:** Results from the proposed method. We label the feature points in different colors according their global appearance model assignment.

instantiations with different spatial-temporal and motion distributions jointly with the video segmentation task. Finally, we show a new task of video retrieval via sketches that is made possible by the exploiting the inferred latent structure of our model.

3.4.5.1 Video Segmentation

In our model we present the video sequence as a collection of image patches distributed over the whole video volume which every image patch contains the appearance feature, spatial-temporal position and also its motion vector. The image patches are sampled according to SIFT feature detector and their appearance information is encoded by SIFT descriptors (Lowe, 2004; Vedaldi and Fulkerson, 2008). These descriptors are then quantized into a codebook so that we can assign a codeword to every image patch. The spatial-temporal location is associated with the image coordinate of the patch center and the index of the video frame it belongs to. For the motion vector, we compute the optical flow vectors in each pair of adjacent frames based on the SIFT-Flow method (Sun *et al.*, 2008).

In order to evaluate the proposed algorithm on a video segmentation task, we use the benchmark of *moseg* dataset introduced in (Brox and Malik, 2010). The dataset provides 26 real world video sequence with 204 frames of annotated ground truth together with the evaluation tool. Here we use the measurements of the *density* for labelled feature points, the pixel-wise *overall error* and the region-wise *average error* in the ground truth.

Applying our algorithm on the video sequences, the feature points are grouped

	Density	overall error	average error
All available frames - 10 car sequences			
our method	0.29%	6.55%	17.83%
no motion	0.29%	35.03%	56.25%
Brox et al (Brox and Malik, 2010)	0.78%	4.08%	24.46%
All available frames - all 26 sequences			
our method	0.55%	11.71%	28.67%
no motion	0.55%	33.13%	54.51%
Brox et al (Brox and Malik, 2010)	3.31%	6.68%	27.7%

Table 3.1: Comparison of performances on different settings. Because of sparse feature extraction, only a small fraction of pixels are labelled therefore having small number in density. Note here the overall error is the number of bad labels over the total number of labels on pixel-wise basis, while the average error is on region-wise basis which depends on the annotation from ground truth.

into clusters which represent local segment instantiation moving within the sequence. Examples can be seen in third column in Fig 3.4. Although the feature points are sparsely distributed, we still can clearly tell that the objects of interest (cars) are labelled differently from the background. Table 3.1 provides a quantitative comparison between our model with two different settings and the results obtained from (Brox and Malik, 2010). The first configuration of the proposed method is to use all the appearance, spatial-temporal and motion information. And the second one is similar but without motion features. We use these two settings to address the importance of motion which helps to cluster the feature points across frames and also resolve the ambiguity coming from the appearance and spatial domain. A significant example can be seen from the second row of Fig 3.4. The white car parking aside and the car moving from right to left are with similar appearance and very close to each other. But the different motion pattern helps to distinguish both of them.

For more precise comparison, we first present the evaluation results only on the car sequences which have higher resolution and more significant moving patterns. Here our method shows improved average error by over 6% at slightly less density. The overall error is slightly worse – but still comparable. For the full dataset our average error is still comparable at 28.67% but with about 5% of degradation in accuracy for overall error. This is mainly because that in the ground truth some small objects are detailedly annotated, for example, phones, skinny chairs and people far away from camera. In those cases, our sparse sampling scheme of the SIFT representation can not get sufficient evidence. Overall, our method achieves competitive results with minimalistic modeling assumption and without any learning of parameters and/or post-processing stages. In particular, we don't use any separate tracking procedure in order to generate feature tracks.

In addition, since the property of globally appearance sharing in the proposed



Figure 3.5: Example of learnt global classes across videos in *moseg* dataset.

method, we learn global appearance classes across all the videos. The inference of global classes for all videos is very beneficial, for the reason that we can collect more evidence to have more stable appearance models which are able to go back to help the segmentation. Examples of the global classes learned from *moseg* dataset are shown in Fig 3.5. We observe clusters with cars, people and the background.

Moreover, since our model will link the objects with similar appearance to the same global class, we are able to handle occlusion case which our competitors cannot recover from as they rely on feature tracks. For instance in Fig 3.6, the lady in white is occlude but our inference procedure can still assign them to the same global appearance class as can be seen from the color coded feature points.

3.4.5.2 Sketch-Based Video Retrieval

After performing inference across all video sequences, the proposed method has extracted a compact representation which can be use to summarize the latent structure of videos. Examples as shown in Fig 3.7 present the object instances with different motion and spatial distributions. Also, since we learn the global appearance classes across video sequences, the segments are linked by the global appearance model. An illustration of the largest clusters found across all the video sequences is shown in Figure 3.5. Each thumbnail corresponds to a whole spatio-temporal segment in the video.

For further retrieval, we manually label the clusters according to the object class - which is easy to perform based on this condensed view on all videos. In contrast to previous work on video segmentation, we have now obtained a video summary that also carries semantic information which is propagated by means of the appearance



Figure 3.6: Illustration of the capability of our model to handle occlusion. Even the lady behind is occluded, our method can link its feature points over frames to the same global appearance model.

clusters.

We showcase our novel representation by performing a sketch based video retrieval. We provide manually generated sketches as shown in the leftmost column of Figure 3.8. They encode simple behaviour in a video sequence like “car moving from left to right”, “person moving from right to left” or “two cars moving in opposite directions”. Given such a query we can match the spatio-temporal distribution of the object classes involved to the video database. For every pair of segments we can compute their Kullback–Leibler–divergence according to spatial–temporal and motion distributions. For example, given two object instance j and k with parameters θ_j and θ_k , their corresponding distributions P and Q can be written as:

$$\begin{aligned} P &= p(s_{ji}, m_{jt} | \theta_j) = p(s_{ji} | \mu_j^s, \Lambda_j^s) p(m_{jt} | \mu_j^m, \Lambda_j^m) = P^s \cdot P^m \\ Q &= p(s_{ki}, m_{ki} | \theta_k) = p(s_{ki} | \mu_k^s, \Lambda_k^s) p(m_{ki} | \mu_k^m, \Lambda_k^m) = Q^s \cdot Q^m \end{aligned} \quad (3.4)$$

As both are given as gaussian, KL measure can be efficiently computed in closed form. The results are presented next to the query in Figure 3.8. We observe that we have indeed found relevant videos based on the sketched action of an object. By means of our latent representation, we have proposed a novel way of performing video retrieval via human sketches.

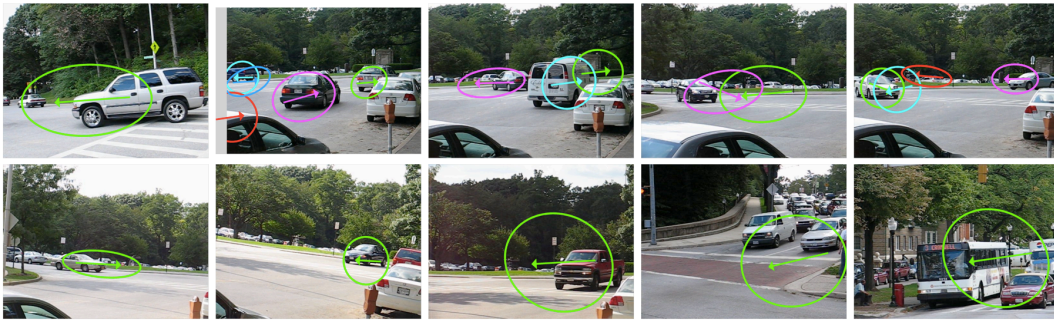


Figure 3.7: Summarization of 10 car sequences from *moseg* dataset. The spatial and motion distributions of every extracted object instance are drawn. Note here we marginalize out the temporal-axis of spatial-temporal distribution for better visualization.

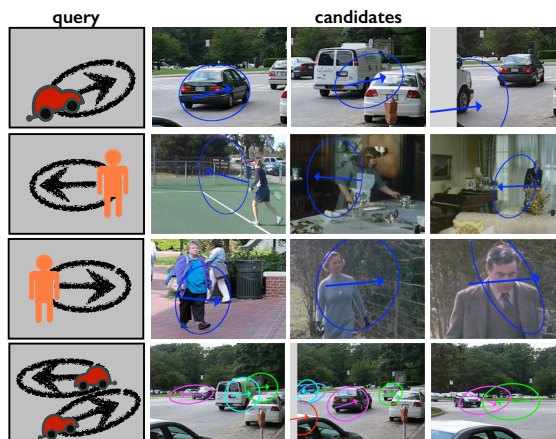


Figure 3.8: Example of sketch-based video retrieval in *moseg* dataset.

3.5 DDCRP-BASED VIDEO CO-SEGMENTATION

3.5.1 Video Representation

In comparison to using sift feature points as the basic data component in the previous sections, here we use superpixels as our data observations, since it can provide dense segmentations while have still less computational demand than pure image pixels. Therefore we describe our video representation as below: Given a set of videos \mathcal{V} , we start by a superpixel segmentation for each frame within the sequence and represent the video as a collection of superpixels. For every video $v \in \mathcal{V}$, we denote its total number of superpixels by N_v , and describe each superpixel i by its appearance feature x_i , spatio-temporal location s_i and motion vector m_i .

3.5.2 Distance Dependent Chinese Restaurant Processes (ddCRP)

For the formulation presented in the previous section, the DPM and CRP based posterior sampling is used to model the appearance, spatial-temporal and motion features of object instances as well as perform the inference. However, we often see in the realistic data that a single Gaussian distribution is not sufficient enough to model the overall motion or spatial-temporal features of a objects, especially for the irregular-shaped or articulated objects in which different parts of a object might have different motions. Therefore, we would like explore the dependencies inside of the structure of the data, for instance, the superpixels close in spatial-temporal positions and with similar motion directions should have a higher chance to come from the same object instance.

Based on same restaurant imagination of Chinese Restaurant Process as we describe in the Section 3.3, while there are dependencies between customers, the table assignment of a customer does not only depend on the number of people sitting on certain table, but also takes the distances/dependencies between customers into consideration. In results, the exchangeability property of CRP is not held any more, and a generalized process allowing non-exchangeable distribution over partitions is needed. The Distance Dependent Chinese Restaurant Processes (ddCRP) was proposed to offer an intuitive way for modelling non-exchangeability and dependency. The main difference between the CRP and ddCRP is that rather than directly linking customers to tables with table assignments, in ddCRP the customers sit down with other customers according to the dependencies between them, which leads to *customer assignments*. Groups of customers sit together at a table only implicitly if they can be connected by traversing the customer assignments. Therefore the i -th customer sits with customer j with a probability inversely proportional to the distance d_{ij} between them or sits alone with a probability proportional to the hyperparameter α :

$$p(c_i = j | D, f, \alpha) \propto \begin{cases} f(d_{ij}) & j \neq i \\ \alpha & j = i \end{cases} \quad (3.5)$$

where c_i is the customer assignment for customer i and $f(d)$ is the decay function and D denotes the set of all distances between customers. The decay function f should be non-increasing, takes non-negative finite values, and satisfies $f(\infty) = 0$. It describes how distances between customers affect the probability of linking them together.

3.5.3 ddCRP Video Segmentation Prior

We use the ddCRP in order to define a video segmentation prior. Customers correspond now to superpixels and tables correspond to object instances. The distance measure D and decay function f is now composed of two parts: $\{D^s, f^s\}$ and $\{D^m, f^m\}$ where the former one comes from the spatio-temporal distance and

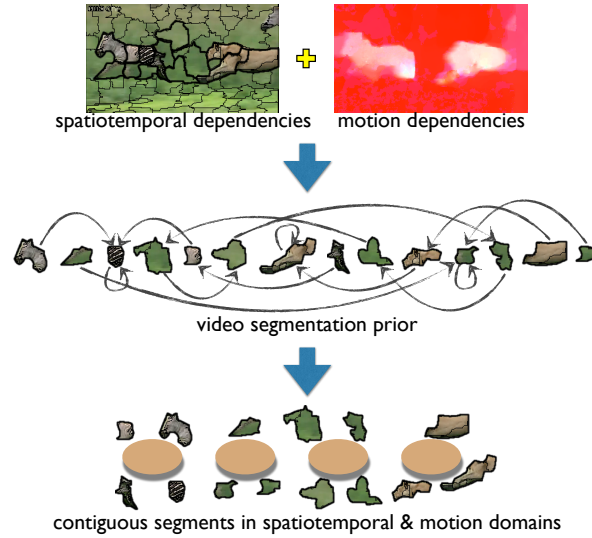


Figure 3.9: The visualization of the idea for the video segmentation prior proposed in our model. The dependencies from spatio-temporal and motion distances between superpixels are incorporated into the ddCRP prior which can control the probability of the customer links as described in Equation 3.6. To compute the connected components of the customer assignments of superpixels, they can produce the clusters which present the contiguous segments in spatio-temporal and motion domains.

the latter one from motion similarities between superpixels.

$$p(c_i = j | D, f, \alpha) \propto \begin{cases} f^s(d_{ij}^s) f^m(d_{ij}^m) & j \neq i \\ \alpha & j = i \end{cases} \quad (3.6)$$

Before measuring the spatio-temporal distance, we first use the optical flow vectors gained from TV-L1 model (Chambolle and Pock, 2011) in each pair of adjacent frames to find the neighbouring superpixels along temporal axis. Then the spatio-temporal distance D^s between superpixels is defined as the number of hops (Ghosh *et al.*, 2011) required to travel from one superpixel to another. For the motion distance D^m between superpixels, we use the euclidean distances between mean motion vectors of superpixels for the motion similarities. For f^s , we use the *window decay* $f(d) = [d < A]$ which determines the probabilities to link only with customers that are at most distance A away. For f^m , we use the *exponential decay* $f(d) = e^{-\frac{d}{B}}$ which decays the probability of linking to customers exponentially with the distance to the current one, where B is the parameter of decay width. With the decay functions f^s and f^m for both spatio-temporal and motion domains, we have defined a distribution over customer (superpixel) assignments which encourages to cluster nearby superpixels with similar motions together, and thus to have contiguous segments in spatio-temporal and motion domains. In Figure 3.10 we show samples from this ddCRP video segmentation prior for different hyperparameters and in

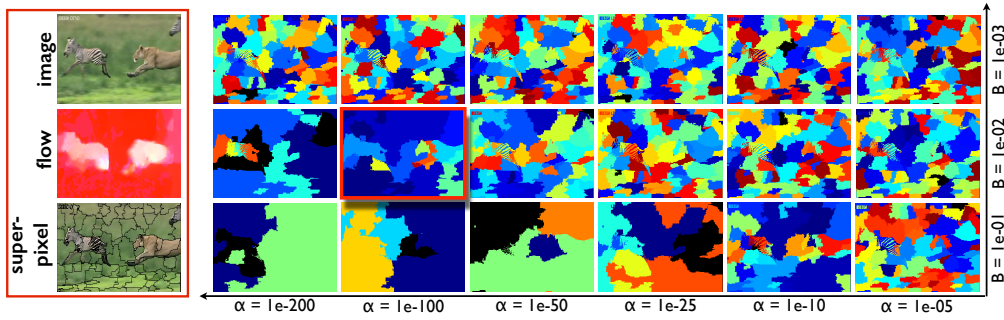


Figure 3.10: The rightmost column: (from top to bottom): original image, motion map from optical flow, superpixel segmentation. Rest columns: samples from ddCRP video cosegmentation prior under different settings between concentration hyperparameter α and width parameter B for exponential decay function of motion f^m .

Figure 3.9 we visualize the concept of the video segmentation prior and how the customer assignments can produce the clusters. The prior proposes segments having contiguous superpixels with similar motion.

3.5.4 Generative Multi-Video Model

In this section we formulate a probabilistic, generative model that links the videos by a global appearance model that is also non-parametric. We consider the following hierarchical generative procedure of multiple video sequences:

Videos consist of multiple global object classes with different appearances, and for every video there are arbitrary number of instances which are located at different locations and possibly move over time. As our model has a hierarchical structure of layers for global classes and local instances which is very similar to the idea of Hierarchical Dirichlet Process (Teh *et al.*, 2006), we use the same metaphor of its Chinese restaurant franchise representation in our case: There is a restaurant franchise (set of videos) with a shared menu of dishes (object classes) across all restaurants (videos). At each table (object instance) of each restaurant one dish (object class) is ordered from the menu by the first customer (superpixel) who sits there, and it is shared among all customers (superpixels) who sit at that table (object instance). Multiple tables (object instances) in multiple restaurants (videos) can serve the same dish (object class). So the analogy is the following: restaurants correspond to videos, dishes correspond to object classes, tables correspond to instances, and customers correspond to superpixels. The visualization of the Chinese restaurant franchise representation and its corresponding metaphors to the proposed multi-video modal can be seen in the Figure 3.11 for better understanding. Here is a summary of the generative process:

1. For each superpixel i_v in video v , draw assignment $c_{i_v} \sim \text{ddCRP}(D, f, \alpha)$ to the

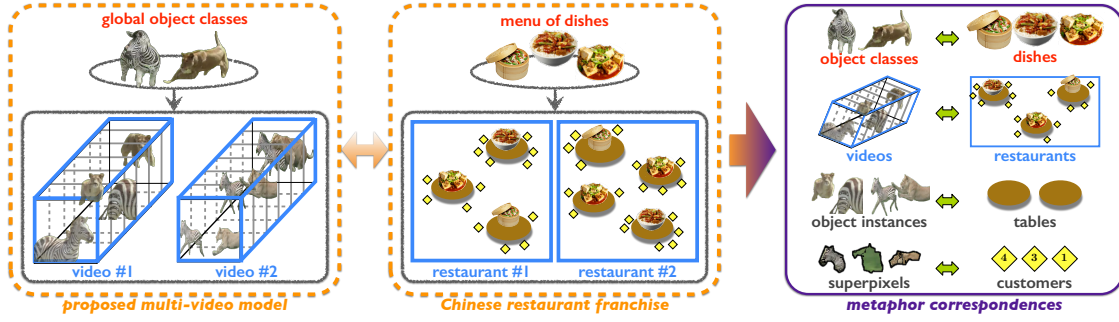


Figure 3.11: The metaphors between our proposed multi-video model with respect to the Hierarchical Dirichlet Process (HDP) (Teh *et al.*, 2006). First block: the visualization of the structure for the proposed model; Second block: the Chinese restaurant franchise representation of the HDP; Third block shows the correspondence between the metaphors of previous two blocks.

object instance

2. For each object instance t_v in video v , draw assignment $k_{t_v} \sim \text{CRP}(\gamma)$ to the object class
3. For each object class k , draw cluster parameters for appearance model $\phi_k \sim G_0$
4. For each superpixel i_v in video v , draw observed feature $x_{i_v} \sim P(\cdot | \phi_{z_{i_v}})$, where $z_{i_v} = k_{t_{i_v}}$ the class assignment for i_v .

where G_0 is drawn from the $\text{DirichletProcess}(\gamma, H_a)$ in order to define an infinite set of appearance models. H_a denote a Dirichlet prior on feature appearance distribution which is used as the base distribution for the process. γ is the concentration parameter for the Dirichlet process. For each global object class k discovered across video sequences, the parameter ϕ_k for its appearance model is sampled from G_0 . We use a multinomial distribution η to describe the appearance model. Therefore given the observed appearance feature x_i for superpixel i , the likelihood of observed appearance feature for global object class k can be denoted as $p(x_i | \phi_k) = \eta_k(x_i)$.

3.5.5 Posterior Inference via Gibbs Sampling

In order to incorporate the ddCRP video segmentation prior with the likelihood of superpixels to object instances whose appearance models are inherited from corresponding global object classes, we can now define a posterior distribution over customer assignments and use it to perform inference.

The goal of posterior inference is to compute posterior distribution for latent

variables given observed data. The posterior for customer assignments $c_{1:N_v}$ is:

$$p(c_{1:N_v} | x_{1:N_v}, D, f, \alpha, \gamma) = \frac{\left(\prod_{i_v=1}^{N_v} p(c_{i_v} | D, f, \alpha) \right) p(x_{1:N_v} | z(c_{1:N_v}), \gamma)}{\sum_{c_{1:N_v}} \left(\prod_{i_v=1}^{N_v} p(c_{i_v} | D, f, \alpha) \right) p(x_{1:N_v} | z(c_{1:N_v}), \gamma)} \quad (3.7)$$

where z are the class assignments for all the tables introduced by the customer assignments.

Here we use ddCRP $p(x_{1:N_v} | z(c_{1:N_v}))$ as prior for all the possible customer configurations such that its combinatorial property makes the posterior intractable wherefore we use sampling techniques. As proposed in original ddCRP paper (Blei and Frazier, 2010), Gibbs sampling is used where samples are iteratively drawn from the conditional distribution of each latent variable given the other latent variables and observations:

$$p(c_{i_v} | c_{-i_v}, x_{1:N_v}, D, f, \alpha, \gamma) \propto p(c_{i_v} | \alpha, D, f) \cdot p(x_{1:N_v} | z(c_{1:N_v}), \gamma) \quad (3.8)$$

The prior term is given in equation 3.6 and the likelihood term for multinomial appearance distribution is

$$\begin{aligned} p(x_{1:N_v} | z(c_{1:N_v}), \gamma) &= \prod_{l=1}^{|z(c_{1:N_v})|} p(x_{z(c_{1:N_v})=l} | z(c_{1:N_v}), \gamma) \\ &= \prod_{l=1}^{|z(c_{1:N_v})|} \eta_l(x_{z(c_{1:N_v})=l}) \end{aligned} \quad (3.9)$$

Resampling the global class (dish) assignment k follows typical Gibbs sampling method for Chinese Restaurant Process but consider all the features $x_{\mathcal{V}}$ and assignments $k^{\mathcal{V}}$ in the video set \mathcal{V} . The class assignment posterior of each table t_v in video v is:

$$p(k_{t_v} = l | k_{-t_v}^{\mathcal{V}}, x^{\mathcal{V}}, \gamma) \propto \begin{cases} m_l^{k_{-t_v}^{\mathcal{V}}} \eta_l^{k_{-t_v}^{\mathcal{V}}}(x_{t_v}) & \text{if } l \text{ is used} \\ \gamma \eta_l(x_{t_v}) & \text{if } l \text{ is new} \end{cases} \quad (3.10)$$

Here $k_{-t_v}^{\mathcal{V}}$ denotes the class assignments for all the tables in the video set \mathcal{V} excluding table t_v , $x^{\mathcal{V}}$ is the appearance features of all superpixels within \mathcal{V} . Given the class assignment setting $k_{-t_v}^{\mathcal{V}}$, $m_l^{k_{-t_v}^{\mathcal{V}}}$ counts the number of tables linked to global class l whose appearance model is $\eta_l^{k_{-t_v}^{\mathcal{V}}}$. x_{t_v} stands for the appearance features of superpixels assigned to the table t_v .

3.5.6 Implementation Details

For computing the appearance feature representation for superpixels, we use the following pipeline: We use the same procedure of dense patch extraction and patch

description as in (Joulin *et al.*, 2012) in order to stay comparable to the image co-segmentation baseline which we will use in the experimental section. These patches are further quantized into a codebook of length 64 so that we can assign a color codeword to every image patch, which is based on a typical Bag-of-Words (BoW) image representation. Now we describe the appearance feature for each superpixel i by using the color codeword histogram x_i computed from the image patches whose centres are located inside that superpixel.

For all our experiments we set the concentration parameter $\gamma = 1$ which is weakly informative. The hyperparameter on multinomial distribution for appearance information is assigned symmetric Dirichlet prior $H_a = \text{Dir}(2e + 2)$ which supports the global classes to contain various codewords. The concentration parameter $\alpha = 1e - 100$ for the proposed video segmentation prior and the width parameter $B = 1e - 1$ for motion decay function f^m are determined by inspecting samples from the prior obtained from equation 3.6. We show examples in Figure 3.10 that displays the effect of the parameters. We set width parameter A for spatial decay function f^s to be 3 for all our experiments.

3.5.7 Experimental Results of ddCRP-Based Video Co-Segmentation

For evaluating our generative video co-segmentation approach, we compare it with various baselines including image co-segmentation, video segmentation and also variants of the proposed model. In addition, we test the proposed method with different settings including different granularities of the superpixels as well as different priors for the object instance layer.

We first present our new dataset and the proposed evaluation criterion. Then we present the results of our method with comparisons to various baselines and a discussion.

3.5.7.1 Dataset

We present the first Multi-Object Video Co-Segmentation (MOVICS) challenge that can be used to provide quantitative comparisons for the multi-class video co-segmentation task. It is based on realistic videos collected from the web (youtube) and exposes several challenges encountered in online or consumer videos.

The first approach to a video co-segmentation benchmark is from the work of (Rubio *et al.*, 2012). The associated dataset is limited as it only consists of one set of 4 videos that are synthetically generated. The same foreground video is pasted into 4 different backgrounds. Accordingly, their task is defined as binary foreground/background segmentation that does not address segmentation of multiple classes and how the segments are linked across videos by the classes.

In contrast to this video co-segmentation approaches which phrases the task as binary foreground/background segmentation problem, we rather propose to tackle the more general problem of multi-class labelling. This change in task is crucial in order to make progress towards more unconstrained video settings as we encounter

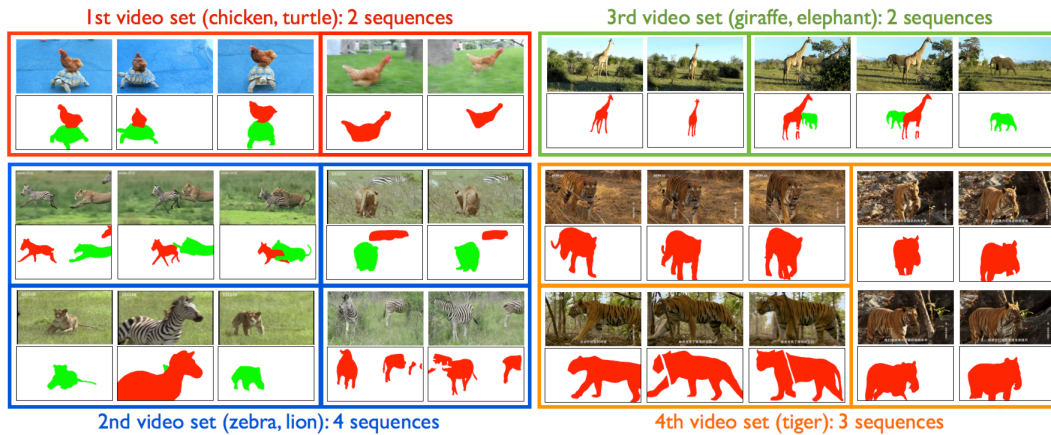


Figure 3.12: Summary of our proposed MOVICS dataset. Different color blocks stand for different video sets and the images within the same block come from the same video sequences.

them in online resources and consumer media collections. Therefore, we propose a new multi-class video co-segmentation task of realistic videos with multiple objects in the scene. This makes a significantly more difficult problem, as not only object have to be correctly segmented but also assigned the same global class across video.

We propose the first benchmark for this task based on realistic video sequences download from Youtube. The dataset has 4 different video sets including 11 videos with 514 frames in total, and we equidistantly sample 5 frames from each video for which we provide ground truth for the object classes of interest. Note that for each video set there are different numbers of common object classes appearing in each video sequence, and all the objects belonging to the same object class will be noted by the same label.

Unlike the image co-segmentation dataset *iCoseg* (Batra *et al.*, 2010) which has similar lighting, image conditions and background or video segmentation dataset *moseg* (Brox and Malik, 2010) with significant motion patterns, our dataset exposes many of the difficulties encountered when processing less constraint sources. In Figure 3.12 we show examples of video frames for the four video sets together with the provided groundtruth annotations. Our sequences show different lighting conditions (e.g. tiger seq.), motion blur (e.g. chicken seq.), varying number of objects moving in and out (e.g. giraffe, elephant seq.), similar appearance between objects and background (e.g. tiger), etc. The MOVICS dataset and our code can be found at <http://www.d2.mpi-inf.mpg.de/datasets>.

3.5.7.2 Follow-Up Datasets

Based on our initial publication (Chiu and Fritz, 2013), there are several following-up research works on this new video object co-segmentation problem to push forward the performance based on our MOVICS challenge and more datasets collections

with different properties. For instance, Zhang et al (Zhang *et al.*, 2014a) reuse some videos from MOVICS and add new ones to build up a new dataset name *Safari* where for each object class there will be one video only contains this class inside, and other videos are with two classes. In (Fu *et al.*, 2014) the authors propose new datasets aiming for multiple foreground video co-segmentation setting. Also in (Wang *et al.*, 2014b) a video object co-segmentation benchmark is collected which emphasizes there will be irrelevant frames in some videos. All these datasets serve different purposes for evaluating the various flexibilities of the video co-segmentation approaches, such as the handling multiple object classes in a video set, discovering the number of object instances, and relaxing the assumption that target objects appear in all frames from all videos. We expect that in the future there will be a new dataset at larger scale which can provide an unified benchmark with different difficulties for this growing research area of video co-segmentation problem.

3.5.7.3 Evaluation Metric

In order to quantify our results, we adopt the *intersection-over-union metric* that is also used in image co-segmentation tasks (e.g. (Kim and Xing, 2012)) as well as the PASCAL challenge.

$$M(S, G) = \frac{S \cap G}{S \cup G} \quad (3.11)$$

where S is a set of segments and G are the groundtruth annotations.

For the problem of video co-segmentation, we aim to not only simultaneously achieve video segmentation for multiple sequences, but also to learn the common object classes across videos. Therefore we define our co-segmentation task as finding for each object class a set of segments that coincide with the object instances in the video frames such that the algorithm has to group the segments by object class. We denote all segments grouped to an object class i by S_i . Accordingly our evaluation assigns the object class to the best matching set of segments predicted by an algorithm:

$$\text{Score}_j = \max_i M(S_i, G_j) \quad (3.12)$$

Please note that this measure is not prone to over-segmentation, as only a single label is associated with each object class for the whole set of videos. We can further condense this performance measure into a single number by averaging over the classes.

$$\text{Score} = \frac{1}{C} \sum_j \text{Score}_j \quad (3.13)$$

where C is the number of object classes in the groundtruth.

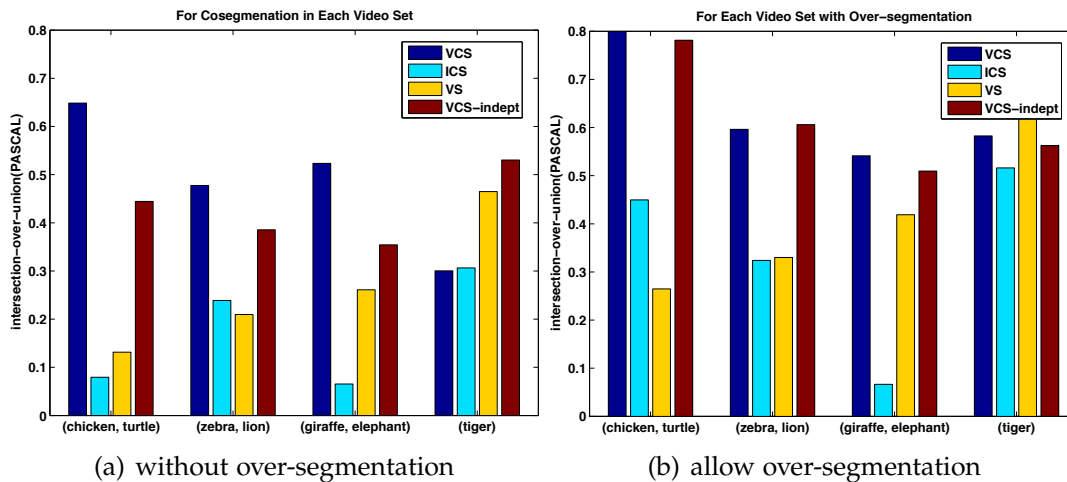


Figure 3.13: Comparison of co-segmentation accuracies between the our method (VCS), image co-segmentation (ICS), video segmentation (VS) and our method with applying on each video independently (VCS-indept) in the proposed MOVICS dataset. (a) Only a single label is assigned per object class in the groundtruth for the whole set of videos. (b) Allow over-segmentation which can assign multiple labels to the same object class in the groundtruth

3.5.7.4 Comparison to Video Segmentation

A comparison to video segmentation methods is not straightforward. As each video is processed independently, there is no linking of segments across the videos. We therefore give the advantage to the video segmentation method that our evaluation links the segments across videos by the groundtruth. All baselines using video segmentation are therefore over optimistic results as they are using a ground truth oracle in order to perform the linking across videos.

3.5.7.5 Results

We evaluate our approach on the MOVICS dataset and compare it to two state-of-the-art baselines from video segmentation and image co-segmentation. The video segmentation baseline (Brox and Malik, 2010; Ochs and Brox, 2011) is denoted by (VS) and the image co-segmentation baseline (Joulin *et al.*, 2012) is denoted by (ICS). (VCS) stands for our proposed multi-class video co-segmentation method. For both baselines we run their publicly available codes on our data.

The performance numbers of the proposed method in comparison to the baselines are shown in Figure 3.13(a). With an overall performance of 48.74% of our method, we outperform VS by 22.07% and ICS by 31.49%.

Figure 3.14 shows a visualization of the results. For each video set the rows sequentially show the video frames, optical flow maps, groundtruth annotations, results of image co-segmentation baseline (ICS), results of video segmentation

approach (VS) and the last row shows the results of the proposed method (VCS).

Here the evaluation is performed per set of video sequences since the algorithm not only have to correctly segment the object instances but also link them to a consistent object class. In theory, for the ideal case there should be exact one-to-one correspondence between object classes from the algorithm and ground truth. Therefore in this experimental setting we restrict to only use one discovered object class for describing each class in ground truth, as described in our evaluation metric. In another word, we do not allow for over-segmentation of the object classes in this experiment.

Also recall that VS does not link objects across videos. Therefore it has no notion of objects linked across videos. As described in section 3.5.7.4 we give an advantage to the VS method by linking the segments across video via the groundtruth. Despite this advantage our method outperforms VS by a large margin for the first 3 video sets. Only on the (*tiger*) sequences VS performs better. It turns out that in this set the appearance is particularly hard to match across videos due to lighting and shadow effects, such that in one sequence of the (*tiger*) video set our VCS method uses another object class to label the tiger objects, see Figure 3.14. The VS gets boosted by the additional information from the groundtruth, in comparison, VCS is purely unsupervised and does not use the groundtruth in any way.

3.5.7.6 Discussion

The video segmentation baseline strongly depends on motion information in order to produce a good segmentation. When the motion map is noisy or there are objects moving together or with similar motion, segmentation errors occur. This issues are particular pronounced in the first video set where the chicken moves together with the turtle and the motion map is noisy due to fast motion in the second video. Our method handles such situations better and maintains a good segmentation despite the noisy motion information.

The image co-segmentation baseline has an assumption which expects a certain number of common object classes for all input images. This often cause problems for the less constraint settings that we are of interest in our study. For example in the second and third video sets in Figure 3.14, there are a varying number of objects moving in and out. The performance of image co-segmentation reduces in these settings. In addition, problems occur with wrongly merged object classes (lion with zebra, and giraffe with elephant). Our non-parametric approach seems to be better suited to deal with this variation on object instances and object classes and shows overall a more consistent segmentation.

Another interesting aspect of our model is how segmentation is supported by jointly considering all the videos of a set and learning a global object class model. Without this global appearance model, the performance decreases by 5.88% - still outperforming the baselines, please see the method noted as (**VCS-indept**) in Figure 3.13(a). Note here for evaluating VCS-indept, the same story of using the groundtruth to link the object classes across videos as for VS is applied. We give

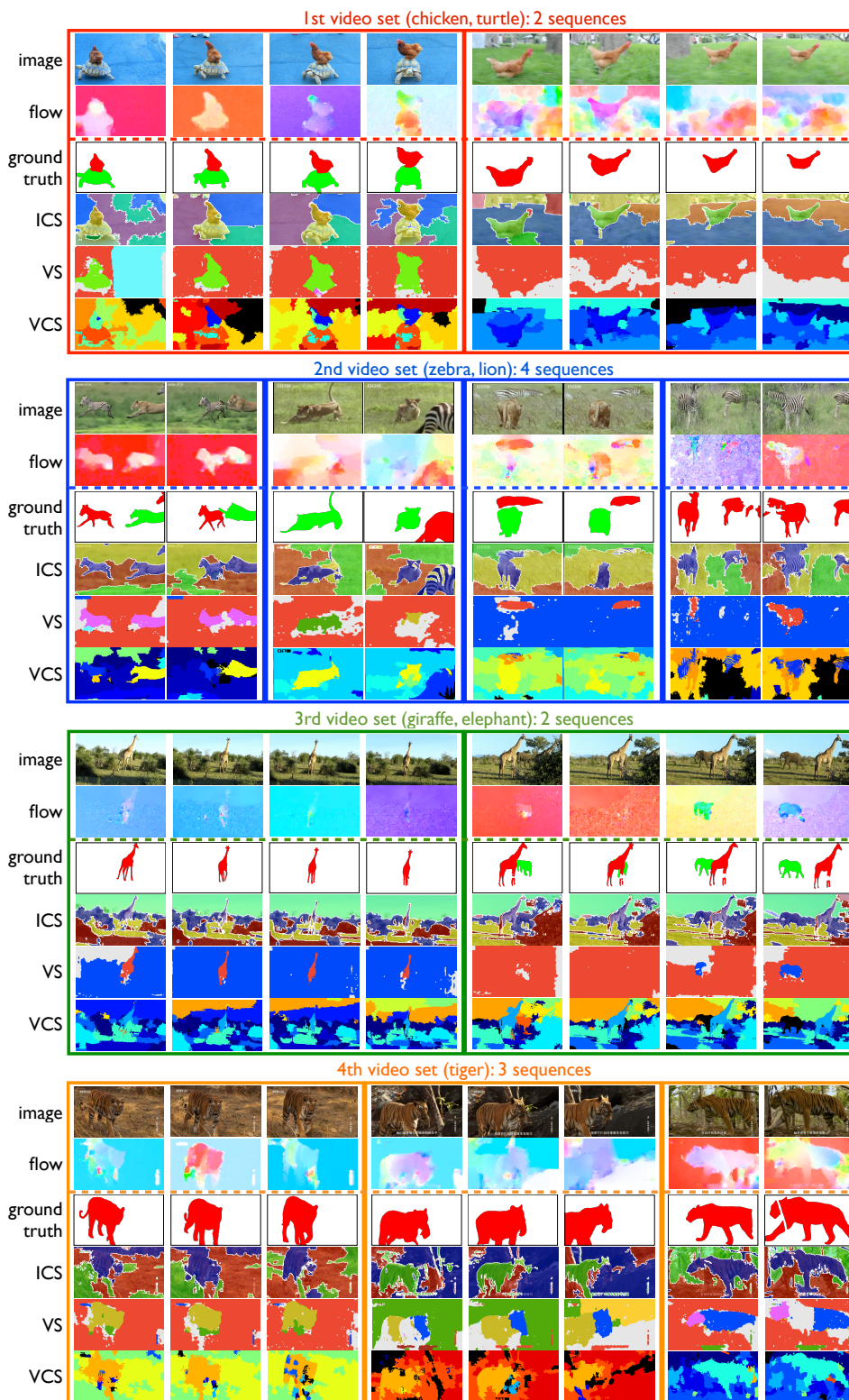


Figure 3.14: Examples of results from the proposed method (VCS), image co-segmentation (ICS) and video segmentation baselines (VS) in MOVICS dataset. For each video set the rows sequentially show the video frames, optical flow maps, groundtruth annotations, results of ICS, results of VS and the last row shows the results of our VCS approach.

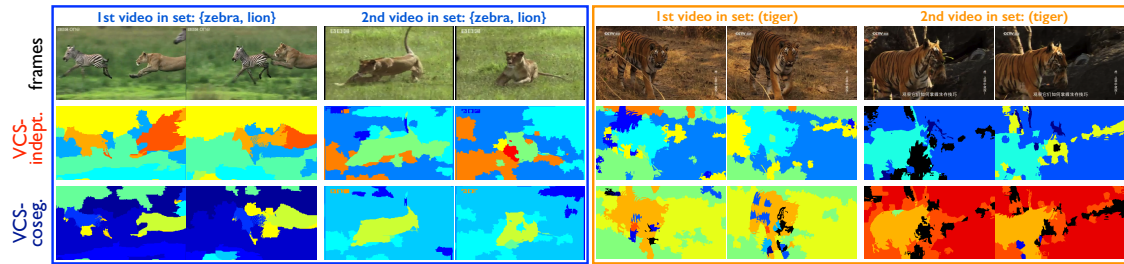


Figure 3.15: Example of improved results by segmenting across videos with a global object class model. First row: frames from video sequences. Second row: results obtained from running the proposed method independently for each sequence. Third row: results from joint segmentation on all videos of the video set.

an example in Figure 3.15 where the first row is the images from video sequences, the second row is the results by applying our proposed method only on each single sequence independently, and the last row is our VCS result while taking all videos in the video set into account. We observe the improved segmentation from the joint segmentation since the global appearance models get richer from various sequences in the set. This observation supports again the importance and potential of the co-segmentation approach since it not only helps to identify the shared object classes across videos but also improves the segmentation on the single video.

3.5.7.7 Analysis with Over-Segmentation

In this analysis we relax the assumption that the sets of segments proposed by the method have to correspond to exactly one groundtruth object class each. Therefore, we now assign multiple set of segments to the same object class in the groundtruth. In Figure 3.13(b) we present the performance comparison under this relaxed setting. Please note that this relaxed measure does not penalize for non-existing links between the videos as well as over segmentation in the spatial domain.

Overall, the performance improves, as over segmentation is not penalized. In average our method achieves a performance of 62.99% which still outperforms VS by 21.71% and ICS by 29.08%. The improvements under this measure are particular prominent on the video sets where appearance is hard to match across sequences. We take the fourth video set (tiger) as an example. In Figure 3.14 we observe that VCS over-segments the tiger. This set of videos is challenging due to varying lighting conditions, shadows and appearance similarities with the background. Both ICS and VS do not match the object correctly across videos, as we can tell by the different coloring across videos. Our method does not show strong over-segmentation artifacts and also matches the object class across the first two videos.

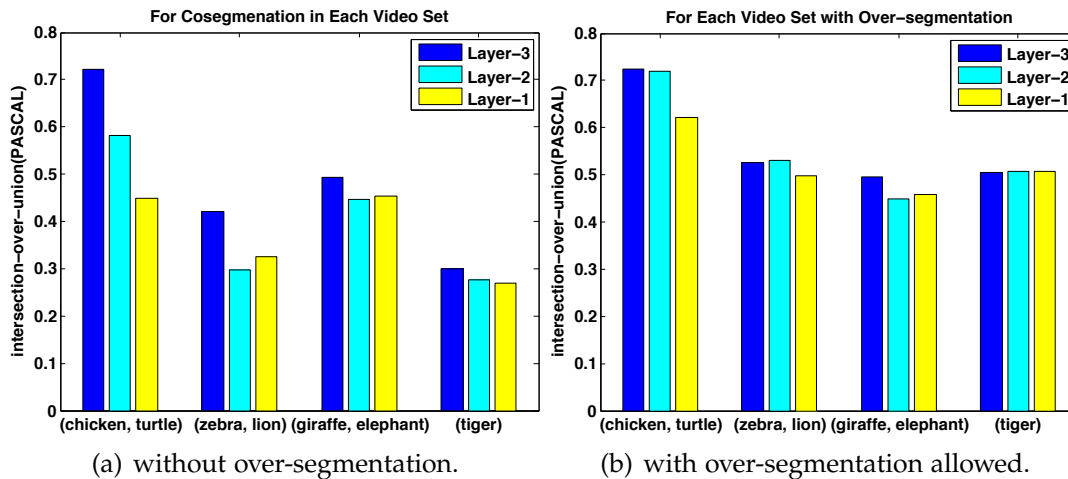


Figure 3.16: Comparison of co-segmentation accuracies between different granularities of the superpixel segmentations obtained by (Arbelaez *et al.*, 2011).

3.5.7.8 Analysis on Different Granularities of Superpixels

As discussed in Section 3.5.5, the posterior probability of customer assignments will depend on both the ddCRP video segmentation prior and also the appearance likelihood. In the ddCRP video segmentation prior the distance function will be affected by the granularity of superpixels, for instance the same threshold value for the window decay function of the spatiotemporal distance will represent different scales in pixel distance between the coarse and fine superpixel segmentations. Also in the appearance likelihood different granularities of the superpixels will contain different levels of statistics in the multinomial distributions for the BoW representations.

To do the analysis on the influences caused by the size of the superpixel segmentation, we use the image segmentation approach from (Arbelaez *et al.*, 2011) to produce the hierarchical superpixel segmentations of 3 coarse-to-fine layers by using the same parameter settings as in the video segmentation baseline (Ochs and Brox, 2011). Each layer of superpixels is utilized as the basic data points in our video co-segmentation model and we perform the inference with the same parameters. From the experiments shown in Figure 3.16 for each of 3 layers we observe that the higher layer with coarser superpixels gives better performance than finer ones under the evaluation scheme which does not allow over-segmentation, whereas different layers perform similarly when over-segmentation is feasible in evaluation. This trend can also be visualized in Figure 3.17.

The experiments here do not conclude that we should always use extremely big superpixels in our scheme because they might also cause the problem of under-segmentation to mix the foreground objects with the background that the algorithm has no way to recover. Instead, this analysis can be associated to the motivation of using higher-order terms in the recent works on graph-based video segmentation (Galasso *et al.*, 2014; Khoreva *et al.*, 2014). In graph-based video segmentation

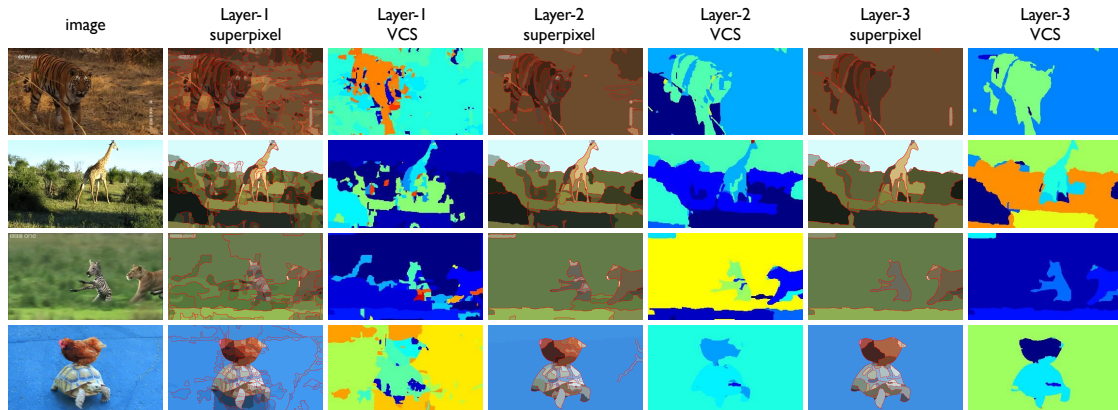


Figure 3.17: Examples of results for running the proposed method on different granularities of superpixel segmentations. The first column are the video frames, the rest ones sequentially shows the superpixels from fine to coarse layers of hierarchical image segmentation (Arbelaez *et al.*, 2011) and their corresponding video co-segmentation results.

approaches the affinity matrix between the nodes (finer superpixels) is built by measuring the distances/similarities from the local neighbors and the coarser superpixels in the hierarchical segmentation can provide more global and higher-order information with linking the nodes in wider spatiotemporal range. Since in our ddCRP video segmentation prior it also depends on the distance functions to bias the customer assignments between superpixels, the way of incorporating higher-order information in graph-based video segmentation methods can provide an feasible extension for our ddCRP prior to leverage. Additionally, in recent works of video co-segmentation such as (Zhang *et al.*, 2014a; Lou and Gevers, 2014; Fu *et al.*, 2014) the object proposals (Endres and Hoiem, 2010) are extracted as candidates of object segments in the pre-processing step. This step can also straightforwardly provide a higher-order information for us to utilize as a distance function in the ddCRP prior where the superpixels located within the same object proposals will have higher probabilities to be clustered together.

3.5.7.9 Variants for Video Segmentation Prior

One of the main contribution of this chapter is the video segmentation prior which incorporate the distance-dependent Chinese Restaurant process idea with the distance functions from both spatiotemporal and motion distances. We analyze several design choices, including: i) combination of both spatiotemporal f^s and motion f^m information as our proposed model, to use ii) only spatiotemporal f^s or iii) motion f^m distances, and iiii) to base on pure hierarchical Dirichlet process (HDP, layers of CRP) (Teh *et al.*, 2006) model which utilizes the Chinese Restaurant process as a prior without taking dependencies between data points into consideration.

We evaluate these design choices of the prior for the local object instance layer in

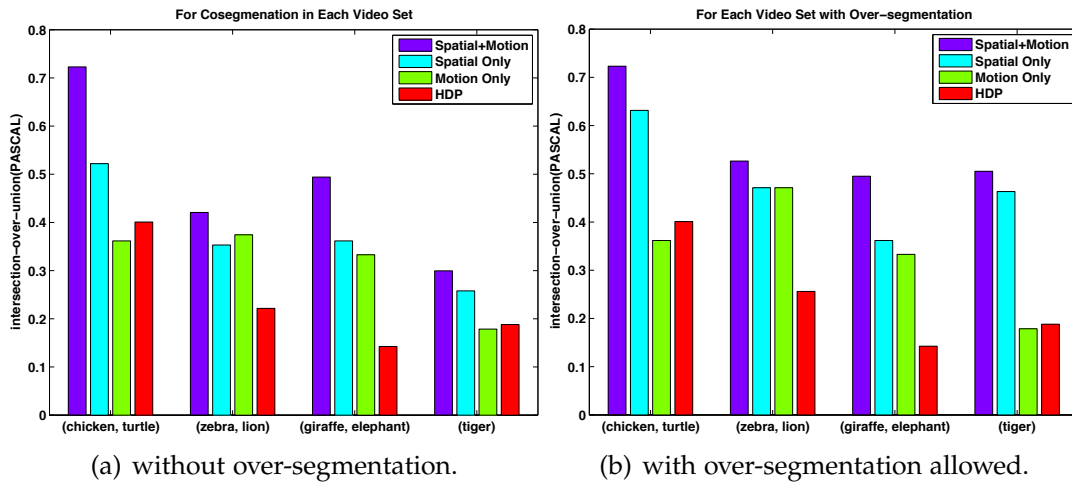


Figure 3.18: Comparison of proposed models under different design choices of the ddCRP prior for the local object-instance layer. There are 4 different choices are evaluated: i) to use both distances in spatiotemporal and motion domains (**Spatial+Motion**), to use only ii) spatiotemporal (**Spatial Only**) or iii) motion (**Motion Only**) informations, and iiiii) to use pure CRP instead of ddCRP as prior (**HDP**).

our model to have the quantitative comparison between them. We use the coarse superpixels obtained from the hierarchical image segmentation as in section 3.5.7.8 to be the basic data points in this experiment. For fair comparison, all the parameters will be kept the same across different designs of priors. Note that for implementation of HDP model we simply set the probabilities of customer assignments from a superpixel to all the others are with the same value, which will make the ddCRP degenerate to CRP.

The experimental results and the visualizations are shown in Figure 3.18 and 3.19. It is clear to see that the choice of ddCRP prior in the local object-instance layer to include both spatiotemporal and motion distance functions in large margins outperforms ii) spatiotemporal-only by 11.07%, iii) motion-only by 17.25% and iiiii) HDP by 24.62% (under the evaluation criterion of not allowing over-segmentation), and it maintains segments to be contiguous in spatial and motion aspect. And with the help from the global appearance models shared across videos, the overall hierarchical model can have better tolerance towards the ambiguities caused by accidental similarities in motion patterns or appearance within a single sequence.

3.5.7.10 Runtime Comparison

We provide the running time for our VCS algorithm and the baselines (ICS and VS) in Table 3.2. In order to have detailed comparison, we further divide the runtime of each method into different parts.

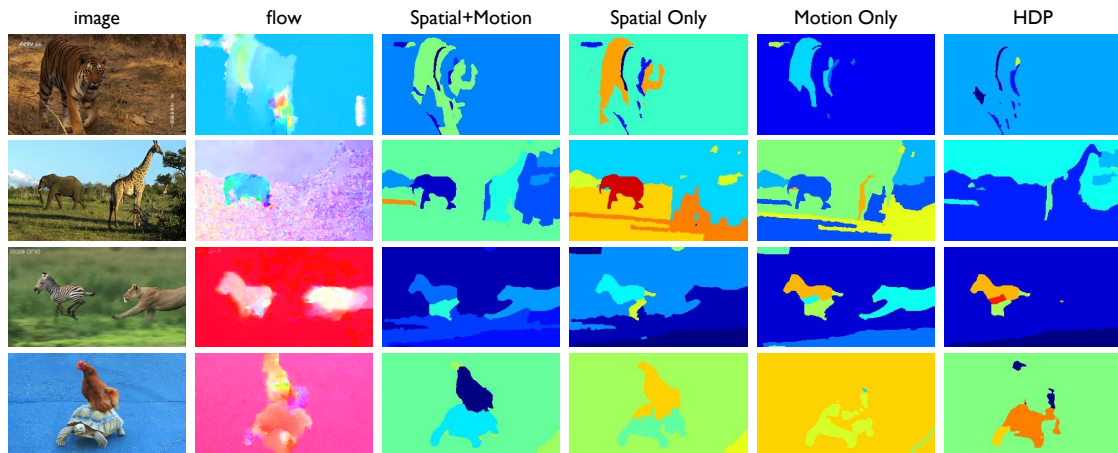


Figure 3.19: Examples of results from using various design choices of the video segmentation ddCRP prior. For the columns from left to right they sequentially show the video frames, optical flow maps, and the co-segmentation results obtained by using i) spatiotemporal+motion distances, ii) spatiotemporal information only, iii) motion distances only and iii) CRP as the prior in local layer of the proposed method.

Video Co-Segmentation (VCS, ours)				
video set	optical flow (Chambolle and Pock, 2011)	superpixels & features	inference	total
1st	1 h 32 m	8 m	1 h 12 m	2 h 52 m
2nd	2 h 59 m	20 m	3 h 13 m	6 h 32 m
3rd	1 h 27 m	6 m	1 h 22 m	2 h 55 m
4th	1 h 7 m	4 m	59 m	2 h 10 m
Image Co-Segmentation (ICS (Joulin <i>et al.</i> , 2012))				
video set	total			
1st	4 h 41 m			
2nd	8 h 38 m			
3rd	3 h 18 m			
4th	6 h 58 m			
Video Segmentation (VS (Brox and Malik, 2010; Ochs and Brox, 2011))				
video set	sparse (Brox and Malik, 2010)	dense (Ochs and Brox, 2011)	total	
1st	30 m	10 h 1 m	10 h 31 m	
2nd	49 m	18 h 17 m	19 h 6 m	
3rd	15 m	6 h 35 m	6 h 50 m	
4th	16 m	7 h 53 m	8 h 9 m	

Table 3.2: Runtime comparison for the proposed method and baselines.

Video Co-Segmentation

Optical Flow: We use TV-L₁ (Chambolle and Pock, 2011) Matlab package¹ for computing dense duality based optical flow between every pair of adjacent frames within the video sequences.

Superpixels and Features: This stage includes the dense image patch extraction, codebook learning, superpixel segmentation by QuickShift (Vedaldi and Soatto, 2008) implementation from VLFEAT toolbox (Vedaldi and Fulkerson, 2008), and the distance computation between superpixels.

Actual Segmentation by Gibbs Sampling: In this part the posterior inference is performed, we run the Gibbs sampling to resample the local object instance assignments of superpixels and the global object class assignments of instances.

Image Co-Segmentation The Matlab package² provided in (Joulin *et al.*, 2012) is used here for evaluating the image cosegmentation baseline.

Video Segmentation

Sparse: We use the executable³ of (Brox and Malik, 2010) to create sparse set of labels.

Dense: Similarly, the program of (Ochs and Brox, 2011) is used to turn the sparse segmentation into dense regions. Table 3.2 shows that our proposed method is faster than both of the considered baselines.

3.6 CONCLUSION

In this chapter we have introduced a non-parametric Bayesian framework of discovering the latent representation of video sequences and also the task of multi-class video co-segmentation. The hierarchical structure of our proposed method utilizes the global appearance model in the global layer to allow shared appearance of object instances as well as discover their object classes across videos, and in the local layer we experiment two variants to add the spatial-temporal and motion information. With explicitly modelling the motion and spatio-temporal distributions in the first variant and defining a probabilistic video segmentation prior based on ddCRP framework for the second one, our approach allows the discovered objects to have distinct motion patterns in the video volume and proposes spatially contiguous segments of similar motion. Therefore, video segments are found and related across to each other across videos.

The value of our algorithm is first demonstrated in applications to video segmentation and video indexing. In particular, we show how our models discover and group objects despite occlusions and highlight the compact, latent representation which is derived from the video sequences that can be interpreted as summarization. We demonstrate a new way to retrieve videos by providing a sketch of the video that can be related to our latent representation. Furthermore, we propose the Multi-Object Video Co-Segmentation (MOVICS) dataset specialized to the multi-class video

¹<http://www.gpu4vision.org/>

²<http://www.di.ens.fr/~joulin/>

³<http://lmb.informatik.uni-freiburg.de/>

co-segmentation problem, which is based on realistic videos and exposes challenges encountered in consumer or online video collections. The proposed approach is demonstrated to resolve the ambiguities of appearance and motion patterns, improve the segmentation results via joint segmentation, and outperform state-of-the-art image co-segmentation and video segmentation baselines. Our method is also shown to be faster in computation which enhances its applicability in real-world applications.

Contents

4.1	Introduction	45
4.2	Previous Works on Activity Discovery	46
4.3	Joint Segmentation and Discovery Approach	47
4.4	Discovery Framework	50
4.4.1	Context Word Extraction	50
4.4.2	Segmenting Context Words into Supersamples	50
4.4.3	Segmentation Priors for Activity Discovery	50
4.4.4	Joint Segmentation and Activity Discovery (ddCRP+CRP)	52
4.5	Evaluation Methodology	52
4.5.1	Dataset	53
4.5.2	Framework Implementation	54
4.5.3	Performance Estimation	54
4.6	Results	55
4.6.1	Semantic Relationships within and between Activities	55
4.6.2	Activity Discovery from Context Word Labels	55
4.6.3	Sensitivity to Context Word Noise	57
4.6.4	Activity Discovery from Sensor Data	57
4.7	Discussion	58
4.8	Conclusion	60

IN the chapter 3 we have addressed the hierarchical non-parametric Bayesian approach for the video co-segmentation task. We now generalize our distance-dependent Chinese Restaurant Process (ddCRP) based formulation to the application of inferring activity routines from sensor streams, which are composed of context words produced by multiple body worn and ambient sensors detecting object usages and mode of locomotion. Our approach does not require labelled data at any stage. Neither does our approach depend on time-invariant sliding windows to sample context word statistics. Context word streams are first segmented into supersamples, which are basic data units for the topic model, and then semantic and temporal features are obtained to construct a segmentation prior that relates supersamples via its context words. Our hierarchical model uses the segmentation prior and ddCRP to group supersamples and the Chinese Restaurant Process (CRP) to discover activities. We evaluate our approach on the dataset that contains activities of daily living, and demonstrate the outperformance of our proposed ddCRP based model with respect to both, classic parametric Latent Dirichlet Allocation (LDA) and

the non-parametric Chinese Restaurant Franchise (CRF).

4.1 INTRODUCTION

Discovery of daily activities and routines from ubiquitous sensor data provides insights into individual behaviour without prior model learning, which is relevant for assisted living, home patient care, and related applications (Seiter *et al.*, 2014b; Aztiria *et al.*, 2012). A commonly considered concept for assessing human behaviour is to partition activity routines into abstraction levels, where regular routine structures, such as *office work* and *lunch* are composed of context symbols with shorter temporal extend, including activity primitives as *sit*, *walk*, locations, such as *home*, or object use, e.g. *computer*. Context symbols could be detected from the continuous acquired data of on-body and ambient sensors, where frequently supervised classification or data clustering were used (Huynh *et al.*, 2008). Discovering activity routines requires methods for analysing context symbol patterns, where often parametric topic models were applied, such as latent Dirichlet allocation (LDA) (Huynh *et al.*, 2008; Farrahi and Gatica-Perez, 2011). Topic models originate from text mining and aim at discovering hidden themes from word statistics in documents. For parametric topic models it is assumed that one document contains a mixture of a finite number of topics and that each topic is described as probabilistic distribution over words from a predefined vocabulary.

In activity discovery, words correspond to context symbols and topics correspond to activities, which we call context words and activity topics respectively. Typically, documents are obtained using a temporal segmentation of the continuous context word stream with a predefined segment size that is large enough to capture context word statistics. Subsequently, discovery results per segment are retrieved. With frequently used segment sizes of 30 min (Huynh *et al.*, 2008; Sun *et al.*, 2014), activity transitions, variations in activity duration, as well as activities that are shorter than the set segment size may not be accurately identified. Moreover, parametric topic models, such as LDA, require to set the expected number of topics. Selecting topic model parameters, including segment size and number of topics, impacts activity discovery performance and highly depends on dataset properties that may be unknown (Seiter *et al.*, 2014a). In addition, dataset properties, i.e., number of daily activity routines and their duration, may vary by individual and due to changing behaviour patterns. Recently, non-parametric Bayesian topic models were proposed for activity discovery to overcome the dependency on predefined topic count (Nguyen, 2014; Sun *et al.*, 2014) (see Section 4.2 for details). Non-parametric models are based on a Dirichlet process and cover an infinite number of activity topics and their distributions over context words. However, existing non-parametric models for activity discovery also depend on fixed observation segment size.

In this chapter, we introduce a novel hierarchical topic model approach that does not depend on manually selecting parameters segment size and number of topics. Instead, segmentation and topic count estimation is performed based on the data

and jointly with the activity topic discovery. We propose a framework that includes context word extraction and activity discovery. Context words are obtained from sensor data without statistical classifier training and thus do not require activity annotations.

We introduce a segmentation prior considering semantic and temporal information and use the non-parametric distance dependent Chinese Restaurant Process (ddCRP) to group context words that belong to one activity. For example, segmentation of activity *lunch* would contain context words such as *spoon* and *plate*, whereas activity *office work* may contain *computer*. Thus, our semantic relationship representation of *spoon* is “closer” to *plate* than to *computer*.

The main contributions are threefold: (1) We introduce a joint segmentation and activity discovery approach that is independent of the number of topics and the segment size. Here we introduce a hierarchical method combining the non-parametric ddCRP and Chinese Restaurant Process (CRP). We formulate a segmentation prior that considers semantic and temporal features of context words, where semantic representations were extracted from a corpus of Wikipedia articles. (2) We evaluate our approach against the parametric LDA and the non-parametric Chinese Restaurant Franchise (CRF) using the Opportunity dataset that contains multi-modal sensor data of daily living activities (Roggen *et al.*, 2010). (3) We compare discovery performances using context word annotations, actual context word detection from raw data, and synthetic noise to demonstrate capabilities of our ddCRP+CRP approach.

4.2 PREVIOUS WORKS ON ACTIVITY DISCOVERY

Several attempts towards activity discovery from sensor data were made. Gu *et al.* extracted characteristic object use fingerprints applying web-mining and discovered contrast patterns for each activity using emerging patterns (Gu *et al.*, 2010). Begole *et al.* applied data clustering to extract and visualize human’s daily rhythms from computer activity (Begole *et al.*, 2003). As clustering-based methods cannot capture uncertainty in the structure of human activities, frequently probabilistic models have been applied. Barger *et al.* used probabilistic mixture models to infer daily life behavior patterns from clusters of sensor events in a smart home (Barger *et al.*, 2005). Probabilistic topic models have been applied to extract user routines from mobile phone data (Farrahi and Gatica-Perez, 2011; Zheng and Ni, 2012). Huynh *et al.* discovered daily routine patterns from activity primitives by applying a topic model (Huynh *et al.*, 2008). However, all of these topic model approaches are parametric and assume a fixed model complexity. Thus, discovery performance critically depends on the number of topics specified and on the segment size used to derive word statistics. In contrast, our approach is non-parametric, thus estimates optimal topic count from the data structure.

Non-parametric models were recently applied for activity discovery. The hierarchical Dirichlet process HDP-HMM was used for abnormal activity detection (Hu *et al.*, 2009) and activity discovery from smartphone sensor (Zhu *et al.*, 2011). Simi-

larly, Nguyen et al. used HDP to discover latent activity topics from acceleration and proximity data (Nguyen, 2014). Sun et al. used HDP to discover patterns of high-level activities (Sun *et al.*, 2014) from data clusters. While non-parametric topic models estimate an optimal number of topics based on the data, their discovery performance remains sensitive to selecting proper segment sizes. The topic model-based discovery frequently used time-invariant segmentation, such as sliding windows (Sun *et al.*, 2014; Huynh *et al.*, 2008). Yet, time-invariant segmentation fails to handle transitions, variations in activity duration, and short activities accurately. Our approach no longer requires selecting a segment size, but performs segmentation dynamically based on the multi-modal sensor data by introducing a segmentation prior which is composed of semantic and temporal features to group context words of the same activity.

4.3 JOINT SEGMENTATION AND DISCOVERY APPROACH

Time-invariant sliding windows cannot adequately handle variations in activity duration. Figure 4.1 illustrates a segmentation problem of variable durations in activity discovery with examples: Large time-invariant windows, e.g. a segment size of $DS=7$, capture context word statistics of activity 1 exactly (see Fig. 4.1(a)). However, context word statistics for activity 3 would be incomplete, as the context word windows of activity 2 and 1 overlap. Contrary, a small segment size (e.g. $DS=3$) does not provide distinct context word statistics for activity 1:*lunch*, as illustrated in Fig. 4.1(b).

We introduce a joint segmentation and discovery approach as depicted in Fig. 4.2 to solve the segmentation problem. The first stage extracts data from multi-modal sensor sources into context words e.g., *sit*, *spoon moved*. The context word extraction relies on basic logic functions, thus avoiding supervised statistical learning and classification. Subsequently, we introduce a data-driven segmentation based on state changes in context words to obtain supersamples (see Fig. 4.1(c)). Supersamples represent short temporal segments of context words with variable size. As state changes in context words may occur within activities, supersamples may not comprehensively capture context word statistics that represent a particular activity, e.g., activity 1:*lunch* in Fig. 4.1(c). Therefore, supersamples will be grouped according to semantic and temporal context word relations.

We assumed that an activity includes semantically similar context words, whereas the semantic relation of context words between different activities is lower. For example, context words x and y in Fig. 4.1 may correspond to *plate* and *spoon*. Then, $x:plate$ and $y:spoon$ are semantically more similar than $x:plate$ and $z:computer$. To group supersamples that belong to the same activity (Fig. 4.1(d)), we introduce a segmentation prior that considers semantic and temporal relationships of supersamples based on the context words in each supersample. We deduced semantic distances between context words based on *word2vec* representations that were extracted from a corpus of Wikipedia articles (Mikolov *et al.*, 2013). For example, activity 1:*lunch*

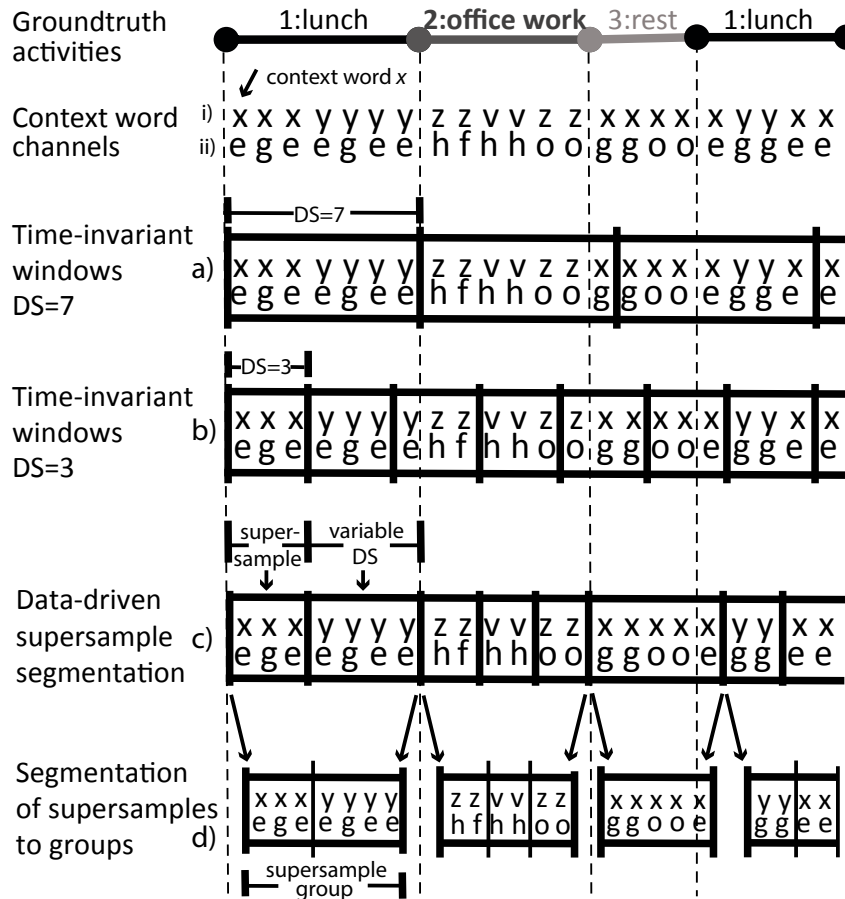


Figure 4.1: Illustration of segmentation methods for activity discovery. Exemplary, 3 activities are shown and segmentations for 2 context word channels $\{i, ii\}$ with the context word vocabulary $\{e, f, g, h, v, x, y, z, o\}$. (a) Time-invariant windowing with segment size $DS=7$. (b) Time-invariant windowing with segment size $DS=3$. (c) Data-driven supersamples segmentation. A new supersample is formed each time a context change occurs in channel (i). (d) supersample groups are segmented by ddCRP with segmentation prior. Whereas in (a) and (b) windows intersect activities, (c,d) perform segmentation according to data.

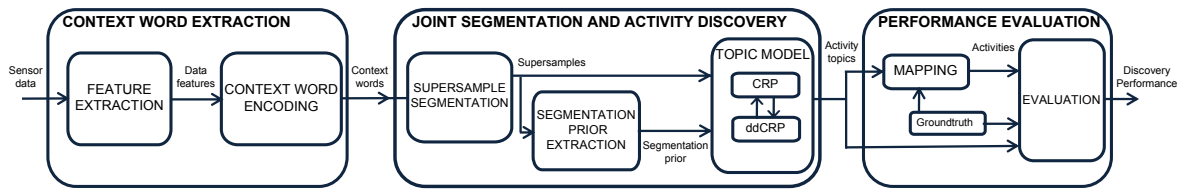


Figure 4.2: Activity discovery framework: Sensor data is processed and encoded in context words using a predefined context vocabulary. Our approach jointly segments context words and performs activity discovery. Initially, a data-driven segmentation transforms context word streams into supersamples as input for a hierarchical, non-parametric topic model. We use semantic and temporal priors to group supersamples of the same activity with ddCRP. The CRP process is then used to infer activity topics from context word statistics of supersample groups. To evaluate performance, activity topics are mapped to a set of activities.

contains a supersample $i=1$ with context words $\{x,x,x,e,g,e\}$ and a supersample $i=2$ with $\{y,y,y,e,g,e,e\}$ (Fig. 4.1(c)). The third supersample $i=3$ belongs to the activity *office work* and includes context words $\{z,z,h,f\}$. We expect higher prior probability to group supersamples 1 and 2 than supersamples 1 and 3 as the semantic and temporal distance of $x:plate$ (supersample 1) and $y:spoon$ (supersample 2) should be smaller than between $x:plate$ (supersample 1) and $z:computer$ (supersample 3). Contrary to the example here, distances for all pairs of context words were considered in the prior (see Sec. 4.4.3 for details).

We then apply a hierarchical, non-parametric topic model for activity topic discovery as depicted in Fig. 4.3: In the local layer, for each data recording supersamples are combined into groups as illustrated in Fig. 4.1(d) and Fig. 4.3 using ddCRP. Grouping by ddCRP depends on the segmentation prior: In our example, supersamples $i=1$ and $i=2$ belong to activity $1:lunch$ and have high prior probability to be grouped contrary to supersamples $i=1$ and $i=3$ that belong to different activities (see Fig. 4.1(c)). We expect supersample groups to provide comprehensive context word statistics describing activities (see Fig. 4.1(d)). Individual data recordings of a dataset likely contain the same activities. Thus, the global layer combines supersample groups that belong to the same activity by CRP to an activity topic group e.g. $q=1:lunch$ (Fig. 4.3). For each activity topic group, the context word distribution is sampled from context word statistics of all assigned supersample groups such that the likelihood of the data is maximized. Retrieved activity topics were finally mapped to activities and discovery performance was analyzed (see Fig. 4.2).

4.4 DISCOVERY FRAMEWORK

The complete discovery framework is illustrated in Fig. 4.2. and detailed below.

4.4.1 Context Word Extraction

The context vocabulary covers X context words $\{e, f, g, \dots\}$ that are extracted from body worn and ambient sensor data. First, features are extracted from raw sensor data (see Fig. 4.2). Each statistical feature from sensor data is transformed to a binary feature using thresholds and subsequently included in logic functions to obtain context words (see Sec. 4.5.2). Parallel operating context word detectors (e.g. *mode of locomotion, object usage*) result in several context word channels. Each context word channel provides either an active context word or a *null class* symbol, when no context word is active.

4.4.2 Segmenting Context Words into Supersamples

We use a data-driven segmentation for context word streams that result in variable sized segments, referred as supersamples similar to superpixels in vision (Chiu and Fritz, 2013). New supersamples are formed each time a context state change occurs (see Fig. 4.1(c) for illustration). As there may be several parallel context channels from different sensors sources, we use the channel that includes the least sparse context word sequence. We consider that supersamples will typically have shorter temporal duration than activities and subsequently need to be grouped. We use a joint segmentation and activity discovery approach, as described below.

4.4.3 Segmentation Priors for Activity Discovery

As the basic units for observations in our framework are supersamples which contain context words inside, we would like to explore the structure of dependencies between observations to infer more semantic-meaningful activities. While the context words have corresponding labels such as *spoon* and *plate*, in this work we explicitly measure the semantic distances between context words based on their *word2vec* vector representations. *Word2vec* (Mikolov *et al.*, 2013) is based on a continuous Skip-gram model that infers word vector representations unsupervised from a corpus of articles, where the word vectors capture semantic relationships between words. Initially, the algorithm constructs a *word2vec* vocabulary of size W from the text corpus and then deduces vector representations based on neural networks. Finally, each word in the *word2vec* vocabulary is represented by the semantic relationship to W other words leading to a $1 \times W$ word vector for each word. We used a *word2vec* vocabulary of dimension $W = 1000$ to extract word vector representations from a corpus of Wikipedia articles (*available at <https://code.google.com/p/word2vec/>*). Context words represented a subset of the *word2vec* vocabulary ($X \ll W$) and were manually

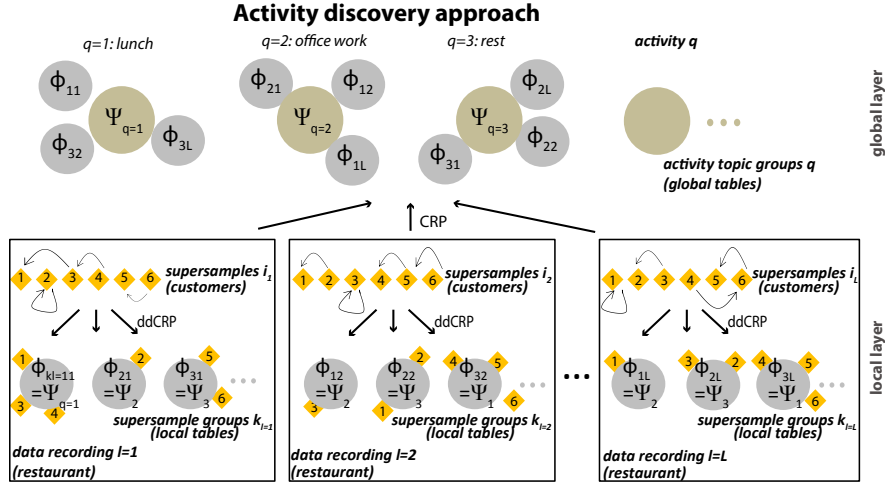


Figure 4.3: Illustration of the hierarchical discovery framework ddCRP+CRP. In the local layer, supersamples from recordings l are combined to supersample groups by ddCRP. Each local supersample group k_l belongs to a local activity topic ϕ_{kl} . In the global layer, local supersample groups from all L recordings are assigned to global activity topic groups q using CRP and share global activity topic Ψ_q . Local activity topics ϕ_{kl} inherit the global activity topic Ψ_q .

mapped to relevant word vectors v_x for X context words by searching the labels of X context words in the *word2vec* vocabulary.

We used the *word2vec*-based semantic as well as temporal distances between supersamples to form a segmentation prior over supersamples for ddCRP that likely groups supersamples belonging to the same activity (see Fig. 4.2). To semantically represent a context word x , we used word vector v_x . To semantically represent supersample i , we calculated the mean word vector v_i across all X_i unique context words x in supersample i : $v_i = \frac{1}{X_i} \sum_{x \in X_i} v_x$. The semantic distance d_{ij}^s of supersamples i and j is the Euclidean distance $d_{ij}^s = d(v_i, v_j)$ of their mean word vectors. The temporal distance d_{ij}^t counts the number of supersamples between supersample i and j . Considering our segmentation prior over supersamples, the supersample assignment c_i can be written as:

$$p(c_i = j | D, f, \alpha) \propto \begin{cases} f^t(d_{ij}^t) f^s(d_{ij}^s) & j \neq i \\ \alpha & j = i \end{cases}. \quad (4.1)$$

The distance measure D and decay function f for ddCRP are composed of a temporal distance measure and decay function (D^t, f^t) and a semantic distance measure and decay function (D^s, f^s) . The window decay function $f^t(d^t) = [d^t < A]$ assigns direct linkage probabilities for supersamples that are at most distance A apart. For the semantic distance d^s , we use an exponential decay function $f^s(d^s) = \exp(-\frac{d^s}{B})$ that decreases linkage probability with increasing semantic distance. B is the width

parameter.

4.4.4 Joint Segmentation and Activity Discovery (ddCRP+CRP)

Our activity discovery approach uses ddCRP in the local layer and CRP in the global layer as illustrated in Fig. 4.3. The ddCRP+CRP approach can be interpreted as follows: There is a set of L data recordings (restaurants) with a shared set of global activity topics Ψ (global dishes) across all recordings (restaurants). For each recording l , supersamples i_l and j_l with small semantic and temporal distances d_{ij_l} are likely grouped to the same supersample group k_l (local table). For example, linked supersamples in Fig. 4.3 (bottom) are assigned to the same supersample group. Each supersample group k_l of all data recordings l is assigned to one global activity topic Ψ_q (global dish) by CRP with the activity topic group q (global table) (see Fig. 4.3, top). The local activity topic ϕ_{kl} (local dish) in recording l inherits the global activity topic Ψ_q (global dish) from the activity topic group q where k_l is assigned to, e.g. Fig. 4.3(c) $\Psi_1 = \phi_{11} = \phi_{3L} = \phi_{32}$. Thus, multiple supersample groups k_l in multiple data recordings l can belong to the same activity topic Ψ_q .

The generative process ddCRP+CRP is described by:

- (1) Each supersample i_l in recording l draws supersample assignment c_{i_l} with supersample group k_{i_l} from ddCRP(D, f, α).
- (2) Each supersample group k_l in recording l draws a global activity topic group assignment q_{kl} from CRP(γ).
- (3) Each global activity topic group q draws activity topic Ψ_q from G_0 .
- (4) For each supersample i_l in recording l , context word statistics u_{i_l} are drawn from η_q , where η_q is a multinomial distribution and $q_{k_{i_l}} = q$.

Given the observed context word statistics u_i for supersample i , the likelihood that u_i is sampled from the global activity topic q is $p(u_i | \Psi_q) = \eta_q(u_i)$. We used Gibbs sampling to infer the probabilities $p(u_i | \Psi_q)$ and thus the most likely activity topic assignment q for each supersample i as detailed in chapter 3.5 (Chiu and Fritz, 2013).

4.5 EVALUATION METHODOLOGY

The evaluation strategy is illustrated in Figure 4.4. We compared performance of our non-parametric ddCRP+CRP approach with data-driven supersamples segmentation and joint segmentation and activity discovery to the parametric LDA-based topic model with time-invariant segmentation (segment size DS) and predefined activity topic count T . We further compared ddCRP+CRP to LDA with supersamples segmentation and predefined T and to the non-parametric model Chinese Restaurant Franchise (CRF) with data-derived T , but time-invariant segmentation DS . CRF (Teh *et al.*, 2006) is a hierarchical method as well. However, CRF uses CRP in the local

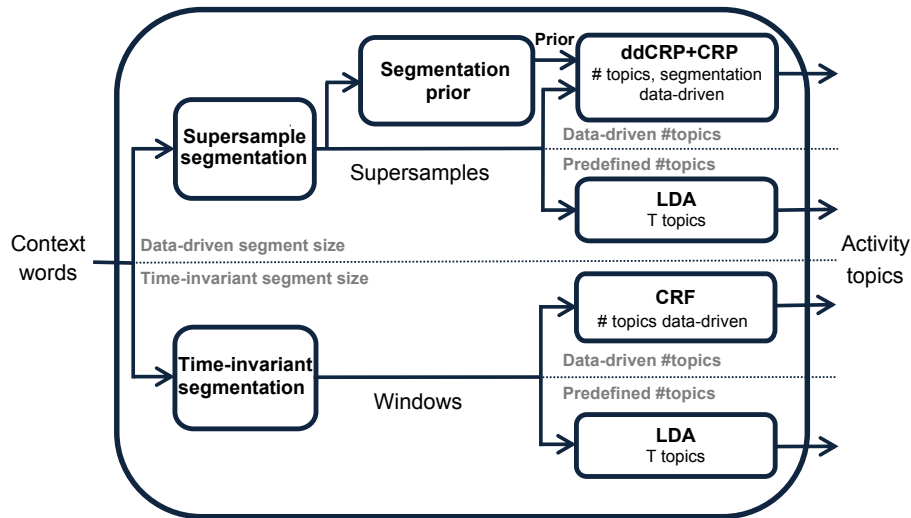


Figure 4.4: Illustration of the evaluation strategy to assess and compare performance of our joint segmentation and activity discovery approach with other variants. We compare performance against LDA with time-invariant segmentation and predefined T , LDA with supersamples segmentation and predefined T , and CRF with time-invariant segmentation.

layer instead of ddCRP and thus does not consider segmentation priors to segment supersample groups. All evaluations were performed per study participant and present average results across all participants and 10 topic model runs. For LDA, we varied DS within $[1, 5]$ min with empirically optimal $T = 10$ as well as varied T within $[5, 20]$ at $DS = 2.5$, as suggested in (Seiter *et al.*, 2014a). We evaluated discovery performance using context word annotations that can be seen as perfect context word detectors, as well as from encoded sensor data using the context vocabulary. We further investigated sensitivity to context word detector noise by adding equally distributed noise to context word annotations.

4.5.1 Dataset

To evaluate our approach, we used the Opportunity dataset that consists of ~ 30 hours of activities of daily living (ADL) recorded at 30 Hz, including annotations for 5 recordings from 4 participants (Roggen *et al.*, 2010). ADLs included *relaxing*, *coffee time*, *early morning*, *cleanup*, *sandwich time* plus a high-level *null class*, in total 120 instances. The dataset further provides annotations for mode of locomotion (4 labels) and object usage (20 labels), plus a low-level *null class*. We considered ADLs as activities, mode of locomotion and object usage corresponded to context words resulting in 25 individual words. To infer context words from sensor data, we used the 3-axis acceleration signals $acc_{x,y,z}$ of the right upper leg sensor SL and the

back-worn sensor SB . We included 3-axis acceleration sensor data of sensors SO_i ($i = 1..15$) attached to 15 objects: *salami, bread, sugar, bottle, milk, spoon, knife cheese, glass, cheese, door1, door2, plate, cup, knife salami, lazychair*. We used binary signals b of reed switches SO_i attached to 5 objects ($i = 16..20$) including *fridge, top drawer, middle drawer, lower drawer, dishwasher*.

4.5.2 Framework Implementation

We extracted a context vocabulary with $X = 25$ context words from the sensor data as mentioned above. Context words included mode of locomotion and object use (O_i) resulting in 21 parallel context word channels (20 object channels, 1 channel for mode of locomotion). Supersamples segmentation from the context word stream was performed using mode of locomotion as context state information, which is the least sparse context word channel of the Opportunity dataset. We used all 20 context word channels with object information to calculate the semantic distance d_{ij}^s between supersamples i and j . For ddCRP+CRP, we used the implementation of (Chiu and Fritz, 2013) with width parameters $B = 0.1$ for f^s and $A = 3$ for f^t , and hyperparameters $\alpha = 50$, $\gamma = 1$ and $\eta = 1$. For CRF, we used hyperparameters $\alpha = 1$, $\gamma = 1$ and $\eta = 1$. For LDA we used the implementation of (Blei *et al.*, 2003) with $\alpha = 1$ and T topics. For time-invariant segmentation, we used sliding windows of size DS and segment step $0.1 * DS$ and applied Borda Count ranking to overlapping segments (Ho *et al.*, 1994).

4.5.3 Performance Estimation

To assess activity discovery performance we mapped discovered activity topics to activities by assigning the most frequent activity per predicted activity topic using the groundtruth. *Null class* data was included for topic discovery, but removed in the performance analysis. As performance measure we used class-normalized accuracy across all 5 activities and the Rand index RI.

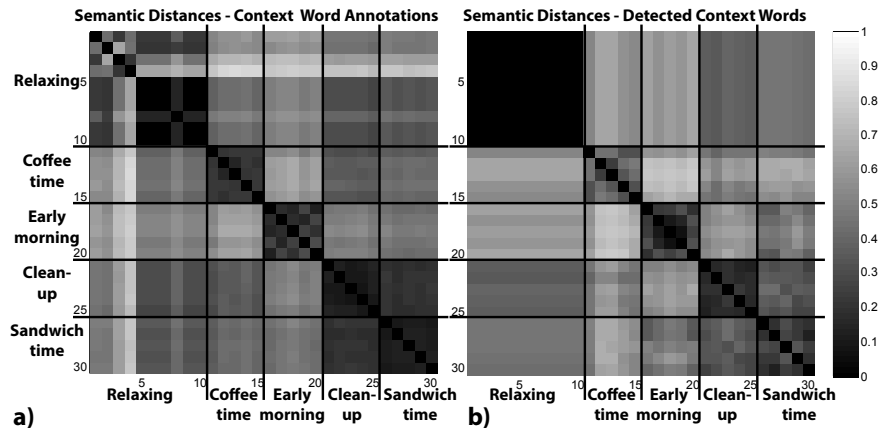


Figure 4.5: Illustration of semantic distances D^s between activity instances of 5 activities in the Opportunity dataset for (a) context word labels and (b) context word detections from sensor data. Semantic distances were calculated from context word vector representations and averaged across all 4 subjects. The graph indicates that context words used within the same activity have close semantic relation.

4.6 RESULTS

4.6.1 Semantic Relationships within and between Activities

Figure 4.5 shows that semantic distances D^s of context word vectors were small among instances of the same activity, e.g. *early morning*. Confirming our approach to use semantic priors, independent activities showed high distances, e.g. *early morning* and *coffee time*. Detector errors may have decreased within-activity similarity of detected context words compared to context labels, e.g. *coffee time*. Nevertheless, we also observed a reverse trend, where context word annotations appeared to be imperfect and incomplete compared to detections, e.g. for *relaxing*, *clean up* and *sandwich time*.

4.6.2 Activity Discovery from Context Word Labels

4.6.2.1 *ddCRP+CRP versus LDA*

Our *ddCRP+CRP* approach yielded 83% accuracy and Rand index $RI = 0.83$, clearly outperforming LDA as depicted in Fig. 4.6. LDA using time-invariant segmentation showed a peak in accuracy and Rand index for $DS = 2.5$ min and $T = 10$ topics.

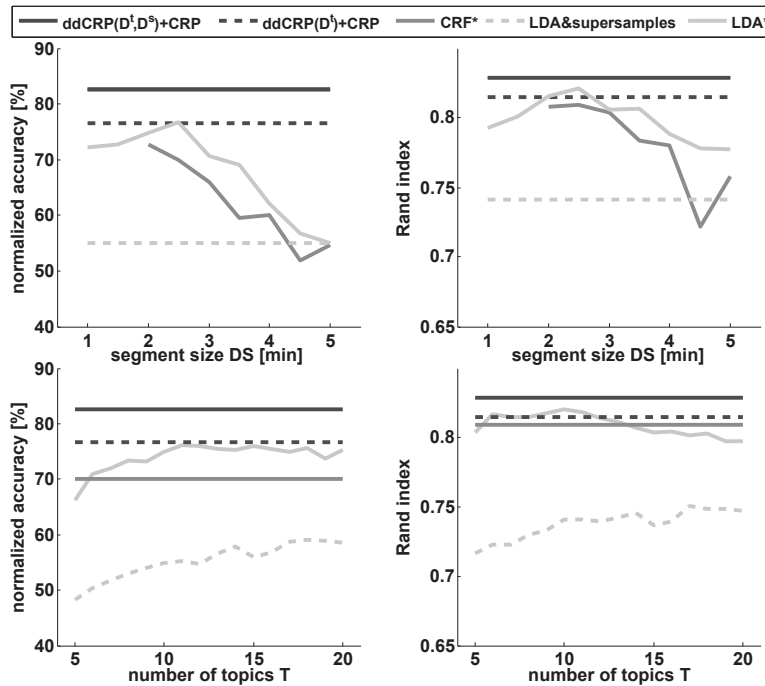


Figure 4.6: Averaged normalized accuracy and Rand index for discovering activities from context word labels. Results are shown for ddCRP+CRP with temporal (D^t) and semantic (D^s) segmentation priors, CRF, and LDA. ddCRP+CRP outperformed non-parametric CRF and parametric LDA. (*) We varied segmentation window and number of topics for CRF and LDA-based methods when parameter dependency was present.

4.6.2.2 *ddCRP+CRP versus CRF*

ddCRP+CRP outperformed non-parametric CRF with optimal segment size by 10% in accuracy and by $RI = 0.02$. With decreasing segment sizes accuracy of CRF increased up to 73%. The Rand index showed a peak at $DS = 2.5$ min with $RI = 0.81$.

4.6.2.3 *Temporal and Semantic Priors*

We assessed the benefit of temporal and semantic priors. ddCRP+CRP with semantic prior increased accuracy by 6% compared to ddCRP+CRP with only temporal prior. The performance of ddCRP+CRP with temporal prior was close to the performance of LDA and CRF with optimal parameters.

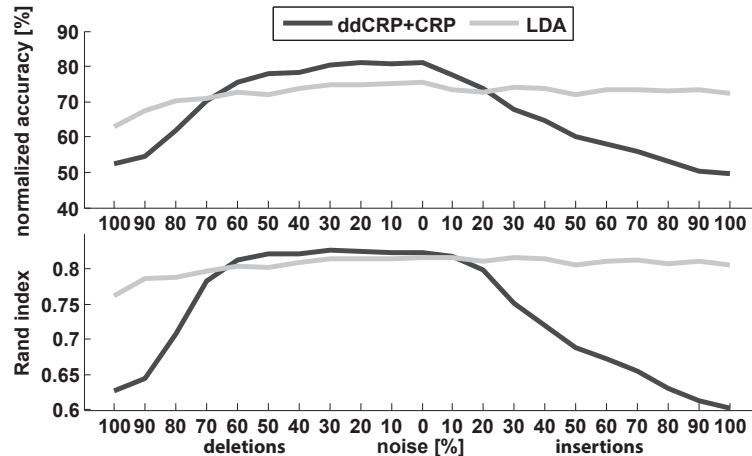


Figure 4.7: Influence of evenly distributed noise over context word detectors on discovery performance. ddCRP+CRP was robust against context word deletions up to 60%, but showed sensitivity to insertions. Our approach outperformed LDA between 60% deletion and 20% insertion noise.

4.6.3 Sensitivity to Context Word Noise

Figure 4.7 shows that ddCRP+CRP was robust against deletion noise with up to 60% deletions and 20% insertion noise, outperforming LDA at optimal parameter settings. In practice, uniformly distributed noise across context word detectors is unlikely to occur. Besides deletions and insertions also timing and substitution errors may hamper discovery. The noise analysis may thus rather illustrate boundaries of our ddCRP+CRP approach: ddCRP+CRP performance depends on the segmentation prior. For uniformly distributed insertion noise, ddCRP likely grouped supersamples of different activities in the local layer leading to less distinct context word statistics of supersample groups at the global CRP layer. Contrary, ddCRP+CRP was less affected by deletions, as they only reduced priors grouping supersamples of the same activity. In contrast, LDA estimates activity topics exclusively from context word statistics in time-invariant segments. Evenly distributed noise offsets all context word statistics and therefore barely changes the context word structure in a segment. The sensitivity of ddCRP+CRP for insertions and robustness against deletions suggests tuning context detectors for high precision.

4.6.4 Activity Discovery from Sensor Data

For activity discovery from detected context words using our context word extraction approach, all methods showed decreased performance compared to discovery from annotations. Figure 4.8 shows that our ddCRP+CRP model outperformed LDA with time-invariant segmentation and optimal parameters ($T = 7$, $DS = 2.5$) by 4.5% accuracy and $\Delta RI = 0.1$. For LDA, optimal activity topic count decreased

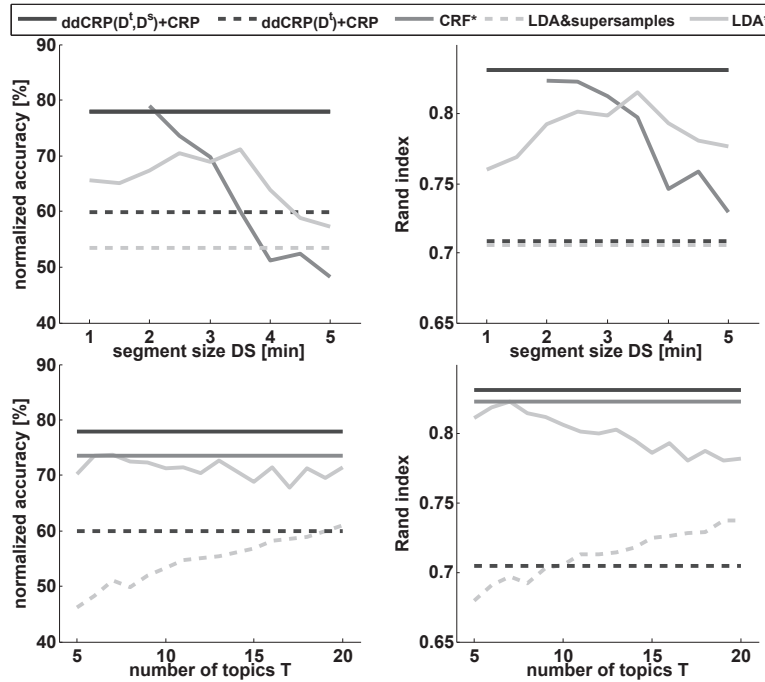


Figure 4.8: Performance of activity discovery from context word detections for ddCRP+CRP including temporal D^t and semantic D^s segmentation priors, CRF and LDA. Our non-parametric ddCRP+CRP approach outperformed parametric LDA and non-parametric CRF at their optimal parameter settings. (*) We varied segmentation window and number of topics for CRF and LDA-based methods when parameter dependency was present.

for detected context words, compared to the discovery from annotations ($T = 7$ vs. $T = 10$). Moreover, optimal segment size changed ($DS = 3.5$ vs. $DS = 2.5$). Our ddCRP+CRP model automatically selected a smaller number of activity topics for activity discovery from detected context words compared to context word labels ($T = 7$ vs. $T = 15$). ddCRP+CRP with just temporal segmentation prior performed with 60% accuracy worse than ddCRP(D^t, D^s)+CRP (78%). CRF with optimal segment size $DS = 2$ min yielded the same accuracy and slightly smaller Rand index $\Delta RI = -0.1$ as ddCRP+CRP, but more activity topics $\Delta T = 5$.

4.7 DISCUSSION

By introducing a framework for joint segmentation and activity discovery in this work, the time-invariant segmentation and parameters used in previous works towards unsupervised activity discovery were removed. Our ddCRP+CRP approach performed supersamples segmentation and activity discovery simultaneously and outperformed the parametric LDA as well as non-parametric CRF, both using time-

invariant segmentation. Our results indicate that the LDA-based approach could not deal with variations in activity duration even for optimal LDA topic model parameters. As topic models are bag-of-word approaches the context word sequences within supersamples was discarded in our approach. Previous work using n-gram topic modeling showed that sequence information could rarely improve performance over an LDA-based approach (Seiter *et al.*, 2014a).

We used basic logic functions to extract context words, thus avoiding statistical classifier learning. Our data-driven context word segmentation generated supersamples that were typically shorter than activities. Hence, individual supersamples did not capture distinct context word statistics to describe activities and may explain the poor performance of LDA using supersamples segmentation. Using a temporal segmentation prior for ddCRP+CRP increased accuracy over the LDA-based approach, but performed less accurate compared to ddCRP+CRP using a temporal and semantic segmentation prior. For discovery from context labels, ddCRP+CRP outperformed CRF by 10% in accuracy. For discovery from detected context words, both methods showed similar peak accuracy. However, CRF yielded 5 additional activity topics compared to ddCRP+CRP (CRF: $T = 12$ at $DS = 2$, ddCRP: $T = 7$).

While the non-parametric model ddCRP+CRP and CRF infer optimal activity topic count T , hyperparameters α, γ determine the expectation over T . There are no established strategies to select α, γ and often their setting was not reported. However, if a range for $\hat{T} \approx T \pm 5$ is estimated, we found that ddCRP+CRP and CRF can automatically choose an optimal T . In our work, we used the same hyperparameter setting for discovery using labels and detected context words. In our tests, ddCRP(D^t, D^s)+CRP showed similar discovery performance even when hyperparameters were varied, indicating robustness of the method. Omitting semantic priors, i.e. ddCRP(D^t)+CRP showed lower robustness to hyperparameter variation that may explain the performance drop from our evaluation based on labels compared to detected context words.

Evaluating activity discovery results often requires interpreting discovered topics and thus mapping of T activity topics to M actual activities. With high T , mapping becomes less intuitive and in practice requires more supervision effort to reliably interpret the topic result. In contrast, $T = M$ may not optimally represent actual activities, as activities could be composed of several activity topics. Our mapping approach favoured higher accuracies for increasing activity topics. At the extreme, if T would equal the number of data segments accuracy would approach 100%. We used the Rand index as second evaluation metric to control for arbitrary large T . Rand index decreases for large T , as many false positives occur. For an intuitive mapping, T in the range of M was desirable. Thus, in this work results were shown for $T < 20$ activity topics to describe the $M = 5$ activities. With CRF, a topic count $T > 20$ was obtained for $DS < 2$ min.

We segmented context words into supersamples and used context state changes to determine supersample bounds. In this work, we only considered context changes in one selected context word channel. State changes could similarly be estimated by combining several context word channels to create a virtual context state. Selecting

and constructing a segmentation source still requires expert knowledge about the targeted discovery objective and context word processing. Similarly, constructing logic functions to derive context words requires knowledge about the sensor modalities and discovery goals. Nevertheless, we consider that such logic functions could be catalogued according to sensor type and scenario, thus become reusable for similar discovery applications without parametric adjustments.

Our approach required context words with semantic meaning, as we used *word2vec* to formulate a segmentation prior. Nevertheless, *word2vec* is flexible and could be applied to a different corpus, e.g. containing data clusters or other symbols extracted from sensor data. In our approach context words corresponded to a small subset of the *word2vec* word vocabulary and we manually extracted context word vectors from the word vocabulary. Instead, string matching could be applied in future to automate the mapping. In our evaluation, all context words could be mapped to a word vocabulary of $W = 1000$. It is nevertheless simple to increase the vocabulary, if corresponding words could not be found or to search synonyms using language processing algorithms. We used a generic text corpus from Wikipedia to extract the *word2vec* vocabulary. Domain specific text corpora might yield even more relevant word coverage and context word vectors.

4.8 CONCLUSION

We introduced a novel non-parametric topic model approach for joint segmentation and activity discovery from sensor data that is independent from topic model parameters, such as segment size and number of topics. We segmented context words into supersamples using context state and formulated a segmentation prior with semantic and temporal information to group supersamples that belong to individual activities using ddCRP and CRP. With this method, segmentation is adjusted to the underlying data. Evaluation results show that our approach can outperform classical parametric LDA and non-parametric CRF even at optimal parameter settings. We concluded that combining segmentation and non-parametric activity discovery by using a segmentation prior and ddCRP+CRP is an adequate technique for activity discovery, and we believe the segmentation prior can be adapted to other datasets with different sensor modalities and discovery objectives.

Contents

5.1	Introduction	61
5.2	Related Work	63
5.3	∇ HOG	65
	5.3.1 Gradients Computation	65
	5.3.2 Weighted Vote into Spatial and Orientation Cells	66
	5.3.3 Contrast Normalization	67
	5.3.4 Implementation	67
5.4	Experimental Results	68
	5.4.1 Reconstruction from HOG Descriptors	68
	5.4.2 Pose Estimation	70
5.5	Conclusions	75

IN this chapter we study the task of 3D model fitting by differentiating the Histogram of Oriented Gradient (HOG) feature representation. Aligning 3D CAD models to 2D images of objects has received research attention and is closely related to many topics: the 3D shape of CAD models together with the appearance features from images provide a informative representation for object detection; the 3D pose estimation of 2D objects enables the 3D volumetric and support reasoning in the 3D scene understanding; the silhouettes and boundaries from the 2D projection of aligned 3D models also helps for getting better image segmentation. Based on the Exemplar LDA models which are popularly used for 2D-3D matching (Lim *et al.*, 2013; Aubry *et al.*, 2014a), we realize that the HOG descriptor utilized for feature representation of exemplar templates is piecewise differentiable hence lends to opportunities of end-to-end optimization for pose parameters. We present our implementation of ∇ HOG based on the auto-differentiation toolbox Chumpy (Loper, 2014) and first test on the application of pre-image visualization as a proof of concept. Then the Exemplar LDA pipeline for pose estimation including ∇ HOG and the existing differentiable renderer OpenDR (Loper and Black, 2014) is presented. The experimental results demonstrate that our proposed pipelines improve on the respective state-of-the-art HOG approaches on both applications.

5.1 INTRODUCTION

Since the original presentation of the Histogram of Oriented Gradients (HOG) descriptor (Dalal and Triggs, 2005) it has seen many use cases beyond its initial target

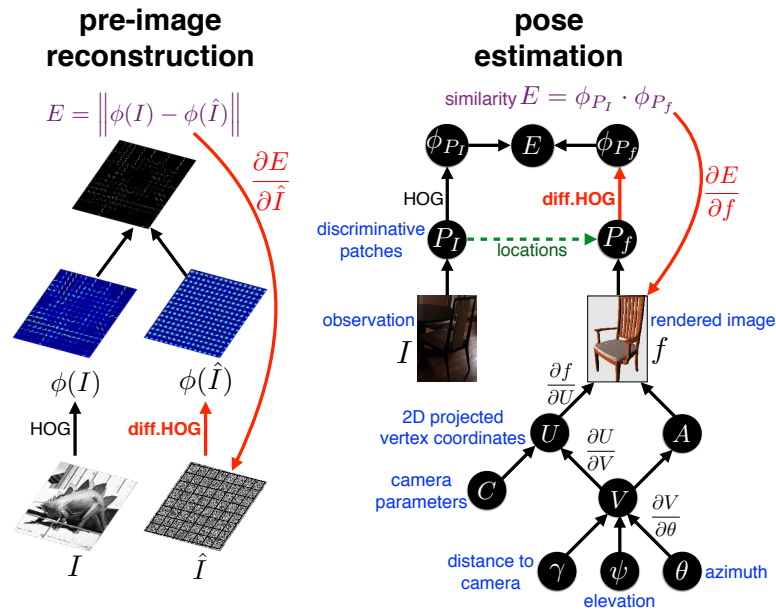


Figure 5.1: We exploit the piecewise differentiability of the popular HOG descriptor for end-to-end optimization. The figure shows applications on the pre-image reconstruction given HOG features as well as the pose estimation task based on the same idea.

application to pedestrian detection. Most prominently it is a core building block of the widely used Deformable Part Model (DPM) object class detector (Felzenszwalb *et al.*, 2010) and exemplar models (Malisiewicz *et al.*, 2011) which both on their own have seen many follow-up approaches. Most recently, HOG-based approaches have repeatedly shown good generalization performance to rendered (Aubry *et al.*, 2014a) and artistic images (Aubry *et al.*, 2014b), while such type of generalizations are non-trivial to achieve in recently very successful deep learning models in vision (Peng *et al.*, 2015).

As all feature representations also HOG seek a reduction of information in order to arrive at a more compact representation of the visual input that is more robust to nuisances such as noise and illumination. It is specified as a mapping of an image into the HOG space. The resulting representation is then typically further used in classification or matching approaches to solve computer vision tasks.

While HOG is only defined as a feed-forward computation and introduces an information bottleneck, sometimes we desire to invert this pipeline for further analysis. E.g. previous work has tried visualize HOG features by solving an pre-image problem (Vondrick *et al.*, 2013; Kato and Harada, 2014). Given a HOG representation of an unobserved input image, the approaches try to estimate an image that produces the same HOG representation and is close to the original image. This has been addressed by sampling approach and approximation of the HOG computation in order to circumvent the problem of the non-invertible HOG

computation. Another example, is pose estimation based on 3D models (Xiang *et al.*, 2014; Aubry *et al.*, 2014a; Pepik *et al.*, 2015; Stark *et al.*, 2010) that exploits renderings of 3D models in order to learn a pose prediction model. Here the HOG computation is followed up by a Deformable Part Model (Felzenszwalb *et al.*, 2010) or simplified versions that related to the Exemplar Model (Malisiewicz *et al.*, 2011). Typically, these methods employ sampling based approaches in order to render discrete view-points that are then used in a learning-based scheme to match to images.

In our work, we investigate directly computing the gradient of the HOG representation which then can be used for end-to-end optimization of the input w.r.t. the desired output. For the visualization via pre-image estimation, we observe the HOG representation and compute the gradient w.r.t. the raw pixels of the input image. For pose estimation we consider the whole pose scoring pipeline of (Aubry *et al.*, 2014a) that constitutes a model with multiple parts and a classifier on top of the HOG representation. Here we show how to directly maximize the pose scoring function by computing the gradient w.r.t. the pose parameters. In contrast to the previous approach, we do not rely on pre-rendering views exhaustively and our pose estimation error is therefore not limited by the initial sampling.

We compare to previous works on HOG visualizations and HOG-based pose estimation using rendered images. By using our approach of end-to-end optimization via differentiation of the HOG computation, we improve over the state of the art on both tasks.

5.2 RELATED WORK

The HOG feature representation is widely used in many computer vision based applications. Despite its popularity, its appearance in the objective functions usually makes the optimization problem hard to operate where it is treated as a non-differentiable function (Huang *et al.*, 2011; Xiong and De la Torre, 2013). How to invert the the feature descriptor to inspect its original observation invokes a line of works of feature inversion and feature visualization (pre-image) problem. There are plenty of works on this interesting topic. Given the HOG features of a test image, Vondrick *et al.* (Vondrick *et al.*, 2013) tried in their baseline to optimize the objective with HOG involved by the numerical derivatives but failed to get reasonable results, thus in their proposed method the inversion is done by learning a paired dictionary of features and the corresponding images. Weinzaepfel *et al.* (Weinzaepfel *et al.*, 2011) attempted to reconstruct the pre-image of the given SIFT descriptors (Lowe, 1999) based on nearest neighbor search in a huge database of images for patches with the closet descriptors. Kato *et al.* (Kato and Harada, 2014) study the problem of pre-image estimation of the bag-of-words features and they rely on a large-scale database to optimize the spatial arrangement of visual words. Although these and other related works provide different ways to approximately illustrate the characteristic of the image features, we nearly have not seen the work directly addressing the differentiable form of the feature extraction procedure. In contrast, our approach contributes to make the differentiation of HOG descriptor practical such that it can

be easily plugged into the computer vision pipeline to enable direct end-to-end optimization and extension to hybrid MCMC schemes (Kulkarni *et al.*, 2015a,b). One most relevant work to ours is from Mahendran *et al.* (Mahendran and Vedaldi, 2015), which inverts feature descriptors (HOG (Felzenszwalb *et al.*, 2010), SIFT (Lowe, 1999), and CNNs (Krizhevsky *et al.*, 2012)) for a direct analysis of the visual information contained in representations, where HOG and SIFT are implemented by Convolutional Neural Networks (CNNs). However, their approach contains an approximation to the orientation binning stage of HOG/SIFT and includes two strong natural image priors in the objective function with some parameters need to be estimated from training set. Instead in our work, we do not have any approximation in the HOG pipeline and no training is needed.

Despite deep-learning based features are in fashion these years, there are plenty of applications using HOG, in particular the Exemplar LDA (Hariharan *et al.*, 2012) for the pose estimation task with rendered/CAD data (Aubry *et al.*, 2014a; Lim *et al.*, 2013; Choy *et al.*, 2015). In (Dong and Soatto, 2015), slightly-modified SIFT (gradient-histogram-based as HOG) can beat CNNs in feature matching task. In this chapter, we specifically demonstrate the application of our ∇ HOG on the pose estimation problem for aligning 3D CAD models to the objects on 2D real images, we briefly review some recent research works here. Lim *et al.* (Lim *et al.*, 2013) assume the accurate 3D CAD model of the target object is given, based on the discretized space of poses for initialization they estimate the poses from the correspondences of LDA patches between the real image and the rendered image of CAD model. Aubry *et al.* (Aubry *et al.*, 2014a) create a large dataset of CAD models of chair objects, with rendering each CAD model from a large set of viewpoints they train the classifiers of discriminative exemplar patches in order to find the alignment between the chair object on the 2D image and the most similar CAD model of the certain rendering pose. In addition to the discrete pose estimation scheme as (Aubry *et al.*, 2014a), there has been works on continuous pose estimation (Song *et al.*, 2011; Choy *et al.*, 2015; Pepik *et al.*, 2015). For instance, Pepik *et al.* (Pepik *et al.*, 2015) train a continuous viewpoint regressor and also the RCNN-based (Girshick *et al.*, 2014) key-point detectors which are used to localize the key-points on 2D images in an object class specific fashion, with the correspondences between the key-points on the 2D image and 3D CAD model, they estimate the pose of the target object. However, for these current state-of-the-art approaches most of them need to collect plenty of data to train the discriminative visual element detectors or key-point detectors for the matching, or to render many images of CAD models of various viewpoints in advance. Instead, our proposed method manages to combine the ∇ HOG based exemplar LDA model with the approximate differentiable renderer from (Loper and Black, 2014) which enable us to have directly end-to-end optimization for the pose parameters of the CAD model in alignment with the target object on the real images.

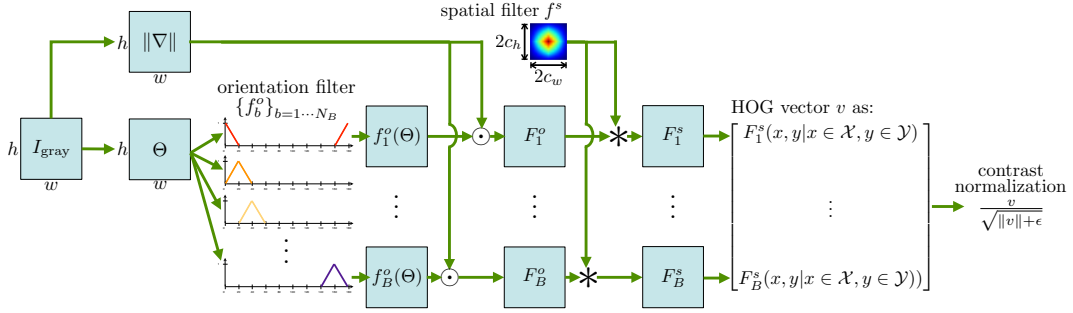


Figure 5.2: Visualization of the implementation procedure for our ∇ HOG method.

5.3 ∇ HOG

Here we describe how we achieve the derivative of the HOG descriptor. In the original HOG computation, there are few sequential key-components, including 1) computing gradients, 2) weighted vote into spatial and orientation cells, and 3) contrast normalization over overlapping spatial blocks. In our implementation we follow the same procedure. For each part we argue for piecewise differentiability. The differentiability of the whole pipeline follows from the chain rule of differentiation. Figure 5.2 shows an overview of the computations involved in the HOG feature computation pipeline which we describe in details in the following.

5.3.1 Gradients Computation

If a color image $I \in \mathbb{R}^{w \times h \times 3}$ is given, we first compute its gray-level image:

$$I_{\text{gray}} = I(:, :, 0) * 0.299 + I(:, :, 1) * 0.587 + I(:, :, 2) * 0.114 \quad (5.1)$$

Then we follow the best setting for gradient computation as in Dalal's approach (Dalal and Triggs, 2005), to apply the discrete derivative 1-D $[-1, 0, 1]$ masks on both horizontal and vertical directions without Gaussian smoothing. We denote the gradient maps on horizontal and vertical directions as G_x and G_y , while the magnitude $\|\nabla\|$ and direction Θ of gradients can be computed by:

$$\begin{aligned} \|\nabla\| &= \sqrt{G_x^2 + G_y^2} \\ \Theta &= \arctan(G_y, G_x) \end{aligned} \quad (5.2)$$

Note that here we use unsigned orientations such that the numerical range of the elements in $\|\nabla\| \in [0, 180]$. The L_2 norm is denoted by $\|\cdot\|$ through this chapter for consistency.

Differentiability: The conversion to gray as well as the derivative computation via linear filtering are linear operations and therefore differentiable. \arctan is differ-

entiable in \mathbb{R} and the gradient magnitude $\|\nabla\|$ is differentiable due to the chaining of the differentiable squaring function and the square root over values in \mathbb{R}^+ .

5.3.2 Weighted Vote into Spatial and Orientation Cells

After we have the magnitude and direction of the gradients, we proceed to do the weighted vote of gradients into spatial and orientation cells which provides the fundamental nonlinearity of the HOG representation. The cells are the local spatial regions where we accumulated the local statistics of gradients by the histogram binning of their orientations. Assume we divide the image region into $N_w^c \times N_h^c$ cells of size $c_w \times c_h$, for each pixel located within the cell we compute the weighted vote of its gradient orientation to an orientation histogram (In our setting we use the same setting as Dalal's to have the histogram of 9 bins spaced over $0^\circ - 180^\circ$ which ignores the sign of the gradients).

Normally for each cell its orientation histogram is represented in a 1-D vector of length B (number of bins), but this operation will miss the positions of the pixels which contribute to the histogram. This does not lead to a formulation that allows for derivation of the HOG representation with respect to different pixel positions. Our main observation here is to view each orientation binning as a filter f_b^o applied to each location in the gradient map. We store the filtered results in $F^o \in \mathbb{R}^{w \times h \times B}$. Analogously, the pixel-wise orientational filters $\{f_b^o\}_{b=1 \dots B}$ are chosen to follow the bi-linear interpolation scheme of the gradients in neighboring orientational bins:

$$f_b^o(\Theta) = \text{clip}_{\min=0}^{\max=1} \left(1 - |\Theta - \mu_b| \times \frac{B}{180} \right) \quad (5.3)$$

$$F_b^o = \|\nabla\| \odot f_b^o(\Theta), \quad \forall b \in 1 \dots B \quad (5.4)$$

where μ_b is the central value of orientation degree for filter f_b^o , $\text{clip}_{\min=0}^{\max=1}$ function clamps the numerical range between 1 and 0, and \odot is an element-wise multiplication. (Note that for the first filter f_1^o we also take care of the numerical range. See the visualization shown in Figure 5.2.)

We have the F^o for orientational binning, we then apply spatial binning for each cell. Here as in the Dalal's method, to reduce the aliasing, for each pixel location it will contribute to its 4 neighboring cells proportional to the distances to the centers of those cells, in another word, the votes are interpolated bilinearly. Following the similar trick as in orientational binning, by creating a $2c_w \times 2c_h$ bilinear filter f^s where its maximum value 1 is in the center with decreasing values toward four corners to minimum value 0, as shown in Figure 5.2, we convolve it with all F_b^o to get the spatial filtered results F_b^s :

$$F_b^s = F_b^o * f^s, \quad \forall b \in 1 \dots B \quad (5.5)$$

then the spatial binning for cells can be easily fetched from:

$$F_b^s(x, y | x \in \mathcal{X}, y \in \mathcal{Y}), \quad \forall b \in 1 \dots B \quad (5.6)$$

where $(\mathcal{X}, \mathcal{Y})$ are the (x, y) coordinates of the centers for all cells.

Note that till here when you concatenate $v = \{F_b^s(x, y | x \in \mathcal{X}, y \in \mathcal{Y})\}_{b=1 \dots B}$ then actually we get exactly the same representation as from original HOG approach.

Differentiability By re-representing the data, we have shown that the histogramming and voting procedure of the HOG approach can be viewed as linear filtering operations followed up by a summation. Both steps are differentiable.

5.3.3 Contrast Normalization

In the original procedure of Dalal's HOG descriptor, contrast normalization is performed on every local region of size 3×3 cells, which they call *blocks*. As many recent applications that we are interested in (Aubry *et al.*, 2014a,b; Kato and Harada, 2014; Vondrick *et al.*, 2013; Felzenszwalb *et al.*, 2010) do not use blocks, we do not consider them either in our implementation. While this step is possible to incorporate, it would also lead to increased computational costs due to multiple representation of the same cell. We instead only use the global normalization by using the robust *L2-norm*. Given the HOG representation v from previous steps, the global contrast normalization can be written as:

$$v_{\text{normalized}} = \frac{v}{\sqrt{\|v\| + \epsilon}} \quad (5.7)$$

where ϵ is a small positive constant.

Differentiability: This is a chain of differentiable functions and therefore the whole expression is differentiable.

Difference to Original HOG While there is a large diversity of HOG implementations available by now, we summarize the two main difference to the standard one as proposed in (Dalal and Triggs, 2005): First, the original HOG compute the the gradients on different color channels and apply the maximum operator on the magnitudes over all channels to get the gradient map. In our implementation we simply first transform the color image into gray scale and compute the gradient map directly. Second, we do not include the local contrast normalization for every overlapping spatial blocks. But we do include the global, robust *L2* normalization.

5.3.4 Implementation

In the above equations (Eq. 5.1, 5.2, 5.3, 5.5, 5.7) all the operations are (piecewise-) differentiable (summation, multiplication, division, square, square root, arc-tangent, clip), with the use of the chain rule, our overall HOG implementation is differentiable on each pixel position. Overall, this is not surprising as visual feature representations are designed to vary smoothly w.r.t. to small changes in the image. We have

implemented this version of the HOG descriptor by using the Python-based auto-differentiation package *Chumpy* (Loper, 2014), which evaluates an expression and its derivatives with respect to its inputs. The package and our extension integrate with the recently proposed Approximate Differentiable Renderer OpenDR (Loper and Black, 2014).

5.4 EXPERIMENTAL RESULTS

5.4.1 Reconstruction from HOG Descriptors

We evaluate our proposed ∇ HOG method on the image reconstruction task based on the feature descriptors. We are interested in this task since it provides a way to visualize the information carried by the feature descriptors and open the opportunity to examine the feature descriptor itself instead of based on the performance measures of certain tasks as proxies. There is already prior work on this problem. (Kato and Harada, 2014; Vondrick *et al.*, 2013; d’Angelo *et al.*, 2012) focus on different feature representations such as Bag-of-Visual-Words (BoVW), Histogram of Orientated Gradients (HOG), and Local Binary Descriptors (LBDs). However, state-of-the-art approaches typically need to use large-scale image bases for learning the reconstruction.

Objective As we have derived the gradient of the HOG feature w.r.t. the input, we can – given a feature vector – directly optimize for the reconstruction of original image without any additional data needed. To define the problem more formally, let $I \in \mathbb{R}^{X \times Y}$ be an image and its HOG representation as $\phi(I)$, we optimize to find the reconstructed image \hat{I} whose HOG features $\phi(\hat{I})$ have the minimum euclidean distance E to $\phi(I)$:

$$\begin{aligned} \hat{I} &= \operatorname{argmin}_{\hat{I} \in \mathbb{R}^{X \times Y}} E \\ &= \operatorname{argmin}_{\hat{I} \in \mathbb{R}^{X \times Y}} \|\phi(I) - \phi(\hat{I})\| \end{aligned} \tag{5.8}$$

The option to approach the problem in this way was mentioned in (Vondrick *et al.*, 2013), however there was no result achieved as numerical differentiation is very computational expensive in this setting. Direct optimization is facilitated for the first time using our ∇ HOG approach.

An overview of our approach is shown in Figure 5.1. We compute derivatives $\frac{\partial E}{\partial i_{x,y}}$ with respect to the intensity values $i_{x,y}$ of all the pixel positions (x,y) on \hat{I} via auto-differentiation. By gradient-based optimization we are able to find a local minimum of E and corresponding reconstructed image \hat{I} . In order to regularize our estimation, we introduce a smoothness prior that penalizes gray value changes of adjacent pixels. Intuitively, this encourages propagation of information into areas

without strong edges for which no signal from the HOG features is available.

$$\hat{I} = \operatorname{argmin}_{\hat{I} \in \mathbb{R}^{X \times Y}} \|\phi(I) - \phi(\hat{I})\| + \zeta \sum_{p,q \in \mathcal{N}} \|i_p - i_q\| \quad (5.9)$$

where $p, q \in \mathcal{N}$ means that pixel p and q are neighbors, and ζ is the weight for the smoothness term which we usually set to a big number as 10^2 in our experiments. Although we give a high weight for the smoothness term, it will only play a key role in the first few iterations of the optimization procedure then the euclidean distance E will dominate to find the local minimum.

The evaluation is based on the image reconstruction dataset proposed in (Kato and Harada, 2014) which contains 101 images for all the categories from Caltech 101 dataset (Fei-Fei *et al.*, 2007) and all have a resolution of 128×128 . We compare our method with few state-of-the-art baselines on image reconstruction from feature descriptions: the **BoVW** method from (Kato and Harada, 2014), the **HOGgles** method from (Vondrick *et al.*, 2013), also **CNN-HOG** and **CNN-HOGb**(CNN-HOG with bilinear orientation assignments) from (Mahendran and Vedaldi, 2015).

Note that our ∇ HOG described in Section 5.3 is based on **Dalal’s**-type HOG (Dalal and Triggs, 2005), while for HOGgles/CNN-HOG/CNN-HOGb baselines they are using **UoCTTI**-type HOG (Felzenszwalb *et al.*, 2010) which additionally considers directed gradients. To have a fair comparison, we also implement UoCTTI HOG under our proposed framework.

We propose two additional variants for reconstruction that exploit multi-scale information as shown in Figure 5.3.

∇ HOG multi-scale We use the single scale HOG descriptor as input, but we first reconstruct $\hat{I}_{\frac{1}{s}}$ with s times smaller resolution than I (the cell size for $\phi(\hat{I}_{\frac{1}{s}})$ is $\frac{1}{\sqrt{s}}$ of the original one used for $\phi(I)$, $s \in \{4, 16, 64\}$ in our experimental setting.). After few iterations of updates in optimization process, we up-sample $\hat{I}_{\frac{1}{s}}$ to higher resolution and continue the reconstruction procedure. These steps are repeated until we reach the initial resolution of I .

∇ HOG multi-scale-more We use the multi scale HOG vectors of the original image I as the input. For the reconstruction on different scale s , the corresponding HOG descriptor $\phi(I_{\frac{1}{s}})$ extracted on the same scale is used in the euclidean distance E , as shown in Figure 5.3(b). As additional HOG descriptors are computed from the original image at different scales, we use more information than in the original setup and therefore the results of the “multi-scale-more” approach cannot be directly compared to prior works.

The optimization is done based on the nonlinear optimization using Powell’s dogleg method (Lourakis and Argyros, 2005) which is implemented in *Chumpy* (Loper, 2014). Example results of the multi scale approaches can be seen in Table 5.1.

Results In order to quantify the performance of image reconstruction, different metrics have been proposed in prior works. For instance, in (Kato and Harada, 2014)

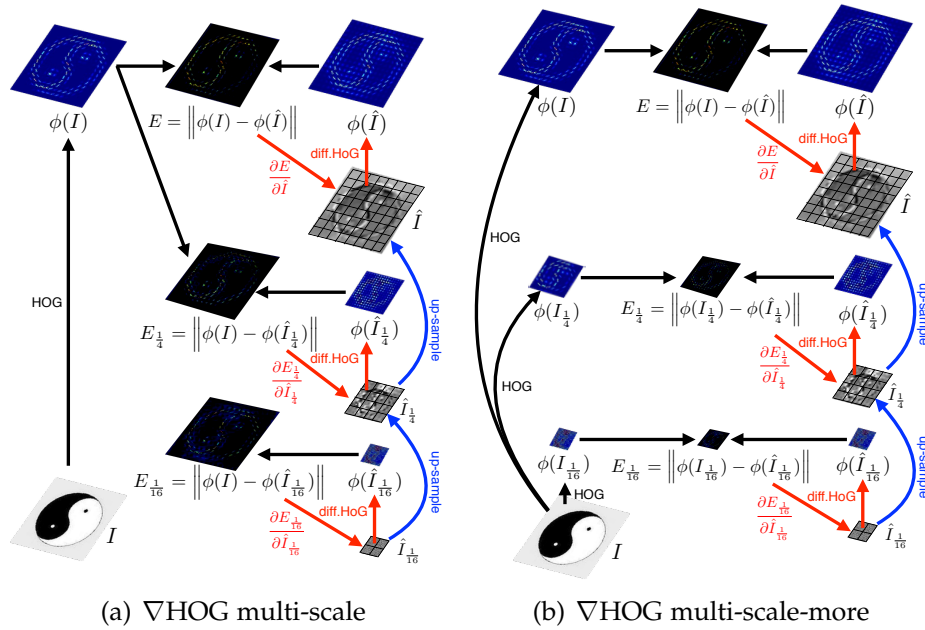


Figure 5.3: Visualizations of variants of the proposed method for the task of image reconstruction from feature descriptors.

the mean squared error of raw pixels is utilized, while in (Vondrick *et al.*, 2013) the cross-correlation is chosen to compare the similarity between the reconstructed image and the original one. In addition to using cross-correlation as the metric for qualitative evaluation, we also investigate different choices used by the research works on the problem of image quality assessment (IQA), including mutual information and **Structural Similarity (SSIM)** (Wang *et al.*, 2004). In particular, mutual information measures the mutual dependencies between images hence gives another metric for similarities, while SSIM measures the degradation of structural information for the distorted/reconstructed image from the original one, under the assumption that human visual perception is adapted to discriminate the structural information from the image.

We report the performance numbers from all the metrics in Table 5.2. The proposed method using UoCTTI-type HOG outperforms the state-of-the-art baselines by a large margins for all metrics. Visually inspected, our proposed method can reconstruct many details in the images and also give accurate estimate on gray-scale values if using UoCTTI HOG. Please note again, our method does not need any additional data for training while in baselines training is necessary.

5.4.2 Pose Estimation

We also evaluate our ∇ HOG approach on a pose estimation task where 3D CAD models have to be aligned to objects in 2D images. We build on openDR (Loper and

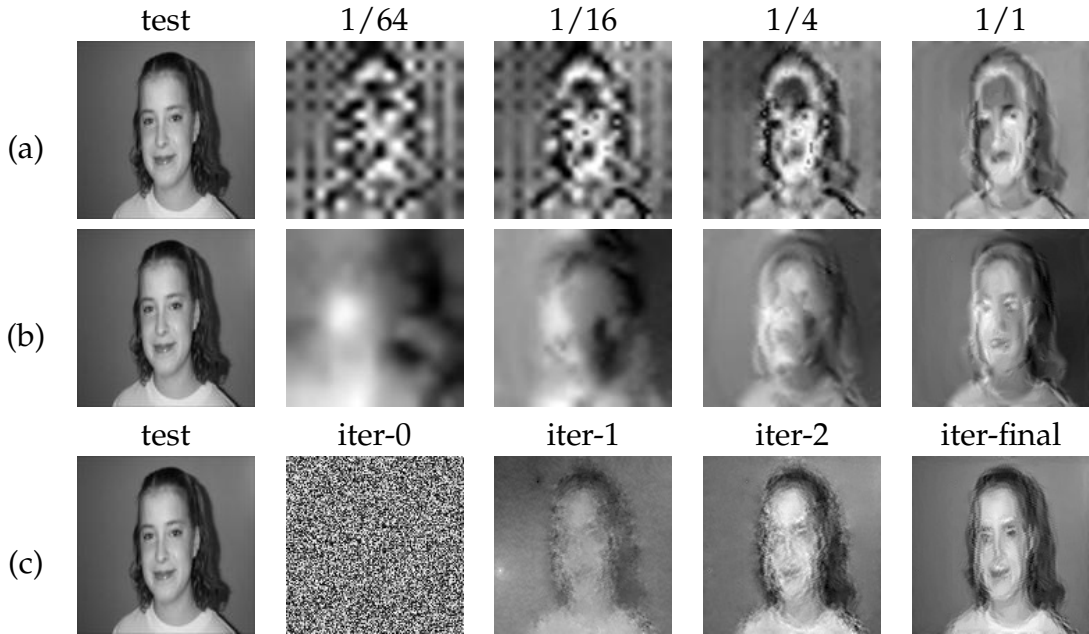


Table 5.1: Example results for (a)(b) ∇ HOG **multi-scale** and ∇ HOG **multi-scale-more** in which both are based on Dalal-HOG(Dalal and Triggs, 2005); and (c) for ∇ HOG on UoCTTI-HOG(Felzenszwalb *et al.*, 2010).

Black, 2014) which is an approximate differentiable renderer. It parameterizes the forward graphics model f based on vertices locations V , per-vertex brightness A and camera parameters C , which is shown on the left part of Figure 5.5, where U is for the $2D$ projected vertex coordinate position. Based on the auto-differentiation techniques, openDR provides a way to derive the derivatives of the rendered image observation with respect to the parameters in the rendering pipeline.

Approach We extend openDR in the following ways as illustrated in Figure 5.5: **1)** We parameterize the vertices locations V of CAD models by three parameters: azimuth θ , elevation ψ , and distance to the camera γ ; **2)** During the pose estimation procedure, as in (Aubry *et al.*, 2014a), the matching between the objects on real images and the rendered images from the CAD models are addressed by the similarities between the HOG descriptors of the visual discriminative elements extracted from them. The detailed procedure of extracting visual discriminative elements is discussed in (Aubry *et al.*, 2014a). In our method, we use our ∇ HOG method $\phi(P_f)$ for the image patches P_f which have the same regions as the visual elements P_I extracted from the test image I , and the similarity between the P_f and P_I is the dot product between HOG descriptors $\phi(P_I)$ of P_I and $\phi(P_f)$. As shown in the right part of Figure 5.5 this similarity can be traversed back to the pose parameters $\{\theta, \psi, \gamma\}$ and the derivatives of the similarity with respect to the pose parameters can be again computed by the auto-differentiation, our method can directly optimize to maximize the similarity to estimate the poses.

	Method	cross correlation	mutual information	structural similarity (Wang <i>et al.</i> , 2004)
	BoVW (Kato and Harada, 2014)	0.287	1.182	0.252
UoCTTI HOG	HOGgles (Vondrick <i>et al.</i> , 2013)	0.409	1.497	0.271
	CNN-HOG (Mahendran and Vedaldi, 2015)	0.632	1.211	0.381
	CNN-HOGb (Mahendran and Vedaldi, 2015)	0.657	1.597	0.387
	our ∇ HOG (single scale)	0.760	1.908	0.433
Dalal's HOG	our ∇ HOG (single scale)	0.170	1.464	0.301
	our ∇ HOG (multi-scale: $\frac{1}{64}$)	0.058	1.444	0.121
	our ∇ HOG (multi-scale: $\frac{1}{16}$)	0.076	1.470	0.147
	our ∇ HOG (multi-scale: $\frac{1}{4}$)	0.108	1.458	0.221
	our ∇ HOG (multi-scale: $\frac{1}{1}$)	0.147	1.478	0.293
	our ∇ HOG (multi-scale-more: $\frac{1}{64}$)	0.147	1.458	0.251
	our ∇ HOG (multi-scale-more: $\frac{1}{16}$)	0.191	1.502	0.291
	our ∇ HOG (multi-scale-more: $\frac{1}{4}$)	0.220	1.565	0.320
	our ∇ HOG (multi-scale-more: $\frac{1}{1}$)	0.236	1.582	0.338

Table 5.2: Comparison on the performance of reconstruction from feature descriptors.

Setup We follow the same experimental setting as (Aubry *et al.*, 2014a), where we test on the images annotated with no-occlusion, no-truncation and not-difficult of the chairs validation set on PASCAL VOC 2012 dataset (Everingham *et al.*, 2015), therefore in total 247 chairs from 179 images are used for the evaluation. To purely focus on evaluation of the pose estimation, we extract the object images based on their bounding boxes annotation, and resize them to have at least length of 100 pixels on the shortest side of image size.

The baseline (Aubry *et al.*, 2014a) is applied on the chair images to search over a chair CAD database of 1393 models which includes the rendered images from 62 different poses relative to camera for each of them, then to detect the chairs, match the styles of the chairs, and simultaneously recover their poses based on rendered images. We select the most confident detection for each chair together with the estimated pose.

We apply our proposed method on pose estimation by using the elevation and azimuth estimates of (Aubry *et al.*, 2014a) as a initialization of pose, and add few more initializations for azimuth (8 equidistantly distribute over 360°). We use gradient descent method with momentum term for optimization in order to optimize for the azimuth parameter and interleave iterations in which we additionally optimize for the distance to camera. In Figure 5.4 we visualize an example of the similarity

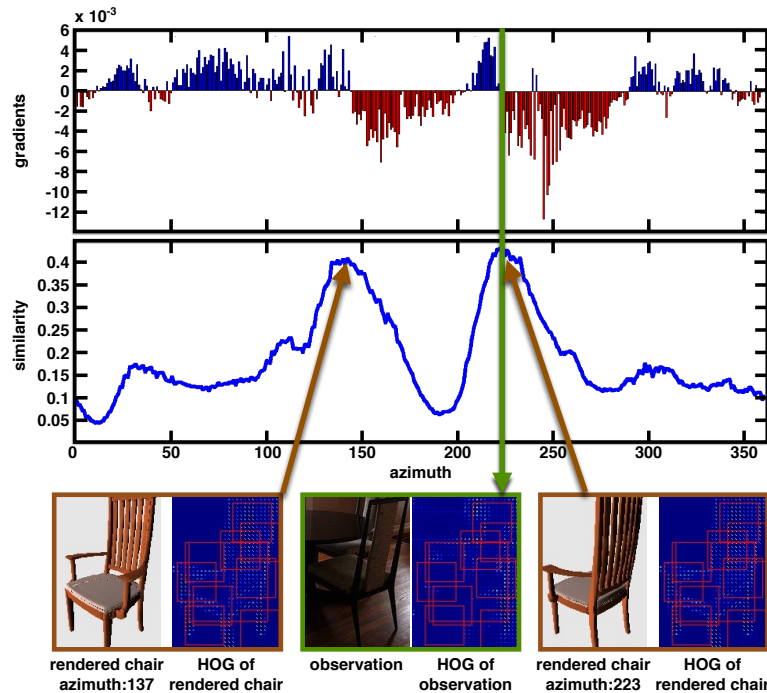


Figure 5.4: Visualization of the similarity and its gradients w.r.t azimuth. The red boxes on the HOG representations are the visual discriminative patches.

between the chair object on the real image and the CAD model on the rendered image, as well as its gradients w.r.t azimuth θ (full 360°). We can see how gradients change related to different local maximums and the corresponding poses of the CAD model.

Results In order to quantify our performance on pose estimation task, we use the continuous 3D pose annotations from PASCAL_{3D+} dataset (Xiang *et al.*, 2014). Following the same evaluation scheme, the view-point estimation is considered to be correct if its estimated viewpoint label is within the same interval of the discrete viewpoint space as the ground-truth annotation, or its difference with ground-truth in continuous viewpoint space is lower than a threshold. We evaluate the performance based on various settings of the intervals and thresholds in viewpoint space: $\{4 \text{ views}/90^\circ, 8 \text{ views}/45^\circ, 16 \text{ views}/22.5^\circ, 24 \text{ views}/15^\circ\}$. In Table 5.3 we report the performance numbers for Aubry’s baseline and our proposed approach. We are outperforming the previous best performance up to 10% points on the coarse and fine measures. Some example results which show improvements of the baseline method are shown in Table 5.4.

Discussion One advantage of our proposed method is that we are able to parameterize the vertex coordinates of the CAD models by the same pose parameters as used in (Aubry *et al.*, 2014a), then the differentiable rendering procedure provided

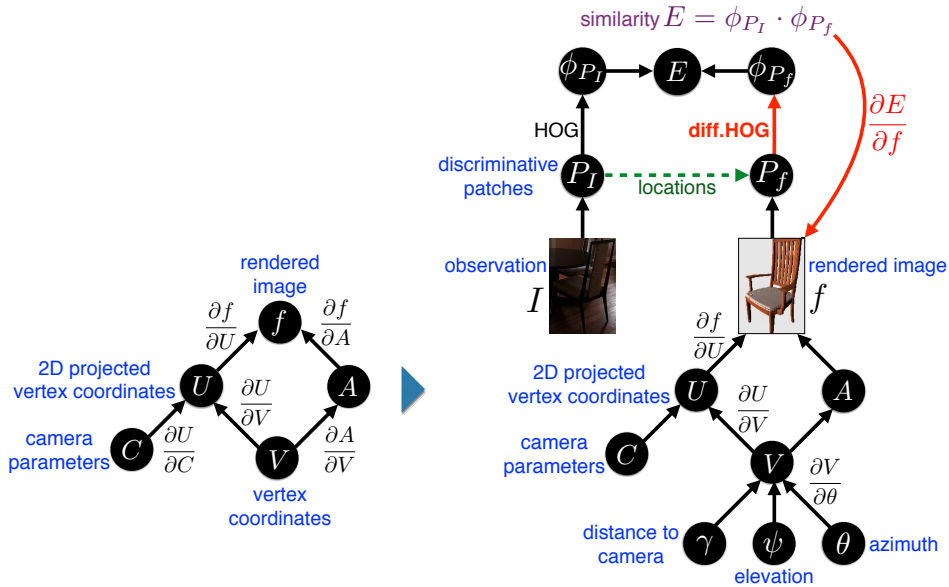


Figure 5.5: (left) The differentiable rendering procedure from openDR (Loper and Black, 2014). (right) The visualization of our model for pose estimation.

	4 views	8 views	16 views	24 views
Aubry et al. (Aubry <i>et al.</i> , 2014a)	47.33	35.39	20.16	15.23
our method	58.85	40.74	22.22	16.87

Table 5.3: Pose estimation results based on the groundtruth annotation from PASCAL3D+ (Xiang *et al.*, 2014).

by openDR (Loper and Black, 2014) and our ∇ HOG representations enable us to directly compute the derivatives of the similarity with respect to the pose parameters, and optimize for continuous pose parameters. In another word, for the proposed approach we do not need to discretize the parameters as (Aubry *et al.*, 2014a) and do not need to render images from many poses in advance for the alignment procedure either.



Table 5.4: Example results for pose estimation.

5.5 CONCLUSIONS

We investigate the feature extraction pipeline of HOG descriptor and exploit its piecewise differentiability. Based on the implementation using auto-differentiation techniques, the derivatives of the HOG representation can be directly computed. We study two problems of image reconstruction from HOG features and HOG-based pose estimation while the direct end-to-end optimization becomes practical with our ∇ HOG. We demonstrate that our ∇ HOG-based approaches outperforms the state-of-the-art baselines for both problems. We have demonstrated that the approach can lead to improved introspection via visualizations and improved performance via direct optimization through a whole vision pipeline. Our implementation is integrated into an existing auto-differentiation package as well as the recently proposed Approximately Differentiable Renderer OpenDR (Loper and Black, 2014) which are both publicly available. It is easy to adopt to new tasks and is applicable to a range of end-to-end optimization problems.








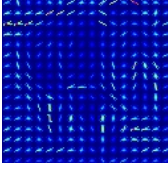
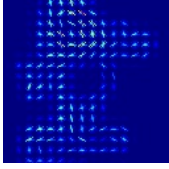



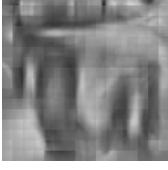

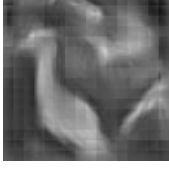

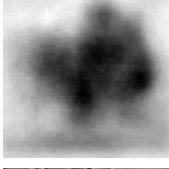






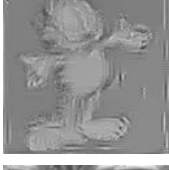





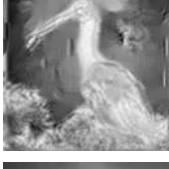






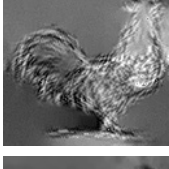
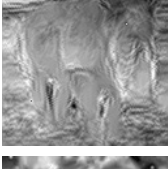

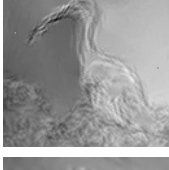
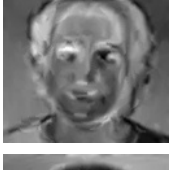
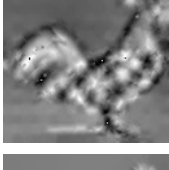

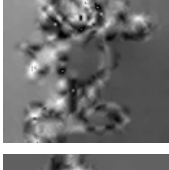
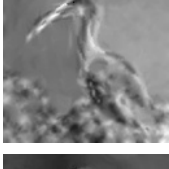


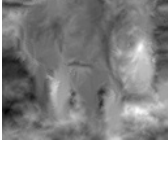


Example					
HOG					
BOVW					
HOGgles UoCTTI-HOG					
CNN-HOG UoCTTI-HOG					
CNN-HOGb UoCTTI-HOG					
Our ∇ HOG (single-scale) UoCTTI-HOG					
Our ∇ HOG (single-scale) Dalal-HOG					
Our ∇ HOG (multi-scale) Dalal-HOG					
Our ∇ HOG (multi-scale-more) Dalal-HOG					

Table 5.5: Example results for image reconstruction from feature descriptors.

Contents

6.1	Introduction	78
6.2	Consumer Stereo Video Segmentation Challenge (CSVSC)	80
6.3	Efficient Adaptive Segmentation of Stereo Videos	81
6.4	Efficient Segmentation Ensemble Model	82
6.4.1	Unifying Graph	82
6.4.2	Improved Efficiency with Graph Reduction	83
6.4.3	Details to Derive the Reduced Graph \mathcal{G}^Q	84
6.4.4	Implementation Details	86
6.5	Performance-Driven Adaptive Combination	87
6.5.1	Adaptive Combination by Regression	87
6.5.2	Performance-Driven Regressor Learning by Differentiable Proxies	87
6.5.3	Implementation Details	90
6.6	Experimental Results	90
6.6.1	Video Segmentations and Their (Static) Ensemble	91
6.6.2	EASVS and the State-of-the-art	91
6.6.3	Deeper Analysis of EASVS	92
6.7	Conclusions	95

IN this chapter we study a segmentation task on the consumer stereo video data, which is a rapidly increasing data type, still largely unexplored, and encompasses rich information of appearance, motion and depth cues. As recent progress in segmentation methods have yield a diverse set of approaches from image to video scenarios, different methods have their own heuristics to combine available cues though no clear winning method method can perform best in all test conditions. Therefore we propose an ensemble method that learns a data-dependent combination scheme to dynamically weight different cues as well as candidate segmentation algorithms in order to maximize the performance metric of segmentation.

In brief, we first propose a new benchmark: videos, annotations and metrics to measure progress on the emerging challenge of segmenting consumer stereo videos. Second, several state of the art segmentation methods as well as a static segmentation combination scheme are evaluated to show the need of adaptive combination scheme in a data-dependent manner. We propose a parametrized similarity graph based on the overlapped superpixels to aggregate the various feature

cues and segmentation information from various segmentation algorithms which can be utilized by spectral clustering technique to produce the final segmentation. Finally, we propose and integrate into this model a novel regressor, learnt to optimize the stereo segmentation performance directly via a differentiable proxy. The regressor makes our segmentation ensemble *adaptive* to each stereo video and outperforms the segmentations of the ensemble as well as a state-of-the-art RGB-D segmentation technique.

6.1 INTRODUCTION

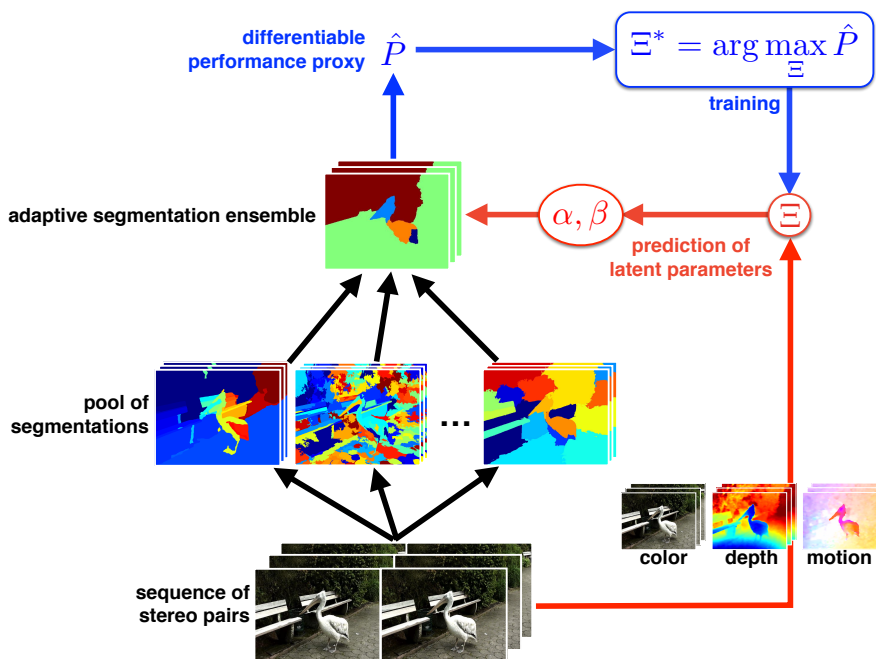


Figure 6.1: Overview of the proposed efficient adaptive stereo segmentation technique. Our proposed segmentation ensemble model leverages the best available image and video segmentation results efficiently. A regressor makes the ensemble model adaptive to each stereo video, based on color, depth and motion features. In our novel learning framework, the segmentation performance is optimized via a differentiable proxy.

We witness a fast growing number of stereo streams on the web, due to the advent of consumer stereo video cameras. Are we ready to exploit the rich cues which stereo videos deliver? Our work focuses on segmentation of such data sources, as it is a common pre-processing step for further analysis such as action (Le *et al.*, 2011; Oneata *et al.*, 2014; Taralova *et al.*, 2014) or scene classification (Raza *et al.*, 2013).

We propose a new consumer stereo video challenge, to understand the opportunities and foster the research in this new area. The new type of data combines the availability of appearance and motion with the possibility of extracting depth

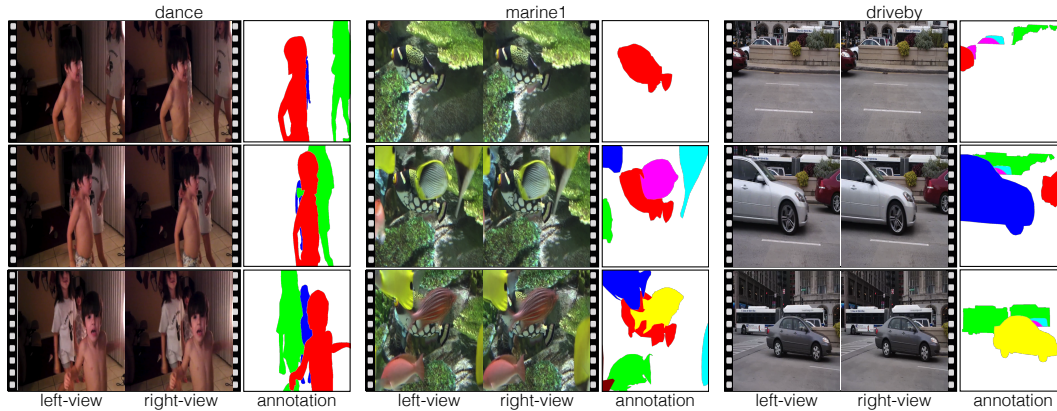


Figure 6.2: Sample frames from the Consumer Stereo Video Segmentation Challenge (CSVSC) dataset (left-right views) with the corresponding annotations. The stereo videos differ in content (appearance, motion, number and type of subjects) and in camera characteristics (intrinsic parameters, zoom, noise).

information. Considering consumer videos means addressing a most abundant web data, which is however also very heterogeneous, due to a variety of consumer cameras.

The new consumer stereo video challenge explicitly concerns the semantics of the video. A number of existing benchmarks have offered ground truth depth and motion, recurring to controlled recordings (Scharstein *et al.*, 2014) or computer graphics simulations (Butler *et al.*, 2012). By contrast, here we address stereo videos *in the wild* and specifically consider the semantics of the data. While this might partly harm analysis (no true depth available), it addresses directly what we are most interested in, the actors and objects in the videos.

We warm-start the challenge with a number of baselines, extending best available image and video segmentations to the consumer stereo videos and their available features, e.g. color, motion and depth. Most baselines perform well on some videos, however none performs well on all. As an example, motion segmentation techniques (Ochs and Brox, 2011) perform well while the object moves, but encounter difficulty with static video shots. On the other hand, camouflaged (but moving) objects impinge appearance-based image (Li *et al.*, 2012) and video (Grundmann *et al.*, 2010) segmentation techniques.

Thus motivated, we introduce in Section 6.4 a new efficient segmentation ensemble model, which leverages existing results where they perform best. Furthermore, we introduce in Section 6.5 the framework to learn a regressor which adapts the ensemble model to each particular stereo video. The proposed technique is overviewed in Section 6.3 and demonstrated in Section 6.6. Although *only* combining optimally existing results, our new algorithm outperforms a most recent RGB-D segmentation technique (Hickson *et al.*, 2014).

The topic of consumer stereo video segmentation is related to several research

areas, such as video segmentation, scene flow estimation, and image/video co-segmentation. The corresponding discussion of these related fields can be found in Section 2.

6.2 CONSUMER STEREO VIDEO SEGMENTATION CHALLENGE (CSVSC)

We launch a Consumer Stereo Video Segmentation Challenge (CSVSC). The new dataset consists of 30 video sequences which we have selected from Youtube based on their heterogeneity. In fact, the footage differs in the number of objects (2-15), the kind of portrayed actors (animals or people), the type of motion (a few challenging stop-and-move scenes and objects entering or exiting the scene), the appearance visual complexity (also in relation to the background, a few objects may be harder to discern) and the distance of the objects from the camera (varying disparity and thus depth). Not less importantly, we have selected videos acquired by different consumer stereo cameras, which implies diverse camera intrinsic parameters, zooms and (as a further challenge) noise degradations such as motion blurs and camera shake. We illustrate a few sample sequences in Figure 6.2.

Benchmark Annotation and Metrics.

We have gathered human annotations and defined metrics to quantify progress on the new benchmark. In particular, we equidistantly sample 5 frames from the left view of the videos to be labelled (for a total of 1738 frames). As for the metrics, we have considered state-of-the-art image (Arbelaez *et al.*, 2011) and video segmentation (Galasso *et al.*, 2013) metrics:

Boundary precision-recall (BPR). This reflects the per-frame boundary alignment between a video segmentation solution and the human annotations. In particular, BPR indicates the F-measure between recall and precision (Arbelaez *et al.*, 2011).

Volume precision-recall (VPR). This measures the video segmentation property of temporal consistency. As for BPR, VPR also indicates the F-measure between recall and precision (Galasso *et al.*, 2013).

It is of research interest to determine which of the metrics is best for learning. We answer this question in Section 6.6, where we consider BPR and VPR alone or combined by their: **arithmetic mean (AM-BVPR)** or **harmonic mean (HM-BVPR)**.

Preparation of Stereo Videos

Not having a ground truth depth may impinge comparison among techniques applied to the dataset. We define therefore an initial set of comparisons among depth-estimation algorithms and make the results available.

We have considered the per-frame rectification of (Fusiello and Irsara, 2008) and the stereo matching algorithm of (Geiger *et al.*, 2010), filling-in the missing correspondences with (Janoch *et al.*, 2011). Furthermore, we have estimated depth by the optical flow algorithm of (Zach *et al.*, 2007) between the right and left views.

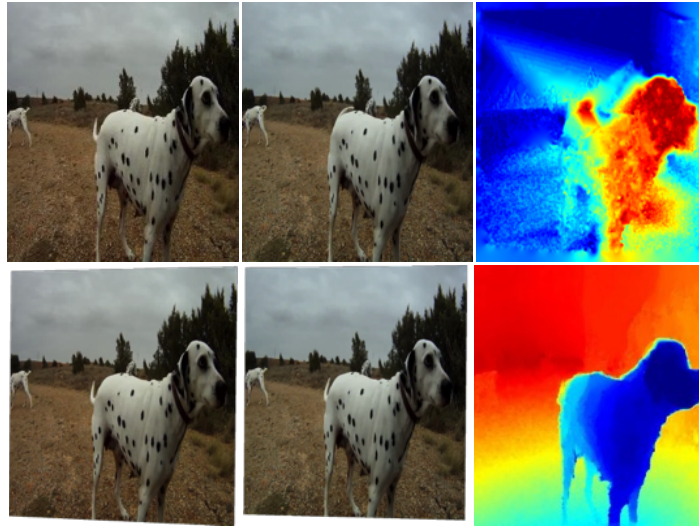


Figure 6.3: Sample disparity estimation. The first two columns are the original stereo pair and their rectified images. The top-right picture is the disparity map computed by (Geiger *et al.*, 2010), the bottom-right is the depth map obtained by optical flow (Zach *et al.*, 2007) between the left and right view.

We illustrate samples in Figure 6.3. Our initial findings are that estimating depth by optical flow leads to best downstream stereo segmentation outputs, which we use therefore in the rest of the chapter.

6.3 EFFICIENT ADAPTIVE SEGMENTATION OF STEREO VIDEOS

We warm start the CSVSC challenge with a basic segmentation ensemble model. To this purpose we first pre-process the stereo videos with a pool of state-of-the-art image and video segmentation algorithms. Then we combine the segmentation outputs with a new efficient segmentation ensemble model (cf. Section 6.4). Finally, we propose the learning framework to adapt the combination parameters of each stereo video (cf. Section 6.5).

Figure 6.1 gives an overview of our ensemble model:

Pool of Image and Video Segmentations. We select most recent algorithms which are available online. These are used to segment the single frames (image segmentations) and the left views of the stereo videos (video segmentations). This results in a pool of segments which are respectively superpixels and supervoxels.

Efficient Segmentation Ensemble Model. We bring together the pool of segments and connect them to the stereo video voxels. The segmentation ensemble model is represented by a graph and parameterized by α and β 's, which weight the contribution of each segmentation method. The model is accurate (voxel-based) but costly. We propose therefore an efficient graph reduction which is exact, i.e. it

provides the same solutions as the voxel-based at a lower computational complexity. **Performance-Driven Adaptive Combination.** We compute stereo video features from the stereo videos based on color, flow and depth. From these features, we regress the combination parameters α and β 's, i.e. we combine optimally the pooled segmentation outputs. To this purpose, we propose a novel regressor Ξ and an inference procedure, to learn from data the optimal regression parameters ζ . For the first time in literature, the regressor parameters ζ are directly optimized according to the final performance measure P (resulting from the graph partitioning and the metric evaluation, cf. Section 6.5.2). We achieve this with a novel differentiable performance proxy \hat{P} .

None of the state-of-the-art segmentation algorithms performs well with all of the challenging consumer stereo videos (cf. experiments in Section 6.6). Both the contributions on the ensemble model (Section 6.4) and the performance-driven adaptive combination (Section 6.5) turn out important for better results.

6.4 EFFICIENT SEGMENTATION ENSEMBLE MODEL

We propose a graph for bringing together the available video segmentation outputs. Additionally, we propose the use of recent spectral techniques to reduce the voxel graph to one based on tailored superpixels/supervoxels, to improve efficiency without any (proven) compromise on performance. The graph partitioning with spectral clustering provides the segmentation output.

6.4.1 Unifying Graph

Given a number of video segmentation outputs, we propose to bring all of them together by defining an unifying graph.

Let us consider Figure 6.4 *left*. Each video segmentation algorithm provides groupings of the video sequence voxels. In the unifying graph, each pixel is therefore linked to the groupings to which it belongs. For example, one algorithm may compute spatio-temporal tubes (supervoxels) (Grundmann *et al.*, 2010), another one may compute image-based superpixels (Arbelaez *et al.*, 2011). The video sequence voxels would then be linked to the tube to which they belong (temporally) and to their superpixels (spatially). Altogether, the outputs from the pool of video segmentation algorithms provide hypotheses of grouping for the video voxels.

More formally, we define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to jointly represent the video and the segmentation outputs. Nodes from the vertex set \mathcal{V} are of two kinds:

Voxels are the video sequence elements which we aim to segment;

Pooled Segmentation Outputs are the computed spatial- and/or temporal-groupings, providing voxel grouping hypotheses.

Further to being connected to the voxels, the pooled groupings from the same output are also connected to their neighbors, which defines the video volume structure. Edges are therefore of two types:

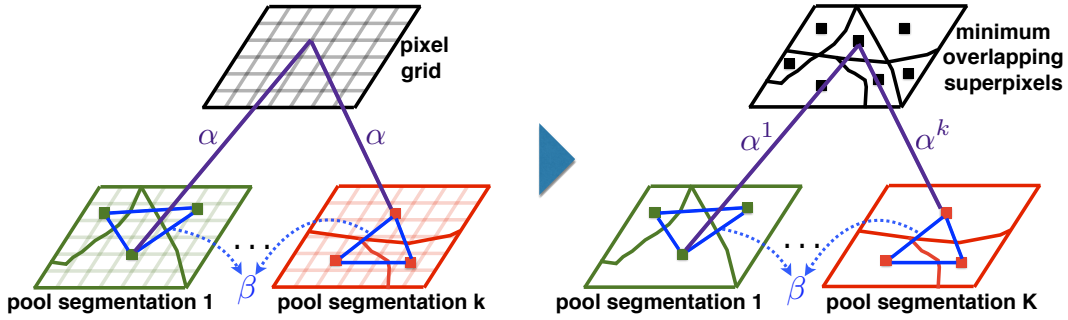


Figure 6.4: Proposed video segmentation model. A number of K pooled image and video segmentation outputs are brought together as hypotheses of grouping for the considered video sequence (cf. Section 6.4 for details). We propose to replace the model of (Li *et al.*, 2012) (*left*) with a new one (*right*) based on minimally overlapping superpixels, which is provably equivalent but yields better efficiency (cf. Section 6.4.1)

β -Edges are between the groupings of each pooled segmentation k ; we assume C features (appearance, motion, etc. cf. Section 6.4) and distances based on β^c -weighted features: $w_{I,J}^k = e^{-(\beta^1 d_{I,J}^{k1} + \dots + \beta^C d_{I,J}^{kC})}$, where $d_{I,J}^{kc}$ is the distance between superpixels I and J from the k -th pooled output based on c -th feature.

α -Edges are between the voxels and the grouping that it belongs to. The α^k 's encode the trust towards the respective K segmentation algorithm, ideally proportional to its accuracy.

Partitioning graph \mathcal{G} with spectral clustering is computationally demanding as the number of nodes (and edges) depends linearly on the video voxels. The theory of (Li *et al.*, 2012) reduces the complexity of a first stage of spectral clustering (the eigendecomposition) but not of the second one (k-means), still of linear complexity in the number of voxels (and thus bottleneck of (Li *et al.*, 2012)). We address both with graph reduction in the following Section.

6.4.2 Improved Efficiency with Graph Reduction

Let us consider again Figure 6.4. A huge number of voxels are similar both in appearance and in motion and are therefore grouped in all segmentation outputs. When partitioning the original graph \mathcal{G} , these voxels are always segmented together. (The trivial proof leverages their equal edges and therefore eigenvectors).

Rather than considering all voxels, we propose to reduce the original graph \mathcal{G} to one of smaller size \mathcal{G}^Q which is equivalent (provides exactly the same clustering solutions). In \mathcal{G}^Q , we basically group all those voxels with equal connections into super-nodes (reweighting their edges equivalently). This reduces the algorithmic complexity, as the spectral clustering (both the eigendecomposition and the k-means) now depends only on the number of super-nodes (which is determined for most pooled segmentation algorithms by the number of objects, rather than voxels).

We identify the voxels with equal connections by intersecting the available segmentation outputs. The result of the intersection is an oversegmentation into superpixels, which can generate all pooled segmentation outputs by merging. We name these *minimally overlapping superpixels*.

More formally, the reduced graph $\mathcal{G}^Q = (\mathcal{V}^Q, \mathcal{E}^Q)$ takes the minimally overlapping superpixels as nodes and the following edge weights

$$w_{IJ}^Q = \begin{cases} \sum_{i \in I} \sum_{j \in J} w_{ij} & \text{if } I \neq J \\ \frac{1}{|I|} \sum_{i \in I} \sum_{j \in J} w_{ij} - \frac{(|I| - 1)}{|I|} \sum_{i \in I} \sum_{j \in \mathcal{V} \setminus I} w_{ij} & \text{if } I = J \end{cases} \quad (6.1)$$

where $|\cdot|$ indicates the number of voxels within the superpixel and I, J are two minimally overlapping superpixels. According to (6.1), two pixels i and j are reduced if belonging to the same superpixel, i.e. if $I = J$. (Since the superpixel connections are equal for the pixels within the same superpixel by construction, the reduction is exact, cf. (Galasso *et al.*, 2014).)

6.4.3 Details to Derive the Reduced Graph \mathcal{G}^Q

In order to express the pairwise affinities between minimally overlapping superpixels in terms of the parameters α and β , we show the details of deriving the reduced graph \mathcal{G}^Q step by step in the following paragraphs.

Review of Transfer-Cut Li *et al.* (2012) By the use of *transfer-cut*, Li *et al.* (2012) connects the pool of segmentations to the pixels of the image (the work is originally defined for image segmentation). In the original procedure, given K pooled segmentation results composed of N^k superpixels, the pairwise distance matrix between superpixels from the k -th pooled result is given by:

$$A^k = \begin{bmatrix} w_{1,1}^k & w_{1,2}^k & \cdots & w_{1,N^k}^k \\ w_{2,1}^k & w_{2,2}^k & \cdots & w_{2,N^k}^k \\ \vdots & \vdots & w_{I,J}^k & \vdots \\ w_{N^k,1}^k & w_{N^k,2}^k & \cdots & w_{N^k,N^k}^k \end{bmatrix} \quad (6.2)$$

Let us define then a block matrix S_Y to stack all pairwise distance matrices from all K algorithms:

$$S_Y = \begin{bmatrix} A^1 & & & \\ & A^2 & & \\ & & \ddots & \\ & & & A^K \end{bmatrix} \quad (6.3)$$

which will have size $N_{sp} \times N_{sp}$, where $N_{sp} = \sum^K N^k$.

Let us then define the distance matrix S_{XY} between the superpixel i and the pixel j for all algorithms as:

$$S_{XY}(i, j) = \alpha^k, \text{ if pixel } j \in \text{ superpixel } i \text{ from algorithm } k \quad (6.4)$$

where its size is in size of $N_{sp} \times N_p$, and N_p is the total number of pixels.

Then the total distance matrix H with size $(N_p + N_{sp}) \times N_{sp}$ can be written in

$$H = [S_{XY} \mid S_Y]^\top \quad (6.5)$$

Let us define

$$D_X = \mathbf{diag}(H1) \quad (6.6)$$

and the graph between all the superpixels from pooled segmentations by transfer-cut:

$$W_Y = H^\top \cdot D_X^{-1} \cdot H \quad (\text{size: } N_{sp} \times N_{sp}) \quad (6.7)$$

To look into details, the parametric forms of elements in W_Y can be written into:

$$\text{diagonal} \Rightarrow \sum \frac{(\alpha^k)^2}{\sum^K \alpha^k} + \sum_{J=1}^{N^k} \frac{(w_{I,J}^k)^2}{\sum_{Z=1}^{N^k} w_{J,Z}^k} \quad (6.8)$$

where M is the number of pixels \in superpixel I ,
and superpixel $I \in k$ -th pool.

$$\text{off-diagonal} \Rightarrow \begin{cases} \sum \frac{(\alpha^k)^2}{\sum^K \alpha^k} + \sum_{Z=1}^{N^k} \frac{w_{I,Z}^k \cdot w_{Z,J}^k}{\sum_{L=1}^{N^k} w_{Z,L}^k} \\ \quad \text{if both superpixels } I \text{ and } J \in k\text{-th pool.} \\ \sum \frac{\alpha^k \cdot \alpha^{k'}}{\sum^K \alpha^k} \\ \quad \text{if superpixel } I \in k\text{-th pool} \\ \quad \text{and } J \in k' \neq k\text{-th pool.} \end{cases} \quad (6.9)$$

where M is the number of pixels $\in I \cap J$

Extend Transfer-Cut to Stereo Videos Let us extend the theory in the previous paragraph to stereo videos (modeled on the left view frames) and voxels, which implies changing the elements of matrices S_{XY} and S_Y from pixels and pooled superpixels to minimally overlapping superpixels.

First, the β edges used to denote the similarities between the pooled superpixels I, J is expanded into the similarities between minimally overlapping superpixels i, j by the graph expansion approach Agarwal *et al.* (2006):

$$w_{ij}^k = \begin{cases} w_{I,J}^k & \text{if } I \neq J, i \in I, j \in J \\ \sum_{Z \neq I} w_{I,Z}^k & \text{if } i, j \in I, i \neq j \\ 0 & \text{if } I = J \end{cases} \quad (6.10)$$

Then we have the parametric forms of the elements for H' as:

$$\begin{aligned} \text{diagonal} &\Rightarrow \frac{(\alpha^k)^2}{\sum^K \alpha^k} + \sum_{j=1}^{N^m} \frac{(w_{ij}^k)^2}{\sum_{z=1}^{N^m} w_{j,z}^k} \\ \text{off-diagonal} &\Rightarrow \frac{(\alpha^k)^2}{\sum^K \alpha^k} + \sum_{z=1}^{N^m} \frac{w_{iz}^k \cdot w_{zj}^k}{\sum_{l=1}^{N^m} w_{zl}^k} \end{aligned} \quad (6.11)$$

where N^m is number of minimally overlapping superpixels. The reduced graph \mathcal{G}^Q is then easily to derived from H' by the graph reduction Galasso *et al.* (2014) to group the edges of identical minimally overlapping superpixels in H' . And we denote the elements in \mathcal{G}^Q by w_{ij}^Q . (Please note that we can also build up the graph H' based on the voxels as nodes then reduced to \mathcal{G}^Q , which will follow the story of Equation 6.1. Here we directly use minimally overlapping superpixels in H' for clarity.)

6.4.4 Implementation Details

The output segmentation is obtained by graph partitioning \mathcal{G}^Q with spectral clustering (Ng *et al.*, 2002; Shi and Malik, 2000; von Luxburg, 2007). In particular, the labels of the minimally overlapping superpixels provide the voxel labels and thus the video segmentation solution.

In this work, we use $K = 6$ image and video segmentation algorithms: (1.) The hierarchical image segmentation of (Arbelaez *et al.*, 2011). We choose one layer from hierarchy based on best performance on a validation set. We take three segmentation outputs by applying the Simple Linear Iterative Clustering (SLIC) (Achanta *et al.*, 2010) respectively on (2.) depth, (3.) optical-flow (Zach *et al.*, 2007) and the (4.) LAB-color coded cues, bilaterally filtered for noise removal and edge preservation (Zhang *et al.*, 2014b). (5.) Hierarchical graph-based video segmentation (GBH) (Grundmann *et al.*, 2010). We choose one layer from the hierarchy on the validation set. (6.) The motion segmentation technique (moseg) of (Ochs and Brox, 2011).

While the features are computed on the stereo video. The graph is constructed on one of the two views (the left one) of the stereo videos, which is then evaluated for the segmentation quality. The contribution of segmentation outputs is weighted by α . β defines the affinities between superpixels/supervoxels from the same pooled segmentation output, weighting $C = 3$ feature cues based on mean Lab-color, depth and motion.

Note the importance of α 's and β 's in the graph \mathcal{G} and therefore \mathcal{G}^Q . These parameters define how much each pooled segmentation output is trusted and how to compute the similarity among superpixels/supervoxels in these outputs. Such parameters can be defined statically (cf. (Li *et al.*, 2012)) or adjusted dynamically in a data-dependent fashion, as we propose in the next Section.

6.5 PERFORMANCE-DRIVEN ADAPTIVE COMBINATION

We propose a regressor Ξ to estimate the optimal segmentation ensemble parameters α and β from the appearance-, motion- and depth-based features of the stereo videos. (Cf. Figure 6.1 where the regressor is given by the red arrows.) Furthermore, we propose a novel inference framework to learn the regressor parameters ξ from the training stereo videos. (Cf. Figure 6.1 where the training is represented with blue arrows.) A new differentiable performance proxy \hat{P} enables optimization driven by the stereo video segmentation performance measure P .

6.5.1 Adaptive Combination by Regression

Let us define a regressor Ξ , with parameters ξ . Ξ takes as input a set of features \mathcal{F} computed from the stereo video and outputs the parameters α and β for the ensemble segmentation model (i.e. the coefficients to optimally combine K segmentation outputs from the pool based on C features, cf. 6.4.1). Intuitively, the regressor should select the best segmentation outputs from the pool, based on the stereo video content. This would imply, for example, a larger trust towards image- rather than motion-segmentation outputs, for those stereo videos where no motion is present.

We employ a second order regressor Ξ which we parameterize by a matrix B . Overall, α and β are computed as:

$$(\alpha^1, \dots, \alpha^K, \beta^1, \dots, \beta^C) = \Xi(\mathcal{F}; \xi) = \mathcal{F}^T B \mathcal{F} \quad (6.12)$$

We consider in \mathcal{F} features based on appearance, motion and depth. A large feature set is important to allow the regressor to understand the type of stereo video (dynamic, static, textured etc.) For each feature, we compute therefore histograms, means, medians, variances and entropies. We would leave the learning framework to choose from the right feature, i.e. training the best regressor Ξ . This should ideally consider the system performance P for optimization or the tractable differentiable proxy which we discuss next.

6.5.2 Performance-Driven Regressor Learning by Differentiable Proxies

Let us consider Figure 6.1. The α and β , regressed by Ξ according to features \mathcal{F} , correspond to a stereo video segmentation performance P . During *training*, we seek to optimize Ξ for the maximum segmentation performance P :

$$\hat{\Xi} = \max_{\Xi} P(\Xi(\mathcal{F})) \quad (6.13)$$

There are two main obstacles to our goal. First, typical video segmentation performance metrics are not differentiable and therefore do not lend themselves to directly optimizing an overall performance. To address this, we propose a differentiable performance proxy \hat{P} in Section 6.5.2.1.

Second, α and β are not part of the objective (6.13) and have to be considered *latent*. In Section 6.5.2.2, we define therefore an EM-based strategy to jointly learn the regressor Ξ , α and β . An overview of our training procedure is given in Algorithm 1.

Algorithm 1 Joint learning of the regressor Ξ and the latent combination weights α, β

Require: \forall training videos with initial set of parameter combinations (α, β) and stereo video features \mathcal{F}

- 1: **repeat**
 - 2: Given the current estimates of (α, β) ,
 - 3: train the Ξ which regresses them from \mathcal{F}
 - 4: **for all** training video **do**
 - 5: predict $(\alpha'', \beta'') = \Xi(\mathcal{F})$;
 - 6: use (α'', β'') as initialization for
 $(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} \hat{P}(\alpha, \beta)$;
 - 7: update (α, β) for the training video by $(\hat{\alpha}, \hat{\beta})$;
 - 8: **end for**
 - 9: **until** Convergence or max. iterations exceeded
-

6.5.2.1 Metric Specific Performance Proxy

In image segmentation, performance is generally measured by boundary precision recall (BPR) and its associated best F-measure (Arbelaez *et al.*, 2011). In video segmentation, benchmarks additionally include volume precision recall (VPR) metrics (Galasso *et al.*, 2013). Both these performance measures are plausible P , but neither of them is differentiable, which complicates optimization. (We experiment on various performance measures in Section 6.6.)

We propose to estimate a differentiable performance proxy \hat{P} which approximates the true performance P . We do so by a second order approximation parameterized by the matrix Y . Taking χ a vector of features which are sufficient to represent the stereo video (at least as far as the estimation of (α, β) is concerned) we have:

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} \hat{P}(\alpha, \beta) = \arg \max_{\alpha, \beta} \chi^\top Y \chi \quad (6.14)$$

We perform training by sampling α and β , computing then vector χ and finally fitting the parameter matrix Y .

Stereo Video Representation by Spectral Properties. We are motivated by prior work on supervised learning in spectral clustering (Meilă *et al.*, 2005; Jordan and Bach, 2004; Ionescu *et al.*, 2015) to represent the stereo videos by their spectral properties. In particular, we draw on (Meilă *et al.*, 2005) and consider the normalized-cut cost NCut (of the similarity graph W which is the reduced graph \mathcal{G}^Q as shown

in Section 6.4.3, based on the training set groundtruth labelling) and its lower bound Trace_R . Our representation vector is therefore $\chi = [\alpha, \beta, \text{NCut}, \text{Trace}_R]^\top$.

In more details, given the indicator matrix $E = \{e_r\}_{r=1\dots R}$ where $e_r \in \mathbb{R}^{N^m}$, $e_r(i) = 1$ if superpixel i belongs to r -th cluster otherwise $= 0$, and N^m is the number of superpixels, we have:

$$\begin{aligned} \text{NCut}(\alpha, \beta, E) &= \sum_{r=1}^R \frac{e_r^\top (D - W) e_r}{e_r^\top D e_r} \\ \text{Trace}_R(\alpha, \beta) &= R - \sum_{r=1}^R \lambda_r(L) \end{aligned} \quad (6.15)$$

where $D = \mathbf{diag}(W\mathbf{1})$ is the degree matrix of W and $\lambda_r(L)$ is the r -th eigenvalues of the generalized Laplacian matrix $L = D^{-1} \cdot W$ of the similarity matrix W .

Derivatives of Performance Proxy. Our performance proxy \hat{P} is now differentiable. For gradient descent optimization, we use its derivatives w.r.t. parameters $\theta \in \{\alpha^k, \beta^c\}$:

$$\frac{\partial \chi^\top Y \chi}{\partial \theta} = \frac{\partial \chi^\top}{\partial \theta} (Y + Y^\top) \chi \quad \forall \theta \in \{\alpha^k, \beta^c\} \quad (6.16)$$

The derivatives of NCut and $\text{Trace}_R \in \chi$ are:

$$\begin{aligned} \frac{\partial(\text{NCut})}{\partial \theta} &= \sum_{r=1}^R \frac{-e_r^\top \frac{\partial W}{\partial \theta} e_r e_r^\top D e_r + e_r^\top W e_r e_r^\top \frac{\partial D}{\partial \theta} e_r}{(e_r^\top D e_r)^2} \\ \frac{\partial(\text{Trace}_R)}{\partial \theta} &= \text{trace}(V^\top \frac{\partial L(\theta)}{\partial \theta} V) \end{aligned} \quad (6.17)$$

where V denotes the subspace spanned by the first R eigenvectors of L .

According to the formulations shown in Section 6.4.3, the derivatives $\frac{\partial w_{ij}^Q}{\partial \alpha^k}$ and $\frac{\partial w_{ij}^Q}{\partial \beta^c}$ for entries of $W (= \mathcal{G}^Q)$ can be derived by a sequence of chain rules. The degree matrix D is diagonal matrix, where we can represent its elements on the diagonal by:

$$D_{II} = \sum_{J=1}^{N^m} w_{I,J}^Q \quad (6.18)$$

Then the derivatives of D 's elements are:

$$\begin{aligned} \frac{\partial D_{II}}{\partial \alpha^k} &= \sum_{J=1}^{N^m} \frac{\partial w_{I,J}^Q}{\partial \alpha^k} \\ \frac{\partial D_{ii}}{\partial \beta^c} &= \sum_{J=1}^{N^m} \frac{\partial w_{I,J}^Q}{\partial \beta^c} \end{aligned} \quad (6.19)$$

Finally the derivative of generalized Laplacian matrix $L = D^{-1} \cdot W$ in equation 6.17 is given by the chain rule:

$$\frac{\partial L}{\partial \theta} = D^{-1} \cdot \frac{\partial W}{\partial \theta} + \frac{\partial D^{-1}}{\partial \theta} \cdot W \quad (6.20)$$

6.5.2.2 Joint Learning of Regressor and Latent Parameter Combinations

As stated in Equation 6.13, we are interested in optimizing the performance P w.r.t. the regressor Ξ and therefore the ensemble combination parameters α and β have to be treated as latent variables. As described in Algorithm 1, we solve this by an EM-type optimization scheme in which we iterate finding optimal parameters α and β and predicting new α and β parameters based on the re-fitted regressor Ξ .

Intuitively, this scheme strikes a balance between the generalization capabilities of the regressor and optimal parameters α and β . We found this to be particularly important, as in many cases a wide range of parameters leads to good results. Fixing the best parameters as a learning target, leads to a more difficult regression and overall worse performance. The metric specific performance proxy is continuously updated by using the samples in a small neighborhood in order to improve the local approximation of the desired metric P .

6.5.3 Implementation Details

As already noted, the computation of NCut at training involves the ground truth annotations. In particular, the NCut for the entire video requires all frames labeled, while ours and most segmentation datasets (Galasso *et al.*, 2013; Ochs *et al.*, 2014) only offer sparse labeling. Aggregating dense optical flow over time allows to connect the sparsely annotated frames. The spatial and temporal connections of these labeled frames are then used for the NCut computation.

Our representation vector χ in (6.14) consists of $[\alpha, \beta, \text{NCut}, \text{Trace}_R]$. We have empirically found that this combination improves of the individual parts and subsets by 5% and therefore we use the full vector in the following experiments. In order to increase the number of examples for our training procedure, we divide each video into subsequences so that each of them contains two frames with groundtruth.

6.6 EXPERIMENTAL RESULTS

We evaluate our proposed **efficient and adaptive stereo video segmentation** algorithm (EASVS) on the CSVSC benchmark. In particular, first we test the pooled segmentation outputs, then we compare EASVS against relevant state-of-the-art on stereo video sequences, finally we present an in-depth analysis of EASVS.

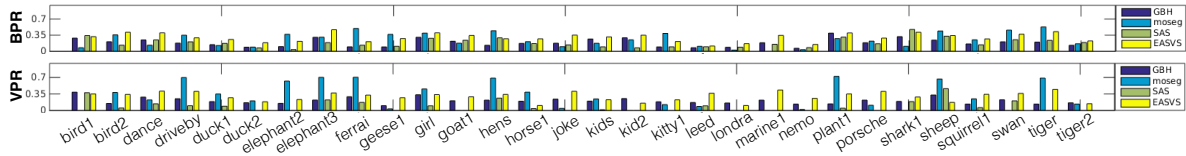


Figure 6.5: Results of the considered video segmentation algorithms (GBH (Grundmann *et al.*, 2010), moseg (Ochs and Brox, 2011), SAS (Li *et al.*, 2012)) and our proposed EASVS on the CSVSC stereo video sequences using BPR and VPR. Both in terms of boundaries and videos, no considered method performs consistently well on all videos. moseg may achieve high performance of stereo videos with large and distinctive motion such as "elephants3" and "hens" but underperforms when motion is not strong, e.g. "marine1". Complementary features are given by GBH. SAS combines statically (cf. segmentation ensemble model of Section 6.4.1) the two video segmentation techniques as well as the pooled image segments but also underperforms, because a static combination cannot address the variety of the stereo videos.

6.6.1 Video Segmentations and Their (Static) Ensemble

Among the pooled segmentations (details in Section 6.4.4), we have included two state-of-the-art video segmentation techniques: the motion segmentation algorithm of (Ochs and Brox, 2011) (moseg) and the graph-based hierarchical video segmentation method of (Grundmann *et al.*, 2010) (GBH).

In Figure 6.5, we illustrate performance of each of moseg and GBH on all stereo video sequences. (Cf. detailed comments in the figure caption.) As expected, none of the two performs satisfactorily on all sequences. Rather, they have in most cases complementary performance, moseg taking the lead on sequences with evident motion and GBH overtaking when spatio-temporal appearance cues are more peculiar in the visual objects.

A third technique illustrated in Figure 6.5 is the segmentation by aggregating superpixel method of (Li *et al.*, 2012) (SAS). This is an interesting baseline for our proposed algorithm. SAS is based on a static combination of pooled segmentation outputs. We extend its original image-based formulation to stereo videos by including into its pool the GBH and moseg video segmentation methods, as we illustrate in Section 6.4.1.

Figure 6.5 clearly states that a static combination does not suffice to address the segmentation of stereo videos. By contrast, quite surprisingly, trying to always pool *all* video and image segmentation output *with the same contributing weights* turns out to harm performance.

6.6.2 EASVS and the State-of-the-art

Our adaptive combination of pooled segmentation outputs poses the question as to which measure to use for learning. As mentioned in Section 6.2, the BPR and VPR

stereo video segmentation	BPR	VPR	AM-BVPR	HM-BVPR
GBH (Grundmann <i>et al.</i> , 2010)	0.187	0.208	0.198	0.198
moseg (Ochs and Brox, 2011)	0.247	0.285	0.266	0.264
SAS (Li <i>et al.</i> , 2012)	0.184	0.087	0.135	0.118
4D-seg (Hickson <i>et al.</i> , 2014)	0.128	0.146	0.137	0.120
VideoCoSeg (Chiu and Fritz, 2013)	0.238	0.140	0.189	0.169
Proposed EASVS	0.301	0.296	0.295	0.288

Table 6.1: Results on the CSVSC benchmark. See the discussion in Section 6.6.2.

	no depth	fixed depth	fixed α	fixed β	proposed EASVS
HM-BVPR	0.254	0.276	0.270	0.276	0.288

Table 6.2: Analysis of the proposed EASVS. See the Section 6.6.3 for the discussion.

measures may push for adaptive algorithms with better boundaries or temporally-consistent volumes. Averaging BPR and VPR may balance the two aspects, which we may achieve by arithmetic (AM-BVPR) or harmonic mean (HM-BVPR).

In Table 6.1, we illustrate performance of EASVS against moseg, GBH and SAS, measured according to the four available metrics (BPR, VPR, AM-BVPR, HM-BVPR). For EASVS, the measured performance statistic has also been respectively used for learning the adaptive ensemble segmentation model. (Since our approach involves learning, our results are averaged on three folds.) The results in the table match the intuition that only an adaptive combination can successfully address all videos. Furthermore, our proposed EASVS outperforms a recent depth video segmentation method (Hickson *et al.*, 2014) (4D-seg) by more than 50% on all measures, as well as a recent video co-segmentation algorithm (Chiu and Fritz, 2013) that we run on each video stereo pair by 65%. This is confirmed by the qualitative examples shown in Table 6.3 and 6.4.

We delve further into the understanding of the potential result improvements within the EASVS framework with an oracle. In more details, we allow our algorithm to estimate the optimal segmentation-pool combination-parameters (α and β) by accessing the ground truth performance measure P for each stereo video sequence. The higher oracle performance by up to 70% (with the current representation and quadratic regressors) anticipate future improvements with richer models and more data.

6.6.3 Deeper Analysis of EASVS

In Table 6.2, we provide additional insights into EASVS. First, we experiment with 1) no depth and 2) fixed depth contribution. The performance drops by 11.5% for 1)

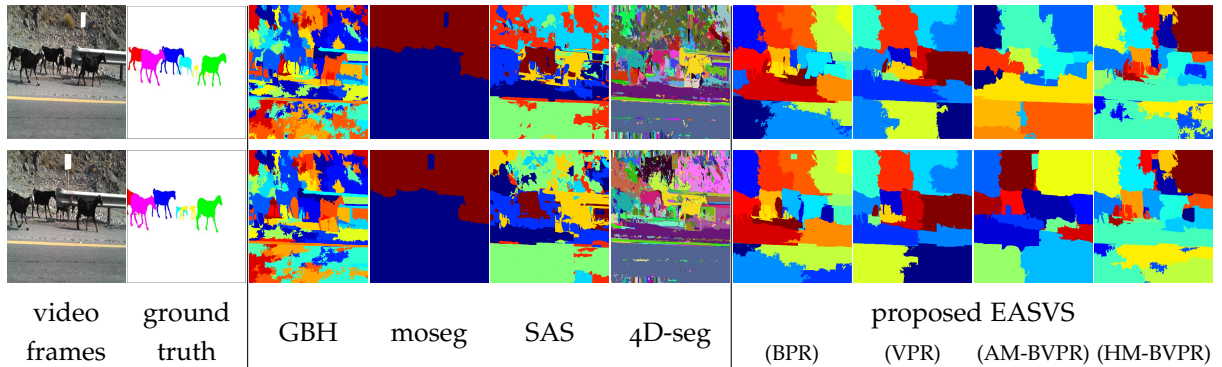


Table 6.3: Examples of the proposed EASVS optimized for different evaluation metrics compared to the state-of-the-art algorithms. Note how GBH (Grundmann *et al.*, 2010) outlines the object boundaries but tends to over-segment, while moseg (Ochs and Brox, 2011) produces under-segmentations and fails to extract objects without significant motion. The static combination scheme SAS (Li *et al.*, 2012) cannot strike good compromised parameters across all videos, which results in degraded results. 4D-seg (Hickson *et al.*, 2014) is a clear leap forward but suffers from some of the drawbacks of GBH. Our proposed EASVS benefits the learning framework and the adaptive nature for a better output.

and 4.2% for 2) in HM-BVPR. This shows the importance of the depth cue within the full system. Additionally, this speaks in favor of the adaptive strategy. (Cf. 4D-seg (Hickson *et al.*, 2014) also leverages depth but cannot reach the same performance as the adaptive depth combination.)

Second, we fix the combination parameters 3) α and 4) β to the single best values determined on the training set, therefore limiting the system adaptivity. The performance drops by 6.3% and 4.2% respectively. Once again, we find that adaptivity is therefore crucial for the performance of our system and that both adaptive aspects are strictly needed: weighting the pooled segmentation (α) and measuring similarity of the resulting superpixels (β).

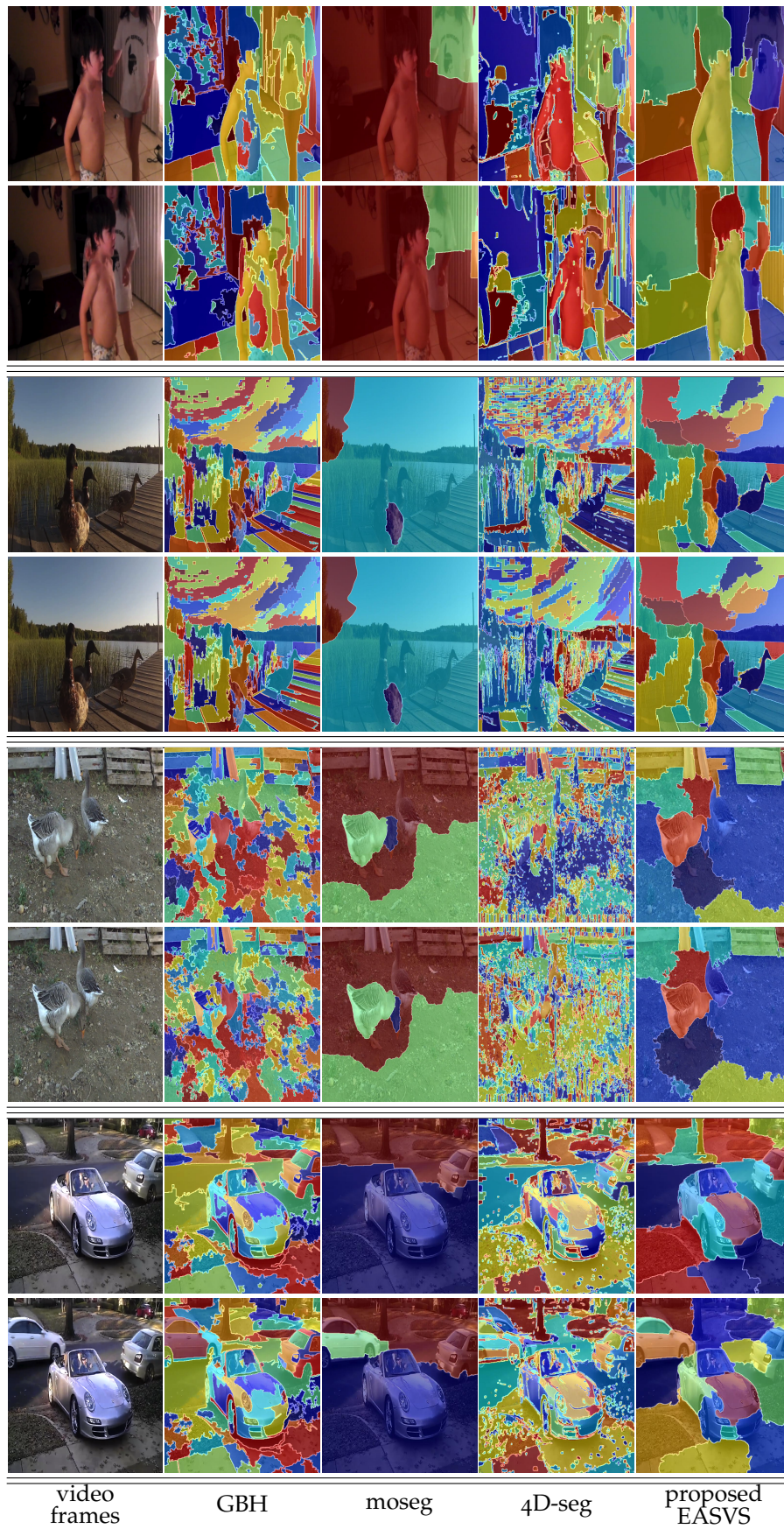


Table 6.4: Additional examples of the proposed EASVS compared to baselines.

6.7 CONCLUSIONS

We have considered the emerging topic of consumer stereo cameras and proposed a benchmark to evaluate progress for the task of segmentation with this interesting type of data. The dataset is challenging and it includes diverse visual cues and camera setups. None of the existing segmentation algorithms can perform well in all conditions.

Furthermore, we have introduced a novel efficient and adaptive stereo video segmentation algorithm. Our method is capable of combining optimally a pool of segmentation outputs from a number of "expert" algorithms. The quality of results highlights that combining single algorithms is promising and that research on such a framework is perfectly orthogonal to pushing performance in the single niches, e.g. motion segmentation, image segmentation, supervoxelization etc.

Contents

7.1	Introduction	97
7.2	Related Work	98
7.3	Cross-Modal Stereo	99
7.3.1	Stereo and Alignment to Kinect Depth	100
7.3.2	Cross-Modal Adaptation for IR-RGB-Stereo	102
7.3.3	Point Cloud Based Object Segmentation	102
7.3.4	Experiments	104
7.4	Learning Optimal Filters to Improve Cross-Modal Stereo	107
7.4.1	Capturing and Analyzing Sensor Characteristics of the Kinect	107
7.4.2	Experiments	111
7.5	Conclusions	114

AQUIREING depth estimate of a vision scenario provides richer geometric information than 2D observations which is helpful for many cases such as occlusion reasoning, 3D reconstruction, and resolving appearance ambiguities for object segmentation. In this chapter we study the problem of object segmentation in the multi-modal 3D data perceived by Kinect, one of most commonly used 3D sensors nowadays. In particular, we focus on improving the depth estimate for better 3D data acquisition in order to boost segmentation performance. We present a cross-modal stereo method which is inspired by the complementary properties between two depth sensing modalities: Kinect performs well on normal surfaces such as skin or cloth yet has failure on transparent and specular surfaces, while stereo vision is capable of estimating the disparities at edges of transparent or reflective objects but has difficulties on homogeneous areas. The depth estimate within the Kinect is compensated by the proposed cross-modal stereo path that we obtain from disparity matching between off-the-shelf IR and RGB sensors of the Kinect. In results, our proposed method produces depth maps that include sufficient evidence for reflective and transparent objects, preserve textureless objects (e.g. tables or walls), and show the improved performance on a point-cloud-based 3D object segmentation task.

7.1 INTRODUCTION

Future mobile robotics rely heavily on robust sensing schemes in order to bring the success of industrial robotic applications in controlled environments to the unstructured and everyday changing scenario in our homes. 3D Perception has been one of the key technologies to provide a rich capture of indoor scenes that facilitates data driven segmentation, grasp planning, and much more. While the steadily improving sensing technology has provided us with more accurate and reliable data, we haven't — and probably will not — see a single sensor performing well across every conceivable condition. This calls for robust integration of multiple sensing schemes to complement for each others' short comings.

With the introduction of the Microsoft Kinect sensor, a highly performant yet low cost 3D sensor was made available that rivals much more costly solutions available to robotics, and it has been widely distributed up to date. Despite being originally designed as a gaming interface, soon after its release there was strong interest from hobbyists over enthusiasts to robotics researchers trying to stretch the envelope of possible application scenarios beyond its original intended use case.

In this chapter we are particularly interested in the robotics scenario and resulting shortcomings when using the Kinect in those settings. In home environments object properties differ beyond the well behaved properties of cloth and skin where the Kinect depth estimation performs admirably well. We realize that in particular specular, transparent and reflective objects cause serious problems — and not rarely lead to a complete failure (See Fig. 7.1). Yet objects like glasses, bottles, tea kettles, pans are objects of daily living and therefore they are in the core set of objects home robotics and assisted living want to address (Choi *et al.*, 2009). We study how the knowledge from different modality helps to eliminate the difficulties met by Kinect. We complement the built-in active depth sensing scheme of Kinect with a passive and cross-modal stereo path which is established by performing stereo matching between RGB and IR sensors. The proposed framework provides a reliable depth estimation scheme using an off-the-shelf Kinect sensor without modifying hardware nor requiring any additional sensors.

We start from investigating several fusion schemes to provide more reliable sensor data and improve particularly on transparent and specular objects (Section 7.3). In order to preserve similarities between modalities, different combination schemes of RGB channels to mimic the image response of the IR sensor are studied. In the experiments we not only provide the qualitative improvements but also the empirical evidence for strong improvement on a data-driven object detection task in a table top scenario. However, as the Kinect's projected patterns for depth sensing introduce the interference observed in the IR images, the performance of the stereo matching across RGB and IR modalities has a significant drop-off from covering the IR projector to making it operated.

Therefore in the subsequent Section 7.4 a more detailed study is conducted. We identify three issues and consequently improves over the previous fusion scheme: First, we take a closer look at the sensor characteristics of the Kinect and realize that

the overlap in the spectral response between the sensors is very small. This argues for a learning based approach that exploits smoothness and correlations in the BRDF function of the materials, as no satisfactory linear reconstruction of the IR channel is possible.

Second, as argued in related literatures, we know from practical considerations that patch based stereo matching is often improved by pre-filter operations. This is a richer class than the pixel based weighting previously investigated. We propose a method for learning optimal filters for improving cross-modal stereo that is rich enough to capture channel-based weighting and filtering like sharpening, smoothing and edge detection.

Third, we realize that for the best results previously obtained, the IR projector had to be covered for capturing the IR-RGB pair. However, this is impracticable as it eliminates the active depth sensing scheme. We show that the proposed method of learning filters can achieve robustness of the stereo algorithm to these nuisances introduced by the projector – and in fact we are able to recover the performance previously only achieved with the covered projector.

7.2 RELATED WORK

Transparent and specular phenomena have been proven notoriously hard to capture (Ihrke *et al.*, 2008), in particular in unconstrained scenarios where prior information about lighting and geometry of the scene can rarely be assumed. Only in recent past some initial success towards practical systems for detecting transparent objects has been reported on visual object detection tasks (Fritz *et al.*, 2009) and multi-view lidar based object detection (Klank *et al.*, 2011).

(Fritz *et al.*, 2009) learns object models for glasses and Klank *et al.* (2011) improves transparent object detection by integrating two sensors of the same type. In contrast, our work uses a single off-the-shelf unit combining an active and a passive approach and we show improved results on wide range of effects like transparency, specularity and highly absorbent surfaces as they occur on many household objects such as tea kettles, mirrors, displays, bottles.

But also stereo algorithms are effected by more complex surface properties. E.g. (Tsin *et al.*, 2006) provides an analysis of such effects and presents a sophisticated model for recovering multi-layered scene structure. In practice, we see stereo correspondences still being preserved at least on borders of objects with complex surface properties. Therefore in order to complete Kinect’s depth estimate on specular, transparent or reflective surfaces, we first attempt to use a simple, computationally efficient block matching algorithm as implemented in OpenCV (Bradski, 2000) for seeking the correspondences between a RGB and IR sensor. However, with regard to the cross-modal stereo, we face a problem of different data domains.

Most recently, related problems have been successfully addressed in a metric learning formulation for visual category recognition from different data sources like images from the web, DSLRs and webcams (Saenko *et al.*, 2010). While this approach

is based on Information Theoretic Metric Learning (ITML) (Davis *et al.*, 2007) much simpler formulations based on large-margin classifiers (Daumé *et al.*, 2010) have been proposed from which we drew some inspiration. However, the latter approach is only applicable for classification while we learn a transformation that is directly applicable to the image without any change to the stereo algorithm.

Previous work investigated fusion techniques for depth measurements originating from time of flight cameras and stereo cameras (J.J. Zhu and Davis., 2008; Kim *et al.*, 2009). In contrast, our main focus in this chapter is to explore cross-modal stereo so that we can have an active and passive depth sensing path in a single sensor unit. Therefore the previous investigations are orthogonal to ours.

Beyond the problem of different data domains, our proposed cross-modal stereo vision encounters the interference on the IR images from Kinect's projected patterns, which complicates the stereo matching more in addition to lighting conditions or specific material property such as transparency and specularity. In practice such variations are typically reduced by filtering techniques (e.g., laplacian of gaussians (Konolige, 1998)), non-parametric matching costs (e.g., census (Zabih and Woodfill, 1994)) or by hand tuning parameters for optimal matching. (Hirschmüller and Scharstein, 2009) provide a thorough comparison of several stereo matching techniques with respect to complex radiometric variations. They compare a large set of filters, and rank them according to performance and computational efficiency.

More recently the path of machine learning is taken to find automatically optimal models for stereo matching (Li and Huttenlocher, 2008). Also (Hirschmüller and Scharstein, 2009) propose to learn pixelwise cost based on mutual information from ground truth data. However, both approaches are global-based matching scheme and prohibit real time applications. Also the sensitivity to local changes (Hirschmüller and Scharstein, 2009) limits its applicability for matching across modalities that exhibit global as well as local variation. In the following sections of our proposed method, we will walk through a succession of stages to study the stereo path of IR and RGB sensors, including: investigate various fusion schemes, experiment a pixel-based optimization based on ground truth, and further focus on learning patch based filters.

In our evaluation we use a object segmentation task based on a support plane assumption, which is not only common to many recent systems using 3D information (e.g. (Gould *et al.*, 2008; Marton *et al.*, 2009; Fritz *et al.*, 2010)) but also with an emphasis on problematic cases containing specular and reflective surfaces to demonstrate the benefits of our proposed cross-modal stereo.

7.3 CROSS-MODAL STEREO

As mentioned previously the Kinect depth estimate fails on specular, transparent or reflective surfaces. Depth is calculated from an IR-pattern that is projected from by the Kinect sensor unit (Fig. 7.2(b)). On reflective objects however the pattern is not visible or being reflected, causing holes in the depth maps or potential interferences



Figure 7.1: (top) Failure cases of Kinect 3D sensor: (red) transparency, (green) specularity, (cyan) dark objects under flat viewing angle, (yellow) reflections of the angle, (violet) interference of dot patterns by reflections. (bottom) We evaluate on a object segmentation task, left: result on Kinect point cloud only, right: strongly improved result on our fused estimate proposed in this chapter. (blue points: kinect; green points: cross-modal stereo)

(Fig. 7.2(d)). In contrast, stereo vision enables to detect disparities at edges of transparent or reflective objects, but has difficulties finding correspondences on textureless areas, such as the wall or the desk (Fig. 7.2(e)). Since the Kinect features two cameras (IR and RGB) we propose a cross-modal stereo approach that we combine with the built-in 3D estimate of the Kinect in order to compensate for the problems of the individual sensors. In the following we describe our method that can be run on an off-the-shelf Kinect sensor without any hardware modifications.

7.3.1 Stereo and Alignment to Kinect Depth

We first briefly describe our stereo calibration algorithm for the IR and RGB camera and the fusion step with Kinect depth information. Then, we describe the alignment procedure that enables the combination of stereo and Kinect depth information by simple union operator on the two point clouds. Since we have to deal with two different data domains, we introduce an optimization step in order to obtain improved stereo correspondences.

Stereo Calibration Given the images from IR and RGB sensors, the extrinsic and intrinsic parameters can be computed with standard stereo calibration technique.

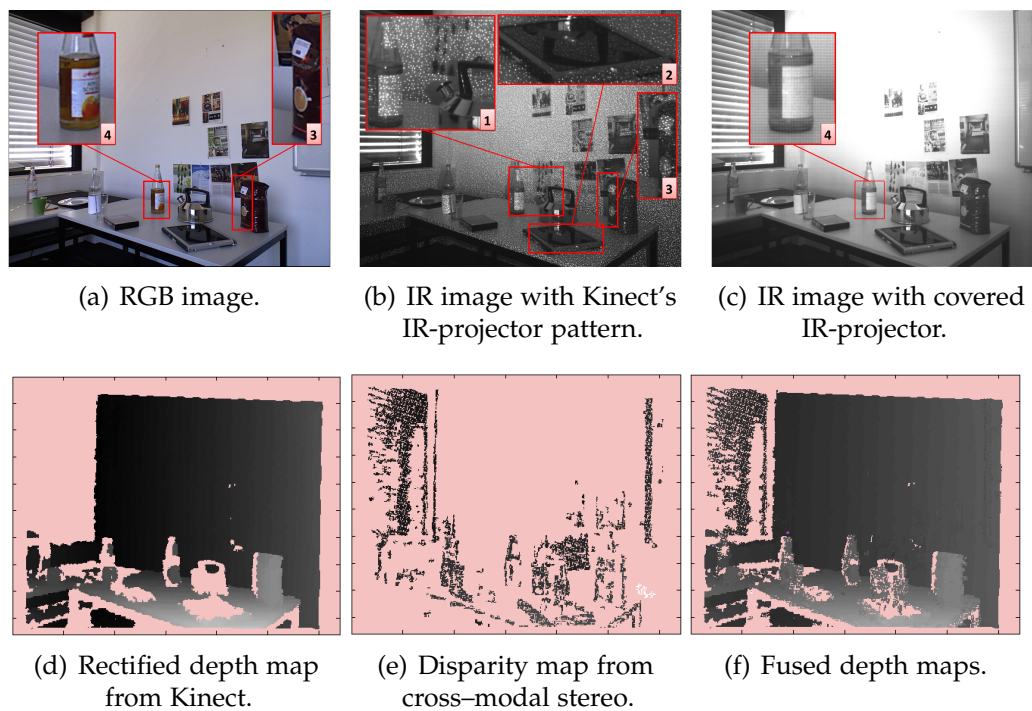


Figure 7.2: (a) *RGB image*. (b) *IR image with Kinect's IR-projector pattern*. 1. pattern not visible on reflective surfaces, 2. reflected pattern on display, 3. Shadow from projector spot differs to RGB lighting condition (c) *IR image with covered IR-projector*. Red texture invisible to IR-sensor (d) *Rectified depth map from Kinect*. (e) *Disparity map from stereo* and (f) *Fused depth maps*.

Once the calibration parameters of IR and RGB cameras are obtained, we apply Bouguet's algorithm to do the rectification toward the stereo image pairs and make them row-aligned. Then we utilize the SAD (Sum of Absolute Difference) block matching to find corresponding pixels between IR and RGB images. The offset between such pixels marks the disparity in image coordinates. As a result, we obtain the disparity maps for every pair of IR and RGB images, for which depth can be calculated as shown in Fig. 7.2(e).

Fusing Stereo Depth with Kinect Depth In order to combine the disparity map from our stereo setting and the depth map from the Kinect, we need to align the image planes. Since the disparity map from the stereo is rectified, we apply the same rectification to the depth map from the Kinect as shown in Figure 7.2(d).

After converting the disparity from stereo into depth measurements, we can directly compare depth values obtained from stereo and the Kinect. From a set of calibration scenes we obtain scaling and offset parameters that align the depth values. We use least squares to estimate these parameters. Figure 7.2(f) shows an example of the aligned depth measurements.

In order to evaluate our depth estimate on an object detection task, we generate

a point cloud by means of the reprojection matrix obtained from stereo calibration. The fusion of stereo and Kinect depth is carried out in 3D by simply taking the union of the point clouds as displayed in Figure 7.3.

7.3.2 Cross-Modal Adaptation for IR-RGB-Stereo

Early Integration As described in the previous section we search for stereo correspondences in the IR-RGB-image pairs. The channels of those images correspond to different sensor characteristics that are only receptive over a range of wavelengths. Due to correlations in the sensor and material characteristics running stereo across the modalities is expected to produce at least some correspondences. Yet, it seems more appropriate to find a better combination of the channels that would make the two signals more similar. We do so by employing a global optimization approach.

Given a IR image I^{ir} and a RGB image $(I_r^{rgb}, I_g^{rgb}, I_b^{rgb})$, we would like to obtain a weighting $w = (w_r, w_g, w_b)$ of the channels such that the converted RGB image is more similar IR image and has more corresponding points during the stereo matching. We evaluate the performance of stereo matching by simply calculating the number of corresponding points we can find, which we denote by: $\text{num_of_stereo_match}(I^{rgb}, I^{ir})$. The resulting optimization problem reads:

$$\begin{aligned} \max_{w_r, w_g, w_b} \quad & \text{num_of_stereo_match}(w_r * I_r^{rgb} + w_g * I_g^{rgb} + w_b * I_b^{rgb}, I^{ir}) \\ \text{subject to} \quad & w_r + w_g + w_b = 1. \end{aligned} \quad (7.1)$$

We use the IR image with covered IR-projector to avoid the effect from the projected pattern. This optimization problem is solved by grid search with uniformly sampling of the plane $w_r + w_g + w_b = 1$. Figure 7.3 shows the disparity map from the RGB image converted with learnt channel weights and compared to the disparity from original RGB image in gray level. We observe that more details are preserved after applying the learnt weighting.

Late Integration We also investigate a late integration scheme where we delay the combination of the different color channels and compute stereo correspondences w.r.t. the IR image independently. We proceed as with the 3D data from the Kinect and fuse the results in 3D space by forming the union over the point clouds.

7.3.3 Point Cloud Based Object Segmentation

In order to quantify the improved depth information for potential detection tasks, we implemented a simple object detection system based on point clouds from Sec. 7.3.1. Since objects of interest are often located on tables or fixed at walls, we first segment space into a support surface and a background surface. We then cluster the residual point cloud, which is neither part of the table's nor the wall's surface, into potential objects.

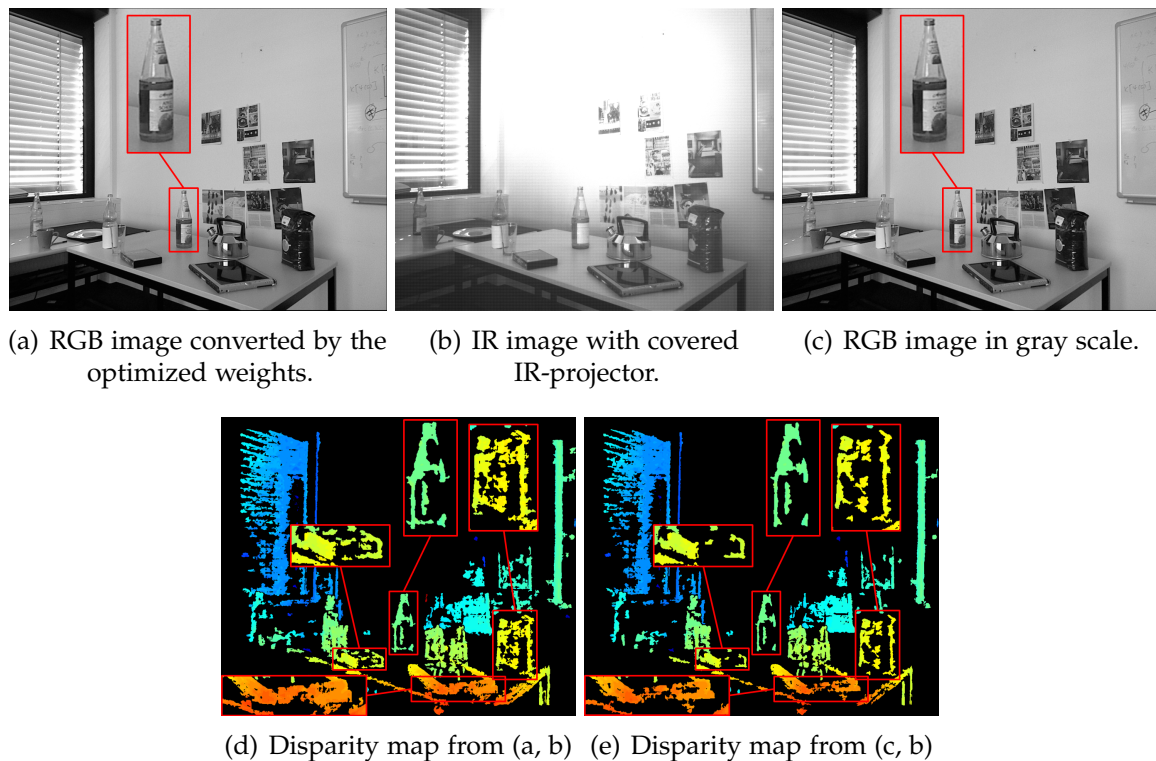


Figure 7.3: (a) RGB image converted according to the optimized color channel weights, $(w_r, w_g, w_b) = (0.368421, 0.473684, 0.157895)$. (b) IR image with covered IR-projector. (c) RGB image in gray scale (averaged channels with equal weight). (d) Disparity map from (a) and (b). (e) Disparity map from (c) and (b).

Support Surface Extraction We apply an iterative RANSAC-algorithm to extract the support surface and the background wall from the point cloud. We assume that the scene contains two surfaces: the background wall and the table, where objects of interest can be spread out. The residual points, which are neither inliers of the background wall and nor the table surface, are then clustered into potential objects.

Point Cloud Clustering We first partition the residual pointcloud from the step above using kmeans-clustering. We set the number of centers to $K = 850$. This effectively reduces the complexity for further calculation. The kmeans-centers are then further grouped by agglomerative clustering. Based on grouped kmeans-centers we calculate a 3D-bounding box as in Fig. 7.1. For the evaluation in the following section, we backproject 3D-coordinates of each group member into the image coordinates using the transformation from Sec. 7.3.1. Upon image pixels of each group a rectangular bounding box is fitted and compare to the ground truth annotations of objects. In order to score each bounding box for precision-recall analysis we use the number of points associated with each box, respectively group, as a score.

7.3.4 Experiments

We evaluate the success of our approach on a new database that we have collected in order to test the kinect sensor on more challenging scenarios. The dataset consists of 106 objects in 19 images. All objects are annotated with 2D bounding boxes. We follow the evaluation criterion of the Pascal challenge (Everingham *et al.*, 2015) and compute precision-recall curves based on the overlap criterion $a_0 = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})}$, where B_p is the predicted bounding box and B_{gt} the ground truth bounding box and where a_0 must exceed 50%.

We evaluate several fusion schemes under different conditions, and settings for the cross-modal stereo matching. Following standard stereo matching, we average RGB into a grayscale image (stereo only: stereo_{rgb} and fused with kinect depth: fused_{rgb} *early*). We also combine each channel individually (stereo only: $\text{stereo}_{\{r,g,b\}}$ and fused with kinect depth: $\text{fused}_{\{r,g,b\}}$). Given a disparity map for each channel individually, we also combine these into a late fusion scheme with the kinect depthmap (fused_{rgb} *late*). For a weighted combination of the R, G and B-channel as presented in Sec 7.3.2 we denote the index c .

7.3.4.1 Results

Table 7.4 shows average precision for different fusion schemes. We expect a strong influence of the IR projector on the correspondence matching of the stereo vision. Therefore, we captured stereo pairs under two conditions, first by covering the emitting IR projector and second under normal operating condition with the IR projector switched on. Note, that since the Kinect depth estimate does not operate without the IR projector, we captured the images consecutively in the first setting.

The built-in Kinect depth estimate achieves 48.8% of average precision. Combining stereo with the Kinect depth results in significant improvement of nearly 30%. Overall the maximum average precision of 76.6% is achieved by fusing all channel-specific depth maps (fusion_{rgb} *late*). When turning the Kinect into normal operation mode that is with emitting IR projector, overall performance of different combination schemes decreases, but still improves the Kinect depth about 10-20%. Interestingly, the best result (68.8%) is achieved by fusing Kinect depth and stereo depth from the green channel, which is closely followed by fusion with channel-specific depth estimates (fusion_{rgb} *late*: 66.5%). As expected in this scenario, stereo only with projected IR pattern, performs worse in all different combination schemes than Kinect only.

Fig. 7.5 shows example depth maps and detections using the pointcloud segmentation from Sec. 7.3.3. Fig. 7.5 (a) shows a comparison between the Kinect-only depthmaps and fusion_{all} *late* scheme. Based on kinect depth estimation, we observe that nearly all transparent or reflective objects are either missed or over-segmented by their opaque parts (e.g. the bottle labels). Also interferences occur, e.g., the reflected IR-pattern from the wall results in false depth on tablet's display.

Fig. 7.5 (b) shows a comparison of fused depth maps with operating IR projector

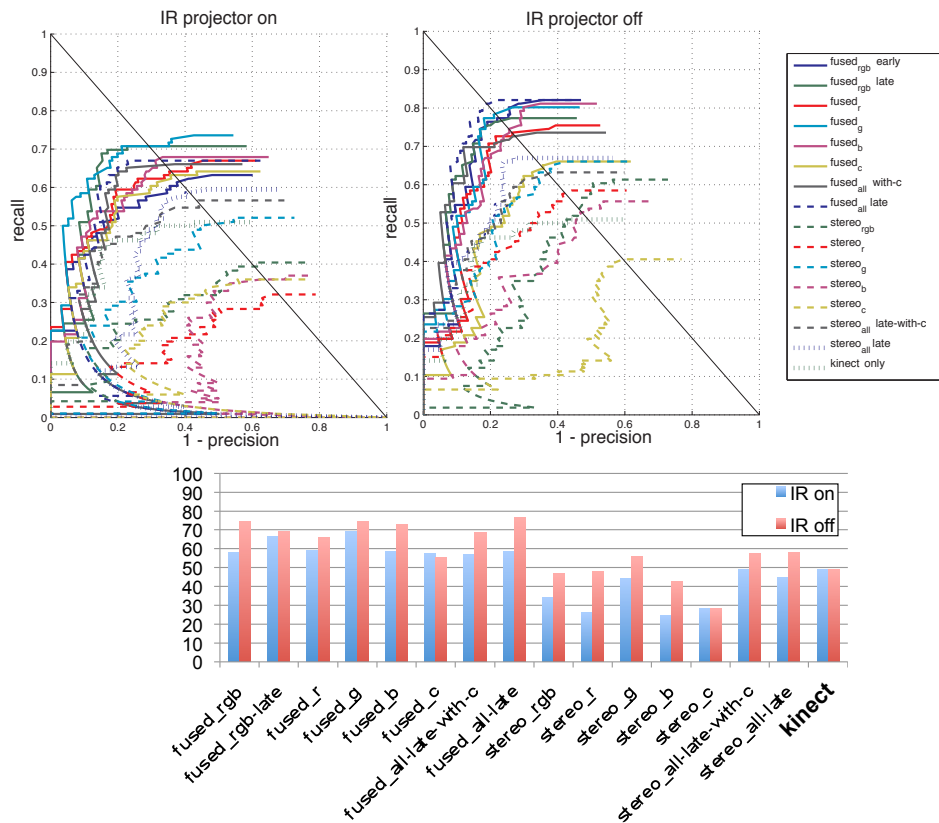


Figure 7.4: Precision Recall and average precision for the table-dataset

(middle) and with covered IR projector (bottom). It can be seen that the object segmentation merges individual objects into one object only with operating IR projector. Besides the interfering IR pattern on the object's edges, lighting conditions differ. While the IR projector behaves similar to a spotlight and causes hard shadows around objects in a scene, the RGB images are effected by environmental illumination only. As a results shadows differ between IR and RGB images (see Fig. 7.2(a) and 7.2(b), box 3). This leads to increased smearing effects and which leads to point cloud connections between nearby objects. Here a more sophisticated segmentation approach or statistical outlier removal techniques can remedy this effect. The left-most depth maps show a rather pathologic case, where the Kinect is directed to a mirror. The emitted IR pattern is reflected back to the camera causing a glare. Only when switching off the IR-pattern depth is revealed by stereo vision.

7.3.4.2 Discussion

We can see that stereo vision across modalities is feasible and improves object detection based on the Kinect's depth estimation up to 30% without any modification of the hardware. Overall, stereo matching between the ir channel and each color channel individually combined with 3D from the Kinect performs best in all considered settings. Using depth maps based on the green channel performs surprisingly

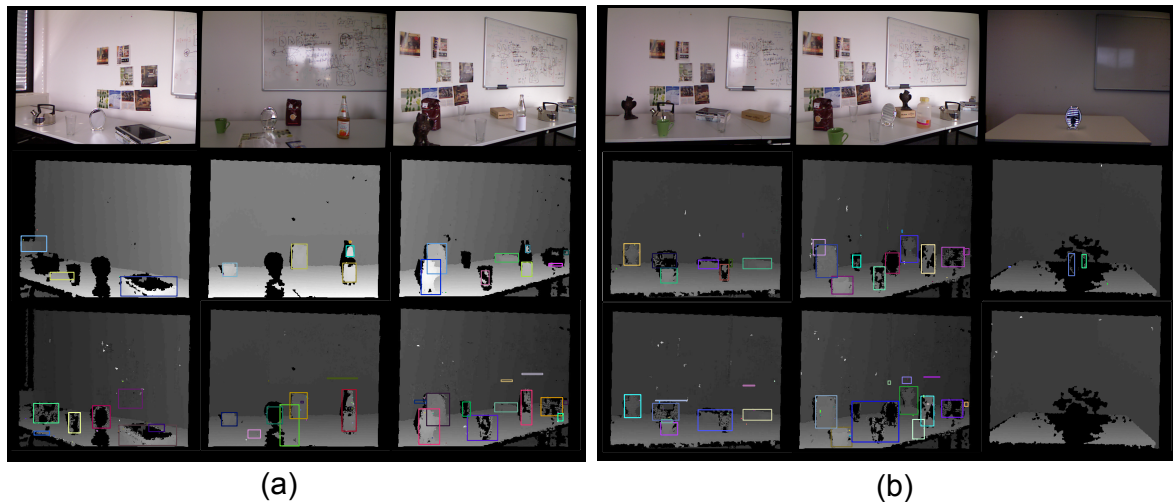


Figure 7.5: Example images from dataset. (a) Top: RGB image, Middle: Kinect only, Bottom *fusion_all_late*, (b) Top: RGB images, Middle: *fusion_all_late* with covered IR projector, Bottom: *fusion_all_late*

well. Since the red and infrared are close in wavelength, the infrared camera has a similar sensitivity. We intuitively expected a good correspondence matching for this channel. In fact however, red textures on white background do not contrast, and red becomes invisible. This effect can be seen on the bottle label in Fig. 7.2(a) and 7.2(c). Green texture however is preserved and represented by low intensities in the IR channel. Then, stronger gradients facilitate the correspondence matching.

Learning a weighted RGB-channel-combination scheme to obtain an “IR-like” image, turns out to be highly sensitive to environmental change. Colors of objects or varying environmental IR exposure influences the choice of weights significantly. Fig. 7.6(c) shows a series of captures during varying daylight conditions and corresponding optimized weights. Although we could find more correspondences using a weighted scheme Fig. 7.6(b) compared to grayscale RGB only Fig. 7.6(a), overall segmentation results did not reflect the improvement. Since, we estimated the parameters on a training set, they did not generalize well to our detection dataset. Dynamic weight adaption based on image statistics, such as illumination and white balance, might lead to improvement, which is subject to further investigation.

Practical Issues Kinect does not allow simultaneous grabbing of the RGB and IR stream. The *OpenNI* framework, as well as *libfreenect* offer functionality of switching the streams programmatically (and asynchronously). The speed depends on the buffer writing speed. When switching too fast the buffer is not entirely written. We did initial stress tests to tune the framerate. We observe that *libfreenect* shows faster performance and yields about 1.5-2fps for taking an IR and RGB pair, which leaves space for estimating stereo disparity maps and provides reasonable update rates for many robotics applications. The *OpenNI* framework achieves far less than 1fps.

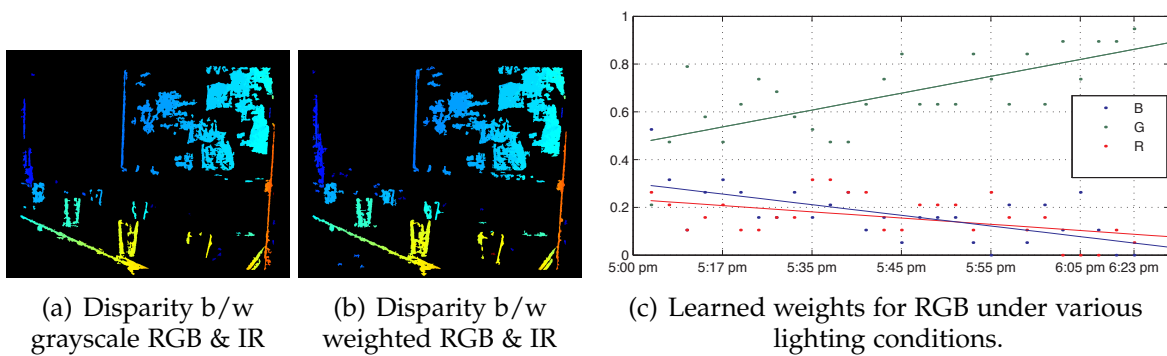


Figure 7.6: (a) Disparity from grayscale RGB and IR, (b) Disparity from weighted RGB and IR, and (c) Learned weights for obtaining IR like image from RGB under different lighting conditions (Date of capture 31. March, sunset 7:56pm).

7.4 LEARNING OPTIMAL FILTERS TO IMPROVE CROSS-MODAL STEREO

7.4.1 Capturing and Analyzing Sensor Characteristics of the Kinect

In the previous section a mapping between IR and RGB was learned from patterns that were illuminated by environmental light. However, there was no justification given if there is any hope to actually recover the sensor response characteristic by a linear combination of the RGB channels. Therefore we provide here a first analysis of the sensor characteristics of the images in the Kinect.

To this end we capture diffracted light which is projected on a “white” surface. This allows us to determine the characteristics of the Kinect cameras by measuring their response to different wavelength (see Fig 7.11).

The setup is depicted in Fig. 7.7. Environmental light is shielded so that we are only capturing the relevant wavelength. A special target that is almost perfectly lambertian ensures that the results are not corrupted by any specular effects. A light source is directed toward two small slits that serve as an aperture for selecting close to parallel light rays. This minimizes overlap between nearby wavelength on our target. Behind the slits a optical grating pattern causes diffraction which separates out the different wavelength. The light source is a 500 Watts Halogen light which – as a black-body-like radiator – emits light across the visual spectrum well into the infrared part, following roughly Planck’s law (Planck, 1901). We do not use a calibrated light source in this study and consider it of lesser importance as we are mostly interested in relative sensitivities under naturally occurring light.

After acquiring reference images in ambient light, we calibrate the images and calculate the response profile across wavelength. We do this for each RGB channel separately, as well for the IR-image. (See Fig 7.8 bottom). This gives us the sensitivity for each channel independently.

Having an estimate for the sensor response characteristics, we can now estimate a

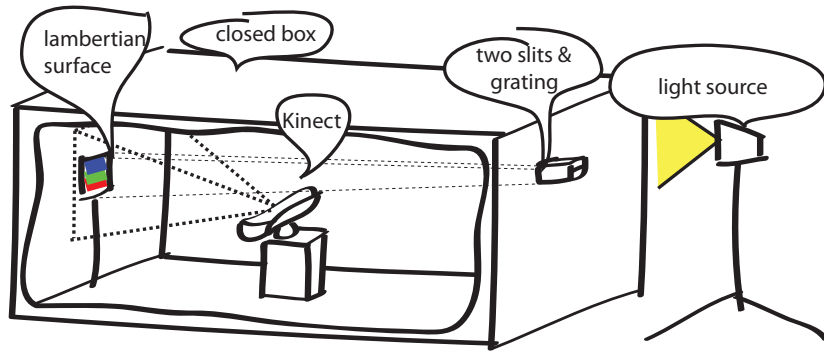


Figure 7.7: Schematic for experimental setup for reading sensor characteristics.

reconstruction of the IR sensor by a linear weighting of the RGB responses. Therefore we find the following least squares solution:

$$\min_w \|R_{ir} - [R_r R_g R_b]w\|_2 \quad (7.2)$$

where R_{ir} is the spectral response of the IR sensor and R_r, R_g, R_b are the responses of the red, blue and green channel respectively.

Results Fig 7.9 depicts the sensor readings we have obtained. The raw sensor data is plotted in pale colors, while the saturated colors show a Gaussian fit. For each channel we subtract the minimum response in order to compensate for sensor noise and residual ambient light and then fit a Gaussian mixture model with 3 modes as we observe 3 maxima of the diffraction pattern. The dominant mode is plotted per channel. There are 4 IR channels as we read the raw IR image from the Kinect that comes in a bayer pattern. We expected slightly different response characteristics for each channel, but they turn out to be almost identical. Furthermore we observe that the overlap between IR and RGB-channels is relatively small. The linear reconstruction of the IR channel from Equation 7.2 results in the cyan line shown in Fig 7.9. As the profile is very flat, we also show an amplified version. The low magnitude indicates that the reconstruction is not working well. The weights for the individual channels are as follows: $w_{red} = 0.0111, w_{green} = -0.0066, w_{blue} = 0.0022$. Obviously, the red channel has the highest weight, as it is closest to the infrared part. Interestingly, we get a negative weight for green which “pushes” the red channel further to the infrared part. The positive weight for blue again compensates partially for the introduced dip in the green to blue wavelength. This is also an interesting parallel to the method presented in the previous section where similar weights were obtained by training pixel correspondences without explicit knowledge of the spectral sensitivities.

In summary, we have to conclude that the overlap of the IR and RGB sensitivity

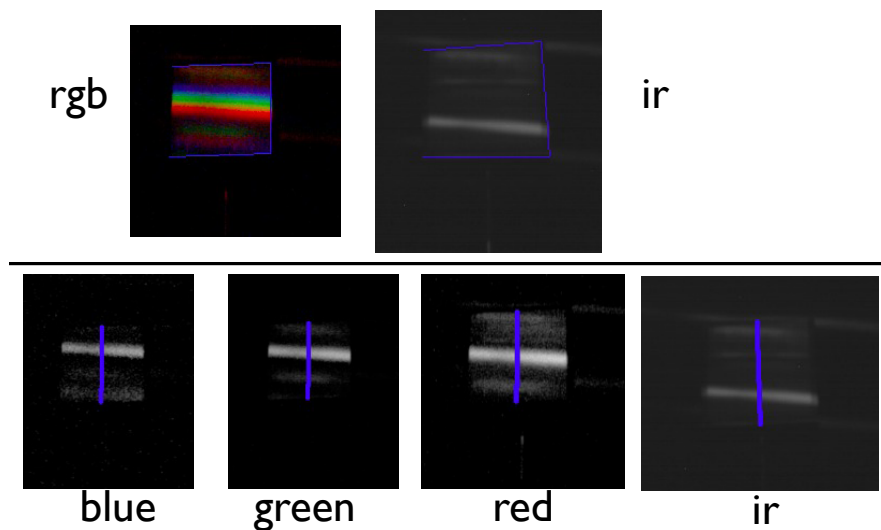


Figure 7.8: Spectrum from Experiment as in Fig 7.7.

of the sensor is indeed smaller than expected, which seems very bad news for any cross-modal matching attempt. However, in practice we do often have light sources that cover a reasonable part of the spectrum – like the above used halogen lamp – and in addition typical materials also reflect in a relative broad and smooth spectrum. This gives a justification to learning-based approaches like the one in previous section that can exploit correlations and smoothness of BRDFs.

Our aim is to increase robustness as well as computational efficiency of cross-modal stereo under projected patterns by learned filters. A simplistic scheme that is exclusively based on pixel-wise re-weighting of IR and RGB seems to be too limited. A learning-based version of this linear scheme was attempted in the previous section (shown as Fig 7.10) and we also derived a weighting based on spectral measurements in the spectrum experiment above.

As we want to stay in the realm of efficient patch-based stereo algorithm, we propose to extend the class of learned transformation to linear filters that leverage a pixel neighborhood in all channels to optimally preserve matches across modalities. These linear filters encompass smoothing, sharpening and edge detection methods that have been shown useful as prefilter in stereo algorithm and can potentially alleviate problems with the projected pattern.

The core idea is to collect corresponding pairs of patches between IR and RGB images into a set S for the training step. Then we use them to determine the weightings of each elements in the IR and RGB patches so that the corresponding patches have a smaller distance after the transformation. We employ an optimization framework to describe this problem.

We denote the s -th corresponding pair of patches by $\{IR^s, C^s\} \in S$ where $C = \{r, g, b\}$ contains three color channels from the RGB image. With the assumption

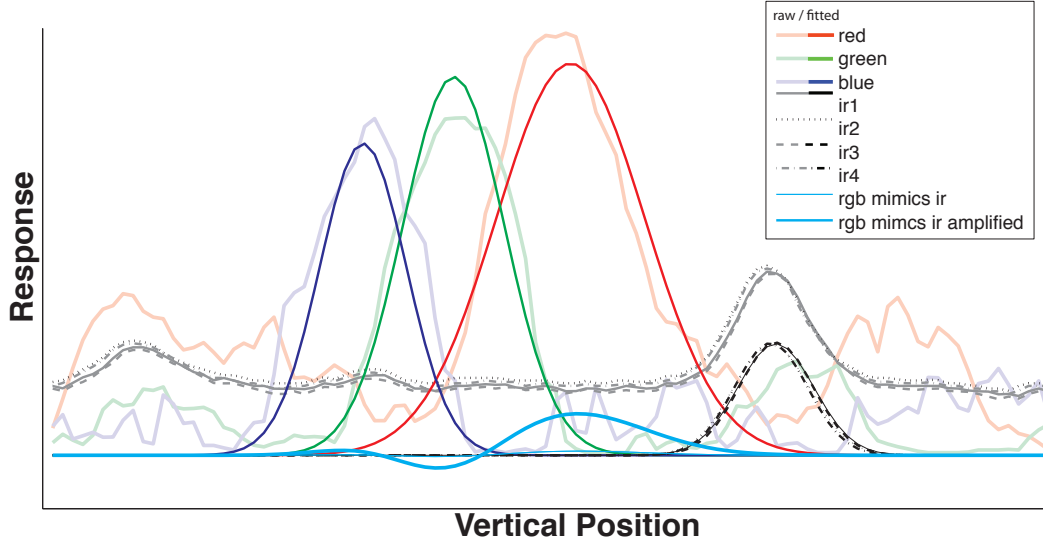


Figure 7.9: Spectrum from Experiment as in Fig 7.7. Light colored plots correspond to the response along blue lines through raw image data (Fig 7.8). Strong colored plots correspond to Gaussian fitted curves.

that the patch is in the size of $n \times n$, we would like to obtain the different weightings $\{w_{i,j}^{IR}, w_{i,j}^C\}$ for every pixels $\{IR_{i,j}^s, C_{i,j}^s\}$ of different positions (i, j) within IR and RGB patches $\{IR^s, C^s\}$. The resulting optimization problem reads:

$$\min_{w^{IR}, w^C} \sum_{s \in S} \left\| \sum_{i=1}^n \sum_{j=1}^n w_{i,j}^{IR} IR_{i,j}^s - \sum_{C=r,g,b} \sum_{i=1}^n \sum_{j=1}^n w_{i,j}^C C_{i,j}^s + b \right\|_1 \quad (7.3)$$

subject to $\sum_{C=r,g,b} \sum_{i=1}^n \sum_{j=1}^n w_{i,j}^C = 1.$

where b is an offset. By applying these weightings for each color channels of RGB

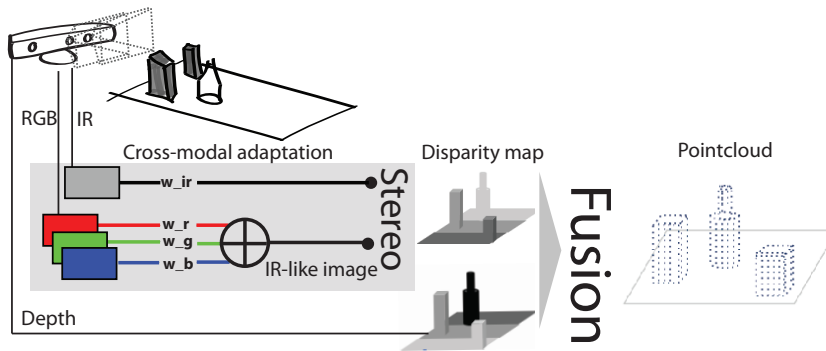


Figure 7.10: Diagrams for weighted fusion scheme.

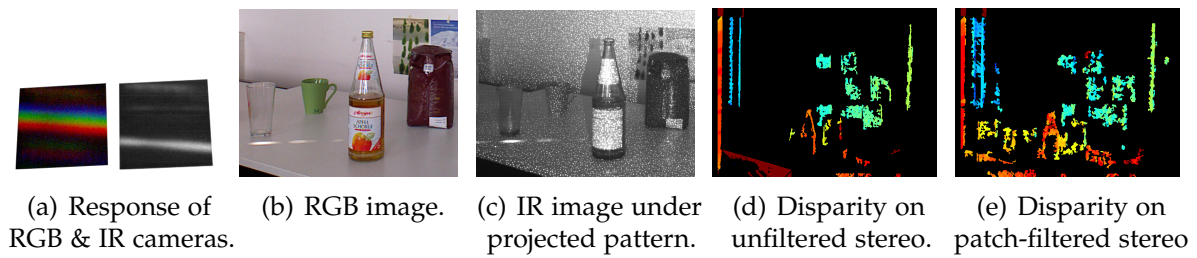


Figure 7.11: (a) Response of RGB camera (left) and IR camera (right). (b) and (c) Image pair obtained by Kinect with projected IR pattern. (d) Disparity map on unfiltered pairs. (e) Disparity map on patch-filtered image pairs.

images and for IR images, we can transform the RGB images into “IR-like” images then use the same stereo matching algorithm to compute the disparity maps as usual. Note that this weighting procedure is the same as utilizing filters for images. We display an instance of our proposed filtering procedure in Figure 7.14.

7.4.2 Experiments

In order to evaluate the effectiveness of our approach we compare to the results from the previous section in the same experimental setting. A clustering approach is used to segment objects in a table-top scenario.

7.4.2.1 Learning Filters

Given image pairs of IR and RGB images, our goal is to learn optimal adaptation between IR and RGB images using the Kinect hardware without any modifications. We manually collect thousands of corresponding pairs of 3×3 patches between low-resolution IR and RGB images under the influence of the IR-projector. The patches are distributed over normal and difficult regions including transparent, specular and reflective surfaces. To solve the optimization problem in Equation 7.3, we use *cvx* (Grant and Boyd, 2011), a matlab-based toolbox for convex optimization.

The resulting filters w^r , w^g , w^b , and w^{IR} are as follows with the offset $b =$

−56.6978:

$$\begin{aligned}
 w^r &= \begin{bmatrix} 0.1451 & 0.1900 & 0.1228 \\ -0.0354 & 0.0089 & 0.1244 \\ 0.1788 & 0.0985 & 0.1809 \end{bmatrix} \\
 w^s &= \begin{bmatrix} 0.1844 & -0.0806 & 0.1249 \\ 0.1866 & -0.1393 & 0.1129 \\ 0.1981 & -0.0841 & 0.0841 \end{bmatrix} \\
 w^b &= \begin{bmatrix} -0.0702 & -0.0260 & -0.0098 \\ -0.1430 & -0.0915 & -0.0600 \\ -0.0984 & -0.0654 & -0.0365 \end{bmatrix} \\
 w^{IR} &= \begin{bmatrix} 0.0049 & 0.0961 & -0.0006 \\ 0.1532 & -1.0000 & 0.1084 \\ -0.0018 & 0.0741 & -0.0062 \end{bmatrix}
 \end{aligned} \tag{7.4}$$

Visualizations are provided in Figure 7.14.

7.4.2.2 Evaluation

Our evaluation uses the same data as shown in previous section and is consistent with the setting in order to ensure comparability. The stereo image pairs from the Kinect were obtained under two conditions. The first one is to cover the IR projector and the second one is under the normal situation with the IR projector.

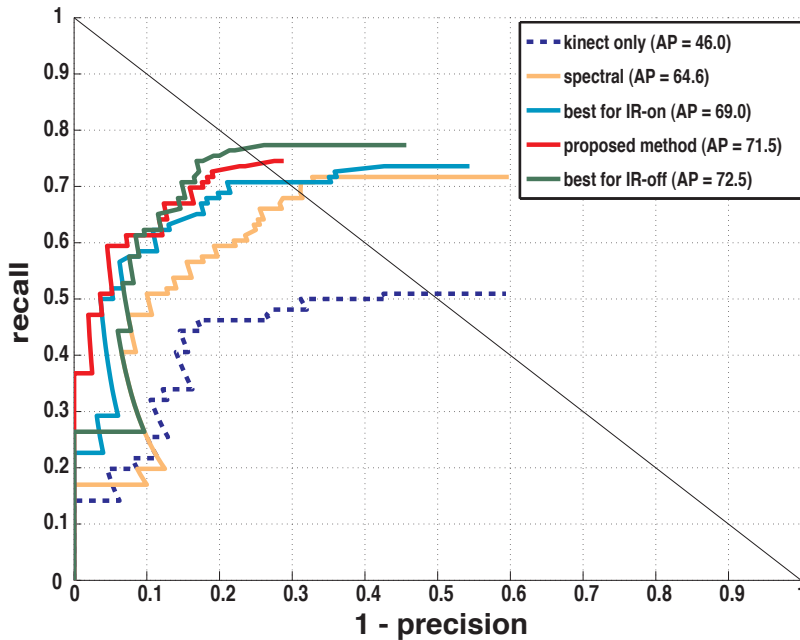


Figure 7.12: Precision–Recall curves of late fusion scheme under IR–projector–off setting, our proposed method and Kinect–only.

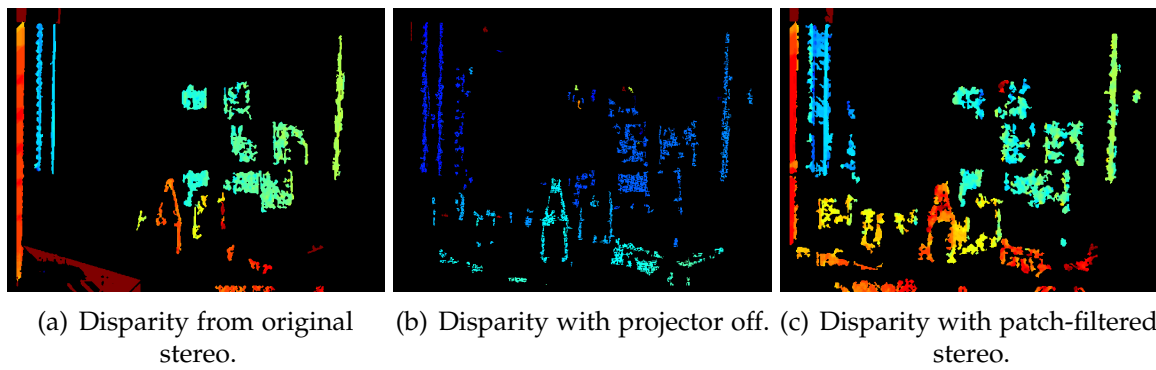


Figure 7.13: Disparity maps computed from (a) original image and (b) IR–RGB–image pair without IR–projector, and from (c) images filtered by our trained filters.

From the evaluation of different fusion schemes under these two settings, the best results previously obtained are an average precision of 69% with IR–projector–on and 72.5% with IR–projector–off by using a late fusion strategy.

Comparing to the average precision 46% of the built-in Kinect depth estimate, the method can improve the Kinect depth over 20% but meet the practical issues as we mentioned in Section 7.4.1. In contrast, our method simply uses the early fusion scheme with applying the filters on IR–RGB images captured under IR–projector–on but still achieves an average precision of 71.5%.

The Precision–Recall curves of the late fusion scheme under IR–projector–off setting, our proposed method, and Kinect–only are plotted in Figure 7.12.

Our approach outperforms all the settings using IR–projector–on and is on par with the previously best result with modified hardware. Also note that our method shows strong improvements in precision and produces the first false positives not until almost 40% recall which is 10% more than the competing methods. In Figure 7.13 we show exemplary disparity maps computed from images in the IR–projector off case, and filtered images by our proposed method.

7.4.2.3 Discussion

In the middle column of Figure 7.14, we present the visualization of the filters obtained from optimization process, and their characteristics can be observed. The filters of red channel and blue channel resemble smoothing operators and the filter of green channel, the smoothing seems to be applied along the y–axis while the x–axis direction resembles a Laplacian operator. The filter of IR channel basically computes a filter similar to a 2–dimensional Laplacian operator. Although at the first glance, people might claim that applying these filters which include the smoothing operations could worsen the images details and make the stereo more difficult. But actually these weighted summations just gather the data around the pixels into themselves then the information required for stereo is still maintained. With the fact that in stereo matching procedure we compare the statistical values (SAD is

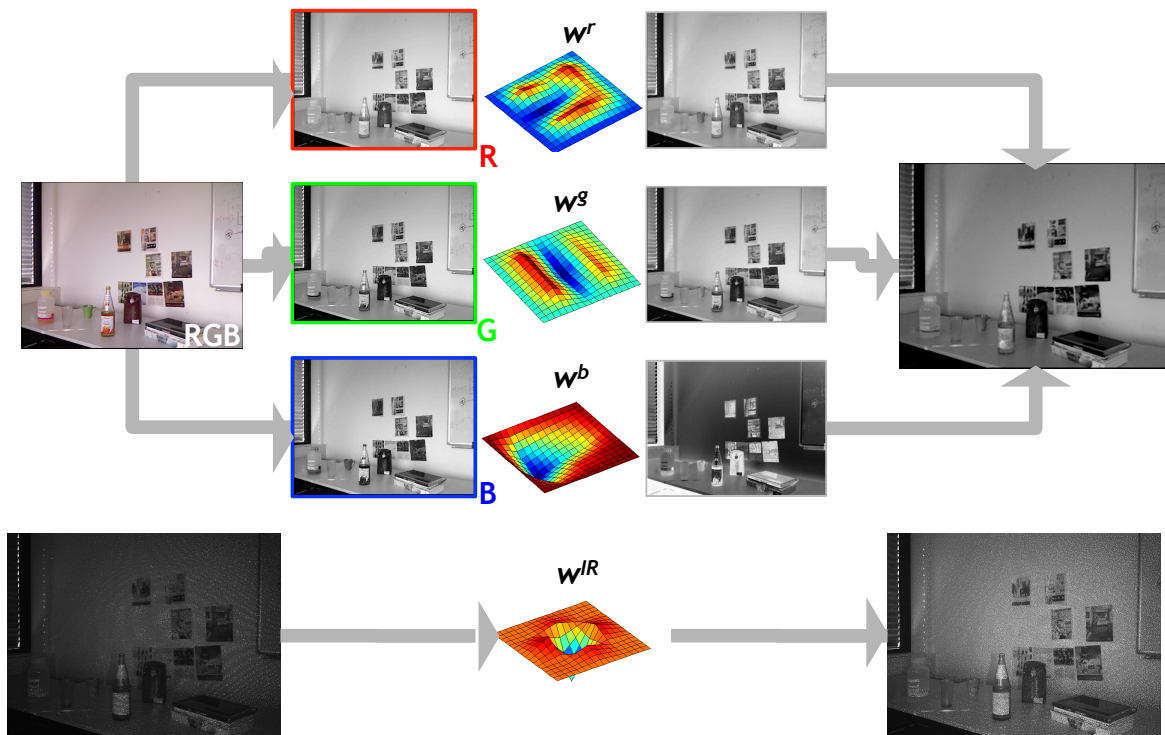


Figure 7.14: (Left) Incoming RGB and IR image with IR-projector-on from Kinect. (Top) color channels from RGB image: red, green, blue and learned filters w_R , w_G , w_B , w_{IR} from optimization. Note here we do the zero-padding and up-sample the filter for a better visualization. The resulting images filtered for each color channels. (Right) transformed RGB (summed filtered images for three color channels) and IR images.

used here) between local patches instead of judging each pixels separately, the filters will not affect the working of stereo. Besides, the Laplacian of Gaussian operations contained in the filters of green and IR channels will enhance the high frequency signals to make the stereo matching more efficient. Therefore, our learned filters can transform the RGB and IR images into more similar ones to improve the stereo across the modalities.

7.5 CONCLUSIONS

We presented a simple and effective cross-modal stereo vision approach for combination with Kinect depth estimates, which can be applied without any further hardware requirements. We provide empirical evidence for drastic improvement to the Kinect 3D sensing capabilities.

We first proposed a cross-modal adaptation scheme that allows for improved correspondence matching between RGB and IR cameras and show general feasibility

of their combination. Our combination method produces depthmaps that include sufficient evidence for reflective and transparent objects, and preserves at the same time textureless objects, such as tables or a walls.

We then presented a method to optimize filters for improved stereo correspondence IR and RGB images that is robust to projected IR patterns. We have experimentally analyzed the spectral characteristics of the Kinect cameras in order to justify such an approach. Adapting RGB in frequency domain to mimic an IR image did not yield improved performance. The small overlap between RGB and IR seems prohibiting this approach. In contrast, learning several filters based on image patches allowed improved stereo vision across modalities. We conclude therefore, that our pre-filtered, cross-modal, SAD-based stereo vision algorithm profits most from combination in the spatial domain, rather than in the frequency domain.

The value of our improved 3D sensing scheme is validated by a generic, data-driven object detection task. Our patch-based approach shows not only the increased performance with respect to Kinect's depth estimate but also the improved robustness against IR-specific interference from the projector.

We expect this work to have a high impact in the robotics community due to the wide spread use of Kinect sensors and the ubiquitous problem of capturing transparent objects for detection, recognition and manipulation.

Contents

8.1	Introduction	116
8.2	Related Work	118
8.3	Object Disambiguation	118
8.4	Experiments	121
	8.4.1 Object Disambiguation DataSet (ObDiDaS)	122
	8.4.2 Object Disambiguation Metrics	122
	8.4.3 Evaluation	125
8.5	Conclusion	127

IN this chapter we propose a multi-part object detection system with a focus on a factory scenario, where a single machine is composed of potentially repetitive machine parts. Given a monocular video, the system not only detects machine parts but also disambiguates them using the 3D context, and leveraging the prior knowledge on the structure of the machine. Specifically, the proposed system combines noisy object detections of 2D machine parts with the 3D depth information obtained from SLAM approach, and produces an accumulated 3D pointcloud of detections over video frames. The matching between 3D detection pointcloud and 3D context disambiguates object detections and enables an augmented reality overlay to assist a maintenance worker in locating machine parts as well as distinguishing their identities within the machine layout. For quantitative evaluation, we proposed an annotated dataset as well as several performance metrics that can be used to quantify the success rate of a variety of 2D and 3D systems for object detection and disambiguation. We verify in the experiments that the 3D context which encodes rich information on various cues such as deformation and scale factors, clearly contributes to the proposed method for the multi-part object detection task.

8.1 INTRODUCTION

The advent of affordable, highly miniaturized wearable camera technology in combination with the latest improvement of head-up display has intensified interest in augmented reality applications. The availability of such devices in the foreseeable future as well as the large scope of use cases in the consumer market (e.g. games) as well as industrial applications (e.g. maintenance) begs the question if current computer vision techniques can shoulder the expectations.

We investigate this question on a task of assisting a maintenance worker in a factory setting. The system has to provide an overlay to the worker so that machines parts are correctly identified. Depending on the application this simply supports successful completion of a task, averts dangers from the worker or prevents damage to the machine. In our study we will focus on sensing by a monocular camera as this is still the most commonly deployed modality in those devices to date.

Object recognition and detection has significantly matured over the last decade. We have seen great progress in instance (Hinterstoisser *et al.*, 2011) as well as category recognition (Felzenszwalb *et al.*, 2008; Jia, 2013). However in many of the aforementioned tasks we are faced with compositionality of objects from potentially repetitive parts. Robust matching of parts with their relational structure is required to detect the object as a whole and give semantics to the individual parts. We denote the task of predicting the identities of parts as object disambiguation.

Although such an object disambiguation task is really at the core of many augmented reality systems and systems for assistance in work environments in particular, there has been little progress in quantifying performance in these settings. In general, computer vision research has a strong tradition in building benchmarks that allow for measuring and comparing performance of object recognition and detection approaches. Most prominently the PASCAL challenge has greatly supported progress in object detection and the ImageNet challenge has played a similar role for object recognition. Therefore we advocate the need of a benchmark for augmented reality settings. We realize that it is very challenging to build such a benchmark in a completely task agnostic manner. In our study, we are focusing on a maintenance work task.

In order to establish a well defined benchmark, a performance metric is needed that allows for automatic evaluation. While there are widely adopted metrics for object recognition and detection, those are not directly applicable to our settings. First, object disambiguation has to deal with potentially repetitive objects whose identities are only resolved in context and therefore it is not captured by previous object detection metrics. Second, we are interest in the actual success of the user of the augmented reality system. Hence we seek a metric that measures the user's success in disambiguating the objects given the observation of the system's output.

We propose the first benchmark for augmented reality systems in maintenance work. Different metrics are evaluated to judge the systems performance in the context of the application. We propose a metric that closely follows the actual performance achieved by the human observer of the system's output. We propose the first system for object disambiguation that leverage 3d context from a SLAM system as well as flexible constraints on the matching procedure in order to robustly interpret the output of state-of-the-art object detectors for the task of object disambiguation.

8.2 RELATED WORK

2D Detection Our approach uses object detectors in order to evidence of machine parts from the image. We evaluate a range of commonly used object detectors (Viola and Jones, 2001; Dalal and Triggs, 2005; Hinterstoisser *et al.*, 2011; Felzenszwalb *et al.*, 2008; Shahbaz Khan *et al.*, 2012). All of them meet real-time constraints. While some of the were already built with efficiency in mind (Viola and Jones, 2001; Dalal and Triggs, 2005; Hinterstoisser *et al.*, 2011), other have seen recent extension by algorithms speed up as well as GPU computation (Song *et al.*, 2012; Dean *et al.*, 2013; Kokkinos, 2013). As we are facing the challenge of reoccurring parts, object detection on it's own is insufficient to resolve ambiguities.

3D Context Previous approach have explored improving object detection based on normal, size, height information extracted from dense 3D data (Gould *et al.*, 2008; Fritz *et al.*, 2010). The most related approach to our work is using 3D layouts of object detectors in order for indoor scene understanding (Choi *et al.*, 2013). Most notably, our approach differs as it uses the layout information for the purpose of disambiguation, we add an expectation over the matched viewpoints as well as address a different task (augmented reality).

Augmented Reality Application and Maintenance Work Augmented reality application have been studied for over two decades (Azuma, 1997). A recent overview of approaches, techniques and datasets can be found in (Uchiyama and Marchand, 2012) and is beyond the scope of exposition. The predominant body of work deals with registration and matching based on markers, low-level features or single objects. We argue for object centric evidence for ease of deployment. In particular, our use case of maintenance work calls for compositional models of multiple objects that allow for object disambiguation within the 3D context. We are not aware of previous efforts of establishing a public dataset for this purpose.

8.3 OBJECT DISAMBIGUATION

As outlined before 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. For disambiguating multiple visual identical parts we fuse the object detector output with a SLAM system that allows us resolve ambiguities by reasoning over the spatial context. Figure 8.1 shows an overview of our system.

2D Object Detection At the core of our model are objects of which a machine is composed off. In order to localize them at test time we investigate a set of recent detectors: LINE-MOD2D (Hinterstoisser *et al.*, 2011), cascade with haar features (Viola and Jones, 2001) as well as HOG features (Dalal and Triggs, 2005), color-DPM (Shahbaz Khan *et al.*, 2012). As such models are all learning-based we can easily

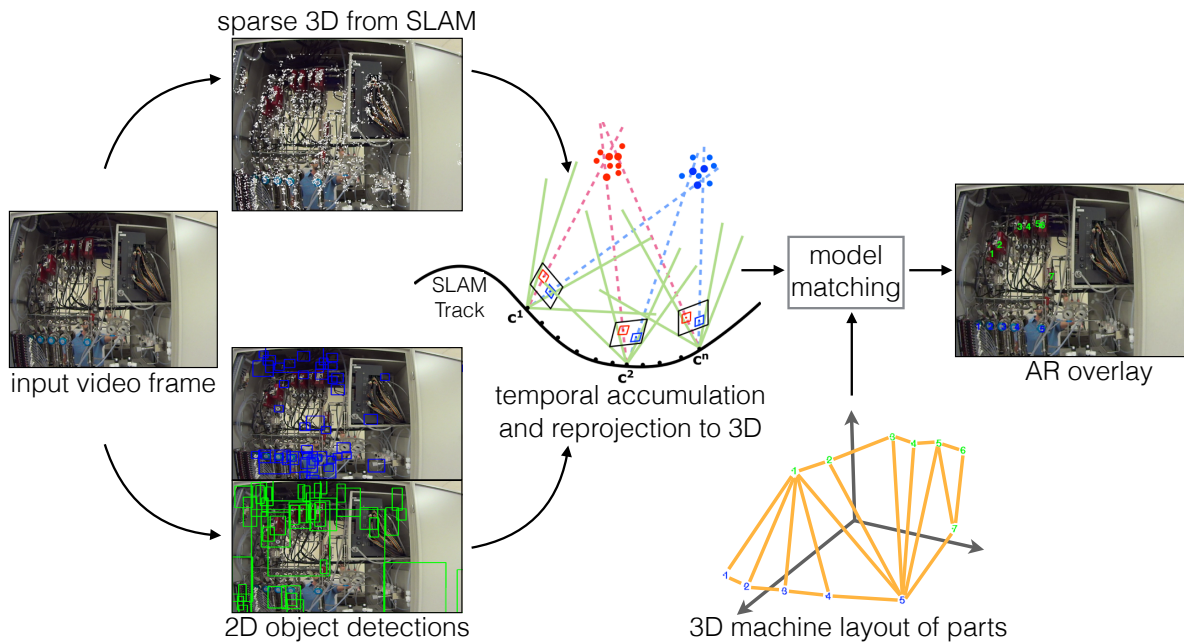


Figure 8.1: Overview of our system for object disambiguation.

adopt our model to new machines and scenarios by training new detectors from examples and plugging them into our model. While the instance based detectors as well as the cascades are fast by design also the more complex detectors have seen recent extension so that they can be computed at interactive rates (Song *et al.*, 2012; Dean *et al.*, 2013; Kokkinos, 2013).

Sparse 3D from SLAM A sparse 3D point cloud is extracted from video using a monocular simultaneous localization and mapping (SLAM) system (Klein and Murray, 2007). An extrinsic camera matrix is estimated based on the set of map points visible in the current frame, and the map is expanded as the camera is moved. The use of monocular image-based SLAM avoids the need for specialized sensors but also introduces challenges. In particular, tracking can fail if insufficient map points are visible (e.g. due to severe motion blur, absence of image features) to reliably triangulate the camera position as well as 3D estimate can be noisy due to complex scene geometry, occlusions and reflective surfaces. Tracking can be reinitialized at the cost of resetting the SLAM coordinate system. Direct use of the sparse 3D information has shown to yield unreliable matches wherefore we opt for integrating 2D and 3D information in the following step.

Temporal Accumulation and Reprojection to 3D We use the sparse 3D information generated by the SLAM system in order to reproject the 2D object detections to 3D. The depth for a particular detection is computed as the average over the covered SLAM features. As all preceding frames are connected by the SLAM track, we

accumulate the reprojected 2D object detections over time. The benefits are threefold. First, object evidence is accumulated over time and can therefore compensate for missing or weak 2D detections in individual frames. Second, potential lag of the detection system can be compensated for as detections from previous frames are already available. Third, partial and ambiguous views of the machine that occur due to zooming in or shifting the viewpoint can be compensated due to previous viewpoints.

3D Machine Layout of Parts We require a 3D machine layout that specifies the relative locations of each object. Such description are often provided by the machine specifications. Please note that the model does not have to be metric – nor do we require a complete 3D model or 3D scan of the machine. This is desirable for easy deployment and adaptation to new scenarios as a complete model can be specified by providing object detectors and a 3D layout.

Model Matching 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 ($E_{\text{appearance}}$), deformation of the layout ($E_{\text{deformation}}$), scale (E_{scale}), viewpoint ($E_{\text{viewpoint}}$) 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, \dots, d_N as well as a projection matrix M so that the following objective:

$$\operatorname{argmin}_{d_1, d_2, \dots, d_N, M} E_{\text{deformation}} + E_{\text{appearance}} + E_{\text{scale}} + E_{\text{viewpoint}} \quad (8.1)$$

where

$$\begin{aligned} E_{\text{deformation}} &= \frac{\sum_{n=1}^N \delta_n}{N} \sum_{n=1}^N \delta_n \cdot \log(\|\bar{M}(P_{g_n}) - P_{d_n}\|) \\ E_{\text{appearance}} &= - \sum_{n=1}^N \delta_n \cdot A_{d_n} \\ E_{\text{scale}} &= \begin{cases} 0, & \bar{s} \in [\mu_s - 2 \cdot \sigma_s, \mu_s + 2 \cdot \sigma_s] \\ \infty, & \text{otherwise} \end{cases} \\ E_{\text{viewpoint}} &= \begin{cases} 0, & \bar{x} \in [\mu_x - 2 \cdot \sigma_x, \mu_x + 2 \cdot \sigma_x], \forall x = \{\alpha, \beta, \gamma\} \\ \infty, & \text{otherwise} \end{cases} \end{aligned} \quad (8.2)$$

P_{g_n} and P_{d_n} denotes the 3D coordinate of g_n and d_n , while A_{d_n} is the detection score of the match d_n . The indicator variable δ_n is for handling the non-matched machine parts, where $\delta_n = 1$ if $\|\bar{M}(P_{g_n}) - P_{d_n}\|$ smaller than a threshold ε , and $\delta_n = 0$ otherwise. The 3D transformation $M(\cdot)$ includes the scale factor s , rotation matrix composed of three rotation angles $\{\alpha, \beta, \gamma\}$ and also a translation vector t . From the training videos of each machine, we compute the distribution of the scale

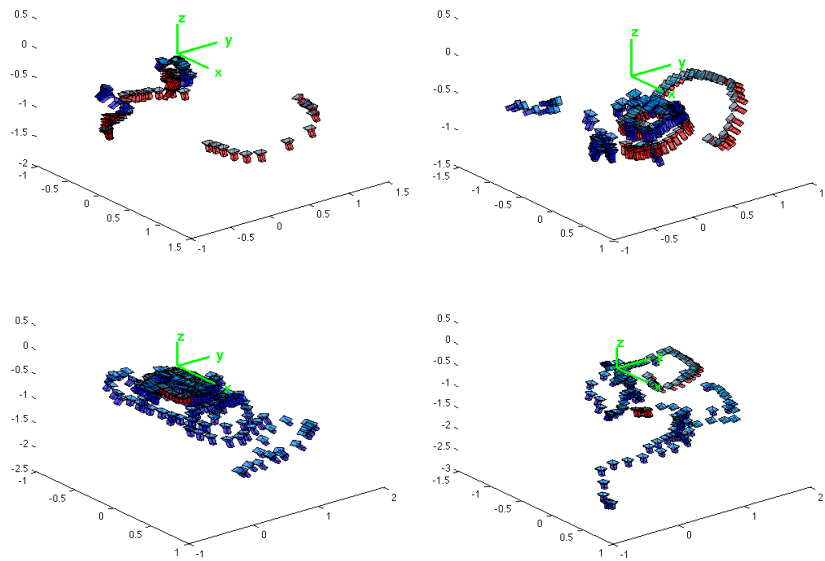


Figure 8.2: Visualization of the distribution for viewpoints in each machine. The red cameras are from the testing videos while the blue ones are from the training sets. The coordinate system is based on the 3D machine layout.

factors and the rotation angles to get their mean μ and standard deviations σ . In the energy terms for both scale E_{scale} and the viewpoint $E_{\text{viewpoint}}$, we hard-constraint the scale factor \bar{s} and rotation angles $\bar{x}, \forall x = \{\alpha, \beta, \gamma\}$ extracted from estimate 3D transformation \bar{M} to be within 2 times of standard deviation from the mean. Figure 8.2 shows the viewpoints of training (blue) and testing (red) in the coordinate system of the machine layout.

In order to minimize the objective, we follow a RANSAC (Fischler and Bolles, 1981) pipeline by randomly selecting candidate alignments between the detections and the machine layout which results in an initial geometric transformation. According to this initial fitting, we iteratively refine the estimate (Besl and McKay, 1992) and re-associate the transformed groundtruth points to the closest detection points.

8.4 EXPERIMENTS

We propose the first benchmark for an object disambiguation task in maintenance work that is composed of an annotated dataset as well as a metric that approximates human judgement. Furthermore, we evaluate our proposed model as well as its components.

8.4.1 Object Disambiguation DataSet (ObDiDaS)

We present the first annotated dataset that allows to quantify performance on a object disambiguation task as it frequently occurs in augmented reality settings and assistance for maintenance work. The dataset captures 4 machines composed of 13 components. Each machine is built of a subset of these potentially repeating components that occur in different spatial arrangements. We provide 14 videos with different viewing scenarios. For each videos we provide human annotation on every 60 frames (at 30fps), with in total 249 frames annotated and 6244 object annotations that specify the type as well as a unique identity. We take one video per machine as testing set and the rest is used for training. Examples are shown in the Figure 8.3. There are various types of difficulties in this dataset, including the wide changes in viewing angles of different object classes, occlusions and motion blur in the videos, reflective surfaces. The dataset allows studies of machine part detection and disambiguation, combination of 2D and 3D cues based on monocular input, generalization between machines and adaptation to new scenarios.

8.4.2 Object Disambiguation Metrics

While object detection metrics assess the performance of object localization in isolation, 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.

Given a video frame with the SLAM extrinsic matrix H and the ground-truth annotation of N visible machine parts by bounding boxes B_{gt} . By using H to project the matches in RANSAC to this frame as bounding boxes, we denote the M visible ones with B_{est} . For each bounding box in ground-truth annotation or RANSAC estimation, they have the labels of their object classes and instance ids. (Note that we define $C(\cdot)$ and $I(\cdot)$ as functions to get the object class label and instance id of the bounding box)

Pascal Object Detection Criterion [Pascal] Inspired by the Pascal Challenge (Everingham *et al.*, 2015), for each $b_{gt}^n, n = 1 \dots N$, we find the corresponding bounding box b_{est}^m with the same class label $C(b_{gt}^n)$ and instance id $I(b_{gt}^n)$ as b_{gt}^n from B_{est} , and measure the intersect-over-union metric between b_{gt}^n and b_{est}^m :

$$O(b_{gt}^n, b_{est}^m) = \frac{b_{gt}^n \cap b_{est}^m}{b_{gt}^n \cup b_{est}^m} \quad (8.3)$$

Then we define the Pascal metric as:

$$S_{pascal} = \frac{1}{N} \sum_{n=1}^N \rho_n, \text{ where } \rho_n = \begin{cases} 1, & O(b_{gt}^n, b_{est}^m) > th \\ 0, & \text{otherwise} \end{cases} \quad (8.4)$$

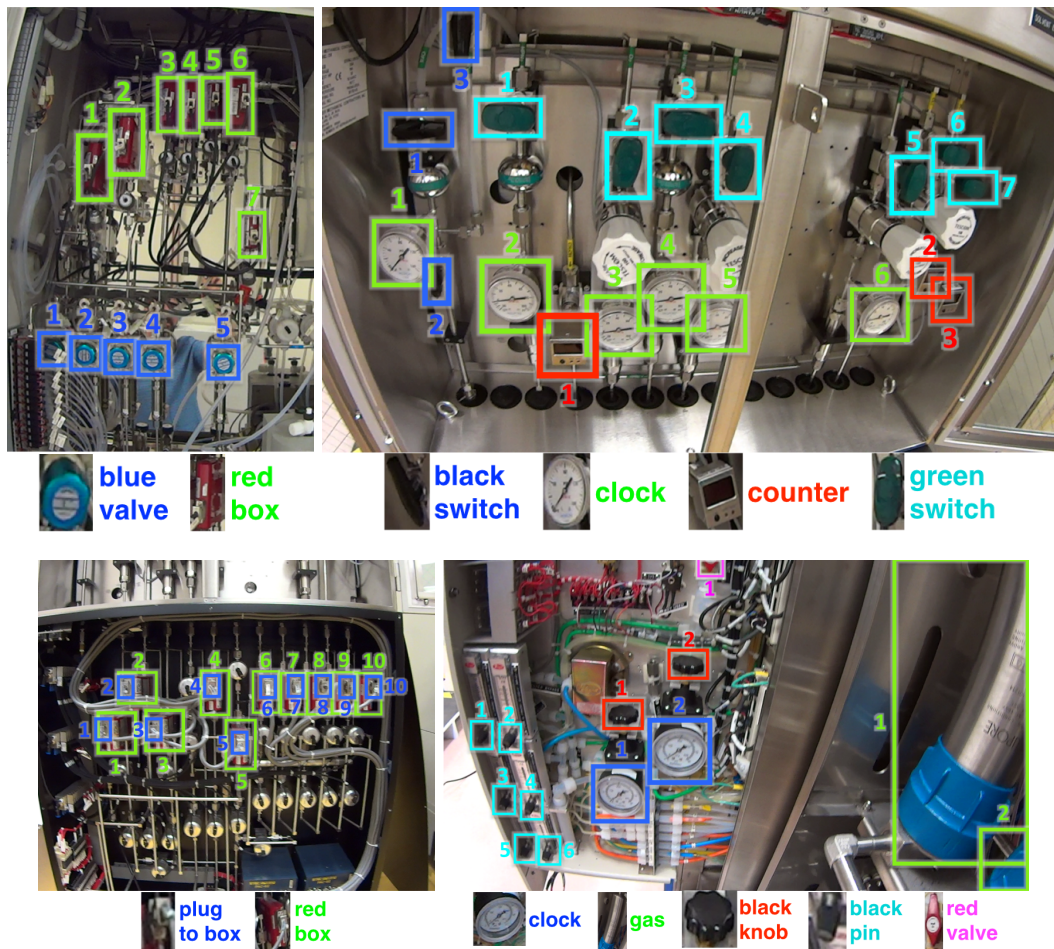


Figure 8.3: Example images for the dataset. In each image we use different color codes for different classes of machine parts. And each instance of the machine parts are labelled with unique identities of the machine.

The variable th is the overlapping threshold, which we set it to be 0.001 in our experiments.

Nearest Neighbor (within/across) We define the pairwise distance between b_{gt}^n and b_{est}^m as the euclidean distance between their box centers in the image coordinate. For each $b_{gt}^n, n = 1 \dots N$, we find its nearest neighbor $b_{est}^{NN_{within}}$ from B_{est} with the same object class label: $B_{est}^C = \{b_{est}^m | C(b_{est}^m) = C(b_{gt}^n)\}$. Then we define the NN_{within} metric as:

$$S_{NN_{within}} = \frac{1}{N} \sum_{n=1}^N \rho_n, \text{ where } \rho_n = \begin{cases} 1, & I(b_{gt}^n) = I(b_{est}^{NN_{within}}) \\ 0, & \text{otherwise} \end{cases} \quad (8.5)$$

Instead of finding the nearest neighbor with the same object class label, in metric NN_{across} we extend to search from all the bounding boxes in B_{est} , we denote the found nearest neighbor as $b_{est}^{NN_{across}}$. Then the metric NN_{across} is represented as:

$$S_{NN_{across}} = \frac{1}{N} \sum_{n=1}^N \rho_n, \text{ where } \rho_n = \begin{cases} 1, & C(b_{gt}^n) = C(b_{est}^{NN_{across}}) \text{ and } I(b_{gt}^n) = I(b_{est}^{NN_{across}}) \\ 0, & \text{otherwise} \end{cases} \quad (8.6)$$

One-to-One (within/across) In comparison to computing the nearest neighbor, we further restrict to have one-to-one matching between b_{gt}^n and b_{est}^m and turn it to be a weighted bipartite matching scenario, where the weights are the $dist(b_{gt}^n, b_{est}^m)$. We use Hungarian method (Kuhn, 2010) to solve this problem. Assume there are in total L object classes shown in this video frame, for each class l we build up the distance matrix by $B_{gt}^l = \{b_{gt}^{n_l} | C(b_{gt}^{n_l}) = l\}$ and $B_{est}^l = \{b_{est}^{m_l} | C(b_{est}^{m_l}) = l\}$. Then for each $b_{gt}^{n_l}$ we have the match $b_{est}^{one_{within}^l}$ after applying Hungarian method. We define the one_{within} metric as:

$$S_{one_{within}} = \frac{1}{N} \sum_{l=1}^L \sum_{n_l=1}^{N_l} \rho_{n_l}, \text{ where } \rho_{n_l} = \begin{cases} 1, & I(b_{gt}^{n_l}) = I(b_{est}^{one_{within}^l}) \\ 0, & \text{otherwise} \end{cases} \quad (8.7)$$

Similar in nearest-neighbor metrics, we can also extend to do the one-to-one matching across classes. Hence we build up the distance matrix between B_{gt} and B_{est} . For each b_{gt}^n we have the match $b_{est}^{one_{across}}$. and the metric one_{across} is written as:

$$S_{one_{across}} = \frac{1}{N} \sum_{n=1}^N \rho_n, \text{ where } \rho_n = \begin{cases} 1, & C(b_{gt}^n) = C(b_{est}^{one_{across}}) \text{ and } I(b_{gt}^n) = I(b_{est}^{one_{across}}) \\ 0, & \text{otherwise} \end{cases} \quad (8.8)$$

Evaluation of Metrics In Table 8.1 we compare the proposed metrics to actual human judgement. We use the output of our full model. For the human judgement,

	machine 1	machine 2	machine 3	machine 4	average
Human Judge.	74.12%	100.00%	99.68%	70.57%	86.09%
Pascal	60.92%	98.68%	95.60%	25.10%	70.08%
NN (within)	57.05%	94.76%	88.06%	72.88%	78.19 %
NN (across)	56.07%	91.97%	65.20%	56.84%	67.52 %
1-to-1 (within)	77.55%	99.18%	99.68%	79.25%	88.92%
1-to-1 (across)	74.63%	96.92%	93.10%	72.45%	84.28 %

Table 8.1: Evaluation of different metrics.

	LINE-MOD	Haar cascade	HoG cascade	LBP cascade	color-DPM
avg. precision	10.81%	8.37%	13.38 %	8.90 %	36.73 %

Table 8.2: Evaluation of 2D object detectors.

we present the produced overlay to a human observer and assess in how many cases the correct object was identified. We observe that pascal metric significantly underestimates the system performance. We attribute this to an implicit matching that the human observer performs between the overlay and the observed machine parts. The nearest neighbor metric narrows the gap – at least for the case of matching within the object types (NN_{within}). The closest match to the true performance is obtained by the one-to-one metric. It takes further into account that the human observer also makes use of the context in order to align the overlay with the observed objects. As the “within” variant overestimates the performance we suggest and use the one-to-one (across) metric in the following experiments. A more detailed analysis of correlation scores on the individual object level has yielded the same ranking of metrics.

8.4.3 Evaluation

2D Detectors in Isolation We compare a range of 2D object recognition/detection algorithms on our new dataset: LINE-MOD_{2D} (Hinterstoisser *et al.*, 2011), cascades with haar features (Viola and Jones, 2001) or histogram of gradient features (Dalal and Triggs, 2005) and color-DPM (Shahbaz Khan *et al.*, 2012). Table 8.2 shows average precision scores for the individual methods averaged across all objects and machines. This evaluation uses the pascal criterion as it evaluates object detection in isolation. We conclude that the color-DPM model outperforms the competitors by a large margin on this task. Therefore we will use it as a object detector throughout our experiments.

Full and Partial Models on Object Disambiguation Task We evaluate our full model as well as switching energy terms off one at a time in order to provide further insights. Table 8.3 shows the individual performance numbers of the object disambiguation task (under the one-to-one-across metric), Figure 8.4 shows example

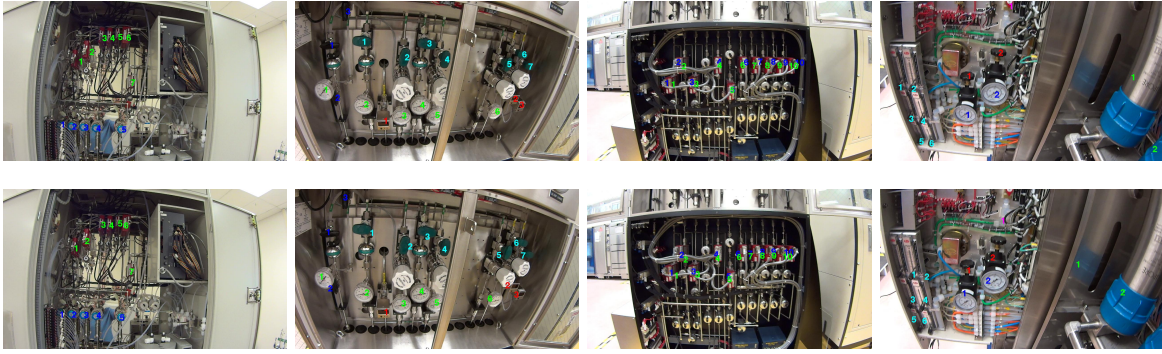


Figure 8.4: Example results. First row are examples for the groundtruth of each machine. Second row are the corresponding results from our proposed method.

	machine 1	machine 2	machine 3	machine 4	average
full model	74.63%	96.92%	93.10%	72.45%	84.28 %
no appearance	67.29%	93.32%	64.05%	51.06%	68.93%
no deformation	83.89%	95.05%	61.44%	40.30%	70.17%
no scale constraint	67.29%	98.53%	53.94%	43.57%	65.84%
no viewpoint constraint	38.01%	88.89%	43.04%	10.21%	45.04%
no scale and viewpoint	38.01%	88.89%	43.04%	10.21%	45.04%
no non-matched objects	74.61%	74.16%	64.10%	55.65%	67.13%

Table 8.3: Evaluation of different model components.

results of our system in comparison to the groundtruth annotations and Figure 8.5 illustrates the effect on the output if parts of the matching energy are not used. We observe the most dramatic drop in performance if the viewpoint and scale constraints are not used, which results in a performance drop of almost 40%. The corresponding visualizations show that disabling this part of our model leads to estimates that exhibit a strong camera roll or suggest a fit beyond working distance. Appearance and the model deformation seem roughly equally important and both boost the performance by over 10%. Also our explicit treatment of non-matched objects is similarly important. Effects can again be observed in Figure 8.5 where a mismatch caused by a partial visible machine is remedied by the full model.

While our full model shows strong performance on machine 2 and 3, there is still a need for improvement on the other two. We attribute the missing performance to reflective surfaces (mirror in the back) that cause problems to the SLAM and detection system, complex 3D structure of machine layout, weak evidence from detector for certain objects and background clutter.

8.5 CONCLUSION

We have investigated a object disambiguation task in a markerless augmented reality scenario, where object identities are inferred from monocular input by exploiting contextual information. To the best of our knowledge, we present the first dataset that allows to quantify the performance of such a system. We propose different metrics and compare them to human judgement. Our proposed metric gives a more realistic estimate of the system performance than a traditional object detection metric that consistently underestimates the system performance. Finally, we present an automatic system for object disambiguation that shows strong performance due to a matching formulation that is based on a composite energy function. We analyse the contribution of each component which underlines in particular the importance of modelling expectations over viewpoints and scales in the matching process.

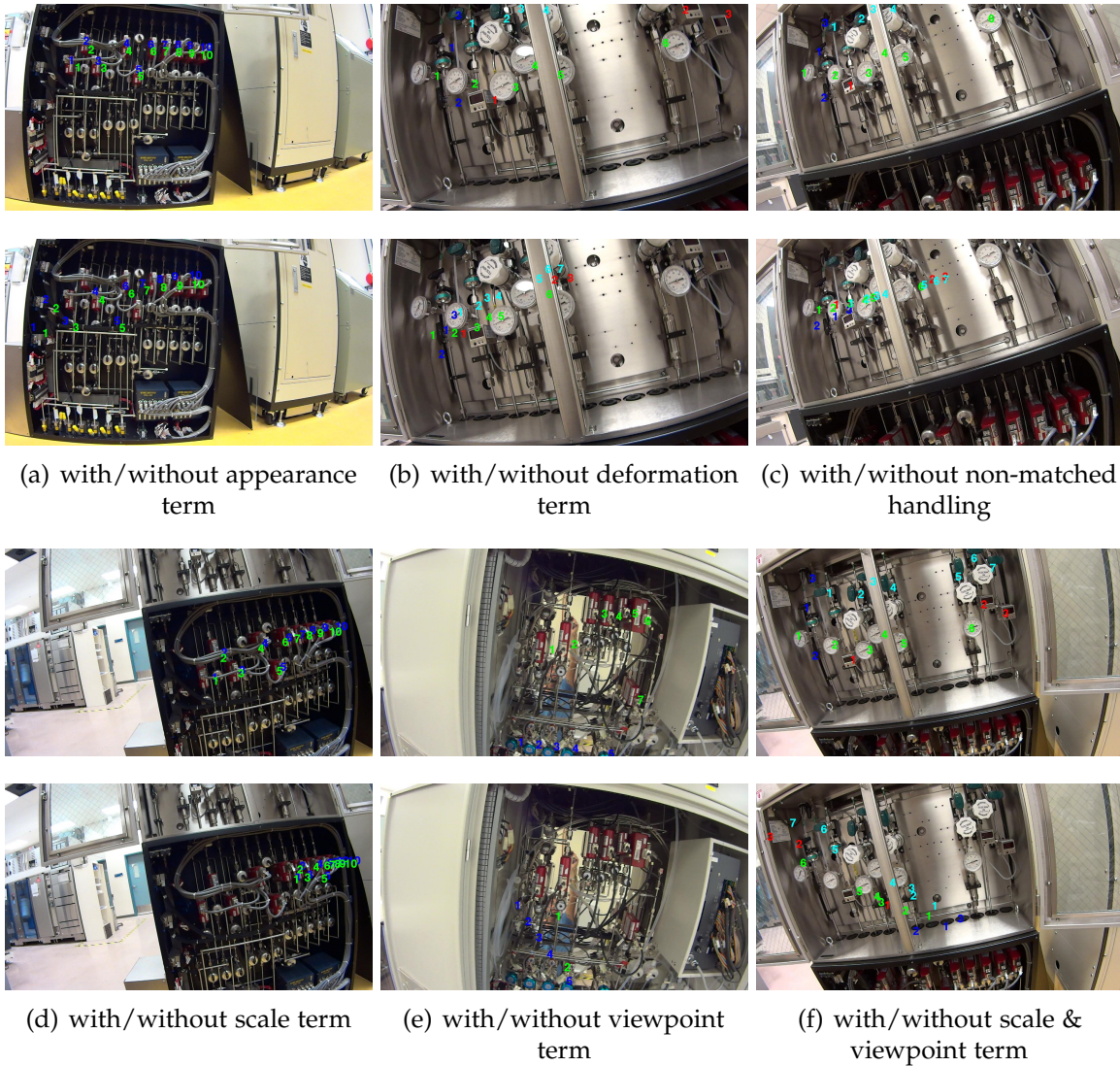


Figure 8.5: Top figure shows output of full model; while in bottom has a particular energy switched off (a)with/without appearance term (b)with/without deformation term (c)with/without non-matched objects handling (d)with/without scale term (a)with/without viewpoint term (a)with/without scale and viewpoint term

Contents

9.1 Future Directions	130
---------------------------------	-----

WE have explored different approaches to multi-modal segmentation. The investigations are organized along 3 axes: video segmentation and object discovery (Chapter 3), activity segmentation and discovery (Chapter 4), and segmentation in 3D data (Chapter 5, 6, 7, 8). Several application scenarios related to these axes are presented, such as: multi-class video co-segmentation, sketch-based video retrieval, 3D pose estimation, stereo video segmentation, and object disambiguation for augmented reality scenario. In the following we summarize them all.

Multi-Modal Video Segmentation and Object Discovery We address multi-class video co-segmentation problem by employing non-parametric Bayesian approaches to model the generative procedure of multiple video sequences, where the complexity of the model, i.e. number of segments, is determined by data. The hierarchical structure manages to unsupervisedly discover the object classes of enriched appearance models across videos, as well as outline object instances with considering spatio-temporal and motion dependencies between basic data units. The experimental results show that the proposed model is managed to resolve the ambiguities of appearance and motion patterns as well as improve the segmentation results via joint segmentation across videos.

Multi-Modal Activity Segmentation and Discovery The non-parametric Bayesian formulation as in video co-segmentation task is generalized to discover activities from context sensor data, by using supersamples composed of context words as the basic data units and modelling their dependencies based on temporal distance together with word2vec semantic information. The data-driven segmentation overcomes the problem of time-invariant sliding windows, the non-parametric framework avoids the manual assignment of cluster numbers which is necessary for the previous state-of-the-art LDA approach, and the semantic distance between context word labels provides a more informative prior for activity discovery in comparison to the non-parametric CRF framework.

Multi-Modal Segmentation in 3D Data We introduce several works for segmentation and manipulation on different sources of 3D information, including 1. 3D CAD models, 2. depth sensors, 3. consumer stereo cameras, and 4. SLAM system, as listed by level of detail. Particularly: First, in Chapter 5 we realize the piecewise differentiability of HOG feature representation, and present an end-to-end optimization

scheme for CAD-to-image alignment task based on the differentiable HOG descriptor and an approximate renderer. Second, in Chapter 6 we present an efficient ensemble model to combine a pool of heterogeneous segmentation algorithms for the stereo video segmentation task. The importances of different pooled segmentations and the weights of feature cues which measure the distances between voxels are learnt discriminatively to be adaptively adjusted in accordance to the data statistics of target stereo video. Third, in Chapter 7 we propose a simple and effective cross-modal stereo path obtained from disparity matching between the IR and RGB sensors on Kinect in order to complement Kinect's difficulties on specular and transparent surfaces. The performance of 3D object segmentation based on the improved 3D sensing is boosted in a large margin. And finally, in Chapter 8 we develop a novel multi-part object detection system that fuses the 3D context with detection outputs of machine parts, which are lifted from 2D to 3D by sparse 3D information SLAM, in order to tackle the object disambiguation problem. The application scenario is to provide a maintenance worker an augmented reality overlay where the potentially repetitive machine components are disambiguated and labelled by unique IDs.

In addition to the performance improvements demonstrating the benefits of our methods on various applications related to 3D data, another main contribution in this axis of works is to resolve the ambiguities shown in multi-modal data. For instance: while there are certain modalities having less informative observations, the adaptive segmentation scheme learns to decrease the importances of those feature cues. In the work of improving Kinect, the cross-modal stereo enables to detect disparities at edges of transparent or reflective objects which are difficult for Kinect, while on textureless area (e.g. the wall or desk) we trust more on the depth information perceived by Kinect instead of stereo vision. In multi-part object detection system the 3D context information is used to disambiguate machine parts of the same category but with different functions.

Overall, this thesis contributes to develop multi-modal segmentation approaches that manage to combine and make better use of information from multiple modalities, resolve ambiguities, and cope with noisy 3D observations. As the segmentation provides an intuitive way to understand the latent structure of the data, where it divides the data into multiple semantic groups based on some characteristics, the works in this thesis can become the building blocks for further applications, such as semantic labelling, scene understanding, and content-based video retrieval. Surely, not all challenges that can be met in segmenting multi-modal data could be fully addressed. There are still many research venues to be explored. The following section discusses several possible directions for future works.

9.1 FUTURE DIRECTIONS

Although in the thesis we divide our works into three axes and tackle different applications, there are opportunities to combine algorithms across the axes into more robust frameworks, and apply them on different scenarios. Additionally, the recent

advances in deep learning and related discriminative models show success in many research areas, such as speech recognition, object recognition, and visual question answering. We propose future research directions that integrate our presented algorithms as well as the research efforts on discriminative learning.

Learning Representations for Non-Parametric Bayesian Models Recent progresses on convolutional neural network and deep learning have shown their power on learning highly discriminative representations (Bengio *et al.*, 2013), especially in comparison to the simple appearance features such as color or SIFT. Although it might seem to be a natural selection of using convnet-based features for the appearance likelihood in the non-parametric Bayesian models, the high dimensionality of convnet-based representations can cause difficulties for some popular non-parametric Bayesian tools (Shah and Ghahramani, 2013), for instance, Dirichlet Process Gaussian Mixture Model (DPGMM). The work from (Socher *et al.*, 2011) proposes a non-parametric clustering model which combines the ddCRP prior based on similarity matrix between the data points on original high-dimensional space, and the likelihood is computed based on the layout of the data points in the low-dimensional space by using spectral methods for dimensionality reduction. This approach denoted as similarity-dependent CRP (sd-CRP) manages to provide a plausible solution for using high-dimensional convnet features in the non-parametric Bayesian framework. An interesting direction for future research would be as follows: Assume we have the a collection of training data with their groundtruth partitions, we would like to learn the specific convnet-based feature representation which is most likely to generate the groundtruth clustering result by applying sd-CRP technique. Recent work (Song *et al.*, 2016) proposes a deep feature embedding and metric learning algorithm with defining a structured loss function based on the lifted *dense pairwise similarity matrix*, which contains the local and global structures of data as required by the prior and likelihood in sd-CRP framework. In which it consequently seems to be a good reference to start with and continue the exploration.

Combining Different Modelling Perspectives While unsupervised clustering is able to discover the structure of large data collection, the algorithms from semantic labelling (Arbeláez *et al.*, 2012; Long *et al.*, 2015) are able to classify each object in the scene, which present more detailed descriptions but are not easily generalized to all the possible classes in the highly complex world. The combination of these two directions would provide a possible scenario: for the common object classes we can utilize the discriminative object detectors to locate them in the videos and even give more fine-grained or 3D representations on the object classes with CAD models available (Chapter 5); while for other classes that are unexplored, rare to see, or hard to collect training data for learning classifiers, the unsupervised clustering as shown in Chapter 3 is capable to partition the visual data into semantic groups for further analysis.

Adaptive Dependencies for ddCRP Learning affinity functions used in the distance-dependent Chinese Restaurant Processes can boost the capacity of our proposed method in Chapter 3 and 4 to model the underlying distribution of data partitions. Recent work from (Ghosh and Sudderth, 2015) explores methods for learning the dependencies in the ddCRP model from human annotated data based on the recent advances in approximate Bayesian computation (ABC). However, their approach does not consider the case that the feature cues are not equally informative over all possible data collections as we have motivated in the Chapter 6, wherefore the dependencies between data units can also have wide variances accordingly. Therefore the adaptive strategy to weight different feature distances in the ddCRP prior depending on the statistical properties of the data becomes an opportunity, which is yet unexplored and expected to improve the segmentation performance.

LIST OF FIGURES

3.1	Illustration for the idea of multi-class video co-segmentation	15
3.2	Visualization of our video co-segmentation model with taking appearance, spatial-temporal and motion features as observed variables . . .	19
	(a) Only observed appearance information is taken into account. . .	19
	(b) Hierarchical Dirichlet Process with modelling spatial-temporal and motion distributions of object instances.	19
3.3	Synthetic sequence to illustrate the proposed approach.	20
3.4	Example of video segmentation results for our model.	21
3.5	Example of learnt global classes across videos in <i>moseg</i> dataset.	23
3.6	Illustration of the capability of our model to handle occlusion.	24
3.7	Summarization of 10 car sequences from <i>moseg</i> dataset.	25
3.8	Example of sketch-based video retrieval in <i>moseg</i> dataset.	25
3.9	Visualization of video segmentation prior.	27
3.10	Samples from video segmentation prior.	28
3.11	Metaphors between the proposed multi-video model w.r.t. Hierarchical Dirichlet Process.	29
3.12	Summary of the proposed MOVICS dataset.	32
3.13	Comparison of co-segmentation accuracies between the proposed method and baselines.	34
	(a) without over-segmentation	34
	(b) allow over-segmentation	34
3.14	Examples of results of the proposed method and multiple baselines in MOVICS dataset.	36
3.15	Example of improved results by the global object class model.	37
3.16	Comparison of co-segmentation accuracies between different granularities of the superpixel segmentations.	38
	(a) without over-segmentation.	38
	(b) with over-segmentation allowed.	38
3.17	Examples of results for running the proposed method on different granularities of superpixel segmentations.	39
3.18	Comparison of proposed models under different design choices of the ddCRP prior for the local object-instance layer. There are 4 different choices are evaluated: i) to use both distances in spatiotemporal and motion domains (Spatial+Motion), to use only ii) spatiotemporal (Spatial Only) or iii) motion (Motion Only) informations, and iii) to use pure CRP instead of ddCRP as prior (HDP).	40
	(a) without over-segmentation.	40
	(b) with over-segmentation allowed.	40

3.19	Examples of results from using various design choices of the video segmentation ddCRP prior.	41
4.1	Illustration of segmentation methods for activity discovery.	48
4.2	Visualization of the proposed framework for activity discovery.	49
4.3	Illustration of the hierarchical discovery framework ddCRP+CRP.	51
4.4	Illustration of the evaluation strategy for the proposed method in comparison to other baselines.	53
4.5	Illustration of semantic distances between activity instances.	55
4.6	Performance evaluation for discovering activities from context word labels.	56
4.7	Influence of evenly distributed noise over context word detectors on discovery performance.	57
4.8	Performance of activity discovery from context word detections.	58
5.1	Piecewise differentiability of HOG for end-to-end optimization on both pre-image reconstruction and pose estimation tasks.	62
5.2	Visualization of the implementation procedure for our ∇ HOG method.	65
5.3	Visualizations of variants of the proposed method for pre-image task.	70
	(a) ∇ HOG multi-scale	70
	(b) ∇ HOG multi-scale-more	70
5.4	Visualization of the similarity and its gradients w.r.t azimuth.	73
5.5	Pipeline visualization of our model for pose estimation.	74
6.1	Overview of proposed efficient adaptive stereo segmentation technique.	78
6.2	Sample frames from Consumer Stereo Video Segmentation Challenge (CSVSC) dataset.	79
6.3	Sample disparity estimation of CSVSC dataset.	81
6.4	Proposed video segmentation model.	83
6.5	Results of the pooled segmentation and the proposed method on CSVSC dataset using BPR and VPR metrics.	91
7.1	Examples for failure cases of Kinect's depth sensing and the visualization of our evaluation scheme on 3D object segmentation task.	100
7.2	Visualization of different stages of the proposed cross-modal stereo.	101
	(a) RGB image.	101
	(b) IR image with Kinect's IR-projector pattern.	101
	(c) IR image with covered IR-projector.	101
	(d) Rectified depth map from Kinect.	101
	(e) Disparity map from cross-modal stereo.	101
	(f) Fused depth maps.	101
7.3	Visualization of improved cross-modal disparity map after applying optimized color channel weights on RGB image.	103
	(a) RGB image converted by the optimized weights.	103
	(b) IR image with covered IR-projector.	103
	(c) RGB image in gray scale.	103
	(d) Disparity map from (a, b)	103
	(e) Disparity map from (c, b)	103

7.4	Precision Recall and average precision for the table-dataset	105
7.5	Example images from dataset.	106
7.6	Impact on cross-modal stereo under different lighting conditions. . .	107
	(a) Disparity b/w grayscale RGB & IR	107
	(b) Disparity b/w weighted RGB & IR	107
	(c) Learned weights for RGB under various lighting conditions. . . .	107
7.7	Schematic for experimental setup for reading sensor characteristics. .	108
7.8	Spectrum from experiment of reading sensor characteristics.	109
7.9	Visualization of the spectrum for different sensor channels.	110
7.10	Diagrams for weighted fusion scheme.	110
7.11	Visualization of improvement based on patch-filtered cross-modal stereo.	111
	(a) Response of RGB & IR cameras.	111
	(b) RGB image.	111
	(c) IR image under projected pattern.	111
	(d) Disparity on unfiltered stereo.	111
	(e) Disparity on patch-filtered stereo	111
7.12	Quantitative comparison between the proposed method of learning optimal filters and several baselines.	112
7.13	Comparison of disparity maps under different settings.	113
	(a) Disparity from original stereo.	113
	(b) Disparity with projector off.	113
	(c) Disparity with patch-filtered stereo.	113
7.14	visualization of the filters obtained from optimization process. . . .	114
8.1	Overview of our system for object disambiguation.	119
8.2	Visualization of the distribution for viewpoints in each machine. . . .	121
8.3	Example images for the dataset.	123
8.4	Example results.	126
8.5	Visualization for the contributions of different energy terms in the proposed model.	128
	(a) with/without appearance term	128
	(b) with/without deformation term	128
	(c) with/without non-matched handling	128
	(d) with/without scale term	128
	(e) with/without viewpoint term	128
	(f) with/without scale & viewpoint term	128

LIST OF TABLES

Tab. 3.1	Performance comparison for video segmentation results.	22
Tab. 3.2	Runtime comparison for the proposed method and baselines. . .	41
Tab. 5.1	Example results for different variants of the proposed pipeline for pre-image task.	71
Tab. 5.2	Comparison on the performance of reconstruction from feature descriptors.	72
Tab. 5.3	Comparison on the performance of pose estimation.	74
Tab. 5.4	Example results for pose estimation.	75
Tab. 5.5	Example results for image reconstruction from feature descriptors.	76
Tab. 6.1	Results on the CSVSC benchmark.	92
Tab. 6.2	Analysis of the proposed EASVS.	92
Tab. 6.3	Examples of the proposed EASVS optimized for different evalua- tion metrics compared to the state-of-the-art algorithms.	93
Tab. 6.4	Additional examples of the proposed EASVS compared to baselines.	94
Tab. 8.1	Evaluation of different metrics.	125
Tab. 8.2	Evaluation of 2D object detectors.	125
Tab. 8.3	Evaluation of different model components.	126

BIBLIOGRAPHY

- R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk (2010). Slic superpixels, Technical report, École polytechnique fédérale de Lausanne. Cited on page 86.
- S. Agarwal, K. Branson, and S. Belongie (2006). Higher order learning with graphs, in *Proceedings of the International Conference on Machine Learning (ICML) 2006*. Cited on page 85.
- P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik (2012). Semantic segmentation using regions and parts, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 131.
- P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik (2011). Contour detection and hierarchical image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33(5), pp. 898–916. Cited on pages 38, 39, 80, 82, 86, and 88.
- M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic (2014a). Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 61, 62, 63, 64, 67, 71, 72, 73, and 74.
- M. Aubry, B. C. Russell, and J. Sivic (2014b). Painting-to-3D model alignment via discriminative visual elements, *ACM Transactions on Graphics (TOG)*, vol. 33(2), p. 14. Cited on pages 62 and 67.
- A. Aztiria, J. C. Augusto, R. Basagoiti, A. Izaguirre, and D. J. Cook (2012). Discovering frequent user–environment interactions in intelligent environments, *Personal and Ubiquitous Computing (PUC)*, vol. 16(1), pp. 91–103. Cited on page 45.
- R. T. Azuma (1997). A survey of augmented reality, *Presence: Teleoperators and virtual environments*, vol. 6(4), pp. 355–385. Cited on page 118.
- T. S. Barger, D. E. Brown, and M. Alwan (2005). Health-status monitoring through analysis of behavioral patterns, *IEEE Transactions on Systems, Man and Cybernetics (SMC)*, vol. 35(1), pp. 22–27. Cited on page 46.
- W. Barhoumi (2015). Detection of highly articulated moving objects by using co-segmentation with application to athletic video sequences, *Signal, Image and Video Processing*, vol. 9(7), pp. 1705–1715. Cited on page 9.
- T. Basha, Y. Moses, and N. Kiryati (2013). Multi-view scene flow estimation: A view centered variational approach, *International Journal of Computer Vision (IJCV)*, vol. 101(1), pp. 6–21. Cited on page 10.

- D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen (2010). iCoseg: Interactive co-segmentation with intelligent scribble guidance, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 32.
- J. Begole, J. Tang, and R. Hill (2003). Rhythm modeling, visualizations and applications, in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST) 2003*. Cited on page 46.
- Y. Bengio, A. Courville, and P. Vincent (2013). Representation learning: A review and new perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35(8), pp. 1798–1828. Cited on page 131.
- P. J. Besl and N. D. McKay (1992). Method for registration of 3-D shapes, in *Robotics-DL tentative 1992*. Cited on page 121.
- D. Blei and P. Frazier (2010). Distance dependent Chinese restaurant processes, in *Proceedings of the International Conference on Machine Learning (ICML) 2010*. Cited on page 30.
- D. Blei, A. Ng, and M. Jordan (2003). Latent dirichlet allocation, *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 993–1022. Cited on page 54.
- M. Bleyer, C. Rhemann, and C. Rother (2012). Extracting 3d scene-consistent object proposals and depth from stereo images, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 10.
- G. Bradski (2000). The OpenCV Library, *Dr. Dobb's Journal of Software Tools*. Cited on page 98.
- T. Brox and J. Malik (2010). Object segmentation by long term analysis of point trajectories, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 13, 21, 22, 32, 34, 41, and 42.
- D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black (2012). A naturalistic open source movie for optical flow evaluation, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 79.
- A. Chambolle and T. Pock (2011). A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision*, vol. 40(1), pp. 120–145. Cited on pages 27, 41, and 42.
- K.-Y. Chang, T.-L. Liu, and S.-H. Lai (2011). From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 8.

- D.-J. Chen, H.-T. Chen, and L.-W. Chang (2012). Video object cosegmentation, in *Proceedings of the ACM international conference on Multimedia 2012*. Cited on pages 8 and 9.
- W.-C. Chiu, U. Blanke, and M. Fritz (2011a). I spy with my little eye: Learning optimal filters for cross-modal stereo under projected patterns, in *IEEE International Conference on Computer Vision (ICCV) Workshop on Consumer Depth Cameras for Computer Vision (CDC4CV) 2011*. Cited on pages 4 and 6.
- W. C. Chiu, U. Blanke, and M. Fritz (2011b). Improving the Kinect by Cross-Modal Stereo., in *Proceedings of the British Machine Vision Conference (BMVC) 2011*. Cited on pages 4 and 6.
- W.-C. Chiu and M. Fritz (2013). Multi-class video co-segmentation with a generative multi-video model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 2, 5, 9, 11, 16, 32, 50, 52, 54, and 92.
- W.-C. Chiu and M. Fritz (2015). See the difference: direct pre-image reconstruction and pose estimation by differentiating HOG, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 3 and 5.
- W.-C. Chiu, F. Galasso, and M. Fritz (2016). Towards Segmenting Consumer Stereo Videos: Benchmark, Baselines and Ensembles, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2016*. Cited on pages 3 and 6.
- W.-C. Chiu, G. S. Johnson, D. Mcculley, O. Grau, and M. Fritz (2014). Object disambiguation for augmented reality applications, in *Proceedings of the British Machine Vision Conference (BMVC) 2014*. Cited on pages 4 and 6.
- W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese (2013). Understanding indoor scenes using 3D geometric phrases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 118.
- Y. Choi, T. Deyle, T. Chen, J. Glass, and C. Kemp (2009). A list of household objects for robotic retrieval prioritized by people with ALS, in *Proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR) 2009*. Cited on page 97.
- C. B. Choy, M. Stark, S. Corbett-Davies, and S. Savarese (2015). Enriching object detection with 2D-3D registration and continuous viewpoint estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 64.
- M. D. Collins, J. Xu, L. Grady, and V. Singh (2012). Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 8.

- D. Comaniciu and P. Meer (2002). Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24(5), pp. 603–619. Cited on page 7.
- N. Dalal and B. Triggs (2005). Histograms of oriented gradients for human detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on pages 61, 65, 67, 69, 71, 118, and 125.
- E. d’Angelo, A. Alahi, and P. Vanderghenst (2012). Beyond bits: Reconstructing images from local binary descriptors, in *Proceedings of the International Conference on Pattern Recognition (ICPR) 2012*. Cited on page 68.
- T. Darrell and A. Pentland (1991). Robust estimation of a multi-layered motion representation, in *IEEE Workshop on Visual Motion 1991*. Cited on page 13.
- H. Daumé, III, A. Kumar, and A. Saha (2010). Frustratingly easy semi-supervised domain adaptation, in *Proceedings of the Workshop on Domain Adaptation for Natural Language Processing 2010*. Cited on page 99.
- J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon (2007). Information-theoretic metric learning, in *Proceedings of the International Conference on Machine Learning (ICML) 2007*. Cited on page 99.
- T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik (2013). Fast, Accurate Detection of 100,000 Object Classes on a Single Machine, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 118 and 119.
- M. V. den Bergh and L. J. V. Gool (2012). Real-time stereo and flow-based video segmentation with superpixels, in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) 2012*. Cited on page 11.
- X. Di, H. Chang, and X. Chen (2013). Multi-layer spectral clustering for video segmentation, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2013*. Cited on page 8.
- J. Dong and S. Soatto (2015). Domain-size pooling in local descriptors: DSP-SIFT, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 64.
- I. Endres and D. Hoiem (2010). Category independent object proposals, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 9 and 39.
- M. D. Escobar and M. West (1995). Bayesian density estimation and inference using mixtures., *Journal of the American Statistical Association*. Cited on page 16.
- M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2015). The Pascal visual object classes challenge: A retrospective, *International Journal of Computer Vision (IJCV)*. Cited on pages 72, 104, and 122.

- K. Farrahi and D. Gatica-Perez (2011). Discovering routines from large-scale human locations using probabilistic topic models, *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, p. 3. Cited on pages 45 and 46.
- L. Fei-Fei, R. Fergus, and P. Perona (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories, *Computer Vision and Image Understanding (CVIU)*, vol. 106(1), pp. 59–70. Cited on page 69.
- P. Felzenszwalb, D. McAllester, and D. Ramanan (2008). A discriminatively trained, multiscale, deformable part model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 117 and 118.
- P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan (2010). Object detection with discriminatively trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32(9), pp. 1627–1645. Cited on pages 62, 63, 64, 67, 69, and 71.
- P. F. Felzenszwalb and D. P. Huttenlocher (2004). Efficient graph-based image segmentation, *International Journal of Computer Vision (IJCV)*, vol. 59(2), pp. 167–181. Cited on page 7.
- M. A. Fischler and R. C. Bolles (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM*, vol. 24(6), pp. 381–395. Cited on page 121.
- K. Fragkiadaki, G. Zhang, and J. Shi (2012). Video segmentation by tracing discontinuities in a trajectory embedding, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 8.
- M. Fritz, M. Black, G. Bradski, S. Karayev, and T. Darrell (2009). An additive latent feature model for transparent object recognition, in *Advances in Neural Information Processing Systems (NIPS) 2009*. Cited on page 98.
- M. Fritz, K. Saenko, and T. Darrell (2010). Size matters: Metric visual search constraints from monocular metadata, in *Advances in Neural Information Processing Systems (NIPS) 2010*. Cited on pages 99 and 118.
- H. Fu, D. Xu, B. Zhang, and S. Lin (2014). Object-based multiple foreground video co-segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 9, 11, 33, and 39.
- A. Fusiello and L. Irsara (2008). Quasi-euclidean uncalibrated epipolar rectification, in *Proceedings of the International Conference on Pattern Recognition (ICPR) 2008*. Cited on page 80.
- F. Galasso, M. Iwasaki, K. Nobori, and R. Cipolla (2011). Spatio-temporal clustering of probabilistic region trajectories, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 8 and 13.

- F. Galasso, M. Keuper, T. Brox, and B. Schiele (2014). Spectral graph reduction for efficient image and streaming video segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 38, 84, and 86.
- F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele (2013). A unified video segmentation benchmark: Annotation, metrics and analysis, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 80, 88, and 90.
- A. Geiger, M. Roser, and R. Urtasun (2010). Efficient large-scale stereo matching, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2010*. Cited on pages 80 and 81.
- S. Ghosh and E. B. Sudderth (2015). Approximate Bayesian computation for distance-dependent learning, in *Advances in Neural Information Processing Systems (NIPS) Workshop on Bayesian Nonparametrics: The Next Generation 2015*. Cited on page 132.
- S. Ghosh, A. Ungureanu, E. Sudderth, and D. Blei (2011). Spatial distance dependent Chinese restaurant processes for image segmentation, in *Advances in Neural Information Processing Systems (NIPS) 2011*. Cited on page 27.
- R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014). Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 64.
- S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller (2008). Integrating Visual and Range Data for Robotic Object Detection, in *European Conference on Computer Vision (ECCV) Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2) 2008*. Cited on pages 99 and 118.
- M. Grant and S. Boyd (2011). *CVX: Matlab Software for Disciplined Convex Programming, version 1.21*, <http://cvxr.com/cvx>. Cited on page 111.
- M. Grundmann, V. Kwatra, M. Han, and I. Essa (2010). Efficient hierarchical graph-based video segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 8, 79, 82, 86, 91, 92, and 93.
- T. Gu, S. Chen, X. Tao, and J. Lu (2010). An unsupervised approach to activity recognition and segmentation based on object-use fingerprints, *Data & Knowledge Engineering (DKE)*, vol. 69(6), pp. 533–544. Cited on page 46.
- J. Guo, L.-F. Cheong, R. T. Tan, and S. Zhiying (2014). Consistent foreground co-segmentation, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2014*. Cited on page 9.

- J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou (2013). Video co-segmentation for meaningful action extraction, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 9.
- B. Hariharan, J. Malik, and D. Ramanan (2012). Discriminative decorrelation for clustering and classification, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 64.
- S. Hickson, S. Birchfield, I. Essa, and H. Christensen (2014). Efficient hierarchical graph-based segmentation of RGBD videos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 11, 79, 92, and 93.
- S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit (2011). Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 117, 118, and 125.
- H. Hirschmüller and D. Scharstein (2009). Evaluation of stereo matching costs on images with radiometric differences, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, pp. 1582–1599. Cited on page 99.
- T. K. Ho, J. J. Hull, and S. N. Srihari (1994). Decision combination in multiple classifier systems, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 16(1), pp. 66–75. Cited on page 54.
- D. H. Hu, X.-X. Zhang, J. Yin, V. W. Zheng, and Q. Yang (2009). Abnormal activity recognition based on HDP-HMM models., in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2009*. Cited on page 46.
- D. Huang, Y. Tian, and F. De la Torre (2011). Local isomorphism to solve the pre-image problem in kernel methods, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 63.
- F. Huguet and F. Devernay (2007). A variational method for scene flow estimation from stereo sequences, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. Cited on page 10.
- T. Huynh, M. Fritz, and B. Schiele (2008). Discovery of activity patterns using topic models, in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2008*. Cited on pages 45, 46, and 47.
- I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich (2008). State of the art in transparent and specular object reconstruction, in *STAR Proceedings of Eurographics 2008*. Cited on page 98.
- C. Ionescu, O. Vantzosy, and C. Sminchisescu (2015). Matrix backpropagation for deep networks with structured layers, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on page 88.

- A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell (2011). A category-level 3-D object dataset: Putting the Kinect to work, in *IEEE International Conference on Computer Vision (ICCV) Workshop on Consumer Depth Cameras for Computer Vision (CDC4CV) 2011*. Cited on page 80.
- Y. Jia (2013). *Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding*, <http://caffe.berkeleyvision.org/>. Cited on page 117.
- R. Y. J.J. Zhu, L. Wang and J. Davis. (2008). Fusion of time-of-flight depth and stereo for high accuracy depth maps, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on page 99.
- F. R. B. M. I. Jordan and F. Bach (2004). Learning spectral clustering, in *Advances in Neural Information Processing Systems (NIPS) 2004*. Cited on page 88.
- A. Joulin, F. Bach, and J. Ponce (2012). Multi-class cosegmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 8, 9, 11, 31, 34, 41, and 42.
- A. Joulin, K. Tang, and L. Fei-Fei (2014). Efficient image and video co-localization with frank-wolfe algorithm, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on page 9.
- T. Kanade and M. Okutomi (1994). A stereo matching algorithm with an adaptive window: Theory and experiment, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 16(9), pp. 920–932. Cited on page 10.
- H. Kato and T. Harada (2014). Image reconstruction from bag-of-visual-words, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 62, 63, 67, 68, 69, and 72.
- A. Khoreva, F. Galasso, M. Hein, and B. Schiele (2014). Learning must-link constraints for video segmentation based on spectral clustering, in *German Conference on Pattern Recognition (GCPR) 2014*. Cited on page 38.
- E. Kim, H. Li, and X. Huang (2012). A hierarchical image clustering cosegmentation framework, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 8.
- G. Kim and E. P. Xing (2012). On multiple foreground cosegmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 8, 11, and 33.
- G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade (2011). Distributed cosegmentation via submodular optimization on anisotropic diffusion, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 8.

- Y. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Matusik, and S. Thrun (2009). Multi-view image and ToF sensor fusion for dense 3D reconstruction, in *IEEE International Conference on Computer Vision (ICCV) Workshop on 3D Digital Imaging and Modeling (3DIM) 2009*. Cited on page 99.
- U. Klank, D. Carton, and M. Beetz (2011). Transparent object detection and reconstruction on a mobile platform, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) 2011*. Cited on page 98.
- G. Klein and D. Murray (2007). Parallel tracking and mapping for small AR workspaces, in *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR) 2007*. Cited on page 119.
- I. Kokkinos (2013). Shufflets: Shared mid-level parts for fast object detection, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 118 and 119.
- K. Konolige (1998). Small vision systems: Hardware and implementation, *Robotics Research*, vol. 8, pp. 203–212. Cited on page 99.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems (NIPS) 2012*. Cited on page 64.
- D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari (2010). What’s going on? discovering spatio-temporal dependencies in dynamic scenes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 9.
- H. W. Kuhn (2010). The hungarian method for the assignment problem, in *50 Years of Integer Programming 1958-2008 2010*, pp. 29–47, Springer. Cited on page 124.
- T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka (2015a). Picture: A probabilistic programming language for scene perception, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 64.
- T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum (2015b). Deep convolutional inverse graphics network, in *Advances in Neural Information Processing Systems (NIPS) 2015*. Cited on page 64.
- Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 78.
- J. Lezama, K. Alahari, J. Sivic, and I. Laptev (2011). Track to the future: Spatio-temporal video segmentation with long-range motion cues, in *Proceedings of the*

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 13.
- Y. Li and D. P. Huttenlocher (2008). Learning for stereo vision using the structured support vector machine, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on page 99.
- Z. Li, X.-M. Wu, and S.-F. Chang (2012). Segmentation using superpixels: a bipartite graph partitioning approach, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 11, 79, 83, 84, 86, 91, 92, and 93.
- J. J. Lim, H. Pirsiavash, and A. Torralba (2013). Parsing ikea objects: Fine pose estimation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 61 and 64.
- J. Long, E. Shelhamer, and T. Darrell (2015). Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 131.
- M. Loper (2014). *Chumpy*, <https://github.com/mattloper/chumpy>. Cited on pages 61, 68, and 69.
- M. M. Loper and M. J. Black (2014). Opendr: An approximate differentiable renderer, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on pages 61, 64, 68, 70, 74, and 75.
- Z. Lou and T. Gevers (2014). Extracting primary objects by video co-segmentation, *IEEE Transactions on Multimedia*, vol. 16(8), pp. 2110–2117. Cited on pages 9 and 39.
- M. I. Lourakis and A. A. Argyros (2005). Is Levenberg-Marquardt the most efficient optimization algorithm for implementing bundle adjustment?, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. Cited on page 69.
- D. Lowe (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision (IJCV)*. Cited on page 21.
- D. G. Lowe (1999). Object recognition from local scale-invariant features, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 1999*. Cited on pages 63 and 64.
- O. Mac Aodha, G. J. Brostow, and M. Pollefeys (2010). Segmenting video into classes of algorithm-suitability, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 11.
- A. Mahendran and A. Vedaldi (2015). Understanding deep image representations by inverting them, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 64, 69, and 72.

- T. Malisiewicz, A. Gupta, and A. A. Efros (2011). Ensemble of Exemplar-SVMs for object detection and beyond, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 62 and 63.
- Z.-C. Marton, R. B. Rusu, D. Jain, U. Klank, and M. Beetz (2009). Probabilistic categorization of kitchen objects in table settings with a composite sensor, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2009*. Cited on page 99.
- M. Meilă, S. Shortreed, and L. Xu (2005). Regularized spectral learning, in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) 2005*. Cited on page 88.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems (NIPS) 2013*. Cited on pages 47 and 50.
- L. Mukherjee, V. Singh, and J. Peng (2011). Scale invariant cosegmentation for image groups, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 8.
- R. Neal (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, pp. 249–265. Cited on page 16.
- A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.* (2002). On spectral clustering: Analysis and an algorithm, in *Advances in Neural Information Processing Systems (NIPS) 2002*. Cited on page 86.
- T. Nguyen (2014). Bayesian nonparametric extraction of hidden contexts from pervasive honest signals, in *IEEE International Conference on Pervasive Computing and Communications (PerCom) Workshops 2014*. Cited on pages 45 and 47.
- P. Ochs and T. Brox (2011). Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 8, 13, 34, 38, 41, 42, 79, 86, 91, 92, and 93.
- P. Ochs, J. Malik, and T. Brox (2014). Segmentation of moving objects by long term video analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 90.
- D. Oneata, J. Revaud, J. Verbeek, and C. Schmid (2014). Spatio-temporal object detection proposals, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on page 78.
- X. Peng, B. Sun, K. Ali, and K. Saenko (2015). Learning deep object detectors from 3D models, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on page 62.

- B. Pepik, M. Stark, P. Gehler, T. Ritschel, and B. Schiele (2015). 3D object class detection in the wild, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on 3D from a Single Image (3DSI) 2015*. Cited on pages 63 and 64.
- J. Pitman (2006). *Combinatorial stochastic processes*, vol. 1875, Springer-Verlag. Cited on page 15.
- M. Planck (1901). On the law of distribution of energy in the normal spectrum, *Annalen der Physik*, vol. 4(553), p. 1. Cited on page 107.
- S. H. Raza, M. Grundmann, and I. Essa (2013). Geometric context from video, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 78.
- D. Roggen, A. Calatroni, M. Rossi, T. Holleczeck, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, *et al.* (2010). Collecting complex activity datasets in highly rich networked sensor environments, in *IEEE International Conference on Networked Sensing Systems (INSS) 2010*. Cited on pages 46 and 53.
- C. Rother, T. Minka, A. Blake, and V. Kolmogorov (2006). Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. Cited on page 8.
- J. C. Rubio, J. Serrat, and A. M. López (2012). Video co-segmentation, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2012*. Cited on pages 8, 9, and 31.
- K. Saenko, B. Kulis, M. Fritz, and T. Darrell (2010). Adapting visual category models to new domains, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on page 98.
- D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nescic, X. Wang, and P. Westling (2014). High-resolution stereo datasets with subpixel-accurate ground truth, in *German Conference on Pattern Recognition 2014*. Cited on page 79.
- D. Scharstein and R. Szeliski (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *International Journal of Computer Vision (IJCV)*, vol. 47(1-3), pp. 7–42. Cited on page 10.
- J. Seiter, O. Amft, M. Rossi, and G. Tröster (2014a). Discovery of activity composites using topic models: An analysis of unsupervised methods, *Pervasive and Mobile Computing (PMC)*, vol. 15(0), pp. 215 – 227. Cited on pages 45, 53, and 59.
- J. Seiter, W.-C. Chiu, M. Fritz, O. Amft, and G. Tröster (2015). Joint segmentation and activity discovery using semantic and temporal priors, in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom) 2015*. Cited on pages 2 and 5.

- J. Seiter, A. Derungs, C. Schuster-Amft, O. Amft, and G. Tröster (2014b). Activity routine discovery in stroke rehabilitation patients without data annotation, in *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) 2014*. Cited on page 45.
- A. Shah and Z. Ghahramani (2013). Determinantal Clustering Process-A Nonparametric Bayesian Approach to Kernel Based Semi-Supervised Clustering, in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI) 2013*. Cited on page 131.
- F. Shahbaz Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez (2012). Color attributes for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 118 and 125.
- J. Shi and J. Malik (2000). Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 7 and 86.
- J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman (2005). Discovering objects and their location in images, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. Cited on page 9.
- R. Socher, A. Maas, and C. D. Manning (2011). Spectral Chinese Restaurant Processes: Nonparametric clustering based on similarities, in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*. Cited on page 131.
- H. O. Song, M. Fritz, C. Gu, and T. Darrell (2011). Visual grasp affordances from appearance-based cues, in *IEEE International Conference on Computer Vision (ICCV) Workshop on Challenges and Opportunities in Robot Perception (CORP) 2011*. Cited on page 64.
- H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese (2016). Deep metric learning via lifted structured feature embedding, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 131.
- H. O. Song, S. Zickler, T. Althoff, R. Girshick, M. Fritz, C. Geyer, P. Felzenszwalb, and T. Darrell (2012). Sparselet models for efficient multiclass object detection, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 118 and 119.
- M. Stark, M. Goesele, and B. Schiele (2010). Back to the future: Learning shape models from 3D CAD data, in *Proceedings of the British Machine Vision Conference (BMVC) 2010*. Cited on page 63.
- E. Sudderth, A. Torralba, W. Freeman, and A. Willsky (2008). Describing visual scenes using transformed objects and parts, *International Journal of Computer Vision (IJCV)*. Cited on page 9.

- D. Sun, S. Roth, J. Lewis, and M. Black (2008). Learning optical flow, *Proceedings of the European Conference on Computer Vision (ECCV)*. Cited on page 21.
- F.-T. Sun, Y.-T. Yeh, H.-T. Cheng, C. Kuo, and M. Griss (2014). Nonparametric discovery of human routines from sensor data, in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom) 2014*. Cited on pages 45 and 47.
- E. Taralova, F. D. la Torre, and M. Hebert (2014). Motion words for videos, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on page 78.
- Y. W. Teh (2004). *Nonparametric Bayesian Mixture Models - release 2.1.*, <http://www.gatsby.ucl.ac.uk/~ywtteh/research/software.html>. Cited on page 20.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes, *Journal of the American Statistical Association*. Cited on pages 16, 17, 28, 29, 39, and 52.
- Y. Tsin, S. B. Kang, and R. Szeliski (2006). Stereo matching with linear superposition of layers, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 28, pp. 290–301. Cited on page 98.
- H. Uchiyama and E. Marchand (2012). Object detection and pose tracking for augmented reality: Recent approaches, in *Korea-Japan Joint Workshop on Frontiers of Computer Vision 2012*. Cited on page 118.
- A. Vedaldi and B. Fulkerson (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*, <http://www.vlfeat.org/>. Cited on pages 21 and 42.
- A. Vedaldi and S. Soatto (2008). Quick shift and kernel methods for mode seeking, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. Cited on page 42.
- S. Vicente, V. Kolmogorov, and C. Rother (2010). Cosegmentation revisited: Models and optimization, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*, pp. 465–479, Springer. Cited on pages 8 and 15.
- P. Viola and M. Jones (2001). Rapid object detection using a boosted cascade of simple features, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2001*. Cited on pages 118 and 125.
- C. Vogel, K. Schindler, and S. Roth (2011). 3D scene flow estimation with a rigid motion prior, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 10.
- U. von Luxburg (2007). A tutorial on spectral clustering, *Statistics and Computing*, vol. 17(4), pp. 395–416. Cited on page 86.

- C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba (2013). HOGgles: Visualizing object detection features, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 62, 63, 67, 68, 69, 70, and 72.
- C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang (2014a). Video Object Co-Segmentation via Subspace Clustering and Quadratic Pseudo-Boolean Optimization in an MRF Framework, *IEEE Transactions on Multimedia*, vol. 16(4), pp. 903–916. Cited on page 9.
- J. Wang and E. H. Adelson (1994). Spatio-temporal segmentation of video data, in *IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology 1994*. Cited on page 13.
- J. Y. A. Wang and E. H. Adelson (1993). Layered representation for motion analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1993*. Cited on page 13.
- L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng (2014b). Video object discovery and co-segmentation with extremely weak supervision, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on pages 9 and 33.
- X. Wang and E. Grimson (2007). Spatial latent dirichlet allocation, *Advances in Neural Information Processing Systems (NIPS)*. Cited on page 9.
- Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli (2004). Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing (TIP)*. Cited on pages 70 and 72.
- D. Weikersdorfer, A. Schick, and D. Cremers (2013). Depth-adaptive superpixels for RGB-D video segmentation, in *Proceedings of the IEEE International Conference on Image Processing (ICIP) 2013*. Cited on page 11.
- P. Weinzaepfel, H. Jégou, and P. Pérez (2011). Reconstructing an image from its local descriptors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 63.
- Y. Xiang, R. Mottaghi, and S. Savarese (2014). Beyond PASCAL: A benchmark for 3D object detection in the wild, in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) 2014*. Cited on pages 63, 73, and 74.
- X. Xiong and F. De la Torre (2013). Supervised descent method and its applications to face alignment, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 63.
- C. Xu and J. J. Corso (2012). Evaluation of super-voxel methods for early video processing, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 8.

- R. Zabih and J. Woodfill (1994). Non-parametric local transforms for computing visual correspondence, in *Proceedings of the European Conference on Computer Vision (ECCV) 1994*. Cited on page 99.
- C. Zach, T. Pock, and H. Bischof (2007). A duality based approach for realtime TV-L 1 optical flow, in *Pattern Recognition 2007*, pp. 214–223, Springer. Cited on pages 80, 81, and 86.
- D. Zhang, O. Javed, and M. Shah (2014a). Video object co-segmentation by regulated maximum weight cliques, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on pages 9, 33, and 39.
- Q. Zhang, X. Shen, L. Xu, and J. Jia (2014b). Rolling guidance filter, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on page 86.
- J. Zheng and L. M. Ni (2012). An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data, in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2012*. Cited on page 46.
- Y. Zhu, Y. Arase, X. Xie, and Q. Yang (2011). Bayesian nonparametric modeling of user activities, in *International Workshop on Trajectory Data Mining and Analysis (TDMA) 2011*. Cited on page 46.

CURRICULUM VITAE

Wei-Chen Chiu

Date of birth:	September 2nd 1986, in Tainan, Taiwan	
Citizenship:	Taiwan	
Education:	08/2011 - 09/2016	PhD student at Max Planck Institute for Informatics, Germany; supervised by Dr. Mario Fritz.
	09/2008 - 07/2009	Master in Computer Science, National Chiao Tung University, Taiwan; supervised by Prof. Jen-Hui Chuang and Prof. Sheng-Jyh Wang.
	07/2006 - 12/2006	Exchange student at Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; supervised by Prof. Tsuhan Chen.
	09/2004 - 07/2008	Bachelor of Science in Electrical Engineering and Computer Science Undergraduate Honors Program (EECSHP), National Chiao Tung University, Taiwan.
	06/2004	High school graduation, Tainan, Taiwan.
Experience:	03/2011 - 06/2011	Internship with Dr. Mario Fritz and Dr. Ulf Blanke, Max Planck Institute for Informatics, Saarbrücken, Germany.
	01/2011 - 03/2011	Research assistant with Prof. Sheng-Jyh Wang, National Chiao Tung University, Taiwan.
	10/2010 - 12/2010	Engineer at Largan Precision Co. Ltd., Taichung, Taiwan.
	07/2007 - 12/2007	Internship at Industrial Technology Research Institute, Hsinchu, Taiwan.
Invited Talk:	08/2013	Multi-Class Video Co-Segmentation with a Generative Multi-Video Model at Research Center for Information Technology Innovation, Academia Sinica, Taiwan.

Academic activities: Reviewer

IEEE Transactions on Circuits and Systems for
Video Technology (2013)

International Conference on Image Processing
(2013)

International Symposium on Wearable Comput-
ers (2014)

European Conference on Computer Vision
(2014)

International Journal of Computer Vision (2015)

IEEE Signal Processing Letters (2016)

International Conference on Pattern Recognition
(2016)

Asian Conference on Computer Vision (2016)

PUBLICATIONS

[6] *Towards Segmenting Consumer Stereo Videos: Benchmark, Baselines and Ensembles.*
Wei-Chen Chiu, Fabio Galasso, and Mario Fritz.
In Asian Conference on Computer Vision (**ACCV**), 2016.

[6] *See the Difference: Direct Pre-Image Reconstruction and Pose Estimation by Differentiating HOG.*
Wei-Chen Chiu and Mario Fritz.
In International Conference on Computer Vision (**ICCV**), 2015.

[5] *Joint Segmentation and Activity Discovery using Semantic and Temporal Priors.*
Julia Seiter, Wei-Chen Chiu, Mario Fritz, Oliver Amft, and Gerhard Tröster.
In IEEE International Conference on Pervasive Computing and Communications (**PerCom**), 2015.

[4] *Object Disambiguation for Augmented Reality Applications.*
Wei-Chen Chiu, Gregory S. Johnson, Daniel Mcculley, Oliver Grau, and Mario Fritz.
In British Machine Vision Conference (**BMVC**) 2014.

[3] *Multi-Class Video Co-Segmentation with a Generative Multi-Video Model.*
Wei-Chen Chiu and Mario Fritz.
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2013.

[2] *I Spy with My Little Eye: Learning Optimal Filters for Cross-Modal Stereo under Projected Patterns.*
Wei-Chen Chiu, Ulf Blanke, and Mario Fritz.
In IEEE Workshop on Consumer Depth Cameras for Computer Vision in conjunction with (**ICCV**), 2011

[1] *Improving the Kinect by Cross-Modal Stereo.*
Wei-Chen Chiu, Ulf Blanke, and Mario Fritz.
In British Machine Vision Conference (**BMVC**) 2011.

