People Detection and Tracking in Crowded Scenes

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by
Siyu Tang, M.Sc.

Saarbrücken
July 2017

Day of Colloquium	29^{th} of September, 2017
Dean of the Faculty	Univ.-Prof. Dr. Frank-Olaf Schreyer

Examination Committee

Chair	Prof. Dr. Holger Hermanns
Reviewer, Advisor	Prof. Dr. Bernt Schiele
Reviewer	Prof. Dr. Michael Black
Reviewer	Prof. Dr. Luc Van Gool
Academic Assistant	Dr. Gerard Pons-Moll

ABSTRACT

People are often a central element of visual scenes, particularly in real-world street scenes. Thus it has been a long-standing goal in Computer Vision to develop methods aiming at analyzing humans in visual data. Due to the complexity of real-world scenes, visual understanding of people remains challenging for machine perception. In this thesis we focus on advancing the techniques for people detection and tracking in crowded street scenes. We also propose new models for human pose estimation and motion segmentation in realistic images and videos.

First, we propose detection models that are jointly trained to detect single person as well as pairs of people under varying degrees of occlusion. The learning algorithm of our joint detector facilitates a tight integration of tracking and detection, because it is designed to address common failure cases during tracking due to long-term inter-object occlusions.

Second, we propose novel multi person tracking models that formulate tracking as a graph partitioning problem. Our models jointly cluster detection hypotheses in space and time, eliminating the need for a heuristic non-maximum suppression. Furthermore, for crowded scenes, our tracking model encodes long-range person re-identification information into the detection clustering process in a unified and rigorous manner.

Third, we explore the visual tracking task in different granularity. We present a tracking model that simultaneously clusters object bounding boxes and pixel level trajectories over time. This approach provides a rich understanding of the motion of objects in the scene.

Last, we extend our tracking model for the multi person pose estimation task. We introduce a joint subset partitioning and labelling model where we simultaneously estimate the poses of all the people in the scene.

In summary, this thesis addresses a number of diverse tasks that aim to enable vision systems to analyze people in realistic images and videos. In particular, the thesis proposes several novel ideas and rigorous mathematical formulations, pushes the boundary of state-of-the-arts and results in superior performance.

ZUSAMMENFASSUNG

Personen sind oft ein zentraler Bestandteil visueller Szenen, besonders in natürlichen Straßenszenen. Daher ist es seit langem ein Ziel der Computer Vision, Methoden zu entwickeln, um Personen in einer Szene zu analysieren. Aufgrund der Komplexität natürlicher Szenen bleibt das visuelle Verständnis von Personen eine Herausforderung für die maschinelle Wahrnehmung. Im Zentrum dieser Arbeit steht die Weiterentwicklung von Verfahren zur Detektion und zum Tracking von Personen in Straßenszenen mit Menschenmengen. Wir erforschen darüber hinaus neue Methoden zur menschlichen Posenschätzung und Bewegungssegmentierung in realistischen Bildern und Videos.

Zunächst schlagen wir Detektionsmodelle vor, die gemeinsam trainiert werden, um sowohl einzelne Personen als auch Personenpaare bei verschiedener Verdeckung zu detektieren. Der Lernalgorithmus unseres gemeinsamen Detektors erleichtert eine enge Integration von Tracking und Detektion, da er darauf konzipiert ist, häufige Fehlerfälle aufgrund langfristiger Verdeckungen zwischen Objekten während des Tracking anzugehen.

Zweitens schlagen wir neue Modelle für das Tracking mehrerer Personen vor, die das Tracking als Problem der Graphenpartitionierung formulieren. Unsere Modelle clustern Detektionshypothesen gemeinsam in Raum und Zeit und eliminieren dadurch die Notwendigkeit einer heuristischen Unterdrückung nicht maximaler Detektionen. Bei Szenen mit Menschenmengen kodiert unser Trackingmodell darüber hinaus einheitlich und genau Informationen zur langfristigen Re-Identifizierung in den Clusteringprozess der Detektionen.

Drittens untersuchen wir die visuelle Trackingaufgabe bei verschiedener Granularität. Wir stellen ein Trackingmodell vor, das im Zeitablauf gleichzeitig Begrenzungsrahmen von Objekten und Trajektorien auf Pixelebene clustert. Diese Herangehensweise ermöglicht ein umfassendes Verständnis der Bewegung der Objekte in der Szene.

Schließlich erweitern wir unser Trackingmodell für die Posenschätzung mehrerer Personen. Wir führen ein Modell zur gemeinsamen Graphzerlegung und Knotenklassifikation ein, mit dem wir gleichzeitig die Posen aller Personen in der Szene schätzen.

Zusammengefasst widmet sich diese Arbeit einer Reihe verschiedener Aufgaben mit dem gemeinsamen Ziel, Bildverarbeitungssystemen die Analyse von Personen in realistischen Bildern und Videos zu ermöglichen. Insbesondere schlägt die Arbeit mehrere neue Ansätze und genaue mathematische Formulierungen vor, und sie zeigt Methoden, welche die Grenze des neuesten Stands der Technik überschreiten und eine höhere Leistung von Bildverarbeitungssystemen ermöglichen.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my supervisor Prof. Bernt Schiele for giving me the opportunity to purse my PhD under your supervision. I feel extremely lucky having the opportunity to work with you. You are always willing to lend an ear and provide me with invaluable guidance, high quality feedback and encouragement on my work. I also truly and deeply appreciate your suggestions and guidance on my career.

I would like to express my gratitude to my thesis reviewers Prof. Michael Black and Prof. Luc Van Gool. Thank you for your effort of reviewing my work. Thank you for your interests and your feedbacks.

I am very thankful to all my collaborators, especially Dr. Bjoern Andres and Dr. Mykhaylo Andriluka. The fruitful discussions and constant supports from you are very valuable for my works.

Special thanks to my colleagues at the Max Planck Institute for Informatics for creating such friendly and excellent working environment. I deeply appreciate and greatly benefit from the time we spent together.

I am grateful to my colleagues at the Max Planck Institute for Intelligent Systems for proof-reading of my thesis and your valuable feedback on my presentation.

Last but not least, I would like to thank my family and my friends for your constant love and support. Especially, I want to thank my husband Di for all the encouragement, unconditional support and love.

CONTENTS

INTRODUCTION 1

Contents

Vision is arguably the most important human sense. It masters immense visual information and allows people to recognize, organize, and interact with their egocentric world. No other human sense is comparable to vision in terms of its versatility and richness. For decades, it has been a fascinating research topic for scientists to understand the human ability to interpret visual scenes and utilize the information to interact with the external world. Computer vision, as the counterpart of human vision, has made significant progress in the past. Together with ubiquitous cameras in our everyday life, image and video understandings become essential for building autonomous and intelligent computer systems.

Computer systems require different levels of visual understanding when performing different tasks. These range from low-level tasks such as super resolution and image de-blurring to high-level tasks such as image classification and face detection. Immense progress has been made in many areas, and in some computers even surpass the quality of human perception, e.g. image classification (He *et al.*, 2016), and are successfully employed in consumer electronics. However, visual understanding of crowded real-world street scenes still remains challenging. People are often a central element of such scenes, thus it is particularly important to develop methods aiming at analysing human movements. Despite the tremendous capability that computer vision systems have today, machine perception is still far from competing with human perception for interpreting people in crowded street scenes. One example could be autonomous driving/road safety, where humans effortlessly identify and localize the positions of pedestrians and predict their motion and intention, while the performance of vision systems is far from satisfactory. The variety and complexity of motions, cluttered backgrounds, or partial occlusions make it difficult for machines to interpret the visual information of crowded real-world street scenes.

The main focus of this thesis is to develop algorithms for tracking multiple people in street scenes, a task that is often referred as **multi person tracking** and sometimes as **multi target tracking**. In general, the multi person tracking task is to identify the location of each person at every time step and to reconstruct the trajectories of individuals in a dynamic scene without manual initialization. Given that pedestrian detectors are rather robust nowadays and produce good detections on non-occluded and reasonably sized pedestrians, a commonly explored strategy for the multi person tracking task is tracking-by-detection, where the tracking task is split into two steps:

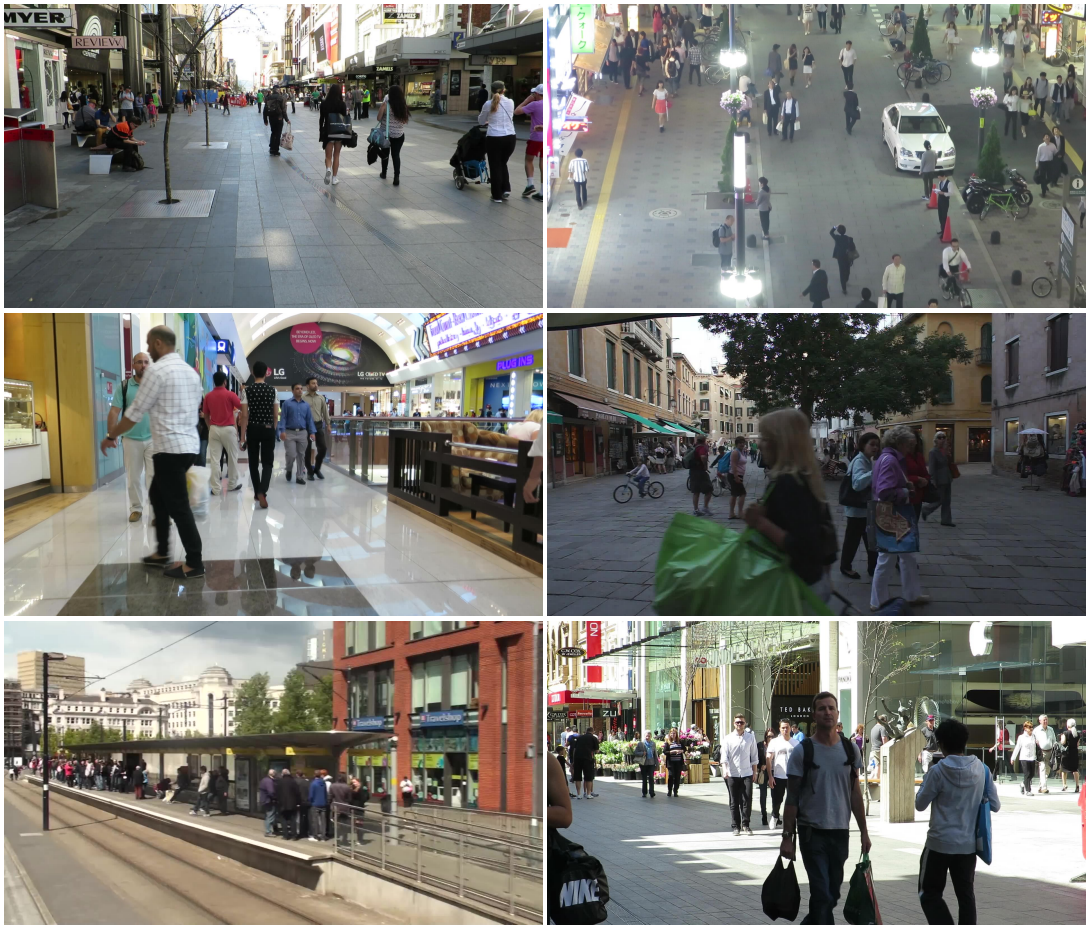


Figure 1.1: Several examples of images of street scenes. Notice that the street scene videos have a large varieties of imaging conditions and camera angles.

generate people hypotheses at every time step (people detection) and associate the hypotheses that describe the same person (people tracking) over time. In this thesis, we explore new techniques and propose new algorithms for both people detection and people tracking, with the focus on (semi-) crowded real-world street scenes.

People detection is an essential component in any tracking-by-detection based method. People detection in street scenes, which is often referred to as pedestrian detection, has been studied intensively. While continuous progress has been made, heavy occlusion remains challenging. The performance of state-of-the-art pedestrian detectors decreases dramatically in the presence of significant partial occlusion, as shown in Tang *et al.*, 2014 and Tang *et al.*, 2012. The challenges of handling heavy occlusion are related to several factors: little image support, under-representation in public pedestrian detection benchmarks, and high diversity of occlusion scenarios. In this thesis, we advance people detection techniques by explicitly addressing these factors. First, we explore person person occlusion patterns. One intuitive way to handle partial occlusion is to first detect the occluding person, remove the image evidence of the detected person and then detect the potentially occluded person

using the remaining image evidence. This type of method treats occlusions as a nuisance and performs inference based on the image information that corresponds to the visible part of the occluded person. We propose an alternative strategy: instead of treating occlusion as a distraction, we explore the person/person occlusion patterns that exhibit regularities which can be used to detect the presence of partial occluded people. Second, we propose a person/person occlusion dataset (MPII-2person dataset) where the occlusion levels are carefully annotated. In popular pedestrian detection benchmarks such as Caltech (Dollár *et al.*, 2012), significant partial occlusion is often under-represented. For evaluation, many works even exclude such cases by mainly reporting the number on a subset like the so called reasonable set of Caltech (Dollár *et al.*, 2012). Our MPII-2Person dataset allows analysing the performance on partially occluded people in detail. Last, to model the diversity of occlusion scenarios, we propose a detector training algorithm which automatically discovers frequent and discriminative occlusion patterns that exhibit variations in several factors, such as people's body articulation, and their position and orientation relative the camera. Furthermore, we combine our occlusion pattern mining algorithm with a multi-person tracker in the loop. To this end, we are able to learn a pedestrian detector that is explicitly optimized for the task of tracking multiple people.

People tracking is another problem intensively studied in computer vision. However the robustness of tracking algorithms is far from satisfactory, especially for crowded street scenes. The challenges come from several sources. First, as discussed in the previous paragraph, the performance of state-of-the-art person detectors decreases significantly in crowded scenes, suffering both from false positives and missing detections. Distinguishing the detections on pedestrians and the detections on background is a hard problem inherited from the detection step. In addition, the bounding box localization becomes less accurate due to occlusion. Handling such noisy detection input is a well-known challenge for data association methods. Second, in crowded street scene videos, it is not only difficult to estimate the exact number of people present but also a large number of people are occluded for more than 50% of the entire video. In particular, for these cases it is difficult to determine the starting and ending moment of the tracks. Last, affinity measures that are robust to camera motion, occlusion, and illumination are essential for improving tracking performance, as tracking through partial occlusion and re-identifying after full occlusion are highly dependent on good similarity measures. We explicitly address all these challenges and advance tracking techniques in this thesis. We propose novel tracking formulations that are rigorous and robust to handle the unknown number of people as well as detection noise. Our tracking formulation is defined as an optimization problem with respect to a graph. The nodes of the graph correspond to people hypotheses or detections. Edges are introduced to connect detections that hypothetically represent the same person. One common approach is to model tracking as finding disjoint paths in the graph, where the paths do not branch or merge, just as one person can not split into two persons. While being intuitive, such models ignore the fact that typical people detectors produce many similar hypotheses for the same person. For the disjoint paths based method, it becomes hard to choose

one best path among many plausible good ones. In this thesis, we propose a novel tracking formulation, the **Minimum Cost Multicut Problem**, that models tracking as a graph decomposition problem. The advantages of this mathematical abstraction are that the number of persons in the video is not fixed or biased, but optimized and determined by the solution of the problem. Meanwhile, it clusters multiple detections of the same targets jointly over space and time, producing a more robust association. However, a rigorous tracking formulation is only one part of the solution. As shown in Fig 1.1, the street scene videos have a large variety of imaging conditions and camera angles. E.g. The camera could be mounted on a moving bus or the video could be captured in standard surveillance setting. The appearance of individual pedestrians also changes significantly depending on where and when the videos are captured. Finally, the size of a pedestrian varies largely within and between videos. All these factors present an exponentially increased state space of appearance patterns which is difficult to compare and associate. In such videos, the commonly explored and simple linear motion model does not work well. To address these problems, in the thesis, we propose robust affinity measures by combining local image patch matching and person re-identification information. The intuition is that the local image patch matching provides reliable measures between detections that are temporally close. The re-identification information provides reliable information for detections that are distant in time. In crowded street scenes, our combined pairwise features are able to produce accurate affinity measures even for irregular camera motions and long-term occlusions.

Incorporating long-range re-identification into the tracking model is not trivial even when the corresponding affinity measures are quite accurate One reason is that similar-looking people may not be identical. Thus, the long-range information should be integrated in a way that it is supported by the majority of the local information. Based on this intuition, we propose another tracking formulation, the **Minimum Cost Lifted Multicut Problem**, which is in particular designed to incorporate the long-term person re-identification information into the tracking model. For the Minimum Cost Lifted Multicut Problem, we introduce two type of edges, regular and lifted edges, in to the graph. The regular edges define the feasible solution of the problem and the lifted edges introduce the long-range person re-identification information into the objective. The model is able to express the fact that similarly looking persons are identical only if there is a valid path formed by the regular edges. By combining the robust affinity measures and the lifted multicut tracking model, we achieve state-of-art tracking performance on the challenging MOT16 benchmark (Milan *et al.*, 2016).

One closely related task to multi person tracking is **motion segmentation** (moving object segmentation), where the goal is to segment point trajectories on each moving object. These two tasks are interrelated problems because their goal is to determine the regions that belong to the same object in an image sequence. The point trajectories carry local and low-level image information which can be robust to partial occlusion. In contrast, the high-level detection bounding box contains semantic information which is robust to articulated motion and coherent motion between

objects. To leverage the cues from both levels, we propose a unified graphical model in which the multi target tracking and motion segmentation tasks are simultaneously solved.

Another closely related problem we explore in this thesis is **multi person pose estimation**. Single-person articulated pose estimation has been studied intensively, but the problem of multi person pose estimation has been largely neglected, particularly because of the inherent challenges related to the fact that it is hard to estimate the number of partially visible people and their poses. The multicut formulation we propose for tracking is able to automatically identify the number of people and robustly associate the detections that belong to the same identity. This property is also desirable for multi person pose estimation. We start with the body part detectors and propose a bottom-up approach where we jointly estimate the relation between body part detections and the part label of individuals. We extend the Minimum Cost Multicut Problem to a joint **Graph Partition and Node Labeling** problem. By solving this optimization problem, we are able to obtain the number of people, their positions, their poses and their partial visibilities simultaneously.

1.1 CONTRIBUTIONS OF THE THESIS

This thesis makes contributions to the tasks of people detection and multi person tracking in street scenes. The thesis also makes contributions to motion segmentation and multi person pose estimation.

The contributions to people detection are as follows:

- We propose a novel approach to simultaneously detect two people that overlap in image. We introduce a joint person detector that is trained to detect the presence of a single person as well as two people that are in close proximity.
- We create a new person/person occlusion dataset (MPII-2Person dataset) where the occlusion level of each image is carefully annotated. The MPII-2Person dataset focuses on partially occluded people which are under-represented in popular pedestrian benchmarks.
- We propose a novel algorithm that optimizes detector performance for tracking. The algorithm automatically mines dominant people/people occlusion patterns and exploit tracking performance on partially occluded people.

The contributions to multi person tracking are as follows:

- We propose a novel multicut tracking model where the number of persons is not fixed or biased by the definition of the problem but is estimated in an unbiased fashion from the video sequence and is determined by the solution of the problem. Multiple detections of the same person in the same frame are effectively clustered, which eliminates the need for heuristic non-maxima suppression.

- We propose various features to measure the similarity between detections. Particularly, for long-range person re-identification, we design a novel deep neural network which fuses the human pose information with deep features. This provides us a mechanism to re-identify people that are distant in time.
- We further propose to model tracking as a minimum cost lifted multicut problem, where two types of edges (regular and lifted edges) are introduced. Our formulation encodes long-range information by the lifted edges which do not define possibilities of directly joining nodes. In order to assign two detections that are far apart in time and similar in appearance to the same cluster (person), there must exist a path (track) along the regular edges, that certifies this decision. The lifted multicut model obtains new state-of-the-art tracking performance on the MOT16 benchmark.

The contributions to joint tracking and motion segmentation are as follows:

- We extend our tracking works by proposing a unified graphical model where multi person tracking and motion segmentation are jointly cast as one graph partitioning problem. The unified model produces consistent identity labels at the bounding box tracks level as well as at pixel-level segmentations.
- We demonstrate experimental results on both tracking and motion segmentation benchmarks, achieving (on MOT16 benchmark) or surpassing (on FBMS benchmark) the state-of-the-art.

The contribution to multi person articulated pose estimation are as follows:

- We extend the multicut model by introducing node labels to the objective function. The novel graph partition and node labeling formulation is able to estimate the number of people, their location, and their poses in a unified manner.

1.2 OUTLINE OF THE THESIS

The thesis is structured as follows:

Chapter 2: Related work. This chapter gives an overview of the related work in people detection and tracking in crowded scenes. In particular, we discuss how these works differ from the approaches presented in this thesis. In addition, we survey the literature related to pose estimation and motion segmentation with a focus on methods that target realistic and challenging scenarios.

Chapter 3: Detection and Tracking of Occluded People. In this chapter we present a novel approach to people detection in realistic street conditions. Given that in street scenes the dominant occlusion cases are person/person occlusions, we leverage these characteristic occlusion patterns in the image domain and propose a joint-person detector to detect single persons as well as pairs of

persons. We integrate the joint person detector into a tracking approach and demonstrate its potential for people detection and tracking of occluded people in challenging benchmarks.

The content of this chapter was first presented at BMVC12 with the title *Detection and Tracking of Occluded People* (Tang *et al.*, 2012). It has been further extended to a journal article that appeared in IJCV14 (Tang *et al.*, 2014).

Chapter 4: Learning People Detectors for Tracking. In this chapter, we improve the joint-person detector presented in chapter 3 by proposing a novel structured loss formulation which combines the VOC loss and the detection type loss. We further propose to train a people detector by optimizing tracking performance.

The content of this chapter was presented at ICCV13 with the title *Learning People Detectors for Tracking in Crowded Scenes* (Tang *et al.*, 2013).

Chapter 5: Subgraph Decomposition for Multi-Target Tracking. In this chapter, we introduce a novel minimum cost subgraph multicut formulation for the multi-target tracking task. Our formulation jointly selects and clusters detections over space and time. In order to conduct a direct comparison with conventional network flow based methods, we also propose a minimum cost disjoint path formulation for the tracking task. We compare both formulations from a theoretical perspective as well as experimentally. We conclude that the multi-cut formulation produces more robust connections and thus is more suitable for the tracking task.

The content of this chapter was presented at CVPR15 with the title *Subgraph Decomposition for Multi-Target Tracking* (Tang *et al.*, 2015).

Chapter 6: Multi-Person Tracking by Multicut and Deep Matching. In this chapter, we present our extensions of the work Tang *et al.*, 2015. We propose a novel local pairwise feature based on local appearance matching that is robust to partial occlusion and camera motion. We also perform extensive experiments to compare different pairwise potentials and to analyze the robustness of the tracking formulation. We demonstrate the effectiveness of our method on the MOT16 benchmark. The work is the winner of the Multiple Object Tracking Challenge at ECCV16.

The content of this chapter was presented at ECCV16 multi target tracking workshop with the title *Multi-Person Tracking by Multicuts and Deep Matching*.

Chapter 7: Lifted Multicuts and Deep Re-identification. In this chapter, we present an advanced clustering based tracking formulation which is called the minimum cost lifted multicut formulation. The novel formulation can be seen as a constrained clustering model, where the detection pairs that are temporally distant from each other are treated differently from the detection pairs that are in neighboring frames. Detection pairs in neighboring frames can be directly clustered into one cluster, but detection pairs that are far away

can only be clustered if there exists a path in the frames between them. This formulation prevents false joints between the distant detections and significantly improves tracking accuracy. Besides the novel tracking formulation, we also propose various deep person re-identification networks for modeling the appearance relations between detections. We demonstrate the effectiveness of our formulation outperforming the state-of-the-art for the MOT16 benchmark.

The content of this chapter is accepted to CVPR2017 with the title *Multi-Person Tracking by Lifted Multicuts and Deep Re-Identification*.

Chapter 8: Joint Segmentation and Tracking of Multiple Objects. In this chapter, we propose to cluster high-level information (detection bounding boxes) and low-level information (motion trajectories) jointly in videos. We utilize the minimum cost multicut formulation, where the nodes in the corresponding graphical model are composed of detections and pixel-level trajectories. By decomposing the graphical model into an optimal number of connected components, the detections and pixel trajectories belong to the same objects in the video are linked and clustered into one component over space and time. The results on motion segmentation benchmarks and multi target tracking benchmarks indicate the effectiveness of the proposed method.

The content of this chapter was published in arXiv with the title *A Multi-cut Formulation for Joint Segmentation and Tracking of Multiple Objects*. Siyu Tang proposed the idea of jointly modeling tracking and motion segmentation with the multicut formulation and contributed the multi person tracking side of the work.

Chapter 9: Multi Person Pose Estimation by DeepCut. In this chapter, we present a novel formulation for human pose estimation problem by jointly clustering and labeling body part detections. We propose two CNN variants to generate body part candidates. Then the pose estimation task is cast as an integer linear program. By optimizing the objective function, we obtain the number of people, their spatial configuration, their articulated poses and occlusion information. The proposed model surpasses state-of-the-art results on four benchmarks.

The content of this chapter was presented at CVPR16 with the title *DeepCut:Joint Subset Partition and labeling for Multi Person Pose Estimation* (Pishchulin *et al.*, 2016)). Siyu Tang proposed the idea of modeling the multi person pose estimation task with the multicut formulation, and contributed and implemented the modeling side of the work.

Chapter 10: Conclusions and future perspectives. In this chapter, we conclude the thesis by summarizing the contributions, results and also the limitations of the proposed methods. Furthermore, we present an outlook on future research directions towards rich video understanding.

RELATED WORK 2

Contents

In this chapter, we mainly discuss recent developments and seminal works in people detection and multi person tracking with a focus on crowded scenes. We will also include related work on motion segmentation and human pose estimation for the complex scene scenarios in section 2.3 and section 2.4.

2.1 PEDESTRIAN DETECTION AND PEOPLE DETECTION

People detection in street scenes is often referred to as pedestrian detection, which has been significantly improved by diverse methods over the last years. In this section, we compare state-of-art pedestrian detectors and people detectors in terms of image representation and learning method, and we discuss their relations to our work.

2.1.1 Pedestrian Detection

Arguably the most popular image representation for pedestrian detection is the Histogram of Oriented Gradients (HOGs) that was introduced by Dalal and Triggs, 2005. By decomposing an image into overlapping cells and carefully aggregating

local gradients information, Dalal and Triggs, 2005 were able to significantly improve the performance of pedestrian detection. Many variants of HOGs have been proposed afterwards: Wang *et al.*, 2009b propose to combine the Local Binary Pattern (LBP) feature that uses the difference sign as signal with HOGs, which is more robust to partial occlusion; Lin and Davis, 2008 propose to learn a global descriptor derived from the original HOGs feature, that is shape invariant as well as articulation insensitive, and is capable of distinguishing human/non-human image patterns.

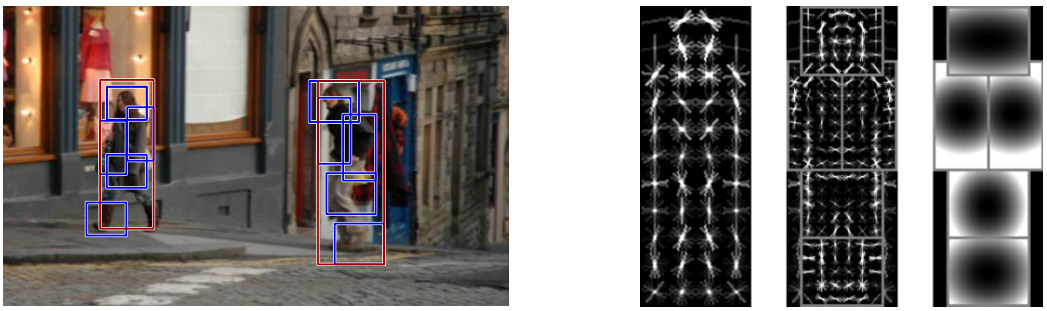
The Integral Channel Feature proposed by Dollár *et al.*, 2009 is another seminal work for pedestrian detection. Dollár *et al.*, 2009 and its many variants are able to reliably detect pedestrians under a variety of imaging conditions, people poses, and appearance. Benenson *et al.*, 2013 show that learning a set of irregular rectangle cells significantly improves over the HOGs style hand-designed patterns. Lim *et al.*, 2013 introduce Sketch token, which is a local contour-based image feature that is learned using middle level image information.

In order to improve detection quality, several works propose to increase and diversify the features computed on the input images. The intuition is that giving higher dimensional and diverse image representations, the classification task (person/non-person in the region of interest) becomes easier. Zhang *et al.*, 2014 employ different types of low-level Haar-like features from multiple rectangle regions to represent various human appearance in street scenes. Another simple yet effective approach for pedestrian detection proposed by Paisitkriangkrai *et al.*, 2014 extracts multiple low-level image features on multiple image regions based on spatial pooling. Daniel Costea and Nedevschi, 2016 propose to use multi-resolution filtered LUV and HOG channels for pedestrian detection as well as semantic segmentation. Furthermore, the resulting semantic segmentation information is employed as additional channels for pedestrian detection, resulting in a more powerful detector.

Recently, several works apply convolutional neural networks (CNNs) to learn image representation for pedestrian detection. Hosang *et al.*, 2015 use people detectors to generate region proposals and then classify the region proposals using the same pipeline as Girshick *et al.*, 2014a. Tian *et al.*, 2015 apply Dollar *et al.*, 2014 to obtain detection proposals and propose a task specific CNN to jointly optimize the detection task as well as semantic attribute recognition. In Zhang *et al.*, 2016, authors combine the Region Proposal Network (RPN) from Ren *et al.*, 2015 with a boosted forest to improve the detection performance on several public benchmarks.

2.1.2 People Detection

General people detection is a challenging task due to the unconstrained setting where people exhibit a large variety of articulated poses, clothing, occlusion, etc. Many approaches to people detection are able to reliably detect people under a variety of imaging conditions, people poses, and appearances. One of the seminal works, which is proposed by Felzenszwalb *et al.*, 2010, introduces the deformable part model (DPM). The DPM model represents people by a set of parts that are spatially deformable. Each part captures the characteristic local appearance of a person, and



(a) (b)

Figure 2.1: Images are from Felzenszwalb *et al.*, 2010. (a) Example image, the red rectangles indicate person detections, the blue rectangles indicate the deformable parts. (b) Visualization of the root filter, the part filters, the spatial configuration of the part filters of the DPM.

the relation between pairs of parts are modeled by spring-like connections. As there is no ground truth information for the part locations, the DPM model adapts a latent variable formulation of support vector machine (LSVM) to learn the model parameters as well as the latent values of the part configuration. Although the DPM model is effective when people are fully visible, its performance degrades when people become partially occluded. Various remedies have been proposed, including a combination of multiple detection components (Felzenszwalb *et al.*, 2010), using a large number of part detectors (Poselets) (Bourdev and Malik, 2009), detection of interactions between people and objects (Desai and Ramanan, 2012), and careful reasoning about association of image evidence to detection hypotheses (Leibe *et al.*, 2012, Barinova *et al.*, 2010, Wang *et al.*, 2009a). Leibe *et al.*, 2012 propose an approach that first aggregates the evidence from the local image features into a probabilistic figure-ground segmentation, and then relies on a Minimal Description Length (MDL) formulation to assign foreground regions to detection hypotheses. Barinova *et al.*, 2010 propose a probabilistic formulation of the generalized hough transform that prevents association of the same image evidence to multiple person hypotheses. These approaches treat partial occlusion as a nuisance and perform decisions based on the image evidence that corresponds to the visible part of the person. This makes them unreliable in cases of severe occlusions (i.e. more than 50% of the person is occluded). Several works have aimed at improving weak detections using information from additional sensing modalities (Enzweiler *et al.*, 2010) or by joint reasoning about people hypotheses and 3D scene layout (Wojek *et al.*, 2011). In Wojek *et al.*, 2011, a bank of partial people detectors is used to generate the initial proposals that are refined based on the 3D scene layout and temporal reasoning. However the detectors used in Wojek *et al.*, 2011 still assume that a significant portion of the person is visible. Besides reasoning about 3D scene layout requires camera calibration and additional assumptions on the scene geometry. Recently, Stewart *et al.*, 2016 propose a model to decode an image into a set of people head detections. They first use a CNN to encode image information, then a recurrent long short-term

memory (LSTM) layer to decode the image feature into a sequence of detections. The whole model is end-to-end trainable with a novel loss function that is defined on the sets of detections.

2.1.3 Synthetic Training Data for People Detection

Using synthetic training data to obtain a better object model in general and for human in particular has been studied in many literature. An early work (Grauman *et al.*, 2003) proposes to use a computer graphics model of articulated human bodies to render a multi-view synthetic images of human silhouettes. The rendered training images are augmented with 3D joint angle locations which enable the authors to learn a joint structure and shape model prior. Marin *et al.*, 2010b show that a HOG-based human detector can be effectively learned from the synthetic examples that are generated by a game engine. Pishchulin *et al.*, 2011 propose to utilize a morphable 3D body model to generate a large number of synthetic training images from a few recorded persons and views.

2.1.4 Joint People Detection and Tracking

As one of the objectives of this thesis is to detect and track people in crowded scenes. It is an interesting direction to explore how we could combine these two tasks in a unified framework. Addressing both detection and tracking as a joint problem has been considered in the literature. In Leibe *et al.*, 2007, the task is formulated as a quadratic Boolean program to combine trajectory estimation and detection. The objective function is optimized locally, by alternating between the two components. In contrast, Wu *et al.*, 2012 formulate a joint integer linear program and allow data association to influence the detector. However, their approach is based on background subtraction on a discretized grid. A slightly different strategy is followed by Yan *et al.*, 2012, where data association not only relies on detector responses, but also on a set of other trackers.

2.1.5 Relations to Our Works

The people detectors proposed in Chapter 3 and 4 build on the general people detector of Felzenszwalb *et al.*, 2010, which we extend in two ways. First, we propose a double-person detector that simultaneously detects two people occluding each other and second, we propose a joint detector that can detect both one person as well as two people due to joint training. Our strategy is possible because overlapping people result in characteristic appearance patterns that are otherwise uncommon. Our approach is related to the "visual phrases" approach which is proposed by Farhadi and Sadeghi, 2011. We train a joint detector for the combination of two object instances. A similar idea is proposed in Desai and Ramanan, 2012 as well that they train mixtures of detectors with some of the mixture components representing

appearance of typical occluders. To capture typical appearance patterns of people occluding each other, we automatically generate a dataset of training images with controlled and varying degrees of occlusion. In this respect our work is also related to recent work combining real and artificially generated images to train people detectors (Marin *et al.*, 2010a, Pishchulin *et al.*, 2011).

Furthermore, in Chapter 4 we propose a detector learning approach tailored to the requirements of people tracking, and in particular propose to train a people detector based on feedback from the tracker. Unlike previous work, in Chapter 4, we do not only consider detection and tracking jointly, but also explicitly adapt the detector to typical tracking failures.

2.2 MULTI PERSON TRACKING

Object tracking is one of the fundamental problems in Computer Vision, the body of related literature is enormous. They range from the early studies about tracking in human perception (Pylyshyn and Storm, 1988), filter-based tracking (E.g. Kalman filter Kalman, 1960) to single and multi object tracking. In recent years, object tracking is often categorized into two main classes. One class is called "model-free" tracking or single object tracking. This line of tracking works does not utilize the prior knowledge of the type of target object. Manual initialization is required to identify the region of interest in the scene. The second class of object tracking method is called multi object tracking, where the object category is known, but the number of objects in the video is unknown. A powerful object detector can be therefore applied to identify the location and size of the objects.

Also note that, multi person tracking is a specialization of multi object tracking. Most of the recent multi object tracking works focus on the people category. In this thesis, we are particularly interested in tracking people in image sequences as they are often the central character in street scenes.

Before we go to the literature review of object tracking, we will first discuss guided filter based tracking approaches, where tracking is modelled as a sequential state estimation from noisy observations.

2.2.1 Guided Filter Based Object Tracking

Kalman filter (Kalman, 1960) proposed in sixties is well known for estimating the state of tracking targets from a sequence of noise observation. The algorithm has recursive two steps. In the prediction step, it generates the estimated states of variables and the corresponding uncertainties. In the evaluation step, the estimated states are updated according to the newly arrived observations. The kalman filter works under certain assumptions, namely, the observation models and the state transition need to be linear functions. Several extensions have been proposed to relax such linearity constraint. Julier and Uhlmann, 1997 proposed extended Kalman filter, which doesn't require the linear transition function and observation model, it only

requires differentiable functions. Another well known state estimation technique is particle filters, which is proposed by Gordon *et al.*, 1993. It uses a set of particles to represent the state given some noisy observations. The particle filter can model non-linear state-space and the observation model can take any form.

To apply particle filter to the object tracking task, Isard and MacCormick, 2001 introduce a particle filter that is used to jointly estimate the number of objects and their state space. Khan *et al.*, 2005 also propose to use a particle filter to jointly model all the objects, additionally, a Markov Random Field (MRF) motion prior is introduced to estimate the object interactions. To efficiently sample for a larger number of objects, a Markov chain Monte Carlo (MCMC) sampling is used to replace the traditional importance sampling step in the particle filter. In a more recent work (Santhoshkumar *et al.*, 2013), a set of two particle filters is proposed for each object in the target video. The local particle filter estimates the local motion of the target object, and the global particle filter models the interaction between the target object and its surrounding objects.

2.2.2 Single Object Tracking

In the visual tracking community, single object tracking is the task of estimating the target location in an image sequence, the tracking target is manually defined by a bounding box in the first frame. Several challenges often occur such as occlusion, illumination changes and abrupt motion. The goal of most single object tracking methods is to construct a robust appearance model for the tracking target, so that it can be easily distinguished from the background of the video, despite the above mentioned challenges.

Early single target tracking works focus on building a matching model between the representation of the tracking target in the past frame and the current frame. Adam *et al.*, 2006 propose to represent the target by multiple image patches, then matching the similarity between the past and current frame is performed by measuring the histograms of the patches. The mean-shift tracker proposed by Comaniciu *et al.*, 2000 is another popular early tracking approach. In this approach, the target is represented by the statistic distribution of its color and texture. The matching model is expressed by a metric that is characterized by the Bhattacharyya distance. Oron, 2012 proposes a Locally Orderless Tracking (LOT) algorithm which estimates a probabilistic model of the variations of the target object over time using the Earth Mover's Distance (EMD). The model is defined in a joint spatial-appearance space to estimate the variations of the tracked object. The above mentioned approaches all work in a generative tracking fashion. The discriminative trackers, on the other hand, often show superior performance, where online learning strategies of the appearance of the tracking target have been extensively investigated (Collins *et al.*, 2005, Avidan, 2007, Grabner *et al.*, 2006, Babenko *et al.*, 2009, Kalal *et al.*, 2010, Dinh *et al.*, 2011, Danelljan *et al.*, 2016, Nam and Han, 2016). Collins *et al.*, 2005 propose to model the tracking as a foreground/background classification task. The feature space is continuously evaluated over time by an online feature tracking ranking

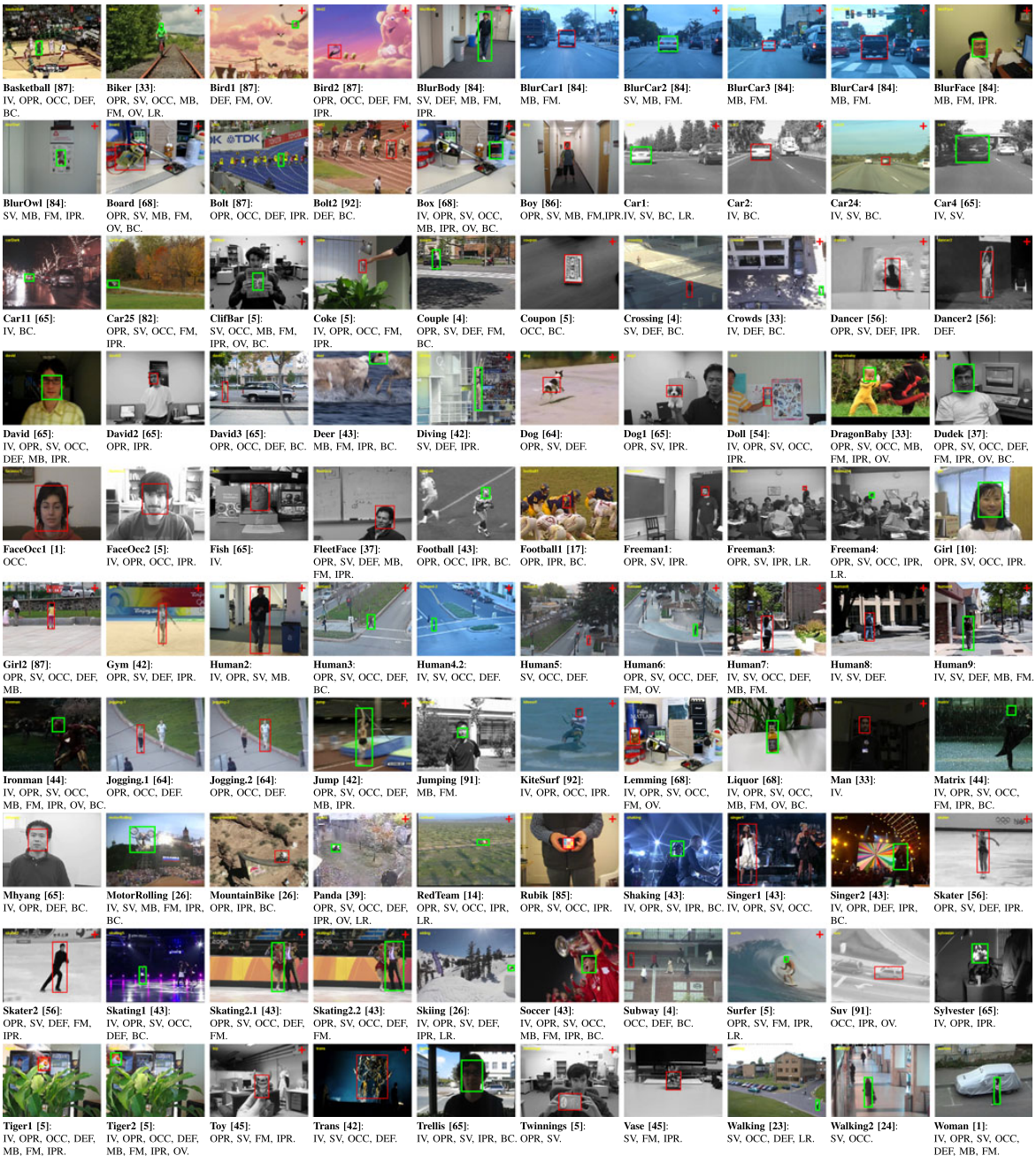


Figure 2.2: Visualization of the sequences used in the object tracking benchmark. The images are from Wu *et al.*, 2015. The bounding box in each frame indicates the tracking target. The green bounding boxes indicate that the corresponding sequence is used for extensive evaluations. The sequences are also annotated with attributes. E.g. IV represents significant illumination variation, OCC represents partial or full occlusion. For the detailed explanation of the attribute annotations, please refer to Wu *et al.*, 2015.

approach which is integrated into their tracking framework. In the work proposed by Avidan, 2007, the tracking task is also modelled as a binary classification problem. The difference comparing to the previous work is that instead of evolving features overtime, a boosting technique is utilized to ensemble weak classifiers to a strong classifier. As a result, each pixel in the next frame is labelled as the object or the background, giving the confidence map obtained by the strong classifier. The new position of the object is obtained by applying the mean shift algorithm on the confidence map. Similar idea is proposed by Grabner *et al.*, 2006, particularly the on-line trained classifier is trained with the negative examples that are selected from the image regions that surround the object. This way of choosing the negative examples makes the classifier more robust against the drifting problem that often occurs in challenging tracking videos. Babenko *et al.*, 2009 propose to model the tracking task as a Multiple Instance Learning (MIL) problem instead of the binary classification problem. The authors argue that the major challenge of the single object tracking task to choose reliable positive and negative examples over time. A slight inaccurate example may degrade the binary classifier and the final tracking performance. The MIL tracker allows the appearance model to be updated with a set of image patches instead of an individual image patch. A set is considered as a positive set if there is at least one positive instance. The classifier has to figure out which examples are the positive ones by itself, which injects flexibility in finding a good decision boundary. Recently, deep learning based single object trackers have been proposed and obtained superior performance. Hong *et al.*, 2015 propose to extract the image features of the tracking target using a deep CNN that is originally trained for large-scale image classification task. The extracted deep features are then used for on-line training of a SVM to distinguish positive and negative examples. The deep features that belong to the positive examples are back-projected to obtain a saliency map of the target on the input image. The saliency map highlights the image regions that discriminating the object from the background. Sequential Bayesian filtering is performed to track the object, where the saliency map is used as the observation. Nam and Han, 2016 propose a Multi-Domain Network (MDNet) which consists of several shared convolutional layers and separate branches, each has a binary classification layer. The shared convolutional layers are meant to capture a generic object representation and each branch is used for modelling the domain-specific information. The network is pre-trained with a large set of videos to learn a general representation of moving objects in videos. The focus of single object tracking task is to distinguish an arbitrary tracking target and the background, whereas for the multi object tracking task, the challenge is to estimate the number of objects and to recover the trajectory of each object over time. Most of the work proposed for the single object tracking task is then not suitable for the multi object tracking task. In the following, we will review the literature on multi object tracking.

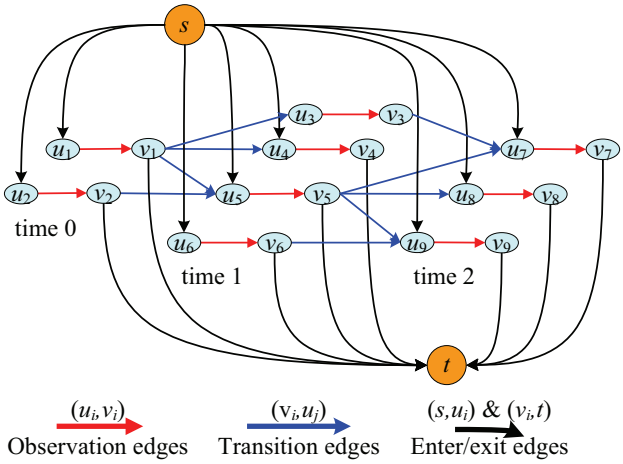


Figure 2.3: Visualization of the min-cost flow algorithm for tracking proposed by Zhang *et al.*, 2008. s and t represent the source and sink of the flow and there are 3 timesteps and 9 observations in this example. Image is from Zhang *et al.*, 2008.

2.2.3 Multi Object Tracking

To obtain the track of each person in a street scene video, which is the main topic of this thesis, multi object tracking methods are often used. Therefore, in this section, we mainly discuss the milestone works for multi object tracking, we also look at the most recent developments and their relations to our work. We categorize the related work by their main underlying concept. Note that it is not possible to clearly summarize each method using one keyword, as many tracking algorithms overlap at some point.

Network flow based tracking approach. A large body of multi target tracking methods are formulated as network-based optimization problems. One early work in this direction is proposed by Jiang *et al.*, 2007, where multi target tracking is cast as a multi-path searching problem. The interactions and mutual occlusions of tracks are explicitly modeled. The optimization is performed for all tracks simultaneously by linear programming relaxation. Another seminal work on network flow based tracking approach is propoed by Zhang *et al.*, 2008. In this work, the data association problem of multi-target tracking is cast as a min-cost flow problem which is illustrated in Fig. 2.3. The model intrinsically solves initialization and termination of tracks, as well as false alarms. The globally optimal solution can be obtained in polynomial time. More recently, Shitrit *et al.*, 2011 propose to model all potential locations over time and find trajectories that produce the minimum cost. Wang *et al.*, 2016 extend the work of Shitrit *et al.*, 2011 to track interacting objects simultaneously by using intertwined flow and imposing linear flow constraints. Pirsiavash *et al.*, 2011 show that their network flow formulation can be solved in polynomial time by a successive shortest path algorithm. A two-frame maximum weight independent set formulation followed by hierarchical merging and linking is proposed for the

tracking task in Brendel *et al.*, 2011.

Hierarchical data association. Hierarchical data association is another popular scheme in the tracking literature. A small batch of frames is first considered and produces local confident tracklets, and then longer tracks are built on such tracklets. In general, tracklet-based association is capable to reduce the state space and recover the trajectory of a target from long-term occlusions. However, such approaches need a separate tracklet generation step, and any mistakes introduced by the tracklet generation are likely to be propagated to the final solution. The early work Kaucic *et al.*, 2005 proposes a tracking-suspending-matching scheme. The areas where the objects are likely to be occluded are first identified. For the occluded area, tracking is suspended and re-initiated. The suspended tracklets are matched by motion appearance similarities. Wu and Nevatia, 2007 proposes a two-stage tracking scheme, combining network-flow and set cover techniques. The local information is aggregated to distinguish objects by bipartite-graph matches. For long-term occlusions, the linking of local tracklets is obtained by a logarithmic approximation solution to the set cover problem. Much of the recent literature on multi-target tracking follows the tracking-by-detection strategy which uses target detectors to establish an initial state-space of detection hypotheses in each frame. Wen *et al.*, 2014 also proposes to group detections into tracklets first, and then in the subsequent stage into tracks. In particular, Wen *et al.*, 2014 find tracks one at a time by relying on the greedy heuristic.

Variable number of targets. Determining the number of target is a well known difficulty for multi-target tracking. Various strategies are proposed to deal with the problem. Pirsiavash *et al.*, 2011 and Zamir *et al.*, 2012 rely on a greedy approach that recovers tracks one at a time by iteratively reducing the state space. Andriyenko *et al.*, 2012 jointly optimizes tracking trajectories and the number of tracking targets. Segal and Reid, 2013 implicitly encode the number of tracks by linking individual detection hypothesis between neighboring frames.

Re-identification for tracking. Several prior works have been proposed to exploit appearance information for multi-person tracking. Kim *et al.*, 2015 propose a target-specific appearance model which integrates long-term information and utilizes features from a generic deep convolutional neural network. Xiang *et al.*, 2015 propose to formulate tracking as a Markov decision process with a policy estimated on the labeled training data and present novel appearance representations that rely on the temporal evolution in appearance of the tracked target. Recently, Leal-Taixé *et al.*, 2016 propose to model the similarity between pairs of detections by CNNs. Several architectures have been explored and they present similar findings to our works, namely that forming a stacked input to CNNs performs best. However, our model proposed in Chapter 7 additionally incorporates human pose information, which improves the similarity measures by a notable margin.

Literature on multi-target tracking is vast, but several key properties reappear in

a number of successful approaches: leveraging long-range associations to prevent ID switches and recover missing detections caused by long-term occlusion (Andriyenko *et al.*, 2012, Zamir *et al.*, 2012); jointly inferring the number of tracks and solving the data association problem (Segal and Reid, 2013, Andriyenko *et al.*, 2012); exploring appearance information and combine it with long-range associations (Zamir *et al.*, 2012, Segal and Reid, 2013); integrate non-maximum suppression with tracking (Andriluka *et al.*, 2008, Pirsiavash *et al.*, 2011). Note that our Multicut based tracking formulations allow to combine all these in one framework.

2.2.4 Relations to Our Works

Approaches of Segal and Reid, 2013 and Zamir *et al.*, 2012 are perhaps the closest to ours. Similarly to Segal and Reid, 2013 we implicitly encode the number of tracks by linking detection hypotheses. However our approach jointly reasons about the connectivity of groups of hypotheses, whereas they connect individual hypotheses only. This allows us to postpone non-maximum suppression until temporal connections are resolved. Our tracking models incorporate long-range connections between hypotheses, and we show that our approaches achieve better experimental results compared to Segal and Reid, 2013. Zamir *et al.*, 2012 also introduce long-range connections between hypotheses and use an iterative greedy procedure finding tracks one at a time, whereas we jointly solve for all tracks. Henschel *et al.*, 2014 aim to delay resolution of local ambiguities by introducing "tree-tracklets" that delay locally ambiguous decision until more information is available. Our approach achieves the same goal by jointly associating groups of detections.

2.3 JOINT MOTION SEGMENTATION AND TRACKING

In this section, we briefly visit the literature of motion segmentation and focus on the recent works that combine tracking and segmentation in a unified framework.

Estimating object-level segmentation from long-term motion information has a long history of research. One popular way is to cast it as a dense point trajectory grouping problem. Related approaches have been suggested in (Brox and Malik, 2010, Lezama *et al.*, 2011, Ochs and Brox, 2012, Li *et al.*, 2013, Shi *et al.*, 2013, Ochs *et al.*, 2014, Rahmati *et al.*, 2014 and Ji *et al.*, 2014). Most of them employ the spectral clustering paradigm to generate segmentations. The seminal work on object segmentation by analysis of point trajectories (Brox and Malik, 2010) proposes an unsupervived object-level segmentation scheme by uitilizing long term motion cues. Ochs and Brox, 2012 further extend this idea by including high-order terms for modeling complex motions such as scaling and out-of-plane rotation. Other approaches (Zografos *et al.*, 2014, Elhamifar and Vidal, 2013) also model higher order motions by different means, where the approaches of (Cheriyadat and Radke, 2009, Dragon and Rosenhahn, 2012, Ochs *et al.*, 2014, Li *et al.*, 2013) base their segmentations on various pairwise affinities.

Combining high-level cues with low-level cues is an established idea in computer

vision and has been used successfully e.g. for image segmentation (Bertasius *et al.*, 2015). Similarly, motion trajectories have been used for tracking (Fragkiadaki and Shi, 2011, Fragkiadaki *et al.*, 2012). Object detections are also employed for segmenting moving objects Fragkiadaki *et al.*, 2015.

In Milan *et al.*, 2015, tracking and video segmentation are approached as one problem. Their approach employs a conditional random field (CRF), and is built upon temporal superpixels (Chang *et al.*, 2013) instead of point trajectories and strongly relies on unary terms on these superpixels learned using support vector machines. The proposed CRF model utilizes the high-level semantic information from object detectors and low-level information from superpixels. As a result, the segmentation as well as the bounding box track of each object in the scene are obtained.

2.3.1 Relations to Our Works

Our proposed method in Chapter 8 is substantially different in that we provide a unified graph structure whose partitioning both solves the low level problem, here, the motion segmentation task, and the high-level problem, i.e. the multi target tracking task, at the same time. In that spirit, the most related previous work is Fragkiadaki *et al.*, 2012, where detectlets, small tracks of detections, are classified in a graphical model that, at the same time, performs trajectory clustering. While we draw from the motivation provided in Fragkiadaki *et al.*, 2012, the key difference to our approach is that we cast both, motion segmentation and multi-target tracking, as clustering problems, allowing for the direct optimization of the Minimum Cost Multicuts (Chopra and Rao, 1993, Deza and Laurent, 2009). Thus, we perform bottom-up segmentation and tracking in a single step.

Furthermore, comparing to Zografos *et al.*, 2014 and Ochs and Brox, 2012, our joint multicut mode does not make use of any higher order motion models. In fact, much of the information these terms carry is already contained in the detections we are using, such that we can leverage this information with pairwise terms.

2.4 MULTI PERSON POSE ESTIMATION

In this section, we present a brief literature review of articulated human pose estimation. Human pose estimation methods traditionally work in the scenarios where the location of the person in the image is given either by cropping the image region that contains the person or by providing a person detection bounding box. The literature on this simplified pose estimation problem is numerous and we mostly discuss the seminal works. Then we review the works that tackle the problem of multi person pose estimation in unconstrained monocular images.

2.4.1 Single Person Pose Estimation

Most work on pose estimation targets the single person case. Methods progressed from simple part detectors and elaborate body models (Ren *et al.*, 2005, Ramanan, 2006, Jiang and Martin, 2009) to tree-structured pictorial structures (PS) models with strong part detectors (Pishchulin *et al.*, 2013, Yang and Ramanan, 2013, Chen and Yuille, 2014, Sapp and Taskar, 2013). Impressive results are obtained predicting locations of parts with convolutional neural networks (CNN) (Toshev and Szegedy, 2014, Tompson *et al.*, 2015). Tompson *et al.*, 2015 proposes a model that does not rely on explicit body modeling, and instead encodes appearance of part configurations via a convolutional multi-scale image representation. While body models are not a necessary component for effective part localization, constraints among parts allow to assemble independent detections into body configurations as demonstrated in Chen and Yuille, 2014 by combining CNN-based body part detectors with a body model (Yang and Ramanan, 2013). Wei *et al.*, 2016 propose a convolution pose machine for pose estimation where the spatial relation of human body part is modelled by a sequential architecture of convolutional networks. The belief maps of the previous steps are incorporated into the current stage, so that the localization of body joints are refined without explicit spatial models. Newell *et al.*, 2016 propose a stacked hourglass network where a sequence of hourglass networks is stacked to capture the body joint location as well as their spatial relations. Intermediate supervision is exploited and the results suggest that such supervision is critical for the final pose estimation performance. Chu *et al.*, 2017 propose to integrate multi-context attention mechanism into the stacked hourglass network. The attention mechanism is used to exploit the contextual information over the whole human body by incorporating CRFs to model the kinematic relations of body joints. The attention model is multi-scale as well, on one side, the holistic attention capture the consistency of the global pose configuration, on the other side, the detailed representation of human body is captured by the body part attention models. The final model presents the state-of-the-art single person pose estimation results on the MPII Human Pose Dataset.

2.4.2 Multi Person Pose Estimation

A popular approach to multi-person pose estimation is to detect people first and then estimate body pose independently (Sun and Savarese, 2011, Pishchulin *et al.*, 2012, Yang and Ramanan, 2013, Gkioxari *et al.*, 2014). Yang and Ramanan, 2013 propose a flexible mixture-of-parts model for detection and pose estimation. Yang and Ramanan, 2013 obtains multiple pose hypotheses corresponding to different root part positions and then performing non-maximum suppression. Gkioxari *et al.*, 2014 detect people using a flexible configuration of poselets and the body pose is predicted as a weighted average of activated poselets. Pishchulin *et al.*, 2012 detects people and then predicts poses of each person using a PS model. Belagiannis *et al.*, 2014 estimates poses of multiple people in 3D by constructing a shared space of

3D body part hypotheses, but use 2D person detections to establish the number of people in the scene. These approaches are limited to cases with people sufficiently far from each other and only limited overlapping body parts.

Similar to our work, Cao *et al.*, 2017 propose to detect body joint locations in the image then assemble the body joints into individual person so that the number of the people as well as the body pose of each person are obtained by the bottom-up assembling approach. They propose a novel non-parametric pairwise representation where the limbs of body are modelled by part affinity field. The final pose estimation of each person is obtained by performing a greedy matching algorithm.

2.4.3 Relations to Our Works

Our multi-person pose estimation work in Chapter 9 is closely related to Eichner and Ferrari, 2010 and Ladicky *et al.*, 2013 who also propose a joint objective to estimate poses of multiple people. Eichner and Ferrari, 2010 propose a multi-person pictorial structure (PS) model that explicitly models depth ordering and person-person occlusions. Our formulation is not limited by a number of occlusion states among people. Ladicky *et al.*, 2013 propose a joint model for pose estimation and body segmentation coupling pose estimates of individuals by image segmentation. Eichner and Ferrari, 2010 and Ladicky *et al.*, 2013 use a person detector to generate initial hypotheses for the joint model. Ladicky *et al.*, 2013 resort to a greedy approach of adding one person hypothesis at a time until the joint objective can be reduced, whereas our formulation can be solved with a certified optimality gap. In addition Ladicky *et al.*, 2013 rely on expensive labelling of body part segmentation, which our proposed approach does not require.

Similarly to Chen and Yuille, 2015, we aim to distinguish between visible and occluded body parts. Chen and Yuille, 2015 primarily focus on the single-person case and handles multi-person scenes akin to Yang and Ramanan, 2013, by performing part-based non-maximum suppression on the set of pose estimates. We consider the more difficult problem of full-body pose estimation, whereas both Eichner and Ferrari, 2010 and Chen and Yuille, 2015 focus on upper-body poses and consider a simplified case of people seen from the front.

Our work is related to early work on pose estimation that also relies on integer linear programming to assemble candidate body part hypotheses into valid configurations (Jiang and Martin, 2009). Their single person method employs a tree graph augmented with weaker non-tree repulsive edges and expects the same number of parts. In contrast, our novel formulation relies on a fully connected model to deal with an unknown number of people per image and visible body parts per person.

DETECTION AND TRACKING OF OCCLUDED PEOPLE 3

Contents

In this chapter, we consider the problem of detecting multiple people in crowded street scenes. State-of-the-art methods perform well in scenes with relatively few people, but are severely challenged by scenes with many subjects that partially occlude each other. This limitation is due to the fact that current people detectors fail when persons are strongly occluded. We observe that typical occlusions are due to overlaps between people and propose a people detector tailored to various occlusion levels. Instead of treating partial occlusions as distractions, we leverage the fact that person/person occlusions result in very characteristic appearance patterns that can help to improve detection results. We demonstrate the performance of our occlusion-aware person detector on a new dataset of people with controlled but severe levels of occlusion and on two challenging publicly available benchmarks outperforming single person detectors in each case.

3.1 INTRODUCTION

Single people detectors such as the powerful deformable part models (DPM, Felzenszwalb *et al.*, 2010) have shown promising results on challenging datasets. However, it is well known that current detectors fail to robustly detect people in the presence of significant partial occlusions. In fact, as we analyze in this chapter, the DPM detector starts to fail already at about 20% of occlusion and beyond 40% of occlusion the detection of occluded people becomes mere chance. Several methods, i.e. tracking and 3D scene reasoning approaches, have been proposed to track people even in the presence of long-term occlusions. Although these approaches allow us to reason across potentially long-term and full occlusions, they still require that each person is sufficiently visible at least for a certain number of frames. In many real

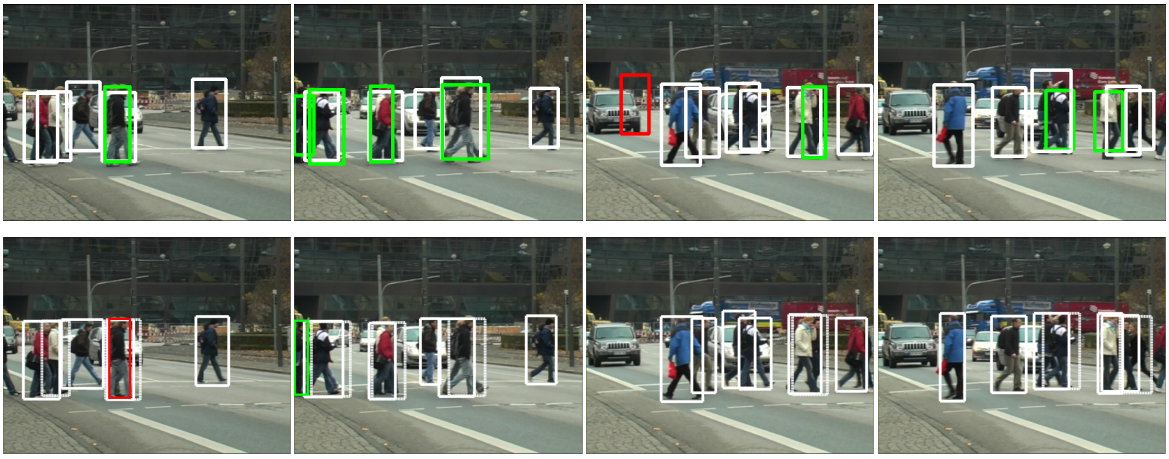


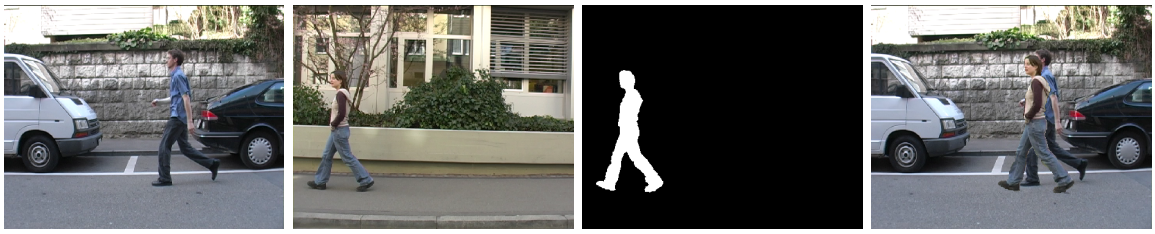
Figure 3.1: Detection results at equal error rate obtained with the approach of Barinova *et al.*, 2010 (top) and our joint detector (bottom) on the TUD-Crossing Andriluka *et al.*, 2008 dataset. False-positive detections are shown in red and missing detections in green. One of the two bounding boxes predicted from the two-person detection is shown with the dotted line.

scenes, however, e.g. when people walk side-by-side across a pedestrian crossing (see Fig. 3.1), a significant number of people will be occluded by 50% and more for the *entire* sequence.

To address this problem this chapter makes three main contributions. First, we propose a new double-person detector that allows us to predict bounding boxes of two people even when they occlude each other by 50% or more as well as a new training method for this detector. This approach outperforms single-person detectors by a large margin in the presence of significant partial occlusions (Sec. 3.2). Second, we propose a joint person detector that is jointly trained to detect single- as well as two-people in the presence of occlusions. This joint detector achieves state-of-the-art performance on challenging and realistic datasets (Sec. 3.3). Last, we integrate the above joint model into a tracking approach to show its potential for people detection and tracking occluded people (Sec. 3.4).

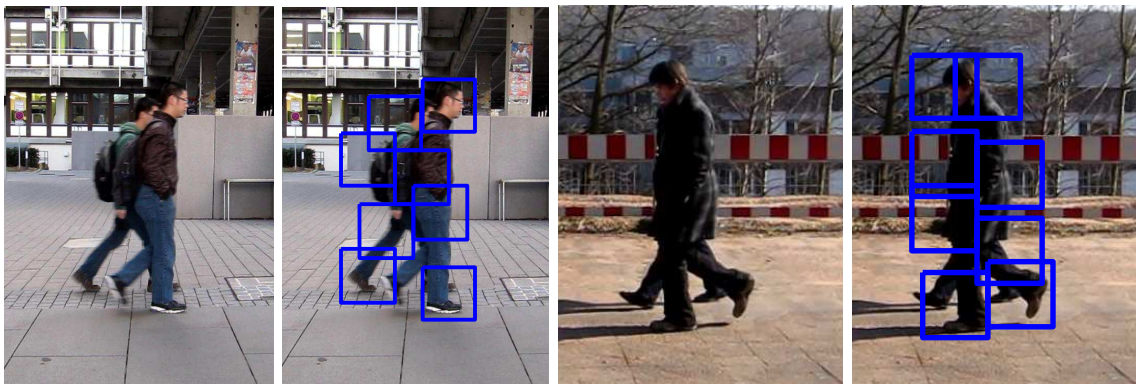
3.2 DOUBLE-PERSON DETECTOR

Our double-person detector builds on the DPM approach Felzenszwalb *et al.*, 2010. The key concept of our double-person model is that person/person occlusion patterns are explicitly used and trained to detect the presence of two people rather than to treat these occlusions as distractions or nuisance as it is typically done. Specifically, our double-person detector shares the deformable parts across two people which belong to the same (two-person) root filter. In that way localizing one person facilitates the localization of the counterpart in the presence of severe occlusions and the deformable parts allow us to improve the localization accuracy of both people



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

Figure 3.2: Procedure to synthetically generate training images for our double-person detector. (a) background person, (b) foreground person, (c) foreground person map, (d) generated synthetic training image.



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

Figure 3.3: Visualization of the deformable parts of the double-person detector. (a) and (c) are the test images from MPII-2person dataset. (b) and (d) are the visualization of the parts locations.

using the above mentioned occlusion patterns whenever appropriate (cf. Fig. 3.3). For this we build on the DPM framework to detect the presence of two people and to predict the bounding boxes of both people, the occluding person as well as the occluded person.

3.2.1 Double-person detector model

In full analogy to DPMs, our double-person detector uses a mixture of components. Each component is a star model consisting of a root filter that defines the coarse location of two people and n deformable part filters that cover representative parts and occlusion patterns of the two people. The vector of latent variables is given by $z = (c, p_0, \ldots, p_n)$, with c denoting the mixture component and p_i specifying the image position of the part and feature pyramid level l_i. The score of a double-person hypothesis is obtained by the score of each filter at the latent position

Figure 3.4: Examples of synthetically generated training images for different levels of occlusion: 5% to 10% (a), 20% to 30% (b), 40% to 50% (c) and 70% to 80% (d).

p_i (unary potentials) minus the deformation cost between root position and part position (pairwise potentials). As in Felzenszwalb *et al.*, 2010, the un-normalized score of a double-person hypothesis is defined by $\langle \beta, \Psi(x,z) \rangle$, where vector β is a concatenation of the root and all part filters and the deformation parameters, and $\Psi(x,z)$ is the stacked HOG features and part displacement features of sample x. $\Psi(x,z)$ is zero except for a certain component c. Therefore, we obtain the construction $\langle \beta, \Psi(x,z) \rangle = \langle \beta_c, \psi_c(x,z) \rangle$. Detection in the test image is done by maximizing over the latent variables z: $\arg\max_{(z)} \langle \beta, \Psi(x,z) \rangle$.

3.2.1.1 *Model training*

Let $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_N, y_N \rangle)$ denote a set of positive and negative training examples, with x_i corresponding to a bounding box enclosing either a pair of people or a background region and $y_i \in \{-1, 1\}$. Given this training set we learn the model parameters β using latent SVM Felzenszwalb *et al.*, 2010. This involves iteratively solving the quadratic program:

$$\begin{aligned} &\min_{\beta,\xi \geq 0} \frac{1}{2} \max_c \|\beta_c\|^2 + C \, \Sigma_{i=1}^{N} \xi_i \\ &\text{sb.t.} \quad y_i \langle \beta, \Psi(x_i, z) \rangle \geq 1 - \xi_i \qquad \xi_i \geq 0, \end{aligned} \tag{3.1}$$

and optimizing for the values of latent parameters z. The optimization objective in Eq. 3.1 includes a regularizer that has been proposed in Girshick *et al.*, 2010 and is

slightly different from the one in Felzenszwalb *et al.*, 2010. Instead of penalizing the norm of the whole parameter vector, it only penalizes maximum over the norms of the parameters of each component. The purpose of such regularization is to prevent one single component from dominating the model, and to make the scores of individual components more comparable. We solve the quadratic program with stochastic gradient descent and employ data-mining of hard-negative examples after each optimization round as proposed in Felzenszwalb *et al.*, 2010.

3.2.1.2 *Initialization*

The objective function of the latent SVM is non-convex, which makes the training algorithm susceptible to local minima. Instead of relying on the bounding box aspect ratio as in Felzenszwalb *et al.*, 2010, we initialize our model using different occlusion levels, which we found to produce slightly better results compared to standard initialization. This follows the intuition that the degree of occlusion is one of the major sources of the appearance variability and can be captured by different components. Other sources of appearance variability such as poses of people and varying clothing are then captured by displacement and appearance parameters of each component. In the experiments reported below we use a three component double-person model. The components are initialized with the occlusion levels 5%–25%, 25%–55%, and 55%–85%. The percentage of occlusion is defined as a percentage of the occluded pixels in the person segmentation.

3.2.1.3 *Bounding box predictions*

Given a double-person detection we predict the bounding boxes of individual people using linear regression. The location of each bounding box is modelled as

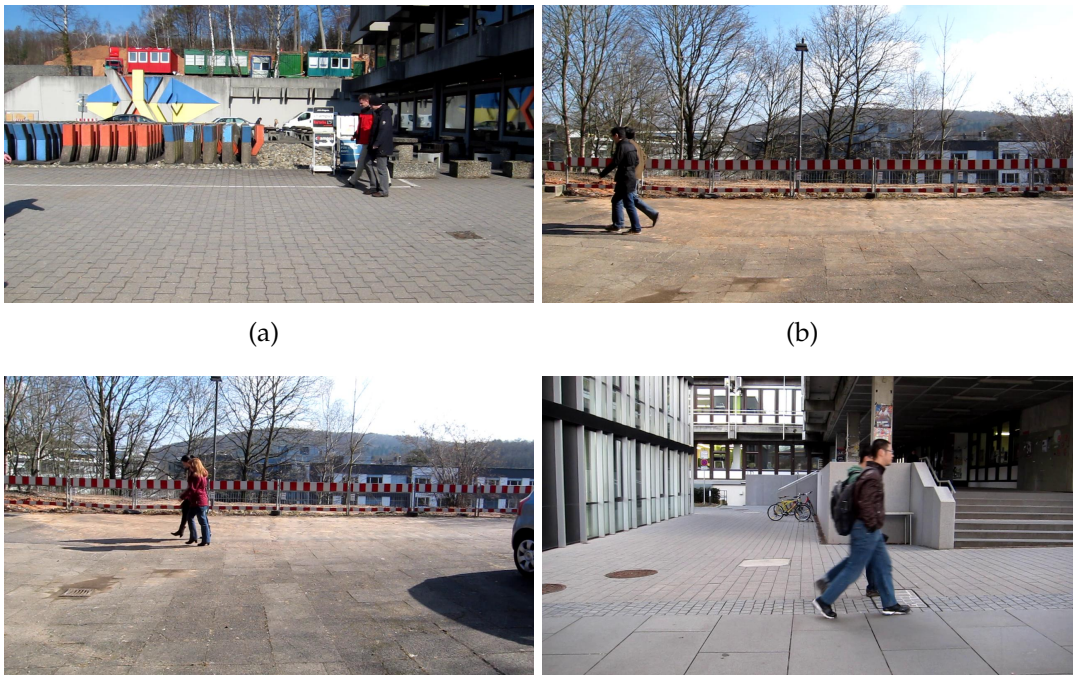
$$B_i = g_i(z)^{\mathrm{T}} \alpha_c + \epsilon_i, \tag{3.2}$$

where B_i is the predicted bounding box for a detection i, c is the index of the DPM component that generated the detection, and $g_i(z)$ is a $2 * n + 3$ dimensional vector that is constructed by the upper left corners of the root filter and the n part filters as well as the width of the root filter. ϵ_i is a Gaussian noise that models deviations between the predicted and observed location of the bounding box.

The regression coefficients α_c are estimated from all positive examples of component c. For each of the model components we estimate two separate regression models that correspond to the two people in the double-person detection. This procedure allows us to accurately localize both people despite severe occlusions, as can be seen e.g. in Fig. 3.6.

3.2.1.4 *Training data generation*

As it is difficult to obtain sufficient training data for the different occlusion levels of our double-person detector, we synthetically generate it. Fig. 3.2 illustrates this



(a) (b)

(c) (d)

Figure 3.5: Example images from the MPII-2Person dataset. The levels of occlusion in (a) to (d) are 30%, 50%, 70% and 80% respectively.

process. For each person we first extract the silhouette based on the annotated foreground person map. Next, another single-person image is selected arbitrarily and combined with the extracted silhouettes. In order to generate a double-person training dataset, we randomly select background images, 2D positions and scale parameters. Each synthetic image provides an accurate occlusion ratio estimated from the two persons' silhouettes. For the experiments reported below we generate 1,300 double-person training images from the 400 TUD training images (Andriluka *et al.*, 2008). For the synthetic dataset we uniformly sample occlusion levels between 0% and 85%, and scale factors between 0.9 and 1.1.

3.2.2 Experimental study

In order to explicitly compare single-person and double-person detector performance for person/person occlusion scenarios, we captured several video sequences and constructed a new double-person dataset (MPII-2Person) where the 850 double-person images are categorized by different occlusion levels[1] (see Fig. 3.5). The person segmentation and occlusion level are estimated from 2D truncated quadrics which are constructed from stick-man annotation.

[1]The training and test datasets are available at www.d2.mpi-inf.mpg.de/datasets

Single-person detector: Fig. 3.7(a) shows the performance of the standard DPM single-person detector on our double-person dataset. In case of little partial occlusion (red curve, below 5%), the single-person detector obtains good performance both in terms of recall (up to 90% recall) and high precision. However, the single-person detector already misses many people when the occlusion level is increased up to 15% (blue curve, maximal recall below 80%), and further decreases in the presence of more occlusion. When the occlusion level is 35% or more, the achieved recall is only slightly above 50%, indicating that in most cases only one of the two people is correctly detected.

Double-person detector: Fig. 3.7(b) shows the performance of our proposed double-person detector. The detector reaches nearly 100% recall with very few false positives, which is a significant improvement over the single-person detector. Interestingly, the performance for the lowest occlusion level (red curve, up to 5%) is lower than for the levels with more occlusion, which can be explained by the difficulty to differentiate a single person that does not occlude a second person from the case that a person occludes a second person significantly (e.g. 80%) (for an example of 80% occlusion see Fig. 3.6). Overall the detection precision is very high for all but the highest occlusion level (green line, up to 85%).

We now compare the double-person detector with two baselines that rely on the single-person detector. The first baseline is obtained by varying the threshold τ used in the non-maximum suppression (NMS) step. This parameter determines the minimum value of the "intersection over union" ratio required for one detection bounding box to suppress the other. The results of this experiment are shown in Fig. 3.8. For each detector we plot the area under the recall-precision curve (AUC) for the range of occlusion levels. For low occlusion levels, the detectors with low NMS thresholds perform reasonably well, however, their performance degrades quickly for higher levels of occlusion. Increasing the NMS threshold improves performance for the higher occlusion levels because the larger number of candidate detections survive NMS, but the performance for the low occlusion levels drops due to an increased number of false positives. The first observation from this experiment is that there is no single NMS threshold which works equally well for all levels of occlusion. The second observation is that our two-person detector (blue dashed line) outperforms all single-person detectors above.

Our second baseline is obtained by predicting the detection bounding boxes for two people based on the output of the single-person detector. To that end the bounding box of the second person is randomly generated in the vicinity of the single-person detection. We purposefully choose a small value of non-maximum suppression parameter $\tau = 0.3$ to prune the detections close to each other and to prevent conflicts between generated and detected bounding boxes. The result of this experiment corresponds to the "Predict double from single" curve in Fig. 3.8. The performance is similar or better than single-person detectors for a full range of NMS thresholds. Recall that the MPII-2Person dataset used in this experiment contains only images of two people walking close to each other, and good performance of the second baseline is not surprising. The performance of the second baseline however

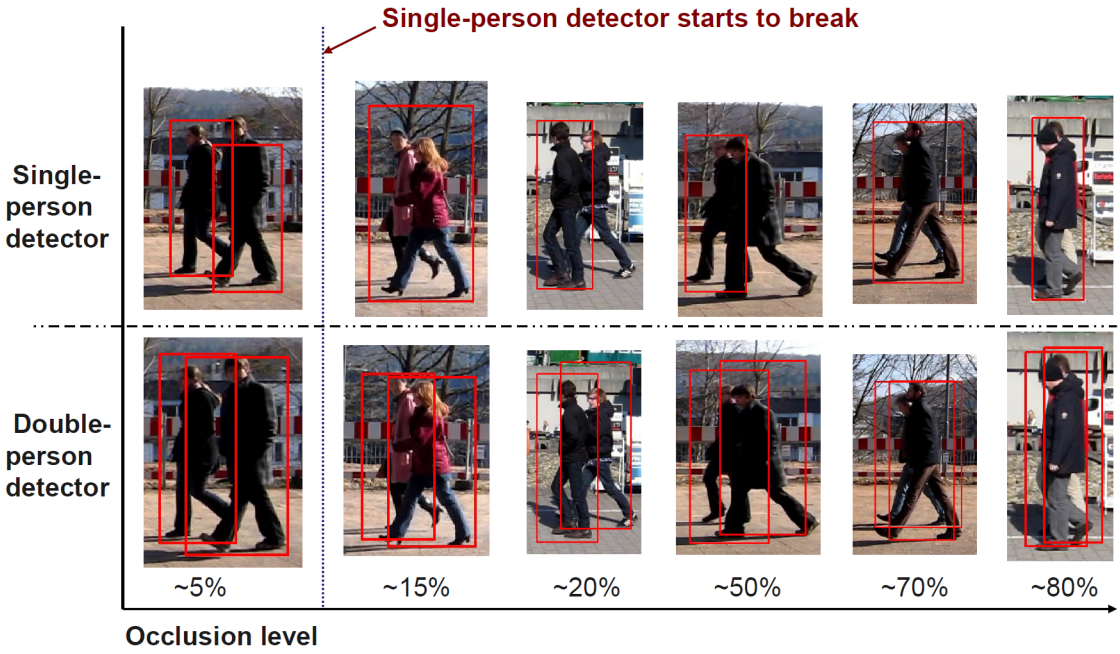
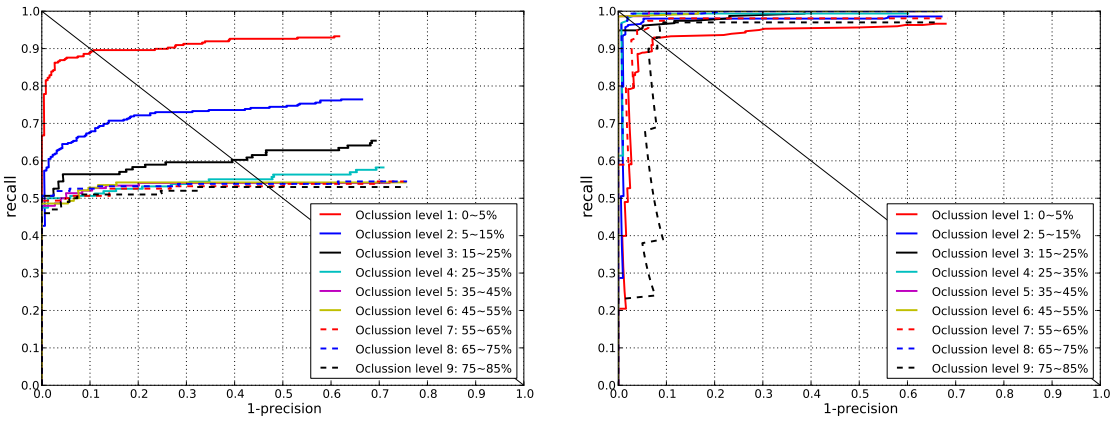


Figure 3.6: Qualitative comparison of single- and double-person detectors for different occlusion levels.

drops on images with small amounts of occlusion (less than 15%). Note that our double-person detector also clearly improves over the second baseline.

From these experiments we conclude that our double-person detector is much more robust than the single-person detector and obtains excellent performance both in terms of recall and precision, even for the heavy occlusion cases. Single person localization (bounding boxes prediction) is not a trivial task, especially for intermediate occlusion level cases $(30\% \sim 60\%)$, because we observe fair evidence from both persons, which can be distracting for single bounding box localization. However, the results show that our double-person detector accurately and robustly predicts the single bounding box for the above mentioned case as well. Fig. 3.6 shows comparative qualitative results. For the same test examples, our double-person detector correctly detects the position of two persons and predicts their respective bounding box with high accuracy.



(a) Single-person detector (b) Double-person detector

Figure 3.7: Detection performance of single- and double-person detectors for different occlusion levels on the MPII-2Person dataset.

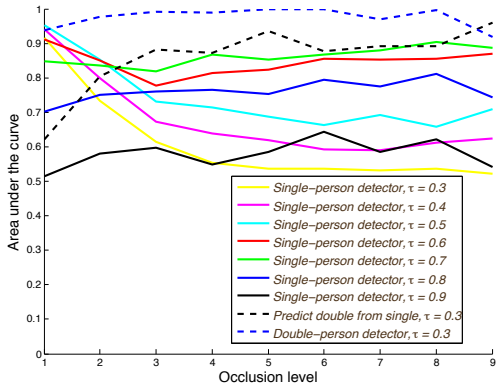


Figure 3.8: Comparison of the double-person detector with various baselines based on the single-person detector on the MPII-2Person dataset. See Fig. 3.7 for the definition of occlusion levels (x-axis).

3.3 MULTI PERSON DETECTOR

The previous section has shown that our double-person detector can indeed outperform a single person detector when people occlude each other by 25% or more. However, the employed dataset was somewhat idealistic as it contained exactly two people that occluded each other at various degrees. In realistic datasets we will have both single people that are fully visible and two and more people that occlude each other. This section therefore proposes a detector that combines both single and two-person detectors into a single model that is jointly trained. The model is again built upon the DPM-approach where the role of the different components is now to differentiate between single and two people as well as between different occlusion levels among two people.

3.3.1 Joint Person Detector

We jointly train single- and double-person detectors by representing them as different components of the DPM. We allocate three components for the double detector and three components for the single-person detector which, after mirroring, results in a 12 component DPM model. Similarly to Sec. 3.2 we initialize the double-person components with training examples corresponding to gradually increasing levels of occlusion. For the single-detector components we rely on the standard initialization based on the bounding box aspect ratio. During learning we allow training examples to be reassigned to other components of the DPM model, but prevent assignments of 2-person examples to 1-person components and vice versa. We found this to be important to improve detection of two people in cases of particularly strong occlusion that are otherwise often incorrectly handled by the single-person components.

The performance of the joint detector strongly depends on its ability to distinguish between single and double-person hypotheses, which requires the scores of single and double person components to be comparable to each other. To achieve such comparability we jointly optimize the parameters of all detection components. The optimization procedure used for learning the DPM parameters described in Sec. 3.2 couples the training of each component in several ways. The components are jointly regularized by penalizing the maximum over the norms of the component parameters (cf. Eq. 3.1). In addition the training examples can be reassigned between components after each optimization round, and hard negative mining and optimization stopping criterion depends on the full model and not on an individual component. Even though we fix the assignment of training examples to single and double-person components, the other coupling mechanisms remain. The empirical evidence suggests that such joint training makes the output scores of each component comparable Girshick *et al.*, 2010. In this chapter we follow this standard practice, but refer to the work in Chapter 4 where we further address this issue by reformulating our joint detector using structural SVM framework and modifying the loss function to penalize detection of single people with double-person components

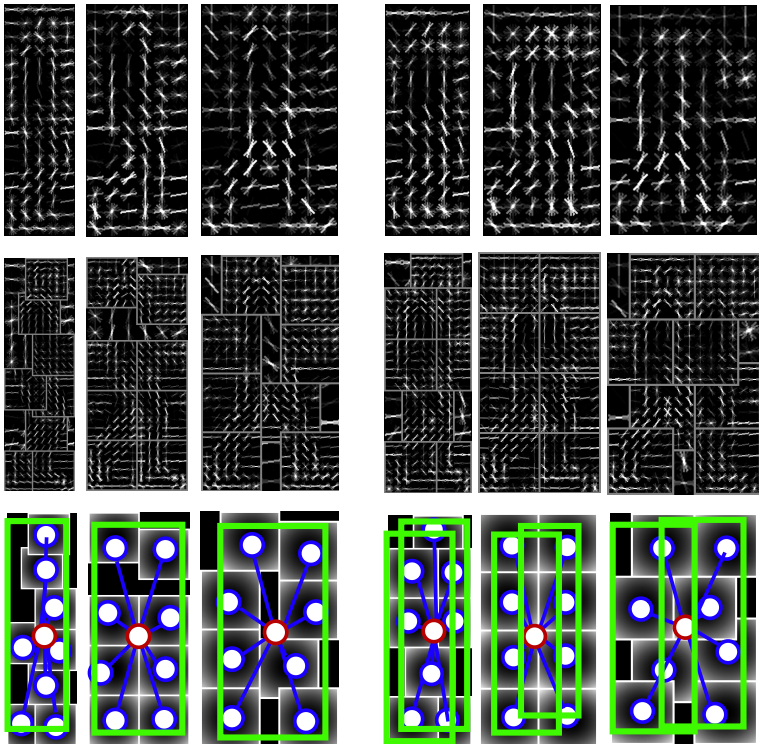
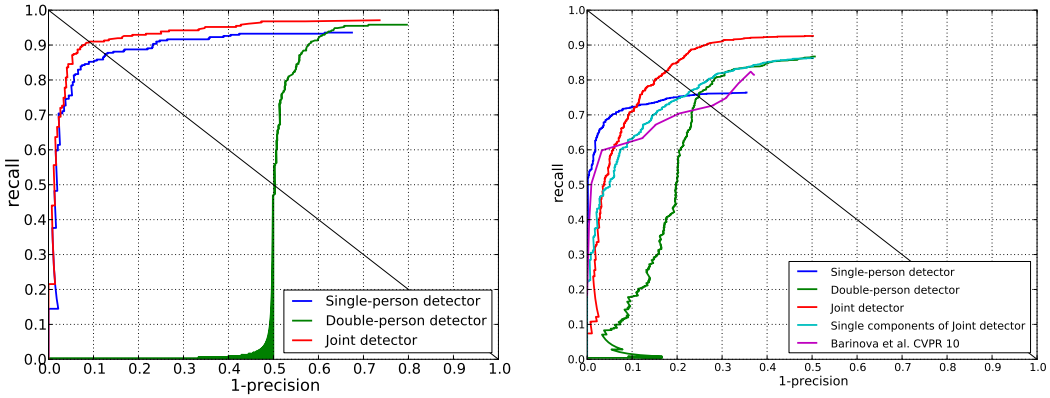


Figure 3.9: Visualization of the root filters (first row), part filters (second row) and mean part locations and detection bounding boxes (third row) of the joint person detector. The first three columns correspond to the single-person and the last three columns to the double-person components.

and vice versa. In Fig. 3.9 we visualize the root and part filters of the joint detector. Note the substantial differences between the filters of the single and double-person components.

Training data: We train our joint detector on the combination of 1-person and 2-person training sets described in Sec. 3.2.1.4, but slightly modify the initial assignment of images to the DPM components. We assign training images with less than 5% occlusion to the single-person training dataset, because in that case the single-person detector already obtains high performance for both people. We initialize the 3 double-person DPM components with images corresponding to occlusion levels: 5%–25%, 25% – 55%, and 55% – 75%.

Non-maximum suppression (NMS): The NMS in the joint detector is more complicated than in the standard DPM since we have bounding box predictions from two different types of detections (single and two-person detections) as well as strongly overlapping bounding box predictions from our two-person components. We thus implement NMS in two steps. The first step is performed prior to bounding box prediction and already removes a large portion of multiple detections on the same person. In this first step two-people detections and single-person detections compete and suppress each other depending on the respective score. The remaining multiple detections are either due to multiple two-person detections in cases when more than



(a) TUD-Pedestrians (b) TUD-Crossing

Figure 3.10: Detection performance on TUD-Pedestrians (a) and TUD-Crossing (b).

two people appear close to each other (e.g. rightmost three people in the fourth image in Fig. 3.1) or detections with significantly different bounding box aspect ratios. Given the reduced set of hypotheses after the first round of NMS, we perform bounding box prediction followed by the second round of NMS. This second step corresponds to the NMS typically performed for DPM Felzenszwalb *et al.*, 2010. The second round is done independently for single-person and two-person components of DPM, as we found that one-person detections may incorrectly suppress two-person detections otherwise. During NMS of detections from the two-person components we additionally prevent two bounding boxes predicted from the same double-person detection from suppressing each other. As an illustrative example, we could correctly detect all three people in the fourth image on Fig. 3.1 despite strong occlusion of the middle person. In that case the single-person detections were predicted from two double-person detections and multiple detections on the middle person were correctly removed by the second stage of the non-maximum suppression.

3.3.2 Results

We evaluate the performance of our joint detector on two publicly available datasets, "TUD-Pedestrians" and "TUD-Crossing", originally introduced in Andriluka *et al.*, 2008. "TUD-Pedestrians" contains 250 images of typical street scenes with 311 people all of which are fully visible. "TUD-Crossing" contains a sequence of 201 images with 1008 annotated people that frequently occlude each other partially or even fully. To capture the full range of occlusions we extended the annotations of the "TUD Crossing" dataset to include also strongly occluded people, which resulted in 1186 annotated people.

We begin our analysis with the "TUD-Pedestrians" dataset. Detection results are shown in Fig. 3.10(a) as recall-precision curves. Since this dataset does not contain

any occluded people our double-person detector (Sec. 3.2) generates numerous false positives, interpreting each person as a pair of people in which one of the persons is severely occluded. As expected the single-person detector performs well on this dataset, achieving an equal error rate (EER) of 87% . The joint detector slightly improves over the single person detector achieving 90.5% EER. This result is a bit surprising because the joint detector is trained to solve a more difficult problem of detecting both fully visible and partially occluded people. We attribute the improvement of the joint detector to the training set that in addition to real images has been augmented with artificial training examples (c.f. Sec. 3.2). This parallels the recent results on using artificially generated data for training of people detection and pose estimation models (Shotton *et al.*, 2011; Pishchulin *et al.*, 2011).

The evaluation on "TUD Pedestrian" demonstrates that integrating single- and double-person detectors in the same model does not result in a performance penalty in the case when people are fully visible. In order to assess the joint detector in realistic scenes that contain both occluded and fully visible people we evaluate its performance on the TUD-Crossing dataset. Quantitative results are shown on Fig. 3.10(b) and a few example images in Fig. 3.1 (bottom row). First we compare the performance of single and double-person detectors, which achieve approximately the same EER of 76%. The double-person detector achieves higher recall compared to the single-person detector, being able to detect even strongly occluded people. However, the precision of the double-person detector suffers from multiple detections of fully visible people. The single-person detector produces fewer false positive detections, but also fails to detect occluded people, saturating at a recall of 76%. Finally, the joint detector significantly improves over both single and double person detectors, achieving an EER of 83%. In order to gain further insight into the workings of our approach, we conduct an additional experiment in which we measure the performance of the detector composed of the single-person components of the joint detector. The results are also shown in Fig. 3.10(b). The single-components detector performs slightly better than the single-person detector (76% vs. 77% EER), but does not reach the performance of the complete joint detector (77% vs. 83% EER).

Note that while demonstrating overall improvement, the joint detector has a somewhat lower performance in the high precision area compared to the single person detector. Inspecting the false positives of the joint detector with highest scores reveals that most of them correspond to cases when one-person and two-person components of the detector fired simultaneously on the same pair of people, but these detections where sufficiently far from each other to persist through the non-maximum suppression step (e.g. false positive detection in the first image on Fig. 3.1).

Finally, we compare the performance of our approach with the Hough transform based detector of Barinova *et al.*, 2010, which is specifically designed to be robust to partial occlusions. The authors of Barinova *et al.*, 2010 kindly provided us their detector output (in terms of bounding boxes) which allows to compare their result on our full ground-truth annotations, making these results directly comparable to the rest of our experiments (Fig. 3.10(b)). The approach of Barinova *et al.*, 2010 improves

over the single-person detector in terms of final recall, but loses some precision, likely because their local features are rather weak compared to the discriminatively trained DPM model. Our joint model outperforms the approach of Barinova *et al.*, 2010 by a large margin. Fig. 3.1 shows a few example frames from the "TUD-Crossing" sequence, comparing our joint detector with the results of Barinova *et al.*, 2010. Note that our approach is able to correctly detect occluded people in the presence of very little image evidence (e.g. three pairs of people in the second image), whereas the approach of Barinova *et al.*, 2010 fails in such cases. At the same time our approach also correctly handles detection of single people (e.g. second and third images).

3.4 MULTI PERSON TRACKING

In this section we compare the performance of the single-person and the joint detectors (Sec. 3.3) in the context of multiple people tracking. To that end we rely on two recently proposed tracking approaches Andriyenko and Schindler, 2011; Pirsiavash *et al.*, 2011. Both of them employ the tracking-by-detection strategy and require output of the person detector as a prerequisite for tracking. In the following we first introduce these approaches and then discuss the experimental results.

The approach of Andriyenko and Schindler, 2011 formulates tracking as a continuous energy minimization problem. Given a set of person detections in each frame it recovers tracks of people by minimizing an objective function of the form

$$E(\mathbf{X}) = E_{obs} + \alpha E_{dyn} + \beta E_{exc} + \gamma E_{per} + \delta E_{reg}, \tag{3.3}$$

where $\mathbf{X}$ is a set of tracks, E_{obs} is a data term that encourages tracks that align well with the person detections, and the terms E_{dyn}, E_{exc}, and E_{per} encode prior assumptions on the tracking trajectories that encourage smooth and persistent trajectories without collisions. The term E_{reg} is a regularizer that penalizes the total number of trajectories. All terms in Eq. 3.3 depend on $\mathbf{X}$, and we omit explicitly stating this dependency for the brevity of notation. We refer to Andriyenko and Schindler, 2011 for the detailed description of the terms in Eq. 3.3.

The approach of Andriyenko and Schindler, 2011 is particularly suited for our task of evaluating different detectors in the context of tracking-by-detection because it relies on a clean formulation that directly accepts object detections as input, and only depends on a handful of free parameters. The only adaptation needed to integrate a particular object detector into the tracking system is to estimate the parameters α, β, γ and δ in Eq. 3.3. In our evaluation we rely on the publicly available implementation provided by the authors[2], but re-estimate the parameters of the objective function by performing a grid search independently for each of the detectors.

As second tracking approach in our experiments we use the multi-person tracker from Pirsiavash *et al.*, 2011. Similarly to Andriyenko and Schindler, 2011 this approach recovers tracks of multiple people by minimizing the joint objective function

[2] http://www.gris.tu-darmstadt.de/~aandriye

that combines the people detection likelihood with the smoothness prior on the track locations. The optimization is performed using an iterative greedy shortest-path algorithm. At each iteration it finds the best track and removes its hypotheses from the search space. The procedure is repeated as long as the newly found tracks have a negative cost and therefore decrease the value of the overall objective function. The objective function optimized in Pirsiavash *et al.*, 2011 is conceptually similar to the one used in Andriyenko and Schindler, 2011, but differs in the details of the likelihood and motion smoothness terms. The approach of Pirsiavash *et al.*, 2011 directly links the people detections across frames, whereas the approach of Andriyenko and Schindler, 2011 has a soft constraint that pulls the tracks towards detections but permits slight deviations. Moreover, the approach of Andriyenko and Schindler, 2011 relies on a constant velocity prior that is more suitable for tracking walking pedestrians compared to constant position prior used in Pirsiavash *et al.*, 2011. Finally, Andriyenko and Schindler, 2011 explicitly discourage multiple explanations of the image detections by several tracks via the exclusion term, whereas Pirsiavash *et al.*, 2011 achieves this using non-maximum suppression. The tracker in Pirsiavash *et al.*, 2011 also incorporates occlusion handling by allowing tracks that skip several consecutive frames with low detection likelihood. In our experiments we rely on the publicly available implementation of Pirsiavash *et al.*, 2011 and use the default tracking parameters provided by the authors[3].
We quantify the tracking performance using the CLEAR MOT metrics Bernardin and Stiefelhagen, 2008. The tracking results are evaluated with respect to the following characteristics: recall, precision, multi-object tracking accuracy, multi-object tracking precision, and the number of mostly tracked and mostly lost targets. Recall and precision are computed in the same way as in the evaluation of the detection performance, but using the ground truth targets and the tracker outputs. Multi-object tracking accuracy *(MOTA)* is the combined metric that takes missed targets, false alarms and identity switches into account. Multi-object tracking precision *(MOTP)* is computed using the average distance between the predicted track and the ground truth trajectory. *MT* is the absolute number of mostly tracked trajectories, and *ML* is the absolute number of mostly lost trajectories. The hit/miss threshold is 50% overlap between the ground truth targets and the tracker outputs in 2D.

We evaluate the full system composed of either our single-person or our joint detector and one of the tracking algorithms Andriyenko and Schindler, 2011; Pirsiavash *et al.*, 2011 on the TUD-Crossing dataset. The results are shown in Tab. 3.1.

First, we present the results obtained with the tracker of Andriyenko and Schindler, 2011. The single-person detector significantly improves over the result of Andriyenko and Schindler, 2011 that was obtained using a detector from Wojek *et al.*, 2010 based on the HOG and optical flow features. The best result is obtained using our joint detector, that improves over the single-person detector both in terms of recall, and with respect to MOTA/MOTP tracking metrics. Fig. 3.11 shows several example frames visualizing the tracking results. Note that the tracker based on the joint detector is able to track people even under significant partial occlusions (e.g.

[3] http://people.csail.mit.edu/hpirsiav

track 2 in the first three images), and is able to track subjects for longer periods of time (e.g. track 10 of the joint detector (third row) corresponds to two tracks of the single-frame detector (second row)). Tracking based on the output of the joint detector also results in fewer identity switches (16 for the single-person detector vs. 11 for the joint detector). Inspection of the output of the single-person detector reveals that in the case of strong partial occlusions the detection output often jumps between occluder and occluded subjects, which results in frequent identity switches in corresponding track. In contrast the joint detector typically includes detections of both subjects into the hypotheses set, which facilitates more consistent tracking.

Note that although the joint detector achieves the best result, the improvement over the single-person detector is only 3.2% of MOTA. This is somewhat surprising given the large improvement of the joint detector on the detection task. This result could be due to the particular choice of the objective function which contains the term E_{exc} which explicitly penalizes tracks which collide with each other in the image space. In the case of strong partial occlusions tracks of both subjects might be rather close to each other, where this exclusion term is likely to be suboptimal. The tracking algorithm does not take advantage of the additional information contained in the output of the joint detector that is able to explicitly label detections as a pair of occluded and occluding people. We envision that a more careful integration of the joint detector into the tracking framework could lead to larger performance gains and leave such integration to the future work.

Next, we evaluate our proposed detectors in combination with the tracking algorithm of Pirsiavash *et al.*, 2011. The results are shown in the last two rows of the Tab. 3.1. The tracking results obtained both with single and joint-person detectors are somewhat lower than with the tracker of Andriyenko and Schindler, 2011. The large difference in tracking recall is particularly striking. For example, in the case of the single-person detector we obtain 79.9% for the tracker of Andriyenko and Schindler, 2011 and 68.3% for the tracker of Pirsiavash *et al.*, 2011. The difference could be due to a more sophisticated design of the objective function in Andriyenko and Schindler, 2011 that explicitly encourages longer tracks by incorporating the persistence term. Importantly, for both trackers we achieve noticeable improvement from substituting the single-person with the joint-person detector. The improvement for the tracker of Pirsiavash *et al.*, 2011 is particularly pronounced. For example, the joint detector is able to improve the aggregated tracking accuracy measure MOTA from 63.3 to 70.7. We hypothesize that the improvement for Pirsiavash *et al.*, 2011 is larger because it operates by linking a discrete set of detection hypotheses over time and is therefore more sensitive to missing detections. In contrast the tracker of Andriyenko and Schindler, 2011 only uses detections as observations for tracking and explicitly reasons about continuous trajectories, which allows it to better handle gaps in detections.

Method	Recall	Precision	MOTA	MOTP	MT	ML
Andriyenko and Schindler, 2011	69.8	92.4	63.0 %	75.5 %	7	1
Our single-person detector + Andriyenko and Schindler, 2011	79.9	96.2	75.2 %	77.7 %	7	0
Our joint detector + Andriyenko and Schindler, 2011	**82.8**	96.2	**78.4** %	**77.9** %	**8**	0
Our single-person detector + Pirsiavash *et al.*, 2011	68.3	**98.4**	63.3 %	76.3 %	5	0
Our joint detector + Pirsiavash *et al.*, 2011	**77.7**	96.2	**70.7** %	**77.1** %	6	0

Table 3.1: 2D tracking evaluation on the TUD-Crossing dataset.

Figure 3.11: Tracking results on the TUD-Crossing dataset obtained with the approach of Andriyenko and Schindler, 2011 (top row), our single-person detector (middle row) and our joint detector (bottom row). Colors and numbers indicate tracks corresponding to different people.

3.5 CONCLUSIONS

Occlusion handling is a notorious problem in computer vision that typically requires careful reasoning about relationships between objects in the scene. Building on the state-of-the-art DPM detector Felzenszwalb *et al.*, 2010, we developed a joint model that is trained to detect single people as well as pairs of people under varying degrees of occlusion. In contrast to standard people detectors that treat occlusions as nuisance and degrade quickly in the presence of strong occlusions, our detector is specifically trained to capture various occlusion patterns. Our joint detector significantly improves over a single-person detector when detecting people in crowded street scenes, without losing performance on images with one person only. On the challenging TUD-Crossing benchmark our joint detector improves the previously best result of Barinova *et al.*, 2010 from 73% to 83% EER. Finally, we demonstrated the effectiveness of our joint detector as a building block for tracking-by-detection. We envision that our approach can be applicable to detection of multiple people in various domains (e.g. surveillance videos or sports scenes) and can be used as a building block for tracking-by-detection, pose estimation, and activity recognition in multi-people scenes.

LEARNING PEOPLE DETECTORS FOR TRACKING 4

Contents

People tracking in crowded real-world scenes is challenging due to frequent and long-term occlusions. Recent tracking methods obtain the image evidence from object (people) detectors, but typically use off-the-shelf detectors and treat them as black box components. In this chapter we argue that for best performance one should explicitly train people detectors on failure cases of the overall tracker instead. To that end, we first propose a novel joint people detector that combines a state-of-the-art single person detector with a detector for pairs of people, which explicitly exploits common patterns of person-person occlusions across multiple viewpoints that are a frequent failure case for tracking in crowded scenes. To explicitly address remaining failure modes of the tracker we explore two methods. First, we analyze typical failures of trackers and train a detector explicitly on these cases. And second, we train the detector with the people tracker in the loop, focusing on the most common tracker failures. We show that our joint multi-person detector significantly improves both detection accuracy as well as tracker performance, improving the performance on standard benchmarks.

4.1 INTRODUCTION

People detection is a key building block of most people tracking methods (Andriyenko *et al.*, 2012; Yang and Nevatia, 2012; Zamir *et al.*, 2012). Although the performance of people detectors has improved tremendously in recent years, detecting partially occluded people remains a weakness of current approaches Dollár

Figure 4.1: Tracking results using the proposed joint detector on four public datasets: (clockwise) TUD-Crossing, ParkingLot, PETS S2.L2 and PETS S1.L2.

et al., 2012. This is also a key limiting factor when tracking people in crowded environments, such as typical street scenes, where many people remain occluded for long periods of time, or may not even become fully visible for the entire duration of the sequence.

The starting point of this chapter is the observation that people detectors used for tracking are typically trained independently from the tracker, and are thus not specifically tailored for best tracking performance. In contrast, the present work aims to train people detectors explicitly to address failure modes of tracking in order to improve overall tracking performance. However, this is not straightforward, since many tracking failures are related to frequent and long-term occlusions – a typical failure case also for people detectors.

We address this problem in two steps. First, we target the limitations of people detection in crowded street scenes with many occlusions. Occlusion handling is a notoriously difficult problem in computer vision and generic solutions are far from being available. Yet for certain cases, successful approaches have been developed that train effective detectors for object compositions, which can then be decoded into individual object detections. Their key rationale is that objects in such compositions exhibit regularities that can be exploited. We build on these ideas, focusing on person-person occlusions, which are the dominant occlusion type in crowded street scenes. Our first contribution is a novel structural loss-based training approach for a joint person detector, based on structured SVMs.

In the second step of our approach, we specifically focus on patterns that are relevant to improving tracking performance. In general, person-person occlusions may result in a large variety of appearance patterns, yet not all of these patterns

are necessarily frequent in typical street scenes. Furthermore, not every pattern will possess a discriminative appearance that can be detected reliably in cluttered images. Finally, some of the person-person occlusion cases are already handled well by existing tracking approaches (e.g., short term occlusions resulting from people passing each other). We argue that the decision about incorporating certain types of occlusion patterns into the detector should be done in a tracking-aware fashion, either by manually observing typical tracking failures or by directly integrating the tracker into the detector training.

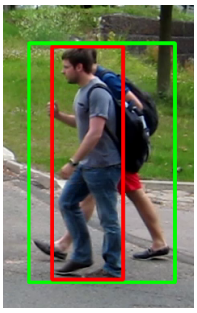
Our second contribution is to propose and evaluate two alternative strategies for the discovery of useful multi-view occlusion patterns. First, we manually define relevant occlusion patterns using a discretization of the mutual arrangement of people. In addition to that, we train the detector with the tracker in the loop, by automatically identifying occlusion patterns based on regularities in the failure modes of the tracker. We demonstrate that this tighter integration of tracker and detector improves tracking results on three challenging benchmark sequences.

4.2 JOINT PEOPLE DETECTION

Before describing our multi-view joint people detector, let us briefly review the deformable parts model (DPM, Felzenszwalb *et al.*, 2010), which forms the basis of our approach. The DPM detector is based on a set of M detection components. Each component is represented by a combination of a rigid root filter F_0, and several part filters $F_1, \ldots, F_n$, which can adjust their positions w.r.t. the root filter in order to capture possible object deformations $p_1, \ldots, p_n$. The detection score of the DPM model is given by the sum of the responses of the root and part filters, a bias b, and the deformation costs between the ideal and the inferred locations of each part (with parameters $d_1, \ldots, d_n$). The positions of the part filters and the component assignment m are assumed to be latent variables $h = (p_1, \ldots, p_n, m)$, which need to be inferred during training and testing. Given training images with ground truth labels, the parameters $\beta = (F_0, F_1, \ldots, F_n, d_1, \ldots, d_n, b)$ are trained by iterating between finding the optimal position of the latent parts in each training example and optimizing the model parameters given the inferred part locations. At test time the model is evaluated densely in the image and each local maximum is used to generate a detection bounding box hypothesis, aided by the model parts. The initial set of detections is then refined by non-maximum suppression.

4.2.1 Overview.

We now use the DPM model to build a joint people detector, which overcomes the limitations imposed by frequent occlusions in real-world street scenes. In doing so, we go beyond the work in Chapter 3 in several significant ways: *(1)* The approach in Chapter 3 focused on side-view occlusion patterns, but crowded street scenes exhibit a large variation of possible person-person occlusions caused by people's



(a) Double person outscores single person with Δ_{VOC}

(b) Double person outscores single person with $\Delta_{\text{VOC+DT}}$

Figure 4.2: Structured training of joint people detectors: Green – correct double-person bounding box. Red – single-person detection whose score should be lower by a margin.

body articulation or their position and orientation relative to the camera. To address this we explicitly integrate *multi-view person/person occlusion patterns* into a joint DPM detector. *(2)* We propose a *structured SVM formulation* for joint person detection, enabling us to incorporate an appropriate structured loss function. Aside from allowing to employ common loss functions for detection (Jaccard index, a.k.a. VOC loss), this allows us to leverage more advanced loss functions as well. *(3)* We model our joint detector as a mixture of components that capture appearance patterns of either a single person, or a person/person occlusion pair. We then introduce an explicit variable modeling the *detection type*, with the goal of enabling the joint detector to distinguish between a single person and a highly occluded person pair. Incorporating the detection type into the structural loss then allows us to force the joint detector to learn the fundamental appearance difference between a single person and a person/person pair.

Before going into detail on learning occlusion patterns in Sec. 4.4, let us first turn to our basic structured SVM formulation for joint person detection.

4.2.2 Structural learning for joint detection.

We adapt the structured SVM formulation for DPMs proposed in Pepik *et al.*, 2012 for our joint person detection model. Given a set of training images $\{I_i | i = 1, \ldots, N\}$ with structured output labels $y_i = (y_i^l, y_i^b)$, which include the class label $y_i^l \in \{1, -1\}$ and the 2D bounding box position y_i^b, we formulate learning the parameters of the DPM, β, as the optimization problem

$$\begin{aligned} \min_{\beta, \xi \geq 0} \quad & \frac{1}{2}\|\beta\|^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \\ \text{sb.t.} \quad & \max_h \langle \beta, \phi(I_i, y_i, h)\rangle - \max_{\hat{h}} \langle \beta, \phi(I_i, \hat{y}, \hat{h})\rangle \\ & \geq \Delta(y_i, \hat{y}) - \xi_i, \quad \forall i \in \{1, \ldots, N\}, \hat{y} \in \mathcal{Y}, \end{aligned} \tag{4.1}$$

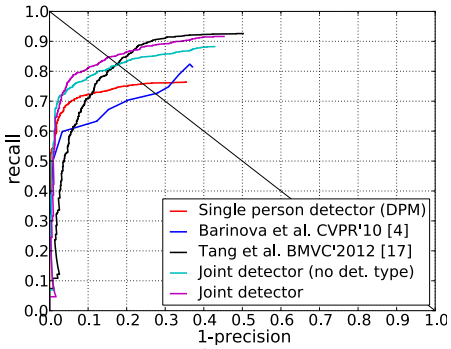


Figure 4.3: Detection performance on TUD-Crossing.

where ξ_i are slack variables modeling the margin violations. For the loss function Δ, we employ the area of the bounding box intersection $A(y_i^b \cap \hat{y}^b)$ over their union $A(y_i^b \cup \hat{y}^b)$

$$\Delta_{\text{VOC}}(y, \hat{y}) = \begin{cases} 0, & \text{if } y^l = \hat{y}^l = -1 \\ 1 - [y^l = \hat{y}^l] \frac{A(y^b \cap \hat{y}^b)}{A(y^b \cup \hat{y}^b)}, & \text{otherwise,} \end{cases} \tag{4.2}$$

as it enables precise 2D bounding box localization. The advantage of the proposed structured learning of a joint people detector is that it learns that a detection with larger overlap with the ground truth bounding box has higher score than a detection with lower overlap. Hence, the single person component should also have a lower score than the double person component on double person examples (see Fig. 4.2(a)).

4.2.3 Introducing detection type.

One limitation of the loss Δ_{VOC} for joint person detection is that it does not encourage the model enough to distinguish between a single person and a highly occluded double person pair. This is due to the large overlap of the ground truth bounding boxes, as illustrated in Fig. 4.2(b). In order to teach the model to distinguish a single person and a highly occluded person pair, we extend the structured output label with a detection type variable $y^{dt} \in \{1, 2\}$, which denotes single person or double person detection. The overall structured output is thus given as $y = (y^l, y^b, y^{dt})$. We can then additionally penalize the wrong detection type using the loss

$$\Delta_{\text{VOC+DT}}(y, \hat{y}) = (1 - \alpha)\Delta_{\text{VOC}}(y, \hat{y}) + \alpha \left[y^{dt} \neq \hat{y}^{dt} \right]. \tag{4.3}$$

4.2.4 Experimental results.

In order to fairly compare our joint detector with the joint detector proposed in Chapter 3, we explicitly train a side-view joint person detector using the same synthetic training images[4] and initialize the single and double person detector

[4] The data is available at www.d2.mpi-inf.mpg.de/datasets.

components in the same way. Fig. 4.3 shows the benefit of the proposed structured training (*Joint detector, no det. type*). By introducing the detection type loss (*Joint detector*, $\alpha = 0.5$), the joint detector further improves precision and achieves similar recall. At 95% precision it outperforms the detector proposed in Chapter 3 by 20.5% recall.

4.3 MULTI-TARGET TRACKING

Our proposed detector learning algorithm (Sec. 4.4) is generic and can, in principle, be employed in combination with any tracking-by-detection method. Here, we use a recent multi-target tracker based on continuous energy minimization (Andriyenko and Schindler, 2011). The tracker requires as input a set of person detections in a video sequence, and infers all trajectories simultaneously by minimizing a high-dimensional, continuous energy function over all trajectories. The energy consists of a data term, measuring the distance between the trajectories and the detections, and several priors that assess the (physical) plausibility of the trajectories. We use a fixed parameter setting throughout all experiments. Note that the employed tracking approach does not include any explicit occlusion handling. It is thus important to consider occlusions directly at the detector level, so as to provide more reliable information to the tracker.

Baseline results. Table 4.1 shows tracking results on the TUD-Crossing sequence (Andriluka *et al.*, 2008), using various detector variants as described above. As expected, tracking based on the output of the joint detector shows improved performance compared to the single-person DPM detector. Note that the side-view joint detector in Chapter 3 was specifically designed to handle the occlusion pattern prevalent in sequences of this type. Even so, structured learning with a detection type variable slightly increases the multi-object tracking accuracy (MOTA, Bernardin and Stiefelhagen, 2008). This experiment is meant to serve as a proof of concept and demonstrate the validity of the joint people detector. Please refer to Sec. 4.5 for an extensive experimental study on more challenging datasets.

Method	Rcll	Prcsn	MOTA	MOTP	MT	ML
single (DPM)	78.0	94.1	72.1 %	78.5 %	4	0
Joint detector (Chapter. 3)	79.9	96.5	75.6 %	79.1 %	6	0
Joint det. (no det. type)	81.9	93.2	75.1 %	79.1 %	8	0
Joint detector	82.7	93.9	76.0 %	78.6 %	7	1

Table 4.1: Tracking performance on TUD-Crossing evaluated by recall *(Rcll)*, precision *(Prcsn)* and standard CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008), including Multi-Object Tracking Accuracy *(MOTA)* and Tracking Precision *(MOTP)*. MT and ML show the number of mostly tracked and mostly lost trajectories, respectively (Wu and Nevatia, 2006).

4.4 LEARNING PEOPLE DETECTORS FOR TRACKING

So far we have shown that the proposed structured learning approach for training joint people detectors shows significant improvements for detection of occluded people in side-view street scenes. This suggests the potential of leveraging characteristic appearance patterns of person/person pairs also for detecting occluded people in more general settings. However, the generalization of this idea to crowded scenes with people walking in arbitrary directions is rather challenging due to the vast amount of possible person-person occlusion situations. This variation may arise from several factors, such as people's body articulation, or their position and orientation relative to the camera. The number of putative occlusion patterns is exponential in the number of factors. The crucial point here is, however, that not all of them are equally relevant for successful tracking. For example, short term occlusions resulting from people crossing each other's way are frequent, but can be often easily resolved by modern tracking algorithms. Therefore, finding occlusion patterns that are relevant in practice in order to reduce the modeling space is essential for applying joint person detectors for tracking in general crowded scenes.

We now propose two methods for discovering occlusion patterns for people walking in arbitrary directions by *(a)* manually designing regular occlusion combinations that appear frequently due to long-term occlusions and are, therefore, most relevant for tracking (Sec. 4.4.1); and *(b)* automatically learning a joint detector that exploits the tracking performance on occluded people and is explicitly optimized for the tracking task (Sec. 4.4.2).

4.4.1 Designing occlusion patterns

For many state-of-the-art trackers, the most important cases for improving tracking performance in crowded scenes correspond to long-term partial occlusions.

Occlusion pattern quantization. We begin by quantizing the space of possible occlusion patterns as shown in Fig. 4.4 (left). Given the position of the front person, we divide the relative position of the occluded person with respect to the occluder into 6 equal angular sectors. We consider the full half circle of the sectors behind the occluder, and do not explicitly quantize the space of possible relative distances between subjects; instead we only consider a fixed threshold, below which the second subject is significantly occluded.

In addition to quantizing the relative position, we also quantize the orientation of the front person with respect to the camera. To keep the number of constellations manageable, we use four discrete directions corresponding to four diagonal views. Independent of the orientation of the front person, the first and last sectors shown in Fig. 4.4 (left, no heavy occlusion) correspond to people walking side-by-side, slightly in front or behind each other. We found that these cases are already handled well by current person detectors. We denote the remaining four sectors as "A", "B", "C" and "D", according to the relative position of the occluded and occluding person.

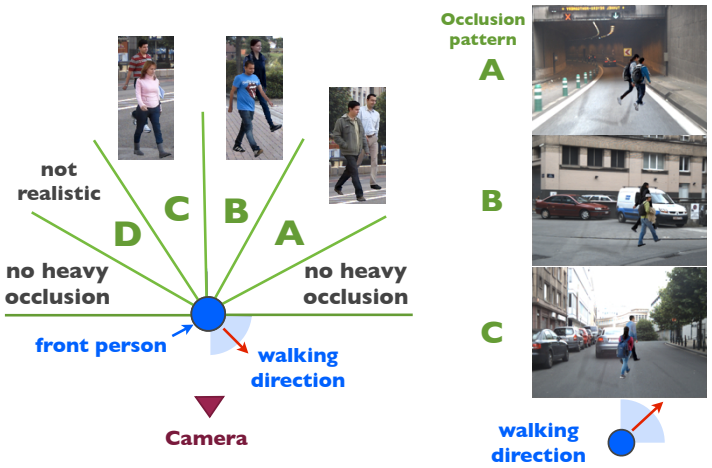


Figure 4.4: Bird's eye view of occluded person's state space (*left*). Synthetically generated training images for different occlusion patterns and walking directions (*right*).

The sector "D" corresponds to a constellation of people walking directly behind each other at close proximity. Although physically possible, this configuration is extremely unlikely in real-world scenes, because people usually tend to leave some space to the person in front when walking. We restrict ourselves to cases in which people walk in the same direction, as they cause long-term occlusions and moreover appear to have sufficient regularity in appearance, which is essential for detection performance in crowded scenes. The occlusion patterns that we consider in the rest of this analysis correspond to a combination of the four walking directions of the subjects and one of the three remaining sectors ("A", "B" or "C").

Joint detector with designed occlusion patterns. Our joint detector uses a mixture of components that capture appearance patterns of either a single person or of a person/person occlusion pair. In case of double person components, we generate two bounding boxes of people instead of one for each of the components' detections. The training procedure in Sec. 4.2 is based on the optimization of a semi-convex objective, thus susceptible to local minima. Therefore, a meaningful initialization of the detector components is important for good performance. One option is to initialize the double-person components with different degrees of occlusion as in Chapter 3. However, in the multi-view setting, the same degree of occlusion can result in very different occlusion patterns. Here, we instead initialize the components from the quantized occlusion patterns from above (Fig. 4.4, left), combining different walking directions with relative positions of the person/person pair; we construct 6 double-person components. The single-person components are initialized with different orientations, clustering appearance into 10 components, and mirroring.

Generating synthetic training examples. Training of our model requires a sufficient amount of training images. As it is very difficult and expensive to collect a representative training dataset with accurate occlusion level annotation for each image, we choose to synthetically generate training data. Most importantly, this allows

us to control the data's variation with respect to viewpoint, degree of occlusion, and variability of backgrounds, as opposed to uncontrolled clutter often present in manually collected datasets.

We collect 2400 images of people walking in 8 different walking directions to construct a synthetic training image pool. We mirror the training images to double the training set. For each captured image, we segment the person and use the segmentation to generate a number of training examples by combining the segmented person with novel backgrounds. In a similar fashion, we are able to generate training examples for different occlusion patterns and walking directions by overlaying people on top of each other in a novel image. In our experiments, we use 4000 synthetic images for training the single-person components, and up to 1200 synthetic images for the double-person components. Fig. 4.4 (right) shows several examples of our synthetically generated training images for different constellations illustrated in Fig. 4.4 (left).

Occlusion-aware NMS. We perform non-maximum suppression in two rounds: First, we consider single-person detections and the predicted occluder bounding box of double-person detections. If the occluder is suppressed by a single-person detection, then the occludee is also removed. For the second round, we allow the predicted individual bounding boxes to suppress each other, except when two bounding boxes are generated by the same double-person component.

4.4.2 Mining occlusion patterns from tracking

As we will see in Sec. 4.5 in detail, carefully analyzing and designing occlusion patterns by hand already allows to train a joint detector that generalizes to more realistic and challenging crowded street scenes. Nonetheless, the question remains which manually designed occlusion patterns are most relevant for successful tracking. Furthermore, it is still unclear whether it is reasonable to harvest difficult cases from tracking failures and explicitly guide the joint detector to concentrate on those. In the following, we describe a method to learn a joint detector specifically for tracking. We employ tracking performance evaluation, occlusion pattern mining, synthetic image generation, and detector training jointly to optimize the detector for tracking multiple targets. The approach is summarized in Alg. 1.

Input: For our study, we use the first half (frames 1–218) of the challenging PETS S2.L2 dataset (Ferryman and Shahrokni, 2009) as our mining sequence. We use the same synthetic training images to train a single-person baseline detector, as we used for training the single-component of our joint detector with manually designed occlusion patterns (see Sec. 4.4.1). Moreover, we employ a recent multi-target tracker (Andriyenko and Schindler, 2011), cf. Sec. 4.3. Note that our algorithm can optimize the joint detector for any target trackers which requires bounding box detection as inputs.

Output: A joint detector that is tailored to detect occlusion patterns that are most relevant for multi-target tracking.

Algorithm 1 Joint detector learning for tracking

Input:

Baseline detector
Multi-target tracker
Synthetic training image pool
Mining sequence

Output:

Joint detector optimized for multi-target tracking

1: run baseline detector on *mining sequence*
2: run target tracker on *mining sequence*, based on the detection result from baseline detector
3: **repeat**
4: collect *missing recall* from the tracking result
5: cluster *occlusion patterns*
6: generate *training images* for mined patterns
7: train a joint detector with *new training images*
8: run the joint detector on *mining sequence*
9: run the target tracker on *mining sequence*
10: **until** tracking results converge

Step 1 and 2 are to perform baseline detector and target tracker on the mining sequence.

Tracking evaluation (step 4): We concentrate on missed targets, which are the main source of failure in crowded scenarios. To that end, we extract all missed targets, evaluated by the standard CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008) for the next step.

Occlusion pattern mining (step 5): The majority of missed targets are occlusion related. For our mining sequence, the total number of missed targets is 1905, only 141 of them are not caused by occlusions (Fig. 4.5(a)). Missed targets can be occluders and/or occludees for a pair of persons (Fig. 4.5(b)), or within a group of multiple people (Fig. 4.5(c)). Here, we concentrate on mining occlusion patterns for pairs of persons and consider the multiple people situation as a special case of a person pair, augmented by distractions from surroundings. Note that our algorithm can be easily generalized to multiple people occlusion patterns given sufficient amount of mining sequences that contain certain distributions of multi-people occlusion patterns. From the missed targets (step 4), we determine the problematic occlusion patterns and cluster them in terms of the relative position of the occluder/occludee pair. We only consider the most dominant cluster. Fig. 4.5(d) and 4.5(e) show the dominant occlusion pattern of the first and second mining iteration. Note that we only mine occlusion patterns and no additional image information (see next step).

Synthetic training example generation (step 6): We generate synthetic training images for the mined occlusion pattern using the same synthetic image pool as in Sec. 4.4.1, which requires the relative position of a person pair, as well as the orientation of each person. To that end, we sample the relative position of a person pair from a Gaussian distribution centered on the dominant relative position cluster

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

Figure 4.5: Missed targets from PETS S2.L2 mining sequence and mined occlusion patterns: *(a)* No person nearby; *(b)* interfered by one person; *(c)* interfered by more persons; *(d)* mined occlusion pattern – 1st iteration; *(e)* mined occlusion pattern – 2nd iteration.

from step 5. We further extract a dominant orientation of the mined examples for occluders and occludees. Training image generation, in principle, thus enables us to model arbitrary occlusion patterns in each iteration. We generate 200 images for every new occlusion pattern, which amounts to the same number of training images as we used in the context of manually designed occlusion patterns. The major benefit of learning these patterns is that more training images can be easily generated for the next iteration, specifically for those relevant cases that still remain unsolved.

Joint detector training with mined occlusion patterns (step 7): The single-person component of the joint detector is initialized with the same training images as the baseline detector. For each iteration, we introduce a new double-person component that models the mined occlusion pattern. Joint training is based on the structured SVM formulation from Sec. 4.2. Learning stops when the tracking performance does not improve further on the mining sequence.

4.5 EXPERIMENTS

We evaluate the performance of the proposed joint person detector with learned occlusion patterns and its application to tracking on three publicly available and particularly challenging sequences: PETS S2.L2 and S1.L2 (Ferryman and Shahrokni, 2009), as well as the recent ParkingLot dataset (Shu *et al.*, 2012). All of them are captured in a typical surveillance setting. S2.L2 and S1.L2 show a substantial amount of person-person occlusions, in particular. We employ the first half of S2.L2 (frames 1–218) as our only *mining sequence* and use the remaining data for testing. Note that our mining algorithm only extracts occlusion patterns and no additional image information. Also note that we do not mine on any of the other sequences, and that the results on the second PETS sequence (S1.L2) and ParkingLot allow to analyze

the generalization performance of our approach to independent sequences.

To quantify the tracking performance on the test sequences, we compute recall and precision, as well as the standard CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008): Multi-Object Tracking Accuracy *(MOTA)*, which combines false alarms, missed targets and identity switches; and Multi-Object Tracking Precision *(MOTP)*, which measures the misalignment of the predicted track with respect to the ground truth trajectory.

Single-person detector. We begin our analysis with the baseline detector, which is a standard DPM single-person detector (Felzenszwalb *et al.*, 2010). For a fair comparison, we use the same synthetic training images and component initialization as for the joint detector. Note that this already yields a rather strong baseline, with far better performance than DPM-INRIA and DPM-VOC2009 (see Fig. 4.6). Tracking results using this baseline detector are also quite competitive and already outperform a state-of-the-art method (Andriyenko and Schindler, 2011) on S1.L2.

Joint detector with designed occlusion patterns (Sec. 4.4.1). Next, we evaluate the performance of our joint detector with manually designed occlusion patterns (see Fig. 4.6). The joint detector (blue) shows its advantage by outperforming the single-person detector on all sequences. It achieves 10% more recall at high precision for S1.L2 and ParkingLot. For the S2.L2 test sequence, the joint detector outperforms the baseline detector by a large margin from 0.9 precision level. These detection results suggest that the joint detection is much more powerful than the single detector; the designed occlusion patterns correspond to compact appearance and can be detected well.

The performance boost is also reflected in the tracking evaluation. Using the joint detector (Joint-Design) yields a remarkable performance boost on the S2.L2 test sequence (reaching 57.6% MOTA), improving *MOTA* by 10.1% points and *MOTP* by 1.7% points at the same time. It also improves *Recall* by 4.2 and *Precision* by 7.9 compared to the single-person detector (Single DPM). On the S1.L2 and the ParkingLot sequences, the joint detector also outperforms the single-person detector with a significantly higher recall achieved by detecting more occluded targets.

By carefully analyzing and designing the occlusion patterns, we obtain very competitive results on publicly available sequences, both in terms of detection and tracking, which shows the advantage of the proposed joint detector for tracking people in crowded scenes.

Joint detector with learned occlusion patterns (Sec. 4.4.2). We report the joint detector performance for one and two mining iterations. As mentioned above, we employ the first half of S2.L2 (frames 1–218) as mining sequence, extracting occlusion patterns, but no further image information.

On the S2.L2 test sequence (frames 219–436), which is more similar to the mining sequence than the other two sequences, our joint detector (black, Joint-Learn 1st, 56,5% MOTA) is nearly on par with the hand-designed patterns after the first iteration, as shown in Fig. 4.6(a). This is because the most dominant occlusion pattern is captured and learned by the joint detector already. For the second iteration (cyan, Joint-Learn 2nd), we also achieve higher recall on the S2.L2 test sequence,

but the precision slightly decreases because the dominant occlusion pattern of the second iteration only contains about 48 missed targets, compared to 5861 ground truth annotations, thus limiting potential performance improvement and introducing potential false positives.

Additionally, we compare our tracking results with Andriyenko and Schindler, 2011 and Breitenstein *et al.*, 2011 on the S2.L2 sequence, as shown in Tab. 4.6(a). They report tracking performance for the whole sequence, ours is for the second half of the sequence. After the second iteration of mining, we obtain a tracking performance of 56.9% MOTA, significantly outperforming the other methods.[5]

Next, we verify the generalization ability of our algorithm on two more sequences: PETS S1.L2, which is extremely crowded, and the ParkingLot sequence, which contains relatively few occlusions. On PETS S1.L2, the learned joint detector (black) is already slightly better than the Joint-Design detector after the first iteration, as shown in Fig. 4.6(b). The second iteration (cyan) once again improves the performance, both in terms of recall and precision. The tracking result is also very promising. Directly mining occlusion patterns from the tracker improves the accuracy (MOTA) with each iteration (from 21.8% over 23.4% to 26.8% MOTA). Note that, similar to the findings above, the tracking performance reaches competitive levels after only one iteration, when compared to manually designed occlusion patterns. This is remarkable, since for the S1.L2 sequence many targets are occluded for long time periods. Our mining algorithm is able to fully recover twice as many trajectories and increase the recall by over 8%.

The ParkingLot sequence contains relatively few occlusions, such that our mining algorithm cannot fully unfold its benefits, and does not improve further after the first iteration. As shown in Fig. 4.6(c), the joint detector from the first iteration outperforms all other detectors, and reaches similar performance for tracking (Fig. 4.6(c)). We also compare our method to two other state-of-the-art multi-person trackers (Shu *et al.*, 2012,Zamir *et al.*, 2012). To enable a fair comparison, we compute the performance of Zamir *et al.*, 2012 using the authors' original results and ground truth. Our joint detector yields state-of-the-art results, both w.r.t. MOTA and MOTP.

Qualitative Comparison. We first demonstrate the qualitative comparison between our joint detector and the joint detector proposed in Chapter 3 using the detector. The results are shown in Fig. 4.7. The joint detector here successfully detects occluded persons in frames 40, 190, and 200 at a high precision level. At the same time, it correctly detects single people even in the presence of background clutter (frame 130), and correctly distinguishes between detection of a single person and two people (frame 70).

We next demonstrate the qualitative comparison between the best performing single-person DPM detector and our joint detector. Fig. 4.8 shows detection results on the PETS S2.L2 sequence (frames 218–436). The results demonstrate the advantages of our joint detector for detection of people in the presence of person-person occlusions. In particular, our joint detector is able to correctly detect partially occluded people

[5]Note that, for the first half of the S2.L2 sequence where we mine the occlusion patterns, we even achieve 63.8% MOTA.

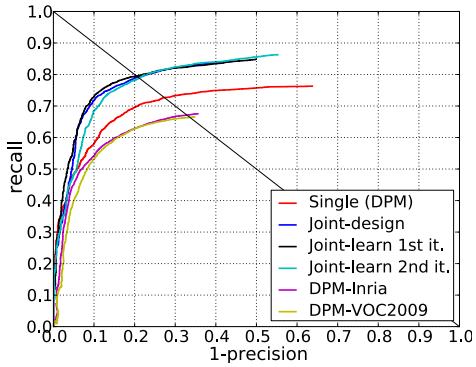
in frames 348, 378, and 436, and also avoid false positives on the occluded people in frames 228, 258, and 436. Our joint detector is able to deal with different occlusion patterns, such as people walking next to each other (frame 378), and people behind each other that are partially visible due to an elevated camera position (frames 348 and 378). Note, that in a few cases our joint detector mistakenly identifies single person detections as detections of two people introducing false positives (frames 258 and 348).

Fig. 4.9 shows detection results on the PETS S1.L2 sequence that depicts a very crowded scene with a large number of partially occluded people. Similarly to the results on PETS S2.L2, the joint detector is more robust to various occlusions and is able to detect people correctly in most of the frames. Note that occluded people are correctly detected despite occlusion by more than one person, as is shown in frames 100 and 130.

Fig. 4.10 shows detection results on the ParkingLot sequence. Although this sequence contains only a moderate amount of occluded people, the joint detectors still demonstrates a reasonable improvement in performance over the single-person detector. In particular the joint detector is able to correctly detect partially occluded people in frames 80, 140, and 160. Interestingly, our joint detector also improves over the single-person detector in detection of single people, as can be seen in frames 120, 140, and 160.

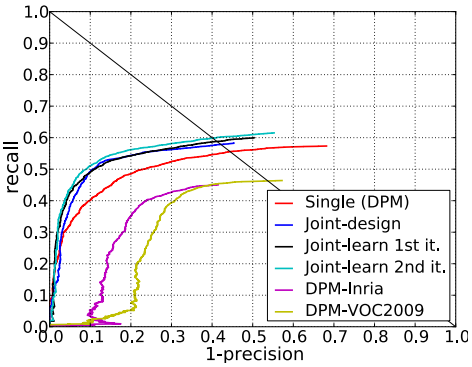
Discussion. We observed that the proposed approach converges already after two iterations; further iterations do not lead to an additional performance boost for detection or tracking. We attribute this mainly to the limited size of the mining sequence and its limited diversity. Still, the experimental results on the S1.L2 and ParkingLot sequences suggest that our detector learning algorithm is not limited to particular occlusion patterns or crowd densities. For more complex scenes such as PETS S1.L2, the performance could be further improved by utilizing a more crowded mining sequence. To that end, we plan to build a large dataset of crowded street scenes to mine a more diverse set of occlusion patterns. Another promising future extension would be to learn a joint upper-body detector on extremely dense scenes, yielding specialized upper-body occlusion patterns.

Method	Rcll	Prcsn	MOTA	MOTP
Single (DPM)	60.8	83.8	47.5 %	73.5 %
Joint-Design	65.0	91.7	57.6 %	75.2 %
Joint-Learn 1st	60.6	95.0	56.5 %	75.7 %
Joint-Learn 2nd	64.0	91.7	56.9 %	74.4 %
Andriyenko and Schindler, 2011	51.0	95.5	47.8 %	73.2 %
Breitenstein *et al.*, 2011	-	-	50.0 %	51.3 %



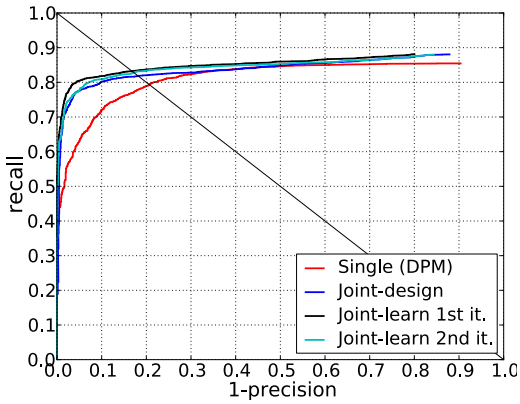
(a) PETS S2.L2 (frames 219–436).

Method	Rcll	Prcsn	MOTA	MOTP
Single (DPM)	24.8	90.1	21.8 %	70.6 %
Joint-Design	28.5	86.3	23.0 %	70.8 %
Joint-Learn 1st	28.9	86.2	23.4 %	69.8 %
Joint-Learn 2nd	32.7	86.7	26.8 %	69.3 %
Andriyenko and Schindler, 2011	24.2	83.8	19.1 %	69.6 %



(b) PETS S1.L2.

Method	Rcll	Prcsn	MOTA	MOTP
Single (DPM)	90.5	97.7	87.9 %	77.2 %
Joint-Design	91.3	97.5	88.6 %	77.6 %
Joint-Learn 1st	91.0	98.5	89.3 %	77.7 %
Joint-Learn 2nd	91.0	98.0	88.7 %	76.9 %
Shu *et al.*, 2012	81.7	91.3	79.3 %	74.1 %
Zamir *et al.*, 2012	95.0	94.2	89.1 %	77.5 %



(c) ParkingLot.

Figure 4.6: Tracking (*top*) and detection (*bottom*) performance on PETS S2.L2, S1.L2, and ParkingLot: *Single (DPM)*: single-person detector; *Joint-Design*: joint detector with designed occlusion patterns; *Joint-Learn 1st*: joint detector with learned occlusion pattern after the first mining iteration; *Joint-Learn 2nd*: joint detector with learned occlusion pattern after the second mining iteration.

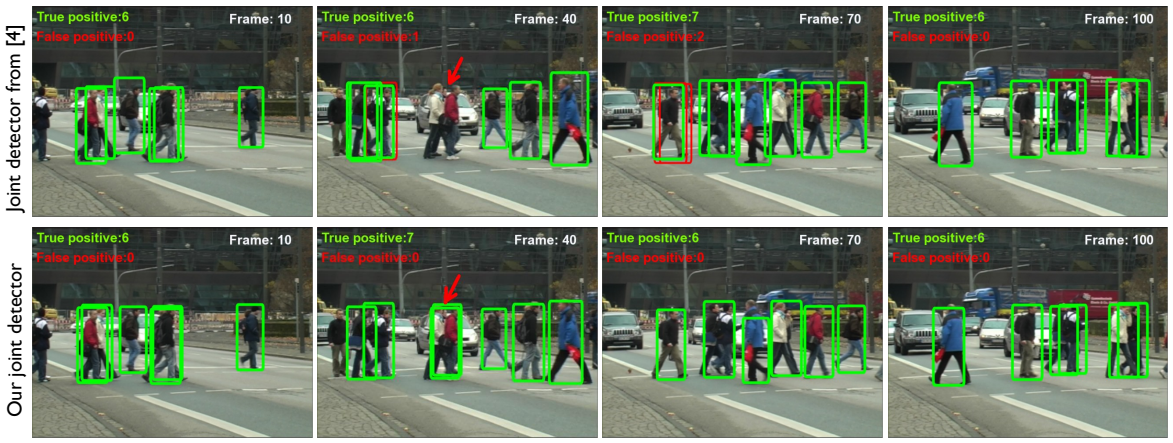


Figure 4.7: Detection results (every 30 frames) on the TUD-Crossing dataset at precision 0.95 obtained with the joint detector from Tang *et al.*, 2012 (top) and our joint detector (bottom). The true positive detections are visualized using green and the false positive detections using red color. The red arrows indicate the detections mentioned in the text.

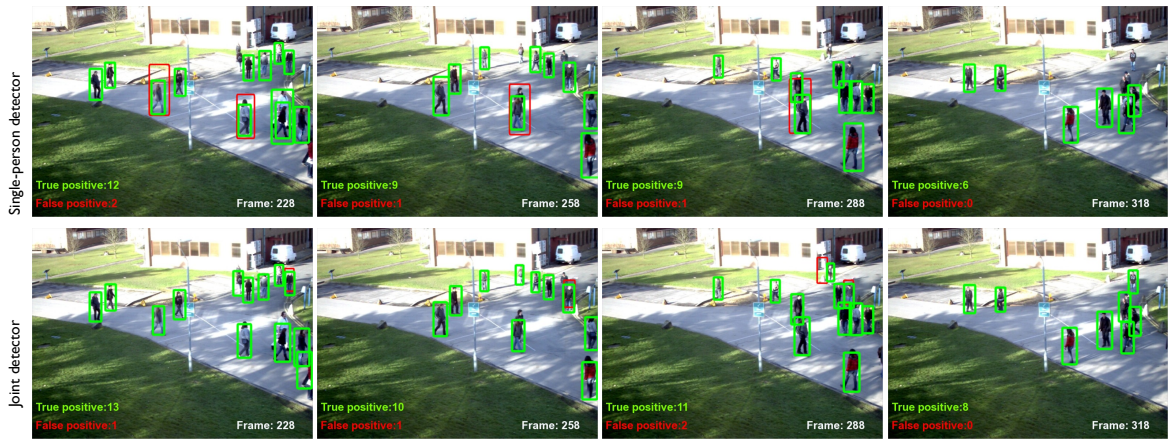


Figure 4.8: Detection results on the PETS S2.L2 (test sequence, frames 228–436) at precision 0.9 obtained with our DPM single-person detector (top) and our joint detector (bottom). The true positive detections are visualized using green and the false positive detections using red color. The red arrows indicate the detections mentioned in the text.

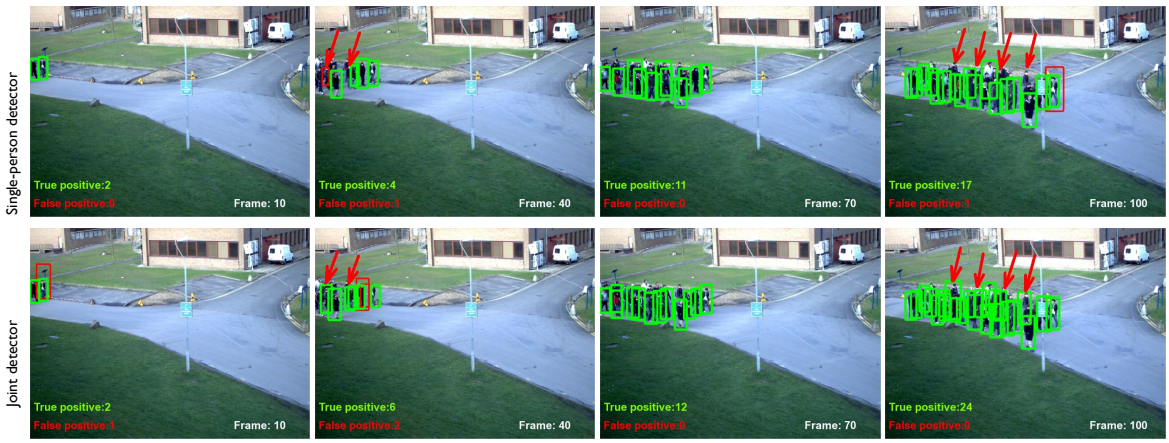


Figure 4.9: Detection results on the PETS S1.L2 dataset at precision 0.9 obtained with our DPM single-person detector (top) and our joint detector (bottom). The true positive detections are visualized using green and the false positive detections using red color. The red arrows indicate the detections mentioned in the text.

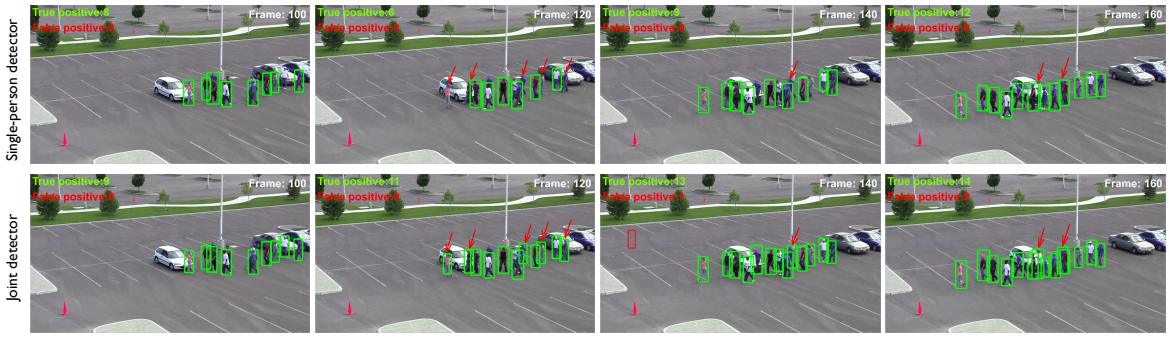


Figure 4.10: Detection results on the ParkingLot dataset at precision 0.95 obtained with our DPM single-person detector (top) and our joint detector (bottom). The true positive detections are visualized using green and the false positive detections using red color. The red arrows indicate the detections mentioned in the text.

4.6 CONCLUSIONS

In this chapter, we presented a novel joint person detector specifically designed to address common failure cases during tracking in crowded street scenes due to long-term inter-object occlusions. First, we showed that the most common occlusion patterns can be designed manually, and second, we proposed to learn reoccurring constellations with the tracker in the loop. The presented method achieves improved performance, surpassing state-of-the-art results at the time of publication of this work on several particularly challenging datasets.

5 SUBGRAPH DECOMPOSITION FOR MULTI-TARGET TRACKING

Contents

Tracking multiple targets in a video, based on a finite set of detection hypotheses, is a persistent problem in computer vision. A common strategy for tracking is to first select hypotheses spatially and then to link these over time while maintaining disjoint path constraints Pirsiavash *et al.*, 2011; Segal and Reid, 2013; Zamir *et al.*, 2012. In crowded scenes multiple hypotheses will often be similar to each other making selection of optimal links an unnecessary hard optimization problem due to the sequential treatment of space and time. Embracing this observation, we propose to link and cluster plausible detections jointly across space and time. Specifically, we state multi-target tracking as a Minimum Cost Subgraph Multicut Problem. Evidence about pairs of detection hypotheses is incorporated whether the detections are in the same frame, neighboring frames or distant frames. This facilitates long-range re-identification and within-frame clustering. Results for published benchmark sequences demonstrate the superiority of this approach.

5.1 INTRODUCTION

Multi-target tracking can be formulated as an optimization problem with respect to a graph whose nodes correspond to detection hypotheses and whose edges connect detection hypotheses that hypothetically describe the same target. A commonly employed objective of the optimization is to select a subset of nodes and edges in such a graph to maximize similarity of connected detection hypotheses, while maintaining constraints that prevent splits and merges of tracks.

By far the most common approach is to choose the initial graph such that detection hypotheses are connected only across time (not within the same time frame) and to constrain the solution such that connected components of selected detection hypotheses are paths (that do not branch). With respect to a linear objective function, this problem is a Minimum Cost Disjoint Paths Problem with respect to the initial graph. It is used, explicitly or implicitly, in many modern tracking algorithms including Pirsiavash *et al.*, 2011, Segal and Reid, 2013, Andriluka *et al.*, 2008, Zhang *et al.*, 2008.

While being intuitive, the Disjoint Paths formulation has a notable caveat: Typical target detectors yield, for each time frame, many similar (and typically equally plausible) detections of the same target. Within the Disjoint Paths formulation, it becomes necessary to choose, for each time frame and target, one best out of many similar (and plausible) hypotheses. Various recipes are proposed in the literature to address this challenge. E.g., Pirsiavash *et al.*, 2011 and Andriluka *et al.*, 2008 rely on a greedy iterative procedure that finds one track at a time and then removes corresponding hypotheses, or Zhang *et al.*, 2008 perform several rounds of optimization that merge detections into tracklets and then into full tracks. Unfortunately, all these methods depend on parameters that need to be tuned carefully, as noted in Pirsiavash *et al.*, 2011, Andriluka *et al.*, 2008, Zhang *et al.*, 2008.

Embracing the possibility of having multiple plausible hypotheses per target and frame motivates us to formulate multi-target tracking as a Minimum Cost Subgraph Multicut Problem. The feasible solutions of this formulation are such that possibly multiple hypotheses per track and time frame are selected and clustered, resulting in an overall rigorous and elegant approach to link, cluster and track targets *jointly* across space and time. To illustrate the similarities and differences to prior work we implement a version of a tracking algorithm based on the Minimum Cost Disjoint Path Problem. Although conceptually simple, its output is already on par with the state of the art for public benchmark sequences, as we show in Sec. 5.5.

This work makes the following contributions: *First*, to our knowledge, our work is the first to propose a Subgraph Multicut model for the multi-target tracking problem jointly solving the spatial *and* temporal associations of detection hypotheses. *Second*, we provide an in-depth analysis and comparison of the Subgraph Multicut and the Disjoint Paths models. Our results suggest that the Subgraph Multicut model has considerable advantages due to the fact that state-of-the-art object detectors output multiple hypotheses per target. *Third*, besides proposing an exact solver, we also provide a heuristic solution based on the Kernighan-Lin algorithm (Kernighan

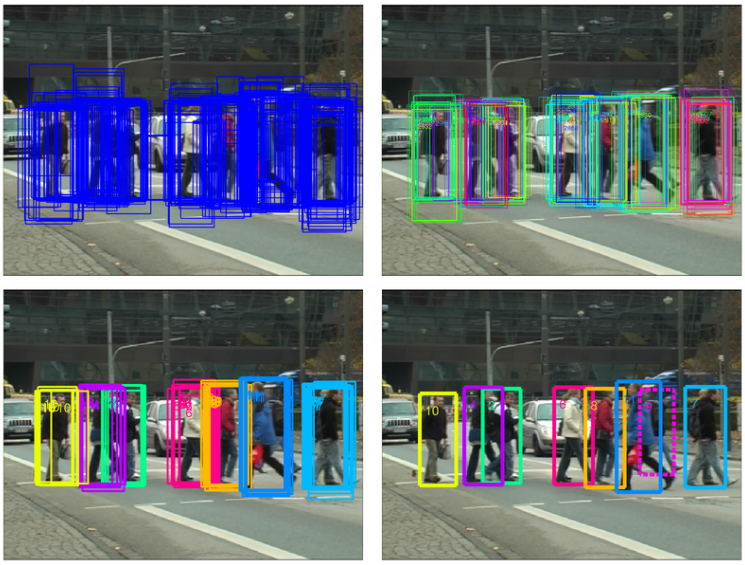


Figure 5.1: Overview of the Subgraph Multicut tracking method: (clockwise) detection hypotheses, overlapping tracklet hypotheses, hypotheses decomposition (clustering jointly across space and time) and final tracks (dotted rectangles are interpolated tracks).

and Lin, 1970), which makes the method applicable to large sequences. Finally we perform extensive experiments and present superior results compared to the state-of-the-art.

5.2 FORMULATION OF MULTI-TARGET TRACKING

Before introducing the formulations for the Subgraph Multicut Problem and the Disjoint Paths Problem, we illustrate the difference between them by visualizing a toy example in Fig. 5.2: (c) shows a solution of the Minimum Cost *Disjoint Paths* problem that finds disjoint trajectories for all targets in a directed graph; and (e) shows a solution to the Minimum Cost *Subgraph Multicut* problem that corresponds to a decomposition of an undirected graph.

5.2.1 Disjoint Paths Problem

We now summarize the formulation of multi-target tracking as a Minimum Cost Disjoint Paths Problem (Def. 1). The formulation is with respect to a *directed* graph $G = (V, E)$ whose nodes V are all hypothesized detections of an entire video and whose edges E connect pairs of detection hypotheses that hypothetically describe the same target in the different frames. More specifically, every edge $vw \in E$ points forward in time, i.e., the frame of the detection v is strictly smaller than the frame of the detection w .

The feasible solutions of the Minimum Cost Disjoint Paths Problem (Def. 1) are

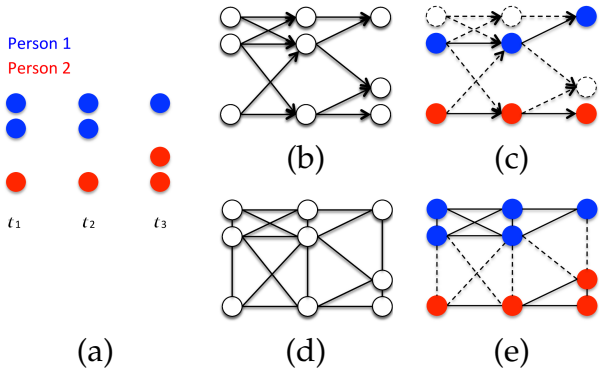


Figure 5.2: Two person detection hypotheses in three consecutive frames (ground truth assignment depicted in color) **(a)**; The disjoint paths **(c)** obtained by solving a Minimum Cost Disjoint Paths Problem with respect to a directed graph **(b)**; The decomposition **(e)** obtained by solving a Minimum Cost Subgraph Multicut Problem with respect to an undirected graph **(d)**.

subgraphs $G' = (V', E')$ of G which are encoded by $x \in \{0,1\}^V$, the characteristic function of the subset $V' = \{v \in V \mid x_v = 1\} \subseteq V$ of nodes, and $y \in \{0,1\}^E$, the characteristic function of the subset $E' = \{vw \in E \mid y_{vw} = 1\} \subseteq E$ of edges. More specifically, the subgraph G' is constrained (by Def. 1) to be a set of disjoint paths in G. The objective function is linear in the coefficients of x and y:

Definition 1. *With respect to a directed graph $G = (V, E)$, $c \in \mathbb{R}^V$ and $d \in \mathbb{R}^E$, the 01-linear program written below is called an instance of the* Minimum Cost Disjoint Paths Problem.

$$\min_{\substack{x \in \{0,1\}^V \\ y \in \{0,1\}^E}} \quad \sum_{v \in V} c_v x_v + \sum_{e \in E} d_e y_e \tag{5.1}$$

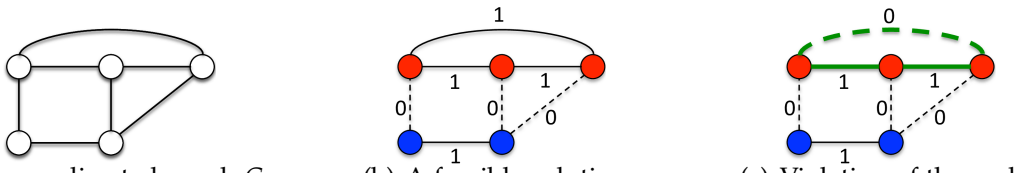
$$\text{subject to} \quad \forall e = vw \in E: \quad y_{vw} \leq x_v \tag{5.2}$$

$$\forall e = vw \in E: \quad y_{vw} \leq x_w \tag{5.3}$$

$$\forall v \in V: \quad \sum_{vw \in E} y_{vw} \leq 1 \tag{5.4}$$

$$\forall w \in V: \quad \sum_{vw \in E} y_{vw} \leq 1 \tag{5.5}$$

Here, c_v and d_e correspond to the unary and pairwise costs. The constraints (5.2) and (5.3) state that an edge can only be selected if both its nodes are selected. The constraints (5.4) and (5.5) state that every node has at most one incoming edge and at most one outgoing edge, respectively, effectively implementing the Disjoint Paths constraint.



(a) An undirected graph G (b) A feasible solution (c) Violation of the cycle constraint

Figure 5.3: **(a)** An undirected graph G; **(b)** A feasible solution of the Minimum Cost Subgraph Multicut Problem (Def. 2) on G, two connected components are in red and blue respectively, the set of edges with value 0 (dotted lines) is a multicut of the graph G; **(c)** The cycle constraint (5.9) is violated for the cycle depicted in green.

5.2.2 Subgraph Multicut Problem

We now formulate multi-target tracking as a Minimum Cost Subgraph Multicut Problem (Def. 2). The formulation is with respect to an *undirected* graph $G = (V, E)$ whose nodes V are all hypothesized detections of an entire video and whose edges E connect pairs of detection hypotheses that hypothetically describe the same target, including pairs in the same video frame.

The feasible solutions of the Minimum Cost Subgraph Multicut Problem (Def. 2) define subgraphs $G' = (V', E')$ of G which are encoded by $x \in \{0,1\}^V$, the characteristic function of the subset $V' = \{v \in V \mid x_v = 1\} \subseteq V$ of nodes, and $y \in \{0,1\}^E$, a characteristic function defining the subset $E' = \{vw \in E \mid y_{vw} = 1\} \subseteq E$ of edges. More specifically, the subgraph G' is constrained (by Def. 2) such that each connected component (V'', E'') of G' contains all edges $E'' = \binom{V''}{2} \cap E$. We show an example graph and a feasible solution in Fig. 5.3.

The objective function of the Minimum Cost Subgraph Multicut Problem is linear in the coefficients of x and y:

Definition 2. *With respect to an undirected graph $G = (V, E)$, $c \in \mathbb{R}^V$ and $d \in \mathbb{R}^E$, the 01-linear program written below is called an instance of the* Minimum Cost Subgraph Multicut Problem.

$$\min_{\substack{x \in \{0,1\}^V \\ y \in \{0,1\}^E}} \quad \sum_{v \in V} c_v x_v + \sum_{e \in E} d_e y_e \tag{5.6}$$

$$\text{subject to} \quad \forall e = vw \in E: \quad y_{vw} \leq x_v \tag{5.7}$$

$$\forall e = vw \in E: \quad y_{vw} \leq x_w \tag{5.8}$$

$$\forall C \in \text{cycles}(G)\ \forall e \in C: \quad (1 - y_e) \leq \sum_{e' \in C \setminus \{e\}} (1 - y_{e'}) \tag{5.9}$$

Here, the constraints (5.7) and (5.8) state that an edge can only be selected if both its nodes are selected. The cycle constraints (5.9) state, firstly, that every component of the selected subgraph G' is also a component of G and, secondly, that every edge of

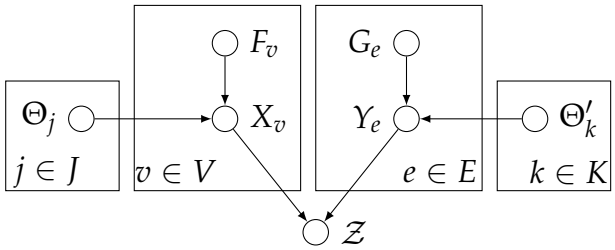


Figure 5.4: A Bayesian Network of probability measures of characteristic functions of subgraphs.

G whose nodes are in the same component of G' is also in G. An example of violation is shown in Fig. 5.3(c). In the context of multi-target tracking this implies that if a detection hypothesis is connected (spatially or temporally) to another detection hypothesis, all neighbors of the first hypothesis have to be connected to all spatial and temporal neighbors of the second hypothesis as well.

5.2.3 Probabilistic Model

Toward the goal of learning and inferring the parameters c and d of both optimization problems (Def. 1 and 2) from video data and toward the goal of comparing the two formulations of the multi-target tracking problem, we now define a probability measure on subgraphs of a graph $G = (V, E)$ such that a maximally probable set of disjoint paths is precisely a solution of the Minimum Cost Disjoint Path Problem (Def. 1) and such that a maximally probable subgraph multicut is precisely a solution of the Minimum Cost Subgraph Multicut Problem (Def. 2).

More specifically, we define a probability measure on the characteristic functions $x \in \{0,1\}^V$ and $y \in \{0,1\}^E$ with respect to the Bayesian Network depicted in Fig. 5.4. Realizations of the random variables X and Y are the characteristic functions x and y. For a finite index set J and every $v \in V$, a realization of the random variable F_v is a vector $f^v \in \mathbb{R}^J$ of features of the node v. For a finite index set K and every $e \in E$, a realization of the random variable G_e is a vector $g^e \in \mathbb{R}^K$ of features of the edge e. A realization of the random variable Θ (Θ') is a vector $\theta \in \mathbb{R}^J$ ($\theta' \in \mathbb{R}^K$) of model parameters. Finally, a realization of the random variable $\mathcal{Z}$ is a set $Z \subseteq \{0,1\}^{V \cup E}$ of feasible characteristic functions.

From the conditional independencies enforced by the Bayesian Network (Fig. 5.4) follows that a probability measure of the conditional probability of characteristic functions x of nodes and y of edges and model parameters θ and θ', given features f and g and given a feasible set Z, factorizes according to

$$\begin{aligned} & p(x, y, \theta, \theta' | f, g, Z) \\ \propto \; & p(Z|x, y) \cdot \prod_{v \in V} p(x_v | f^v, \theta) \cdot \prod_{j \in J} p(\theta_j) \\ & \cdot \prod_{e \in E} p(y_e | g^e, \theta') \cdot \prod_{k \in K} p(\theta'_k) \; . \end{aligned} \tag{5.10}$$

In order to constrain the characteristic functions x and y jointly to the feasible set Z, the first term, the probability density of a feasible set Z, given x and y, is defined to be 0 if $(x, y) \notin Z$; It is defined to be positive and constant, otherwise:

$$p(Z|x, y) \propto \begin{cases} 1 & \text{if } (x, y) \in Z \\ 0 & \text{otherwise} \end{cases} . \tag{5.11}$$

The second and third term in Eq. 5.10 are a probabilistic model for the independent 01-classification of nodes (detections). The fourth and fifth term are a probabilistic model for the independent 01-classification of edges (pairs of detections). Specifically, we consider a linear logistic model and a Gaussian prior with $\sigma \in \mathbb{R}^+$. These are stated below for nodes. The definition for edges is independent and analogous.

$$p(x_v = 1|f^v, \theta) = \frac{1}{1 + \exp(-\langle \theta, f^v \rangle)} \tag{5.12}$$

$$p(\theta_j) = \mathcal{N}(0, \sigma^2) \tag{5.13}$$

Estimation (Learning and Inference). Estimating maximally probable model parameters θ, θ' from training data x, y, f, g requires the solution of two (convex) logistic regression problems, one for nodes and one for edges.

Estimating maximally probable characteristic functions x and y for previously unseen data f, g, given a feasible set Z and given (learned) model parameters θ, θ' amounts to solving the 01-linear problem stated in Lemma 5.2.1. This problem specializes to the problems in Definitions 1 and 2 for the respective feasible sets Z and motivates our choice of the parameters c and d.

Lemma 5.2.1. *Given a graph $G = (V, E)$, a feasible set Z, feature vectors f, g, and model parameters θ, θ', all as defined above with respect to G, a pair (x, y) with $x \in \{0,1\}^V$ and $y \in \{0,1\}^E$ is maximally probable with respect to the measure defined above if and only if it is a solution of the 01-linear program written below, with $c_v = -\langle \theta, f^v \rangle$ and $d_e = -\langle \theta', g^e \rangle$.*

$$\min_{\substack{x \in \{0,1\}^V \\ y \in \{0,1\}^E}} \quad \sum_{v \in V} c_v x_v + \sum_{e \in E} d_e y_e \tag{5.14}$$

$$\text{subject to} \quad (x, y) \in Z \tag{5.15}$$

Proof. Let $G = (V, E)$ be a graph and $Z \subseteq \{0,1\}^{V \cup E}$. For every $v \in V$, let $f^v \in \mathbb{R}^J$. For every $e \in E$, let $g^e \in \mathbb{R}^K$. Finally, let $\theta \in \mathbb{R}^J$ and $\theta' \in \mathbb{R}^K$. Moreover, recall from (10)–(13) in the main text the definition of the probability measure

$$\begin{aligned} & p(x, y, \theta, \theta'|f, g, Z) \\ \propto \quad & p(Z|x, y) \cdot \prod_{v \in V} p(x_v|f^v, \theta) \cdot \prod_{j \in J} p(\theta_j) \\ & \cdot \prod_{e \in E} p(y_e|g^e, \theta') \cdot \prod_{k \in K} p(\theta'_k) \end{aligned} \tag{5.16}$$

with

$$p(Z|x,y) \propto \begin{cases} 1 & \text{if } (x,y) \in Z \\ 0 & \text{otherwise} \end{cases} \tag{5.17}$$

$$p(x_v = 1|f^v,\theta) = \frac{1}{1+\exp(-\langle\theta, f^v\rangle)} \tag{5.18}$$

$$p(y_e = 1|g^e,\theta') = \frac{1}{1+\exp(-\langle\theta', g^e\rangle)} \tag{5.19}$$

$$p(\theta_j) = \mathcal{N}(0,\sigma^2) \tag{5.20}$$

$$p(\theta'_k) = \mathcal{N}(0,\rho^2)\ . \tag{5.21}$$

Although $p(Z|x,y)$ can be zero, the probability measure is well-defined for any $Z \neq \emptyset$ because $p(x_v|f^v,\theta) > 0$, by (5.18), and $p(y_e|g^e,\theta') > 0$, by (5.19).

It is such that

$$\begin{aligned} & p(x,y|\theta,\theta',f,g,Z) \\ = \ & p(Z|x,y) \prod_{v\in V} p(x_v|f^v,\theta) \prod_{e\in E} p(y_e|g^e,\theta') \end{aligned} \tag{5.22}$$

because $p(\theta_j) > 0$, by (5.20), and $p(\theta'_k) > 0$, by (5.21), and by conditioning on θ and θ'.

Moreover, it is such that

$$\begin{aligned} & \underset{\substack{x\in\{0,1\}^V \\ y\in\{0,1\}^E}}{\operatorname{argmax}} \ p(x,y|\theta,\theta',f,g,Z) \\ = \ & \underset{(x,y)\in Z}{\operatorname{argmax}} \ \prod_{v\in V} p(x_v|f^v,\theta) \prod_{e\in E} p(y_e|g^e,\theta') && (5.23) \\ = \ & \underset{(x,y)\in Z}{\operatorname{argmax}} \ \sum_{v\in V} \log p(x_v|f^v,\theta) + \sum_{e\in E} \log p(y_e|g^e,\theta') && (5.24) \\ = \ & \underset{(x,y)\in Z}{\operatorname{argmax}} \ \sum_{v\in V} \langle\theta, f^v\rangle x_v + \sum_{e\in E} \langle\theta', g^e\rangle y_e\ . && (5.25) \end{aligned}$$

In this statement that concludes the proof, (5.23) holds by (5.22) and (5.17). (5.24) follows by the strict monotonicity of the logarithmic function, and (5.25) follows by the arithmetic transformations stated below for $p(x_v|f^v,\theta)$. The transformations for $p(y_e|g^e,\theta')$ are analogous.

$$\begin{aligned} & \log p(x_v|f^v,\theta) \\ = \ & x_v \log p(x_v = 1|f^v,\theta) + (1-x_v) \log p(x_v = 0|f^v,\theta) \\ = \ & x_v \log \frac{p(x_v = 1|f^v,\theta)}{p(x_v = 0|f^v,\theta)} + \log p(x_v = 0|f^v,\theta) \\ = \ & x_v \langle f^v,\theta\rangle - \log(1+\exp(-\langle f^v,\theta\rangle)) \end{aligned} \tag{5.26}$$

□

Here, (5.26) follows by (5.18). Note that the second term in (5.26) does not depend on x_v and can hence be dropped from the objective function (5.25).

Certified Optimal Solutions. The Minimum Cost Disjoint Paths Problem has polynomial time complexity. We solve instances of this problem by Branch-and-Cut. The Minimum Cost Subgraph Multicut Problem is NP-hard (Bansal *et al.*, 2004) and APX-hard (Demaine *et al.*, 2006). In order to solve instances of this problem exactly, we make use of the Branch-and-Cut loop of the closed-source commercial software Gurobi[6] which represents the state of the art in integer linear programming (ILP).

In every iteration of an outer cutting plane loop, we consider a relaxed ILP with the full objective function and a subset of the cycle inequalities (none in the first iteration). In order to solve this relaxed ILP to optimality, in an inner loop, we resort to the general classes of branches and cuts implemented in Gurobi. Once a solution of the relaxed ILP is found, we separate violated cycle inequalities, by breadth-first-search, and add these to the relaxed ILP, thus tightening the relaxation. The procedure stops when all cycle inequalities are satisfied and, thus, the full problem has been solved to optimality.

Heuristic Solutions. Alternatively, we propose a heuristic solution for the unconstrained set partition problem by making use of the Kernighan Lin (KL) algorithm as defined in Kernighan and Lin, 1970, which uses the KL for the bi-partition problem, also defined in Kernighan and Lin, 1970, as a subroutine. The procedure starts from an initial decomposition defined, in our case, by the components of the graph containing precisely the edges $e \in E$ for which $d_e > 0$. In every iteration, an attempt is made to strictly improve the current decomposition via a sequence of transformations: In an outer loop, every pair of adjacent components is considered. For any such pair, it is assessed, in an inner loop, whether moving nodes from one set to the other improves the objective value. In every iteration of this inner loop, an optimal move is chosen and saved, together with the difference of the objective value caused by this move. Having ordered all possible moves in this way, the smallest k is chosen such that the first k moves, carried out in order, improve the objective value maximally. If the improvement is positive, the moves are made and thus, the current decomposition is improved. If the improvement is not positive, the procedure terminates.

5.3 TRACKING DETAILS

In this section, we describe our tracklet hypotheses generation method in Sec. 5.3.1, definitions of the unary feature f and the pairwise feature g in Sec. 5.3.2 and further implementation details about the Disjoint Paths and Subgraph Multicut tracking model in Sec. 5.3.3.

[6] Version 6.0, http://www.gurobi.com

5.3.1 Tracklet Generation

We start with person detections produced by the Deformable Part Model (DPM) (Felzenszwalb *et al.*, 2010). Instead of using the detections as person hypotheses directly, we generate overlapping tracklet hypotheses by the method proposed in Andriluka *et al.*, 2008. Let the length of a tracklet be M, the set of all detections in frame t is denoted by $h^t = [h_1^t, \ldots, h_{nt}^t]$. Then a tracklet $H = [h_{t1}^1, \ldots, h_{tM}^M]$ is optimal given all the detections in M frames if H maximizes the following probability:

$$p(H) = p(h_{t1}^1) \cdot \prod_{k=2}^{M} p(h_{tk}^k) \cdot p(h_{tk}^k, h_{tk-1}^{k-1}). \tag{5.27}$$

where $p(h_{tk}^k)$ denotes the probability of detection h_{tk} being true, and $p(h_{tk}^k, h_{tk-1}^{k-1})$ is the transition probability which models a simple Gaussian position dynamics. In our implementation, $M = 5$ for sequences which are shorter than 300 frames and $M = 9$ for others due to the computation cost.

Overlapping Tracklets. For all the detections in every M consecutive frames, we apply the Viterbi algorithm to maximize Eq. (5.27) to obtain the optimal sequence of detections - our tracklet hypotheses. We remove the selected detections from the set of detections and maximize Eq. (5.27) iteratively until all the detections are considered. Our tracklets are obtained in an over-complete fashion in two aspects (1) Non-Maximum Suppression (NMS) is not applied for the detections and (2) we compute overlapping tracklets starting at every frame of the sequence. Each strong detection contributes M times to different tracklets (which have different starting frames). Our overlapping tracklets contain a sufficient number of good ones which is arguably a good basis for a tracking algorithm.

5.3.2 Unary and Pairwise Features

Each tracklet contains the following information: spatial-temporal location, speed, scale, appearance and confidence (tracklet score). Here, with respect to the detection in the middle frame of a tracklet, we use x and y to denote the tracklet center; t is the frame index; v_x and v_y is the velocity fo the tracklet along x and y coordinate respectively; h and a denotes the scale and appearance of the tracklet; s is the tracklet score. Given two tracklets $(x, y, t, v_x, v_y, h, a, s)$ and $(x', y', t', v_x', v_y', h', a', s')$, the unary feature is simply the tracklet score and we define the following auxiliary variables for the pairwise feature:

$$\begin{aligned} m_1 &= x' - x \quad m_2 = v_x(t' - t) \quad m_3 = v_x'(t' - t) \\ n_1 &= y' - y \quad n_2 = v_y(t' - t) \quad n_3 = v_y'(t' - t) \end{aligned}$$

which are all further normalized by $\bar{h}$ where $\bar{h} = \max(h, h')$. The pairwise features are defined as

$$\begin{aligned} g_1 &= |t - t'| & g_4 &= |m_1 - m_2| & g_7 &= |n_1 - n_2| \\ g_2 &= \frac{|h - h'|}{\bar{h}} & g_5 &= |m_2 - m_3| & g_8 &= |n_2 - n_3| \\ g_3 &= D(a, a') & g_6 &= |m_1 - m_3| & g_9 &= |n_1 - n_3| \end{aligned} \tag{5.28}$$

where g_1 denotes temporal distance between two tracklets, g_2 is the normalized scale difference, $g_4 \dots g_9$ describe the relations between speed and temporal-spatial locations of two tracklets, g_3 is the euclidean distance between two tracklets' dColorSIFT features proposed in Zhao *et al.*, 2013.

We introduce a non-linear mapping from the feature space to the cost space by extending our unary and pairwise features to quadratic and exponential terms. Unary feature f^v is extended as $(f_1, f_1^2, e^{(-f_1)})$ and pairwise feature g^e is $(g_1, \dots, g_9, g_1^2 \dots g_9^2, e^{(-g_1)} \dots e^{(-g_9)})$.

5.3.3 Further Details

NMS for the Disjoint Paths Model. The above technical details are identical for the Subgraph Multicut model and the Disjoint Paths model. However, pre-selection of tracklet hypotheses (tracklet NMS) and post-processing of the final tracks (tracks NMS) are necessary steps for the Disjoint Paths model. In our implementation, these two steps are performed in a standard way: the tracklet NMS is performed in full analogy to a greedy NMS for people detection, with respect to the middle frame of the tracklet. For the NMS of the final tracks, the suppression is performed on the overlapping fragment of each track, which means that if the optimal track of a target is obtained, it suppresses all other suboptimal redundant tracks of the target. The extensive evaluation described in Sec. 5.5.3 shows that our Disjoint Paths model with the standard NMS technics achieves results which are on par with state-of-the-art, indicating that our Disjoint Paths model is a good baseline to conduct valid analyses and comparisons.

Tracks from the Subgraph Multicut Model. While the Disjoint Paths model directly produces tracks for each target by its definition, our Subgraph Multicut model produces a connected component for each target. Generating tracks from connected components is straight-forward: in each frame, for all the hypotheses which belong to the same component, we obtain representative location x,y and scale s in this frame by averaging all the connected hypotheses weighted by their probability defined in Eq.5.12. The final track of the target is a smoothed trajectory which links the representative hypotheses across all the frames.

5.4 SUBGRAPH MULTICUT FOR DETECTION NMS

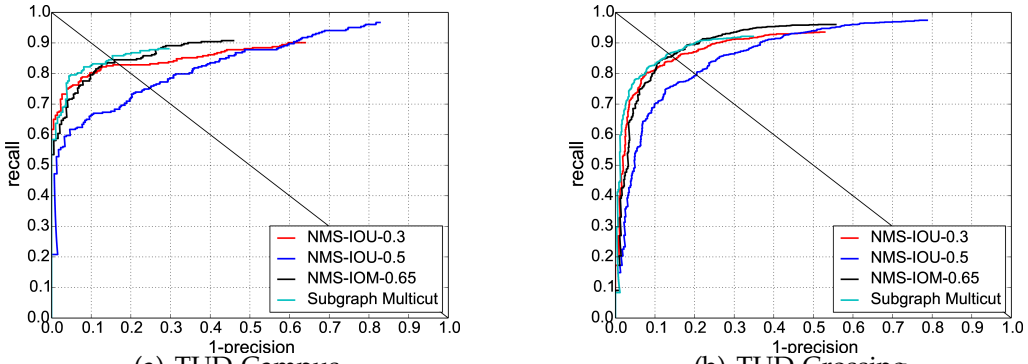
Our Subgraph Multicut model has the property of jointly addressing the problem of spatial (within-frame) *and* temporal (across-frame) associations. Non-Maximum Suppression (NMS) for detections in single frames, on the other hand, is a spatial association problem. Therefore, it is straight-forward to apply the Subgraph Multicut model to the NMS problem.

In full analogy to our Subgraph Multicut tracking model, for detection hypotheses, the unary feature is the detection score, the pairwise feature is derived from Eq.5.28. Given that we have $|t' - t| = 0$, the pairwise feature g is defined as $(|h - h'|, |m_1|, |n_1|)$. The final representation of each target is obtained by weighted averaging of all the detections which are associated together.

Results. We evaluate the Subgraph Multicut NMS method on the TUD-Campus and TUD-Crossing datasets (Andriluka *et al.*, 2008), which are challenging for pedestrian detection due to partial occlusions. Given the detections obtained from DPM (Felzenszwalb *et al.*, 2010), two state-of-the-art NMS methods are used as baselines. *(1)* NMS intersection over union (NMS-IOU) (Girshick *et al.*, 2014b) and *(2)* NMS intersection over minimum (NMS-IOM) (Dollár *et al.*, 2009).

In Fig. 5.5(a), NMS-IOU with threshold 0.3 gets better precision and NMS-IOU with threshold 0.5 obtains higher recall. For NMS-IOM, we use threshold 0.65 which is the best setting for this method (Dollár *et al.*, 2009). Our Subgraph Multicut model is able to improve the performance comparing to all the NMS methods evaluated here. In Fig. 5.5(b), our Subgraph Multicut model is on par with NMS-IOM at equal-error-rate, and outperforms others at high precision. The parameters used in the Subgraph Multicut NMS model for TUD-Crossing are learned from TUD-Campus and vice versa.

Summary. Only spatial relations between two detections are considered in the current pairwise feature, which is a fair comparison between our Subgraph Multicut model and local greedy NMS methods. Our model performs better because *(1)* associations of detections are obtained in a globally optimal fashion and *(2)* different spatial relations between two detections are learned for associations. Note that, our Subgraph Multicut model has the potential of leveraging other information in the pairwise term, e.g., appearance and prior knowledge about object layout.



(a) TUD-Campus (b) TUD-Crossing

Figure 5.5: Performance comparison between the Subgraph Multicut model and local greedy NMS methods.

5.5 TRACKING EVALUATION

We evaluate the performance of the proposed Subgraph Multicut model on three publicly available sequences: TUD-Campus, TUD-Crossing (Andriluka *et al.*, 2008) and ParkingLot (Zamir *et al.*, 2012). We perform extensive experiments and analysis on TUD-Crossing and present quantitative, superior results compared to other competitive methods on three sequences.

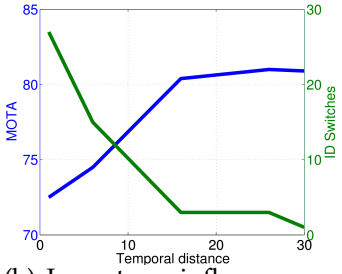
We use standard CLEAR MOT as evaluation metrics that include recall (Rcll), precision (Prcsn), multiple object tracking accuracy (MOTA), and multiple object tracking precision (MOTP) (Bernardin and Stiefelhagen, 2008). MOTA is a cumulative measure that combines missed targets (FN), false alarms (FP), and identity switches (IDs). MOTP measures overlap between the ground truth and estimated trajectory. We also report mostly tracked (MT), partly tracked (PT), mostly lost (ML) and fragmentation (FM) for measuring track completeness.

We analyze the performance of the proposed methods in four aspects. *(1)* We compare the exact integer linear programming (ILP) solver and the heuristic Kernighan Lin (KL) solver in terms of run time and MOTA. For the same tracklet hypotheses, KL obtains nearly the same MOTA compared to ILP, but much faster (Sec. 5.5.1). *(2)* We evaluate the influence of long-term associations both for the Disjoint Paths model and the Subgraph Multicut model. By associating tracklet hypotheses that are temporally far from each other (up to 30 frames), MOTA is improved for both models and the number of ID switches is substantially reduced (Sec. 5.5.2). *(3)* We provide an in-depth analysis of the Disjoint Paths model and the Subgraph Mutlicut model. Extensive experimental results indicate that the properties of leveraging multiple hypotheses per target within and across frames facilitate the Subgraph MultiCut model to obtain a more robust association (Sec. 5.5.3). *(4)* We show that our Subgraph Multicut model obtains superior results over the state-of-the-art (Sec. 5.5.4).

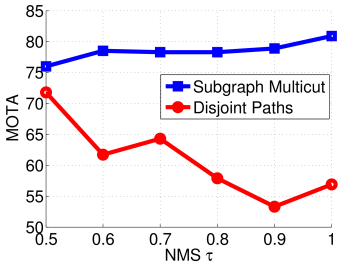
Training sequences. For the Subgraph Multicut and Disjoint Paths models, we need training data to learn the model parameters θ and θ' (Sec 5.2.3). In our

		KL solver		ILP solver	
$\|V\|$	$\|E\|$	Run time (s)	MOTA	Run time (s)	MOTA
277	4835	0.86	79.4	0.48	79.4
616	35424	1.82	80.8	76.39	83.3
1453	199333	12.49	83.3	79986.01	83.3

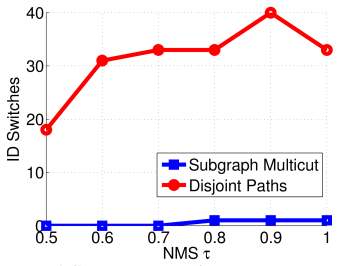
(a) Solver comparison



(b) Long-term influence



(c) MOTA



(d) ID switches

Figure 5.6: **(a)** Comparison of tracking performance and convergence speed of KL and ILP solvers on TUD-Campus; **(b)** Long-term association for the Subgraph Multicut model on TUD-Crossing; MOTA **(c)** and ID switches **(d)** comparison for the Subgraph Multicut model and the Disjoint Paths model on TUD-Crossing.

experiments, we use the parameters learned from TUD-Crossing for the experiments on TUD-Campus and ParkingLot. For TUD-Crossing, we use the parameters learned on TUD-Campus.

5.5.1 Solver Comparison

We start by comparing the performance of the Subgraph Multicut model optimized by the KL and ILP solvers on TUD-Campus. In this experiment we vary the number of initial person hypotheses $|V|$ by adjusting the threshold τ of NMS and report tracking performance and convergence speed of each solver. Results are shown in Tab. 5.6(a).

Setting τ to 0.5 results in 277 tracklet hypotheses and 4835 pairwise terms. Both solvers achieve the same MOTA (79.4%) within comparable runtime (0.86 sec. vs. 0.48 sec.). Increasing τ to 0.7 results in 616 tracklet hypotheses. In this regime ILP achieves better MOTA, but is 40 times slower than KL. Omitting NMS further increases the number of tracklet hypotheses to 1453. KL achieves the same MOTA as ILP in 12.5 seconds, compared to ILP that takes 22 hours. These results indicate that the KL algorithm achieves results comparable to ILP but significantly faster. For efficiency, we apply the KL solver for the Subgraph Multicut Problem in the following experiments. Note that reducing amount of NMS leads to improved performance, likely because NMS makes local decisions on the level of individual frames that are potentially suboptimal in the context of global optimization.

Method	Rcll	Prcsn	FAR	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	MOTAL
Pirsiavash *et al.*, 2011	66.6	95.5	0.15	8	3	4	1	11	118	10	13	60.6	78.2	63.2
Breitenstein *et al.*, 2011	-	-	-	-	-	-	-	-	-	2	-	73.3	67.0	-
Segal and Reid, 2013	-	-	-	-	5	-	-	-	-	0	3	82.0	74.0	-
Subgraph Multicut	83.8	99.3	0.03	8	5	2	1	2	58	0	1	83.3	76.9	83.3

Table 5.1: Tracking performance on TUD-Campus.

Method	Rcll	Prcsn	FAR	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	MOTAL
Pirsiavash *et al.*, 2011	73.9	95.0	0.21	13	6	7	0	43	286	50	42	65.5	76.8	69.9
Breitenstein *et al.*, 2011	-	-	-	-	-	-	-	-	-	2	-	84.3	71.0	-
Segal and Reid, 2013	-	-	-	-	7	-	-	-	-	2	12	74.0	76.0	-
Tang *et al.*, 2013	82.7	93.9	-	-	7	-	1	-	-	-	-	76.0	78.6	-
Zamir *et al.*, 2012	88.4	96.2	0.19	13	9	4	0	38	128	2	5	84.8	74.5	84.9
Disjoint Paths	74.5	98.6	0.06	13	6	7	0	12	281	18	18	71.8	77.7	73.3
Subgraph Multicut	82.0	98.8	0.05	13	8	3	2	11	198	1	1	80.9	78.0	81.0

Table 5.2: Tracking performance on TUD-Crossing.

Method	Rcll	Prcsn	FAR	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	MOTAL
Pirsiavash *et al.*, 2011	69.4	97.8	0.16	14	2	1	-	39	754	52	60	65.7	75.3	-
Shu *et al.*, 2012	-	-	-	-	-	-	-	-	-	-	-	79.3	74.1	-
Wen *et al.*, 2014	90.8	98.4	0.16	14	11	-	0	39	227	21	23	88.4	81.9	-
Tang *et al.*, 2013	91.0	98.5	-	-	-	-	-	-	-	-	-	89.3	77.7	-
Zamir *et al.*, 2012	85.3	98.2	0.02	14	-	-	-	-	-	-	-	90.4	74.1	-
Disjoint Paths	89.0	98.5	0.14	14	11	3	0	34	272	25	24	86.6	76.7	87.5
Subgraph MultiCut	96.1	95.4	0.45	14	13	1	0	113	95	5	18	91.4	77.4	91.5
Subgraph MultiCut[7]	96.9	97.0	0.37	14	13	1	0	46	47	1	6	93.8	78.3	93.9

Table 5.3: Tracking performance on ParkingLot.

5.5.2 Long-Term Association

Next, we evaluate the robustness of the Disjoint Paths and Subgraph Multicut models with respect to long-term associations between hypotheses. To that end we apply both models to graphs that connect each tracklet hypothesis to every other tracklet hypothesis within a neighborhood of 30 frames. Intuitively enabling such long-range connectivity should be helpful for misdetection and occlusion cases that otherwise result in ID switches. We conduct this experiment on TUD-Crossing that has a large number of people that frequently occlude each other.

The baseline model in this comparison corresponds to a graph in which each hypothesis is connected to hypotheses in the next and previous frames only. This baseline model for the Disjoint Paths formulation results in 66.8% MOTA. Adding long-range connections improves the performance of the Disjoint Paths model to 71.8% and reduces the number of ID switches from 34 to 18. The results for the Subgraph Multicut model are shown in Fig. 5.6(b). The performance improves from

72.5% to 80.9% MOTA, and the number of ID switches is reduced from 27 to 1. This result indicates the importance of long-term associations across frames which the Subgraph Multicut model can leverage.

5.5.3 Subgraph Multicut vs. Disjoint Paths Models

The Disjoint Paths model achieves results on par with the state-of-the-art, as shown in Tab. 5.2 and Tab. 5.3 (71.8% MOTA for TUD-Crossing, 86.6% MOTA for ParkingLot). This suggests that the Disjoint Paths model is a good baseline to conduct a detailed analysis. Note that both models are based on the same set of tracklet hypotheses as well as unary and pairwise terms as detailed above.

An important difference between the Disjoint Paths and Subgraph Multicut models is that the Disjoint path model imposes mutual exclusion constraints when connecting tracklet hypotheses. This is in contrast to the Subgraph Multicut model that allows each tracklet hypothesis to associate with an appropriate number of tracklet hypotheses in the same and other frames resulting in more robust associations.

When the tracklet hypotheses are pre-selected by performing NMS, as shown in Fig 5.6(c), with $\tau = 0.5$, the Disjoint Paths model performs best. However, the model is sensitive to the NMS threshold. Decreasing the level of NMS or skipping the NMS step altogether results in a substantial performance drop for MOTA (from 71.8% to 56.9%). Additionally, the number of ID switches increases from 18 to 33 (red line in Fig 5.6(d)). This is an inherent limitation of the Disjoint Paths model resulting from the mutual exclusion constraints. This and similar models require both pre-processing of person hypotheses (detection/tracklets-NMS) as well as post-processing of tracks (tracks-NMS) to obtain good performance.

In contrast, decreasing the level of NMS improves the performance of the Subgraph Multicut model constantly from 76.0% MOTA to 80.9% (blue curve in Fig. 5.6(c)). This is due to the ability of the Subgraph Multicut model to associate hypotheses jointly across space and time, thereby aggregating information about the targets which results in more robust associations over the whole sequence.

With respect to ID switches, the Subgraph Multicut model constantly outperforms the Disjoint Paths model for all NMS thresholds by a large margin as shown in Fig. 5.6(d). This performance difference is explained by the fact that finding a disjoint path for a target precisely in a graph across all frames is a substantially harder problem than clustering nodes that correspond to the same target.

5.5.4 Comparison to the State-of-the-art

We now compare our approach to recent approaches on TUD-Crossing, TUD-Campus, and "Parking Lot" datasets. TUD-Campus and TUD-Crossing show people from the camera at low viewpoint resulting in frequent occlusions, and TUD-Campus also includes substantial variation in people scale. The Parking Lot

sequence is captured in a surveillance setting with a camera elevated above the ground that results in pedestrians' walking patterns substantially different compared to TUD-Campus and TUD-Crossing. Tables 5.1, 5.2 and 5.3 show results for TUD-Campus, TUD-Crossing and "Parking Lot" respectively. The ground truth tracks used in all experiments are from Andriyenko *et al.*, 2012. Our Subgraph Multicut model achieves state-of-the-art MOTA results overall. In particular the number of ID switches is substantially improved compared to other approaches. Zamir *et al.*, 2012 also report tracking results on TUD-Crossing. Based on the tracking results they provided to us, we obtain 84.8% MOTA and 2 ID Switches on the ground truth from Andriyenko *et al.*, 2012.

On the ParkingLot sequence, the Disjoint Paths model again performs on par with state-of-the-art (86.6% MOTA), suggesting that it is a good baseline to conduct comparison and analysis. With our Subgraph Multicut Model and parameters learned from TUD-Crossing, we achieve 91.4% MOTA and 5 ID Switches. To evaluate sensitivity of our model to particular training set we split the "Parking Lot" sequence into to training(1-345) and testing(346-989) sequences, and retrain parameters of our pairwise and unary terms on the training subset. This results in slight improvement in performance compared to the model with parameters trained on TUD-Crossing. We obtain 93.8% MOTA and ID switches are reduced to 1, as shown in the last row of Tab. 5.3.

5.6 CONCLUSIONS

In this chapter, we propose to formulate multi-target tracking as a Minimum Cost Subgraph Multicut Problem. In contrast to the Minimum Cost Disjoint Paths formulation, which selects a set of disjoint paths as tracks and which is similar in spirit to many state-of-the-art methods, the Subgraph Multicut model selects and clusters all suitable hypotheses for each target jointly in space and time. Experiments show that our Subgraph Multicut model improves the multi-target tracking performance on several datasets underlying both the usefulness as well as the applicability of the proposed formulation. We also show initial results to the classic problem of Non-Maximum Suppression that without any changes achieves performance on par with top-performing NMS-schemes. In the future we will explore more powerful unary and pairwise terms to further improve NMS and tracking performance.

6 MULTI-PERSON TRACKING BY MULTICUT AND DEEP MATCHING

Contents

In the previous chapter, we proposed a graph-based formulation that links and clusters person hypotheses over time by solving a minimum cost subgraph multicut problem. In this chapter, we modify and extend the previous chapter in three ways: *1)* We introduce a novel local pairwise feature based on local appearance matching that is robust to partial occlusion and camera motion. *2)* We perform extensive experiments to compare different pairwise potentials and to analyze the robustness of the tracking formulation. *3)* We consider a plain multicut problem and remove outlying clusters from its solution. This allows us to employ an efficient primal feasible optimization algorithm that is not applicable to the subgraph multicut problem proposed in the previous chapter. Unlike the branch-and-cut algorithm used there, this efficient algorithm used here is applicable to long videos and many detections. Together with the novel pairwise feature, it eliminates the need for the intermediate tracklet representation.

6.1 INTRODUCTION

Multi person tracking is a problem studied intensively in computer vision. While continuous progress has been made, false positive detections, long-term occlusions and camera motion remain challenging, especially for people tracking in crowded scenes. Tracking-by-detection is commonly used for multi person tracking where a state-of-the-art person detector is employed to generate detection hypotheses for a video sequence. In this case tracking essentially reduces to an association task between detection hypotheses across video frames. This detection association task

is often formulated as an optimization problem with respect to a graph: every detection is represented by a node; edges connect detections across time frames. The most commonly employed algorithms aim to find disjoint paths in such a graph (Pirsiavash *et al.*, 2011, Segal and Reid, 2013, Andriluka *et al.*, 2008; Zhang *et al.*, 2008). The feasible solutions of such problems are sets of disjoint paths which do not branch or merge. While being intuitive, such formulations cannot handle the multiple plausible detections per person, which are generated from typical person detectors. Therefore, pre- and/or post-processing such as non maximum suppression (NMS) on the detections and/or the final tracks is performed, which often requires careful fine-tuning of parameters.

The minimum cost subgraph multicut problem proposed in Chapter 5 is an abstraction of the tracking problem that differs conceptually from disjoint path methods. It has two main advantages: *1)* Instead of finding a path for each person in the graph, it links and clusters multiple plausible person hypotheses (detections) jointly over time and space. The feasible solutions of this formulation are components of the graph instead of paths. All detections that correspond to the same person are clustered jointly within and across frames. No NMS is required, neither on the level of detections nor on the level of tracks. *2)* For the multicut formulation, the costs assigned to edges can be positive, to encourage the incident nodes to be in the same track, or negative, to encourage the incident nodes to be in distinct tracks. Thus, the number and size of tracks does not need to be specified, constrained or penalized and is instead defined by the solution. This is fundamentally different also from distance-based clustering approaches, e.g. Wen *et al.*, 2014 where the cost of joining two detections is non-negative and thus, a non-uniform prior on the number or size of tracks is required to avoid a trivial solution. Defining or estimating this prior is a well-known difficulty. We illustrate these advantages in the example depicted in Fig. 6.1: We build a graph based on the detections on three consecutive frames, where detection hypotheses within and between frames are all connected. The costs assigned to the edges encourage the incident node to be in the same or distinct clusters. For simplicity, we only visualize the graph built on the detections of two persons instead of all. By solving the minimum cost subgraph multicut problem, a multicut of the edges is found (depicted as dotted lines). It partitions the graph into distinct components (depicted in yellow and magenta, resp.), each representing one person's track. Note that multiple plausible detections of the same person are clustered jointly, within and across frames.

The effectiveness of the multicut formulation for the multi person tracking task is driven by different factors: computing reliable affinity measures for pairs of detections; handling noisy input detections and utilizing efficient optimization methods. In this work, we extend Chapter 5 on those fronts. First, for a pair of detections, we propose a reliable affinity measure that is based an effective image matching method DeepMatching (Weinzaepfel *et al.*, 2013). As this method matches appearance of local image regions, it is robust to camera motion and partial occlusion. In contrast, the pairwise feature proposed in Chapter 5 relies heavily on the spatio-temporal relations of tracklets (a short-term tracklet is used

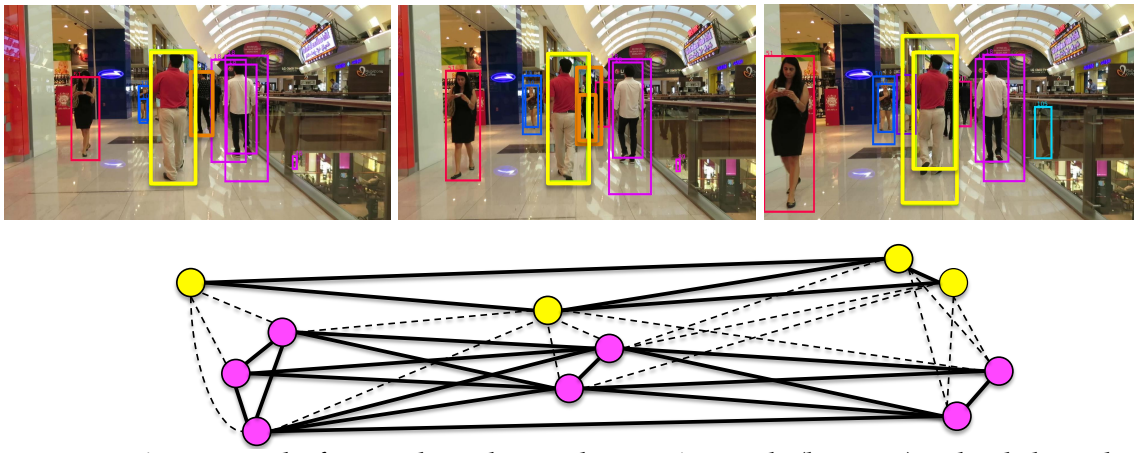


Figure 6.1: An example for tracking by multicut. A graph (bottom) is built based on the detections in three frames (top). The connected components that are obtained by solving the multicut problem indicate the number of tracks (there are two tracks, depicted in yellow and magenta respectively) as well as the membership of every detection.

to estimate the speed of a person) which works well only for a static camera and when people walk with constant speed. By introducing the DeepMatching pairwise feature, we make the multicut formulation applicable to more general moving-camera videos with arbitrary motion of persons. Secondly, we eliminate the unary variables which are introduced in Chapter 5 to integrate the detection confidence into the multicut formulation. By doing so, we simplify the optimization problem and make it amenable to the fast Kernighan-Lin-type algorithm of Keuper *et al.*, 2015b. The efficiency of this algorithm eliminates the need for an intermediate tracklet representation, which greatly simplifies the tracking pipeline. Thirdly, we integrate the detection confidence into the pairwise terms such that detections with low confidence simply have a low probability to be clustered with any other detection, most likely ending up as singletons that we remove in a post-processing step. With the above mentioned extensions, we are able to achieve competitive performance on the challenging MOT16 benchmark.

6.2 MULTI-PERSON TRACKING AS A MULTICUT PROBLEM

In Section 6.2.1, we recall the minimum cost multicut problem that we employ as a mathematical abstraction for multi person tracking. We emphasize differences compared to the minimum cost subgraph multicut problem proposed in the previous chapter. In Section 6.2.2, we define the novel DeepMatching feature and its incorporation into the objective function. In Section 6.2.3, we present implementation details.

6.2.1 Minimum Cost Multicut Problem

In this work, multi person tracking is cast as a minimum cost multicut problem (Chopra and Rao, 1993) w.r.t. a graph $G = (V, E)$ whose node V are a finite set of *detections*, i.e., bounding boxes that possibly identify people in a video sequence. Edges within and across frames connect detections that possibly identify the same person. For every edge $vw \in E$, a cost or reward $c_{vw} \in \mathbb{R}$ is to be payed if and only if the detections v and w are assigned to distinct tracks. Multi person tracking is then cast as a binary linear program

$$\min_{x \in \{0,1\}^E} \quad \sum_{e \in E} c_e x_e \tag{6.1}$$

$$\text{subject to} \quad \forall C \in \text{cycles}(G)\ \forall e \in C: \ x_e \leq \sum_{e' \in C \setminus \{e\}} x_{e'} \ . \tag{6.2}$$

Note that the costs c_e can be both positive or negative. For detections $v, w \in V$ connected by an edge $e = \{v, w\}$, the assignment $x_e = 0$ indicates that v and w belong to the same track. Thus, the constraints (6.2) can be understood as follows: If, for any neighboring nodes v and w, there exists a path in G from v to w along which all edges are labeled 0 (indicating that v and w belong to the same track), then the edge vw cannot be labeled 1 (which would indicate the opposite). In fact, (6.2) are generalized transitivity constraints which guarantee that a feasible solution x well-defines a decomposition of the graph G into tracks.

We construct the graph G such that edges connect detections not only between neighboring frames but also across longer distances in time. Such edges $vw \in E$ allow to assign the detections v and w to the same track even if there would otherwise not exist a vw-path of detections, one in each frame. This is essential for tracking people correctly in the presence of occlusion and missing detections.

Differences compared to Chapter 5. The minimum cost multicut problem (6.1)–(6.2), we consider here differs from the minimum cost subgraph multicut problem proposed in Chapter 5. In order to handle false positive detections, in Chapter 5, we introduces additional binary variables at the nodes, switching detections on or off. A cost of switching a dectection on is defined w.r.t. a confidence score of that detection. Here, we do not consider binary variables at nodes and incorporate a detection confidence into the costs of edges. In order to remove false positive detections, we remove small clusters from the solution in a post-processing step. A major advantage of this modification is that our minimum cost multicut problem (6.1)–(6.2), unlike the minimum cost subgraph multicut problem, is amenable to efficient approximate optimization by means of the KLj algorithm (Keuper *et al.*, 2015b), without any modification. This algorithm, unlike the branch-and-cut algorithm of Chapter 5, can be applied in practice directly to the graph of detections defined above, thus eliminating the need for the smaller intermediate representation by tracklets.

Optimization. Here, we solve instances of the minimum cost multicut problem approximatively with the KLj algorithm (Keuper *et al.*, 2015b). This algorithm iteratively updates bipartitions of a subgraph. The worst-case time complexity of any

Figure 6.2: Visualization of the DeepMatching results on the MOT16 sequences

such update is $O(|V||E|)$. The number of updates is not known to be polynomially bounded but is small in practice (less than 30 in our experiments). Moreover, the bound $O(|V||E|)$ is almost never attained in practice, as shown by the more detailed analysis in Keuper *et al.*, 2015b.

6.2.2 Deep Matching based Pairwise Costs

In order to specify the costs of the optimization problem introduced above for tracking, we need to define, for any pair of detection bounding boxes, a cost or reward to be payed if these bounding boxes are assigned to the same person. For that, we wish to quantify how likely it is that a pair of bounding boxes identify the same person. In Chapter 5, this is done w.r.t. an estimation of velocity that requires an intermediate tracklet representation and is not robust to camera motion. Here, we define these costs exclusively w.r.t. image content. More specifically, we build on the significant improvements in image matching made by DeepMatching (Weinzaepfel *et al.*, 2013).

DeepMatching applies a multi-layer deep convolutional architecture to yield possibly non-rigid matchings between a pair of images. Fig. 6.2 shows results of DeepMatching for two pairs of images from the MOT16 sequences[8]. The first pair of images is taken by a moving camera; the second pair of images is taken by a static camera. Between both pairs of images, matched points (blue arrows) relate a person visible in one image to the same person in the second image.

Next, we describe our features defined w.r.t. a matching of points between a pair of detection bounding boxes. Each detection bounding box $v \in V$ has the following properties: its spatio-temporal location (t_v, x_v, y_v), scale h_v, detection confidence ξ_v and, finally, a set of keypoints M_v inside v. Given two detection bounding boxes v and w connected by the edge $\{v, w\} = e \in E$, we define $MU = |M_v \cup M_w|$ and

[8]We use the visualization code provided by the authors of Weinzaepfel *et al.*, 2013

$MI = |M_v \cap M_w|$ and the five features

$$f_1^{(e)} := MI/MU \tag{6.3}$$
$$f_2^{(e)} := \min\{\xi_v, \xi_w\} \tag{6.4}$$
$$f_3^{(e)} := f_1^{(e)} f_2^{(e)} \tag{6.5}$$
$$f_4^{(e)} := (f_1^{(e)})^2 \tag{6.6}$$
$$f_5^{(e)} := (f_2^{(e)})^2 \tag{6.7}$$

Given, for any edge $e = \{v, w\} \in E$ between two detection bounding boxes v and w, the feature vector $f^{(e)}$ for this pair, we learn a probability $p_e \in (0, 1)$ of these detection bounding boxes to identify the same person. More specifically, we assume that p_e depends on the features $f^{(e)}$ by a logistic form

$$p_e := \frac{1}{1 + \exp(-\langle \theta, f^{(e)} \rangle)} \tag{6.8}$$

with parameters θ. We estimate these parameters from training data by means of logistic regression. Finally, we define the cost c_e in the objective function (6.1) as

$$c_e := \log \frac{p_e}{1 - p_e} = \langle \theta, f^{(e)} \rangle \ . \tag{6.9}$$

Two remarks are in order: Firstly, the feature $f_2^{(e)}$ incorporates the detection confidences of v and w that defined unary costs in Chapter 5 into the feature $f^{(e)}$ of the pair $\{v, w\}$ here. Consequently, detections with low confidence will be assigned with low probability to any other detection. Secondly, the features $f_3^{(e)}, f_4^{(e)}, f_5^{(e)}$ are to learn a non-linear map from features $f_1^{(e)}, f_2^{(e)}$ to edge probabilities by means of linear logistic regression.

6.2.3 Implementation Details

Clusters to tracks. The multicut formulation clusters detections jointly over space and time for each target. It is straight-forward to generate tracks from such clusters: In each frame, we obtain a representative location (x, y) and scale h by averaging all detections that belong to the same person (cluster). A smooth track of the person is thus obtained by connecting these averages across all frames. Thanks to the pairwise potential incorporating a detection confidence, low confidence detections typically end up as singletons or in small clusters which are deleted from the final solution. Specifically, we eliminate all clusters of size less than 5 in all experiments.

Maximum temporal connection. Introducing edges that connect detections across longer distance in time is essential to track people in the presence of occlusion. However, with the increase of the distance in time, the pairwise feature becomes

less reliable. Thus, when we construct the graph, it is necessary to set a maximum distance in time. In all the experiments, we introduce edges for the detections that are at most 10 frames apart. This parameter is based on the experimental analysis on the training sequences and is explained in more detail in Section 6.3.1.

6.3 EXPERIMENTS

We analyze our approach experimentally and compare to prior work on the MOT16 Benchmark (Milan *et al.*, 2016). The benchmark includes training and test sets composed of 7 sequences each. We learn the model parameters for the test sequences based on the corresponding training sequences. We first conduct an experimental analysis that validates the effectiveness of the DeepMatching based affinity measure in Sec. 6.3.1. In Sec. 6.3.2 we demonstrate that the multicut formulation is robust to detection noise. In Sec. 6.3.3 we compare our method with the best published results on the MOT16 Benchmark.

6.3.1 Comparison of Pairwise Potentials

Setup. In this section we compare the DeepMatching (DM) based pairwise potential with a conventional spatio-temporal relation (ST) based pairwise potential. More concretely, given two detections v and w, each has the following properties: spatio-temporal location (t, x, y), scale h, detection confidence ξ. Based on these properties the following auxiliary variables are introduced to capture geometric relations between the bounding boxes: $\Delta x = \frac{|x_v - x_w|}{\bar{h}}, \Delta y = \frac{|y_v - y_w|}{\bar{h}}, \Delta h = \frac{|h_v - h_w|}{\bar{h}}, y = \frac{|y_v - y_w|}{\bar{h}}, IOU = \frac{|B_v \cap B_w|}{|B_v \cup B_w|}, t = t_v - t_w$, where $\bar{h} = \frac{(h_v + h_w)}{2}$, IOU is the intersection over union of the two detection bounding box areas and ξ_{min} is the minimum detection score between ξ_v and ξ_w. The pairwise feature $f^{(e)}$ for the spatio-temporal relations (ST) is then defined as $(\Delta t, \Delta x, \Delta y, \Delta h, IOU, \xi_{min})$. Intuitively, the ST features are able to provide useful information within a short temporal window, because they only model the geometric relations between bounding boxes. DM is built upon matching of local image features that is reliable for camera motion and partial occlusion in longer temporal window.

We collect test examples from the MOT16-09 and MOT16-10 sequences which are recorded with a static camera and a moving camera respectively. The positive (negative) pairs of test examples are the detections that are matched to the same (different) persons' ground truth track over time. The negative pairs also include the false positive detections on the background.

Metric. The metric is the verification accuracy, the accuracy or rate of correctly classified pairs. For a pair of images belong to the same (different) person, if the estimated joint probability is larger (smaller) than 0.5, the estimation is considered as correct. Otherwise, it is a false prediction.

Results. We conduct a comparison between the accuracy of the DM feature and

MOT16-09: Static camera						
Feature	$\triangle t = 1$	$\triangle t = 2$	$\triangle t = 5$	$\triangle t = 10$	$\triangle t = 15$	$\triangle t = 20$
ST	0.972	0.961	0.926	0.856	0.807	0.781
DM	0.970 (-0.2%)	0.963 (+0.2%)	0.946 (+2%)	0.906 (+5%)	0.867 (+6%)	0.820 (+3.9%)
MOT16-10: Moving camera						
Feature	$\triangle t = 1$	$\triangle t = 2$	$\triangle t = 5$	$\triangle t = 10$	$\triangle t = 15$	$\triangle t = 20$
ST	0.985	0.977	0.942	0.903	0.872	0.828
DM	0.985	0.984 (+0.7%)	0.975 (+3.3%)	0.957 (+5.4%)	0.939 (+6.7%)	0.925 (+9.7%)

Table 6.1: Comparison of tracking results based on the DM and the ST feature. The metic is the accuracy or rate of correctly classified pairs on the MOT16-09 and the MOT16-10 sequences.

the accuracy of the ST feature as a function of distance in time. It can be seen from Tab. 6.1 that the ST feature achieves comparable accuracy only up to 2 frames distance. Its performance deteriorates rapidly for connections at longer time. In contrast, the DM feature is effective and maintains superior accuracy over time. For example on the MOT16-10 sequence which contains rapid camera motion, the DM feature improves over the ST feature by a large margin after 10 frames and it provides stable affinity measure even at 20 frames distance (accuracy = 0.925). On the MOT16-09 sequence, the DM feature again shows superior accuracy than the ST feature starting from $\triangle t = 2$. However, the accuracy of the DM feature on the MOT16-09 is worse than the one on MOT16-10, suggesting the quite different statistic among the sequences from the MOT16 benchmark. As discussed in Sec. 6.2.3, it is necessary to set a maximum distance in time to exclude unreliable pairwise costs. Aiming at a unique setting for all sequences, we introduce edges for the detections that are maximumly 10 frames apart in the rest experiments of this work.

6.3.2 Robustness to Input Detections

Handling noisy detection is a well-known difficulty for tracking algorithms. To assess the impact of the input detections on the tracking result, we conduct tracking experiments based on different sets of input detections that are obtained by varying a minimum detection score threshold ($Score_{min}$). For example, in Tab. 6.2, $Score_{min} = -\infty$ indicates that all the detections are used as tracking input; whereas $Score_{min} = 1$ means that only the detections whose score are equal or larger than 1 are considered. Given the fact that the input detections are obtained from a DPM detector (Felzenszwalb *et al.*, 2010), $Score_{min} = -\infty$ and $Score_{min} = 1$ are the two extreme cases, where the recall is maximized for the former one and high precision is obtained for the latter one.

Metric. We evaluate the tracking performance of the multicut model that operates on different sets of input detections. We use the standard CLEAR MOT metrics. For simplicity, in Tab. 6.2 we report the Multiple Object Tracking Accuracy (MOTA)

MOT16-09							
$Score_{\min}$	$-\infty$	-0.3	-0.2	-0.1	0	0.1	1
$\|V\|$	5377	4636	4320	3985	3658	3405	1713
$\|E\|$	565979	422725	367998	314320	265174	229845	61440
Run time (s)	30.48	19.28	13.46	11.88	8.39	6.76	1.71
MOTA	37.9	43.1	43.1	44.9	45.8	44.1	34.1
MOT16-10							
$Score_{\min}$	$-\infty$	-0.3	-0.2	-0.1	0	0.1	1
$\|V\|$	8769	6959	6299	5710	5221	4823	2349
$\|E\|$	1190074	755678	621024	511790	427847	365949	88673
Run time (s)	88.34	39.28	30.08	21.99	16.13	13.66	1.94
MOTA	26.8	32.4	34.4	34.5	34.5	33.9	23.3

Table 6.2: Tracking performance on different sets of input detections. $Score_{\min}$ indicates the minimum detection score threshold. $|V|$ and $|E|$ are the number of nodes (detections) and edges respectively.

that is a cumulative measure that combines the number of False Positives (FP), the number of False Negatives (FN) and the number of ID Switches (IDs).

Results. On the MOT16-09 sequence, when the minimum detection score threshold ($Score_{min}$) is changed from 0.1 to -0.3, the number of detection is largely increased (from 3405 to 4636), however the MOTA is only decreased by 1 percent (from 44.1% to 43.1%). Even for the extreme cases, where the detections are either rather noisy ($Score_{min} = -\infty$) or sparse ($Score_{min} = 1$), the MOTAs are still in the reasonable range. The same results are found on the MOT16-10 sequence as well. Note that, for all the experiments, we use the same parameters, we delete the clusters whose size is smaller than 5 and no further tracks splitting/merging is performed.

These experiments suggest that the multicut formulation is very robust to the noisy detection input. This nice property is driven by the fact that the multicut formulation allows us to jointly cluster multiple plausible detections that belong to the same target over time and space.

We also report run time in Tab. 6.2. The KLj multicut solver provides arguably fast solution for our tracking problem. E.g. for the problem with more than one million edges, the solution is obtained in 88.34 second. Detailed run time analysis of the KLj algorithm are shown in Keuper *et al.*, 2015b.

6.3.3 Results on MOT16

We test our tracking model on all the MOT16 sequences and submitted our results to the ECCV 2016 MOT Challenge [9] for evaluation. The performance is shown in Tab. 7.2. The detailed performance and comparison on each sequence will be

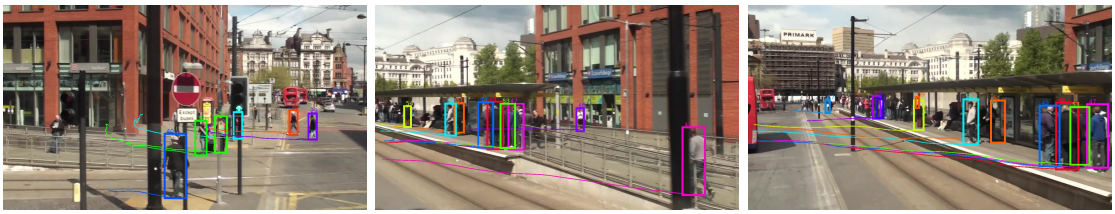
[9] https://motchallenge.net/workshops/bmtt2016/eccvchallenge.html

Method	MOTA	MOTP	FAF	MT	ML	FP	FN	ID Sw	Frag	Hz
NOMT (Choi, 2015)	46.4	76.6	1.6	18.3%	41.4%	9753	87565	359	504	2.6
MHT (Kim *et al.*, 2015)	42.8	76.4	1.2	14.6%	49.0%	7278	96607	462	625	0.8
CEM (Milan *et al.*, 2014)	33.2	75.8	1.2	7.8%	54.4%	6837	114322	642	731	0.3
TBD (Geiger *et al.*, 2014)	33.7	76.5	1.0	7.2%	54.2%	5804	112587	2418	2252	1.3
Ours	46.3	75.7	1.09	15.5%	39.7%	6449	90713	663	1115	0.8

Table 6.3: Tracking Performance on MOT16.

(a) MOT16-06 (b) MOT16-12 (c) MOT16-03

(d) MOT16-08 (e) MOT16-07 (f) MOT16-01



(g) MOT16-14 (frame 290) (h) MOT16-14 (frame 360) (i) MOT16-14 (frame 390)

Figure 6.3: Qualitative results for all the sequences from the MOT16 Benchmark. The first and second rows are the results from the MOT16-01,MOT16-03,MOT16-06, MOT16-07, MOT16-08 and MOT16-12 sequence. The third row is the result from the MOT16-14 sequence when the camera mounted on a bus is turning fast at a street intersection.

revealed at the ECCV 2016 MOT Challenge Workshop. We compare our method with the best reported results including NOMT (Choi, 2015), MHT-DAM (Kim *et al.*, 2015), TBD (Geiger *et al.*, 2014) and CEM (Milan *et al.*, 2014). Overall, we achieve the second best performance in terms of MOTA with 0.1 point below the best performed one (Choi, 2015). We visualize our results in Fig. 6.3. On the MOT16-12 and MOT16-07 sequences, the camera motion is irregular; whereas on the MOT16-03 and MOT16-08 sequences, scenes are crowded. Despite these challenges, we are still able to link people through occlusions and produce long-lived tracks. The third row of Fig. 6.3 shows images captured by a fast moving camera mounted on a bus turning at a street intersection. Under such extreme circumstance, our model is able to track people in a stable and persistent way, demonstrating the reliability of the multicut formulation for multi-person tracking task.

6.4 CONCLUSIONS

In this work, we revisit the multi-cut approach for multi-target tracking that is proposed in the previous chapter. We propose a novel pairwise potential that is built based on local image patch appearance matching. We demonstrate extensive experimental analysis and show state-of-art tracking performance on the MOT16 Benchmark. In the future we plan to further develop our approach by incorporating long-range temporal connections in order to deal with longer-term occlusions, and will extend the model with more powerful pairwise terms capable of matching person hypothesis over longer temporal gaps.

LIFTED MULTICUTS AND DEEP RE-IDENTIFICATION 7

Contents

In this chapter we revisit the modeling and representation of long-range dependencies for multi-person tracking. Humans can master it even if they loose track of a person locally by re-identifying the same person based on their appearance. Care must be taken across long distances, as similar-looking persons need not be identical. In this work, we propose a novel graph-based formulation that links and clusters person hypotheses over time by solving an instance of a minimum cost lifted multicut problem. The model generalizes the multicut tracking model proposed in the previous chapter by introducing a mechanism for adding long-range attractive connections between nodes in the graph without modifying the original set of feasible solutions. This allows us to reward tracks that assign detections of similar appearance to the same person in a way that does not introduce implausible solutions. To effectively match hypotheses over longer temporal gaps we develop new deep architectures for re-identification of people. They combine holistic representations extracted with deep networks and body pose layout obtained with a state-of-the-art pose estimation model.

7.1 INTRODUCTION

Multiple people tracking has improved considerably in the last two years, driven also by the MOT challenges (Leal-Taixé *et al.*, 2015, Milan *et al.*, 2016). One trend in this area of research has been to develop CNN-based feature representations for people appearance to effectively model relations between detection hypotheses

Figure 7.1: Qualitative results on the MOT16 Benchmark. The solid line under each bounding box indicates the life time of the track. The lifted multicut tracking model is able to link people through occlusions and produces persistent long-lived tracks

(Kim *et al.*, 2015, Leal-Taixé *et al.*, 2016). This trend has two advantages: Firstly, representations of people appearance can be learned for varying camera position and motion, a goal less easy to achieve with simple motion models, especially for monocular video due to the complexity of motion under perspective projection. Secondly, appearance facilitates the re-identification of people across long distances, unlike motion models that become asymptotically uncorrelated. Yet, incorporating long-range re-identification into algorithms for multiple people tracking remains challenging. One reason is the simple fact that similar looking people are not necessarily identical. To address these challenges, in this chapter, we generalize the mathematical models proposed in Chapter 5 and 6 so as to express the fact that similar looking people are considered as the same person only if they are connected by at least one feasible track (possibly skipping occlusion). In 5, multi-person tracking is cast as a minimum cost multicut problem (Grötschel and Wakabayashi, 1989, Chopra and Rao, 1993). There and in this chapter, every detection is represented by a node in a graph; edges connect detections within and across time frames, and costs assigned to edges can be positive, to encourage the incident nodes to be in the same track, or negative, to encourage the incident nodes to be in distinct tracks. Such mathematical abstraction has several advantages. Firstly, the number of persons is not fixed or biased by definition of the problem, but is estimated in an unbiased fashion from the video sequence and is determined by the solution of the problem. Secondly, multiple detections of the same person in the same frame are effectively clustered, which eliminates the need for heuristic non-maxima suppression. In order to avoid that distinct but similar looking people are assigned to the same track, a distinction must be made between edges that define possible connections (i.e., a feasible set) and edges that define the costs or rewards for assigning the incident nodes to distinct tracks (i.e., an objective function). We achieve this by casting the

multi-person tracking problem as a minimum cost *lifted* multicut problem (Andres, 2015).

Specifically, we make three contributions:

Firstly, we design and train deep networks for re-identifying persons by fusing human pose information. This provides a mechanism for associating person hypotheses that are temporally distant and allows to obtain correspondence before and after occlusion.

Secondly, we propose a novel formulation of multi-person tracking as the *minimum cost lifted multicut* problem. We introduce two types of edges (regular and lifted edges) into the graph. The regular edges define the set of feasible solutions in the graph, namely, which pair of nodes can be joint/cut. The lifted edges add additional long range information to the objective on which nodes should be joint/cut without modifying the set of feasible solutions. Our formulation encodes long-range information, yet penalizes long-term false joints (e.g., similar looking people) by forcing valid paths in the feasible solution in a unified and rigorous manner.

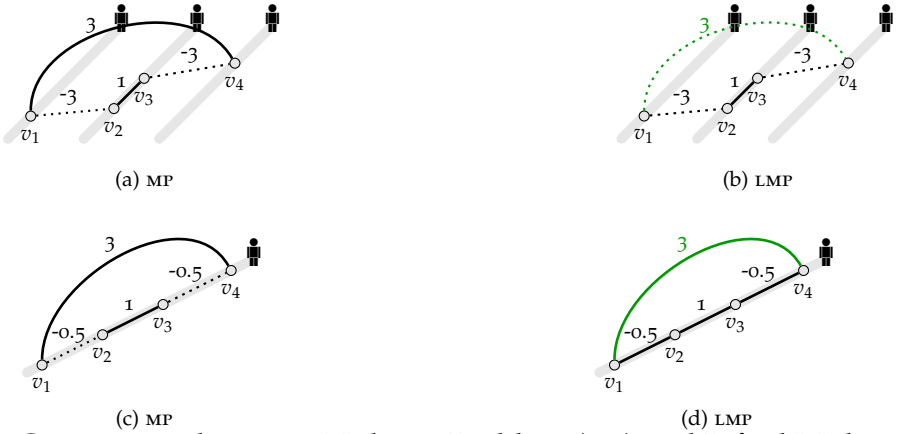
Thirdly, we show that tracks defined by local optima of this optimization problem define a new state-of-the-art for the MOT16 benchmark.

7.2 MULTI-PERSON TRACKING AS AN OPTIMIZATION PROBLEM

We now turn to our mathematical abstraction of multiple people tracking as a minimum cost lifted multicut problem (LMP). The LMP is an optimization problem whose feasible solutions can be identified with decompositions of a graph. Comparing to the minimum cost multicut problem (MP), which is defined w.r.t. a graph whose edges define possbilities of joining nodes directly into the same track, the LMP is defined, in addition, w.r.t. additional *lifted* edges that do not define possibilities of directly joining nodes. The decision of joining the nodes needs to be supported by the regular edges.

Our motivation for modeling the *lifted* edges comes from the simple fact that persons of similar appearance are not necessarily identical. Given two detections that are far apart in time and similar in appearance, it is more likely a priori that they represent the same person. At the same time, this decision has to be certified a posteriori by a track connecting the two. We achieve precisely this by introducing the two classes of edges: In order to assign two detections that are far apart in time and similar in appearance to the same cluster (person), there must exist a path (track) along the regular edges, that certifies this decision.

Two intuitive examples are given in Fig. 7.2. In (**a**) and (**b**) there are three persons in the scene, v_1 is the detection on the first person, v_2 and v_3 are the detections on the second, v_4 is on the third. The costs on the edges v_1v_2 and v_3v_4 are -3, suggesting strong rewards towards cutting the edges, and this is correct. However, the cost on the edge v_1v_4 suggests that the first and the third person look similar and introduces a strong reward towards connecting them. As a result, the MP incorrectly connects v_1 and v_4 as the same person; the LMP does not connect v_1 and v_4, as such long-range join is not supported by the local edges. (**c**) and (**d**) is another example



(a) MP (b) LMP

(c) MP (d) LMP

Figure 7.2: Comparison between Multicut Problem (MP) and Lifted Multicut Problem (LMP). Ground truth track of each person is depicted in gray. Regular edges are depicted in black, lifted edges are in green. Solid lines indicate joints, dotted lines indicate cuts. Costs of cutting edges are indicated by the numbers on the corresponding edges. (Best view in color)

where all the detections are on the same person, namely, a track that connects all the nodes in the graph is desirable. Due to partial occlusion or inaccurate bounding box localization, the costs on the local edges v_1v_2 and v_3v_4 could be ambiguous, sometimes even reverse. The long-range edge v_1v_4 correctly re-identifies the person. The MP, however, produces two clusters for a single person because the long-range edge does not have influence on the local connections. In contrast, the LMP allows us to influence an entire chain of connections between person hypotheses with a single confident long-range observation.

In the following, we discuss in detail first the parameters, then the feasible set, and finally the objective function.

Parameters. Given an image sequence, we consider an instance of the LMP with respect to the parameters defined below. The estimation of these parameters from the image sequence is discussed in the next section.

- A finite set V in which every element $v \in V$ represents a *detection* of one person in one image, i.e., a bounding box. For every detection $v \in V$, we also define its scale $s_v \in \mathbb{N}$ and the coordinates $x_v, y_v, t_v \in \mathbb{N}$ of its center in the image sequence.
- For every pair $v, w \in V$: a conditional probability $p_{vw} \in (0, 1)$ of v and w to represent distinct persons, given their scales, coordinates and appearance.
- A graph $G = (V, E)$ whose edges are regular edges that connect detections v, w in the same image $t_v = t_w$ and also connect detections v, w in distinct images $t_v \neq t_w$ that are *close in time*, i.e., for some fixed upper bounds $\delta_t \in \mathbb{N}$: $|t_v - t_w| \leq \delta_t$.
- A graph $G' = (V, E')$ with $E \subseteq E'$ whose additional edges $\{v, w\} \in E' \setminus E$ are

lifted edges which connect detections v, w that are *far apart in time* and *similar in appearance*, i.e., for some fixed $p_0 \in (0, \frac{1}{2})$:

$$|t_v - t_w| > \delta_t \tag{7.1}$$

$$p_{vw} \leq p_0 \tag{7.2}$$

The graph G defines the decomposition space, and the graph G' adds lifted edges $E' \setminus E$ on top of G and defines the structure of the cost function.

Feasible Set. The feasible solutions of the LMP can be identified with the decompositions (clusterings) of the graph G. Here, in the context of tracking, every component (cluster) of detections defines a track of one person. It is therefore reasonable to think of our approach as tracking by clustering.

Formally, any feasible solution of the LMP is a 01-vector $x \in \{0,1\}^{E'}$ in which $x_{vw} = 1$ indicates that the nodes v and w are in distinct components. In order to ensure that x well-defines a decomposition of G, it is further constrained to the set $X_{GG'} \subseteq \{0,1\}^{E'}$ of those $x \in \{0,1\}^{E'}$ that satisfy the system of linear inequalities written below.

$$\forall C \in \text{cycles}(G)\, \forall e \in C :$$

$$x_e \leq \sum_{e' \in C \setminus \{e\}} x_{e'} \tag{7.3}$$

$$\forall vw \in E' \setminus E\, \forall P \in vw\text{-paths}(G) :$$

$$x_{vw} \leq \sum_{e \in P} x_e \tag{7.4}$$

$$\forall vw \in E' \setminus E\, \forall C \in vw\text{-cuts}(G) :$$

$$1 - x_{vw} \leq \sum_{e \in C} (1 - x_e) \tag{7.5}$$

The constraints (7.3) are generalized transitivity constraints which mean: For any neighboring nodes v and w, if there exists a path from v to w in G along which all edges are labeled as 0, then the edge vw can only be labeled as 0. The constraints (7.4) and (7.5) guarantee, for every feasible solution and every lifted edge $vw \in E' \setminus E$, that the label x_{vw} of this edge is 0 (indicating that v and w belong to the same track) if (7.4) and only if (7.5) v and w are connected in the smaller graph G by a path of edges labeled 0. By assigning a cost or reward $c_{vw} \in \mathbb{R}$ to a lifted edge $vw \in E' \setminus E$, we can thus assign this cost or reward precisely to those feasible solutions for which v and w belong to distinct tracks, *without* introducing the additional possibility of joining v and w directly.

Objective function. We consider instances of the LMP of the form

$$\min_{x \in X_{GG'}} \sum_{e \in E'} c_e x_e \tag{7.6}$$

with the costs c_e defined as

$$c_e = \log \frac{1 - p_e}{p_e} \quad . \tag{7.7}$$

The objective function is chosen such that solutions are decompositions of G into tracks that maximize the probability of detections representing the same or distinct persons. More specifically, we define p_e as a logistic form:

$$p_e := \frac{1}{1 + \exp(-\langle \theta, f^{(e)} \rangle)}. \tag{7.8}$$

Then the cost c_e has the form:

$$c_e := \log \frac{1 - p_e}{p_e} = -\langle \theta, f^{(e)} \rangle \quad . \tag{7.9}$$

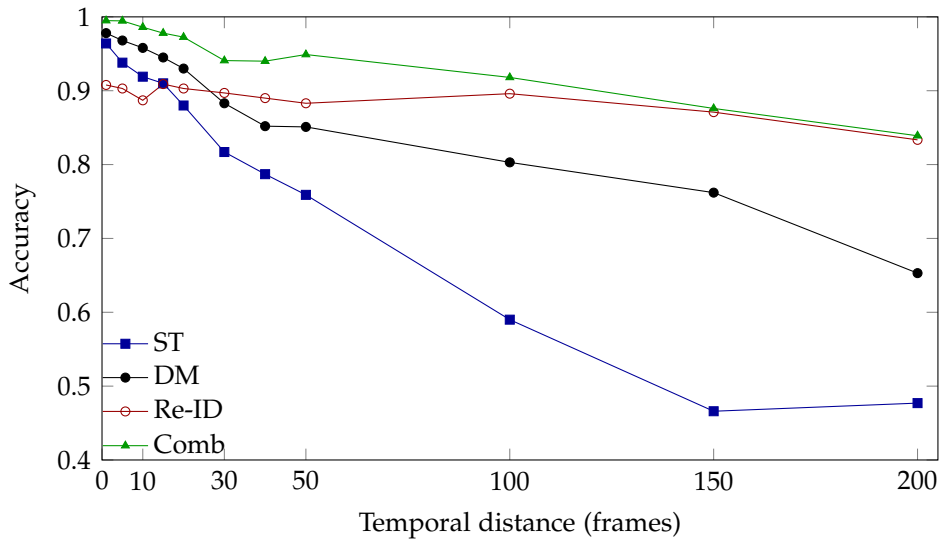
The model parameter θ is estimated on the training set by means of logistic regression. The feature $f^{(e)}$ describes the similarity between detections. In this work, $f^{(e)}$ is defined as a combination of person re-identification confidence (Sec. 7.4), deep correspondence matching, and spatio-temporal relations, which is discussed in Sec. 7.3

Optimization. The minimum cost lifted multicut problem defined by (7.6) is APX-hard (Demaine *et al.*, 2006). Given the size of instances of our tracking problems, solving to optimality or within tight bounds using branch and cut is beyond feasibility. In this work, we exploit a primal heuristic proposed by Keuper *et al.*, 2015b, where the bi-partitions of a subgraph are updated by a set of sequences of transformations. The update has the worst-case complexity of $O(|V||E|)$ which is almost never reached in practice. Detailed run time analysis can be found in Keuper *et al.*, 2015b.

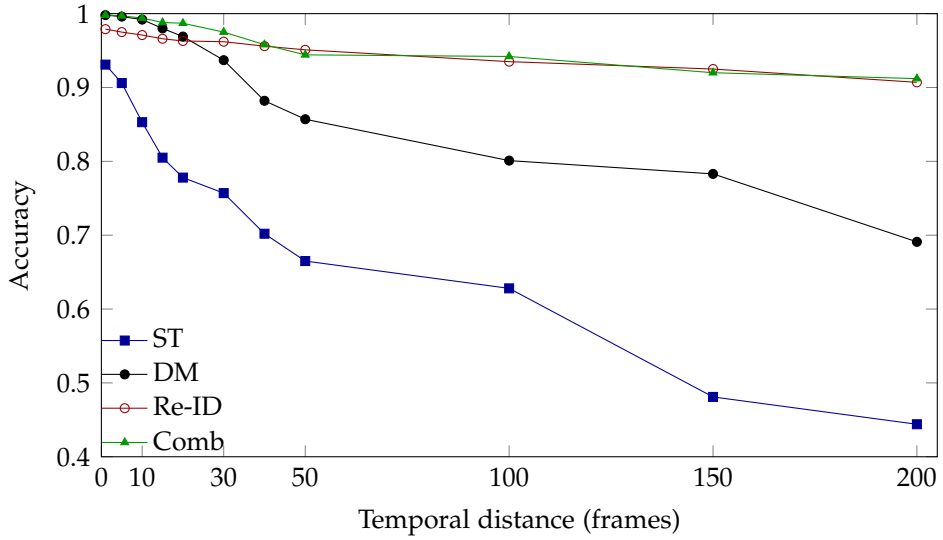
7.3 PAIRWISE POTENTIALS

As discussed in Sec. 7.2, the cost c_e in the objective function (7.6) is defined as $c_e = -\langle \theta, f^{(e)} \rangle$. In this section, we introduce the feature $f^{(e)}$, which is based on three information sources: spatio-temporal relations (ST), dense correspondence matching (DM) and person re-identification confidence (Re-ID) that is described in the previous section.

ST. The spatio-temporal relation based feature is commonly used in many multi-person tracking works (Pirsiavash *et al.*, 2011, Xiang *et al.*, 2015, Choi, 2015), as it is a good affinity measure for pairs of detections that are in close proximity. Given two detections v and w, each has spatio-temporal locations (x, y, t) and scale s. The ST feature is defined as $f_{st} = \frac{\sqrt{(x_v - x_w)^2 + (y_v - y_w)^2}}{\bar{s}}$, where $\bar{s} = \frac{(s_v + s_w)}{2}$. Intuitively, the ST features are able to provide useful information within a short temporal window.



(a)MOT16-02



(a)MOT16-11

Figure 7.3: Accuracy of pairwise affinity measures on the MOT16-02 **(a)** and MOT16-11 **(b)** sequences.

They model the geometric relations between bounding boxes but do not take image content into account.

DM. DeepMatching (Weinzaepfel *et al.*, 2013) is introduced as a powerful pairwise affinity for multi-person tracking in the previous chapter. We apply it in this work as well. Given two detections v and w, each has a set of matched keypoints M. We define $MU = |M_v \cup M_w|$, and $MI = |M_v \cap M_w|$ between the set M_v and M_w. Then the pairwise feature between the two detections is defined as $f_{dm} = MI/MU$.

Re-ID. The DM feature is based on local image patch matching, which makes it robust to irregular camera motion and to partial occlusion in short temporal distance. As shown in Chapter 6 and in the experiment section of this chapter, the performance of the DM feature drops dramatically when increasing temporal distance. ReID is explicitly trained for the task of person re-identification. It is robust with respect to large temporal and spatial distance and allows long-range association. We utilize our deep re-identification model (StackNetPose) for modeling the long-range connections. Our final pairwise feature $f^{(e)}$ is defined as $(f_{st}, f_{dm}, f_{reID}, \xi_{min}, f_{st}^2, f_{st} \cdot f_{dm}, \ldots, \xi_{min}^2)$, where ξ_{min} is the lower detection confidence within the pair, and f_{reID} is the probability estimated by our StackNetPose. The quadratic terms introduce a non-linear mapping from the feature space to the cost space.

7.3.1 Experimental Analysis

In this section, we present an analysis of the pairwise features. We also choose MOT16-02 and MOT16-11 from the MOT16 training set for the analysis, as the imaging conditions and camera motion are largely different between these two sequences. The test example collection and the evaluation metric are the same as for evaluating the person re-identification networks, namely for every test pair, we estimate the probability of the pair of images containing the same person. For the positive (negative) pairs, if the estimated probabilities are larger (smaller) than 0.5, they are considered as correctly classified examples. Any bias toward cut or joint decreases the tracking performance. A higher accuracy leads to a better tracking performance. We conduct a comparison between features as a function of temporal distance. Unlike the previous work where the temporal distance is only up to 20 frames (e.g. Choi, 2015), we demonstrate much longer temporal distance (200 frames), as our model is able to incorporate such information.

Results. It can be seen from Fig. 7.3 that the DM feature achieves the best accuracy up to 20 frames, but its performance deteriorates for connections at longer time span. The performance of the ST feature drops quickly after 5 frames. This is especially pronounced on the MOT16-11 sequence that has rapid camera motion. In contrast, the Re-ID feature is effective and maintains high accuracy over time. For example on the MOT16-11 sequence the Re-ID (red line) improves over DM (black line) by a notable margin for the temporal distances that are larger than 50 frames. When we combine the three features (Comb, green line in Fig. 7.3), we obtain the best accuracy at all the temporal distances. The reason is that, at different temporal

distance, our combined feature is able to take advantage from different information sources. E.g., when the temporal distance is smaller than 30 frames (1 sec. for these two sequences), the DM and ReID features combine both low-level (local image patch matching) and high-level (person-specific appearance similarity) to produce high accuracy pairwise affinity measures. When the temporal distance increases gradually, the ReID feature becomes more and more informative. However, still adding the ST and DM feature improves the overall accuracy, because they act as a regularizer, that forbids physically impossible associations. Based on these results, we use the combined feature in our tracking experiments.

7.4 PERSON RE-IDENTIFICATION FOR TRACKING

Traditionally, person re-identification is the task to associate observed pedestrians in non-overlapping camera views. In the context of multi-person tracking, linking the detected pedestrians across the whole video can be viewed as re-identification with special challenges: occlusions, cluttered background, large difference in image resolution and inaccurate bounding box localization. In this section, we investigate several CNN architectures for re-identification for the multi-person tracking task. Our basic CNN architecture is VGG-16 Net (Simonyan and Zisserman, 2014). Particularly, we propose a novel person re-identification model that combines the body pose layout obtained with state-of-the-art pose estimation methods.

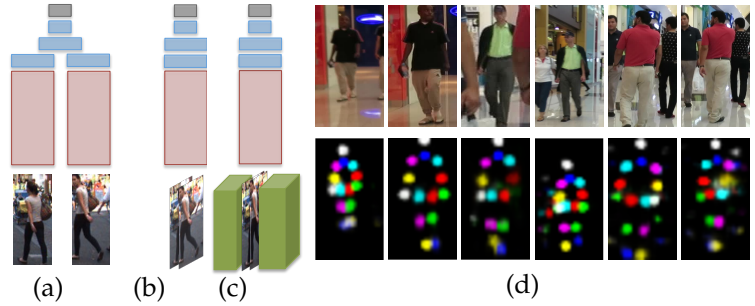
Data Collection. One of the key ingredients of deep CNNs is the availability of large amounts of training data. To apply to re-identification for tracking, we collect images from 8 training sequences of the MOT15 benchmark (Leal-Taixé *et al.*, 2015) and 5 sequences of the MOT16 benchmark (Milan *et al.*, 2016). We also collect person identity examples from the CUHK03 (Li *et al.*, 2014), Market-1501 (Zheng *et al.*, 2015) datasets that are captured by 6 surveillance cameras. We use the MOT16-02 and MOT16-11 sequences from the MOT16 training set as test sets. Overall a total of 2511 identities is used for training and 123 identities for testing.

7.4.1 Architectures

In this chapter, we explore three architectures, namely ID-Net, SiameseNet, and StackNet.

ID-Net. We first learn a VGG net Φ to recognize $N = 2511$ unique identities from our data collection as an N-way classification problem. We re-size the training images to $112 \times 224 \times 3$. Each image $x_i, i = 1, ..., M$ associates to a ground truth identity label $y_i \in \{1, ..., N\}$. The VGG estimates the probability of each image being each label as $p_i = \Phi(x_i)$ by a forward pass. The network is trained by the softmax loss.

During testing, given an image from unseen identities, the final softmax layer is removed and the output of the fully-connected layer Φ_{f7} is used as the identity feature. Given a pair of images, the Euclidean distance between the two identity



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

Model	Acc. (%)
ID-Net	80.4
SiameseNet	84.7
StackNet	86.9
StackNetPose	90.0

(e)

Figure 7.4: (a) SiameseNet. (b) StackNet. (c) StackNetPose. Red rectangles indicate the convolutional, relu and pooling layers of VGG16. Blue rectangles indicate the fully-connected layers. Grey rectangles on the top of each network are the loss layers. Green boxes are the stacked body part score maps. (d) Example results from StackNetPose. (e) Comparison of person re-identification models.

features can be used to decide whether the pair contains the same identity. In the experiments we observe that this identity feature already provides good accuracy. However, the performance is boosted by turning to a Siamese architecture and a StackNet, explained next.

SiameseNet. A Siamese architecture means the network contains two symmetry CNNs which share the parameters. We start with a commonly used Siamese architecture (Chopra *et al.*, 2005) as shown in Fig. 7.4(a). To model the similarity we use fully connected layers on top of the twin CNNs. More specifically, the features $FC_6(x_i), FC_6(x_j)$ from a pair of images are extracted from the first fully-connected layer of the VGG-based Siamese network that shares the weights. Then the features are concatenated and transformed by two fully-connected layers (FC_7, FC_8), where FC_7 are followed by a ReLU non-linearity. FC_8 uses a softmax function to produce a probability estimation over a binary decision, namely the same identity or different identities.

StackNet. The most effective architecture we explored is the StackNet, where we stack a pair of images together along the RGB channel. The input to the network becomes $112 \times 224 \times 6$. Then the filter size of the first convolutional layer is changed from $3 \times 3 \times 3$ to $3 \times 3 \times 6$, and for the rest of the network we follow the VGG architecture. The last fully-connected layer models a 2-way classification problem, namely the same identity or different identities. During testing, given a pair of images, both SiameseNet and StackNet produce the probability of the pair being the same/different identities by a forward pass.

The StackNet allows a pair of images to communicate at the early stage of the network, but it is still limited by the lack of ability to incorporate body part correspondence between the images. Next, we propose a body part fusing method to explicitly allow modeling the semantic body part information within the network.

7.4.2 Fusing Body Part Information

A desirable property of the network is to localize the corresponding regions of the body parts, and to reason about the similarity of a pair of pedestrian images based on localized body regions as well as the full images. We implement such a model by fusing body part detections into the CNN. More specifically, we utilize the body part detectors proposed in Pishchulin *et al.*, 2016 to produce individual score maps for 14 body parts, namely, head, shoulders, elbows, wrists, hips, knees, and ankles, each with left/right symmetry body parts except the head which is indicated by head top and head bottom. We combine the score maps from every two symmetry body parts which results in 7 scores maps; each has the same size as the input image. We stack the pair of images as well as the 14 score maps together to form a $112 \times 224 \times 20$ input volume. Now the filter size of the first convolutional layer is set as $3 \times 3 \times 20$, and the rest of the network follows the VGG16 architecture with a 2-way classification layer in the end. In Fig. 7.4(d) we show several examples of estimated body poses on our dataset. Note that augmenting the network with body layout information can be interpreted as an attention mechanism that allows us to focus representation on the relevant portions on the input. It can also be seen as a mechanism to highlight the foreground and enable the network to establish corresponding regions between input images.

7.4.3 Experimental Analysis

Training. Our implementation is based on the Caffe deep learning framework (Jia *et al.*, 2014). To learn the ID-Net, our VGG model is pre-trained on the ImageNet Classification task. Following a common practice in face recognition/person ReID literature (Parkhi *et al.*, 2015), we use our ID-Net as initialization for learning the SiameseNet, StackNet and StackNetPose, which makes the training faster and produces better results.

Setup. We have 123 person identities as test examples which are collected from the MOT16-02 and the MOT16-11 sequence. More specifically, on these two sequences, detections that are considered as true positives for a certain identity are those whose intersection-over-union (IOU) with the ground truth of the identity is larger than 0.5. Given the true positive detections for all the identities, we randomly select 1,000 positive pairs from the detections assigned to the same identity and 4000 negative pairs from the detections assigned to different identities as our test set. A larger ratio of negative pairs in the test set is to simulate the real positive/negative distribution during the tracking. For every test pair, we estimate the probability of the pair of images containing the same person. For the positive (negative) pairs, if the estimated probabilities are larger (smaller) than 0.5, they are considered as correctly classified examples. The metric is the verification accuracy, the accuracy or rate of correctly classified pairs. For the ID-Net, the verification result (same/different identities) of pairs of images is obtained by testing whether the distance between the extracted features is smaller than a threshold. The threshold is obtained on a

separate validation data to maximize the verification accuracy.

Results. It can be seen from Fig. 7.4(e) that the l^2 distance of the Φ_{f7} features from the ID-Net already produces reasonable accuracy. The performance is improved by applying the SiameseNet, from 80.4% to 84.7%. The accuracy is further improved when using the StackNet, achieving 86.9% accuracy. Fusing the body part information (StackNetPose) outperforms all other models by a large margin, achieving 90.0% accuracy. For our tracking task, we use the StackNetPose model to generate person re-identification confidence. We show three pairs of detections that are correctly estimated by StackNetPose in Fig. 7.4(d). It can be seen that the body part maps enable the network to localize the person despite the inaccurate bounding boxes (the first/second pairs) and cluttered background (the third pair).

7.5 EXPERIMENTS

We perform our tracking experiments and compare to prior works on the MOT16 Benchmark (Milan *et al.*, 2016). The test set contains 7 sequences, where camera motion, camera angle, and imaging condition are largely different. For each test sequence, the benchmark also provides a training sequence that is captured in the similar setting. Therefore, we learn the model parameter θ (defined in Eq. (7.9)) for the test sequences on the corresponding training sequences.

For analyzing our tracking models, we use MOT16-02 and MOT16-11 from the training set as validation sequences, the same as previous sections. The model parameter θ trained on MOT16-02 is used for MOT16-11 and vice versa. To obtain the final tracks from the clusters generated by MP or LMP, we estimate a smoothed trajectory from the detections that belongs to the same cluster. We do not consider any clusters whose size are less than 5 in all the experiments.

Evaluation Metric. We follow the standard CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008) for evaluating multi-person tracking performance. The metrics includes multiple object tracking accuracy (MOTA), which combines identity switches (IDs), false positives (FP), and false negatives (FN). Beside we also report multiple object tracking precision (MOTP), mostly tracked (MT), mostly lost (ML) and fragmentation (FM).

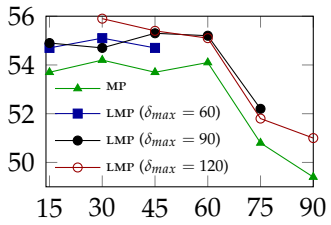
7.5.1 Lifted Edges versus Regular Edges

The graph for the lifted multicut (LMP) includes two types of edges: regular edges and lifted edges. The regular edges define the decomposition of the graph. The lifted edges introduce long-range information on which nodes should be joint/cut without modifying the set of feasible solutions. They penalize long-term false joint (e.g. similar looking people) by forcing valid paths in the feasible solution. As shown in Fig. 7.3, even beyond 50 frames, the accuracy of our pairwise affinity measure is still above 90%, Such good pairwise affinity should be leveraged into the tracking model. However, if we encode them by regular edges, we have 10% chances of

MOT16-11					
δ_{max}	30	60	90	120	150
MP	54.2	54.1	49.4	43.9	32.1
LMP	54.5 (+0.3)	55.1 (+1)	55.3 (+5.9)	55.0 (+11.1)	51.1 (+19.1)

MOT16-02					
δ_{max}	30	60	90	120	150
MP	19.9	21.5	21.2	19.1	17.2
LMP	21.3 (+1.4)	22.4 (+0.9)	21.3 (+0.1)	22.3 (+3.2)	19 (+1.8)

(a) Varying δ_{max}



(b) Varying δ_t

Figure 7.5: Comparison of Multicut model (MP) and Lifted Multicut model (LMP) with different δ_{max} values (a) and different δ_t values (b).

Method	MOTA	MOTP	FAF	MT	ML	FP	FN	ID Sw	Frag	Hz	Detector
Milan *et al.*, 2014	33.2	75.8	1.2	7.8%	54.4%	6837	114322	642	731	0.3	Public
Geiger *et al.*, 2014	33.7	76.5	**1.0**	7.2%	54.2%	5804	112587	2418	2252	1.3	Public
Le *et al.*, 2016	37.6	75.9	2.0	9.6%	55.2%	11,969	101,343	481	1,012	0.6	Public
Ban *et al.*, 2016	38.4	75.4	1.9	7.5%	47.3%	11,517	99,463	1,321	2,140	0.3	Public
Fagot-Bouquet *et al.*, 2016	41.0	74.8	1.3	11.6%	51.3%	7896	99224	430	963	1.1	Public
Kim *et al.*, 2015	42.9	76.6	**1.0**	13.6%	46.9%	**5668**	97919	499	659	0.8	Public
Choi, 2015	46.4	76.6	1.6	**18.3%**	41.4%	9753	87565	**359**	**504**	**2.6**	Public
Multicut (MP)	46.3	75.7	1.1	15.5%	**39.7%**	6373	90914	657	1114	0.8	Public
Lifted Multicut (LMP)	**48.8**	**79.0**	1.1	18.2%	40.1%	6654	**86245**	481	595	0.5	Public

Table 7.1: Tracking Performance on the MOT16 test set. Best in bold, second best in blue.

making a false joint, such errors directly produce long false-positive tracks. If they are lifted edges, connecting those detections must be certified by the majority of the local regular edges. Two intuitive examples are shown in Fig. 7.2. In this section we perform experimental analysis on the two graph variants: Multicut (MP) and Lifted Multicut (LMP), to validate the effectiveness of the proposed methods.

Given a tracking instance, intuitively, we would connect detections with regular edges up to a certain temporal distance to overcome potential missing detections due to occlusion. For the further distant detections, we would connect them with lifted edges to incorporate person re-identification information into the model to gain better tracking performance. Following the intuition, our MP is constructed in the way that besides having the regular edges between neighboring frames, we also introduce regular edges between all pairs of detections whose temporal distance are up to δ_{max}. The LMP has a combination of regular edges and lifted edges, we denote the temporal distance where we start to change the regular edges to the lifted edges as δ_t.

Varying δ_{max}. In our first analysis, we gradually change the value of δ_{max} from 1 to 150 frames. As shown in Fig. 7.5(a), on the MOT16-11 sequence, the MP achieves competitive MOTA (54.2%) when δ_{max} equals 30 frames, but the performance decreases

significantly when δ_{max} is increased to 150 frames (5 sec on the MOT16-11). The reason is that the long-range regular edges change the feasible solution of the MP. Although the accuracy of the pairwise affinity at 150 frames is near 90%, the model can still make catastrophic false joint, which introduces long-term false positive tracks. Similar results are obtained on the MOT16-02 sequence, MOTA drops to 17.2% when $\delta_{max} = 150$.

For the LMP, we also change δ_{max} from 1 to 150 frames and we set $\delta_t = \delta_{max}/2$. Comparing to the MP, the LMP obtains the best MOTA on the MOT16-11 sequence (55.3%) as well as on the MOT16-02 sequence (22.4%). Moreover, it presents a superior performance in all the settings. Particularly for the long-range connections, the margin between the MP and the LMP is more than 10% on the MOT16-11 sequence. Note that, these experiment results reveal a very desirable property of the LMP: stability with respect to the range of connections. Given a new tracking instance, due to unknown camera motion and imaging condition, it is not trivial to build a proper graph for the MP. As to the LMP, due to its robustness and stability, we are free to choose any sensible range of connections. In the next experiment, we further reveal the stability of the LMP by varying δ_t.

Varying δ_t. As shown in Fig. 7.5(b), we evaluate the influence of δ_t on LMP under 3 different δ_{max} settings, namely $\delta_{max} = 60, 90, 120$. As a baseline, the tracking performance of MP with $\delta_{max} = 15, 30, 45, 60, 75, 90$ is also shown in the Fig. 7.5(b), depicted as the green line. It can be seen that at all the temporal distances, adding lifted edges improves the tracking performance over MP, suggesting that long-range person re-identification information is useful for the tracking task. Furthermore, for the longer temporal distance (e.g. $\delta_{max} = 90$), MOTA of the MP drops significantly (49.4%); however, for the LMP with $\delta_{max} = 90$, MOTA maintains at higher levels for $\delta_t = 15, 30, 45, 60$ (black line), indicating that LMP is also robust to a large range of δ_t. Overall, the results show that our LMP is able to encode long-range information in a more rigorous manner, such that it produces much more stable and robust tracking results.

7.5.2 Results on the MOT16 Benchmark

Here we present our results on the MOT16 test set. We first use the public set of detections and compare our method with the best published results on the benchmark, including NOMT (Choi, 2015), MHT-DAM (Kim *et al.*, 2015), OVBT (Ban *et al.*, 2016), LTTSC-CRF (Le *et al.*, 2016), CEM (Milan *et al.*, 2014), TBD (Geiger *et al.*, 2014) and Multicut (Chapter 6). Multicut (Chapter 6) is the most relevant approach comparing to the lifted multicut model, where the deep matching feature is employed and tracking is cast as the minimum cost multicut problem. It can be seen from Tab. 7.2 that our method establishes a new state-of-the-art performance in terms of MOTA, MOTP and false negative (FN). Comparing to the previous best result, we improve MOTA by 2.5% and MOTP by 3.1%. For FAF, MT, ML and FM, our method achieves the second best performance. The improvement over Multicut demonstrates the advantage of incorporating the long-range person re-identification

Method	MOTA	MOTP	FAF	MT	ML	FP	FN	ID Sw	Frag	Hz	detector
Choi, 2015	62.2	79.6	**0.9**	32.5%	31.1%	5119	63352	**406**	642	3.1	-
Lee *et al.*, 2016	62.4	78.3	1.7	31.5%	24.2%	9855	57257	1394	1318	**34.9**	-
Yu *et al.*, 2016	66.1	79.5	**0.9**	34.0%	20.8%	**5061**	55914	805	3093	9.9	Yu *et al.*, 2016
Yu *et al.*, 2016	68.2	79.4	1.9	41.0%	**19.0%**	11479	45605	933	1093	0.7	Yu *et al.*, 2016
LMP (our)	**71.0**	**80.2**	1.3	**46.9%**	21.9%	7880	**44564**	434	**587**	0.5	Yu *et al.*, 2016

Table 7.2: Comparison of our lifted multicut (LMP) approach with the best-performing methods on the MOT16 benchmark. In this comparison we use the people detections privided by Yu *et al.*, 2016. We achieve the best result over all approaches using either private or public set of detections.

Method	MOT16-01	MOT16-03	MOT16-06	MOT16-07	MOT16-08	MOT16-12	MOT16-14	Average MOTA
Choi, 2015	54.1	72.7	61.3	49.9	**42.9**	50.3	39.8	53.0
Lee *et al.*, 2016	54.7	73.2	57.7	54.4	41.9	42.3	42.0	52.3
Yu *et al.*, 2016	57.7	78.1	63.6	57.0	36.2	48.0	46.2	55.3
Yu *et al.*, 2016	51.7	81.4	63.6	58.4	39.6	48.2	45.4	55.5
LMP (our)	**65.6**	**82.8**	**65.7**	**62.1**	38.0	**51.6**	**55.6**	**60.2**

Table 7.3: Tracking performance (MOTA) on each of the test sequences from the MOT16 benchmark.

information with the lifted multicut formulation.

Furthermore, we provide the tracking result obtained using the detections provided by Yu *et al.*, 2016. The comparison of our approach with the best published results, including NOMT (Choi, 2015), MCMOT (Lee *et al.*, 2016), POI (Yu *et al.*, 2016) and KDNT (Yu *et al.*, 2016) is shown in Tab. 7.2. Note that our results are directly comparable with POI (Yu *et al.*, 2016) and KDNT (Yu *et al.*, 2016) since we use the same set of detections provided by Yu *et al.*, 2016. The accuracy of detections from Yu *et al.*, 2016 is considerably higher compared to the public detections. This translates to improvements in tracking accuracy and allows us to achieve the best result on the MOT16 benchmark accross all methods. The results in Tab. 7.2 show that our approach achieves the best performance in terms of MOTA, MOTP, MT, false negatives (FN) and track fragmentation (Frag). Compared to the previously best result of Yu *et al.*, 2016 we improve MOTA by 2.8 percent points.

In Tab. 7.3 we evaluate the tracking results individually for each of the test

Method	MOT16-01	MOT16-03	MOT16-06	MOT16-07	MOT16-08	MOT16-12	MOT16-14	Total
#Boxes	6,395	104,556	11,538	16,322	16,737	8,295	18,483	182,326
#Tracks	23	148	221	54	63	86	164	759
Density	14.2	69.7	9.7	32.6	26.8	9.2	24.6	30.8

Table 7.4: Statistics of the test sequences from the MOT16 benchmark.

sequences. Notice that our approach achieves the best result on 6 out or 7 sequences improving on some of the sequences by a large margin. For example on the sequences "MOT16-01" and "MOT16-14" the improvement is 7.9 and 9.4 MOTA points respectively, whereas the improvement on the easier sequence "MOT16-02" is smaller (1.4 MOTA). The improvement on the full banchmark (c.f. Tab 7.2) is less pronounced since the overall MOTA is obtained by jointly considering all the sequences. Therefore sequences with larger number of people have more influence on the overall MOTA. We show the number of annotated bounding boxes and tracks in each sequence in Tab. 7.4. Notice that more than half of all annotated bounding boxes belong to the sequence "MOT16-03", which biases the overall result towards the results on this sequence.

7.6 CONCLUSIONS

Incorporating long-range information for multi-person tracking is challenging. In this work, we propose to model such long-range information by pose aided deep neural networks. Given the fact that similar looking people are not necessarily identical, we propose a minimum cost lifted multicut formulation where the long-range person re-identification information is encoded in the way that it forces valid paths along the local edges. In the end, we show that the proposed tracking method outperforms previous works on the challenging MOT16 benchmark.

8 JOINT SEGMENTATION AND TRACKING OF MULTIPLE OBJECTS

Contents

In previous chapters, the Minimum Cost Multicut Formulations have been proposed and proven to be successful for the multi target tracking task. In this chapter, we show that such formulation can be used for jointly solving the motion segmentation and the multi-target tracking tasks. Both tasks benefit from decomposing a graphical model into an optimal number of connected components based on attractive and repulsive pairwise terms. The two tasks are formulated on different levels of granularity and, accordingly, leverage mostly local information for motion segmentation and mostly high-level information for multi-target tracking. In this chapter we argue that point trajectories and their local relationships can contribute to the high-level task of multi-target tracking and also argue that high-level cues from object detection and tracking are helpful to solve motion segmentation. We propose a joint graphical model for point trajectories and object detections whose Multicuts are solutions to motion segmentation *and* multi target tracking problems at once. Results on the FBMS59 motion segmentation benchmark (Ochs *et al.*, 2014) as well as on pedestrian tracking sequences demonstrate the promise of this joint approach.

8.1 INTRODUCTION

Several problems in computer vision, such as image segmentation or motion segmentation in video, are traditionally approached in a low-level, bottom-up way while other tasks like object detection, multi-target tracking, and action recognition often require previously learned model information and are therefore traditionally approached from a high-level perspective.

In this chapter, we propose a joint formulation for one such classical high-level

problem (multi-target tracking) and a low-level problem (moving object segmentation). Multi-target tracking and motion segmentation are both active fields in computer vision (Segal and Reid, 2013, Huang *et al.*, 2008, Wojek *et al.*, 2010, Andriluka *et al.*, 2010, Fragkiadaki *et al.*, 2012, Zamir *et al.*, 2012, Wojek *et al.*, 2013, Henschel *et al.*, 2014, Tang *et al.*, 2014, Shi *et al.*, 2013, Ochs *et al.*, 2014, Rahmati *et al.*, 2014, Ji *et al.*, 2014, Keuper *et al.*, 2015a). These two problems are clearly related in the sense that their goal is to determine those regions that belong to the same moving object in an image sequence.

We argue that these interrelated problems can and should be addressed jointly so as to leverage the advantages of both. In particular, the low-level information contained in point trajectories and in their relation to one another form important cues for the high-level task of multi-target tracking. They carry the information where single, well localized points are moving and can thus help to disambiguate partial occlusions and motion speed changes, both of which are key challenges for multi-target tracking. For motion segmentation, challenges are presented by (1) articulated motion, where purely local cues lead to over-segmentation and (2) coherently moving objects where motion cues cannot tell the objects apart. High level information from an object detector or even an object tracking system is beneficial as it provides information about the rough object location, extent, and possibly re-identification after occlusion.

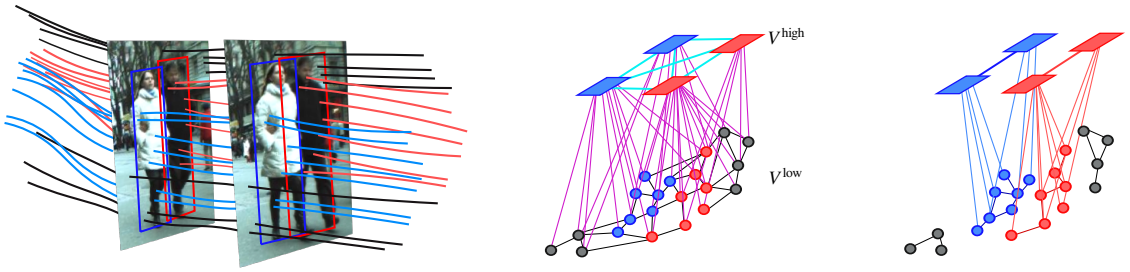
Ideally, employing such pairwise information between detections may replace higher-order terms on trajectories as proposed in Ochs and Brox, 2012. While it is impossible to tell two rotational or scaling motions apart from only pairs of trajectories, pairs of detection bounding boxes contain enough points to distinguish their motion. With sufficiently complex detection models, even articulated motion can be disambiguated.

To leverage high-level spatial information as well as low-level motion cues in both scenarios, we propose a unified graphical model in which multi-target tracking and motion segmentation are both cast in one graph partitioning problem. As a result, the method provides consistent identity labels in conjunction with accurate segmentations of moving objects.

We show that this joint graphical model improves over the individual, task specific models. Our results improve over the state of the art in motion segmentation evaluated on the FBMS59 (Ochs *et al.*, 2014) motion segmentation benchmark as well as over the state of the art in multiple object tracking evaluated on the 2D MOT 2015 (Leal-Taixé *et al.*, 2015) and the MOT 2016 (Milan *et al.*, 2016) benchmarks, while additionally providing fine-grained motion segmentations.

8.2 JOINT MULTICUT PROBLEM FORMULATION

Here, we describe the proposed joint high-level - low-level Minimum Cost Multicut Problem formulation which we want to jointly apply to multi-target tracking and moving object segmentation. Our aim is to build a graphical model representing



(a) high-level and low-level entities (b) proposed graph (c) feasible Multicut

Figure 8.1: (a) While pedestrian detections, here drawn as bounding boxes, represent frame-wise high-level information, point trajectories computed on the same sequence represent spatio-temporal low-level cues. Both can be represented as vertices in a joint graphical model (b). The optimal decomposition of this graph into connected components yields both a motion trajectory segmentation of the sequence as well as the tracking of moving objects represented by the detections (c).

detection and point trajectory nodes and their relationships between one another in a simple, unified way such that the Multicut Problem on this graph directly yields a joint clustering of these high-level and low-level nodes into an optimal number of motion segments and according object tracks.

We define an undirected graph $G = (V, E)$, where $V = \{V^{\text{high}}, V^{\text{low}}\}$ is composed of nodes $v^{\text{high}} \in V^{\text{high}}$ representing high-level entities (detections), and nodes $v^{\text{low}} \in V^{\text{low}}$ representing fine-grained, low-level entities (point trajectories) as depicted in Fig. 8.1 (b).

To represent the three different types of pairwise relations between these nodes, we define three different kinds of edges. The edge set $E = \{E^{\text{high}}, E^{\text{low}}, E^{\text{hl}}\}$ consists of edges $e^{\text{high}} \in E^{\text{high}}$ defining the pairwise relations between detections (depicted in cyan in Fig. 8.1 (b)). These can provide pairwise information computed from strong, very specific object features, reflected in the real-valued edge costs $c_{e^{\text{high}}}$. The edges $e^{\text{low}} \in E^{\text{low}}$ represent pairwise relations between point trajectories (depicted in black in Fig. 8.1 (b)). The according costs $c_{e^{\text{low}}}$ are mostly based on local information. The edges $e^{\text{hl}} \in E^{\text{hl}}$ represent the pairwise relations between these two levels of granularity (depicted in magenta in Fig. 8.1 (b)). The Minimum Cost Multicut Problem on this graph defines a binary *edge* labeling problem:

$$\begin{aligned} &\min_{y \in \{0,1\}^E} \\ &\sum_{e^{\text{high}} \in E^{\text{high}}} c_{e^{\text{high}}} y_{e^{\text{high}}} + \sum_{e^{\text{low}} \in E^{\text{low}}} c_{e^{\text{low}}} y_{e^{\text{low}}} + \sum_{e^{\text{hl}} \in E^{\text{hl}}} c_{e^{\text{hl}}} y_{e^{\text{hl}}} \\ &\text{subject to} \qquad y \in \text{MC}, \end{aligned} \tag{8.1}$$

where MC is the set of exactly all edge labelings $y \in \{0,1\}^E$ that decompose the graph into connected components. Thus, the feasible solutions to the optimization problem from Eq. 8.1 are exactly all *partitionings* of the graph G. In the optimal case, each partition describes either the entire backgroud or exactly one object throughout

the whole video at two levels of granularity: the tracked bounding boxes of this object and the point trajectories of all points on the object. In Fig. 8.1 (c), the proposed solution to the Multicut problem on the graph in Fig. 8.1 (b) contains four clusters: one for each pedestrian tracked over time, and two background clusters in which no detections are contained.

Formally, the feasible set of all multicuts of G can be defined by the cycle inequalities (Chopra and Rao, 1993) $\forall C \in \text{cycles}(G), \forall e \in C: \quad y_e \leq \sum_{e' \in C \setminus \{e\}} y_{e'}$, making the optimization problem APX-hard (Demaine *et al.*, 2006). Yet, the benefit of this formulation is that (1) it contains exactly the right set of feasible solutions, and (2) if p_e denotes the probability of an edge $e \in E$ to be cut, then an optimal solution of the Minimum Cost Multicut Problem with the edge weights computed as $c_e = \text{logit}(p_e) = \log \frac{p_e}{1-p_e}$ is a maximally likely decomposition of G. Note that the *logit* function generates real valued costs c_e such that trivial solutions are avoided.

8.2.1 Pairwise Potentials

In this section, we describe the computation of the pairwise potentials c_e we use in our model. Ideally, one would like to learn terms from training data. However, since the available training datasets for motion segmentation are quite small, we choose to rather define intuitive pairwise terms whose parameters have been validated on training data.

8.2.1.1 *Low-level Nodes and Edges*

In our problem setup, low-level information for motion segmentation and multi-target tracking is built upon point trajectory nodes v^{low} over time and their respective pairwise relations are represented by edge costs $c_{e^{\text{low}}}$.

Low-level Nodes v^{low}: Motion Trajectory Computation A motion trajectory is a spatio-temporal curve that describes the long-term motion of a single tracked point. We compute the motion trajectories according to the method proposed in Ochs *et al.*, 2014. For a given point sampling rate, all points in the first video frame having some underlying image structure are tracked based on large displacement optical flow (Brox and Malik, 2011) until they are occluded or lost.

The decision about ending a trajectory is made by considering the consistency between forward and backward optical flow. In case of large inconsistencies, a point is assumed to be occluded in one of the two frames. Whenever trajectories end, new trajectories are inserted to maintain the desired sampling rate, unless the underlying region is too homogeneous such that accurate point tracking fails.

Trajectory Edge Potentials c_e^{low} The edge potentials $c_{e^{\text{low}}}$ between point trajectories v_i^{low} and v_j^{low} are all computed from low-level image and motion information. Motion distances $d^{\text{m}}(v_i^{\text{low}}, v_j^{\text{low}})$ are computed from the maximum motion difference between

two trajectories during their common life-time as in Ochs *et al.*, 2014. Additionally, we compute color and spatial distances $d^{\text{c}}(v_i^{\text{low}}, v_j^{\text{low}})$ and $d^{\text{sp}}(v_i^{\text{low}}, v_j^{\text{low}})$ between each pair of trajectories with a common life-time and spatial distances for trajectories without temporal overlap as in Keuper *et al.*, 2015a and combine them non-linearly to $z := c_e^{\text{low}} = \max(\bar{\theta}_0 + \theta_1 d^{\text{m}} + \theta_2 d^{\text{c}} + \theta_3 d^{\text{sp}}, \theta_0 + \theta_1 d^{\text{m}})$. The model parameters θ are set as in Keuper *et al.*, 2015a. These costs can be mapped to cut probabilities p_e by the logistic function $p_e = \frac{1}{1+\exp(-z)}$.

8.2.1.2 *High-level Nodes and Edges*

The high-level nodes v^{high} we consider represent object detections. Since these build upon strong underlying object models, the choice of the object detector is task dependent. the faster R-CNN (Ren *et al.*, 2015). Details on the specific detectors and resulting vertex sets V^{high} are given in the experimental section (Sec. 8.3).

We assume, the safest information we can draw from any kind of object detection represented by a node v_i^{high} is its spatio-temporal center position $\text{pos}_{v_i^{\text{high}}} = (x_{v_i^{\text{high}}}, y_{v_i^{\text{high}}}, t_{v_i^{\text{high}}})^\top$ and size $(w_{v_i^{\text{high}}}, h_{v_i^{\text{high}}})^\top$. Ideally, the underlying object model allows to produce a tentative frame-wise object segmentation or template $T_{v_i^{\text{high}}}$ of the detected object. Such a segmentation template can provide far more information than the bounding box alone. Potentially, a template indicates uncertainties and enables to find regions within each bounding box, where points most likely belong to the detected object.

Detection Edge Potentials c_e^{high} Depending on the employed object detector and the specific task, a variety of different object features could be used to compute high-level pairwise potentials. In our setup, we compute the high-level pairwise terms from simple features based on the intersection over union (IoU) of their bounding boxes, their normalized distances d^{sp}

$$d^{\text{sp}}(v_i^{\text{high}}, v_j^{\text{high}}) = 2 \left\| \begin{pmatrix} \frac{x_{v_i^{\text{high}}} - x_{v_j^{\text{high}}}}{w_{v_i^{\text{high}}} + w_{v_j^{\text{high}}}} \\ \frac{y_{v_i^{\text{high}}} - y_{v_j^{\text{high}}}}{h_{v_i^{\text{high}}} + h_{v_j^{\text{high}}}} \end{pmatrix} \right\| \tag{8.2}$$

and their confidence measures.

While for nearby frames, the IoU of two detections can be directly computed from the detection bounding boxes or template masks, this is error prone for larger temporal distances. To introduce robustness, in Chapter 6 we proposed to compute the IoU based on Deep Matching (Weinzaepfel *et al.*, 2013). Deep Matching is based on a deep, multi-layer convolutional architecture and performs dense image patch matching. For every pair of frames t_a and t_b and every detection v_i^{high} in t_a, Deep Matching generates a set of matched keypoints M_{i,t_b} inside the detection. For

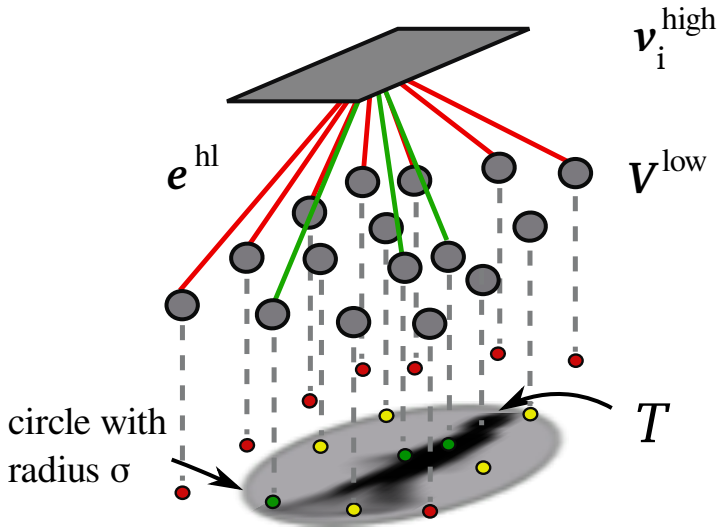


Figure 8.2: Edges e^{hl} between high and low level nodes. For every detection v_i^{high}, an edge with an attractive cost c_e^{hl} is introduced when it hits the according template T (green edges). If the template is not hit and the distance is larger than a threshold σ (indicated by the gray circle), an edge with repulsive edge cost is introduced (red). If the template is not hit but the distance is smaller than σ, no edge is introduced.

every pair of detections v_i^{high} in t_a and v_j^{high} in t_b with $t_a \neq t_b$, we can compute the intersection as $\text{MI}_{ij} = |\text{M}_{i,t_b} \cap \text{M}_{j,ta}|$ and the union as $\text{MU}_{ij} = |\text{M}_{i,t_b} \cup \text{M}_{j,t_a}|$. Then, the Deep Matching based IoU can be computed as

$$\text{IoU}_{ij}^{\text{M}} = \frac{\text{MI}_{ij}}{\text{MU}_{ij}} \tag{8.3}$$

The exact implementation details of the edge potentials c_e^{high} depend on the used detector and will be specified in the experimental section (Sec. 8.3).

8.2.1.3 *Pairwise Potentials c_e^{hl} between High-level and Low-level Nodes*

For point trajectory nodes v_j^{low}, the spatio-temporal location $(x_{v_j^{\text{low}}}^t, y_{v_j^{\text{low}}}^t)^\top$ is the most reliable property. Therefore, we compute pairwise relations between detections and trajectories according to their spatio-temporal relationship, computed from the normalized spatial distance

$$d^{\text{sp}}(v_i^{\text{high}}, v_j^{\text{low}}) = 2 \left\| \begin{pmatrix} \frac{x_{v_i^{\text{high}}} - x_{v_j^{\text{low}}}^t}{w_{v_i^{\text{high}}}} \\ \frac{y_{v_i^{\text{high}}} - y_{v_j^{\text{low}}}^t}{h_{v_i^{\text{high}}}} \end{pmatrix} \right\| \quad \text{for} \quad t = t_{v_i^{\text{high}}} \tag{8.4}$$

and the template value at the trajectory position $T_{v_i^{\text{high}}}(x_{v_j^{\text{low}}}^t, y_{v_j^{\text{low}}}^t)$. If a trajectory passes through a detected object in frame t, it probably belongs to that object. If it

passes far outside the objects bounding box in a certain frame, it is probably not part of this object.

Thus, we compute edge cut probabilities $p_{e^{\text{hl}}}$ from the above described measures as

$$p_{e^{\text{hl}}_{ij}} = \begin{cases} 1 - T_{v_i^{\text{high}}}(x^t_{v_j^{\text{low}}}, y^t_{v_j^{\text{low}}}), & \text{if } T_{v_i^{\text{high}}}(x^t_{v_j^{\text{low}}}, y^t_{v_j^{\text{low}}}) > 0.5 \\ 1, & \text{if } d^{\text{sp}}(v_i^{\text{high}}, v_j^{\text{low}}) > \sigma \\ 0.5, & \text{otherwise} \end{cases} \tag{8.5}$$

using an application dependent threshold σ large enough in order not to conflict with the first case. See Fig. 8.2 for an illustration.

8.2.2 Solving Minimum Cost Multicut Problems

The Minimum Cost Multicut problem defined by the integer linear program in Eq. (8.1) is APX-hard (Demaine *et al.*, 2006). Still, instances of sizes relevant for computer vision can potentially be solved to optimallity or within tight bounds using branch and cut (Andres *et al.*, 2012). However, finding the optimal solution is not necessary for many real world applications. Recently, the primal heuristic proposed by Kernighan and Lin (Kernighan and Lin, 1970) has shown to provide very reasonable results on image and motion segmentation tasks (Keuper *et al.*, 2015b,a). Alternative heuristics were in Beier *et al.*, 2014, 2015. In our experiments, we employ the Kernighan and Lin (Kernighan and Lin, 1970) because of its computation speed and robust behavior.

8.3 EXPERIMENTS

We evaluate the proposed Joint Multicut Formulation on motion segmentation and multi target tracking benchmarks. First, we show our results on the FBMS59 (Ochs *et al.*, 2014) motion segmentation dataset containing sequences with various object categories and motion patterns. Then, the tracking performance is evaluated on the 2D MOT 2015 benchmark (Leal-Taixé *et al.*, 2015) and the MOT 2016 benchmark (Milan *et al.*, 2016) for multi target pedestrian tracking.

8.3.1 Motion Segmentation Dataset

The FBMS59 (Ochs *et al.*, 2014) motion segmentation dataset consists of 59 sequences split into a training set of 29 and a test set of 30 sequences. The videos are of varying length (19 to about 500 frames) and show diverse types of moving objects such as cars, persons and different types of animals.

To exploit the Joint Multicut model for this data, the very first question is how to obtain reliable detections in a video sequence without knowing the category of the object of interest. Here, we evaluate on detections from two different methods

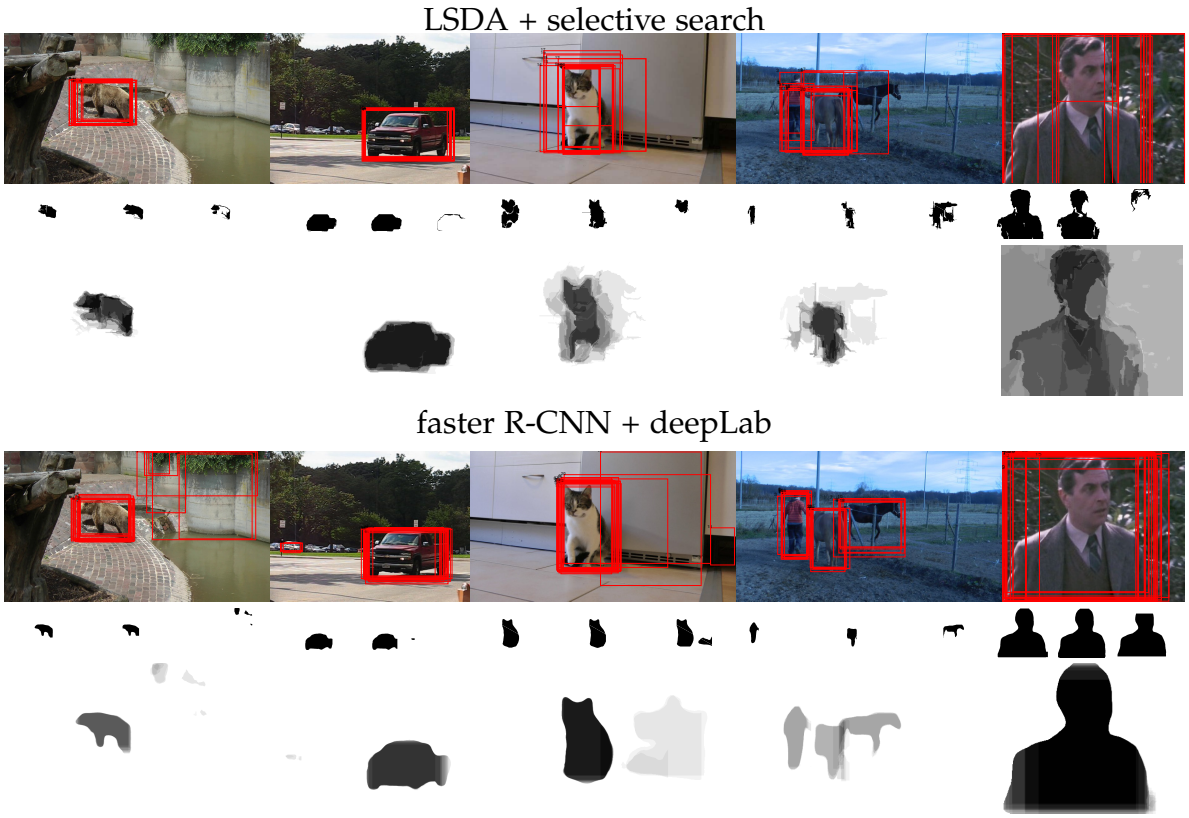
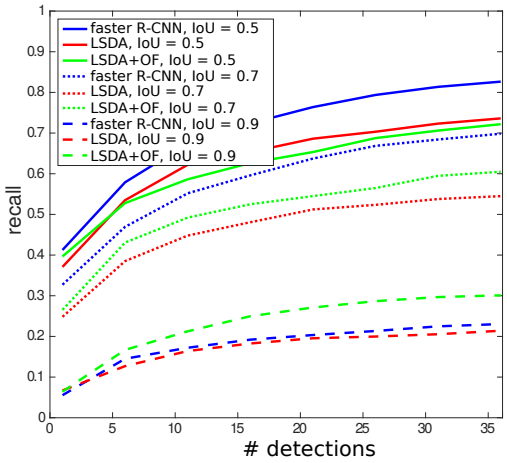


Figure 8.3: Examples of the object detections and according segmentations. Top: LSDA detections on images from FBMS59 sequences (Ochs *et al.*, 2014). The first row shows the best 20 detections. The second row shows three exemplary selective search proposals and third row visualizes the average segmentation of all proposals. Bottom: The corresponding faster R-CNN detections. The first row shows the best 20 detections with a minimum detection score of 0.2. The second row shows three exemplary segmentations from deepLab (Chen *et al.*, 2015; Papandreou *et al.*, 2015) on these detections and third row visualizes the average segmentation.

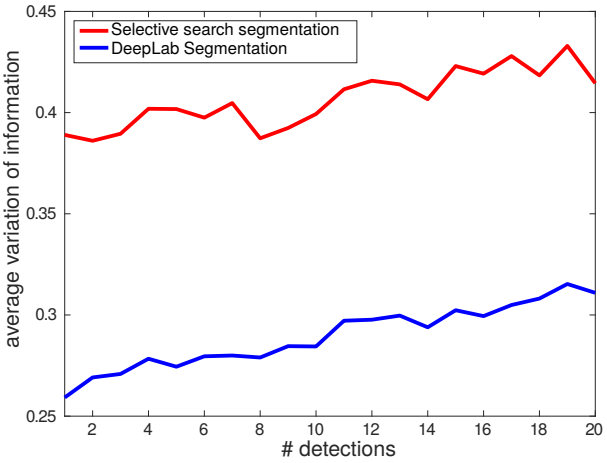
: Large Scale Detection through Adaptation (LSDA) (Hoffman *et al.*, 2014) and the Faster R-CNN (Ren *et al.*, 2015).

Large Scale Detection through Adaptation. The LSDA is a general object detector, trained to detect 7602 object categories (Hoffman *et al.*, 2014). In our experiments, we directly use the code and model deployed with their paper. It operates on a set of object proposals, which is produced by selective search (Uijlings *et al.*, 2013b). The selective search method operates on hierarchical segmentations, which means that we obtain a segmentation mask for each detection bounding box. This segmentation provides a rough spatial and appearance estimation of the object of interest.

To better capture the moving objects in the video, we additionally generate selective search proposals from optical flow images and pass them to the LSDA framework. Example results for the detections and according frame-wise segmentations are given in Fig. 8.3 (top).



(a) The detection performance of LSDA (Hoffman *et al.*, 2014) and faster R-CNN (Ren *et al.*, 2015). We compare the recall for three different IoU thresholds 0.9, 0.7, and 0.5.



(b) The Variation of Information for the proposed object masks over the number of detections (lower is better).

Figure 8.4: Evaluation of the detection and segment proposals on the annotated frames of the FBMS59 (Ochs *et al.*, 2014) training set.

Faster R-CNN. Faster R-CNN is an object detector that integrates a region proposal network with the Fast R-CNN (Girshick, 2015) network. It achieves state-of-the-art object detection accuracy on several benchmark datasets including PASCAL VOC 2012 and MS COCO with only 300 proposals per image (Ren *et al.*, 2015). In our experiments, we directly used the code and model deployed with their work.

On the detections, we generate segmentation proposals using DeepLab (Chen *et al.*, 2015, Papandreou *et al.*, 2015), again by directly using their implementation. Example results for the detections and according frame-wise segmentations are given in Fig. 8.3 (bottom).

Evaluation. Fig. 8.4(a) shows the achieved recall over the number of detections for LSDA (Hoffman *et al.*, 2014) and faster R-CNN (Ren *et al.*, 2015) for different thresholds on the intersection over union (IoU) on the FBMS59 (Ochs *et al.*, 2014) training set. For the higher thresholds, the performance of LSDA is improved when proposals from optical flow images are used (LSDA+OF) and for IoU $\geq$ 0.9, this setup yields best recall. However, for smaller IoU thresholds, faster R-CNN yields highest recall even without considering optical flow. The comparison of the segment mask proposals from selective search (for LSDA) and deepLab (for faster R-CNN) (Fig. 8.4 b(b)) shows the potential benefit of DeepLab. The visual comparison on the examples given in Fig. 8.3 shows that the selective search segmentation proposals selected by LSDA are more diverse than the DeepLab segmentations on the faster R-CNN detection. However, the overall localization quality is worse. We further evaluate detections from both methods in the Joint Multicut model.

Implementation Details. In our graphical model, high-level nodes represent detections from either of the above described methods. For both detectors, we use

	Training set (29 sequences)				**Test set** (30 sequences)			
	P	R	F	O	P	R	F	O
Ochs *et al.*, 2014	85.10%	62.40%	72.0%	17/65	79.61%	60.91%	69.02%	24/69
Ochs and Brox, 2012	81.55%	59.33%	68.68%	16/65	82.11%	64.67%	72.35%	27/69
MCe Keuper *et al.*, 2015a	**86.73%**	73.08%	79.32%	**31/65**	**87.88%**	67.7%	76.48%	25/69
MCe + det. (LSDA)	86.43%	75.79%	80.7617%	31/65	-	-	-	-
JointMulticut (LSDA)	86.43%	75.79%	80.7634%	31/65	87.46%	70.80%	78.25%	29/69
MCe + det. (f. R-CNN)	83.46%	79.46%	81.41%	35/65	-	-	-	-
JointMulticut (f. R-CNN)	84.85%	**80.17%**	**82.44%**	**35/65**	84.52%	**77.36%**	**80.78%**	**35/69**

Table 8.1: Results on the FBMS-59 dataset on training (left) and test set (right). We report **P**: average precision, **R**: average recall, **F**: F-measure and **O**: extracted objects with F $\geq$ 75%. All results are computed for sparse trajectory sampling at 8 pixel distance.

the same setup. First, we select the most confident detections [10]. From those, we discard some detections according to the statistics of their respective segmentations. Especially masks from the selective search proposals sometimes only cover object outlines or leak to the image boundaries. Thus, if such a mask covers less than 20% of its bounding box or more than 60% of the whole image area, the respective detections are not used as nodes in our graph.

The pairwise terms between detections are computed from the IoU and the normalized distances d^{sp} of their bounding boxes

$$d^{\text{sp}}(v_i^{\text{high}}, v_j^{\text{high}}) = 2 \left\| \begin{pmatrix} \frac{x_{v_i^{\text{high}}} - x_{v_j^{\text{high}}}}{w_{v_i^{\text{high}}} + w_{v_j^{\text{high}}}} \\ \frac{y_{v_i^{\text{high}}} - y_{v_j^{\text{high}}}}{h_{v_i^{\text{high}}} + h_{v_j^{\text{high}}}} \end{pmatrix} \right\|,$$

where $\text{pos}_{v_i^{\text{high}}}$, $w_{v_i^{\text{high}}}$, and $h_{v_i^{\text{high}}}$ are defined as in Eq. (8.4). For all pairs of detections within one frame and in neighboring frames, the pseudo cut probability is computed as

$$p_{e_{ij}^{\text{high}}} = \begin{cases} 1 - \frac{1}{1+\exp(20*(0.7-\text{IoU}(v_i^{\text{high}}, v_j^{\text{high}})))}, & \text{if IoU}(v_i^{\text{high}}, v_j^{\text{high}}) > 0.7 \\ \frac{1}{1+\exp(5*(1.2-d^{\text{sp}}(v_i^{\text{high}}, v_j^{\text{high}})))}, & \text{if } d^{\text{sp}}(v_i^{\text{high}}, v_j^{\text{high}}) > 1.2 \\ 0.5, & \text{otherwise} \end{cases} \tag{8.6}$$

The parameters have been set such as to produce reasonable results on the FBMS59 training set. Admittedly, parameter optimization on the training set might further improve our results.

The pairwise terms $c_{e^{\text{hl}}}$ are computed from $p_{e^{\text{hl}}}$ as defined in Eq. (8.6) with $\sigma = 2$. This large threshold accounts for the uncertainty in the bounding box localizations.

[10] above 0.47 for LSDA and 0.97 for faster R-CNN - on a scale between 0 and 1.

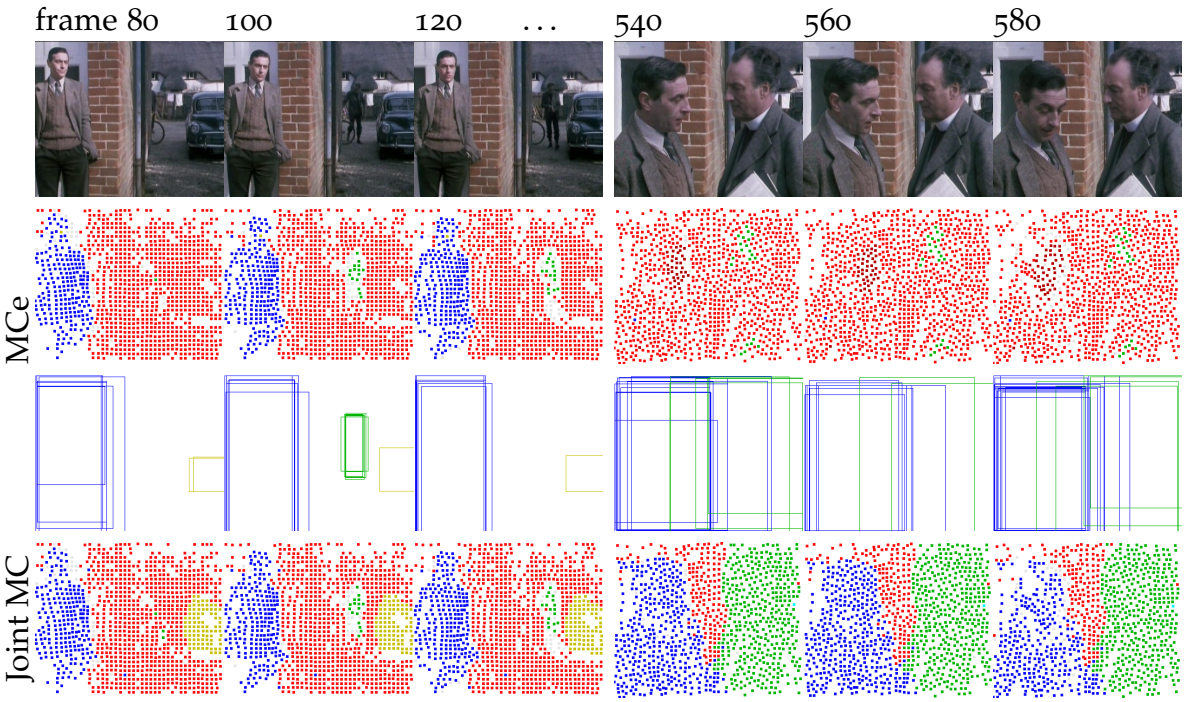


Figure 8.5: Comparison of the proposed Joint Multicut model and the multicut on trajectories (MCe) (Keuper *et al.*, 2015a) on the *marple6* sequence of FBMS59. While with MCe the segmentation breaks between the shown frames, the tracking information from the bounding box subgraph helps our joint model to segment the two men throughout the sequence. Additionally, static, consistently detected objects like the car in the first part of the sequence are segmented as well. As these are not annotated, this causes over-segmentation on the FBMS59 benchmark evaluation.

Results. Our results are given in Tab. 8.1. The motion segmentation considering only the trajectory information from Keuper *et al.*, 2015a performs already well on the FBMS59 benchmark. However, the Joint Multicut model improves over the previous state of the art for both types of object detectors. Note that not only the baseline method of Keuper *et al.*, 2015a is outperformed with quite a margin on the test set - also the motion segmentation based on higher-order potentials (Ochs and Brox, 2012) can not compete with the proposed joint model.

To assess the impact of the joint model components, we evaluate not only the full model but also its performance if pairwise terms between detection nodes are omitted (denoted by MCe + detections). For LSDA detections, this result is pretty close to the Joint Multicut model, implying that the pairwise information we currently employ between the bounding boxes is quite weak. However, for the better localized faster R-CNN detections, the high-level pairwise terms contribute significantly to the overall performance of the joint model.

Qualitative examples of the motion segmentation and object tracking results using the faster R-CNN detections are given in Fig. 8.5 and 8.6. Due to the detection information and the repulsive terms between those object detections and point

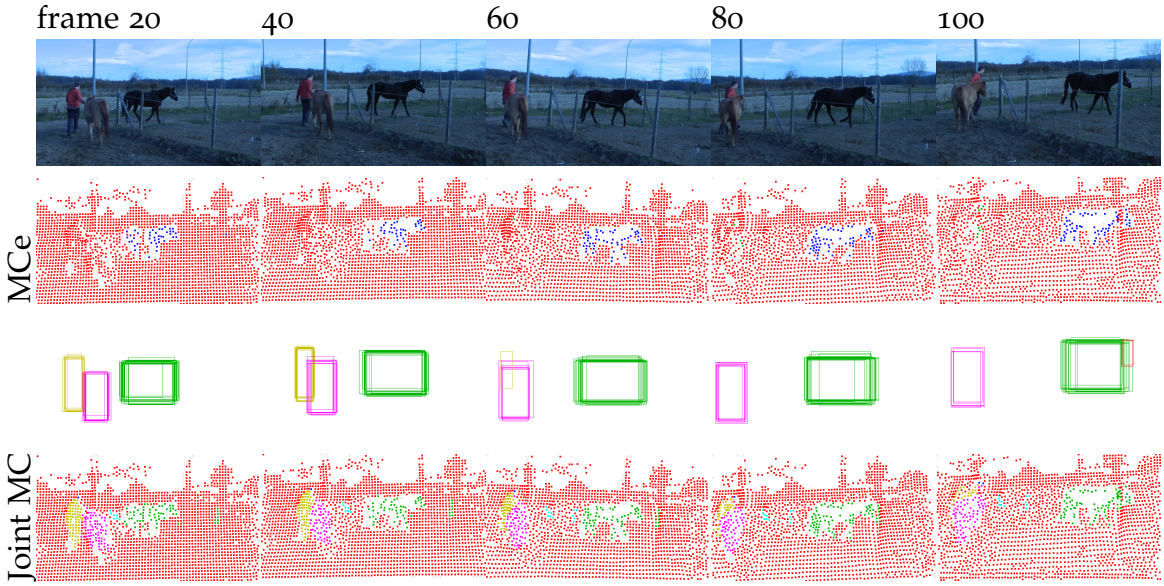


Figure 8.6: Comparison of the proposed Joint Multicut model and the multicut on trajectories (MCe) (Keuper *et al.*, 2015a) on the *horseso6* sequence of FBMS59.

trajectories not passing through them, static objects like the car in the *marple6* sequence (yellow cluster) can be segmented. The man approaching the camera in the same sequence can be tracked and segmented (green cluster) throughout the sequence despite the scaling motion. Similarly, in the *horses* sequence, all three moving objects can be tracked and segmented through strong partial occlusions.

Since the ground truth annotations are sparse and only contain moving objects, this dataset was not used to quantitatively evaluate the multi-target tracking performance.

8.3.2 Multi Target Tracking

Here, we evaluate our Joint Multicut model on the pedestrian tracking task and show the benefit of the Joint Multicut model for the tracking performance. To allow for a comparison to other state-of-the-art multi-target tracking methods, we evaluate our joint multicut approach on 2D MOT 2015 (Leal-Taixé *et al.*, 2015) and MOT 2016 (Milan *et al.*, 2016). Both benchmarks contain videos from static and moving camera recorded in unconstrained environments. MOT15 contains 11 training and 11 test sequences, MOT16 contains 7 sequences each in training and test. In both benchmarks, detections for all sequences are provided and allow for direct comparison to other tracking methods. While the detections in MOT15 are computed using the Aggregate Channel Features pedestrian detector (Dollár *et al.*, 2009), DMP (Felzenszwalb *et al.*, 2010) detections are provided for MOT16.

	MT	ML	FP	FN	IDs	FM	MOTA
Choi, 2015	12.2%	44%	**7,762**	32,547	**442**	823	33.7
Milan *et al.*, 2015	5.8%	63.9%	7,890	39,020	697	**737**	22.5
JointMulticut	**23.2%**	**39.3%**	10,580	**28,508**	457	969	**35.6**

Table 8.2: Multi target tracking results on the 2D MOT 2015 benchmark.

Implementation Details. We link every detection node v_i^{high} to every other detection node v_j^{high} within 3 frames. From a feature vector $f = (\text{IoU}_{ij}^{\text{M}}, \mathit{minConf}, \text{IoU}_{ij}^{\text{M}} \cdot \mathit{minConf}, \text{IoU}_{ij}^{\text{M}^2}, \mathit{minConf}^2)^\top$, where *minConf* is the minimum of the detection scores of v_i^{high} and v_j^{high} and $\text{IoU}_{ij}^{\text{M}}$ is computed according to eq. (8.3), we learn cut propabilities using logistic regression.

The computation of pairwise terms between detections and trajectories is done using an undirected template computed as the average pedestrian shape from the shape prior training data provided in Cremers *et al.*, 2008 and its horizontally flipped analogon.

The cut propability $p(v_i^{\text{high}}, v_j^{\text{low}})$ between a detection v_i^{high} and a trajectory v_j^{low} is computed according to Eq. (8.6) with $\sigma = 1.2$. Since the publicly provided detections are relatively sparse (provided after non-maximum suppression), the statistics of the graph are altered. We compensate for this fact by weighting the costs c_e^{hl} by a constant factor [11]. Detections before non-maximum suppression are unfortunately not provided.

Results. Our tracking performance is evaluated on the official MOT15 (Leal-Taixé *et al.*, 2015) and MOT16 (Milan *et al.*, 2016) benchmarks in terms of the CLEAR MOT evaluation metrics. Here, we report MOTA (multiple object tracking accuracy), which is a cumulative measure combining missed targets (FN), false alarms (FP), and identity switches (IDs), as well as the number of mostly tracked (MT) and mostly lost (ML) objects, and the fragmentation (FM). Results on the MOT15 benchmark are given in Tab. 8.2. Compared to the state-of-the-art multi-target tracking method (Choi, 2015), we have an overall improvement in MOTA. We observe a decrease in the number of false negatives while false positives increase. Also, we show a clear improvement over the performance of the previously proposed method for joint tracking and segmentation (Milan *et al.*, 2015). Our final results on the MOT16 benchmark are given in Tab. 8.3. Here, we also compare to the baseline minimum cost multicuts model proposed in Chapter 6. Our joint model can improve over the baseline in particular by reducing the number of identity switches and fragmentations while keeping the number of false alarms low, resulting in a better MOTA.

Last, we evaluate our sparse segmentations on the pedestrian tracking sequence *tud-crossing* from the MOT15 benchmark. For this sequence, segmentation annotations in every 10th frame have been published by E. Horbert, 2011. The pedestrian

[11] By factor 20 for MOT15 and factor 4 for MOT16.

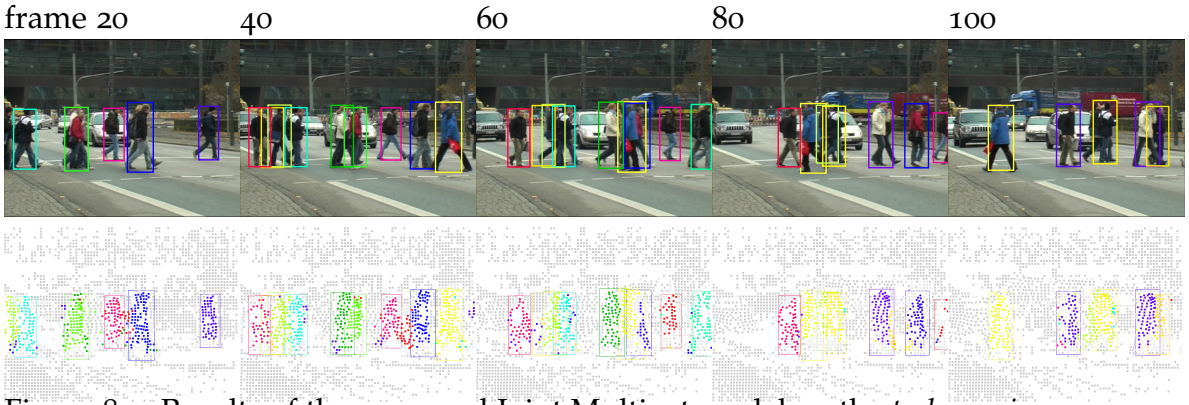


Figure 8.7: Results of the proposed Joint Multicut model on the *tud-crossing* sequence from MOT15.

	MT	ML	FP	FN	IDs	FM	MOTA
Choi, 2015	18.3%	41.4%	9,753	**87,565**	**359**	**504**	46.4
Tang *et al.*, 2016	15.5%	**39.7%**	**6,373**	90,914	657	1.114	46.3
JointMulticut	**20,4%**	46.9%	6,703	89,368	370	598	**47.1**

Table 8.3: Multi target tracking results on the MOT16 benchmark.

motion segmentation is evaluated with the metrics precision (P), recall (R), f-measure (F) and number of retrieved objects (O) as proposed for the FBMS59 motion segmentation benchmark (Ochs *et al.*, 2014).

A qualitative result is given in Fig. 8.7. The bounding boxes overlayed on the image sequence are, for every frame and cluster, the ones with the highest detection score. These were also used for the tracking evaluation. The second row visualizes the trajectory segmentation. Both detection and trajectory clusters look very reasonable. Segmentations provide better localizations for the tracked pedestrians. The quantitative results and a comparison to the motion segmentation methods from Ochs *et al.*, 2014 and Keuper *et al.*, 2015a is given in Tab. 8.4. To assess the importance of the model parts, we not only evaluate the full Joint Model but

TUD-Crossing					
	P	R	F	O ($\geq$ 75)	O ($\geq$ 60)
SC Ochs *et al.*, 2014	67.92	20.16	31.09	0/15	1/15
MCe Keuper *et al.*, 2015a	43.78	38.53	40.99	1/15	1/15
MCe + det.	62.05	54.72	58.15	1/15	**9/15**
MC + traj.	69.37	48.88	57.35	**2/15**	**9/15**
JointMulticut	67.22	55.11	**60.57**	**2/15**	**9/15**

Table 8.4: Motion Segmentation on the Multi-Target Tracking sequence *tud-crossing*. We report **P**: average precision, **R**: average recall, **F**: F-measure (all numbers in %) and **O**: extracted objects with F $\geq$ 75% and with F $\geq$ 60%. All results are computed for sparse trajectory sampling at 8 pixel distance, leading to an average region density of 0.85%.

also the performance of the Multicut formulation when not considering pairwise terms between trajectories (Tracklet MC + traj.) as well as the performance when omitting the pairwise terms between tracklet nodes (MCe + det.). The comparison confirms that the full, joint model performs better than any of its parts. On the important f-measure, the proposed Joint Multicut model improves over the previous state-of-the-art in motion segmentation on this sequence.

8.4 CONCLUSIONS

This chapter proposes a Multicut Model that jointly addresses multi target tracking and motion segmentation so as to leverage the advantages of both. Motion segmentation allows for precise local motion cues and correspondences that support robust multi target tracking results with high recall. Object detection and tracking allows a more reliable grouping of motion trajectories on the same physical object. Promising experimental results are obtained in both domains with a strong improvement over the state of the art in motion segmentation.

9 MULTI PERSON POSE ESTIMATION BY DEEPCUT

Contents

This chapter considers the task of articulated human pose estimation of multiple people in real-world images. We propose an approach that jointly solves the tasks of detection and pose estimation: it infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other. This joint formulation is in contrast to previous strategies, that address the problem by first detecting people and subsequently estimating their body pose. We propose a partitioning and labeling formulation of a set of body-part hypotheses generated with CNN-based part detectors. Our formulation, an instance of an integer linear program, implicitly performs non-maximum suppression on the set of part candidates and groups them to form configurations of body parts respecting geometric and appearance constraints. Experiments on four different datasets demonstrate the effectiveness of our model for both single person and multi person pose estimation.

9.1 INTRODUCTION

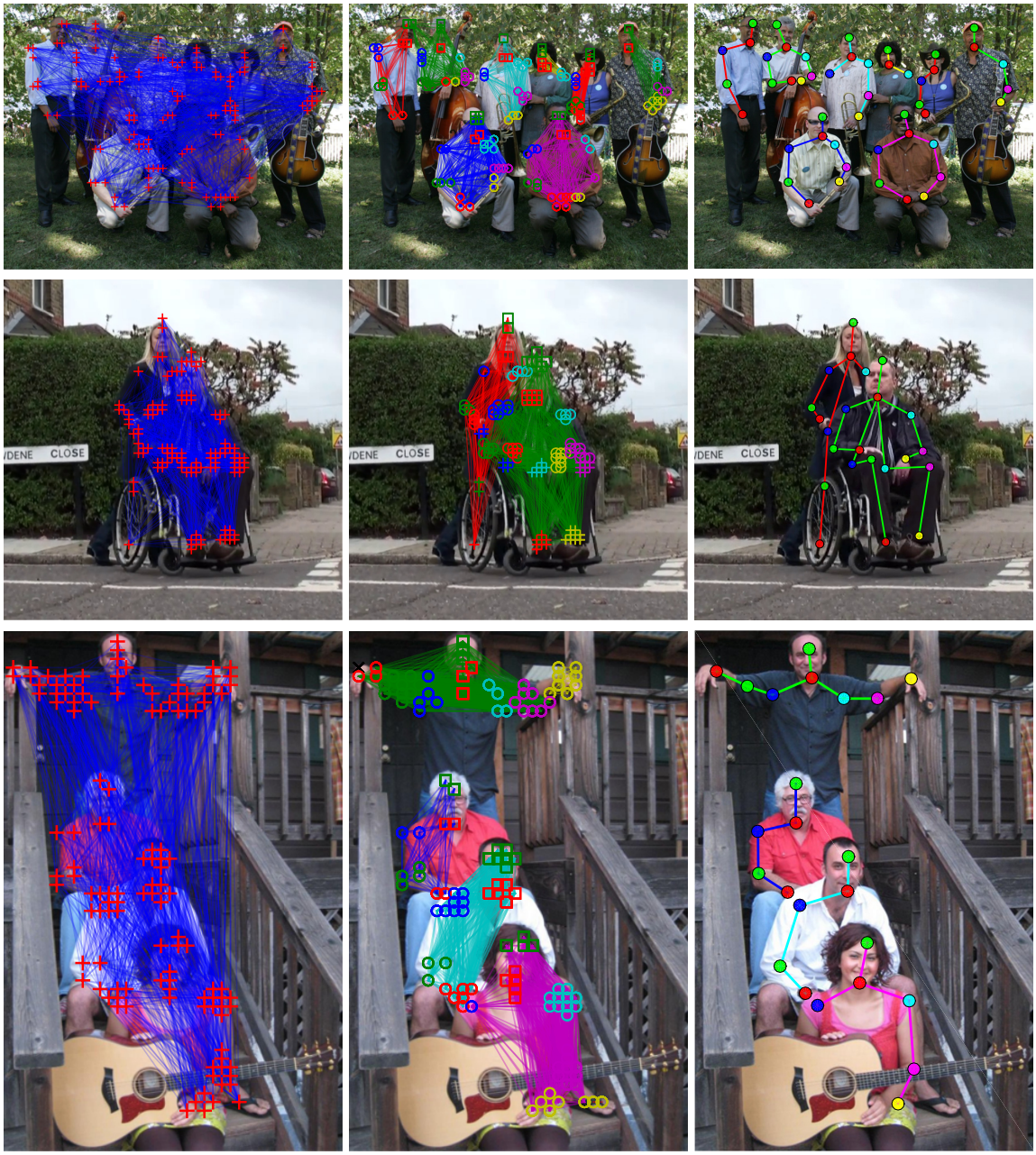
Human body pose estimation methods have become increasingly reliable. Powerful body part detectors (Tompson *et al.*, 2015) in combination with tree-structured body

models (Tompson *et al.*, 2014, Chen and Yuille, 2014) show impressive results on diverse datasets (Johnson and Everingham, 2011, Andriluka *et al.*, 2014, Sapp and Taskar, 2013). These benchmarks promote pose estimation of single pre-localized persons but exclude scenes with multiple persons. This problem definition has been a driver for progress, but also falls short on representing a realistic sample of real-world images. Many photographs contain multiple people of interest (see Fig 9.1) and it is unclear whether single pose approaches generalize directly. We argue that the multi person case deserves more attention since it is an important real-world task.

Key challenges inherent to multi person pose estimation are the partial visibility of some people, significant overlap of bounding box regions of people, and the a-priori unknown number of people in an image. The problem thus is to infer the number of persons, assign part detections to person instances while respecting geometric and appearance constraints. Most strategies use a two-stage inference process (Pishchulin *et al.*, 2012, Gkioxari *et al.*, 2014, Sun and Savarese, 2011) to first detect and then independently estimate poses. This is unsuited for cases when people are in close proximity since they permit simultaneous assignment of the same body-part candidates to multiple people hypotheses.

As a principled solution for multi person pose estimation a model is proposed that jointly estimates poses of all people present in an image by minimizing a joint objective. The formulation is based on partitioning and labeling an initial pool of body part candidates into subsets that correspond to sets of mutually consistent body-part candidates and abide to mutual consistency and exclusion constraints. The proposed method has a number of appealing properties. (1) The formulation is able to deal with an unknown number of people, and also infers this number by linking part hypotheses. (2) The formulation allows to either deactivate or merge part hypotheses in the initial set of part candidates hence effectively performing non-maximum suppression (NMS). In contrast to NMS performed on individual part candidates, the model incorporates evidence from all other parts making the process more reliable. (3) The problem is cast in the form of an Integer Linear Program (ILP). Although the problem is NP-hard, the ILP formulation facilitates the computation of bounds and feasible solutions with a certified optimality gap.

The chapter makes the following contributions. The main contribution is the derivation of a joint detection and pose estimation formulation cast as an integer linear program. Further two CNN variants are proposed to generate representative sets of body part candidates. These, combined with the model, obtain state-of-the-art results for both single-person and multi-person pose estimation on different datasets.



(a) (b) (c)

Figure 9.1: Method overview: (a) initial detections (= part candidates) and pairwise terms (graph) between all detections that (b) are jointly clustered belonging to one person (one colored subgraph = one person) and each part is labeled corresponding to its part class (different colors and symbols correspond to different body parts); (c) shows the predicted pose sticks.

9.2 PROBLEM FORMULATION

In this section, the problem of estimating articulated poses of an unknown number of people in an image is cast as an optimization problem. The goal of this formulation is to state three problems jointly: 1. The selection of a subset of body parts from a set D of *body part candidates*, estimated from an image as described in Section 9.4.1 and depicted as nodes of a graph in Fig. 9.1(a). 2. The *labeling* of each selected body part with one of C *body part classes*, e.g., "arm", "leg", "torso", as depicted in Fig. 9.1(c). 3. The *partitioning* of body parts that belong to the same person, as depicted in Fig. 9.1(b).

9.2.1 Feasible Solutions

We encode labelings of the three problems jointly through triples (x, y, z) of binary random variables with domains $x \in \{0,1\}^{D\times C}, y \in \{0,1\}^{\binom{D}{2}}$ and $z \in \{0,1\}^{\binom{D}{2}\times C^2}$. Here, $x_{dc} = 1$ indicates that body part candidate d is of class c, $y_{dd'} = 1$ indicates that the body part candidates d and d' belong to the same person, and $z_{dd'cc'}$ are auxiliary variables to relate x and y through $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$. Thus, $z_{dd'cc'} = 1$ indicates that body part candidate d is of class c ($x_{dc} = 1$), body part candidate d' is of class c' ($x_{d'c'} = 1$), and body part candidates d and d' belong to the same person ($y_{dd'} = 1$).

In order to constrain the 01-labelings (x, y, z) to well-defined articulated poses of one or more people, we impose the linear inequalities (9.1)–(9.3) stated below. Here, the inequalities (9.1) guarantee that every body part is labeled with at most one body part class. (If it is labeled with no body part class, it is suppressed). The inequalities (9.2) guarantee that distinct body parts d and d' belong to the same person only if neither d nor d' is suppressed. The inequalities (9.3) guarantee, for any three pairwise distinct body parts, d, d' and d'', if d and d' are the same person (as indicated by $y_{dd'} = 1$) and d' and d'' are the same person (as indicated by $y_{d'd''} = 1$), then also d and d'' are the same person ($y_{dd''} = 1$), that is, transitivity, cf. Chopra and Rao, 1993. Finally, the inequalities (9.4) guarantee, for any $dd' \in \binom{D}{2}$ and any $cc' \in C^2$ that $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$. These constraints allow us to write an objective function as a linear form in z that would otherwise be written as a cubic form in x and y. We denote by X_{DC} the set of all (x, y, z) that satisfy all inequalities, i.e., the

set of feasible solutions.

$$\forall d \in D \forall cc' \in \binom{C}{2}: \quad x_{dc} + x_{dc'} \leq 1 \tag{9.1}$$

$$\forall dd' \in \binom{D}{2}: \quad y_{dd'} \leq \sum_{c \in C} x_{dc}$$

$$y_{dd'} \leq \sum_{c \in C} x_{d'c} \tag{9.2}$$

$$\forall dd'd'' \in \binom{D}{3}: \quad y_{dd'} + y_{d'd''} - 1 \leq y_{dd''} \tag{9.3}$$

$$\forall dd' \in \binom{D}{2} \forall cc' \in C^2: \quad x_{dc} + x_{d'c'} + y_{dd'} - 2 \leq z_{dd'cc'}$$

$$z_{dd'cc'} \leq x_{dc}$$

$$z_{dd'cc'} \leq x_{d'c'}$$

$$z_{dd'cc'} \leq y_{dd'} \tag{9.4}$$

When at most one person is in an image, we further constrain the feasible solutions to a well-defined pose of a single person. This is achieved by an additional class of inequalities which guarantee, for any two distinct body parts that are not suppressed, that they must be clustered together:

$$\forall dd' \in \binom{D}{2} \forall cc' \in C^2: \quad x_{dc} + x_{d'c'} - 1 \leq y_{dd'} \tag{9.5}$$

9.2.2 Objective Function

For every pair $(d, c) \in D \times C$, we will estimate a probability $p_{dc} \in [0, 1]$ of the body part d being of class c. In the context of CRFs, these probabilities are called *part unaries* and we will detail their estimation in Section 9.4.

For every $dd' \in \binom{D}{2}$ and every $cc' \in C^2$, we consider a probability $p_{dd'cc'} \in (0, 1)$ of the conditional probability of d and d' belonging to the same person, given that d and d' are body parts of classes c and c', respectively. For $c \neq c'$, these probabilities $p_{dd'cc'}$ are the *pairwise terms* in a graphical model of the human body. In contrast to the classic pictorial structures model, our model allows for a *fully connected graph* where each body part is connected to all other parts in the entire set D by a pairwise term. For $c = c'$, $p_{dd'cc'}$ is the probability of the part candidates d and d' representing the same body part of the same person. This facilitates *clustering* of multiple body part candidates of the same body part of the same person and a *repulsive* property that prevents nearby part candidates of the same type to be associated to different people.

The optimization problem that we call the *subset partition and labeling problem* is the ILP that minimizes over the set of feasible solutions X_{DC}:

$$\min_{(x,y,z) \in X_{DC}} \quad \langle \alpha, x \rangle + \langle \beta, z \rangle, \tag{9.6}$$

where we used the short-hand notation

$$\alpha_{dc} := \log \frac{1 - p_{dc}}{p_{dc}} \tag{9.7}$$

$$\beta_{dd'cc'} := \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} \tag{9.8}$$

$$\langle \alpha, x \rangle := \sum_{d \in D} \sum_{c \in C} \alpha_{dc}\, x_{dc} \tag{9.9}$$

$$\langle \beta, z \rangle := \sum_{dd' \in \binom{D}{2}} \sum_{c,c' \in C} \beta_{dd'cc'}\, z_{dd'cc'} \ . \tag{9.10}$$

9.2.3 Optimization

In order to obtain feasible solutions of the ILP (9.6) with guaranteed bounds, we separate the inequalities (9.1)–(9.5) in the branch-and-cut loop of the state-of-the-art ILP solver Gurobi. More precisely, we solve a sequence of relaxations of the problem (9.6), starting with the (trivial) unconstrained problem. Each problem is solved using the cuts proposed by Gurobi. Once an integer feasible solution is found, we identify violated inequalities (9.1)–(9.5), if any, by breadth-first-search, add these to the constraint pool and re-solve the tightened relaxation. Once an integer solution satisfying all inequalities is found, together with a lower bound that certifies an optimality gap below 1%, we terminate.

9.3 PAIRWISE PROBABILITIES

Here we describe the estimation of the pairwise terms. We define pairwise features $f_{dd'}$ for the variable $z_{dd'cc'}$ (Sec. 9.2). Each part detection d includes the probabilities $f_{p_{dc}}$ (Sec. 9.4.4), its location (x_d, y_d), scale h_d and bounding box B_d coordinates. Given two detections d and d', and the corresponding features $(f_{p_{dc}}, x_d, y_d, h_d, B_d)$ and $(f_{p_{d'c}}, x_{d'}, y_{d'}, h_{d'}, B_{d'})$, we define two sets of auxiliary variables for $z_{dd'cc'}$, one set for $c = c'$ (same body part class clustering) and one for $c \neq c'$ (across two body part classes labeling). These features capture the proximity, kinematic relation and appearance similarity between body parts.

The same body part class ($c = c'$). Two detections denoting the same body part of the same person should be in close proximity to each other. We introduce the following auxiliary variables that capture the spatial relations: $\Delta x = |x_d - x_{d'}|/\bar{h}$, $\Delta y = |y_d - y_{d'}|/\bar{h}$, $\Delta h = |h_d - h_{d'}|/\bar{h}$, $IOUnion$, $IOMin$, $IOMax$. The latter three are intersections over union/minimum/maximum of the two detection boxes, respectively, and $\bar{h} = (h_d + h_{d'})/2$.

Non-linear Mapping. We augment the feature representation by appending quadratic and exponential terms. The final pairwise feature $f_{dd'}$ for the variable $z_{dd'cc}$ is $(\Delta x, \Delta y, \Delta h, IOUnion, IOMin, IOMax, (\Delta x)^2,$
$\ldots, (IOMax)^2, \exp(-\Delta x), \ldots, \exp(-IOMax))$.

Two different body part classes ($c \neq c'$). We encode the kinematic body constraints into the pairwise feature by introducing auxiliary variables $S_{dd'}$ and $R_{dd'}$, where $S_{dd'}$ and $R_{dd'}$ are the Euclidean distance and the angle between two detections, respectively. To capture the joint distribution of $S_{dd'}$ and $R_{dd'}$, instead of using $S_{dd'}$ and $R_{dd'}$ directly, we employ the posterior probability $p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'})$ as pairwise feature for $z_{dd'cc'}$ to encode the geometric relations between the body part class c and c'. More specifically, assuming the prior probability $p(z_{dd'cc'} = 1) = p(z_{dd'cc'} = 0) = 0.5$, the posterior probability of detection d and d' have the body part label c and c', namely $z_{dd'cc'} = 1$, is

$$\begin{aligned} & p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'}) \\ &= \frac{p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1)}{p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1) + p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 0)}, \end{aligned}$$

where $p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1)$ is obtained by conducting a normalized 2D histogram of $S_{dd'}$ and $R_{dd'}$ from positive training examples, analogous to the negative likelihood $p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 0)$. In Sec. 9.5.1 we also experiment with encoding the appearance into the pairwise feature by concatenating the feature $f_{p_{dc}}$ from d and $f_{p_{d'c}}$ from d', as $f_{p_{dc}}$ is the output of the CNN-based part detectors. The final pairwise feature is $(p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'}), f_{p_{dc}}, f_{p_{d'c}})$.

9.3.1 Probability Estimation

The coefficients α and β of the objective function (Eq. 9.6) are defined by the probability ratio in the log space (Eq. 9.7 and Eq. 9.8). Here we describe the estimation of the corresponding probability density: *(1)* For every pair of detection and part classes, namely for any $(d, c) \in D \times C$, we estimate a probability $p_{dc} \in (0, 1)$ of the detection d being a body part of class c. *(2)* For every combination of two distinct detections and two body part classes, namely for any $dd' \in \binom{D}{2}$ and any $cc' \in C^2$, we estimate a probability $p_{dd'cc'} \in (0, 1)$ of d and d' belonging to the same person, meanwhile d and d' are body parts of classes c and c', respectively.

Learning. Given the features $f_{dd'}$ and a Gaussian prior $p(\theta_{cc'}) = \mathcal{N}(0, \sigma^2)$ on the parameters, logistic model is

$$p(z_{dd'cc'} = 1|f_{dd'}, \theta_{cc'}) = \frac{1}{1 + \exp(-\langle \theta_{cc'}, f_{dd'} \rangle)}. \tag{9.11}$$

$(|C| \times (|C| + 1))/2$ parameters are estimated using ML.

Inference. Given two detections d and d', the coefficients α_{dc} for x_{dc} and $\alpha_{d'c}$ for $x_{d'c}$ are obtained by Eq. 9.7, the coefficient $\beta_{dd'cc'}$ for $z_{dd'cc'}$ has the form

$$\beta_{dd'cc'} = \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} = -\langle f_{dd'}, \theta_{cc'} \rangle. \tag{9.12}$$

Model parameters $\theta_{cc'}$ are learned using logistic regression.

9.4 BODY PART DETECTORS

We first introduce our deep learning-based part detection models and then evaluate them on two prominent benchmarks thereby significantly outperforming state of the art.

9.4.1 Adapted Fast R-CNN (*AFR-CNN*)

To obtain strong part detectors we adapt Fast R-CNN (Girshick, 2015). FR-CNN takes as input an image and set of class-independent region proposals (Uijlings *et al.*, 2013a) and outputs the softmax probabilities over all classes and refined bounding boxes. To adapt FR-CNN for part detection we alter it in two ways: *1)* proposal generation and *2)* detection region size. The adapted version is called *AFR-CNN* throughout the chapter.

Detection proposals. Generating object proposals is essential for FR-CNN, meanwhile detecting body parts is challenging due to their small size and high intra-class variability. We use DPM-based part detectors (Pishchulin *et al.*, 2013) for proposal generation. We collect K top-scoring detections by each part detector in a common pool of N part-independent region proposals and use these proposals as input to *AFR-CNN*. N is 2K in case of single and 20K in case of multiple people..

Larger context. Increasing the size of DPM detections by upscaling every bounding box by a fixed factor allows to capture more context around each part. In Sec. 9.4.3 we evaluate the influence of upscaling and show that using larger context around parts is crucial for best performance.

Details. Following standard FR-CNN training procedure ImageNet models are finetuned on pose estimation task. Center of a predicted bounding box is used for body part location prediction. See supplemental for detailed parameter analysis.

9.4.2 Dense architecture (*Dense-CNN*)

Using detection proposals for body part detection may be sub-optimal. We thus develop a fully convolutional architecture for computing part probability scoremaps.

Stride. We use VGG (Simonyan and Zisserman, 2014) as our basis architecture. Converting VGG to fully convolutional mode leads to 32 px stride which is too coarse for precise part localization. We thus use hole algorithm (Chen *et al.*, 2015) to reduce the stride to 8 px.

Scale. Selecting the scale at which CNN is applied is crucial. We empirically found that scaling an image such that an upright standing person is 340 px high leads to

best results. This way 224×224 VGG receptive field sees sufficiently large portion of human to disambiguate body parts.

Loss function. Similar to *AFR-CNN* we start with a softmax loss function that outputs probabilities for each body part and background. The downside is its inability to assign probabilities above 0.5 to several close-by body parts. We thus re-formulate the part detection as multi-label classification problem, where at each location a separate set of probability distributions is estimated for each part. We use sigmoid activation function on the output neurons along with cross entropy loss. We found this loss to perform better than softmax and converge much faster compared to MSE (Tompson *et al.*, 2014). During training a target scoremap is constructed as follows: at each location for each joint a positive label 1 is assigned if the location is within 15 px to the ground truth, and negative label 0 otherwise. Locations with all 0 are the negatives.

Location refinement. While scoremaps provide sufficient resolution, location precision can be improved. Tompson *et al.*, 2015 train additional net to produce fine scoremaps. We follow an alternative and simpler route (Girshick, 2015): we add a location refinement FC layer after the FC7 and use the relative offsets $(\Delta x, \Delta y)$ from a scoremap location to the ground truth as targets.

Regression to other parts. Similar to location refinement we add an extra term to the objective function where for each part we regress onto all other part locations. We empirically found this auxiliary task to improve the unary performance (c.f. Sec. 9.4.3). We envision these predictions to improve the spatial model as well and leave this for the future work.

Training. We follow best practices and use SGD for CNN training. In each iteration we forward-pass a single image. After FC6 we select all positive and random negative samples to keep the pos/neg ratio as 25%/75%. We finetune VGG from Imagenet model to pose estimation task and use training data augmentation. We train for 430k iterations with the following learning rates (lr): 10k at lr=0.001, 180k at lr=0.002, 120k at lr=0.0002 and 120k at lr=0.0001. Pre-training at smaller lr prevents the gradients from diverging.

9.4.3 Evaluation of part detectors

Datasets. For training and evaluation we use three public benchmarks: "Leeds Sports Poses" (LSP) (person-centric (PC) annotations) including 1000 training and 1000 testing images of people doing sports; "LSP Extended" (LSPET) (Johnson and Everingham, 2011) consisting of 10000 training images; "MPII Human Pose" ("Single Person") (Andriluka *et al.*, 2014) consisting of 19185 training and 7247 testing people in every day activities. The MPII training set is used as default. In some cases LSP training *and* LSPET is included, this is denoted as MPII+LSPET in the

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
oracle 2000	98.8	98.8	97.4	96.4	97.4	98.3	97.7	97.8	84.0
DPM scale 1	48.8	25.1	14.4	10.2	13.6	21.8	27.1	23.0	13.6
AlexNet scale 1	82.2	67.0	49.6	45.4	53.1	52.9	48.2	56.9	35.9
AlexNet scale 4	85.7	74.4	61.3	53.2	64.1	63.1	53.8	65.1	39.0
+ optimal params	88.1	79.3	68.9	62.6	73.5	69.3	64.7	72.4	44.6
VGG scale 4 optimal params	91.0	84.2	74.6	67.7	77.4	77.3	72.8	77.9	50.0
+ finetune LSP	**95.4**	**86.5**	**77.8**	**74.0**	**84.5**	**78.8**	**82.6**	**82.8**	**57.0**

Table 9.1: Unary only performance (PCK) of *AFR-CNN* on the LSP (Person-Centric) dataset. *AFR-CNN* is finetuned from ImageNet to MPII (lines 3-6), and then finetuned to LSP (line 7).

experiments. As LSPET has severe labeling noise, all original high-resolution images were re-annotated.

Evaluation measures. We use the standard "Percentage of Correct Keypoints (PCK)" evaluation metric (Sapp and Taskar, 2013; Toshev and Szegedy, 2014; Tompson *et al.*, 2014). We use evaluation scripts available on the web page of (Andriluka *et al.*, 2014) and thus are directly comparable to other methods. In addition to PCK at fixed threshold, we report "Area under Curve" (AUC) computed for the entire range of PCK thresholds.

AFR-CNN. Evaluation of *AFR-CNN* on LSP is shown in Tab. 9.1. Oracle selecting per part the closest from 2000 proposals achieves 97.8% PCK, as proposals cover majority of the ground truth locations. Choosing a single proposal per part using DPM score achieves 23% PCK – not surprising given the difficulty of the body part detection problem. Re-scoring the proposals using *AFR-CNN* with AlexNet (Krizhevsky *et al.*, 2012) dramatically improves the performance to 56.9% PCK, as CNN learns richer image representations. Extending the regions by 4x (1x $\approx$ head size) achieves 65.1% PCK, as it incorporates more context including the information about symmetric parts and allows to implicitly encode higher-order part relations. Using data augmentation and slightly tuning training parameters improves the performance to 72.4% PCK. We refer to the supplementary material for detailed analysis. Deeper VGG architecture improves over smaller AlexNet reaching 77.9% PCK. All results so far are achieved by finetuning the ImageNet models on MPII. Further finetuning to LSP leads to remarkable 82.8% PCK: network learns LSP-specific image representations. Strong increase in AUC (57.0 vs. 50%) is due to improvements for smaller PCK thresholds. No bounding box regression leads to performance drop (81.3% PCK, 53.2% AUC): location refinement is crucial for better part localization. Overall *AFR-CNN* obtains very good results on LSP by far outperforming the state of the art (c.f. Tab. 9.3, rows $7-9$). Evaluation on MPII Single Person shows competitive

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
MPII softmax	91.5	85.3	78.0	72.4	81.7	80.7	75.7	80.8	51.9
+ LSPET	94.6	86.8	79.9	75.4	83.5	82.8	77.9	83.0	54.7
+ sigmoid	93.5	87.2	81.0	77.0	85.5	83.3	79.3	83.8	55.6
+ location refinement	95.0	88.4	81.5	76.4	88.0	83.3	80.8	84.8	61.5
+ auxiliary task	95.1	89.6	82.8	78.9	89.0	85.9	81.2	86.1	61.6
+ finetune LSP	**97.2**	**90.8**	**83.0**	**79.3**	**90.6**	**85.6**	**83.1**	**87.1**	**63.6**

Table 9.2: Unary only performance (PCK) of *Dense-CNN* VGG on LSP (PC) dataset. *Dense-CNN* is finetuned from ImageNet to MPII (line 1), to MPII+LSPET (lines 2-5), and finally to LSP (line 6).

performance (Tab. 9.4, row 1).

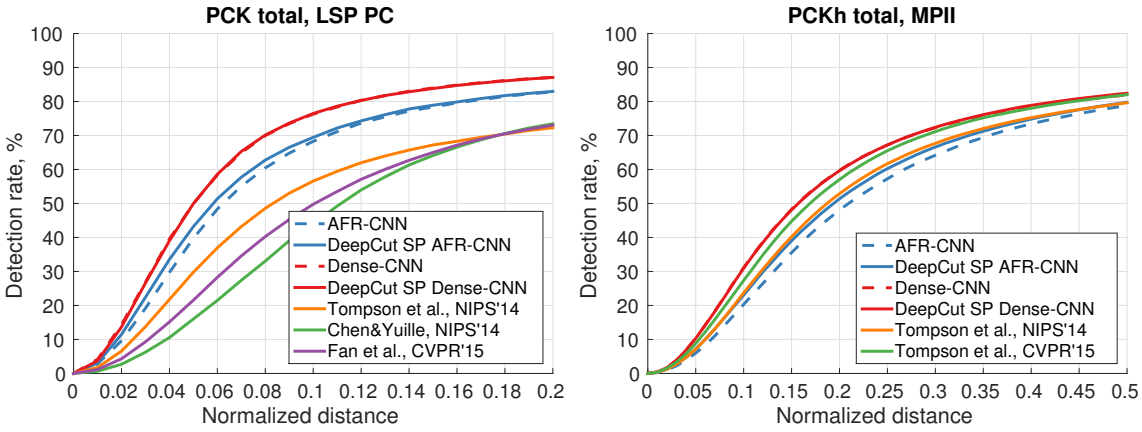
Dense-CNN. The results are in Tab. 9.2. Training with VGG on MPII with softmax loss achieves 80.8% PCK thereby outperforming *AFR-CNN* (c.f. Tab. 9.1, row 6). This shows the advantages of fully convolutional training and evaluation. Expectedly, training on larger MPII+LSPET dataset improves the results (83.0 vs. 80.8% PCK). Using cross-entropy loss with sigmoid activations improves the results to 83.8% PCK, as it better models the appearance of close-by parts. Location refinement improves localization accuracy (84.8% PCK), which becomes more clear when analyzing AUC (61.5 vs. 55.6%). Interestingly, regressing to other parts further improves PCK to 86.1% showing a value of training with the auxiliary task. Finally, finetuning to LSP achieves the best result of 87.1% PCK, which is significantly higher than the best published results (c.f. Tab. 9.3, rows $7-9$). Unary-only evaluation on MPII reveals slightly higher AUC results compared to the state of the art (Tab. 9.4, row $3-4$).

9.4.4 Using detections in DeepCut models

The SPLP problem is NP-hard, to solve instances of it efficiently we select a subset of representative detections from the entire set produced by a model. In our experiments we use $|D| = 100$ as default detection set size. In case of the *AFR-CNN* we directly use the softmax output as unary probabilities: $f_{p_{dc}} = (p_{d1}, \ldots, p_{dc})$, where p_{dc} is the probability of the detection d being the part class c. For *Dense-CNN* detection model we use the sigmoid detection unary scores.

9.5 DEEPCUT RESULTS

The aim of this work is to tackle the multi-person case. To that end, we evaluate the proposed *DeepCut* models on four diverse benchmarks. We confirm that both single person (*SP*) and multi-person (*MP*) variants (Sec. 9.2) are effective on standard *SP* pose estimation datasets (Andriluka *et al.*, 2014). Then, we demonstrate superior performance of *DeepCut MP* on the multi-person pose estimation task.



(a) LSP (PC) (b) MPII Single Person

Figure 9.2: Pose estimation results over all PCK thresholds.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AFR-CNN (unary)	95.4	86.5	77.8	74.0	84.5	82.6	78.8	82.8	57.0
+ *DeepCut SP*	95.4	86.7	78.3	74.0	84.3	82.9	79.2	83.0	58.4
+ appearance pairwise	95.4	87.2	78.6	73.7	84.7	82.8	78.8	83.0	58.5
+ *DeepCut MP*	95.2	86.7	78.2	73.5	84.6	82.8	79.0	82.9	58.0
Dense-CNN (unary)	**97.2**	90.8	83.0	**79.3**	90.6	85.6	**83.1**	**87.1**	**63.6**
+ *DeepCut SP*	97.0	91.0	**83.8**	78.1	91.0	86.7	82.0	**87.1**	63.5
+ *DeepCut MP*	96.2	**91.2**	83.3	77.6	**91.3**	**87.0**	80.4	86.7	62.6
Tompson *et al.*, 2014	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3
Chen and Yuille, 2014	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	40.1
X. Fan and Wang, 2015*	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2

* re-evaluated using the standard protocol, for details see project page of X. Fan and Wang, 2015

Table 9.3: Pose estimation results (PCK) on LSP (PC) dataset.

9.5.1 Single person pose estimation

We now evaluate single person (*SP*) and more general multi-person (*MP*) *DeepCut* models on LSP and MPII *SP* benchmarks described in Sec. 9.4. Since this evaluation setting implicitly relies on the knowledge that all parts are present in the image we always output the full number of parts.

Results on LSP. We report per-part PCK results (Tab. 9.3) and results for a variable distance threshold (Fig. 9.2 (a)). *DeepCut SP AFR-CNN* model using 100 detections improves over unary only (83.0 vs. 82.8% PCK, 58.4 vs. 57% AUC), as pairwise connections filter out some of the high-scoring detections on the background. The improvement is clear in Fig. 9.2 (a) for smaller thresholds. Using part appearance scores in addition to geometrical features in $c \neq c'$ pairwise terms only slightly

improves AUC, as the appearance of neighboring parts is mostly captured by a relatively large region centered at each part. As geometrical only pairwise lead to faster experiments. The performance of *DeepCut MP AFR-CNN* matches the *SP* and improves over *AFR-CNN* alone: *DeepCut MP* correctly handles the *SP* case. Performance of *DeepCut SP Dense-CNN* is almost identical to unary only, unlike the results for *AFR-CNN*. *Dense-CNN* performance is noticeably higher compared to *AFR-CNN*, and "easy" cases that could have been corrected by a spatial model are resolved by stronger part detectors alone.

Comparison to the state of the art (LSP). Tab. 9.3 compares results of *DeepCut* models to other deep learning methods specifically designed for single person pose estimation. All *DeepCuts* significantly outperform the state of the art, with *DeepCut SP Dense-CNN* model improving by 13.7% PCK over the best known result (Chen and Yuille, 2014). The improvement is even more dramatic for lower thresholds (Fig. 9.2 (a)): for PCK @ 0.1 the best model improves by 19.9% over Tompson et al. (Tompson *et al.*, 2014), by 26.7% over Fan et al. (X. Fan and Wang, 2015), and by 32.4% PCK over Chen&Yuille (Chen and Yuille, 2014). The latter is interesting, as (Chen and Yuille, 2014) use a stronger spatial model that predicts the pairwise conditioned on the CNN features, whereas *DeepCuts* use geometric-only pairwise connectivity. Including body part orientation information into *DeepCuts* should further improve the results.

Results on MPII Single Person. Results are shown in Tab. 9.4 and Fig. 9.2 (b). *DeepCut SP AFR-CNN* noticeably improves over *AFR-CNN* alone (79.8 vs. 78.8% PCK, 51.1 vs. 49.0% AUC). The improvement is stronger for smaller thresholds (c.f. Fig. 9.2), as spatial model improves part localization. *Dense-CNN* alone trained on MPII outperforms *AFR-CNN* (81.6 vs. 78.8% PCK), which shows the advantages of dense training and evaluation. As expected, *Dense-CNN* performs slightly better when trained on the larger MPII+LSPET. Finally, *DeepCut Dense-CNN SP* is slightly better than *Dense-CNN* alone leading to the best result on MPII dataset (82.4% PCK).

Comparison to the state of the art (MPII). We compare the performance of *DeepCut* models to the best deep learning approaches from the literature (Tompson *et al.*, 2014, 2015)[12]. *DeepCut SP Dense-CNN* outperforms both (Tompson *et al.*, 2014, 2015) (82.4 vs 79.6 and 82.0% PCK, respectively). Similar to them *DeepCuts* rely on dense training and evaluation of part detectors, but unlike them use single size receptive field and do not include multi-resolution context information. Also, appearance and spatial components of *DeepCuts* are trained piece-wise, unlike (Tompson *et al.*, 2014). We observe that performance differences are higher for smaller thresholds (c.f. Fig. 9.2 (b)). This is remarkable, as a much simpler strategy for location refinement is used compared to (Tompson *et al.*, 2015). Using multi-resolution filters and joint training should improve the performance.

[12]Tompson *et al.*, 2014 was re-trained and evaluated on MPII dataset by the authors.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK_h	AUC
AFR-CNN (unary)	91.5	89.7	80.5	74.4	76.9	69.6	63.1	78.8	49.0
+ *DeepCut SP*	92.3	90.6	81.7	74.9	79.2	70.4	63.0	79.8	51.1
Dense-CNN (unary)	93.5	88.6	82.2	77.1	81.7	74.4	**68.9**	81.6	56.0
+LSPET	94.0	89.4	82.3	77.5	82.0	74.4	68.7	81.9	**56.5**
+*DeepCut SP*	94.1	90.2	83.4	77.3	**82.6**	**75.7**	68.6	**82.4**	**56.5**
Tompson *et al.*, 2014	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6	51.8
Tompson *et al.*, 2015	**96.1**	**91.9**	**83.9**	**77.8**	80.9	72.3	64.8	82.0	54.9

Table 9.4: Pose estimation results (PCK_h) on MPII Single Person.

9.5.2 Multi-person pose estimation

We now evaluate *DeepCut MP* models on the challenging task of *MP* pose estimation with an unknown number of people per image and visible body parts per person.

Datasets. For evaluation we use two public *MP* benchmarks: "We Are Family" (WAF) (Eichner and Ferrari, 2010) with 350 training and 175 testing group shots of people; "MPII Human Pose" ("Multi-Person") (Andriluka *et al.*, 2014) consisting of 3844 training and 1758 testing images of multiple interacting individuals in highly articulated poses with variable number of parts. We use a representative subset of 288 testing images for evaluation. We first pre-finetune both *AFR-CNN* and *Dense-CNN* from ImageNet to MPII and MPII+LSPET, respectively, and further finetune each model to WAF and MPII Multi-Person. For WAF, we re-train the spatial model on WAF training set.

WAF evaluation measure. Approaches are evaluated using the official toolkit (Eichner and Ferrari, 2010), thus results are directly comparable to prior work. The toolkit implements occlusion-aware "Percentage of Correct Parts (*m*PCP)" metric. In addition, we report "Accuracy of Occlusion Prediction (AOP)" (Chen and Yuille, 2015).

MPII Multi-Person evaluation measure. PCK metric is suitable for *SP* pose estimation with known number of parts and does not penalize for false positives that are not a part of the ground truth. Thus, for *MP* pose estimation we use "Mean Average Precision (mAP)" measure, similar to (Sun and Savarese, 2011; Yang and Ramanan, 2013). In contrast to (Sun and Savarese, 2011; Yang and Ramanan, 2013) evaluating the detection of *any* part instance in the image disrespecting inconsistent pose predictions, we evaluate consistent part configurations. First, multiple body pose predictions are generated and then assigned to the ground truth (GT) based on the highest PCK_h (Andriluka *et al.*, 2014). Only single pose can be assigned to GT. Unassigned predictions are counted as false positives. Finally, AP for each body part is computed and mAP is reported.

Setting	Head	U Arms	L Arms	Torso	*m*PCP	AOP
AFR-CNN det ROI	69.8	46.0	36.7	83.7	53.1	73.9
DeepCut MP AFR-CNN	99.0	79.5	74.3	87.1	82.2	85.6
Dense-CNN det ROI	76.0	46.0	40.2	83.7	55.3	73.8
DeepCut MP Dense-CNN	**99.3**	**81.5**	**79.5**	87.1	**84.7**	**86.5**
Ghiasi *et al.*, 2014	-	-	-	-	63.6	74.0
Eichner and Ferrari, 2010	97.6	68.2	48.1	86.1	69.4	80.0
Chen and Yuille, 2015	98.5	77.2	71.3	**88.5**	80.7	84.9

Table 9.5: Pose estimation results (*m*PCP) on WAF dataset.

Baselines. To assess the performance of *AFR-CNN* and *Dense-CNN* we follow a traditional route from the literature based on two stage approach: first a set of regions of interest (*ROI*) is generated and then the *SP* pose estimation is performed in the *ROIs*. This corresponds to unary only performance by *DeepCuts*. *ROI* are either based on a ground truth (*GT ROI*) or on the people detector output (*det ROI*).

Results on WAF. Results are shown in Tab. 9.5. *det ROI* is obtained by extending provided upper body detection boxes. *AFR-CNN det ROI* achieves 57.6% *m*PCP and 73.9% AOP. *DeepCut MP AFR-CNN* significantly improves over *AFR-CNN det ROI* achieving 82.2% *m*PCP. This improvement is stronger compared to LSP and MPII due to several reasons. First, *m*PCP requires consistent prediction of body sticks as opposite to body joints, and including spatial model enforces consistency. Second, *m*PCP metric is occlusion-aware. *DeepCuts* can deactivate detections for the occluded parts thus effectively reasoning about occlusion. This is supported by strong increase in AOP (85.6 vs. 73.9%). Results by *DeepCut MP Dense-CNN* follow the same tendency achieving the best performance of 84.7% *m*PCP and 86.5% AOP. Both increase in *m*PCP and AOP show the advantages of *DeepCuts* over traditional *det ROI* approaches.

Tab. 9.5 shows that *DeepCuts* outperform all prior methods. Deep learning method (Chen and Yuille, 2015) is outperformed both for *m*PCP (84.7 vs. 80.7%) and AOP (86.5 vs. 84.9%) measures. This is remarkable, as *DeepCuts* reason about part interactions across several people, whereas (Chen and Yuille, 2015) primarily focuses on the single-person case and handles multi-person scenes akin to (Yang and Ramanan, 2013). In contrast to (Chen and Yuille, 2015), *DeepCuts* are not limited by the number of possible occlusion patterns and cover person-person occlusions and other types as truncation and occlusion by objects in one formulation. *DeepCuts* significantly outperform (Eichner and Ferrari, 2010) while being more general: unlike (Eichner and Ferrari, 2010) *DeepCuts* do not require person detector and not limited by a number of occlusion states among people. Qualitative comparison to (Chen and Yuille, 2015) is provided in Fig. 9.3.

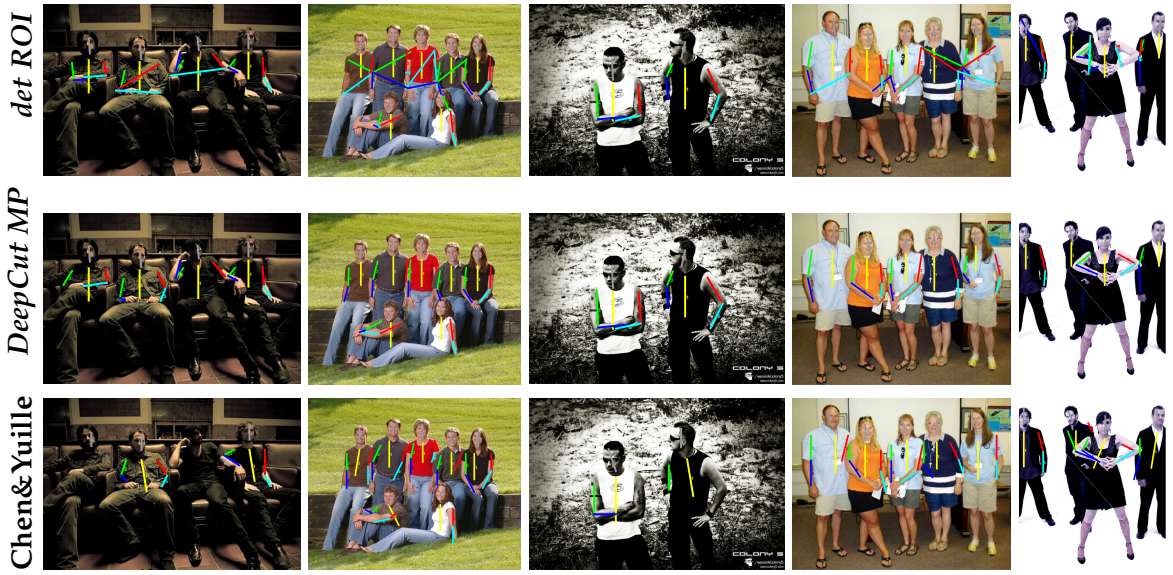


Figure 9.3: Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (middle) to the traditional two-stage approach *Dense-CNN det ROI* (top) and the approach of Chen&Yuille Chen and Yuille, 2015 (bottom) on WAF dataset. In contrast to *det ROI*, *DeepCut MP* is able to disambiguate multiple and potentially overlapping persons and correctly assemble independent detections into plausible body part configurations. In contrast to Chen and Yuille, 2015, *DeepCut MP* can better predict occlusions (image 2 person $1-4$ from the left, top row; image 4 person 1, 4; image 5, person 2) and better cope with strong articulations and foreshortenings (image 1, person 1, 3; image 2 person 1 bottom row; image 3, person 1-2). See supplementary material for more examples.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	UBody	FBody
AFR-CNN det ROI	71.1	65.8	49.8	34.0	47.7	36.6	20.6	55.2	47.1
AFR-CNN MP	71.8	67.8	54.9	38.1	52.0	41.2	30.4	58.2	51.4
AFR-CNN MP UB	75.2	71.0	56.4	39.6	-	-	-	60.5	-
Dense-CNN det ROI	77.2	71.8	55.9	42.1	53.8	39.9	27.4	61.8	53.2
Dense-CNN MP	73.4	71.8	57.9	39.9	**56.7**	**44.0**	**32.0**	60.7	**54.1**
Dense-CNN MP UB	**81.5**	**77.3**	**65.8**	**50.0**	-	-	-	**68.7**	-
AFR-CNN GT ROI	73.2	66.5	54.6	42.3	50.1	44.3	37.8	59.1	53.1
Dense-CNN GT ROI	78.1	74.1	62.2	52.0	56.9	48.7	46.1	66.6	60.2
Chen&Yuille SP GT ROI	65.0	34.2	22.0	15.7	19.2	15.8	14.2	34.2	27.1

Table 9.6: Pose estimation results (AP) on MPII Multi-Person.

Results on MPII Multi-Person. Obtaining a strong detector of highly articulated people having strong occlusions and truncations is difficult. We employ a neck detector as a person detector as it turned out to be the most reliable part. Full body bounding box is created around a neck detection and used as *det ROI*. *GT ROI*s were provided by the authors (Andriluka *et al.*, 2014). As the *MP* approach (Chen and Yuille, 2015) is not public, we compare to *SP* state-of-the-art method (Chen and Yuille, 2014) applied to *GT ROI* image crops.

As shown in Tab. 9.6. *DeepCut MP AFR-CNN* improves over *AFR-CNN det ROI* by 4.3% achieving 51.4% AP. The largest differences are observed for the ankle, knee, elbow and wrist, as those parts benefit more from the connections to other parts. *DeepCut MP UB AFR-CNN* using upper body parts only slightly improves over the full body model when compared on common parts (60.5 vs 58.2% AP). Similar tendencies are observed for *Dense-CNN*s, though improvements of *MP UB* over *MP* are more significant.

All *DeepCuts* outperform *Chen&Yuille SP GT ROI*, partially due to stronger part detectors compared to (Chen and Yuille, 2014) (c.f. Tab. 9.3). Another reason is that *Chen&Yuille SP GT ROI* does not model body part occlusion and truncation always predicting the full set of parts, which is penalized by the AP measure. In contrast, our formulation allows to deactivate the part hypothesis in the initial set of part candidates thus effectively performing non-maximum suppression. In *DeepCuts* part hypotheses are suppressed based on the evidence from all other body parts making this process more reliable.

9.6 CONCLUSIONS

Articulated pose estimation of multiple people in uncontrolled real world images is challenging but of real world interest. In this work, we proposed a new formulation as a joint subset partitioning and labeling problem (SPLP). Different to previous two-stage strategies that separate the detection and pose estimation steps, the SPLP

model jointly infers the number of people, their poses, spatial proximity, and part level occlusions. Empirical results on four diverse and challenging datasets show significant improvements over all previous methods not only for the multi-person, but also for the single-person pose estimation problem. On multi-person WAF dataset we improve by 30% PCP over the traditional two-stage approach. This shows that a joint formulation is crucial to disambiguate multiple and potentially overlapping persons.

CONCLUSIONS AND FUTURE PERSPECTIVES 10

Contents

In this chapter, we summarize the contributions of this thesis and discuss potential directions for future work.

10.1 DISCUSSION OF CONTRIBUTIONS

Visual understanding of people in unconstrained monocular images and videos, especially street scene videos, has been extensive studied and significantly advanced over the past years. Still machine perception is far below human quality level. In order to handle the complexity of this problem, it has been decomposed into well defined and highly correlated sub-problems, such as people detection, tracking, human pose estimation. The focus of this thesis is to propose novel algorithms and models for each sub-problems, as well as to bridge the gaps between them by proposing joint formulations that simultaneously solve two or more sub-tasks. In the following, we summarize the challenges of visual understanding of people in realistic images and videos and summarize the contributions of thesis.

Occlusion handling is a well-known challenge for **people detection** in crowded scenes and generic solutions are far from being available. People detectors such as the deformable part models (Felzenszwalb *et al.*, 2010) and the faster RCNN (Ren *et al.*, 2015) have demonstrated good detection results on challenging datasets. However, the performance degrades quickly in the presence of heavy occlusions. Such detectors treat occlusions as distractions or nuisance. The dominant occlusions in crowed street scenes are due to overlaps between people. Our intuition is that person/person occlusion patterns are characteristic and could be explicitly used to detect the presence of occlusions. Localizing the person in front facilitates the localization of the occluded person and vice versa.

As our first contribution towards people detection, we propose joint detection models that are trained to detect single person as well as pairs of people under varying degrees of occlusion. As presented in Chapter 3, the joint detectors significantly improve over single-person detectors for detecting people in crowded street

scenes, without losing performance on images with one person only. To deal with the difficulties of obtaining sufficient training data for varying degrees of occlusion, we propose to generate synthetic training images. The focus of Chapter 3 is to detect people in the side-view images, where the configuration of occlusion patterns can be characterized by the relative scale, postion of the occluding and occluded people. The results are very promising and suggest the potential of leveraging characteristic appearance patterns of person/person pairs also for detecting occluded people in more general settings. However, the generalization of this idea to crowded scenes with people walking in arbitrary directions is rather challenging due to the vast amount of possible person-person occlusion situations. This variation may arise from several factors, such as people's body articulation, or their position and orientation relative to the camera. The number of putative occlusion patterns is exponential in the number of factors. The crucial point here is, however, that not all of them are equally relevant for successful tracking. Therefore, finding occlusion patterns that are relevant in practice in order to reduce the modeling space is essential for applying joint person detectors for tracking in general crowded scenes. As our second contribution towards people detection, the joint models proposed in Chapter 3 are further extended in Chapter 4 where the learning of the joint models is designed to address common failure cases during tracking due to long-term inter-object occlusions. The reoccurring constellations of person/person occlusions are learned with a multi-person tracker in the loop. The tighter integration of tracker and detector that is proposed in the thesis improves tracking results on several challenging benchmark sequences.

Tracking multiple people in a sequence of images is often modelled as a data association problem. Given person detections in each frame, by far the most common approach is to define a graph whose nodes represent the detections and edges link detections that hypothetically describe the same person. With respect to a linear objective function, the tracks of multiple people can be obtained by solving a **Minimum Cost Disjoint Paths Problem**. Disjoint paths mean that the paths do not branch or merge, which is rather intuitive as a single person can not either occupy the same physical space or split into two persons. However, typical people detectors generate many similar detections for the same person. Non-maximum Suppression (NMS) is often applied to pre-select a single detection for each target. Such approaches have a notable caveat: NMS could remove the detections for partially occluded person. Acknowledging the fact that target detectors produce multiple equally plausible detections per target and frame, we propose a novel mathematical abstraction for multi person tracking, namely, **the Minimum Cost Subgraph Multicut Problem** (Chapter 5). As our first contribution towards multi person tracking, the Minumum Cost Subgraph Multicut model is superior to the conventional Minimum Cost Disjoint Paths formation in the way that the Subgraph Multicut model selects and clusters all suitable hypotheses for each target jointly in space and time, eliminating the needs for the heuristic Non-maximum Suppression. Besides, the number of persons is not fixed or biased by definition of the problem, but is estimated in an unbiased fashion from the video sequence and determined by

the solution of the problem. A notable challenge for tracking in crowded scenes is to associate detections before and after long-term occlusions. One trend in the research area of tracking is to utilize the deep person re-identification networks to effectively model the relations between detections that are far in time. However, incorporating such long-term information into tracking algorithms is not trivial. Because, one has to model the fact that similar looking people are not necessarily identical. As our second contribution towards multi person tracking, we propose another novel formulation for multi person tracking, namely, **Minimum Cost Lifted Multicut Problem** (Chapter 7). It has the advantage that distinction can be made between the edges that define possible connections and the edges that define the costs or rewards for assigning the incident nodes to distinct tracks. This allows us to avoid assigning the distinct but similar looking people to the same track. The Minimum Cost Lifted Multicut Model encodes the long-range person re-identification information, and at the same time penalizes long-term false joints by forcing valid paths along the regular edges in a rigorous manner. The tracking model achieves the top performance on the challenging MOT16 benchmark.

The tracking summarized in the previous paragraph requires previously learned semantic knowledge of the target object category and is therefore traditionally approached from a high-level perspective. Motion segmentation, which aims at tracking pixels in a sequence of images, could be considered as a low-level visual understanding of the motion in a video. These two problems are highly related in the sense that their goal is to determine the image regions that belong to the same object in the video. In Chapter 8, we further explore visual tracking in these two different granularities. More specifically, we aim to understand the motion of the scene and the objects in the scene both in the pixel trajectory level and the detection bounding box level. To that end, we propose a joint multicut model that simultaneously addresses **multi object tracking and motion segmentation** so as to leverage the advantages of both. The pixel trajectories carry the information that how each single, well localized points are moving and can be used to disambiguate partial occlusion and motion speed changes. Therefore, motion segmentation allows for precise local motion cues and correspondences that support robust multi-object tracking results with high recall. Object detection and tracking allows a more reliable grouping of motion trajectories on the same physical object. These high-level cues also provide the information about the rough object location and re-identify the object after occlusion. The experimental results are obtained in both domains with a strong improvement over the state of the art in motion segmentation.

Towards the goal of a richer understanding of people in realistic images and videos, we further extend the multicut tracking model for **multi person pose estimation** in Chapter 9. By far the most common approach for the multi person pose estimation task is to separate the detection and pose estimation steps. This is unsuited for the cases when people are in close proximity. The detection errors are inherently propagated into the pose estimation step. In contrast, we propose to model the task as a joint subset partitioning and labelling problem (SPLP) where we jointly estimate poses of all people present in an image by minimizing a joint

objective. The formulation is based on partitioning and labelling an initial pool of body part candidates into subsets that correspond to sets of mutually consistent body-part candidates and abide to mutual consistency and exclusion constraints. The proposed method has several advantages. Firstly, the fomulation is able to deal with an unknown number of people, and also infers this number by linking part hypotheses. Secondly, the formulation allows to either deactivate or merge part hypotheses in the initial set of part candidates hence effectively performing non-maximum suppression (NMS). In contrast to NMS performed on individual part candidates, the model incorporates evidence from all other parts making the process more reliable. And last, the problem is cast in the form of an Integer Linear Program (ILP). Although the problem is NP-hard, the ILP formulation facilitates the computation of bounds and feasible solutions with a certified optimality gap. Empirical results on four diverse and challenging datasets show significant improvements over all previous methods not only for the multi-person, but also for the single-person pose estimation problem.

In thesis, we address a number of diverse tasks that aim to enable vision systems to understand people in realistic images and videos, at human perception level or even better. In particular, we propose several novel ideas and rigorous mathematical formulations for each task, push the boundary of state-of-the-arts and result in superior performance. However, some of the ideas proposed in this thesis are not fully explored. Next, we discuss how to further advance the techniques for visually understanding of people in images and videos in the future.

10.2 PERSPECTIVES FOR PEOPLE DETECTION

The people detection methods proposed in this thesis are mainly for crowded street scenes. Despite the heavy occlusions, recent techniques in this area have shown significant detection performance on large scale pedestrian detection benchmarks. In the future, we would like to develop detection models for more general scenarios, where people are no longer restricted to the upright orientation. The general people detection task is much more challenging due to several factors, such as large pose changes, appearance variance and partial occlusions. To address these challenges, we plan to focus on the following directions.

Modelling interaction between people and their environments. In real-world scenes, people often interact with objects. E.g. A person could sits on a chair and be heavily occluded by a table that is in front of him. In this case, the visual recognition of the chair and the table provides useful information for detecting the significantly occluded person. Meanwhile, the detection for the person could help us to reason about the existence of the table and the chair. Therefore how to jointly model people, their surrounding objects and the relations between them is an interesting direction to explore in the future. In particular, we would like to focus on designing structure models which take the object recognition information into account for the people

detection task.

Learning non-maximum suppression (NMS). Modern object detectors yield multiple equally plausible detections for a single object, which often requires a separate post-processing step to select a single detection hypothesis for the object. Several works have been proposed to learn a selection mechanism (Hosang *et al.*, 2017, Henderson and Ferrari, 2016, Hosang *et al.*, 2016). However none of them produce significant improvement over the traditional NMS. We believe that for a superior performance, the loss function should be designed in a way that instead of suppressing non-maximum detections, it should allow multiple detections and encourage a proper clustering of all plausible detections.

People detection by pose estimation and instance segmentation. The joint partitioning and labelling model proposed in Chapter 9 can be viewed as a person detector that is able to detect people under significant occlusions. The key element is that the detection task is operated on pre-defined body joints (key points) level, which is different from the typical bounding box to a full person extent. A straightforward way to further improve the detection performance in crowded scenes is to utilize the instance segmentation technique for the detection task. The Mask RCNN, recently proposed by He *et al.*, 2017, is a deep convolutional neural network architecture that is jointly trained for detection, instance segmentation and key points localization tasks. The results presented in the paper are encouraging and suggest that segmentation and pose aided detection model is a very promising future direction to explore.

10.3 PERSPECTIVES FOR MULTI PERSON TRACKING

In this thesis, we have focused on rigorous mathematical abstractions for the multi person tracking task. Our minimum cost multicut formulation (Chapter 5) and minimum cost lifted multicut formulation (Chapter 7) produce robust detection association results and define the state-of-the-art tracking performance on the challenging multi object tracking benchmark (Milan *et al.*, 2016). However, there are several limitations. First, the tracking approaches proposed in this thesis are stage-wise approaches. The training is performed in a piece-wise fashion, which does not necessarily result in a good estimation of model parameters. Second, we follow the traditional tracking-by-detection framework, where detection and tracking are considered as well-defined, separate tasks (An exception is the detector learning algorithm proposed in Chapter 4). Such decomposition is not ideal as these two tasks are highly correlated. A joint solution could be much more desirable. Third, in this thesis, multi person tracking is performed on 2D image domain and 3D scene information is completely ignored. We believe that jointly estimating 3D scene estimation and 3D object tracking is a very interesting direction to explore. In the following, we discuss several concrete directions for the multi person tracking task.

End-to-end modelling. Deep learning techniques have been used in many computer vision tasks, often obtaining superior performance. Researchers have been exploring deep learning based tracking algorithms as well (Sadeghian *et al.*, 2017, Milan *et al.*, 2017). However, there is arguably no convincing work yet. Reasons are two-fold: first, deep learning based methods often need a large amount of training data, which is not the case for multi person tracking, where annotating tracking data is very expensive; second, the multi person tracking task involves detection, association, re-identification, and counting. The deep features learned for one task are not necessarily suitable for another task. E.g. the features for the detection task should be characteristic to distinguish people and background. However, features for associating people detections should be conscious of fine-grained differences between people. Therefore designing an end-to-end deep architecture to accomplish the above-mentioned multiple tasks is an interesting and challenging topic to work on.

Tracking across multiple cameras. Another interesting direction for multi person tracking is to explore multi-camera setting. The tracking approaches proposed in this thesis are dedicated to single camera settings. However, multiple camera systems are largely used in video surveillance. Therefore, visual tracking across multiple cameras that without overlaps or small overlap regions is crucial for certain application, e.g. security. The main challenge of tracking across multiple cameras is to re-identify the target under various lighting, background, viewpoints. Given the powerful person re-identification model proposed in Chapter 7, extending the current tracking framework to the multiple camera setting is a straightforward direction for the future work.

Combing tracking and flow. Tracking and flow are typically considered as two separate problems. But to certain extent, they are not. Flow can be considered as "tracking" of pixels and tracking is the crude "flow" on the object level. Sevilla-Lara *et al.*, 2016 propose a semantic flow approach, where the objects in a video are first identified and tracked. The flow is then computed relative to the translation and scale that are obtained from the tracking result. We could further extend the semantic flow to utilize tracking to reason about the occlusion relation between objects, which is one of the toughest problems in flow estimation. The flow on the rigid part of the scene can be explored to estimate the depth of the scene across time, and then the tracks could be further extended with the depth information. Overall, we could obtain a detailed motion representation of both rigid and non-rigid scenes over time.

10.4 PERSPECTIVES FOR HUMAN POSE ESTIMATION

In the following, we discuss two future directions towards human pose estimation.

End-to-end multi person pose estimation approach. The Minimum Cost Multicut Problems that are intensively explored in this thesis present an interesting area of research in Computer Vision. The plain multicut formulation and the joint partitioning and node labelling formulation often consist of four individual steps. E.g. in the case of multi person pose estimation, the individual steps are detecting key points, computing similarity measures (costs on the edges), solving the optimization problem, and converting clusters to poses. The learning of the key point detector and the similarity measures are independent, which could lead to a suboptimal model parameters for the later optimization step. One important aspect that is not addressed by the thesis is the joint training of detection and similarity measures. Although lots of methods for learning structural model have been proposed, they typically rely on conditional random field models, which only concerns the labelling of node variables. Our models are applicable to the node labelling problem and the edge labelling problem. To our best knowledge, an end-to-end model for the combined problem has not yet been proposed. In the future, we would like to explore the end-to-end multi person pose estimation approach.

Human pose estimation in 3D. Another important aspect of future directions is to lift the human pose estimation from 2D image coordinates to 3D space. Single person 3D pose estimation has been studied and many works have been proposed (Ionescu *et al.*, 2014, Ramakrishna *et al.*, 2012, Li and Chan, 2014). But there are only a few works on 3D multi person pose estimation. One example is the work proposed by Rogez *et al.*, 2017, where the 2D and 3D pose estimation of multiple people are simultaneously obtained from an end-to-end trainable deep neural network. The network uses a pose proposal generator to produce a set of person-level bounding boxes and possible poses at different locations on the image. The estimated poses are later refined both in 2D and 3D. Instead of operating on the full person extent, a very interesting future direction for us is to extend our bottom-up 2D pose estimation approach to simultaneously estimate 3D locations and rotations of the key points. We believe that such bottom-up pose estimation approach could be more suitable for heavily occluded people. Another way of performing 3D human pose estimation is to utilize generative 3D human models and combine them with discriminative models. Such hybrid models could bring the complementary information from the both sides, resulting in a generic and scalable solution to 3D human pose estimation task.

10.5 THE BIGGER PICTURE

In the previous section, we have discussed the future works that are highly related to each contribution of this thesis. In the following, we will present a broader view and long-term directions towards a holistic understanding of visual scenes and in particular, visual understanding of people in the scenes.

Closing the gap. Due to the complexity and diversity of realistic visual scenes, people detection, tracking, articulated pose estimation are typically considered as well-defined and isolated research areas. While such simplifications seem to make each task easier, there are several notable problems. First, a large portion of image content is ignored, which normally contains information about 3D scene structure, surrounding objects, interaction between people. We argue that instead of discarding the rich visual information, we should utilize them as complementary observation when we infer the location, pose and motion of people in the scene. Second, detections errors introduced in the detection phase could directly influence the performance of tracking and pose estimation. As discussed in previous paragraphs, we believe that jointly modelling multiple tasks could be beneficial and is a promising direction for future work. In this thesis, we move towards the idea of jointly modeling different tasks. For instance, the detection approach proposed in Chapter 4 is optimized for the tracking performance, which facilitates a tighter relationship between people detector and tracker. Another example is the multi person pose estimation model proposed in Chapter 9. Instead of a two-stage approach where detection and pose estimation are performed separately, we utilize a single objective function to obtain the location and the pose estimation of each person in the scene. In our very recent work (Insafutdinov *et al.*, 2017), we propose a pose tracking approach where pose estimation as well as temporal association are jointly modelled. Last, higher-level vision tasks, such as activity recognition, body language, and social relation are not explored in this thesis, however these tasks could facilitate the recognition tasks such as detection, tracking and pose estimation. We argue that closing the gap between the higher level and lower level recognition tasks is a very promising long-term direction for future exploration.

Holistic understanding of real world videos. Visual understanding of general scenes is arguably the most important goal of many vision systems. The visual information contains many aspects, such as 3D scene geometry, object recognition, people recognition, the intersection between people and objects, and even physical constraints. When we go beyond single images, visual information also includes motion of the scene. As a long-term direction, we aim to obtain a holistic understanding of realistic videos. We would like to have a rich representation about how each pixel moves in the scene. By analysing the scene geometry and semantic information, we could split the scene into rigid and non-rigid parts. The motion of the rigid part of the scene could be the cues for uncovering the camera pose and motion. For the non-rigid part, we could explore tracking to understand their motion. As people are often the central characters for real world videos, it is particular interesting to understand their motion and behavior. We could utilize generative and discriminative human models to uncover their pose, shape and motion. We could also explore physical constraints that are inherent in the natural world to model the interaction between humans and objects. We argue that such holistic understanding of realistic images and videos is essential for building autonomous and intelligent computer systems and it is our long-term direction for future research.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

A. Adam, E. Rivlin, and I. Shimshoni (2006). Robust fragments-based tracking using the integral histogram, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. Cited on page 14.

B. Andres (2015). Lifting of Multicuts, *arXiv:1503.03791*. Cited on page 91.

B. Andres, T. Kröger, K. L. Briggman, W. Denk, N. Korogod, G. Knott, U. Köthe, and F. A. Hamprecht (2012). Globally Optimal Closed-surface Segmentation for Connectomics, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 111.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 122, 129, 130, 131, 134, and 137.

M. Andriluka, S. Roth, and B. Schiele (2008). People-Tracking-by-Detection and People-Detection-by-Tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 19, 24, 28, 34, 46, 60, 68, 70, 71, 78, and 147.

M. Andriluka, S. Roth, and B. Schiele (2010). Monocular 3D Pose Estimation and Tracking by Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 106.

A. Andriyenko and K. Schindler (2011). Multi-target tracking by continuous energy minimization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 36, 37, 38, 39, 46, 49, 52, 53, 55, and 148.

A. Andriyenko, K. Schindler, and S. Roth (2012). Discrete-Continuous Optimization for Multi-Target Tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 18, 19, 41, and 75.

S. Avidan (2007). Ensemble Tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 14 and 16.

B. Babenko, M.-H. Yang, and S. Belongie (2009). Visual tracking with online multiple instance learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 14 and 16.

Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud (2016). Tracking Multiple Persons Based on a Variational Bayesian Model, in *Workshop on Benchmarking Multi-target Tracking: MOTChallenge 2016*. Cited on pages 101 and 102.

N. Bansal, A. Blum, and S. Chawla (2004). Correlation Clustering, *Machine Learning*, vol. 56(1–3), pp. 89–113. Cited on page 67.

O. Barinova, V. Lempitsky, and P. Kohli (2010). On detection of Multiple object instances using Hough transform, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 11, 24, 35, 36, 40, and 147.

T. Beier, F. A. Hamprecht, and J. H. Kappes (2015). Fusion moves for correlation clustering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 111.

T. Beier, T. Kroeger, J. Kappes, U. Kothe, and F. Hamprecht (2014). Cut, Glue, & Cut: A Fast, Approximate Solver for Multicut Partitioning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 111.

V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic (2014). 3D Pictorial Structures for Multiple Human Pose Estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 21.

R. Benenson, M. Mathias, T. Tuytelaars, and L. J. V. Gool (2013). Seeking the Strongest Rigid Detector., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 10.

K. Bernardin and R. Stiefelhagen (2008). Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics, *Image and Video Processing*, vol. 2008. Cited on pages 37, 46, 50, 52, 71, 100, and 153.

G. Bertasius, J. Shi, and L. Torresani (2015). High-for-Low and Low-for-High: Efficient Boundary Detection from Deep Object Features and its Applications to High-Level Vision, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on page 20.

L. Bourdev and J. Malik (2009). Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on page 11.

M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool (2011). Online Multiperson Tracking-by-Detection from a Single Uncalibrated Camera, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 53, 55, and 73.

W. Brendel, M. Amer, and S. Todorovic (2011). Multiobject Tracking as Maximum Weight Independent Set, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 18.

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 page 19.

T. Brox and J. Malik (2011). Large displacement optical flow: descriptor matching in variational motion estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 108.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh (2017). Realtime multi-person 2d pose estimation using part affinity fields, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 22.

J. Chang, D. Wei, and J. W. F. III (2013). A Video Representation Using Temporal Superpixels, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 20.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2015). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, in *International Conference on Learning Representations (ICLR) 2015*. Cited on pages 112, 113, 128, and 151.

X. Chen and A. Yuille (2014). Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations, in *Advances in Neural Information Processing Systems (NIPS) 2014*. Cited on pages 21, 122, 132, 133, and 137.

X. Chen and A. Yuille (2015). Parsing Occluded People by Flexible Compositions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 22, 134, 135, 136, 137, and 152.

A. Cheriyadat and R. Radke (2009). Non-negative matrix factorization of partial track data for motion segmentation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on page 19.

W. Choi (2015). Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 86, 87, 94, 96, 101, 102, 103, 117, and 118.

S. Chopra, R. Hadsell, and Y. Lecun (2005). Learning a similarity metric discriminatively, with application to face verification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on page 98.

S. Chopra and M. Rao (1993). The partition problem, *Mathematical Programming*, vol. 59(1–3), pp. 87–115. Cited on pages 20, 80, 90, 108, and 124.

X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang (2017). Multi-context attention for human pose estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 21.

R. T. Collins, Y. Liu, and M. Leordeanu (2005). Online selection of discriminative tracking features, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 14.

D. Comaniciu, V. Ramesh, and P. Meer (2000). Real-time tracking of non-rigid objects using mean shift, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*. Cited on page 14.

D. Cremers, F. R. Schmidt, and F. Barthel (2008). Shape Priors in Variational Image Segmentation: Convexity, Lipschitz Continuity and Globally Optimal Solutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on page 117.

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 9 and 10.

M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking, in *Proceedings of the European Conference on Computer Vision (ECCV) 2016*. Cited on page 14.

A. Daniel Costea and S. Nedevschi (2016). Semantic Channels for Fast Pedestrian Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 10.

E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica (2006). Correlation clustering in general weighted graphs, *Theoretical Computer Science*, vol. 361(2–3), pp. 172–187. Cited on pages 67, 94, 108, and 111.

C. Desai and D. Ramanan (2012). Detecting Actions, Poses, and Objects with Relational Phraselets, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 11 and 12.

M. M. Deza and M. Laurent (2009). *Geometry of Cuts and Metrics*, Springer Publishing Company, Incorporated, 1st edn. Cited on page 20.

T. B. Dinh, N. Vo, and G. Medioni (2011). Context tracker: Exploring supporters and distracters in unconstrained environments, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 14.

P. Dollar, R. Appel, S. Belongie, and P. Perona (2014). Fast Feature Pyramids for Object Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 10.

P. Dollár, Z. Tu, P. Perona, and S. Belongie (2009). Integral Channel Features, in *Proceedings of the British Machine Vision Conference (BMVC) 2009*. Cited on pages 10, 70, and 116.

P. Dollár, C. Wojek, B. Schiele, and P. Perona (2012). Pedestrian Detection: An Evaluation of the State of the Art, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 3 and 41.

R. Dragon and J. Rosenhahn, B.and Ostermann (2012). Multi-scale clustering of frame-to-frame correspondences for motion segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 19.

B. L. E. Horbert, K. Rematas (2011). Level-Set Person Segmentation and Tracking with Multi-Region Appearance Models and Top-Down Shape Information, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 117.

M. Eichner and V. Ferrari (2010). We Are Family: Joint Pose Estimation of Multiple Persons, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 22, 134, and 135.

E. Elhamifar and R. Vidal (2013). Sparse subspace clustering, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 19.

M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila (2010). Multi-cue pedestrian classification with partial occlusion handling, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 11.

L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle (2016). Improving Multi-frame Data Association with Sparse Representations for Robust Near-online Multi-object Tracking, in *Proceedings of the European Conference on Computer Vision (ECCV) 2016*. Cited on page 101.

A. Farhadi and M. Sadeghi (2011). Recognition using Visual Phrases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 12.

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)*. Cited on pages 10, 11, 12, 23, 24, 26, 27, 34, 40, 43, 52, 68, 70, 84, 116, 139, and 147.

J. M. Ferryman and A. Shahrokni (2009). PETS2009: Dataset and challenge, in *Winter-PETS 2009*. Cited on pages 49 and 51.

K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik (2015). Learning to Segment Moving Objects in Videos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 20.

K. Fragkiadaki and J. Shi (2011). Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement, in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011. Cited on page 20.

K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi (2012). Two-Granularity Tracking: Mediating Trajectory and Detection Graphs for Tracking under Occlusions, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 20 and 106.

A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun (2014). 3D Traffic Scene Understanding from Movable Platforms, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36(5), pp. 1012–1025. Cited on pages 86, 87, 101, and 102.

G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes (2014). Parsing Occluded People, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 135.

R. Girshick (2015). Fast R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 113, 128, and 129.

R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014a). 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 10.

R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014b). 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 70.

R. B. Girshick, P. F. Felzenszwalb, and D. McAllester (2010). LSVM-MDPM Release 4 Notes. Cited on pages 26 and 32.

G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik (2014). Using k-poselets for detecting people and localizing their keypoints, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 21 and 122.

N. Gordon, D. Salmond, and A. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEEE Proceedings F, Radar and Signal Processing*. Cited on page 14.

H. Grabner, M. Grabner, and H. Bischof (2006). Real-Time Tracking via On-line Boosting, in *Proceedings of the British Machine Vision Conference (BMVC) 2006*. Cited on pages 14 and 16.

K. Grauman, G. Shakhnarovich, and T. Darrell (2003). Inferring 3D Structure with a Statistical Image-Based Shape Model., in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2003*. Cited on page 12.

M. Grötschel and Y. Wakabayashi (1989). A cutting plane algorithm for a clustering problem, *Mathematical Programming*, vol. 45(1), pp. 59–96. Cited on page 90.

K. He, G. Gkioxari, P. Dollár, and R. Girshick (2017). Mask R-CNN, *arXiv:1703.06870*. Cited on page 143.

K. He, X. Zhang, S. Ren, and J. Sun (2016). Deep Residual Learning for Image Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 1.

P. Henderson and V. Ferrari (2016). End-to-end training of object class detectors for mean average precision, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2016*. Cited on page 143.

R. Henschel, L. Leal-Taixe, and B. Rosenhahn (2014). Efficient Multiple People Tracking Using Minimum Cost Arborescences, in *Proceedings of the German Conference on Pattern Recognition (GCPR) 2014*. Cited on pages 19 and 106.

J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko (2014). LSDA: Large Scale Detection through Adaptation, in *Advances in Neural Information Processing Systems (NIPS) 2014*. Cited on pages 112 and 113.

S. Hong, T. You, S. Kwak, and B. Han (2015). Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network, in *Proceedings of the International Conference on Machine Learning (ICML) 2015*. Cited on page 16.

J. Hosang, R. Benenson, and B. Schiele (2016). A Convnet for Non-maximum Suppression, in *Proceedings of the German Conference on Pattern Recognition (GCPR) 2016*. Cited on page 143.

J. Hosang, R. Benenson, and B. Schiele (2017). Learning Non-maximum Suppression, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 143.

J. Hosang, M. Omran, R. Benenson, and B. Schiele (2015). Taking a Deeper Look at Pedestrians, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 10.

C. Huang, B. Wu, and R. Nevatia (2008). Robust Object Tracking by Hierarchical Association of Detection Responses, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. Cited on page 106.

E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele (2017). Articulated Multi-person Tracking in the Wild, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 146.

C. Ionescu, J. Carreira, and C. Sminchisescu (2014). Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 145.

M. Isard and J. MacCormick (2001). BraMBLe: a Bayesian multiple-blob tracker, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2001*. Cited on page 14.

P. Ji, H. Li, M. Salzmann, and Y. Dai (2014). Robust Motion Segmentation with Unknown Correspondences, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on pages 19 and 106.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell (2014). Caffe: Convolutional Architecture for Fast Feature Embedding, *arXiv:1408.5093*. Cited on page 99.

H. Jiang, S. Fels, and J. J. Little (2007). A linear programming approach for multiple object tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on page 17.

H. Jiang and D. R. Martin (2009). Global Pose Estimation using Non-tree Models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 21 and 22.

S. Johnson and M. Everingham (2011). Learning Effective Human Pose Estimation from Inaccurate Annotation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 122 and 129.

S. Julier and J. Uhlmann (1997). A new extension of the Kalman filter to nonlinear systems, in *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defense Sensing, Simulations and Controls 1997*. Cited on page 13.

Z. Kalal, J. Matas, and K. Mikolajczyk (2010). P-N learning: Bootstrapping binary classifiers by structural constraints, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 14.

R. E. Kalman (1960). A New Approach to Linear Filtering And Prediction Problems, *ASME Journal of Basic Engineering*. Cited on page 13.

R. Kaucic, A. G. A. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs (2005). A Unified Framework for Tracking Through Occlusions and Across Sensor Gaps, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on page 18.

B. W. Kernighan and S. Lin (1970). An efficient heuristic procedure for partitioning graphs, *Bell Systems Technical Journal*, vol. 49. Cited on pages 60, 67, and 111.

M. Keuper, B. Andres, and T. Brox (2015a). Motion Trajectory Segmentation via Minimum Cost Multicuts, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 106, 109, 111, 114, 115, 116, 118, and 151.

M. Keuper, E. Levinkov, N. Bonneel, G. Lavoué, T. Brox, and B. Andres (2015b). Efficient Decomposition of Image and Mesh Graphs by Lifted Multicuts, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 79, 80, 81, 85, 94, and 111.

Z. Khan, T. Balch, and F. Dellaert (2005). MCMC-based particle filtering for tracking a variable number of interacting targets, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 14.

C. Kim, F. Li, A. Ciptadi, and J. M. Rehg (2015). Multiple Hypothesis Tracking Revisited, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 18, 86, 87, 90, 101, and 102.

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 130.

L. Ladicky, P. H. Torr, and A. Zisserman (2013). Human Pose Estimation using a Joint Pixel-wise and Part-wise Formulation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 22.

N. Le, A. Heili, and J.-M. Odobez (2016). Long-Term Time-Sensitive Costs for CRF-Based Tracking by Detection, in *Workshop on Benchmarking Multi-target Tracking: MOTChallenge 2016*. Cited on pages 101 and 102.

L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler (2016). Learning by tracking: siamese CNN for robust target association, *DeepVision: Deep Learning for Computer Vision Workshop*. Cited on pages 18 and 90.

L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler (2015). MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking, *arXiv:1504.01942*. Cited on pages 89, 97, 106, 111, 116, and 117.

B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee (2016). Multi-class Multi-object Tracking Using Changing Point Detection, in *Workshop on Benchmarking Multi-target Tracking: MOTChallenge 2016*. Cited on page 103.

B. Leibe, K. Schindler, and L. Van Gool (2007). Coupled Detection and Trajectory Estimation for Multi-Object Tracking, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. Cited on page 12.

B. Leibe, E. Seemann, and B. Schiele (2012). Pedestrian Detection in Crowded Scenes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 11.

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 19.

S. Li and A. B. Chan (2014). 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2014*. Cited on page 145.

W. Li, R. Zhao, T. Xiao, and X. Wang (2014). DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 97.

Z. Li, J. Guo, L. Cheong, and S. Zhou (2013). Perspective motion segmentation via collaborative clustering, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 19.

J. J. Lim, C. L. Zitnick, and P. Dollar (2013). Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 10.

Z. Lin and L. S. Davis (2008). A pose-invariant descriptor for human detection and segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. Cited on page 10.

J. Marin, D. Vazquez, D. Geronimo, and A. Lopez (2010a). Learning appearance in virtual scenarios for pedestrian detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 13.

J. Marin, D. Vázquez, D. Gerónimo, and A. M. López (2010b). Learning appearance in virtual scenarios for pedestrian detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 12.

A. Milan, L. Leal-Taix, K. Schindler, and I. Reid (2015). Joint Tracking and Segmentation of Multiple Targets, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 20 and 117.

A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler (2016). MOT16: A Benchmark for Multi-Object Tracking, *arXiv:1603.00831*. Cited on pages 4, 83, 89, 97, 100, 106, 111, 116, 117, and 143.

A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler (2017). Online Multi-Target Tracking using Recurrent Neural Networks, in *AAAI 2017*. Cited on page 144.

A. Milan, S. Roth, and K. Schindler (2014). Continuous Energy Minimization for Multitarget Tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 86, 87, 101, and 102.

H. Nam and B. Han (2016). Learning multi-domain convolutional neural networks for visual tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on pages 14 and 16.

A. Newell, K. Yang, and J. Deng (2016). Stacked Hourglass Networks for Human Pose Estimation, in *Proceedings of the European Conference on Computer Vision (ECCV) 2016*. Cited on page 21.

P. Ochs and T. Brox (2012). Higher Order Motion Models and Spectral Clustering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 19, 20, 106, 114, and 115.

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)*, vol. 36(6), pp. 1187 – 1200. Cited on pages 19, 105, 106, 108, 109, 111, 112, 113, 114, 118, and 151.

S. Oron (2012). Locally Orderless Tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 14.

S. Paisitkriangkrai, C. Shen, and A. van den Hengel (2014). Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features, in *Proceedings of the European Conference on Computer Vision (ECCV) 2014*. Cited on page 10.

G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille (2015). Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation. Cited on pages 112, 113, and 151.

O. M. Parkhi, A. Vedaldi, and A. Zisserman (2015). Deep face recognition, in *Proceedings of the British Machine Vision Conference (BMVC) 2015*. Cited on page 99.

B. Pepik, M. Stark, P. Gehler, and B. Schiele (2012). Teaching 3D Geometry to Deformable Part Models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 44.

H. Pirsiavash, D. Ramanan, and C. C. Fowlkes (2011). Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 17, 18, 19, 36, 37, 38, 39, 59, 60, 73, 78, and 94.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele (2013). Strong Appearance and Expressive Spatial Models for Human Pose Estimation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 21 and 128.

L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele (2016). DeepCut: Joint Subset Partition and Labeling for Multi Person

Pose Estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on pages 8 and 99.

L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele (2012). Articulated People Detection and Pose Estimation: Reshaping the Future, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 21 and 122.

L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele (2011). Learning People Detection Models from Few Training Samples, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 12, 13, and 35.

Z. W. Pylyshyn and R. W. Storm (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism, *Spatial vision*, vol. 3(3), pp. 179–197. Cited on page 13.

H. Rahmati, R. Dragon, O. M. Aamo, L. V. Gool, and L. Adde (2014). Motion Segmentation with Weak Labeling Priors, in *Proceedings of the German Conference on Pattern Recognition (GCPR) 2014*. Cited on pages 19 and 106.

V. Ramakrishna, T. Kanade, and Y. Sheikh (2012). Reconstructing 3d human pose from 2d image landmarks, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 145.

D. Ramanan (2006). Learning to Parse Images of Articulated Objects, in *Advances in Neural Information Processing Systems (NIPS) 2006*. Cited on page 21.

S. Ren, K. He, R. Girshick, and J. Sun (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in *Advances in Neural Information Processing Systems (NIPS) 2015*. Cited on pages 10, 109, 112, 113, and 139.

X. Ren, A. C. Berg, and J. Malik (2005). Recovering Human Body Configurations using Pairwise Constraints between Parts, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. Cited on page 21.

G. Rogez, P. Weinzaepfel, and C. Schmid (2017). LCR-Net: Localization-Classification-Regression for Human Pose, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 145.

A. Sadeghian, A. Alahi, and S. Savarese (2017). Tracking The Untrackable: Learning To Track Multiple Cues with Long-Term Dependencies, *CoRR*, vol. abs/1701.01909. Cited on page 144.

S. Santhoshkumar, S. Karthikeyan, and B. Manjunath (2013). Robust multiple object tracking by detection with interacting markov chain monte carlo, in *Proceedings of the IEEE International Conference on Image Processing (ICIP) 2013*. Cited on page 14.

B. Sapp and B. Taskar (2013). Multimodal Decomposable Models for Human Pose Estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 21, 122, and 130.

A. V. Segal and I. Reid (2013). Latent Data Association: Bayesian Model Selection for Multi-target Tracking, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 18, 19, 59, 60, 73, 78, and 106.

L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black (2016). Optical Flow with Semantic Segmentation and Localized Layers, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 144.

F. Shi, Z. Zhou, J. Xiao, and W. Wu (2013). Robust trajectory clustering for motion segmentation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 19 and 106.

H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua (2011). Tracking Multiple People Under Global Appearance Constraints, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 17.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake (2011). Real-Time Human Pose Recognition in Parts from a Single Depth Image, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 35.

G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah (2012). Part-based multiple-person tracking with partial occlusion handling, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 51, 53, 55, and 73.

K. Simonyan and A. Zisserman (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, in *International Conference on Learning Representations (ICLR) 2014*. Cited on pages 97 and 128.

R. Stewart, M. Andriluka, and A. Y. Ng (2016). End-To-End People Detection in Crowded Scenes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 11.

M. Sun and S. Savarese (2011). Articulated Part-based Model for Joint Object Detection and Pose Estimation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 21, 122, and 134.

S. Tang, B. Andres, M. Andriluka, and B. Schiele (2015). Subgraph Decomposition for Multi-Target Tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 7.

S. Tang, B. Andres, M. Andriluka, and B. Schiele (2016). Multi-Person Tracking by Multicuts and Deep Matching, in *Workshop on Benchmarking Multi-target Tracking: MOTChallenge 2016*. Cited on page 118.

S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele (2013). Learning People Detectors for Tracking in Crowded Scenes, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 7 and 73.

S. Tang, M. Andriluka, and B. Schiele (2012). Detection and Tracking of Occluded People, in *Proceedings of the British Machine Vision Conference (BMVC) 2012*. Cited on pages 2, 7, 56, and 148.

S. Tang, M. Andriluka, and B. Schiele (2014). Detection and Tracking of Occluded People, *International Journal of Computer Vision (IJCV)*. Cited on pages 2, 7, and 106.

Y. Tian, P. Luo, X. Wang, and X. Tang (2015). Pedestrian detection aided by deep learning semantic tasks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 10.

J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler (2015). Efficient Object Localization Using Convolutional Networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 21, 121, 129, 133, and 134.

J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler (2014). Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation, in *Advances in Neural Information Processing Systems (NIPS) 2014*. Cited on pages 122, 129, 130, 132, 133, and 134.

A. Toshev and C. Szegedy (2014). DeepPose: Human Pose Estimation via Deep Neural Networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 21 and 130.

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders (2013a). Selective Search for Object Recognition, *International Journal of Computer Vision (IJCV)*. Cited on page 128.

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders (2013b). Selective Search for Object Recognition, *International Journal of Computer Vision (IJCV)*, vol. 104(2), pp. 154–171. Cited on page 112.

X. Wang, T. Han, and S. Yan (2009a). An HOG-LBP Human Detector with Partial Occlusion Handling, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on page 11.

X. Wang, T. X. Han, and S. Yan (2009b). An HOG-LBP human detector with partial occlusion handling, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on page 10.

X. Wang, E. Turetken, F. Fleuret, and P. Fua (2016). Tracking Interacting Objects Using Intertwined Flows, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 17.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh (2016). Convolutional Pose Machines, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 21.

P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid (2013). DeepFlow: Large displacement optical flow with deep matching, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 78, 81, 96, and 109.

L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Li (2014). Multiple Target Tracking Based on Undirected Hierarchical Relation Hypergraph, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 18, 73, and 78.

C. Wojek, S. Roth, K. Schindler, and B. Schiele (2010). Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 37 and 106.

C. Wojek, S. Walk, S. Roth, and B. Schiele (2011). Monocular 3D Scene Understanding with Explicit Occlusion Reasoning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 11.

C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele (2013). Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 106.

B. Wu and R. Nevatia (2006). Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. Cited on pages 46 and 153.

B. Wu and R. Nevatia (2007). Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors, in *International Journal of Computer Vision (IJCV) 2007*. Cited on page 18.

Y. Wu, J. Lim, and M. H. Yang (2015). Object Tracking Benchmark, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 15 and 147.

Z. Wu, A. Thangali, S. Sclaroff, and M. Betke (2012). Coupling Detection and Data Association for Multiple Object Tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 12.

Y. L. X. Fan, K. Zheng and S. Wang (2015). Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 132 and 133.

Y. Xiang, A. Alahi, and S. Savarese (2015). Learning to Track: Online Multi-Object Tracking by Decision Making, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2015*. Cited on pages 18 and 94.

X. Yan, X. Wu, I. A. Kakadiaris, and S. Shah (2012). To Track or To Detect? An Ensemble Framework for Optimal Selection, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 12.

B. Yang and R. Nevatia (2012). Online Learned Discriminative Part-Based Appearance Models for Multi-Human Tracking, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 41.

Y. Yang and D. Ramanan (2013). Articulated Human Detection with Flexible Mixtures of Parts, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on pages 21, 22, 134, and 135.

F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan (2016). POI: Multiple Object Tracking with High Performance Detection and Appearance Feature, in *Workshop on Benchmarking Multi-target Tracking: MOTChallenge 2016*. Cited on pages 103 and 153.

A. R. Zamir, A. Dehghan, and M. Shah (2012). GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 18, 19, 41, 53, 55, 59, 71, 73, 75, and 106.

L. Zhang, Y. Li, and R. Nevatia (2008). Global data association for multi-object tracking using network flows, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 17, 60, 78, and 147.

L. Zhang, L. Lin, X. Liang, and K. He (2016). Is Faster R-CNN Doing Well for Pedestrian Detection?, in *Proceedings of the European Conference on Computer Vision (ECCV) 2016*. Cited on page 10.

S. Zhang, C. Bauckhage, and A. B. Cremers (2014). Informed Haar-Like Features Improve Pedestrian Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 10.

R. Zhao, W. Ouyang, and X. Wang (2013). Unsupervised Salience Learning for Person Re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 69.

L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian (2015). Scalable Person Re-identification: A Benchmark, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 97.

V. Zografos, R. Lenz, E. Ringaby, M. Felsberg, and K. Nordberg (2014). Fast segmentation of sparse 3D point trajectories using group theoretical invariants, in *Proceedings of the Asian Conference on Computer Vision (ACCV) 2014*. Cited on pages 19 and 20.

PUBLICATIONS

[12] *Multiple People Tracking with Lifted Multicut and Person Re-identification.*
Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele.
IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2017.

[11] *Joint Graph Decomposition and Node Labeling by Local Search.*
Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox and Bernt Schiele.
IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2017.

[10] *Articulated Multi-person Tracking in the Wild.*
Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele.
IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2017.

[9] *Generating Descriptions with Grounded and Co-Referenced People.*
Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh and Bernt Schiele.
IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2017.

[8] *Multi-Person Tracking by Multicuts and Deep Matching.*
Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele.
ECCV Workshop on Benchmarking Mutliple Object Tracking, 2016.
Winner of the ECCV Multi-Object Tracking Challenge

[7] *A Multi-cut Formulation for Joint Segmentation and Tracking of Multiple Objects.*
Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox and Bernt Schiele.
In arXiv:1607.06317, 2016.

[6] *DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation.*
Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehker and Bernt Schiele.
IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2016.

[5] *Subgraph decomposition for multi-target tracking.*
Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele.
IEEE Conf. on Computer Vision and Pattern Recognition (**CVPR**), 2015.

[4] *Detection and Tracking of Occluded People.*
Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele.
The International Journal of Computer Vision (**IJCV**), 2014.

[3] *Learning People Detectors for Tracking in Crowded Scenes.*
Siyu Tang, Mykhaylo Andriluka, Anton Milan, Konrad Schindler, Stefan Roth and Bernt Schiele.
In International Conference on Computer Vision (**ICCV**), 2013.

[2] *Detection and Tracking of Occluded People.*
Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele.
In British Machine Vision Conference (**BMVC**), 2012.
BMVC Best Paper Award

[1] *OpenBioSafetyLab: A virtual world based biosafety training application for medical students.*
Arturo Nakasone, Siyu Tang, Mika Shigematsu, Berthold Heinecke, Shuji Fujimoto, and Helmut Prendinger.
International Conference on Information Technology: New Generations (**ITNG**), 2011.