Detection, Tracking and Pose Estimation of People in Challenging Real-World Scenes

A dissertation submitted to the
TECHNISCHE UNIVERSITÄT DARMSTADT
Fachbereich 20

for the degree of
Dr. rer. nat.

presented by

MYKHAYLO ANDRILUKA
Dipl.–Math.

born on 26th of June, 1980
in Odessa, Ukraine

Prof. Stefan Roth, Ph.D., examiner
Prof. Daniel Huttenlocher, Ph.D., co-examiner
Prof. Dr. Bernt Schiele, co-examiner

Date of Submission: 13th of September, 2010
Date of Defense: 22nd of Oktober, 2010

Darmstadt, 2011
D17

Abstract

In this thesis, we consider three challenging and longstanding problems in computer vision: people detection, people tracking and articulated pose estimation. Generic solutions to these problems are essential building blocks for understanding images containing people, an exciting and challenging task with numerous applications in automotive safety, robotic navigation, human-computer interaction, and automatic image indexing and retrieval. Indeed, human actions, intentions and emotions can often be inferred from accurate estimates of human body poses and their movement over time. However, untill recently, accurate estimation of body poses has been possible only in controlled laboratory conditions, typically requiring multiple cameras and specialized motion capture equipment. In order to address this shortcoming, we propose algorithms capable of automatically finding people in uncontrolled outdoor environments, tracking them over time and estimating their body configurations. In the process, we also tackle several important technical challenges, including the large appearance variability of humans, the full and partial occlusions that frequently occur in typical street scenes, and ambiguities in 2D to 3D lifting and data association.

Humans appear in images wearing a large variety of clothing, in a large number of possible body poses and visible from various viewpoints. Jointly, these factors create very complex appearance patterns that are hard to model and detect well. In order to deal with the large appearance variability, we propose an approach based on the pictorial structures paradigm in which we represent the human body as a flexible configuration of rigid body parts and model the appearance of each body part using local image descriptors and discriminative classifiers. We demonstrate the generality of our approach by successfully applying it to various human detection and pose estimation problems.

One of the goals of this work is to demonstrate the advantages of a tight coupling of people detection, pose estimation and tracking. Tracking of people in uncontrolled conditions is difficult not only due to appearance variability, but also to frequent full and partial occlusions, which often happen when multiple people are present in the scene. Presence of multiple people also severely complicates data association between frames of the sequence. In order to address this challenge, we propose a tracking-by-detection framework that combines evidence from single-frame detections over several subsequent frames using a dynamical model of body articulations. We demonstrate the effectiveness of our tracking-by-detection approach by applying it to the problem of monocular 3D pose estimation of people in uncontrolled street environments.

Zusammenfassung

In dieser Dissertation untersuchen wir drei komplexe und zusammenhängende Fragestellungen aus dem Bereich des maschinellen Sehens: Menschenerkennung, Menschen-Tracking, und Posenschätzung. Generische Lösungen für diese Aufgaben sind wichtige Bausteine für das automatische Bildverstehen und haben viele Anwendungen in der Automobilindustrie, Roboter Navigation, Mensch-Maschiene Interaktion, Bild Indizierung und Retrieval. In der tat, ist es in vielen Fällen möglich die Posen und Bewegungen von Menschen zu verwenden um auf die Aktivitäten, Intentionen und Emotionen von Menschen Rückschlüsse zu ziehen. Allerdings war das präzise Schätzen von menschlichen Posen bisher nur unter kontrollierten Labor-Bedingungen möglich, und erforderte den Einsatz mehrerer Kameras und spezieller Ausrüstung. Um diesen Mangel zu beheben, schlagen wir in dieser Arbeit Algorithmen vor, die es erlauben die Menschen unter unkontrollierten Bedingungen zu erkennen, über die Zeit zu verfolgen und ihre Posen zu schätzen. Dabei befassen wir uns mit solchen wichtigen technischen Herausforderungen wie grosse Variabilität im Aussehen von Menschen, volle und partielle Verdeckungen und Mehrdeutigkeiten in der Daten-Assoziierung und 3D Rekonstruktion.

Menschen erscheinen in Bildern in unterschiedlicher Bekleidung, nehmen unterschiedliche Posen an, und können aus unterschiedlichen Bildwinkeln dargestellt werden. Gemeinsam tragen diese Faktoren dazu bei, dass daraus resultierende Muster schwer zu repräsentieren und zu erkennen sind. Um mit solche grosser Variabilität umgehen zu können, schlagen wir den Ansatz, der auf dem "pictorial structures" Modell basiert und in dem der menschliche Körper durch eine flexible Konfiguration aus den starren Körperteilen repräsentiert wird vor. Dabei modellieren wir das Aussehen jedes Körperteils mit Hilfe von lokalen Discriptoren und diskriminiativen Klassifikatoren. Wir demonstrieren die Allgemeinheit unseres Ansatzes, indem wir ihn erfolgreich auf unterschiedliche Aufgaben in der Menschenerkennung und Posenschätzung anwenden.

Eines der wichtigen Ziele dieser Arbeit ist, die Vorteile von enger Kopplung zwischen Menschenerkennung, Tracking und Posenschätzung zu demonstrieren. Tracking von Menschen unter unkontrollierten Bedingungen ist nicht nur wegen der komplexen visuellen Muster schwer, sondern auch wegen häufiger voller und partieller Verdeckungen. Die Präsenz von mehreren Leuten in einer Szene macht es auch schwierig die Hypothesen zwischen einzelnen Bildern zu assoziieren. Um dieser Herausforderung zu begegnen, schlagen wir den "tracking-by-detection" Ansatz vor, in dem die Beobachtungen über mehrere einzelne Bilder anhand des dynamischen Körpermodells kombiniert und verfeinert werden. Um die Effektivität unseres "tracking-by-detection" Ansatzes zu demonstrieren wenden wir Ihn auf das Problem der monokularer 3D Posenschätzung von mehreren Menschen in unkontrollierten Strassenbedingungen an.

Acknowledgments

This thesis would not be possible without the support of many people. First, I would like to thank my supervisors, Prof. Bernt Schiele and Prof. Stefan Roth, for their invaluable advice and encouragement during my work. Being supervised by two senior researchers allowed me to obtain timely and high quality feedback on my work and created an extremely motivating research atmosphere. In particular, I would like to thank Prof. Bernt Schiele for an excellent thesis topic and for always pushing me in the right direction. Many times, his energy and support have helped me continue working where I would have otherwise given up. I am also very thankful to Prof. Stefan Roth for his rigorous attitude towards research and his unrivaled ability to spot weak points in my formulations, which has saved me from later embarrassment so many times.

During my work on this thesis, I was a member of the Multimodal Interactive Systems group at TU Darmstadt. I would like to thank my colleagues Dr. Gyuri Dorko, Mario Fritz, Ulrich Steinhoff, Tam Huynh, Maja Stikic, Andreas Zinnen, Victoria Carlsson, Edgar Seemann, Dr. Kristof Van Laerhoven, Dr. Diane Larlus, Ulf Blanke, Sandra Ebert, Stefan Walk, Marcus Rohrbach, Nikodem Majer, Michael Stark, Christian Wojek and Paul Schnitzpan for creating an excellent research environment, cheering me up, and having many discussions over coffee at “603 qm” and on the way to the bakery. Edgar Seemann helped me a lot in the first months of my work and guided me through the jungles of TUDVision. Christian Wojek spent a lot of his time maintaining the computing infrastructure and introduced the PBS cluster management system that reduced the fights over computing time in our lab to a minimum. He also contributed the boosting code that we use in our pose estimation approach. Nikodem Majer was as excellent discussion partner throughout my work and introduced me to the prior work on iterative image parsing that I discuss in Chapter 4. Many thanks go to my lab roommates, Nikodem, Michael and Christian, for all of the discussions about research and life that we had over the years. I am very proud to have worked with you all!

I am very thankful to a secretary of our group, Ursula Paeckel, for her continuous support and for reminding us that life outside of the lab is still there. Without her excellent organizational skills, my life in Darmstadt definitely would have been much harder.

Many thanks also go to my colleagues from the research training group “Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments”, and especially to Paul Schnitzspan who was always there to back me up.

Pursuing a PhD is a hard work, and I would like to thank people who have provided me with opportunities to see the world outside of the lab. Here, I would like to thank Giovanni Maria Farinella for organizing an excellent computer vision summer school in Sicily, which was definitely one of the most enjoyable experiences during my PhD time. I would also like to thank the Darmstadt Dragons softball team, in particular people who have coached the team and kept it alive over the years: Chris Urbanik, Bernt Ost, Andy Wirth, Nate Wasser, Andre Spiess, Bjoern Schumacher and Christian Dick.

I would also like to thank people who were involved in my graduate studies at TU Darmstadt. In particular, I would like to thank Prof. Thomas Hofmann for introducing me to the field of machine learning and Lorenz Weizsaecker for his help during my work on the diploma thesis and in creating the dataset used in Chapter 3. In my undergraduate studies, I learned a lot from Prof. Ulrich Reif and Prof. Klaus Keimel, and I am very proud to have been one of their students. Dr. Werner Nickel provided invaluable help in my first months in Germany, Dr. Nicole Marx taught me German in less than a year, and Tanja Griesshammer from DAAD provided initial support and introduced me to the MCS program at TU Darmstadt. I am also thankful for my colleagues from Advancis Software and Services, especially to Peter Arndt for all of the help during my first experience of working in a software development company.

Finally, I would like to thank my friends Pavol Safarik, Bjoern Schumacher and Igor Shevchenko for being there over many years, and my parents, Valentina Skobeeva and Vladimir Andriluka and my aunt Galina Skobeeva for their encouragement and support.

Contents

1 Introduction **1**

1.1 Contributions . . . 3

1.2 Outline . . . 6

2 Related Work **9**

2.1 Pictorial Structures Approach . . . 11

2.2 People Detection and Tracking . . . 13

2.2.1 Pedestrian Detection . . . 13

2.2.2 People Detection . . . 16

2.2.3 People Tracking . . . 18

2.3 2D Human Pose Estimation . . . 19

2.4 3D Human Pose Estimation . . . 23

2.5 Other Related Work . . . 26

3 Combining People Detection, Tracking and Pose Estimation **27**

3.1 Introduction . . . 28

3.2 Pedestrian Detector . . . 30

3.2.1 Part-based model for pedestrian detection . . . 30

3.2.2 Model training . . . 32

3.2.3 Inference and Results . . . 32

3.3 Detection of Tracks in Image Sequences . . . 35

3.3.1 Gaussian process latent variable model . . . 36

3.3.2 Reconstruction of poses in short sequence . . . 37

3.3.3 Optimal track selection based on overlapping tracklets . . . 39

3.4 Experiments . . . 40

3.5 Conclusion . . . 42

4 Discriminative Appearance Model for Pictorial Structures **45**

4.1 Introduction . . . 45

4.2 Generic Model for People Detection and Pose Estimation . . . 48

4.2.1 Kinematic tree prior . . . 48

4.2.2 Discriminatively trained part models 50
4.2.3 Exact model inference 54
4.2.4 Incorporation of additional relationships between body parts . 54
4.3 Evaluation of the Model Components 56
4.3.1 Evaluation of appearance model parameters 56
4.3.2 Extension of the model with repulsive factor 59
4.3.3 Comparison of sum-product and max-product belief propagation 61
4.3.4 Performance for varying training and test sets 62
4.3.5 Run-time performance 63
4.4 Comparison to State of the Art 64
4.4.1 Pedestrian detection 64
4.4.2 Upper body pose estimation 68
4.4.3 Full body pose estimation 73
4.5 Conclusion 76

5 Monocular 3D Pose Estimation and Tracking by Detection 79
5.1 Introduction 80
5.2 Multiview People Detection in Single Frames 83
5.2.1 Basic pictorial structures model 83
5.2.2 Multiview Extensions 84
5.3 2D Tracking and Viewpoint Estimation 87
5.4 3D Pose Estimation 89
5.4.1 3D pose estimation for longer sequences 91
5.5 Experiments 91
5.5.1 Evaluation on the “HumanEva II” dataset 92
5.5.2 3D pose estimation in street scenes 94
5.6 Conclusions 94

6 Vision Based Victim Detection from Unmanned Aerial Vehicles 97
6.1 Introduction 97
6.2 System Overview 99
6.3 Vision-based People Detection 101
6.3.1 Proposed extensions 103
6.4 Experiments 105
6.4.1 Experimental setup 105
6.4.2 Comparison of people detection methods 106
6.4.3 Integration of scale prior 108
6.4.4 Combination of multiple detectors 110
6.5 Conclusion 112

7 Conclusion and Future Work **113**
7.1 Summary . . . 113
7.2 Perspectives for People Detection . . . 114
7.3 Perspectives for 2D Pose Estimation . . . 115
7.4 Perspectives for Monocular 3D Pose Estimation . . . 117

List of Figures **123**

List of Tables **126**

Bibliography **127**

1

Introduction

Es hört doch jeder nur, was er versteht.

Johann Wolfgang von Goethe,
from *Maximen und Reflexionen*

Computer systems are ubiquitous in our everyday life. These systems take on different forms, including traditional personal computers, mobile phones, robotic assistants and embedded devices. However, what is common to the majority of these systems is that they are designed to help us achieve more in less time and make our life more enjoyable and fulfilling. The ability of these systems to realize their purpose depends on how well they can interact with and perceive the human environment and how well they can interpret the motivations, intentions and actions of people. It is therefore an important research objective to develop theories and methods that enable computer systems to see patterns and recover high-level representations of the world from low level sensor measurements, such as sound and camera images.

Building computers that see is therefore a fundamental goal and an important milestone on the way to building autonomous and intelligent computer systems. Seeing may mean many things in this context. It may mean that computers can reliably recover structures contained in the image at various levels of detail. At the coarse level, this could include objects in the scene, important scene surfaces such as walls and pathways, and the rough 3D scene geometry. At the finer scale, the objective is to obtain more detailed descriptions of each of the scene's objects to understand what parts they are built from and what physical properties they might have. Seeing might also mean understanding the visual scene, which allows the system to predict its evolution over time, and the ability to explain these predictions.

This thesis is about developing the models and algorithms necessary for enabling computers to understand images that contain people. In general, the goal of image understanding is difficult to define precisely. This is because the type of the problem one is trying to address often dictates the kind of scene interpretation one would like to achieve. For example, in an automotive scenario, it might be sufficient to recover only positions of pedestrians and the drivable surface in the image, while correctly interpreting social situations such as those shown in Fig. 1.1 might require both knowing where people are in the image and what poses they assume. Furthermore, a household robotic assistant could be interested in understanding the emotions and intentions of the surrounding people, which might require detailed information about their 3D pose and facial expression. Depending on the task, different levels of detail may be necessary. In the automotive example, it might be helpful to know a person orientation with respect to the camera in addition to his or her position in order to infer whether the person is attempting to cross the street and when and where that might occur. In this thesis, in order to address these diverse requirements, we develop methods for people detection and human pose estimation at different levels of detail both in still images and in image sequences.

Visual people detection and pose estimation are rather challenging problems. Fig. 1.2 shows several example images containing people, which are ordered from left to right by their complexity for pose estimation. The ordering also reflects our intuition about the relative difficulty of the challenges encountered in visual people detection. People typically wear a large variety of clothing and are capable of assuming numerous body poses. Complex body poses mean that body parts will often be fully or partially occluded. The appearance and dimensions of individual body parts also change significantly depending on the body shape, pose, viewpoint, and imaging conditions. Jointly, these factors result in complex appearance patterns that are difficult to represent and detect. In addition, the detection of people is a special case of generic object category detection, so it inherits all the difficulties common to this problem, such as the necessity to deal with large amount of background clutter, the limited number of training examples and the variability of lighting conditions. Due to the complex layout and large appearance variability of body parts, it is very difficult to determine which image regions belong to body parts of the person and which are merely background clutter. This problem becomes even more complex when multiple people are present in the image. Finally, even assuming that this problem is solved and we managed to locate the image positions of body parts, we are still facing the ambiguities inherent to recovering a 3D body pose given its image projection. Indeed, we have to get many things right in order to correctly interpret images containing people in unconstrained settings.

As is common in cases where the ultimate problem is very difficult, the immediate research often is focused on various special cases for which at least some progress seems possible in the near future. Such specialization brings various additional constraints which, while unapplicable in general, often lead to interesting practical results. One important special case of generic people detection is **pedestrian detection**. In this thesis, we consider this problem in Chapters 3 and 5. While people walking on the street still vary greatly in their appearance, we can safely assume that they appear in the images in the upright position with their body part configurations constrained by their walking motion. This allows us to learn compact models for both the global appearance of pedestrians and their body parts as well as for the spatial layout of the body configurations. As we demonstrate, we can use such models to go beyond simply detecting people in images and employ them for **3D pose estimation** and tracking of specific people in image sequences that contain multiple people frequently occluding each other.

In practice, any constraints on the possible body configurations and viewpoints of people appear to be helpful. For example when looking for injured people laying on the ground in **search and rescue** scenarios, the position and orientation of each person is restricted by the ground plane. This restriction makes people detection in this setting tractable. At the same time, this setting is still very interesting from the research standpoint because it exhibits such challenges as arbitrary body articulations and body part occlusions, neither of which is handled satisfactorily by state-of-the-art people detectors. We explore this setting in Chapter 6. We demonstrate that algorithms based on learning the global appearance of people, which are successful for pedestrian detection do not perform well in this setting. Instead, local approaches based on the assembly of partial body configurations are necessary. To this end, we apply an approach based on our model for people detection and **2D pose estimation** that we present in Chapter 4. This approach is the furthest we got on our quest towards generic people detection and pose estimation. It still operates only with 2D part configurations and does not explicitly model foreshortening or occlusions of body parts [1]. Nonetheless, we show that it improves over the previous methods that were developed specifically for special cases such as pedestrian detection and **upper body pose estimation**.

One of the general themes pursued in this thesis is the development of algorithms capable of recovering increasingly fine-grained descriptions of people seen in images. As a first step, in Chapter 3, we have developed a new method for pedestrian detection that provided estimates of image scale and position for each detected person. In Chapter 4, we developed a new approach to 2D pose estimation that is also capable of estimating the 2D positions and orientations of the body parts. Finally, in Chapter 5, we proposed an approach that in addition to detection also recovers viewpoints of people, estimates their 3D poses and tracks individual people in crowded street scenes.

Another theme underlying our work was to move from restricted to more generic scenarios for people detection and pose estimation. Our initial approach, introduced in Chapter 3, was applicable to pedestrians seen from the predefined viewpoint only. In the course of our work we have developed an approach for pose estimation of people in arbitrary environments (Chapter 4) and an approach to multi-view pedestrian detection (Chapter 5). In Chapter 6, we have also developed an approach for people detection in “search and rescue” scenarios, which is significantly more difficult and general setting compared to pedestrian detection.

1.1 Contributions

This thesis makes several contributions in the area of people detection as well as in 2D human pose estimation, 3D human pose estimation and multi-person tracking.

---

[1] In the case of people detection in search and rescue scenarios we manage to escape these issues by employing several realizations of our model, each with different number of body parts.

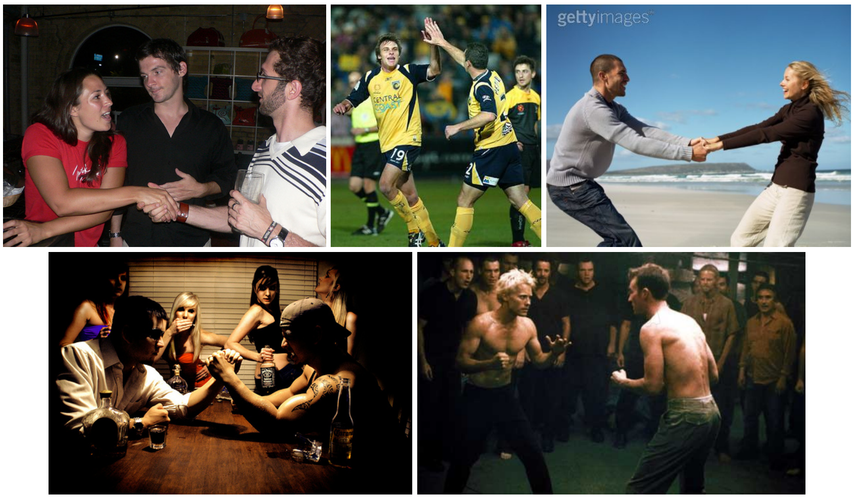


Figure 1.1: Several examples of images containing two people interacting with each other. Notice that the poses of the people often provide essential cues for determining the type of interaction.

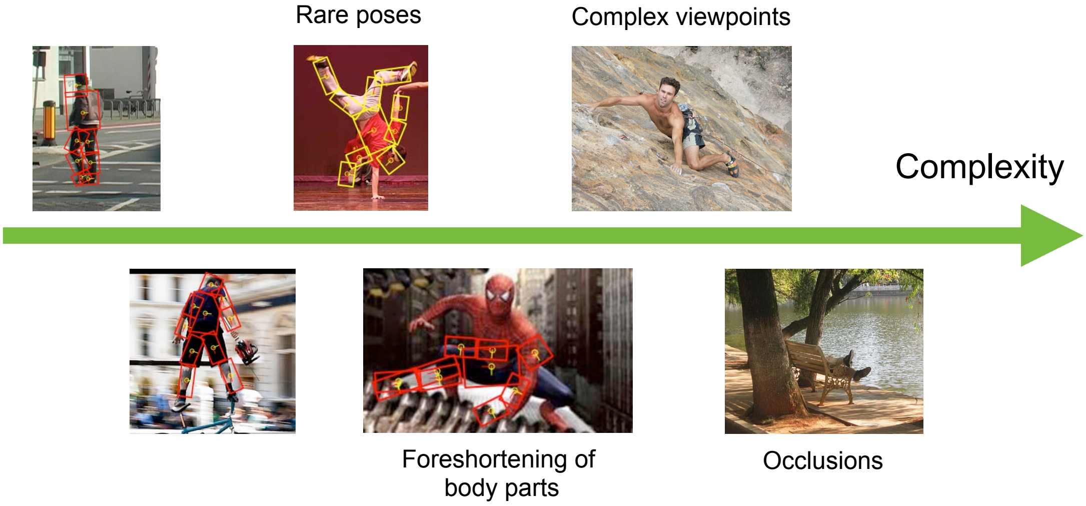


Figure 1.2: Several examples of images with people that illustrate some of the challenges for people detection and articulated pose estimation. The images are ordered from left to right by their complexity for 2D pose estimation as measured by the percentage of body parts correctly localized by the approach of Andriluka et al. (2009).

We propose two novel approaches for people detection and 2D pose estimation. Both of them build on the *pictorial structures* framework of Fischler and Elschlager (1973) and Felzenszwalb and Huttenlocher (2005), but they differ in the appearance model and the inference algorithm. In the first approach, partISM, we employ the implicit shape model (ISM) (Leibe et al., 2004) in order to represent the appearance of body parts, and we subsequently combine part evidence using a generalized distance transform. We evaluate our approach on the task of pedestrian detection, where it shows improved performance over the previously proposed approach of Dalal and Triggs (2005), which uses a global appearance representation, and over the approach of Seemann and Schiele (2006), which uses a similar ISM-based appearance representation, but does not rely on a part-based model. The important advantage of our approach is that, in addition to detecting people, it also provides evidence for their body configurations. By employing long range dependencies between local image features and positions of body parts, our approach is also capable of handling the partial occlusions that frequently occur in crowded street scenes.

Our second approach to people detection builds on a tree-structured body model and represents the appearance of individual body parts with densely sampled local features and discriminative classifiers. Our approach further pushes the state-of-the-art in pedestrian detection by improving over the previous partISM approach and the more recent approach of Gall and Lempitsky (2009). The important advantage of our second approach is its generality. In addition to pedestrian detection, it is also applicable to upper-body pose estimation and full-body pose estimation, and it significantly improves over results of Ferrari et al. (2008) and Ramanan (2006) on standard benchmark datasets. In chapter 6, we also apply this approach to the detection of arbitrarily articulated people from the on-board camera of an unmanned aerial vehicle. In order to foster further research on human detection and pose estimation, we have made the source code of our implementation publicly available[2], which has enabled other researchers to use it in their work (Freifeld et al., 2010; Yao and Fei-Fei, 2010; Eichner and Ferrari, 2010; Stark et al., 2010*a*) or directly compare their results to ours (Karlinsky et al., 2010; Sapp, Jordan and Taskar, 2010; Sapp, Toshev and Taskar, 2010; Tian and Sclaroff, 2010*a*,*b*).

As a second major contribution of this thesis, we have developed a *tracking-by-detection* framework that builds on our results in single frame people detection and 2D pose estimation. Given a monocular image sequence obtained in uncontrolled street conditions with an uncalibrated and potentially moving camera, our framework is capable of automatically finding people, tracking them over time and estimating their body poses. In this framework, we combine the initial position and articulation estimates with the non-parametric dynamical model of Lawrence and Moore (2007) in order to obtain short-term tracking hypotheses (tracklets), and we subsequently use them to infer tracks of multiple people in crowded street scenes. The important property of our approach is its ability to deal with long-term occlusions of people, which we handle with the help of person-specific appearance models

---

[2] `http://www.d2.mpi-inf.mpg.de/code`

obtained using estimates of body configurations.

Our third contribution is an approach to 3D pose estimation of pedestrians in monocular image sequences, which builds both on our tracking-by-detection framework and on our results in 2D pose estimation. We propose a novel formulation of the 3D pose likelihood that relies on the estimates of the 2D body position, the 2D body pose, and the viewpoint obtained from a bank of viewpoint-specific pictorial structures models. Contrary to many approaches in the literature (e.g., (Ormoneit et al., 2001; Sigal et al., 2004)), we avoid the assumption of conditional independence of 2D evidence in each frame of the sequence. Instead, we proceed by integrating 2D evidence over multiple frames using temporal consistency, subsequently using the improved 2D estimates in order to define the 3D pose likelihood. This approach allows us to significantly reduce the number of spurious local optima in the 3D likelihood and also provides an effective means for initializing the 3D pose estimation procedure. We evaluate our approach on a standard benchmark in 3D human pose estimation where it compares favorably to the state of the art and demonstrate its applicability to images taken in uncontrolled street conditions with static and moving cameras.

1.2 Outline

In this section, we briefly outline the structure of this dissertation.

Chapter 2 reviews the literature on people detection, tracking and human pose estimation. We present the most prominent approaches relevant to our work, highlight commonalities and differences between them, and discuss various paradigms in representation, modelling and detection of people in images and image sequences.

Chapter 3 presents a partISM approach to people detection and pose estimation based on the *implicit shape model* (Leibe et al., 2004) and *pictorial structures model* (Felzenszwalb and Huttenlocher, 2005). This chapter also introduces a *tracking-by-detection* framework that combines detection, tracking and pose estimation in order to reliably track people in crowded street scenes with frequent full and partial occlusions.

Chapter 4 introduces a generic model for people detection and articulated pose estimation. Similarly to partISM, it builds on the pictorial structures model. However, it uses discriminative appearance representations and a more flexible kinematic tree prior on human poses. We perform a detailed evaluation of various aspects of the appearance model, such as the type and parameters of the local descriptors, type of belief propagation algorithm, and compare tree and non-tree versions of our approach. We directly compare our appearance model to the template-based appearance model of Ramanan (2006) by integrating it into our system and demonstrating that our approach to appearance modeling leads to significant improvements in performance. Due to its flexible appearance and pose representation, our approach is applicable to a wide variety of tasks. We demonstrate this by applying it to pedestrian detection and frontal upper body and generic full body pose estimation. In each case, we use previously proposed datasets and show that our approach improves over the state of the art, outperforming methods designed specifically for each of these tasks.

Chapter 5 describes an approach to 3D pose estimation of pedestrians in realistic street conditions. Leveraging the results from Chapter 4, we develop a novel approach to pedestrian detection and viewpoint estimation that builds on a bank of part-based viewpoint-specific detectors. We integrate position and viewpoint estimates over time and use them within the *tracking-by-detection* framework introduced in Chapter 3 in order to recover 3D poses of people.

Chapter 6 presents the application of our approach from Chapter 4 to the task of people detection from unmanned aerial vehicles. In order to further improve the performance of our approach in this novel setting, we augment it with on-board sensor measurements, which we use to obtain an informative prior on the scale of people in images and cope with the effects of perspective distortion. We also demonstrate the effectiveness of combining multiple models for coping with partial occlusions.

Chapter 7 presents concluding remarks and sketches possible directions for future work.

2

Related Work

> *Fifteen years of research he'd filed from one planet alone and they'd cut it to two words. "Mostly Harmless."*
>
> *Douglas Adams*, from *Mostly Harmless*

The *pictorial structures* approach is central to the work presented in this thesis. Therefore, we begin this chapter by introducing it in Sec. 2.1, and then proceed by describing its applications in the literature and comparing the pictorial structures approach with other approaches to people detection in Sec. 2.2 and 2D human pose estimation in Sec. 2.3. In Sec. 2.4 we review the state of the art in 3D pose estimation, and conclude this chapter with Sec. 2.5, in which we survey the literature on people detection using other sensors and sensor fusion. When applicable, we also refer to the literature on people tracking, both in the 2D and 3D setting, and relate it to our work on people tracking presented in Chapters 3 and 5.

In addition to reviewing the literature related to the pictorial structures model and the state of the art in pose estimation, detection, and tracking, we pursue two additional goals in this chapter. On the one hand we highlight the currently established *benchmarks* in computer vision that are relevant for people detection and pose estimation. For each of the tasks considered in this thesis, there are established datasets and evaluation criteria that allow for the direct comparison of various approaches with each other. Such benchmarks play an important role in the computer vision research, as they allow the research community to focus their efforts and monitor their achieved progress.

As a second goal, we highlight the relationship of the described approaches to various paradigms in visual data modeling that have been established over the years in computer vision literature. In particular, we indicate whether the presented approaches are *model-based* or *model-free*, employ *top-down* or *bottom-up* inference, and *generative* or *discriminative* modeling. Whenever possible, we also point out how these paradigms are combined in various settings.

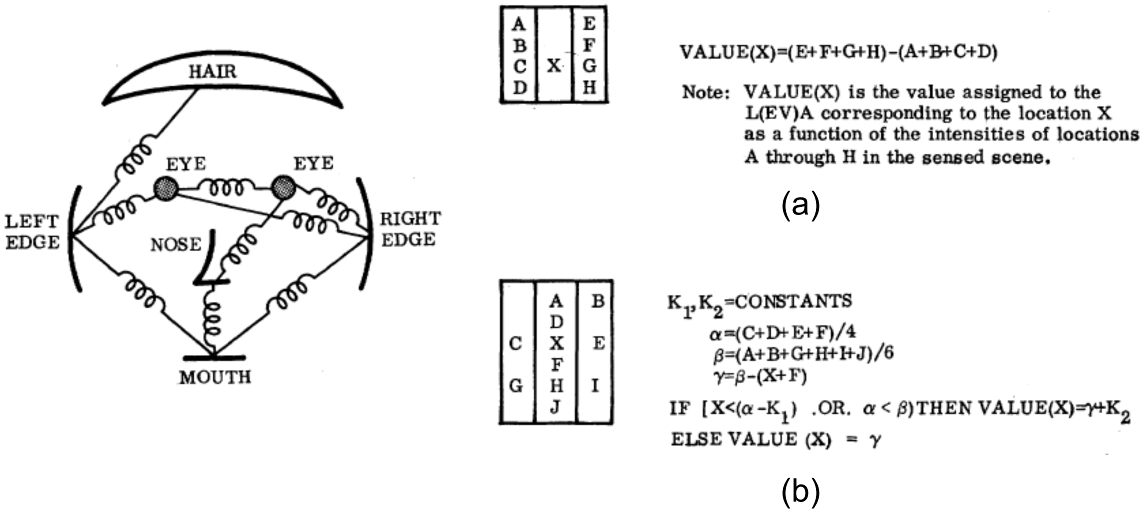

HAIR
EYE
EYE
LEFT EDGE
NOSE
RIGHT EDGE
MOUTH
A B C D
X
E F G H
VALUE(X)=(E+F+G+H)-(A+B+C+D)
Note: VALUE(X) is the value assigned to the L(EV)A corresponding to the location X as a function of the intensities of locations A through H in the sensed scene.
(a)
C G
A D X F H J
B E I
K1,K2=CONSTANTS
α=(C+D+E+F)/4
β=(A+B+G+H+I+J)/6
γ=β-(X+F)
IF [X<(α-K1) .OR. α<β)THEN VALUE(X)=γ+K2
ELSE VALUE (X) = γ
(b)

Figure 2.1: Diagram showing spatial relations between parts of the face, and likelihood functions (or reference description) for the left edge of the face (a) and the eye (b). This figure is reproduced from the original pictorial structures work of Fischler and Elschlager (1973).

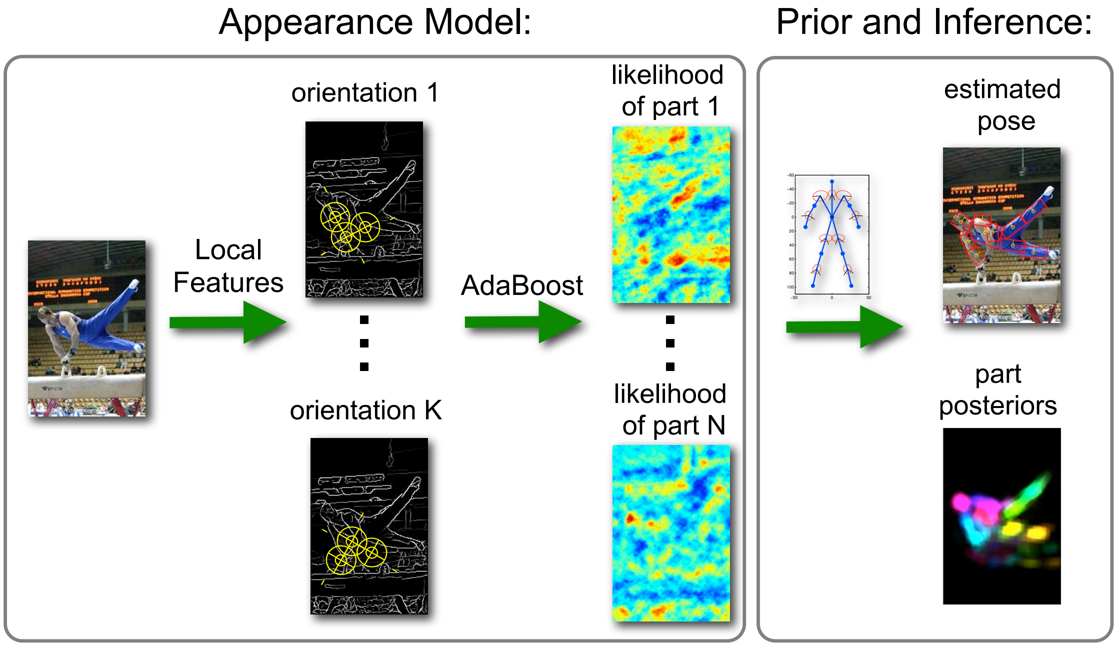


Figure 2.2: Diagram showing the main components of the pictorial structures based people detection and pose estimation system proposed here.

2.1 Pictorial Structures Approach

The pictorial structures model, originally introduced by Fischler and Elschlager (1973), has a long history in computer vision. The key ideas of representing objects as a set of rigid components with pairwise constraints on their spatial layout, and matching algorithms based on dynamic programming were introduced already in the original work of Fischler and Elschlager (1973). Fig. 2.1 shows an example of the face model considered in their work, along with the corresponding likelihood functions (referred to as reference description) for two of the model parts. Fischler and Elschlager discuss the principal difficulties of finding the arrangement of parts that maximizes the criteria combining the likelihood of individual parts and pairwise terms representing spatial constraints on part positions. As they point out, the structure of the part dependencies has strong implications on the complexity of the inference. In the case of part dependencies without loops optimal solution can be found in polynomial time with the algorithm based on the dynamic programming. This algorithm exploits conditional independencies between part positions in order to perform computations efficiently and is not applicable in cases where such dependencies have a loopy structure (e.g., as in the case of the face model in Fig. 2.1). In order to deal with the generic case, Fischler and Elschlager propose a heuristic approach that, while not guaranteed to find the optimal solution, appears to work well in their experiments.

The pictorial structures approach naturally lends itself to a *probabilistic interpretation*. Denoting the location of part i by $\mathbf{l}_i$, the joint configuration of parts by L, and the evidence in the form of image features by E, the problem of finding the optimal configuration of parts corresponds to finding the maximum of the posterior distribution

$$p(L|E) \propto P(E|L)p(L). \tag{2.1}$$

The assumption that the appearance of a part is independent of the other parts corresponds to the decomposition of the likelihood term into the product of individual part likelihoods

$$p(E|L) = \prod_i p(E|\mathbf{l}_i). \tag{2.2}$$

This probabilistic interpretation of pictorial structures served as a basis for the development of *constellation models* (Burl et al., 1998; Fergus et al., 2003) and *tree-structured pictorial structures* models (Felzenszwalb and Huttenlocher, 2005). The latter, in addition to the likelihood decomposition in Eq. 2.2, assumes that prior distribution $p(L)$ has a tree structure and can be decomposed into the product of unary and pairwise terms

$$P(L) = p(\mathbf{l}_0) \prod_{(i,j)\in G} p(\mathbf{l}_i|\mathbf{l}_j). \tag{2.3}$$

The probabilistic interpretation of pictorial structures also suggests inference algorithms for finding optimal locations of individual parts and their joint configurations.

The best location of a particular part i is given by the maximum of the marginal posterior distribution

$$p(\mathbf{l}_i|E) = \sum_{L \setminus \mathbf{l}_i} p(L|E), \tag{2.4}$$

which in the case of a tree-structured model can be computed in polynomial time with sum-product belief propagation algorithm, and the optimal configuration of body parts can be found efficiently with max-product belief propagation (Pearl, 1988; Bishop, 2006). In the more generic case, the posterior marginals can be approximated using loopy belief propagation (Frey, 1998), or a number of alternative approximate inference techniques (Jordan et al., 1999; Yedidia et al., 2000).

The domain of the part locations $\mathbf{l}_i$ plays an important role in the choice of inference algorithm. In the case where the set of possible part locations is finite and relatively small and the model has only a few parts, exhaustive enumeration of the state space of each part is possible. An example of such a case is the constellation model that identifies the set of potential part locations using an interest point detector. In the case where the domain of $\mathbf{l}_i$ is continuous, the solution can be found using sampling-based techniques (e.g. non-parametric belief propagation (Sudderth et al., 2003; Isard, 2003)). Sampling-based techniques are also applicable in the case of a discrete representation and allow tractable approximate inference even when the number of candidate part locations becomes large (Stark et al., 2010*b*).

In many cases it is also possible to represent the domain of $\mathbf{l}_i$ as a finite multidimensional grid by discretizing it along each dimension. This approach avoids the need to preselect a set of candidate part locations as is necessary in the constellation models and allows to perform inference exactly in contrast to sampling-based techniques. Felzenszwalb and Huttenlocher (2005) have employed such a grid-based representation in conjunction with a tree-structured prior distribution on part configurations and a Gaussian assumption on the relationship between body parts in order to further reduce inference cost. Given these additional assumptions, the MAP estimate of the body configuration can be found with generalized distance transform in the time linear in the number of discrete grid locations (Felzenszwalb and Huttenlocher, 2004). Similarly, one can also efficiently compute the marginal posterior distribution for the position of each body part using Gaussian convolution. Another important contribution of Felzenszwalb and Huttenlocher (2005) is the application of the pictorial structures model to 2D articulated pose estimation, and the description of a joint transformation that permits to perform linear time inference, even when part dependencies are non-Gaussian in the spatial domain. In the context of belief propagation, the approach to linear time inference introduced by Felzenszwalb and Huttenlocher (2005) can be interpreted as a way to quickly compute messages between model variables during message passing.

The pictorial structures model specifies the decomposition of the object into parts, constraints on the relative locations of parts, and criteria for judging the quality of the part assembly. For example, in the case of 2D human pose estimation, the representation frequently used in the literature consists of 10 parts, corresponding to torso, head, and left/right legs, upper arms and forearms. Approaches of this type are often referred to as being model-based. At the same time, the pictorial structures approach is flexible with respect to the choice of parameter learning and inference algorithm. In the approach of Felzenszwalb and Huttenlocher (2005) spatial parameters are learned in a generative way, by maximizing the likelihood of the training data, and inference proceeds by combining a bottom-up inference step to compute the part marginals, and a top-down step in which the best sample from the marginal is selected using a Chamfer measure. This procedure effectively combines local foreground segmentation and global shape cues and mitigates effects of the evidence overcounting.

The appearance representation used in the work of Fischler and Elschlager is rather simple and employs local image filters and heuristically constructed part detectors, which are shown in Fig. 2.1 (a,b). Fig. 2.2 shows the main components of our people detection and pose estimation approach based on the pictorial structures model described in more detail in Chapter 4. While the key concepts of the pictorial structures model have remained the same since its original introduction, new appearance representations and inference algorithms have been developed over the years, incorporating advances in generic object-class recognition and machine learning. In the remainder of this chapter we will introduce some of these developments and discuss them in more detail.

2.2 People Detection and Tracking

Much of the recent work on people detection has focused on detection of pedestrians and people in upright orientation. Pedestrians exhibit significantly more regularities in pose and appearance compared to people in general, which makes their detection more tractable. At the same time, the detection of pedestrians is also of great practical importance, with the numerous applications in automotive safety, robotic navigation and surveillance. In this section we proceed by first reviewing the literature on pedestrian detection and then discussing recent approaches to generic people detection and tracking.

2.2.1 Pedestrian Detection

Early progress in the detection of pedestrians was made using global silhouette matching methods, performing detection by exhaustively comparing an image with a hierarchy of silhouette exemplars (Gavrila, 1999, 2000). These methods relied on the distance transform in order to efficiently compute the Chamfer distance between image edges and silhouettes for every possible image position, allowing them to operate in real time. Although computationally efficient, these exemplar matching methods did not exhibit several important properties which underlie the success of more recent detection approaches. Firstly, they did not employ any form of local pooling of gradient information. Local pooling is important for making the appearance model robust to image noise and small local image deformations and was shown to be an important component of optical character recognition systems (LeCun et al., 1989, 1998), and local image descriptors (Lowe, 2004; Mikolajczyk and Schmid, 2005). Silhouette matching methods have to encode such variability either by means of a large collection of exemplars or by directly modeling the variability of silhouettes, as is for example done in the active shape model (Cootes et al., 1995). While both of these methods are applicable in controlled laboratory conditions, they are difficult to implement in realistic street environments due to large amounts of background clutter, a large variability in pedestrian silhouettes and a limited number of training examples. Another drawback of a silhouette-based approach turned out to be the difficulty of combining it with discriminative machine learning techniques. Although Felzenszwalb (2001) has developed an approach to learn discriminative shape-based models, this approach was not further explored in the literature.

Robust image representations and discriminative learning algorithms appear to be the key to recent advances in pedestrian detection. Many of the recently proposed approaches follow the top-down *sliding-window* paradigm to object detection. In this approach an image is exhaustively scanned with a detection window of fixed aspect ratio at a fixed step in scale and position. For each window an independent decision is made about the presence or absence of the person. Since the system is required to examine large numbers of overlapping windows, the features describing each window are typically designed so that they can be computed quickly or can be reused for representation of the windows that are close to each other. Early work in this direction (Oren et al., 1997; Papageorgiou and Poggio, 1999) relied on Haar wavelet features to describe window contents and support vector machine for window classification (Vapnik, 1999). The Haar wavelet features correspond to differences between the sums of pixels in the rectangular image areas and can be computed quickly using the integral image representation. At the same time, these features are also robust to noise and image transformations since image intensities are aggregated from the entire rectangular region. Similar rectangular filters were used in (Viola et al., 2003), who build on their earlier work on face detection, and employed AdaBoost (Freund and Schapire, 1997) in order to identify the important features and combine them into strong classifiers.

An approach based on the histograms of oriented gradients (HOG), introduced by Dalal and Triggs (2005), follows a similar sliding-window paradigm. However, the careful design of image representation allowed it to achieve significant improvements in performance over the previously proposed approaches of this type. The HOG representation is based on the local gradient histograms, which are computed for a set of rectangular regions denoted as *cells* that densely conver the entire area of the detection window. The gradient of image intensity at each pixel contributes to each of the histograms of its adjacent cells through trilinear interpolation. The cells are grouped into overlapping *blocks*, and cell histograms are normalized jointly for each block. There are several aspects of this approach that made the HOG representation successful. The representation is based on the local distributions of gradient locations and orientations which appear to be more effective for modeling object appearance compared to Haar features, while at the same being robust to noise and intra-class variability. The HOG representation encodes the contents of the entire window, allowing the classifier decide which features are important. Interpolation of the gradient information between histogram bins makes the representation robust to in-plane translations and rotations. Block normalization makes the HOG descriptor robust to illumination changes and encodes the relative strength of gradient in each cell with respect to different subsets of its neighbors.

The approach of Dalal and Triggs (2005) achieved near perfect results on the *MIT Pedestrian* dataset (Oren et al., 1997), which contains images of pedestrians predominantly in front and back views rescaled to the same scale and cropped around the person bounding box. In order to evaluate their approach in a more challenging setting Dalal and Triggs (2005) introduced the *INRIA Person* dataset, which remains and important benchmark in pedestrian detection to date. This dataset contains full images of pedestrians at multiple scales and seen from a variety of viewpoints, and in addition includes a separate set of images without pedestrians to asses the robustness of the detector to background clutter.

Much of the recent work on pedestrian detection has been using HOG-based detector of Dalal and Triggs (2005) as a baseline, and focused on the development of new image features (Dollár, Tu, Perona and Belongie, 2009), combining appearance with new sources of information, such as optical flow (Dalal et al., 2006) and image disparity (Walk, Schindler and Schiele, 2010), and more powerful approaches to classification (Wojek et al., 2009). These developments led to introduction of new datasets such as “ETH” (Ess et al., 2007) and “TUD-Brussels” (Wojek et al., 2009) that permit computation of stereo- and motion-based features. Recent datasets also address important applications of pedestrian detection. For example, *Caltech Pedestrian* (Dollar, Wojek, Schiele and Perona, 2009*a*), and *Daimler Detection Benchmark* datasets (Enzweiler and Gavrila, 2009) specifically address the automotive scenario and focus on challenges encountered in this setting, such as detection of partially occluded people and detection of people at very small scales.

While demonstrating excellent performance, top-down approaches based on monolithic representations (such as HOG) have difficulties handling detection of people in the presence of occlusions and in cases where people exhibit especially large pose variability. In order to handle these challenges, several approaches have been developed that build on more flexible part-based representations that represent object via a collection of rigid parts and their spatial relationships. Part-based approaches differ significantly in what is considered to be a model part, how the appearance of each part is represented, and what structure is imposed on part configurations (Murphy, 2005). These approaches include ad-hoc decomposition of the detection bounding box in rectangular regions (Mohan et al., 2001), association of model parts with anatomically meaningful parts such as torso, arms and legs (Andriluka et al., 2009) and automatic mining of parts using image statistics (Leibe et al., 2005) or their discriminative properties (Felzenszwalb et al., 2008; Bourdev and Malik, 2009).

An important example of the part-based approach to pedestrian detection is the Implicit Shape Model (ISM) of Leibe et al. (2005). In this approach visual appearance of a person is represented using a codebook of *visual words*, along with the learned spatial relationships between each visual word and the center of the person bounding box. Inference proceeds in bottom-up fashion using the generalized Hough transform. In the first stage, local regions found with the interest point detector are matched to the codebook, and each match is used to cast probabilistic votes for the center of the bounding box according to the learned spatial distribution. People hypothesis in the image correspond to the local maxima in the voting space and can be effciently found with the mean-shift algorithm. The ISM approach is able to deal with intra-class variability and changes in appearance due to articulation by encoding them via a large collection of visual words. Since each visual word captures only local information and each image feature contributes to the overall hypothesis score independently this approach is robust to partial occlusions. The original ISM approach introduced in (Leibe et al., 2004, 2005), has been further developed by Seemann and Schiele (2006), who introduced additional articulation-based constraints into the model, and by Gall and Lempitsky (2009) and Maji and Malik (2009*b*), who have developed methods to discriminatively learn the weights of the visual words. The ISM approach also generalizes to situations when multiple people are present in the image by explicitly modeling relationships between people hypothesis and image features and performing additional inference steps that ensure that each feature contributes to one person hypothesis only (Leibe et al., 2008; Barinova et al., 2010).

2.2.2 People Detection

Perhaps due to its difficulty, the problem of generic people detection in unconstrained environments has only recently shifted into the focus of attention of the computer vision literature. People in this setting exhibit a much larger pose variability compared to pedestrians, making it challenging to learn a single template-based model for their appearance.

Recall that the pictorial structures model introduced in Sec. 2.1, permits the representation of complex appearance patterns in terms of appearance of a fixed number of components and their spatial arrangement. Such representation seems appealing for generic people detection, since people can be viewed as a set of rigid components corresponding to anatomic body parts, connected with each other by body joints. However, anatomic body parts are not guaranteed to be easily detected in the images. For example, parts such as hands and forearms are relatively small and will be frequently occluded or foreshortened. It appears that better models can be obtained using generic parts that do not carry any semantic information and are selected purely by their visual saliency. Recently Felzenszwalb et al. (2009) have proposed an approach for generic people detection that builds on the pictorial structures model and is capable of discovering model parts in an unsupervised manner. Given a rough initialization, their algorithm employs an iterative procedure which interchangingably estimates the locations of parts and the parameters of the spa-

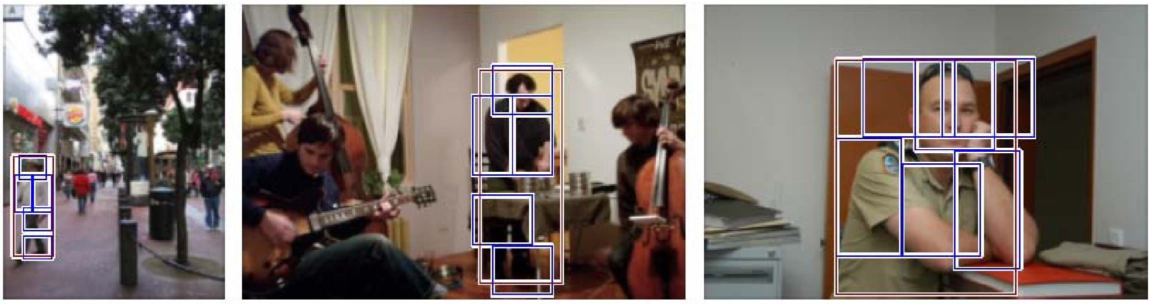


Figure 2.3: Several examples of detections obtained with the model of Felzenszwalb et al. (2009).

tial and appearance models. Note that according to Eq. 2.2 the pictorial structures model represents object likelihood as a product of local part likelihoods. Consequently, low likelihood of even one of the parts may lead to a low likelihood for the whole configuration. Such low part likelihood may be the result of occlusion, image noise or bad contrast. By choosing parts automatically, these effects can be taken into account at the training stage, which leads to a more robust model. In addition, automatic part discovery allows the application of the model to various other object classes without the need for a human expert to determine the model parts.

The pictorial structures model does not explicitly take occlusion of parts into account and expects that all model parts are simultaneously present in the image. In practice, this results in models with relatively few generic parts where each part appears to be rather insensitive to fine details of the object's appearance. Fig. 2.3 shows several examples of detections obtained with the model of Felzenszwalb et al. (2009). Note how in the first and second images the same part is matched to two different configurations of the upper legs. Also note that the part corresponding to head and shoulders captures various head poses and viewpoints. Although such generic parts are robust to image noise and pose variations, they are also uninformative for particular body configuration and hence are not suitable as evidence for pose estimation. In addition, appearance models of such generic parts disregard important features if they appear in only a few body configurations.

In order to address these shortcomings Bourdev and Malik (2009) have proposed a part-based model which, in contrast to six parts used in (Felzenszwalb et al., 2009), relies on hundreds parts. These parts, denoted as *poselets*, correspond to regions with stable appearance that are also informative for the underlying partial body configurations. Individual poselets are discovered in an unsupervised manner, using a dataset of images of people with detailed annotations of 3D body poses. There is a strong similarity between the model of Bourdev and Malik (2009) and the ISM model of Leibe et al. (2005). In particular, similarly to ISM, scores of individual poselets are combined using the generalized Hough transform, and the poselets themselves can be seen as a kind of appearance codebook. The crucial difference lies in the appearance model for individual poselets, which relies on HOG representation and discriminative learning. In addition poselets typically cover a significantly larger area then local features used in (Leibe et al., 2005), and can therefore be detected more reliably.

It is interesting to note that the models proposed in (Bourdev and Malik, 2009) and (Felzenszwalb et al., 2009), both build on the components that have proven to be effective for detection of pedestrians. In particular, both of them use the HOG appearance representation, SVM classifiers, and sliding-window paradigm to detect model parts. At the same time they rely on deformable part models, such as pictorial structures and ISM, to represent part configurations. In this thesis, following a similar strategy, we proceed by first developing a part-based approach to pedestrian detection. We subsequently use the components of this approach for human pose estimation and tracking.

2.2.3 People Tracking

Visual object tracking deals with the problem of detection and trajectory estimation of objects in image sequences. Many of the tracking algorithms are online, i.e. they sequentially process the incoming images, maintaining states of the tracked objects and updating these states over time. Although such online algorithms are essential in real-time applications such as automotive safety and robotic navigation, there also exist numerous cases in which the whole image sequence is available in advance. For example, in surveillance or video indexing one typically has access to large volumes of image data and can jointly infer object states over several subsequent images. The principal components of any tracking algorithm area an *appearance* or *likelihood model*, which reflects the probability of observing an object in the image given its state, and a *dynamical* model, which allows for the prediction of the object's position in future frames and also gives the criterion for the temporal compatibility of object states in several subsequent frames.

A large body of the tracking literature is dedicated to a class of *model-free tracking* algorithms that avoid using a-priori assumptions on the type of the objects being tracked. The algorithms of this type frequently rely on manual initialization, although other methods for initialization such as background subtraction or visual and motion coherence exist (Schiele, 2005; Yilmaz and Shah, 2006). Since the appearance of the tracked object is likely to change significantly over time, model-free tracking algorithms maintain an on-line appearance model that is dynamically updated with each incoming image.

Handling of rapid changes in appearance and recovery from full and partial occlusions are among the key difficulties for tracking. In order to address them in a principled manner, several authors have proposed tracking methods that combine static and dynamically updated models of appearance (Grabner et al., 2008), or rely entirely on appearance models in the form of object detectors. The latter strategy is often referred to as *tracking-by-detection*, and is effective when the class of tracked objects is known a-priori as, for example, in the case of people tracking-by-detection (Okuma et al., 2004; Wu and Nevatia, 2007; Leibe et al., 2007; Andriluka et al., 2008; Breitenstein et al., 2009*b*).

Approaches to people tracking vary significantly with respect to their representation of the person state, ranging from a simple bounding box up to a realistic model of the body shape (Balan, Sigal, Black, Davis and Haussecker, 2007). The bounding box representation is often used in the context of visual surveillance in urban environments, where the focus is mainly on tracking of walking and standing people. Due to recent progress in the detection of people in such conditions (see Sec. 2.2.1), tracking can often be implemented by simply linking the detection hypotheses throughout the image sequence. Therefore, the active research in this area has shifted towards tracking of multiple people and handling of occlusions (Isard and MacCormick, 2001; Leibe et al., 2007; Breitenstein et al., 2009*b*,*a*; Kuo et al., 2010). There is also a large body of literature on *articulated people tracking* that aims at the recovery of both positions and poses of people in image sequences. We will revisit this literature in Sections 2.3 and 2.4 together with other approaches to human pose estimation.

2.3 2D Human Pose Estimation

We begin the review of computer vision literature on 2D human pose estimation by discussing the applications of pictorial structures to this problem, following it with a brief survey of other related work and a discussion of applications of pose estimation to high-level image understanding tasks.

As we have discussed in Sec. 2.1, the pictorial structures approach consists of two main components: an appearance model for individual parts and a spatial model that describes pairwise part relationships. Much of the research since the original publication of Fischler and Elschlager (1973), has focused on the improvement of these components, but also on the development of novel learning and inference algorithms and approaches that go beyond pairwise dependencies between parts.

The appearance model used for human pose estimation in (Felzenszwalb and Huttenlocher, 2005) consisted of simple rectangular part templates and relied on background subtraction. In static images with a cluttered background, or in videos in which the background is constantly changing due to camera motion and multiple moving objects, reliable background subtraction is difficult. Therefore, the method of Felzenszwalb and Huttenlocher (2005) has been applied to images with relatively clean backgrounds taken under laboratory conditions. In order to address this shortcoming, Ronfard et al. (2002) have developed a discriminative appearance model, while still building on the simple image features based on Gaussian derivatives. Ramanan (2006) proposed to extract more powerful part templates using an iterative parsing approach. His approach was later extended by Ferrari et al. (2008, 2009*b*) and Eichner and Ferrari (2009), who furthermore integrated features from an automatic foreground segmentation step to improve performance. The latter two approaches iteratively build more powerful detectors to reduce the search space of valid articulations but use relatively weak edge cues at the initial detection stage.

In (Andriluka et al., 2009), we have proposed an appearance model built on strong discriminatively-trained part detectors that do not require iterative parsing or other ways of reducing the search space, other than of course an articulated body model. In our model, we employ dense appearance representations based on the local image descriptors (Belongie et al., 2000; Lowe, 2004; Mikolajczyk and Schmid, 2005), and use AdaBoost (Freund and Schapire, 1997) to train discriminative part classifiers. Strong detectors of this type have been commonplace in the pedestrian detection literature (Viola et al., 2003; Leibe et al., 2005; Mikolajczyk et al., 2006; Andriluka et al., 2008; Gall and Lempitsky, 2009). In these cases, however, the employed body models are often simplistic. A simple star model for representing part articulations is, for example, used in our work (Andriluka et al., 2008) and in (Felzenszwalb et al., 2008), whereas Leibe et al. (2005) and Gall and Lempitsky (2009) do not use an explicit part representation at all. This precludes the applicability of these approaches to strongly articulated people and, consequently, they have been applied to upright people detection only.

There also exists prior work on appearance models for pictorial structures that specifically addresses tracking and pose estimation of people in image sequences. Ramanan et al. (2007) have developed a *tracking by model building and detection* approach. This approach proceeds by detecting people in characteristic poses (e.g. people in side-view with legs far apart), learning discriminative color-based appearance models of body parts from such detections and then relying on these models to recover poses in other frames of the sequence. Color based features used in this model are demonstrated to generalize well to unseen poses of people and are able to capture the characteristic differences between foreground and background color distributions. One of the limitations of this approach is that the people are not guaranteed to appear in predefined characteristic poses, e.g. in many scenes people are never seen in a side-view orientation as is expected by the key-pose detector. The proposed color-based appearance model is also not always effective for discrimination among multiple people present in the scene. In Chapters 3 and 5 of this thesis we build on the ideas similar to Ramanan et al. (2007). However, we also build on our previous results in 2D pose estimation and tracking, which allows us to estimate appearance of subjects from detections in multiple frames and rely on much broader class of body poses as considered in their approach.

In addition to the appearance model, another important aspect of the pictorial structures approach is the representation and structure of the part dependencies. Although kinematic constraints between parts are well captured by the tree-structured model, incorporating appearance and coordination constraints requires a more complex model structure. The focus of recent research has therefore been on the development of models that are capable of incorporating such additional constraints and, at the same time, allowing for efficient inference. In order to address this challenge, several authors have proposed approaches that rely on families of pictorial structure models. Lan and Huttenlocher (2005) have developed the *common factor model* that augments the tree-structured model with an additional latent variable responsible for capturing activity-induced correlations between arms and legs. Common factor models are similar in spirit to k-fan models proposed in (Crandall et al., 2005). Both of them contain just a few core variables that explain most of the non-tree dependencies within the model. Conditioning the common factor model on a value of the latent factor transforms it into a tree, which allows efficient inference. In the test image, the inference proceeds by taking the maximum over the output of multiple tree-structured models where each such model corresponds to a particular discretization value of the common factor.

Recently, Sapp, Jordan and Taskar (2010) have proposed the *adaptive pictorial structures model* (APS), which dynamically chooses the parameters of the prior distribution on part configurations depending on the content of the test image. The key intuition behind this approach is that it is often possible to estimate the rough layout of the body configuration by global comparison of the test image with the dataset of annotated training images. More specifically, the adaptive pictorial structures model employs kernel regression in order to infer the parameters of the spatial prior given image features, and then localizes the parts using standard inference in the tree-structured model. In comparison to the common factor model, APS is more efficient, since inference in the tree model has to be performed only once. On the other hand, the APS model is also dependent on the ability of the kernel regressor to identify a pictorial structure model that is sufficiently close to the true body configuration.

The approach employed in the APS model can be seen as a way to combine global and local appearance models, in which the global model is used to select prior and local appearance model is subsequently used to localize body parts. Interestingly, similar combination of global and local appearance information has been previously used by Felzenszwalb and Huttenlocher (2005). There the local appearance information is used during the first stage to compute distribution over poses, and the global appearance model is subsequently employed to select the best sample from this distribution. The adaptive pictorial structures model is also related to the local probabilistic regression (Urtasun and Darrell, 2008) and to the mixture of experts model (Jacobs et al., 1991). These models approach complex prediction problems by combining output from several local predictors denoted as experts using a data dependent gating function that controls the influence of experts in different regions of the input domain. In the case of APS, the role of experts is played by the pictorial structures models, and the gating function corresponds to the kernel regression component, which selects the most appropriate pictorial structures model given the input image.

Besides the already mentioned related work on 2D human pose estimation with pictorial structures, there exists a large body of literature relevant to this problem. The proposed approaches differ in various aspects, such as whether they are model-free or model-based, discriminative or generative, bottom-up or top-down. These paradigms are often combined within a single approach. For example, discriminative part models have been previously used in conjunction with generative body models. Lee and Cohen (2004) and Sigal and Black (2006*c*) use them as proposal distributions ("shouters") for MCMC or non-parametric belief propagation. In our approach to

2D pose estimation (see Chapter 4), we also rely on discriminative part detectors and generatively learned body model. However, in contrast to the approaches of Lee and Cohen (2004) and Sigal and Black (2006*c*) we use discriminative part detectors directly as the appearance model.

A large body of work has focused on strong body models and inference algorithms, necessary for the reliable assembly of a bottom-up part hypothesis. This work however frequently relied on rather simple appearance models based on edge templates, silhouettes and background subtraction. Such strong body models have appeared in various forms. A certain focus has been the development of non-tree models. Ren et al. (2005) propose a model which imposes constraints not only between limbs on the same extremity, but also between extremities, and rely on integer programming for inference. Another approach incorporating self-occlusion in a non-tree model was proposed by Jiang and Martin (2008). Both approaches rely on the matching of simple line features and have only been shown to work on relatively clean backgrounds. Sigal and Black (2006*b*) also use non-tree models to improve occlusion handling but still rely on simple features such as color. A fully connected graphical model for representing articulations was proposed by Bergtholdt et al. (2009), who also used discriminative part detectors. However, the method has several restrictions, such as relying on absolute part orientations, which makes it applicable to people in upright poses only. Moreover, the fully connected graph complicates inference. Ramanan and Sminchisescu (2006) have proposed discriminative tree-structured models, but due to the use of simple features their method was not shown to perform well for cluttered scenes.

The 2D human pose estimation is also being increasingly more often considered in the context of high level image understanding tasks such as image annotation (Jie et al., 2009) and action recognition (Yang, Wang and Mori, 2010; Yao and Fei-Fei, 2010; Schindler et al., 2008). Poses of people can provide essential cues for their actions, and recognizing actions of people is in turn helpful for understanding the entire visual scene. At the same time, high level knowledge about the scene provides useful context for solving other recognition tasks, such as object detection, and might in fact be useful for narrowing down the space of possible poses of people. An example of such an integrated approach is the work of Jie et al. (2009), where face detection, people detection, pose estimation, and text parsing are brought together in order to automatically learn appearances of specific people and the correspondence between verbs and poses from a large collection of images and text annotations. Learning appearance of people from noisy text annotations is difficult because images often contain multiple people, and it is unclear which tag in the annotation corresponds to which person in the image. The authors show that the estimated poses can be often helpful in resolving such ambiguities because names of people are often used in conjunction with verbs describing their actions.

A somewhat different approach to integration of pose estimation into a larger image understanding system was proposed by Yao and Fei-Fei (2010). The focus of this work is on modeling interaction between activities, objects, and poses of people. In sports scenes, considered in this work, objects of interest (such as a tennis racket or volleyball) are often difficult to detect due to their small size, a cluttered background and out-of-plane rotations. However, poses of people interacting with objects provide strong bias for the object location, which results in better object detection. At the same time, object detection can be used to constrain the space of body poses and can be used as additional evidence for body limbs interacting with the object. Finally both objects and poses have strong connection to the activities of people present in the images. As demonstrated by Yao and Fei-Fei (2010), one can approach action recognition, object detection, and pose estimation in a unified manner and exploit statistical dependencies between them in order to improve recognition performance for each task.

2.4 3D Human Pose Estimation

Chapter 5 of this thesis deals with the problem of 3D human pose estimation in image sequences taken in uncontrolled street environments with an uncalibrated, monocular, and potentially moving camera. Images obtained under such conditions often feature cluttered, dynamically-changing backgrounds and contain multiple people who frequently occlude each other. This is one of the most difficult settings considered in the 3D human pose estimation literature. In contrast, due to the inherent difficulties of reliable *2D to 3D lifting* and *data association*, this problem has often been considered in controlled laboratory conditions (Balan, Sigal, Black, Davis and Haussecker, 2007; Deutscher and Reid, 2005; Gall et al., 2010; Sigal and Black, 2006*b*; Vondrak et al., 2008), with solutions frequently relying on background subtraction and simple image evidence, such as silhouettes or edge-maps. In order to constrain the search in high-dimensional pose space these approaches often use multiple calibrated cameras (Gall et al., 2010), complex dynamic motion priors (Vondrak et al., 2008), or detailed body models (Balan, Sigal, Black, Davis and Haussecker, 2007). The combination of these components allows one to achieve impressive results (Gall et al., 2010; Hasler et al., 2009), similar in performance to commercial marker-based motion capture systems.

However, realistic street scenes, such as those considered in this thesis, do not satisfy many of the assumptions made by these systems. For such scenes, multiple synchronized video streams are difficult to obtain; the appearance of people is significantly more complex; and robust extraction of evidence is challenged by frequent full and partial occlusions, clutter, and camera motion. In order to address these challenges, a number of methods leverage recent advances in people detection and rely on them for prefiltering, initialization and as evidence for pose estimation (Fossati et al., 2007; Gammeter et al., 2008; Andriluka et al., 2010).

Due to the complexity of the problem, scenarios considered in realistic uncontrolled environments are typically limited to walking people only. Fossati et al. (2007) build on the bank of silhouette-based people detectors designed to detect people appearing in the characteristic phase of the walking cycle with legs far apart. This bank of detectors is capable of detecting people seen from various viewpoints and also provides viewpoint estimates. The initial detections are projected into the world coordinates where the position of the person is tracked by combining single-frame detections using constraints given by motion speed, viewpoint and walking cycle. Such tracks, in turn, provide strong constraints on the possible 3D body configurations. Initialization for poses of people can be obtained using snippets of 3D pose sequences from the training data. Homography estimates are used to reconstruct the scene background and rely on pose initializations to obtain a rough appearance model of the person. Jointly these components allow one to define a generative model for the image sequence, which is used to refine the 3D pose estimates using a stochastic local search.

Gammeter et al. (2008) have proposed a system for 3D pose estimation in crowded street scenes. They rely on a stereo camera setup and structure-from-motion in order to estimate the scene geometry and camera trajectory. These estimates allow one to back-project detections of people into world coordinates, where they are associated with tracks corresponding to individual people and are integrated over time with a Kalman filter. For each of these tracks, the segmentation of people is computed by combining top-down cues obtained from the person detector and bottom-up cues, such as disparity and image edges. Subsequently, 3D poses of people are estimated using a particle filtering framework in which each particle corresponds to low-dimensional representation of the human pose obtained with locally linear embedding (LLE). The particles are propagated through time using an auto-regressive dynamical model for the human walking motion, and are resampled after each iteration according to their compatibility with inferred segmentation.

In our work presented in (Andriluka et al., 2010) and described in detail in Chapter 5, we propose an approach to 3D pose estimation which relies on our detection and 2D pose estimation approach (Andriluka et al., 2009). At the same time we also build on dynamic motion priors (Urtasun et al., 2006; Lawrence and Moore, 2007) and on our approach to articulated tracking-by-detection that we introduced in (Andriluka et al., 2008). Although estimation of 3D poses from 2D body part positions was previously proposed in (Sigal and Black, 2006*c*), this approach was evaluated only in laboratory conditions for a single subject, and it remains unclear how well it generalizes to more complex settings with multiple people as considered here. Compared to the approach of Gammeter et al. (2008), we are able to estimate poses in monocular images without relying on estimates of world geometry, camera calibration and stereo cues. At the same time we do not rely on the ability to reconstruct the scene background and detect people in characteristic poses as is required in (Fossati et al., 2007).

Summarizing the results obtained by Fossati et al. (2007); Gammeter et al. (2008), and Andriluka et al. (2010), it appears that once the detections of people in subsequent frames are associated with each other and the viewpoints of people are known, the activity-specific dynamic models provide sufficient temporal constraints to infer 3D poses. This fits well with the observation previously made in (Forsyth et al., 2006): “There is a body of evidence, strongly suggestive though not absolutely conclusive, that 3D reconstruction from 2D frames has few ambiguities that persist over any but the shortest timescales ... it suggests that the proper approach to tracking in 3D is to track in 2D, and then report an estimate of 3D by matching the 2D track to 3D body configuration snippets". Furthermore, it appears that in most cases the parameters, such as the image position of the person, overall scale, and relative orientation to the camera, can be estimated reliably prior to the actual estimation of the 3D pose. These parameters, in turn, significantly reduce the search space, and also allow one to obtain much better initializations during 3D pose estimation. In (Andriluka et al., 2010) we estimate these parameters from monocular image sequences by combining the evidence obtained from multiple subsequent frames. Similar results were obtained in (Fossati et al., 2007), where the walking phase and viewpoint were estimated using temporal integration of a single-frame hypothesis.

Recently, Ferrari et al. (2008) have used a similar coarse-to-fine strategy, relying on people detection and tracking in order to determine the image position and scale of people, and then used these estimates to restrict the search during 2D pose estimation. While being effective in practice, such approaches have a major drawback in that the consistency of people poses cannot be used as an additional constraint in order to assist with data association. While data association is easy in scenes with only few people, it becomes non-trivial in crowded street scenes with multiple people who are frequently occluding each other. The most effective approach would be to delay the resolution of ambiguities in data association until the pose estimation stage, where they can be resolved using the combination of temporal, appearance and pose consistency constraints. In (Andriluka et al., 2008), we have implemented this idea within a tracking-by-detection framework, where we have combined single-frame people detections into *tracklets* that can be seen as *data association hypotheses*. This, on one hand, allows pose estimation using combined evidence from multiple frames and, on the other hand, allow us to delay the final data association step until the later stages of the inference.

Apart from the approaches discussed so far, there is also an extensive body of literature on predicting 3D poses directly from image features using regression (Agarwal and Triggs, 2006; Ionescu et al., 2009; Urtasun and Darrell, 2008), classification (Rogez et al., 2008), or a search over a database of exemplars (Mori and Malik, 2006; Shakhnarovich et al., 2003). The general theme underlying these approaches is to formulate human pose estimation as an end-to-end prediction problem that can be directly solved using generic machine learning techniques. This is in contrast to methods previously discussed in this section, which typically consist of multiple stages and integrate the output of various independently developed algorithms. Hypothetically, end-to-end approaches have the advantage of training all components of the model jointly, which should allow them to automatically determine the relative importance of various image cues and dependencies between outputs. In practice, they were shown to perform well in cases when poses and the appearance of people in training and test images are sufficiently similar and require large databases of training examples to achieve good performance. The application of these methods has been limited to the prediction of poses of individual people where the data association problem has already been solved.

2.5 Other Related Work

In addition to the work on vision-based people detection discussed above, there exists a large body of literature on people detection using other types of sensors. Most of this work has been done in the robotics community and is often oriented towards tasks such as collision avoidance, person following, and search and rescue. Since many mobile robotic systems are equipped with laser range scanners, a significant effort has been dedicated to using them for people detection and tracking (Arras et al., 2007, 2008; Carballo et al., 2009; Fod et al., 2002; Schulz et al., 2003). The complementarity between visual- and laser-based detectors has been explored in (Gate et al., 2009), where a laser range scanner is used both to extract regions of interest in camera images and to improve the confidence of an AdaBoost-based visual detector. Thermal images have also been extensively used for people detection using either specifically designed methods (Davis and Sharma, 2004; Pham et al., 2007) or by directly applying methods originally designed for detection of people in daylight images (Suard et al., 2006). People detectors used in robotics are frequently built using a combination of several sensing modalities (Doherty and Rudol, 2007; Kleiner and Kuemmerle, 2007; Meyer et al., 2010), and rely on complementary measurements from multiple sensors to improve the robustness, but also for prefiltering in order to speed up computations.

In Chapter 6 of this thesis we develop an approach to visual people detection from unmanned aerial vehicles. Our approach is related to the work of Doherty and Rudol (2007), who address a similar problem, but in contrast to our approach rely on a thermal camera to prefilter promising image locations. While in (Doherty and Rudol, 2007) people lying on the ground are assumed to be in upright poses, our approach addresses the significantly more complex problem of detecting arbitrarily articulated people. Note that the results of our work can still be used in combination with thermal camera images which, similar to (Doherty and Rudol, 2007), can be used to restrict the search to image locations likely to contain people or to prune false positives when they contain no thermal evidence. While combining multiple sensors for people detection is clearly beneficial in many scenarios it comes at the cost – in particular for unmanned aerial vehicles – of an increased payload for the additional sensors. Our approach to people detection from UAVs, therefore, aims to evaluate and push the state-of-the-art in visual people detection in order to minimize sensor requirements for this task.

3

Combining People Detection, Tracking and Pose Estimation

Tracking people in images of uncontrolled street environments, such as those shown in Fig. 3.1, is a challenging task. Such images typically contain multiple people moving in front of complex dynamic backgrounds where the people often become fully or partially occluded. Occlusions lead to lost tracks, which means that in such scenes automatic re-initialization is an essential part of any successful tracking algorithm. Due to a large number of possible poses, background clutter and large appearance variability people detection and pose estimation in such scenes is challenging as well.

In this chapter we present a *tracking-by-detection* approach that simultaneously address people detection, tracking and pose estimation. Using the output of our people detector as evidence for tracking allows us to cope with occlusions and avoid manual initialization. In addition, pose estimates from our part-based people detector allow us to employ more powerful dynamical model during tracking, which results in improved data association between frames. Finally, we rely on the obtained pose estimates in order to construct appearance models of individual persons on the fly, and use these models in order to locate individuals after long-term occlusions.

While core ideas of this thesis are recognizable in the approach presented in this chapter, it still contains significant limitations. In particular, we consciously omit the problems associated with ambiguities of 2D to 3D lifting by assuming that people appear in the images in a side view. The appearance model used by the people detector is also both viewpoint- and activity-specific, which is also true for the dynamical model for body articulations used for tracking. In the subsequent parts of this thesis we work towards relaxing these assumptions by introducing more generic part-based object detection approach in Chapter 4 and estimating poses of people using a 3D body model in Chapter 5.

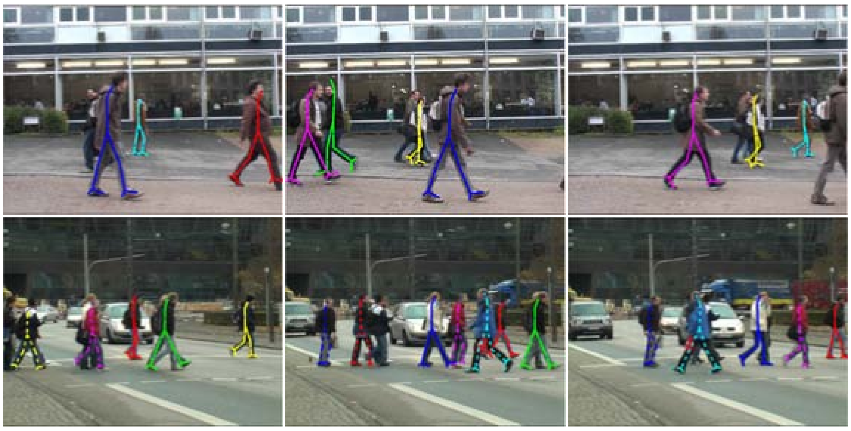


Figure 3.1: Examples of detection and tracking of *specific* persons in image sequences of crowded street scenes.

3.1 Introduction

Probably the most fundamental difficulty in detection and tracking of multiple people in cluttered scenes is that many people will be partially and also fully occluded for longer periods of time. Consequently, both the detection of people in individual frames as well as the association between people detections in different frames are highly challenging and ambiguous. To address this, we exploit temporal coherency, extract people-tracklets from a small number of consecutive frames and from those tracklets build appearance models of the individual people. As any single person might be detectable only for a small number of frames the extraction of people-tracklets has to be highly robust. At the same time the extracted model of the individual has to be discriminative enough in order to enable tracking and data-association across long periods of partial and full occlusions.

To achieve reliable extraction of people-tracklets as well as data-association across long periods of occlusion, our approach combines recent advances in people detection with the power of dynamical models for tracking. Rather than to simply determine the position and scale of a person as is common for state-of-the-art people detectors (Dalal and Triggs, 2005; Leibe et al., 2005), we also extract the position and articulation of the limbs. This allows us to use a more powerful dynamical model that extends people detection to the problem of reliably extracting people-tracklets – people detections consistent over a small number of frames. In particular, we use a hierarchical Gaussian process latent variable model (hGPLVM) (Lawrence and Moore, 2007) to model the dynamics of the individual limbs. As we will show in the experiments this enables us to detect people more reliably than it would be possible from single frames alone. We combine this with a hidden Markov model (HMM) that allows to extend the people-tracklets, which cover only a small number of frames at a time, to possibly longer people-tracks. These people-tracks identify individuals over longer sequences of consecutive frames when that is appropriate, such as between major occlusion events. Tracking people over even longer

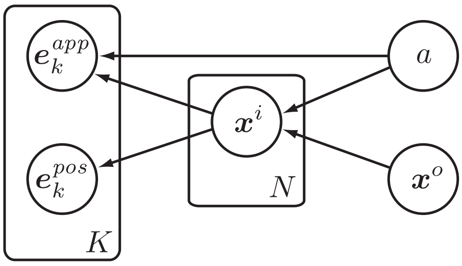


Figure 3.2: Graphical model structure describing the relation between articulation, parts, and features.

periods of time is then achieved by associating people-tracks across potentially long periods of occlusion using both the dynamical model and an extracted appearance model, which allows identifying specific individuals throughout the sequence.

In Sec. 3.2 of this chapter we introduce a novel people detection approach based on the Pictorial Structures (PS) framework (Felzenszwalb and Huttenlocher, 2005), in which we represent object as flexible configuration of rigid body parts and model pat appearance using Implicit Shape Model (ISM) (Leibe et al., 2005). We demonstrate that this novel detector outperforms two state-of-the-art detectors on a challenging dataset of street images.

Our detector has two important properties that make it particularly suited for the detection and tracking of multiple people in crowded scenes: First, it allows to detect people in the presence of significant partial occlusions, and second, the output of the detector includes the positions of individual limbs, which are used as input for the dynamical model at the next stage. We integrate the people detection model with a dynamical limb-model to enable reliable detection of people-tracklets over small number of consecutive frames, which as we show further improves the detection performance. The extracted people-tracklets are then used to generate a detailed appearance model of each person on the fly.

Finally we link short people-tracklets to longer tracks of the various individuals in the scene. In this, we take advantage of the tracklet detector, which allows us to perform inference in a simple discrete state space of tracklet hypothesis. This is in contrast to typical tracking approaches that need to perform stochastic search in high-dimensional, continuous spaces (Deutscher and Reid, 2005), which is well known to suffer from many problems (Demirdjian et al., 2005; Sminchisescu et al., 2007). Note that while a precise recovery of the full articulation is not the main focus of this work, we can still quite accurately recover the articulation even in complex scenes. We link the people-tracks in scenes with multiple people and over periods of long occlusions using the learned appearance model, which allows us to identify and tracked individuals even through complex occlusions without requiring any manual initialization or manual intervention at *any* stage of the process.

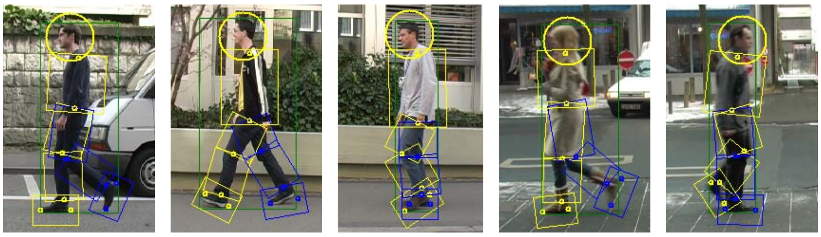


Figure 3.3: Examples of images from our training set.

3.2 Pedestrian Detector

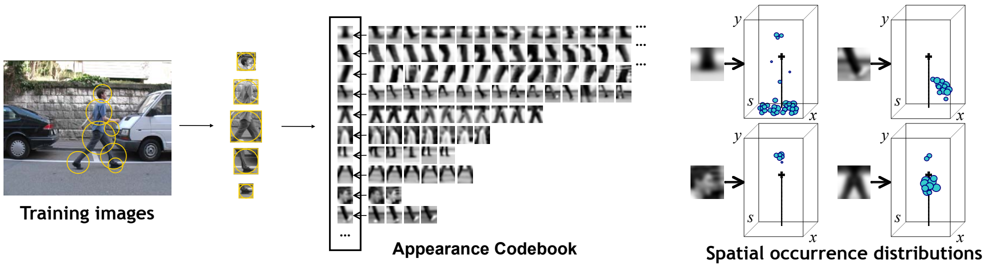
Before introducing temporal constraints into the detection process, we first propose a novel part-based object detection model that is capable of detecting pedestrians in single images of real-world scenes. The model is inspired by the pictorial structures model proposed by Fischler and Elschlager (1973) and Felzenszwalb and Huttenlocher (2005), but uses more powerful part representations and detection, and as we will show outperforms recent pedestrian detectors (Dalal and Triggs, 2005; Seemann and Schiele, 2006).

3.2.1 Part-based model for pedestrian detection

Following the general pictorial structures idea, an object is represented as a joint configuration of its parts. In such a model the problem of locating an object from a specific class in a test image is formulated as search for the modes of the posterior probability distribution $p(L|E)$ of the object configuration L given the image evidence E and (implicit) class-dependent model parameters θ.

In our model, the configuration is described as $L = \{\mathbf{x}^o, \mathbf{x}^1, \ldots, \mathbf{x}^N\}$, where $\mathbf{x}^o$ is the position of the object center and its scale, and $\mathbf{x}^i$ is the position and scale of part i. The image evidence, which here is defined as a set of local features observed in the test image, will be denoted as $E = \{\mathbf{e}_k^{app}, \mathbf{e}_k^{pos} | k = 1, \ldots, K\}$, where $\mathbf{e}_k^{app}$ is an appearance descriptor, and $\mathbf{e}_k^{pos}$ is the position and scale of the local image feature with index k. We will denote the combination of position, scale, and appearance of a local feature as $\mathbf{e}_k = (\mathbf{e}_k^{app}, \mathbf{e}_k^{pos})$.

An important component of the pictorial structures model is a-priori distribution of the possible object configurations, which must be expressive enough to capture all important dependencies between parts. Part positions are mutually dependent in general, which can make inference difficult. However, for particular object categories, such as walking people, we can introduce auxiliary state variables that represent the *articulation state* or an *aspect* of the object, such as different phases in the walking cycle of a person (Lan and Huttenlocher, 2005), and make the parts conditionally independent. If the articulation state is observed, the model becomes a star model (or tree model in general) and efficient algorithms based on dynamic programming can be used for inference (Felzenszwalb and Huttenlocher, 2005). If we are not



(a) (b)

Figure 3.4: The main components of the Implicit Shape Model (Leibe et al., 2008): appearance codebook (a) and learned spatial occurrence distribution corresponding to each the codebook entries (b).

interested in knowing the articulation state, but only the object and limb positions, then articulation state a can be marginalized out:

$$p(L|E) = \sum_a p(L|a,E)p(a). \tag{3.1}$$

From decomposing $p(L|a,E) \propto p(E|L,a)p(L|a)$, assuming that the configuration likelihood can be approximated with product of individual part likelihoods (Felzenszwalb and Huttenlocher, 2005) $p(E|L,a) \approx \prod_i p(E|\mathbf{x}^i,a)$, and assuming uniform $p(\mathbf{x}^i|a)$, it follows that

$$p(L|a,E) \approx p(\mathbf{x}^o) \prod_i p(\mathbf{x}^i|a,E)p(\mathbf{x}^i|\mathbf{x}^o,a). \tag{3.2}$$

If we assume that a particular image feature $\mathbf{e}_k$ belongs to part i of an object instance in the image with probability α, then it holds that

$$p(\mathbf{x}^i|a,E) = c_0 + c_1 \sum_{\mathbf{e}_k} p(\mathbf{x}^i|a,\mathbf{e}_k) + O(\alpha^2), \tag{3.3}$$

where c_0 and c_1 depend only on the image features E (Williams and Allan, 2006). If α is sufficiently small, which is true for street scenes in which a particular person usually represents only a small portion of the image, we obtain

$$p(L|a,E) \approx \prod_i p(\mathbf{x}^i|\mathbf{x}^o,a) \left[\beta + \sum_{\mathbf{e}_k} p(\mathbf{x}^i|a,\mathbf{e}_k)\right], \tag{3.4}$$

where β can be seen as a regularizer for the evidence obtained from the individual image features, and we have additionally assumed uniform $p(\mathbf{x}^o)$.

In order to obtain the probability of object part given local image feature we rely on the Implicit Shape Model (Leibe et al., 2008), which consists of the object-specific *codebook*, and spatial *occurrence distribution* of the object center with respect to each codebook entry (see Fig. 3.4 for illustration).

The part posterior with respect to a single image feature is computed by marginalization over the codebook entries:

$$p(\mathbf{x}^i|a, \mathbf{e}_k) = \sum_{\mathbf{c}_j \in \mathcal{C}} p(\mathbf{x}^i|a, \mathbf{c}_j, \mathbf{e}_k^{pos}) p(\mathbf{c}_j|\mathbf{e}_k^{app}), \tag{3.5}$$

where $\mathcal{C} = \{\mathbf{c}_j | j = 1, \ldots, J\}$ denotes the codebook, $p(\mathbf{c}_j|\mathbf{e}_k^{app})$ is a discrete distribution over codebooks based on a Gaussian similarity measure, and $p(\mathbf{x}^i|a, \mathbf{c}_j, \mathbf{e}_k^{pos})$ is occurrence distribution of part i with respect to codebook entry $\mathbf{c}_j$ translated according to the location of image feature $\mathbf{e}_k^{pos}$. The structure of the dependencies between the variables in the model is shown in Fig. 3.2.

Fig. 3.6 shows an example of the posterior distribution of the feet, lower leg and head parts given positions of several local features, selected from the set of all local features shown on Fig. 3.5. Note, how the posterior distribution reflects the uncertainty in the interpretation of local appearance information, e.g.multiple modes in the posterior of the feature detected on the foot of the person appear since this feature could occur on both left and right feet. Also note, that in the case of pedestrians local features even far from the body part position, turn out to be useful for its localization, e.g. as in the posterior of head given features on the legs shown on Fig. 3.6(d). Using these long-rage spatial relations makes our model more robust to partial occlusions and also allows to localize small body parts, such as feet.

3.2.2 Model training

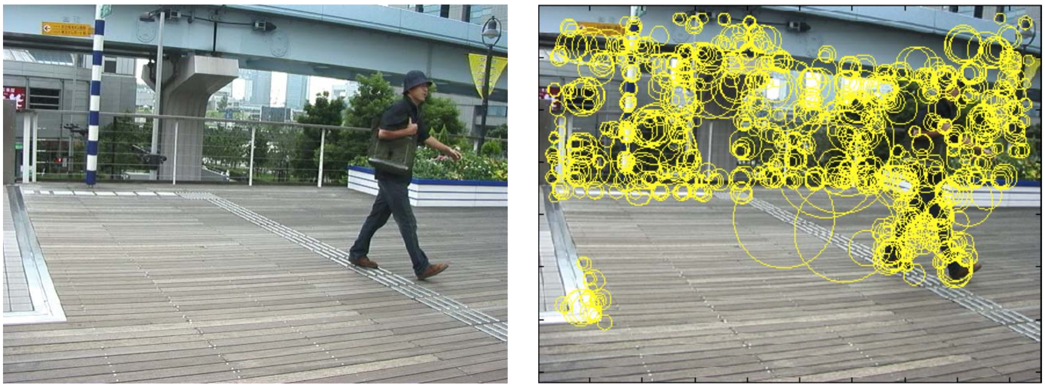
In all presented experiments we use shape context feature descriptors (Belongie et al., 2000) and the Hessian-Laplace interest point operator (Mikolajczyk and Schmid, 2004) as detector. The object-specific codebook is constructed by clustering local features extracted from the set of training images.

For each codebook cluster $\mathbf{c}_j$ we compute its occurrence distribution, which corresponds to a set of jointly observed relative position and scale of the cluster with respect to the part centers computed for each occurrence of the cluster in the training set.

In order to compute the occurrence distributions we annotated each person along with its parts in all training images. Fig. 3.3 shows several images from our training set. From the same annotation we also learn a Gaussian distribution of the position of each part relative to the object center, $p(\mathbf{x}^i|\mathbf{x}^o, a)$. While the position components are learned, the scale component is taken to be relatively broad and chosen empirically. This appears to be sufficient for pedestrians, particularly since we are not differentiating between left and right legs at the detection stage.

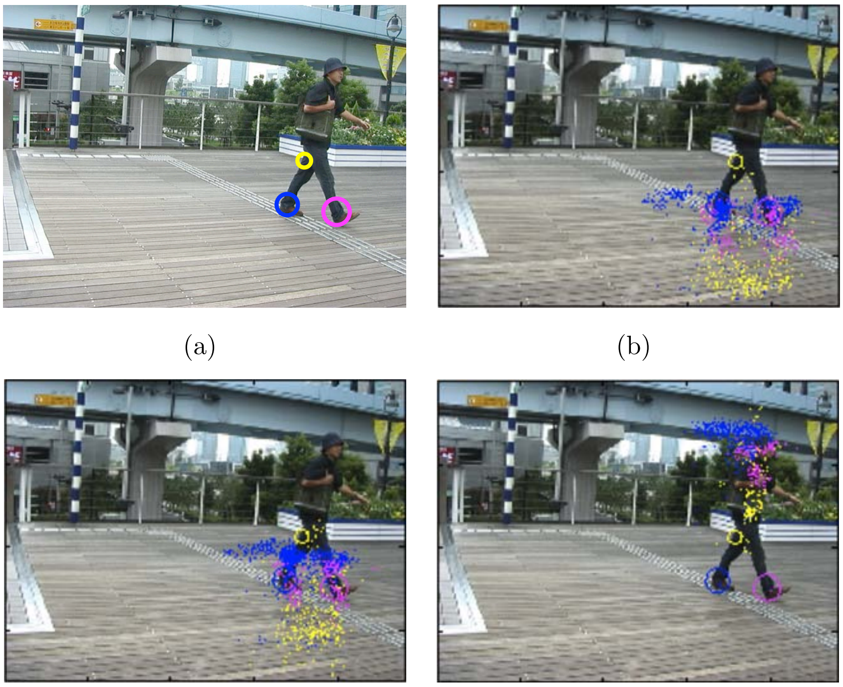
3.2.3 Inference and Results

In the first step of inference we accumulate $p(\mathbf{x}^i|a, \mathbf{e}_k)$ in a 3 dimensional array of discretized image positions and scales. After that Eq. (3.4) can be maximized efficiently using the generalized distance transform (Felzenszwalb and Huttenlocher,



(a) (b)

Figure 3.5: (a) Test image, and (b) scale-invariant Hessian-Laplace interest points detected on the image.



(a) (b)

(c) (d)

Figure 3.6: (a) Several detected interest points, and conditional posterior distributions $p(\mathbf{x}^i|a, \mathbf{e}_k)$ of the foot (b), lower leg (c), and head (d) body parts given position $\mathbf{e}_k^{pos}$ and local image feature $\mathbf{e}_k^{app}$ corresponding to each of the interest points. Distributions corresponding to different interest points are visualized using different colors (see text for description).

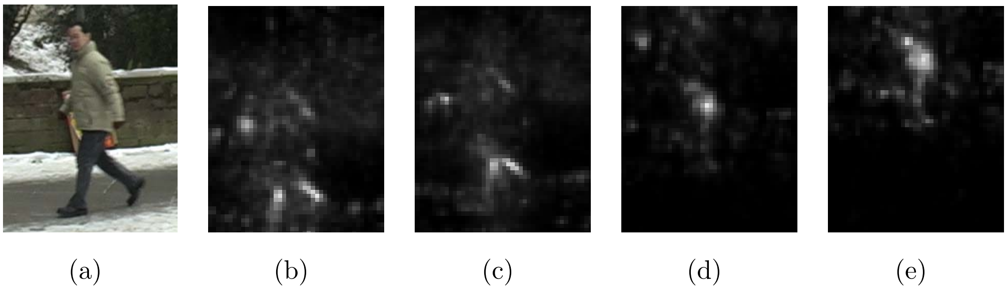


Figure 3.7: Person hypothesis with corresponding probabilities of feet (b), lower legs (c), torso (d), and head (e).

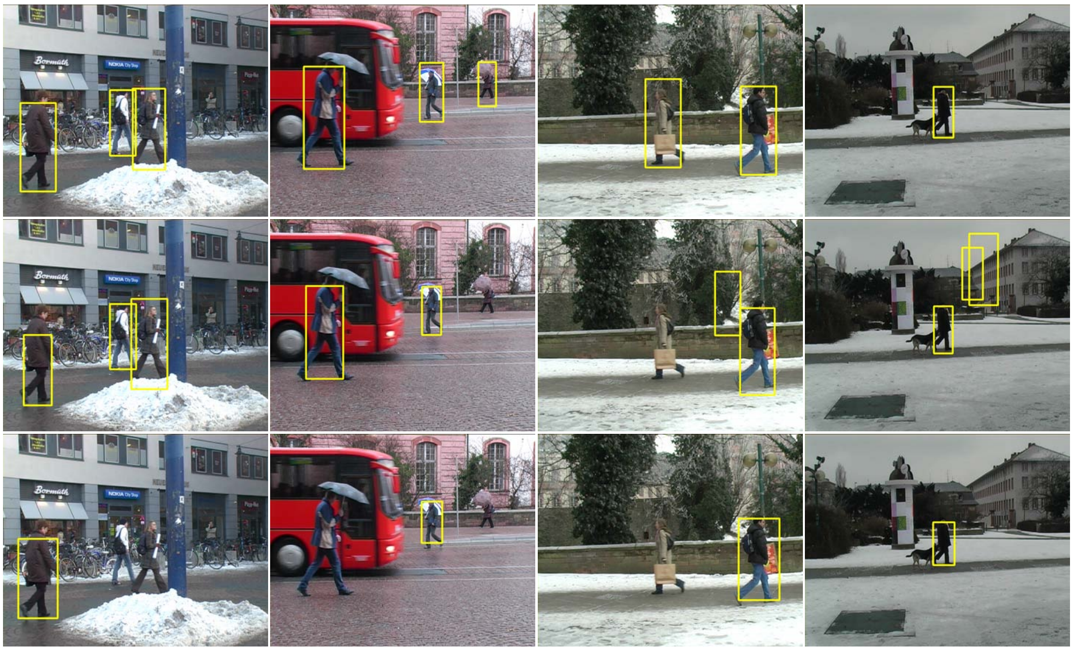


Figure 3.8: Example detections at equal error rate of our detector (top), 4D-ISM (middle) and HOG (bottom) on the "TUD-Pedestrians" dataset.

2005). This is possible since part dependencies have a tree (star-) structure, appearance components are computed separately for each part, and $p(\mathbf{x}^i|\mathbf{x}^o, a)$ is Gaussian. Fig. 3.7 visualizes $\sum_{\mathbf{e}_k} p(\mathbf{x}^i|a, \mathbf{e}_k)$ in the region of a person hypothesis for different body parts.

In the following we evaluate the novel detector on a challenging dataset of 250 images of street scenes containing 311 side-view pedestrians with significant variation in clothing and articulation, which we denote as "TUD-Pedestrians" [1]. Fig. 3.10(a) shows the comparison of our detector with two state of the art detectors. Using the same training set as Seemann and Schiele (2006) our detector outperforms the 4D-ISM approach (Seemann and Schiele, 2006) as well as the HOG-detector (Dalal and Triggs, 2005). Increasing the size of the training set further improves performance

---

[1] Available at `www.mis.informatik.tu-darmstadt.de`.

significantly.

Fig. 3.8 shows sample detections of the 3 methods on test images. The 4D-ISM detector is specifically designed to detect people in cluttered scenes with partial occlusions. Its drawback is that it tends to produce hypotheses even when little image evidence is available (image 3 and 4), which results in increased number of false positives. The HOG detector seems to have difficulties with the high variety in articulations and appearance present in out dataset. However, we should note that it is a multi-view detector designed to solve a more general problem than we consider here.

In addition to the high precision of our detector, we observe an improved scale estimation of the hypotheses as can be seen on the leftmost image of Fig. 3.8. We attribute this to the fact that the influence of imprecise scale-estimates of the local feature detector are reduced using local object parts.

3.3 Detection of Tracks in Image Sequences

The person detector described in Sec. 3.2 provides hypotheses for position, scale, and body articulation in single frames based on the detection of individual body parts or limbs. To further improve the detection performance in image sequences several authors have proposed to incorporate temporal consistency among subsequent frames, typically using a simple motion model based on position and velocity of the person. In contrast, we propose to use a more expressive kinematic limb model thereby leveraging the articulated tracking literature, e.g., (Deutscher and Reid, 2005; Demirdjian et al., 2005; Sigal and Black, 2006*b*; Urtasun et al., 2006; Ramanan et al., 2007; Sminchisescu et al., 2007). Clearly, the expressiveness and the robustness of the kinematic limb model are crucial as it has to be powerful enough to reduce the number of false positives significantly, and at the same time robust enough to enable people detection in crowded scenes.

Given the image evidence $E = [E_1, E_2, \ldots, E_m]^T$ in a sequence of m subsequent frames, we would like to recover the positions $\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \mathbf{x}_2^{o*}, \ldots, \mathbf{x}_m^{o*}]^T$ of the person as well as the configurations of the limbs in each frame $\mathbf{Y}^* = [\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_m^*]^T$ with $\mathbf{y}_j^*$ denoting the recovered limb orientations in the j-th frame. Assuming independence of the detections in each frame, the posterior factorizes as:

$$\begin{aligned} p(\mathbf{Y}^*, \mathbf{X}^{o*}|E) &\propto p(\mathbf{Y}^*)p(\mathbf{X}^{o*})p(E|\mathbf{Y}^*, \mathbf{X}^{o*}) \\ &\propto p(\mathbf{Y}^*)p(\mathbf{X}^{o*})\prod_{j=1}^{m} p(E_j|\mathbf{y}_j^*, \mathbf{x}_j^{o*}). \end{aligned} \tag{3.6}$$

$p(E_j|\mathbf{y}_j^*, \mathbf{x}_j^{o*})$ is the likelihood of the image evidence E_j, and is given by the detection model described in the previous section. $p(\mathbf{X}^{o*})$ corresponds to a prior of human body speed, which we model as a broad Gaussian. Probably the most interesting term is $p(\mathbf{Y}^*)$, which denotes the prior over the kinematic limb-motions, and in general is difficult to estimate reliably due to the high dimensionality of the pose space. Instead of modelling the pose dynamics directly in an high-dimensional space, several authors (Urtasun et al., 2006; Wang et al., 2005; Sminchisescu et al., 2007) have argued and shown that a low-dimensional representation is sufficient to approximate the dynamics. In the following we use a Gaussian process latent variable model (GPLVM) to obtain such a low-dimensional representation and discuss how it can be used to obtain reliable people detections in image sequences.

3.3.1 Gaussian process latent variable model

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]^T$ be a sequence of D-dimensional observations (here describing the relative joint angles of body limbs). GPLVMs model the D-dimensional observation space as the output of D Gaussian processes with an input space of dimensionality q, where $q < D$. Each observation $\mathbf{y}_i$ is associated with a q-dimensional latent point $\mathbf{z}_i$. The likelihood of the observation sequence $\mathbf{Y}$ given the latent sequence $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m]^T$ and model parameters θ is given by Lawrence (2005):

$$p(\mathbf{Y}|\mathbf{Z}, \theta) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{Y}_{:,i}|0, \mathbf{K}_\mathbf{z}), \tag{3.7}$$

where $\mathbf{Y}_{:,i}$ is the vector of values of feature i across all observations, and $\mathbf{K}_\mathbf{z}$ is the covariance matrix with elements given by a covariance function $k(\mathbf{z}_i, \mathbf{z}_j)$. In this work we use a squared exponential covariance function augmented by Gaussian white noise. For a given $\mathbf{Y}$ we can find the positions of the latent points $\mathbf{Z}$ along with the model parameters θ by maximizing their likelihood from Eq. (3.7).

In addition to the low-dimensional latent representation of the data, GPLVMs provide a probabilistic mapping from the latent space to the observation space. One possibility to define a dynamic model in the latent space is to place a suitable prior on the elements of $\mathbf{Z}$. Such a prior can be given by a Gaussian process with time as input variable (Lawrence and Moore, 2007). Given the sequence of points in time, $\mathbf{T} = [t_1, t_2, \ldots, t_m]^T$ at which the observations $\mathbf{Y}$ were made, the prior over $\mathbf{Z}$ is given by

$$p(\mathbf{Z}|\mathbf{T}) = \prod_{i=1}^{q} \mathcal{N}(\mathbf{Z}_{:,i}|0, \mathbf{K}_\mathbf{T}) \tag{3.8}$$

where $\mathbf{K}_\mathbf{T}$ is the covariance matrix of the time points. The covariance function in the time space can again be taken as squared exponential, which ensures smoothness of the trajectories.

We now combine this prior with the likelihood from Eq. (3.7), and maximize w.r.t. $\mathbf{Z}$ and θ. Fig. 3.9 shows the 2 dimensional latent space obtained by applying this model to 11 walking sequences of different subjects, each containing one complete walking cycle. Walking cycles in each sequence are manually aligned so that we can interpret the frame number in each sequence as phase of the walking cycle. This hierarchical approach to GPLVM dynamics has several advantages over the auto-regressive prior proposed in (Wang et al., 2005). In particular, it allows us to evaluate the likelihood of a sequence of poses, even if the poses occurred at unequally spaced time intervals. This arises, e.g., when the subject was occluded or not detected for several frames. Additionally, for a given pose the model allows us to hypothesize both successive and previous poses, which we use to produce good initial hypotheses for the whole image sequence from a few good detections.

3.3.2 Reconstruction of poses in short sequence

Given limb likelihoods and the hGPLVM prior, we can maximize Eq. (3.6) to find the best pose sequence. This is equivalent to jointly solving the inverse kinematics in each frame of the sequence under soft constraints given by limb likelihoods and is similar to (Grochow et al., 2004), except that in our case hints about limb positions are provided by a person detector instead of being manually given by the user. If we denote the training observations, their latent representation and model parameters by $\mathcal{M} = [\mathbf{Y}, \mathbf{T}, \mathbf{Z}, \boldsymbol{\theta}]$, the probability of the unknown pose sequence $\mathbf{Y}^*$, its latent representation $\mathbf{Z}^*$, and the person positions $\mathbf{X}^{o*}$ is given by

$$\begin{aligned} & p(\mathbf{Y}^*, \mathbf{X}^{o*}, \mathbf{Z}^* | \mathcal{M}, E, \mathbf{T}^*) \propto \\ & p(E|\mathbf{Y}^*, \mathbf{X}^{o*}) p(\mathbf{Y}^*|\mathbf{Z}^*, \mathcal{M}) p(\mathbf{Z}^*|\mathbf{T}^*, \mathcal{M}) p(\mathbf{X}^{o*}). \end{aligned} \tag{3.9}$$

The first term is the detection likelihood from single-frame detections (see Eq. (3.6)). The second term is given by

$$p(\mathbf{Y}^*|\mathbf{Z}^*, \mathcal{M}) = \prod_{i=1}^{D} p(\mathbf{Y}^*_{:,i}|\mathbf{Z}^*, \mathbf{Y}_{:,i}, \mathbf{Z}), \tag{3.10}$$

where $p(\mathbf{Y}^*_{:,i}|\mathbf{Z}^*, \mathbf{Y}_{:,i}, \mathbf{Z})$ is a Gaussian process prediction of the pose sequence given a sequence of latent positions. The third term is given by the dynamics prior on the latent space:

$$p(\mathbf{Z}^*|\mathbf{T}^*, \mathcal{M}) = \prod_{i=1}^{q} p(\mathbf{Z}^*_{:,i}|\mathbf{T}^*, \mathbf{Z}_{:,i}, \mathbf{T}). \tag{3.11}$$

In our formulation, detecting people in a series of m frames therefore corresponds to finding pose sequences $\mathbf{Y}^*$ and people positions $\mathbf{X}^{o*}$ that maximize Eq. (3.9). We use the following strategy to efficiently obtain such maxima: Each hypothesis obtained from the people detector from Sec. 3.2 contains an estimate of the person's position $\mathbf{x}^o$, limbs' positions $\mathbf{x}^i$ and articulation a. From these parameters we can directly estimate both the limb-orientations $\mathbf{y}$, and position in the walking cycle t. Using those parameters and propagating them to neighboring frames using the kinematic limb model has proven to yield good initializations for optimization. In the experiments described below we use a sufficient but small number of detections in each frame to obtain initialization values for $\mathbf{Y}^*$, $\mathbf{T}^*$, and $\mathbf{X}^{o*}$, and then use a conjugate gradient method to find local maxima of Eq. (3.9). The gradients of the second, third, and fourth term in Eq. (3.9) can be computed analytically,

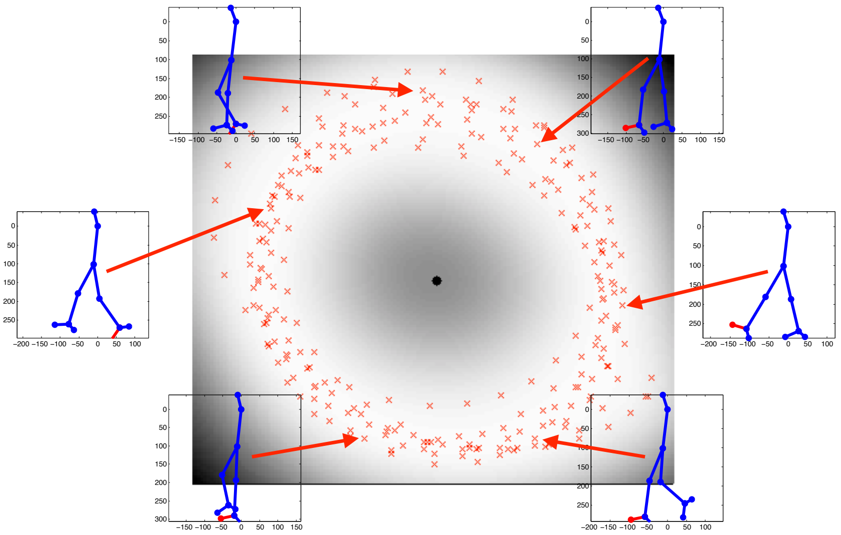
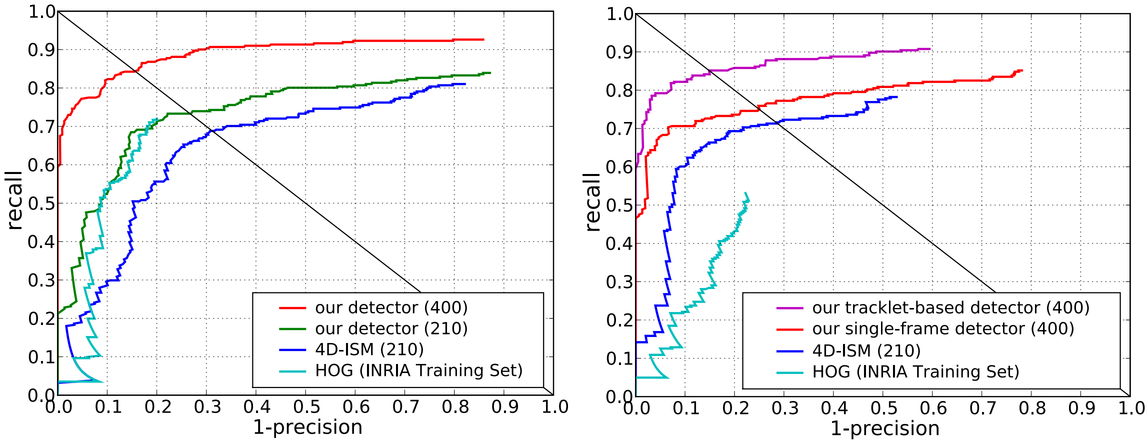


Figure 3.9: Representation of articulations in the latent space.



(a) (b)

Figure 3.10: Comparison of our pedestrian detector with 4D-ISM (Seemann and Schiele, 2006) and HOG (Dalal and Triggs, 2005) on (a) the "TUD-Pedestrians" and (b) "TUD-Campus" datasets. Numbers in parenthesis indicate number of training images.

while we use a finite difference approximation for gradient of $p(E|\mathbf{Y}^*, \mathbf{X}^{o*})$. As the experiments show, this enables us to efficiently obtain people detections in image sequences.

Quantitatively, Fig. 3.10(b) shows how the extracted tracklets lead to increased precision and recall compared to the detector alone (note that the recall does not reach 1 in either case due to the use of non-maxima suppression.)

3.3.3 Optimal track selection based on overlapping tracklets

The optimization procedure just described is suited to reliably detect people in short frame sequences. We found that in order to reconstruct tracks of people over longer periods of time, it is more reliable to merge hypotheses from different short tracklets rather than increasing the length of the tracklets itself. First, we compute overlapping fixed-length tracklets ($m = 6$) starting at every frame of the sequence. As tracklet optimization relies on good initialization, the use of overlapping sequences ensures that each strong detection is used multiple times for initialization.

We then exploit that every frame is overlapped by several different tracklets (including ones with different starting frames), which provide competing hypotheses that explain the same image evidence. In principle it would be possible to choose the best sequence of hypotheses using their joint posterior (i.e., an extension of Eq. (3.9)). This is, however, computationally prohibitive since the large state-space of all possible combinations of hypotheses cannot be searched efficiently without making simplifying assumptions. Instead, we select hypotheses using pairwise relations between them by introducing a first-order Markov assumption on the hypothesis sequence.

Let the length of the complete image sequence be equal to M. We denote the set of all hypotheses obtained from individual tracklets in frame j by $\mathbf{h}^j = [h_1^j, \ldots, h_{n_j}^j]$. We will call the track given by a set of hypotheses $\mathcal{H} = [h_{j_1}^1, \ldots, h_{j_M}^M]$ optimal if it maximizes the joint sequence probability according to the hidden Markov model:

$$p(\mathcal{H}) = p_{img}(h_{j_1}^1) \prod_{k=2}^{M} p_{img}(h_{j_k}^k) p_{trans}(h_{j_k}^k, h_{j_{k-1}}^{k-1}). \tag{3.12}$$

In this expression $p_{img}(h_{j_k}^k)$ is computed using the people detection model from Sec. 3.2. The transition probability is p_{trans} is given by

$$\begin{aligned} p_{trans}(h_{j_k}^k, h_{j_{k-1}}^{k-1}) &= p_{dynamic}(h_{j_k}^k, h_{j_{k-1}}^{k-1}) \cdot \\ &\quad\; p_{app}(h_{j_k}^k, h_{j_{k-1}}^{k-1}), \end{aligned} \tag{3.13}$$

where $p_{dynamic}(\cdot, \cdot)$ is our dynamical model consisting of Gaussian position dynamics and the GPLVM articulation dynamics, and $p_{app}(\cdot, \cdot)$ is computed using an appearance model of the hypothesis, which can be based on color histograms of person parts, oriented edge features, or any other appearance description. We use color histograms extracted from the detected parts, and use the Bhattacharyya distance to model the appearance compatibility.

The optimal sequence can be efficiently found by maximizing Eq. (3.12) using the Viterbi algorithm. If a given image sequence contains only one person that is never occluded, the proposed approach is able to reconstruct its track from the individual tracklets.

For the case of more complex image sequences we have adopted the following strategy that has proven to be quite effective (see Sec. 3.4 for results): Let i be the

Dataset	HOG		4D-ISM		single-frame		tracklets	
TUD-Pedestrians	0.53	-	0.28	0.68	0.81	**0.84**	-	-
TUD-Campus	0.22	-	0.6	0.71	0.7	0.75	0.82	**0.85**

Table 3.1: Recall of 4D-ISM, HOG, and our detectors at precision equal to 0.9 and at equal error rate on "TUD-Pedestrians" and "TUD-Campus" datasets.

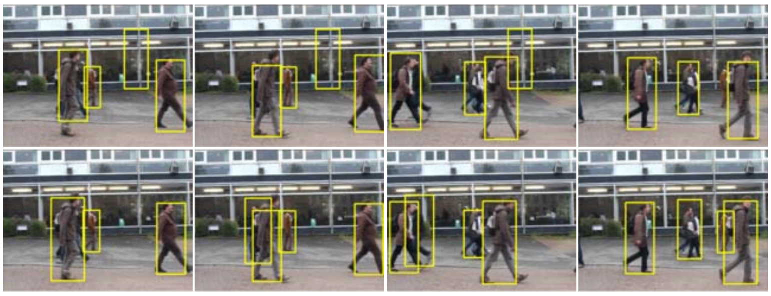


Figure 3.11: Examples of detector hypotheses (top row) and tracklet hypotheses (bottom row) at equal error rate on the "TUD-Campus" dataset.

number of the current frame in the sequence (in the beginning $i = 1$). We proceed by iteratively computing the optimal track starting with i. At the nth iteration of the Viterbi algorithm, we compute the transition probabilities between the hypotheses of the optimal track at frame $i+n$ and each of the hypotheses in the frame $i+n+1$. If the person either walks out of the image or becomes heavily occluded, all of the transition probabilities will be low, which means that we can end the track. In that case all its hypotheses are removed from the sets of hypotheses $\mathbf{h}^j$, $j = i, \dots, i+n$. We then repeat the procedure again starting from frame i until $\mathbf{h}^i$ becomes empty. In this case we set $i = i + 1$ and repeat the process. As a result of this iterative computation we obtain a set of tracks with hypotheses that are consistent in both motion and articulation. To connect such tracks across long-term occlusions we again use the appearance of the person as well as a coarse motion model to decide if two tracks correspond to the same person. The appearance model is the same as used for modeling the appearance of individual hypotheses. For the motion model we only require tracks to have consistent movement direction (i.e. left or right). In practice, we found that even such a simplistic method is sufficient since only few possible tracks are available after the initial detection stage.

3.4 Experiments

We evaluate our approach both quantitatively and qualitatively. In the first experiment, we compare the detection performance of the single-frame person detector proposed in Sec. 3.2 with the tracklet-based detector. Tracklet-based detections are obtained by first detecting tracklets in the image sequence as described in Sec. 3.3.2, grouping together hypotheses from all tracklets corresponding to a particular image

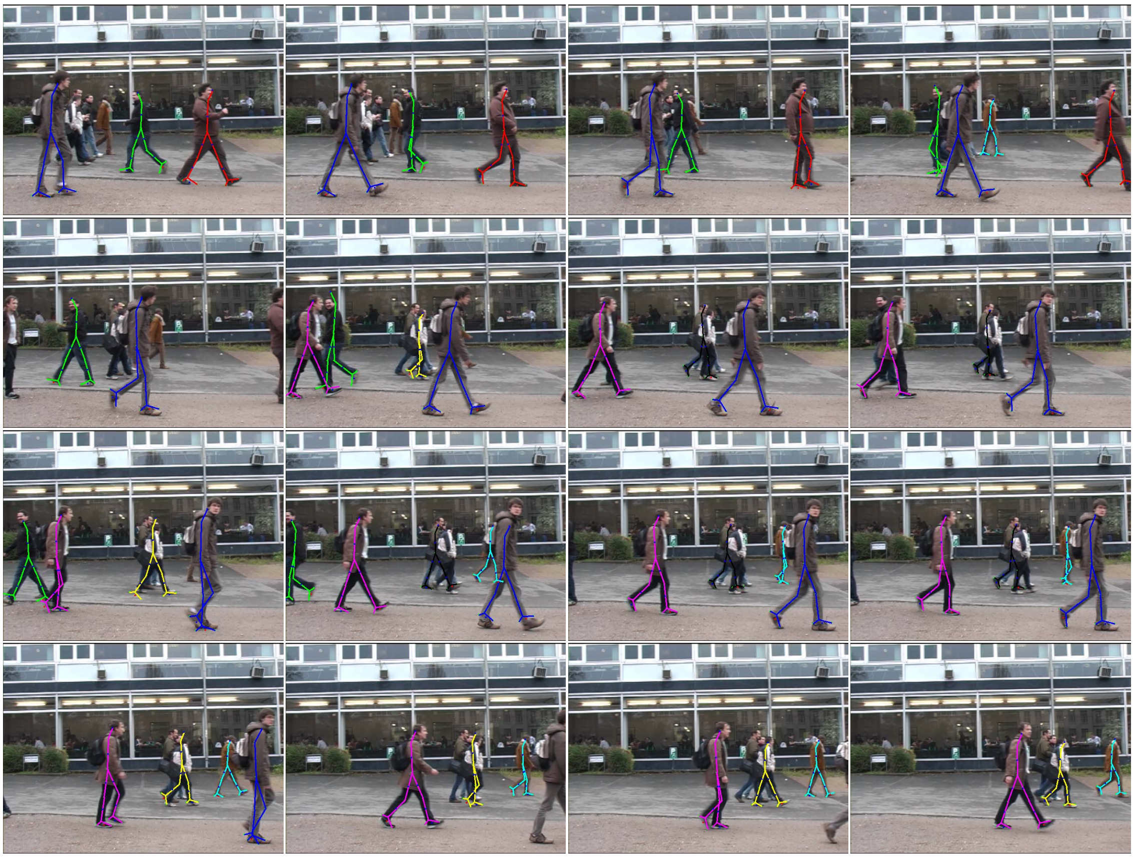


Figure 3.12: Detection and tracking on "TUD-Campus" dataset (see supplementary material).

of the sequence, and performing non-maxima suppression on this set of hypotheses. In this process the score of a hypothesis is set to score of the corresponding tracklet, which is given by Eq. 3.9.

The comparison of both detectors is done on a sequence with multiple full and partial occlusions. Fig. 3.12 shows several example images from the sequence. Note that in such cluttered sequences ground truth annotation is difficult as it is unclear how to decide when a partially or fully occluded person should be included. In order to have a fair evaluation for the single-frame person detector we decided to annotate people when they are at least 50% visible. The quantitative comparison of the single-frame detector and the tracklet-based detector is given in Fig. 3.10. The tracklet-based detector improves the precision considerably. Fig. 3.11 shows sample detections. As can be seen, the single-frame detector obtains false positives on the background (images 1, 2, and 3), whereas the tracklet-based detector can successfully filter those false-positives. At the same time the tracklet-based detector is capable of detecting several partially occluded people (e.g.in image 2 and 4) that cannot be detected in a single frame alone.

In the second experiment we evaluate the tracks produced by our system on the

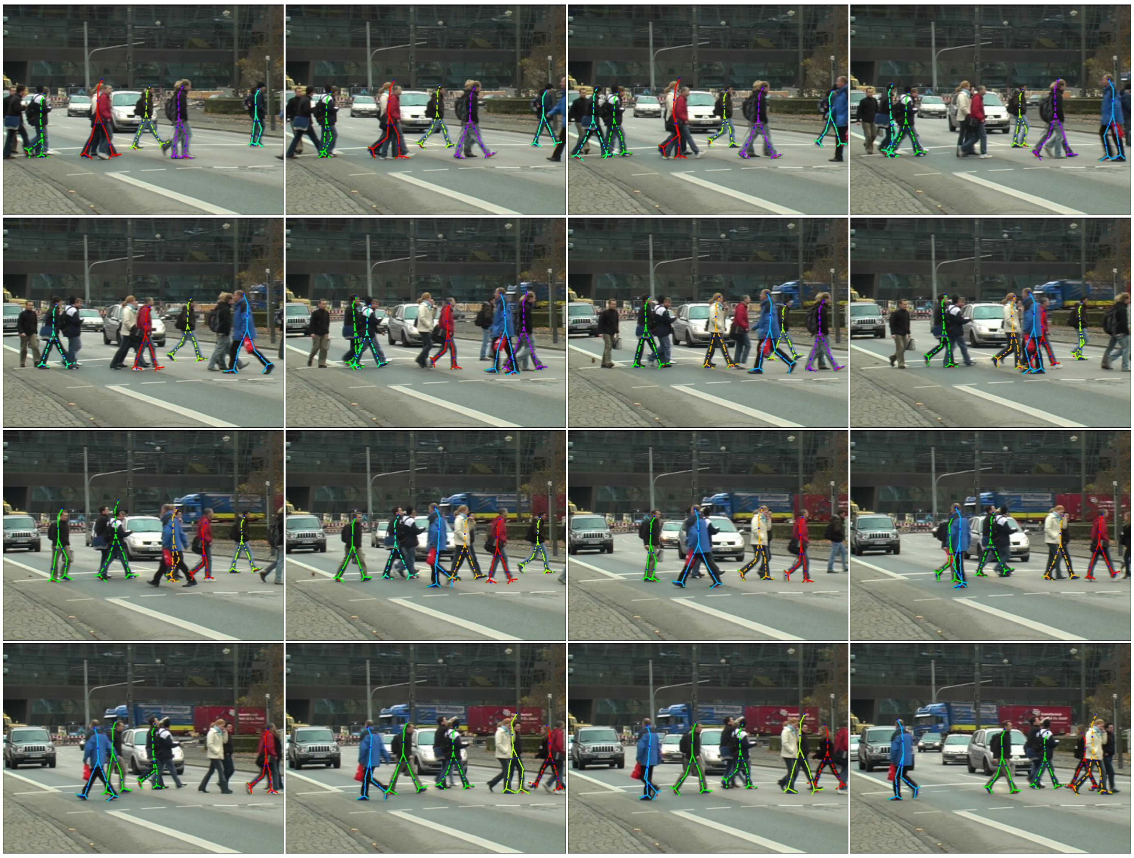


Figure 3.13: Detection and tracking on “TUD-Crossing” dataset (see supplementary material).

sequence from the first experiment and an additional sequence with significantly larger number of people. Example snapshots from the resulting tracks are shown in figures 3.12 and 3.13 respectively. Clearly, in any single frame a significant number of people is detected and their limb-configuration is correctly inferred. Interestingly, we obtain tracks for nearly all people in these sequences, and in particular in sequence 1 we obtain tracks for all people even though some of them become fully occluded over significant time intervals. Quite importantly, on this sequence we can also differentiate between hypotheses of two people walking side by side, with one person occluding the other most of the time. The complete videos of both sequences can be found in the supplementary material.

3.5 Conclusion

In this chapter we have introduced a novel method capable of detecting and tracking people in cluttered real-world scenes with many people and changing backgrounds. For this we have introduced pedestrian detector based on the articulated limb-based model, which outperforms the state-of-the-art on single frame person detection, and also delivers estimates for body part articulations. A dynamic model based on a hierarchical Gaussian process latent variable model is used to further improve people-detection by people-tracklet detection in image sequences. Those tracklets are then used to enable people-tracking in complex scenes with many people and long-term occlusions.

Notice that despite its good performance the model introduced in this chapter is purely generative and does not rely on any form of discriminative learning. Such good performance can in part be explained by a strong spatial prior on body configurations and motions of people. The reliance on such a strong prior is perhaps one of the weaknesses of the framework described in this chapter. In order to alleviate the strong *a-priori* assumptions, in the next chapter we introduce a novel part-based people detector that relies on discriminative appearance model and is capable of detecting and estimating pose of people in more generic settings than we have considered so far.

4

Discriminative Appearance Model for Pictorial Structures

In this chapter we present a new approach to people detection and articulated pose estimation. At the core of our approach is a novel appearance model based on densely sampled local features and discriminative classifiers. We combine this appearance model with a loose-limbed articulated body model capable of capturing relative positions and orientations of the main body parts. In contrast to the body model used in Chapter 3 this model is more general and is not limited to representing articulations of pedestrians.

Our approach builds on the *pictorial structures* model. Similarly to the original pictorial structures approach of Fischler and Elschlager (1973) and Felzenszwalb and Huttenlocher (2005) we model the appearance of each part independently and represent object as a collection of rigid body parts coupled by pairwise spatial relationships. We identify the components of the model that make it applicable to the uncontrolled real-world scenes considered in this thesis. As a first important ingredient of our approach, we propose a discriminative appearance model based on densely sampled local descriptors and AdaBoost classifiers. Secondly, we interpret the normalized margin of each classifier as a likelihood in a generative model and compute marginal posteriors for each part using belief propagation. Thirdly, non-Gaussian relationships between parts are represented as Gaussians in the coordinate system of the joint between parts. Additionally, in order to cope with shortcomings of the tree-structured pictorial structures model, we augment the model with an additional repulsive factor to discourage over-counting of image evidence. We demonstrate that the combination of these components within the pictorial structures framework results in a generic model that achieves state-of-the-art performance for several datasets on different tasks: people detection, upper body pose estimation, as well as of full body pose estimation.

4.1 Introduction

People detection and articulated pose estimation are two challenging and long-standing problems in computer vision. Addressing these problems in real-world

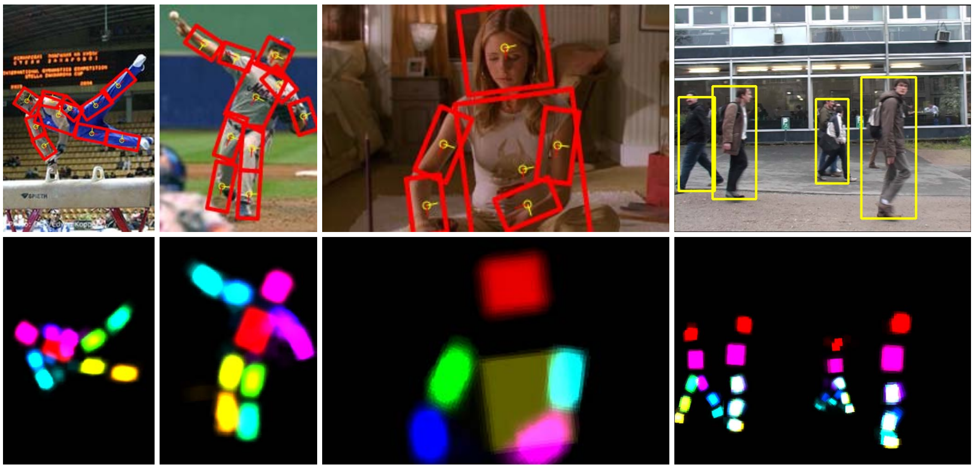


Figure 4.1: **Sample results** (left to right): Full body pose estimation (1st and 2nd column), upper-body pose estimation, and pedestrian detection. *Bottom:* Part posteriors.

scenes is difficult because a successful approach should be both discriminative enough in order to distinguish people from the large amount of background clutter, and representative and flexible enough to capture the large variation in human appearance and poses. While numerous approaches have been proposed over the years, none of them has been demonstrated to be equally applicable for both people detection and pose estimation. Interestingly, quite a number of them (Andriluka et al., 2008; Ferrari et al., 2008; Ramanan, 2006; Ramanan and Sminchisescu, 2006; Zhang et al., 2006), while specializing on one of these tasks only, build on the same basic pictorial structures model (Felzenszwalb and Huttenlocher, 2005; Fischler and Elschlager, 1973). In this chapter we show that given an appropriate representation for appearance and spatial components, the pictorial structures model obtains equal or even superior performance compared to many specialized approaches. This results in a generic model equally applicable for human detection and pose estimation, which allows to detect upright people (i.e., pedestrians (Leibe et al., 2005)), as well as highly articulated people (e.g., in sports scenes (Ramanan, 2006)), and to estimate their poses (see Fig. 4.1 for examples).

The *pictorial structures* model is a powerful and general, yet simple generative object/body model that allows for exact and efficient inference of (body) part constellations. In order to model the appearance of individual body parts we build upon *strong part detectors* (Andriluka et al., 2008; Mikolajczyk et al., 2006; Viola et al., 2003), which have shown to enable object and people detection in challenging scenes, but have not yet proven to enable state-of-the-art articulated pose estimation. While previous work has either focused on strong part detectors or on powerful body models, our work combines the strengths of both.

We combine a discriminative appearance model with a generative pictorial structures model in a *hybrid approach* (c.f. Tu et al. (2005)) by interpreting the normalized classifier margin as the likelihood of the image evidence. As a result, we obtain a

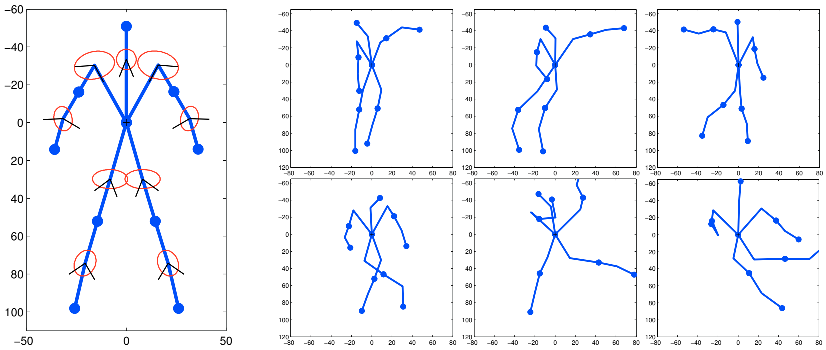


Figure 4.2: Left: **Kinematic prior** learned on a multi-view and multi-articulation dataset (Ramanan, 2006). Mean part positions (blue dots) and covariances of the part relations in the transformed space (red ellipses) are shown. Right: **Several independent samples** from the learned prior (to ease visualization visualized assuming a fixed torso position and orientation).

generic model for people detection and pose estimation, which not only outperforms recent work in both areas by a large margin, but is also surprisingly simple and allows for exact and efficient inference.

We perform a careful performance analysis of the model, and quantify changes in model performance due to the type of local descriptor, discretization step between evaluated part locations, the number of boosting rounds, and due to variants of belief propagation.

While demonstrating state-of-the-art performance, our initial model has a tree structure, which does not allow to incorporate complex dependencies between arbitrary body parts. The underlying assumption is that detector responses of different body parts are independent from each other. While this often works well in practice, it can lead to overcounting of image evidence for parts with similar appearance (such as the two lower legs). Previously proposed solutions to this problem either assume availability of foreground segmentation masks (Jiang, 2009; Zhang et al., 2009) or require explicit modeling of foreground, background and depth ordering of body parts (Buehler et al., 2008; Sudderth et al., 2004; Wang and Mori, 2008), and are not directly applicable to our discriminative appearance representation. Therefore, in this work we pursue a different avenue and introduce an additional repulsive factor (Ferrari et al., 2009*b*) into the model that discourages configurations with overlapping parts. We quantify the improvements in body-part localization due to this additional factor on the pedestrian and full body pose estimation tasks. We also demonstrate that even though the repulsive factor introduces loops into the model structure, we can still perform inference efficiently by first filtering the search space with the original tree-structured model and then applying the full model to a subset of most promising part configurations, which can be seen as a form of *probabilistic search space reduction*.

4.2 Generic Model for People Detection and Pose Estimation

To facilitate reliable detection of people across a wide variety of poses, we follow Felzenszwalb and Huttenlocher (2005) and assume that the body model is decomposed into a set of parts. The body part configuration is denoted as $L = \{\mathbf{l}_0, \mathbf{l}_1, \ldots, \mathbf{l}_N\}$, where the state of part i is given by $\mathbf{l}_i = (x_i, y_i, \theta_i, s_i)$. (x_i, y_i) denotes the part position in image coordinates, θ_i the absolute part orientation, and s_i the part scale, defined to be relative to the part size in the training set.

Depending on the task, the number of body parts may vary (see Figs. 4.2 and 4.3, and Sec. 4.4). For upper body detection and pose estimation we use six parts: head, torso, and left/right lower/upper arms. In case of full body pose estimation, we additionally consider four lower body parts (left/right upper/lower legs) resulting in a 10-part model. For pedestrian detection we do not use arms, but add feet, leading to an 8-part model.

Given the image evidence E, the posterior of the part configuration L is given by

$$p(L|E) \propto p(E|L) \cdot p(L) \tag{4.1}$$

where $p(E|L)$ is the likelihood of the image evidence given a particular body part configuration and $p(L)$ corresponds to a kinematic tree prior. Here, both terms are learned from training data, either from generic data or trained more specifically for the application at hand. To make such a seemingly generic and simple approach work well, and to compete with more specialized models on a variety of tasks, it is necessary to carefully design the appropriate prior $p(L)$ and an appropriate image likelihood $p(E|L)$. In Sec. 4.2.1, we will first introduce our generative kinematic model $p(L)$, which closely follows the approach of (Felzenszwalb and Huttenlocher, 2005). In Sec. 4.2.2, we will then introduce our discriminatively trained appearance model $p(E|L)$.

Given such a model, we estimate articulated poses by finding the most probable configuration for each part given the image evidence through maximizing the marginal posterior $p(\mathbf{l}_i|E)$. In case of multiple people this generalizes to finding the modes of the marginal posterior density.

We use our articulated body model also for people detection in order to cope with the large variety of possible body poses. This is similar in spirit, e.g., to (Andriluka et al., 2008; Felzenszwalb et al., 2008; Ramanan, 2006). To that end we first compute the marginal distribution of the torso configuration and then use its modes to deterministically predict bounding box detections.

4.2.1 Kinematic tree prior

The first important component in our pictorial structures approach is the prior $p(L)$, which encodes probabilistic constraints on part configurations. A common source of such constraints are kinematic dependencies between parts. Such kinematic constraints can be captured probabilistically using a tree-structured graphical model. In such a model the prior on part configurations factorizes into the product of unary and pairwise terms:

$$p(L) = p(\mathbf{l}_0) \prod_{(i,j) \in G} p(\mathbf{l}_i | \mathbf{l}_j), \tag{4.2}$$

where we let G denote the set of all directed edges in the kinematic tree and assign $\mathbf{l}_0$ to be the root node (torso).

It is possible to incorporate action specific constraints into the prior and to combine them with articulation dynamics to enable tracking (see e.g. (Lan and Huttenlocher, 2005; Urtasun et al., 2006)). However, we omit such extensions, as they would restrict the applicability of the model to rather specific scenarios.

Part relations. To fully specify the prior of Eq. (4.2) we have to define the various components. The prior for the root part configuration $p(\mathbf{l}_0)$ is assumed to be uniform to allow for a wide range of possible configurations. The part relations are modeled using Gaussian distributions (c.f. (Felzenszwalb and Huttenlocher, 2005; Ramanan and Sminchisescu, 2006)), which enable efficient inference (see below). This may seem like a significant limitation as, for example, the distribution of the forearm configuration given the upper arm configuration follows a semi-circular rather than a Gaussian shape. It was pointed out in (Felzenszwalb and Huttenlocher, 2005) that such a distribution is not Gaussian in the image coordinates, but that it is possible to transform it to a different space, in which the spatial distribution between parts is captured well by a Gaussian distribution. More specifically, to model $p(\mathbf{l}_i|\mathbf{l}_j)$, we transform the part configuration $\mathbf{l}_i = (x_i, y_i, \theta_i, s_i)$ into the coordinate system of the joint between the two parts using the transformation:

$$T_{ji}(\mathbf{l}_i) = \begin{pmatrix} x_i + s_i d_x^{ji} \cos\theta_i - s_i d_y^{ji} \sin\theta_i \\ y_i + s_i d_x^{ji} \sin\theta_i + s_i d_y^{ji} \cos\theta_i \\ \theta_i \\ s_i \end{pmatrix}. \tag{4.3}$$

Here, the $d^{ji} = (d_x^{ji}, d_y^{ji})^T$ is the position of the joint between parts i and j represented in the coordinate system associated with $\mathbf{l}_i$. Note that d^{ij} will vary across people and will also depend on the out-of-plane rotation of $\mathbf{l}_i$. In order to accommodate such variability, the value of d^{ji} is regarded as a model parameter and estimated from training data.

The part relation is modeled as a Gaussian with respect to the transformed part configurations:

$$p(\mathbf{l}_i|\mathbf{l}_j) \propto \mathcal{N}(T_{ji}(\mathbf{l}_i) - T_{ij}(\mathbf{l}_j) | \mu^{ji}, \Sigma^{ji}), \tag{4.4}$$

where T_{ij} is the transformation that maps the configuration of the parent part $\mathbf{l}_j$ to the position of the joint between parts i and j, $\mu^{ji} = (0, 0, \mu_\theta^{ji}, 0)$ represents the

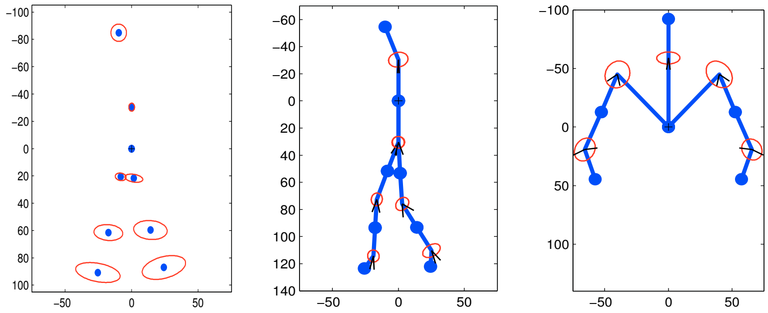


Figure 4.3: **Priors on the part configurations** (left to right): Pedestrian detection (star vs. tree model) and upper body detection.

preferred relative orientation of the parts, and Σ^{ji} is the covariance matrix determining the stiffness of the joint. The parameters d^{ji}, μ_{θ}^{ji} and Σ^{ji} can be learned directly using maximum likelihood estimation. It is worth noting that this corresponds to a so-called "loose limbed" model (c.f. Sigal and Black (2006*c*)) as the limbs do not rigidly rotate around the joints. Instead, the parts are only loosely attached to the joint by the Gaussian distribution from Eq. (4.4), which helps reducing brittle behavior. In our experiments, we found this procedure to work much better than the non-parametric part relation model used in Ramanan (2006).

Learned prior. Fig. 4.2 shows the prior learned from the multi-view and multi-articulation **People** dataset from (Ramanan, 2006). This dataset includes people performing a large variety of activities ranging from simple walking to acrobatic exercises. Samples from this model exhibit a large variety of poses (see Fig. 4.2). Fig. 4.3 shows priors learned on the TUD-Pedestrians dataset (Andriluka et al., 2008), which contains upright pedestrians in street scenes and from the **Buffy** dataset (Ferrari et al., 2008) containing upper body configurations in TV footage.

4.2.2 Discriminatively trained part models

Before introducing our formulation of the likelihood $p(E|L)$ of image evidence given the part configuration, we would like to discuss several considerations that motivate our approach. Our principal aim is to detect people in unconstrained environments and arbitrary poses. As the search space over all possible poses is typically high-dimensional, successful approaches often resort to techniques for search space reduction (Ferrari et al., 2008). As has been argued before (Tu et al., 2005), discriminatively learned detectors allow to reduce the search space for generative models significantly, thereby enabling both efficient as well as effective inference in challenging real world scenes. Following this avenue, we rely on discriminatively learned part detectors to effectively reduce the search space as much as possible. At the same time we aim to postpone the final decision on the part configuration to the final stage when evidence from all body parts is available, rather than to deterministically prefilter possible part configurations at the individual part detection level. Therefore, we densely evaluate all possible part positions, orientations, and scales, which is in contrast to bottom-up appearance models (e.g. (Andriluka et al., 2008; Mikolajczyk et al., 2006)) based on a sparse set of interest points. We believe and also show in the experimental section that dense sampling is better suited for detecting body parts, especially in cases of low contrast and partial occlusion.

Assuming that each part is independently generating the portion of the image evidence within its bounding box, it follows that the likelihood can be decomposed into the product of individual part likelihoods:

$$p(E|L) = \prod_{i=0}^{N} p(E|\mathbf{l}_i) = \prod_{i=0}^{N} p(\mathbf{e}_i(\mathbf{l}_i)) p(E \backslash \mathbf{e}_i(\mathbf{l}_i)), \tag{4.5}$$

where $\mathbf{e}_i(\mathbf{l}_i)$ denotes the portion of the evidence for part i covered by the part bounding box in configuration $\mathbf{l}_i$. As is common in the literature (Felzenszwalb and Huttenlocher, 2005; Sigal and Black, 2006*b*), we model the image evidence outside of the part bounding box with a uniform background distribution, so that the term $p(E \backslash \mathbf{e}_i(\mathbf{l}_i))$ depends only on the image size and the size of the part bounding box in configuration $\mathbf{l}_i$. Recall that in our model the only paramter that influences the dimensions of the part bounding box is a part scale. During inference we assume that the person is generating the evidence E within its bounding box only and that the scale of the body parts is fixed relative to this bounding box. Under this assumption the term $p(E \backslash \mathbf{e}_i(\mathbf{l}_i))$ is independent of the part configuration and does not influence the position of the modes in the posterior, and we ignore it in the rest of the presentation.

It should be noted that the naive Bayes assumption made in Eq. 4.5 disregards potential correlations in the appearance of body parts and assumes that parts do not occlude each other. Nonetheless, it significantly simplifies computation and has been found to work well when augmented with a subsequent verification step (c.f. Felzenszwalb and Huttenlocher (2005)). Importantly, this enables efficient and exact inference and leads to very competitive experimental results as we demonstrate in Sec. 4.4.

Using Eq. (4.5) the posterior over the configuration of parts (Eq. (4.1)) factorizes as:

$$p(L|E) \propto p(\mathbf{l}_0) \cdot \prod_{i=0}^{N} p(\mathbf{e}_i(\mathbf{l}_i)) \cdot \prod_{(j,k) \in G} p(\mathbf{l}_j|\mathbf{l}_k). \tag{4.6}$$

Boosted part detectors. In our model we represent image evidence E by a densely computed grid of local image descriptors (e.g., shape context (Belongie et al., 2000) or SIFT (Lowe, 2004) – see Sec. 4.3.1 for a comparison)

Individual part likelihoods $p(\mathbf{e}_i(\mathbf{l}_i))$ are based on discriminatively trained part classifiers. The feature vector used for classification is obtained by concatenating all local descriptors at given orientation and scale whose centers fall inside the part bounding box, so that some dimensions of the feature vector also capture the surrounding context. During detection we discretize the range of positions, scales, and orientations and exhaustively scan the discretized part configurations in a sliding window fashion. To predict the presence of a part, we train an AdaBoost classifier (Freund and Schapire, 1997) with simple decision stumps that consider whether one of the dimensions of the feature vector is above or below a threshold.

Denoting the feature vector by $\mathbf{e}$, the stump with index t is given by $h_t(\mathbf{e}) = \text{sign}(\xi_t(e_{n(t)} - \varphi_t))$, where φ_t is a threshold, $\xi_t \in \{-1, +1\}$, and $n(t)$ indexes the descriptor bin chosen by the stump. Training of the AdaBoost classifier proceeds as usual, yielding a strong classifier $H_i(\mathbf{e}) = \text{sign}\left(\sum_t \alpha_{i,t} h_{i,t}(\mathbf{e})\right)$ for each part i. Here, $\alpha_{i,t}$ are the learned weights of the weak classifiers.

To integrate the discriminative classifiers into the generative probabilistic framework described above, it is necessary to give a probabilistic meaning to the classifier outputs.

We denote the normalized margin of the boosted classifier as

$$m(\mathbf{e}_i(\mathbf{l}_i)) = \frac{\sum_t \alpha_{i,t} h_{i,t}(\mathbf{e}_i(\mathbf{l}_i))}{\sum_t \alpha_{i,t}}, \tag{4.7}$$

where $\mathbf{e}_i(\mathbf{l}_i)$ is the vector of concatenated local descriptors corresponding to part i in configuration $\mathbf{l}_i$. The likelihood function of the part configuration $\mathbf{l}_i$ is defined as

$$p(\mathbf{e}_i(\mathbf{l}_i)) = \frac{1}{Z_i} \max\left(m(\mathbf{e}_i), \varepsilon_0\right), \tag{4.8}$$

where ε_0 is a small positive constant, which makes the model more robust to missing and occluded parts[1] and Z_i is a normalization constant given by

$$Z_i = \int \max\left(m(\mathbf{e}_i(\mathbf{l}_i)), \varepsilon_0\right) d\mathbf{e}_i. \tag{4.9}$$

Z_i is finite because in our model individual feature dimensions are bounded to the unit interval and the expression under the integral corresponds to a piecewise continuous function given by the classifier margin thresholded from below by the constant ε_0. Note, that the functional form of the likelihood in Eq. 4.8 is not directly related to any generative probabilistic model of the image evidence. Instead, it models the probability of a feature vector based on the confidence of the boosted part detector. We see such a definition of the likelihood as a shortcut that allows us to avoid the complexity of properly modeling the image statistics of the body part regions and the background. Although there has been considerable interest in generative models for natural images (Roth and Black, 2009), such models are still subject of active research and are often too complex to be applicable to our task. At the same time, in Sec. 4.3.1 we experimentally demonstrate that our likelihood is superior to

---

[1] In our experiments we set $\varepsilon_0 = 10^{-4}$.

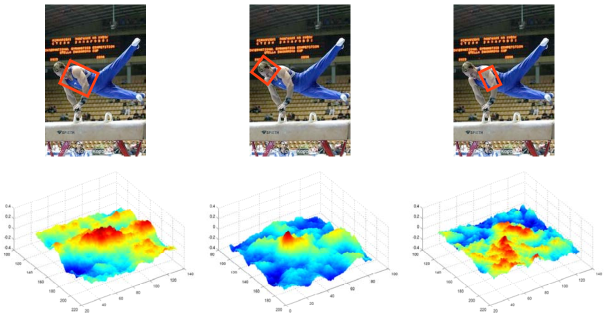


Figure 4.4: Likelihood maps of the torso, head and upper arm visualized for the neighborhood around the ground-truth part location.

the simple likelihood of Ramanan (2006), which is based on a product of pixel-wise terms and similar to our likelihood is trained discriminatively.

Note, that in our model we do not attempt to learn the relative importance of the likelihoods of different parts. Typically, such learning is implemented by formulating the model in log-linear form in which the likelihood of each part has an additional weighting parameter that makes the part likelihoods more peaky or more flat depending on the importance of the part (Ramanan, 2006; Sapp, Jordan and Taskar, 2010). We observe that our likelihoods appear to implicitly contain a bias towards more salient parts and therefore explicit learning of likelihood weights may not be essential for good performance. Consider the visualization of likelihoods of several parts shown on Fig. 4.4. Note that more salient parts, such as head or torso, have likelihoods with clearly distinguishable local maxima, while less salient parts, such as the upper arm, have likelihoods that are more flat and have positive responses at multiple locations. During inference parts with more peaky likelihoods will have more influence on the outcome, because the cost of shifting the position of such part away from the location of the peak in the likelihood will be high. Parts with flat likelihoods will be easier to shift, since their likelihood has similar values for multiple part configurations. This is similar to a behavior that is achieved by assigning higher weights to likelihoods of parts that can be detected more reliably. Nonetheless, we expect that the explicit learning of part likelihood weights may still be beneficial and plan to address this in future work.

Training. For training, each annotated part is scaled and rotated to a canonical pose prior to learning. Note that this in-plane rotation normalization significantly simplifies the classification task. Additionally, we extend the training set by adding small scale, rotation, and position transformations to the original images. The transformation parameters are obtained by sampling from Gaussians with $\sigma_{scale} = 0.05$, $\sigma_{rot} = 1°$ and $\sigma_{pos} = 2$ pixels. Negative feature vectors are obtained by uniformly sampling image regions outside of the object bounding box. After initial training, the classifiers are re-trained with the negative training set augmented with false positives produced by the initial classifier. This is commonly referred to as bootstrapping. We have found that bootstrapping is essential to obtain good performance with our discriminative part detectors for the task of people detection. Please refer to Fig. 4.11(a) for a comparison of results with and without bootstrapping.

4.2.3 Exact model inference

An important property of our tree-based model is that optimal inference is tractable. Specifically, we can compute the globally optimal body configuration by MAP inference using the max-product algorithm (Felzenszwalb and Huttenlocher, 2005). We can also compute exact marginal distributions using the sum-product algorithm (Pearl, 1988), which is used here, for example, to obtain marginals for pedestrian detection. To that end, we interpret the underlying directed graphical model as a factor graph and apply standard factor graph belief propagation (Kschischang et al., 2001). Interestingly, the sum-product algorithm turns out to be more robust w.r.t. the discretization of the part configuration space (see Sec. 4.3.3).

It has been shown that expensive summations necessary in the sum-product algorithm can be efficiently computed using Gaussian convolutions whenever the part dependencies are modeled using Gaussian distributions (Crandall et al., 2005; Ramanan, 2006). However, care has to be taken when doing so, as the part relations in our model are Gaussian not in the image space, but rather in the transformed space of the joint. To apply the efficient algorithm nonetheless, we rely on the approach suggested in (Felzenszwalb and Huttenlocher, 2005): we transform the messages into the coordinate system of the joint using Eq. (4.3), then apply Gaussian convolutions there, and finally transform the result into the coordinate system of the target part, which is possible since the transformation from the part configuration to the position of the joint is invertible. If the Gaussian distribution in the transformed space is separable these computations are even more efficient.

Similarly, it is also possible to efficiently compute the maxima of $p(L|E)$ using the max-product algorithm. In this case the messages are computed using a generalized distance transform in time linear in the number of possible states of the part variable (see (Felzenszwalb and Huttenlocher, 2005) for details).

4.2.4 Incorporation of additional relationships between body parts

While the original pictorial structures model (Fischler and Elschlager, 1973), did not impose any restrictions on the structure of the relationships between model parts, the exact and efficient inference with generalized distance transform or Gaussian convolutions is only possible if these relationships are assumed to have a tree structure (Felzenszwalb and Huttenlocher, 2005). This in turn requires that the image likelihood can be factorized into a product of local part likelihoods that are computed independently of each other.

While these assumptions often lead to competitive results, they can obviously be

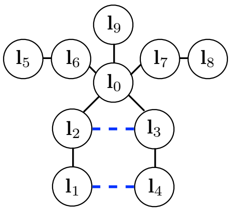


Figure 4.5: Visualization of the factors between body parts in our model. Solid lines correspond kinematic factors and dashed lines correspond to repulsive factors.

violated in practice as, for example, in the case of body parts occluding each other. In such cases it is desirable to extend the model with additional dependencies between parts. These, however, introduce loops in the model structure and complicate inference. The avenue we are exploring in this work consists of a two stage procedure: first we perform inference using the original model with kinematic constraints only, then we sample from the posterior marginals of each part and perform inference with the full model using the samples as the new state-space. As before, we apply factor graph belief propagation for inference, however here loopy belief propagation (LBP), for which we use the implementation of Mooij (2009). Clearly, there are no guarantees that the true positive hypothesis will always be sampled from the posterior of the tree model. In practice, our approximative two-stage inference procedure improves over exact inference in the tree model, which suggests that true positives are included in the reduced state-space most of the time.

One of the cases where this two stage procedure becomes necessary is the model shown on Fig. 4.5. This model includes additional repulsive factors between the lower extremities that compensate for correlations between responses of part detectors. Since these correlations are not taken into account by the tree-structured model discussed before, it might prefer to explain the same image region multiple times by different parts when the local likelihood of this region is sufficiently strong. We take this factor to have the form similar to the one used in Ferrari et al. (2009*b*):

$$f(\mathbf{l}_i, \mathbf{l}_j) = \begin{cases} \exp(-\beta) & : \quad \mathrm{IoU}(\mathbf{l}_i, \mathbf{l}_j) > \delta \\ 1 & : \quad \text{otherwise,} \end{cases} \tag{4.10}$$

where $\mathrm{IoU}(\mathbf{l}_i, \mathbf{l}_j)$ is the ratio of intersection and union of the bounding boxes of parts i and j, the parameter β controls how strong the parts are pushed away from each other, and δ defines the minimal relative intersection between parts required for the repulsive factor to take effect. In our experiments we set $\beta = 1.25$ and $\delta = 0.5$ and keep their values constant in all trials.

We compare tree-structured and loopy versions of our model in Sec. 4.3.2 and demonstrate that the loopy model can significantly reduce the effect of over-counting of image evidence and achieves better pose estimation results. In addition, we have also performed a series of experiments with a loopy model that contains repulsive factors between the upper arms and forearms. Such additional factors did not achieve any improvement of the pose estimation results, which is most likely due to the form of the prior distribution (c.f. Fig. 4.2), which already discourages placement of upper body limbs at the same image location.

4.3 Evaluation of the Model Components

We begin the experimental part of this chapter with the evaluation of various aspects our model. In particular, we perform a detailed evaluation of our discriminative appearance representation, compare the performance of tree and non-tree models extended with a repulsive factor, and contrast inference with sum-product and max-product algorithms. For the experiments in this section we use the full-body pose estimation task as testbed, since this is the most complex task considered in our work and includes other tasks as a special case. Our experimentation relies on the **People** dataset of Ramanan (2006), which includes annotated humans across a variety of views, articulations, and activities.

In our experiments we change only one aspect or parameter of the model at a time, keeping the rest of the model parameters fixed. The appearance of body parts is represented using a shape context descriptor with 4 angular, 3 radial bins, and 8 discretized gradient directions, ignoring the sign of the gradient vector. AdaBoost part detectors are trained for 500 rounds. Inference is performed with sum-product belief propagation. Bootstrapping of part detectors is performed only in the pedestrian detection case, since it did not help to improve performance in the other cases. The discretization step size is 15 degrees for the part orientation, and 0.1 for the object scale. At the training scale the part detectors are evaluated with a step size of 8 pixels in case of full-body pose estimation and pedestrian detection, and with step size of 12 pixels for the upper body pose estimation[2].

In our experiments, we quantify the pose estimation performance of the model by computing the average percentage of correctly localized body parts. A body part is considered to be localized correctly if the endpoints of its segment lie within 50% of the ground-truth segment length from its true position. This measure, often referred to as "percentage of correct parts" (PCP), was originally introduced in Ferrari et al. (2008), and was subsequently used in Ferrari et al. (2009*b*); Eichner and Ferrari (2009); Johnson and Everingham (2009) for performance evaluation on both upper- and full-body pose estimation tasks.

4.3.1 Evaluation of appearance model parameters

Image descriptor.

---

[2] The step size is made larger in the upper body case to keep the ratio between average size of the body parts and step between evaluated locations comparable across all datasets.

Image descriptor. We compare the performance of shape context descriptors as previously used in Andriluka et al. (2009) with SIFT descriptor (Lowe, 2004), and edge templates obtained using the code from (Ramanan, 2006) and integrated into our pose estimation framework. For reference we also include results of Johnson and Everingham (2009), where HOG descriptors (Dalal and Triggs, 2005) are used for appearance representation. We use the implementation of shape context from (Mikolajczyk and Schmid, 2005), which is similar to the original shape context descriptor (Belongie et al., 2000), but captures the distribution of gradient orientations instead of distribution of the edge pixels. We rely on mean/variance normalization to make the descriptor invariant to illumination changes and use the Canny edge detector to find edges in the image patch.

The results of the comparison of different descriptors are shown in Tab. 4.1. The first interesting outcome of this experiment is that the original SIFT descriptor did not perform well compared to the results obtained with shape context. The performance difference is significant with 45.9% of correctly localized parts for SIFT and 55.6% for shape context. Apart from the fact that shape context is based on edges, the main differences between the SIFT and shape context descriptors are the layout of spatial bins, size of the gradient discretization bins and the way image patches are normalized prior to descriptor computation. The SIFT descriptor uses a 4×4 uniform spatial grid with 8 bins for the gradient orientation, which results in a 128 dimensional descriptor; shape context uses 4 angular bins and 3 radial bins with 8 bins for the gradient orientation, which results in a 96 dimensional descriptor. However, while SIFT discretizes the full gradient orientation $[0, 2\pi]$, the shape context used in Andriluka et al. (2009) neglects the sign of the gradient vector by mapping it to $[0, \pi]$ prior to discretization, thus discretizing gradients twice as finely compared to SIFT. As the results in Tab. 4.1 show, fine gradient discretization is important for good performance. In particular, increasing the number of gradient orientation bins from 8 to 16 resulted in an improvement from 45.9% to 52.1%. Similar results are obtained by ignoring the gradient orientation in the original SIFT descriptor (performance increases to 52.3%). Interestingly, the shape context descriptor appears to be more robust to coarse gradient discretization. Doubling the number of gradient discretization bins in shape context improves the performance from 50.2% to 53.4%, which is significantly less than the improvement for the SIFT descriptor. Ignoring the sign of the gradient vector resulted in improvements for both shape context and SIFT, which is explained by the intuition that, in the context of human detection, we do not want to distinguish between light clothes on dark background and vice versa. However, the improvement is rather small (less than 1%) which suggests that invariance to gradient orientation can be successfully learned by the classifiers during training. The HOG descriptor in Johnson and Everingham (2009) performed very similar to the modified SIFT, which is understandable, since both of them are gradient-based descriptors with rectangular spatial binning, and both ignore the gradient sign.

In our comparison we observe that the purely edge-based shape context descriptor performs on par or better than the appearance-based SIFT and HOG descriptors.

Appearance Descriptor	Torso	U. leg	L. leg	U. arm	Forearm	Head	Total
edge features (Ramanan, 2006)	63.4	48.0	37.9	26.8	20.5	45.4	37.5
HOG (Johnson and Everingham, 2009)	73.2	58.5	52.2	47.8	**32.5**	62.4	51.8
original SIFT	68.8	53.6	49.0	37.5	26.9	55.6	45.9
SIFT, 8 bins, $[0, \pi]$	78.1	58.5	54.6	44.6	31.2	66.8	52.3
SIFT, 16 bins, $[0, \pi]$	76.1	56.4	52.0	44.6	29.8	69.8	51.1
SIFT, 16 bins, $[0, 2\pi]$	77.1	59.3	53.1	46.8	30.5	64.4	52.1
SC, 8 bins, $[0, \pi]$, (Andriluka et al., 2009)	**84.9**	**64.2**	54.4	**48.3**	31.0	**75.6**	**55.6**
SC, 8 bins, $[0, 2\pi]$	75.1	57.3	53.4	40.0	31.2	62.9	50.2
SC, 16 bins, $[0, \pi]$	82.4	62.5	**55.4**	47.5	32.5	71.2	54.9
SC, 16 bins, $[0, 2\pi]$	78.1	60.8	53.6	47.1	31.5	69.8	53.4

Table 4.1: **Evaluation of model parameters:** 2D pose estimation results on the **"People"** dataset for different image descriptors and discretizations of gradient orientation.

On one hand, this demonstrates that the normalization scheme employed in our shape context implementation is sufficiently robust to capture important image edges across a wide range of illumination conditions. On the other hand, it shows that SIFT- and HOG-based detectors fail to benefit from a richer image description, which is perhaps due to the fact that properties such as texture do not generalize well across object instances.

Note that all considered descriptors significantly outperform the edge based template features from (Ramanan, 2006), showing the importance of a strong shape/appearance representation. The best results were achieved with the shape context descriptor using 8 orientation bins and ignoring the sign of gradient vector. At the same time we have shown that comparable performance can be achieved with SIFT, if the gradient discretization is fine enough. The improvement obtained with shape context is most likely due to the log-polar spatial binning, which allows the descriptor to be more robust to small rotations of the body parts.

Number of boosting rounds. One of the parameters of the part detectors introduced in Sec. 4.2.2 is the number of boosting rounds used for training the AdaBoost classifiers. In Tab. 4.2 we show the influence of this parameter on the part localiza-

Number of rounds	Torso	U. leg	L. leg	U. arm	Forearm	Head	Total
8	72.2	53.0	46.1	38.0	19.8	55.6	44.2
16	74.2	58.5	47.8	41.0	23.1	60.5	47.6
32	78.1	59.2	48.8	43.6	28.3	68.3	50.6
63	82.4	63.2	54.1	44.9	30.8	71.2	54.0
125	80.5	**66.1**	53.6	46.3	31.5	71.2	54.7
250	81.5	64.7	54.8	47.3	**33.0**	73.2	55.4
500	**84.9**	64.2	54.4	**48.3**	31.0	**75.6**	55.6
750	**84.9**	63.6	**54.9**	**48.3**	32.9	72.2	**55.7**

Table 4.2: **Evaluation of model parameters:** 2D pose estimation results on the **"People"** dataset for varying numbers of boosting rounds.

tion performance. Interestingly, performance of the model trained for as few as 8 rounds is already reasonable. This due to the fact that each body part has several very characteristic features sufficient to separate it from the background. On the other hand, performance of the model increases with the number of boosting rounds. This means that the part detectors are able to learn more complex representations for each part and are not prone to overfitting.

Discretization step size between evaluated part configurations. In all experiments on the "People" dataset, the default value for the step between evaluated part locations was 8 pixels in the image position and 15° degrees for orientation. This discretization is rather coarse given that the average height of the human head in this dataset is 33 pixels and the width of the forearm is around 8 pixels. Tab. 4.3 shows the improvements due to a denser evaluation of part detectors both spatially and orientation-wise. The improvement in performance is significant: for example, the head was localized 75% of the time at the coarse and 79% at the finest discretization level. Interestingly, fine discretization is more important if inference is done with max-product than with sum-product belief propagation (see Sec. 4.3.3 for comparison and discussion).

4.3.2 Extension of the model with repulsive factor

Next we report a set of experiments comparing our model with and without the repulsive factors. We apply the two stage inference procedure as described in Sec. 4.2.4. In these experiments, the reduced state space is obtained by generating 1000 samples from the marginal posterior distribution of each part.

Results on the **TUD-UprightPeople** dataset Andriluka et al. (2009) are presented in Tab. 4.4 and for the "People" dataset Ramanan (2006) in Tab. 4.5. In each case, we are extending the model with repulsive factors between the corresponding

Discretization step	Torso	U. leg	L. leg	U. arm	Forearm	Head	Total
8 px., 15°	**84.9**	64.2	54.4	48.3	31.0	75.6	55.6
4 px., 15°	83.9	64.6	56.4	49.8	32.7	76.1	56.7
2 px., 15°	83.9	65.7	56.5	50.2	33.7	76.1	57.2
2 px., 7.5°	**84.9**	65.2	58.5	**51.7**	34.1	78.5	58.2
1 px., 7.5°	**84.9**	**66.3**	**58.5**	51.0	**35.4**	**79.0**	**58.6**

Table 4.3: **Evaluation of model parameters:** 2D pose estimation results on the **"People"** dataset for different step between evaluated part locations.

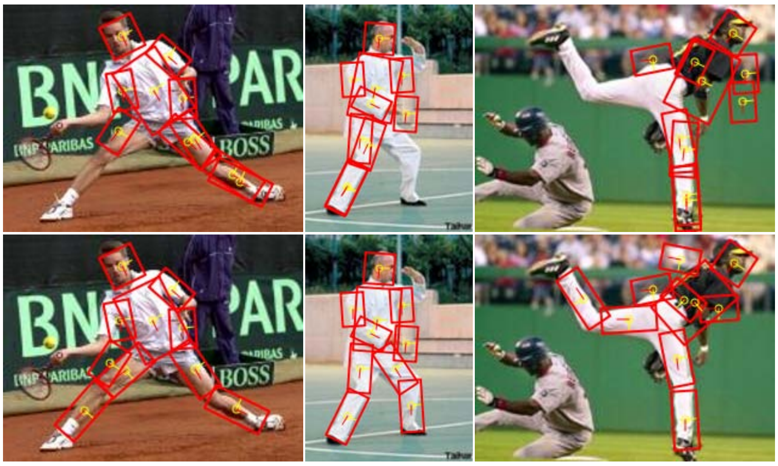


Figure 4.6: Example images from **"People"** dataset, with poses estimated by a tree model (top) and a non-tree model with repulsive factors (bottom).

left and right upper and lower legs. On both datasets the model with repulsive factors shows better performance than the baseline tree model (88.1% vs. 86% on "TUD-UprightPeople" and 60.1% vs. 58.6% on the "People" dataset). This improvement is significant, especially considering that in both cases the kinematic tree prior already encourages configurations with non-overlapping limbs (see Fig. 4.2 and Fig. 4.3). Fig. 4.6(top) shows a few example images where the true part configuration is either too far from the mean of the prior distribution or one of the part hypotheses is too weak, which results in the incorrect estimate. The repulsive factors allow to influence the trade-off between explaining strong hypotheses and staying close to the prior distribution, and lead to better pose estimation results as shown in Fig. 4.6(bottom).

Method	Foot	L. leg	U. leg	Torso	Head	Total
tree model	82.0	85.0	88.3	93.3	84.1	86.0
non-tree model with repulsive factor	**84.8**	**88.7**	**89.8**	**93.8**	**84.4**	**88.1**

Table 4.4: Comparison of models with and without repulsive factors on the **"TUD-UprightPeople"** dataset.

Method	Torso	U. leg	L. leg	U. arm	Forearm	Head	Total
tree model	**84.9**	66.3	58.5	**51.0**	**35.4**	79.0	58.6
non-tree model with repulsive factor	**84.9**	**69.2**	**63.0**	50.7	35.2	**79.5**	**60.1**

Table 4.5: Comparison of models with and without repulsive factors on the **"People"** dataset. Numbers indicate the percentage of the correctly detected parts. The total number of part segments is $10 \times 205 = 2050$. Note that in this experiment part detectors are evaluated at the resolution of 1 px. in image and 7.5° in the orientation domain.

4.3.3 Comparison of sum-product and max-product belief propagation

Several recently proposed 2D pose estimation approaches use sum-product belief propagation to estimate body part configurations (Andriluka et al., 2009; Ferrari et al., 2008; Ramanan, 2006). Note that part configurations estimated with the sum-product algorithm might contain body parts that are inconsistent with each other. Such behavior, however, is often not penalized and localization performance is measured on a per-part basis. In this section we quantify the differences in performance when we require that the model produces consistent part configurations. For that purpose we switch from sum-product to max-product belief propagation, while still measuring how often each of the parts is correctly localized.

Tab. 4.6 compares the max- and sum-product algorithms on the "People" dataset. Interestingly, the sum-product algorithm turns out to be more robust to the step size between evaluated part configurations and is able to find significantly more body parts than the max-product algorithm. For example, when part detectors are evaluated on the coarse grid the sum-product algorithm obtains 55.9% vs. 51.8% for the max-product algorithm. When the grid is sufficiently fine both algorithms show comparable performance (58.6% for sum-product vs. 58.7% for max-product). One explanation for this result is that max-product is looking for a globally consistent part configuration and in that case the whole configuration can have a low probability even when only one body part has a weak response in the relevant area of the search space. In the case of sum-product, such weak responses are less of a bottleneck since the algorithm does not enforce global consistency, and might still find partially

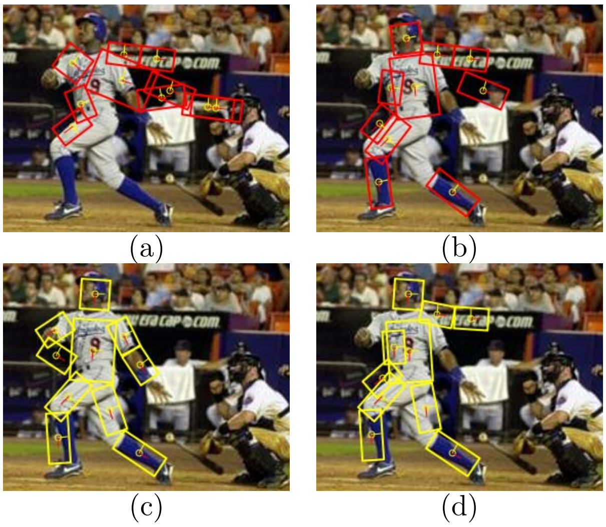


Figure 4.7: Results obtained with sum-product (b, d) and max-product (a, c) algorithms for coarse 8 px. (top row) and fine 1 px. (bottom row) grid of evaluated part locations.

Method	Torso	U. leg	L. leg	U. arm	Forearm	Head	Total
sum-product, 8 px., 15°	82.0	65.4	55.9	47.8	31.2	76.1	55.9
max-product, 8 px., 15°	79.0	60.7	50.2	42.0	31.5	69.8	51.8
sum-product, 1 px., 7.5°	**84.9**	66.3	58.5	51.0	**35.4**	**79.0**	58.6
max-product, 1 px., 7.5°	83.9	**67.8**	**59.2**	**52.0**	35.1	75.1	**58.7**

Table 4.6: Comparison of sum-product and max-product belief propagation on **“People”** dataset.

correct configurations.

Fig. 4.7 shows an example of part estimates obtained with the max- and sum-product algorithms for coarse and fine discretizations. Note, how for the coarse discretization sum-product is able to estimate a few body parts correctly (subplot (b)), while max-product produces a consistent but incorrect configuration (subplot (a)). For a finer discretization estimates obtained with both algorithms are similar with max-product localizing several more parts correctly than sum-product (subplots (c,d)).

4.3.4 Performance for varying training and test sets

So far, all of the results reported in this chapter were obtained for the standard split of the “People” dataset into training and test sets, in which 100 images are allocated for training and 205 for testing.

In this section we quantify the changes in performance depending on the particular selection of the training and test data. To that end we perform a series of experiments, in which we first independently vary the training and test sets, and finally evaluate performance for different splits of the "People" dataset. The performance is reported for the tree-structured version of our approach with parameters set as described in the beginning of Sec. 4.3.

In the first experiment we use the standard test set from the "People" dataset and generate training sets by randomly selecting 100 images from the **Leeds Sports Poses** (LSP) dataset (Johnson and Everingham, 2010). The images in the "LSP" dataset are qualitatively similar to the "People" dataset, but exhibit somewhat larger variability in poses of people and imaging conditions. The interesting outcome from this experiment is that the pose estimation performance varies significantly with the particular training set. The average PCP score over 10 runs is 51.8% with the standard deviation of 1.4%. Depending on the training set the result vary between 49.7% and 54%.

In the second experiment we fix the training set to the one that resulted in the best performance in the first experiment and vary the test set. Each test set is generated by randomly choosing 205 images from the "People" dataset. The average over 10 runs is 53% with the standard deviation of 0.9%. The results vary between 51.8% and 54.1%.

In both experiments the numbers appear somewhat lower than those for the standard split of the "People" dataset. In order to clarify that, we have randomly generated 5 splits of the "People" dataset into training and test data. The results varied between 50.1% and 56.1%, with the average of 52.9% and the standard deviation of 2%. This result suggest that the standard split of the "People" dataset is somewhat easier than average.

The overall conclusion from these experiments is that while being visually similar the particular choice of test and training data can still result in significant difference in performance (up to 5% PCP in our case). This should be kept in mind when indirectly comparing the performance of different approaches. The variability in performance could be due to relatively small size of the training and test sets, which suggests that future work should focus on datasets with larger number of training and test images.

4.3.5 Run-time performance

The run-time of our approach is mainly influenced by the discretization of the part configuration space and the number of body parts. For a 10-part body model and a discretization step size of 8 pixels our implementation requires approximately 0.5 seconds (Intel Xeon 2.4 GHz, C++ implementation) to compute the part likelihoods for a single orientation on a typical image from the "People" dataset (167×251 pixels). When discretizing part orientations with a 15° step size, this translates to about $12-20$ seconds for the entire likelihood computation. Passing a single message between two body parts during inference requires approx. 2 seconds. Overall, our approach takes about 50 seconds to compute the max-marginal estimate for all body parts. Note that these times can likely be significantly improved by parallelizing the computation of Gaussian convolution and image features, which we leave for future work.

Approximate inference in case of repulsive factors (c.f. Sec. 4.2.4) is performed on a reduced state-space and does not require additional likelihood computations. Full BP inference including sampling from the marginal posteriors of the tree model takes from 5 to 10 seconds in addition.

4.4 Comparison to State of the Art

In this section we evaluate our model on three related tasks of increasing complexity: pedestrian detection, upper body pose estimation, and multi-view full body pose estimation. For each task we use publicly available datasets and directly compare to the respective methods designed to work specifically on each of the tasks. The main goal in these experiments is to demonstrate the generic applicability of our model and to quantify the performance gains over results previously published in the literature.

Since in our work we do not focus on identifying the optimal model parts, we rely on a canonical set of parts corresponding to the body limbs, torso and head. The number of parts in each experiment is imposed by the task at hand, that is for upper-body pose estimation where the task is to localize the upper-body, head and upper/lower arms, we rely on a model with these six parts. For full-body pose estimation we use a model that also includes legs. In case of pedestrian detection the upper and lower arms were not annotated in the training set and are therefore not included in the model.

In this section we choose model components and parameters, which we found to perform best in the experiments in Sec. 4.3. In particular we use shape context descriptors that are densely computed over the image and perform inference with sum-product belief propagation. We also augment our model with repulsive factors when applying it to the full body pose estimation task.

4.4.1 Pedestrian detection

To detect pedestrians, we compute the marginal distribution of the torso location and use its modes to predict the pedestrians' bounding boxes. To deal with multiple peaks in the posterior corresponding to the same detection hypothesis, we perform a non-maximum suppression step. For each detection we remove all detections with smaller probability and more than 50% cover and overlap.

In this section we use three publicly available datasets to evaluate our model.

The **TUD-Pedestrians**[3] dataset Andriluka et al. (2008) contains 250 images

---

[3] www.d2.mpi-inf.mpg.de/datasets

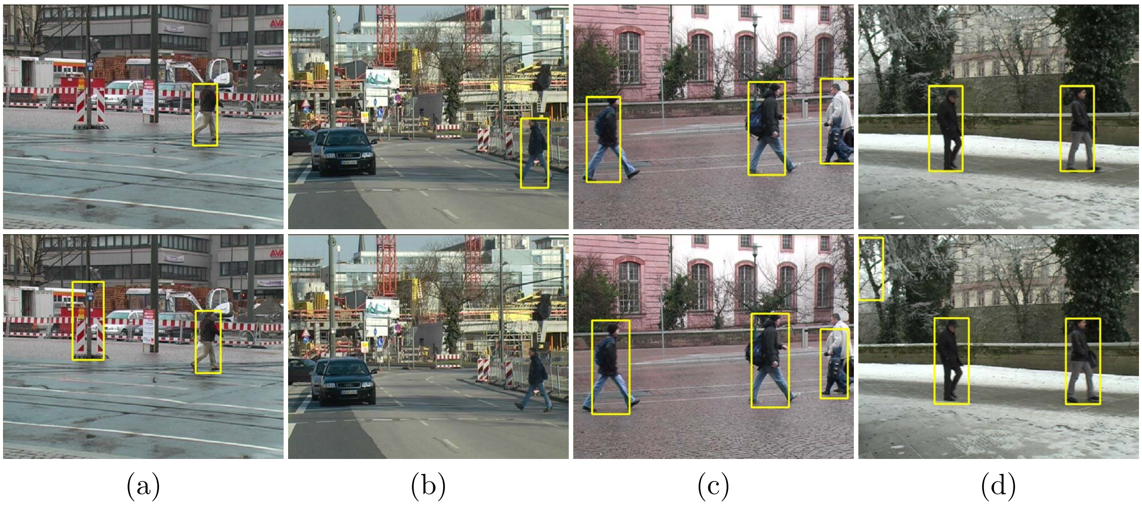


Figure 4.8: Several examples of detection at equal error rate obtained with our model (8 parts and tree prior, top) and partISM (bottom) on the **"TUD-Pedestrians"** dataset.

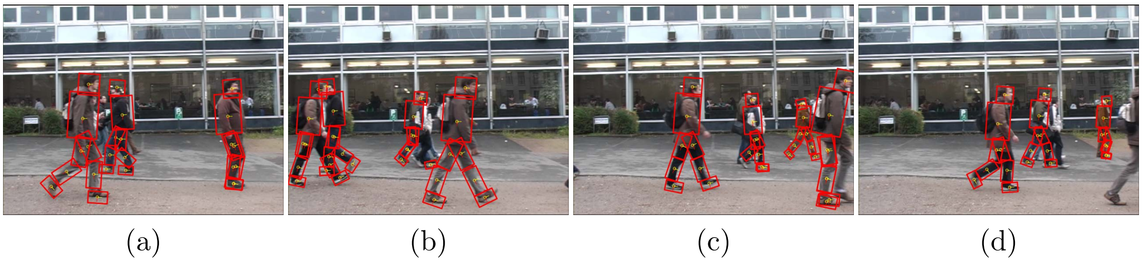
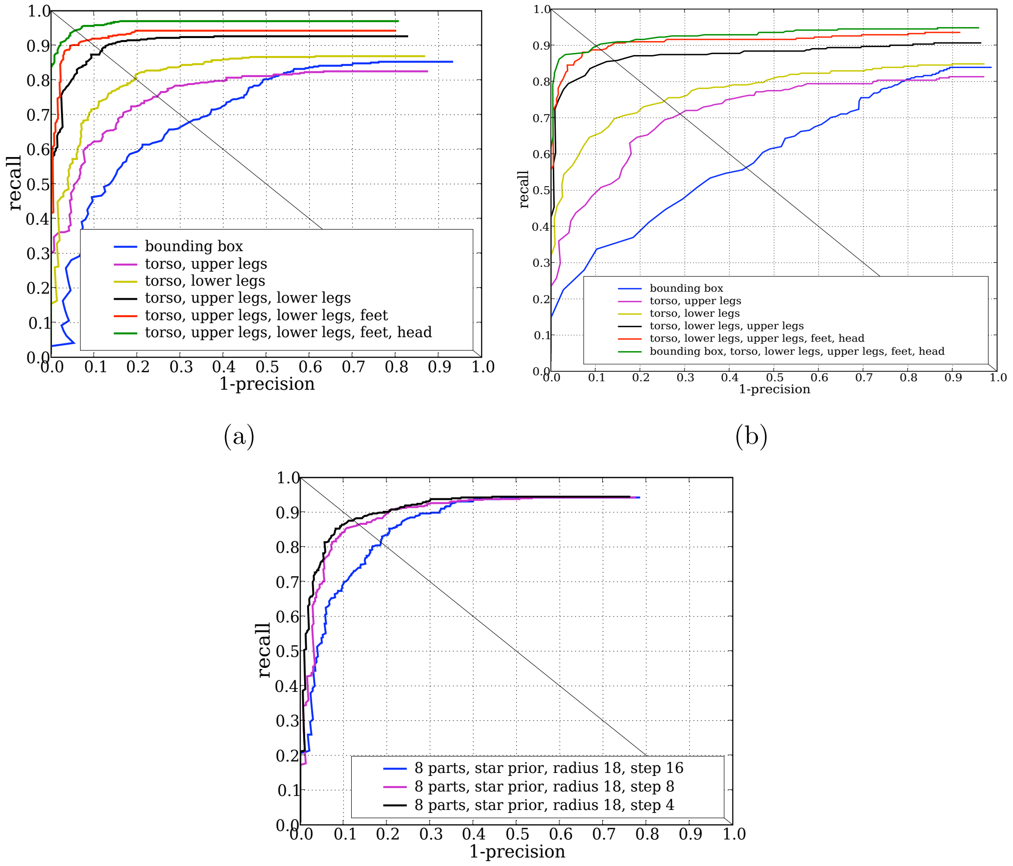


Figure 4.9: Several examples of detection and pose estimation obtained with our model (8 parts and tree prior) on the **"TUD-Campus"** sequence.

with 311 pedestrians (mostly side-views) with large variability in clothing and articulation. The corresponding training set contains 400 images. Additionally, we created a new dataset called **TUD-UprightPeople**[3], which contains images from "TUD-Pedestrians" and additional images taken under various illumination conditions, for which we also annotated body part configurations. This new dataset is used to evaluate different aspects of our model including part localization performance. The dataset contains 435 images with one person per image. The image sequence **TUD-Campus**[3] introduced in Andriluka et al. (2008) is used to evaluate the performance of our method on images with multiple people. We compare our approach to previous work on all three datasets.

In all pedestrian detection experiments we use the 400 training images provided with "TUD-Pedestrians" to train the part detectors. We used two different priors on the part configuration: (1) a star prior in which all parts are connected directly to the center part; and (2) a kinematic tree prior (both are shown in Fig. 4.3).

First, we compare our approach to results from the literature on the "TUD-Pedestrians" dataset as shown in Fig. 4.11(b). We use 8 parts either with a star



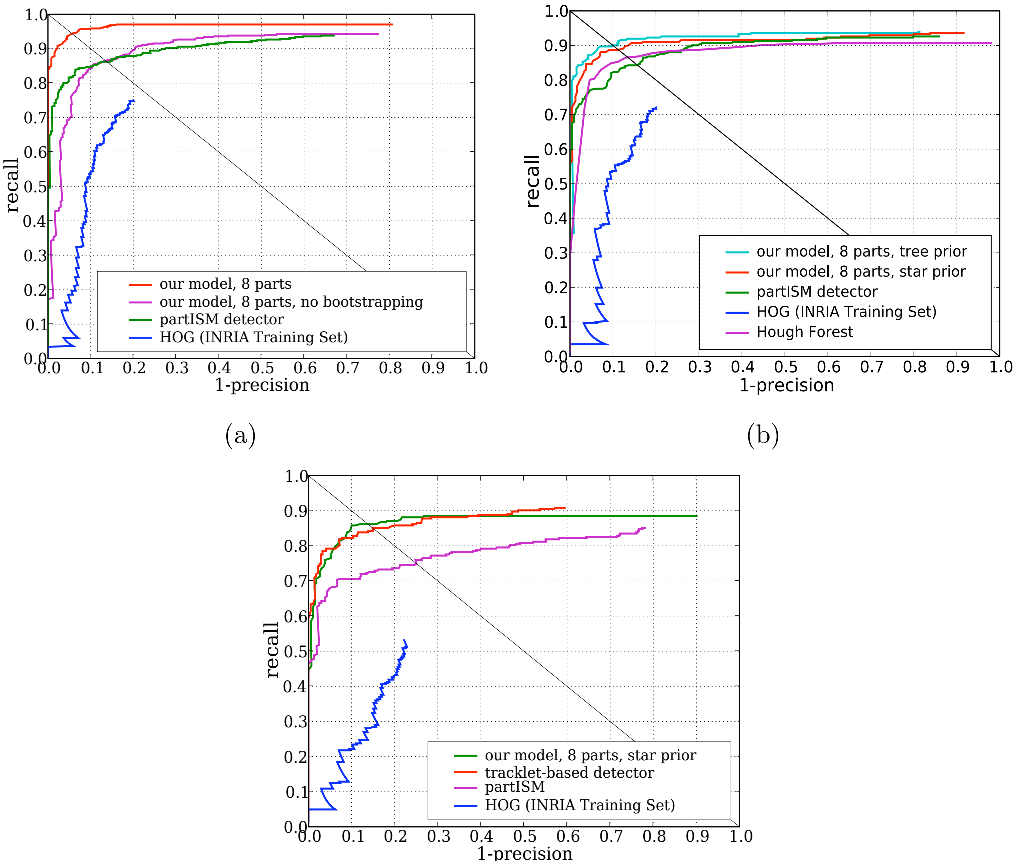
(a) (b)

(c)

Figure 4.10: **Pedestrian detection results:** Performance with varying number of parts on (a) **TUD-UprightPeople** and (b) **TUD-Pedestrians**. (c) Comparison of part detectors with varying step sizes between local descriptors on **TUD-UprightPeople**.

prior or with a kinematic tree prior, whose parameters were estimated on the training images of "TUD-Pedestrians".

While the tree-based prior slightly outperforms the star prior, both outperform the partISM-model (Andriluka et al., 2008) by 4% and 5% equal error rate (EER) respectively, and slightly improves over recent work of Gall and Lempitsky (2009). All three approaches significantly outperform the publicly available HOG binary (Dalal and Triggs, 2005), which compared to our approach needs less supervision as it does not require part annotations during training. Similarly, the same performance ordering can be observed in Fig. 4.11(a) for the "TUD-UprightPeople" dataset. We attribute the improved performance over partISM to our dense part representation and to the discriminatively learned appearance model. PartISM, in contrast, uses a generative appearance model based on sparse interest points. Fig. 4.8 shows example detections of our model and partISM. Our model is flexible enough to capture diverse articulations of people as for example in Fig. 4.8(b,c,d), which typically



(a)

(b)

(c)

Figure 4.11: **Pedestrian detection results:** Comparison with previously proposed approaches partISM (Andriluka et al., 2008), HOG (Dalal and Triggs, 2005) on (a) **TUD-UprightPeople**, (b) **TUD-Pedestrians**, and (c) **TUD-Campus**. On the **TUD-Pedestrians** (b) we also compare with the "Hough Forests" approach of Gall and Lempitsky (2009).

are problematic for monolithic detectors such as HOG. However, since our model is built on top of discriminative classifiers, it can avoid false positives in cluttered backgrounds, which plague the generative partISM model (e.g., Fig. 4.8(a) and (d)).

To gain more insight into the role of the different components, we conducted a series of experiments on the "TUD-UprightPeople" dataset. Here, we report results for the star prior only, as it allows to arbitrarily add and remove body parts from the model. Fig. 4.10 shows the influence of varying numbers of parts on the detection performance for "TUD-UprightPeople" and "TUD-Pedestrians". As expected, using more body parts generally improves detection performance. We also compare to a monolithic detector that consists of a single part defined by the person's bounding box (as is typical, e.g., for the HOG detector (Dalal and Triggs, 2005)). This monolithic detector did not perform well even compared to detectors with as few as 3 parts.

In Fig. 4.10(c) we evaluate how the density of local feature sampling affects performance. The distance between evaluated part configurations is kept constant, but the distance between features included in the feature vector presented to the AdaBoost classifier is varied from 4, 8 to 16 pixels. The denser version consistently obtains better results and outperforms the sparser version by about 6% EER. This confirms our intuition that an over-complete encoding of the information at different spatial resolutions is important for classification. Note that the overall performance difference between Fig. 4.10(a) and (c) is due to not performing bootstrapping on the part detectors in Fig. 4.10(c). Also, note that denser sampling of local features (Fig. 4.10(c)) and bootstrapping (Fig. 4.11(a)) result in improvements of 6% and 8% EER respectively.

Both "TUD-Pedestrians" and "TUD-UprightPeople" contain only one person per image. However, it is also interesting to analyze the performance on images with multiple people frequently occluding each other. For that purpose we use the publicly available image sequence "TUD-Campus" (Andriluka et al., 2008). Fig. 4.11(c) shows the comparison of our model both with the tracklet-based and the partISM detectors from (Andriluka et al., 2008). Our detector significantly outperforms partISM and even slightly improves over the tracklet based detector, while operating on single frames only. This result is remarkable since the tracklet-based detector requires associations of hypotheses across frames and assumes that people in the scene are walking. Our model avoids these assumptions without sacrificing performance. Fig. 4.9 shows a few detections and the estimated body poses obtained with our approach on "TUD-Campus". Note that since the evidence in our model is aggregated locally it can still cope with partial occlusion as in Fig. 4.9(b) and is robust to clutter and overlaps between people as in Fig. 4.9(a,c,d). At the same time, as our model represents a person as a flexible configuration of parts, it can successfully cope with large variations in articulation present in the dataset.

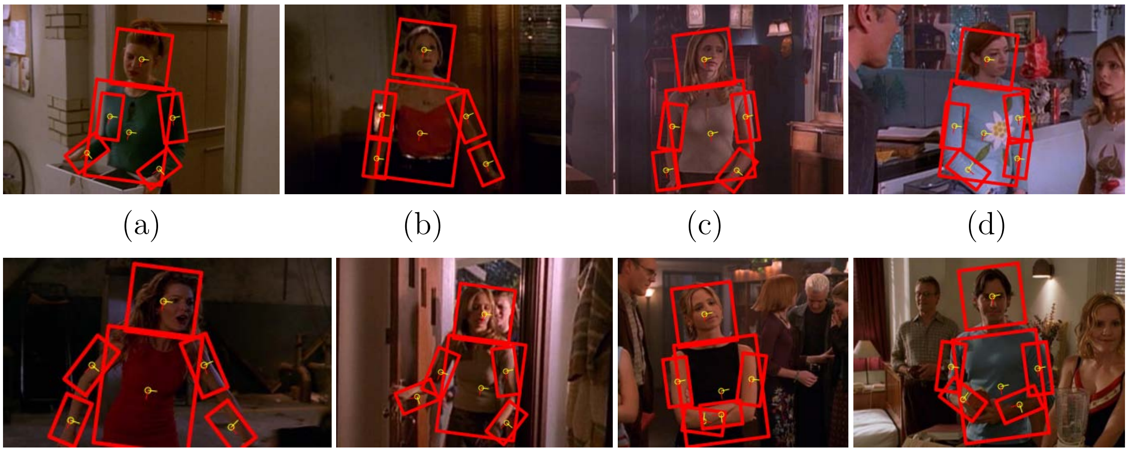
4.4.2 Upper body pose estimation

In order to evaluate our method on the task of upper-body pose estimation, we use two recently proposed and publicly available datasets: **Buffy Stickmen** (Ferrari et al., 2008)[4] and **ETHZ PASCAL Stickmen** (Eichner and Ferrari, 2009)[5]. "Buffy Stickmen" consists of frames extracted from 5 episodes of the popular TV show "Buffy the Vampire Slayer", with episodes 2, 5, and 6 containing 276 annotated people used for testing. The "ETHZ PASCAL Stickmen" dataset is a subset of the PASCAL VOC 2008 (Everingham et al., n.d.) dataset containing 549 annotated people in front/back view. Note that we are using the latest revision of the "Buffy Stickmen" dataset, released after the publication of Eichner and Ferrari (2009), which includes a slightly different set of images than the previous release and also contains the Matlab code for evaluation of localization performance.

---

[4] `www.robots.ox.ac.uk/~vgg/data/stickmen`

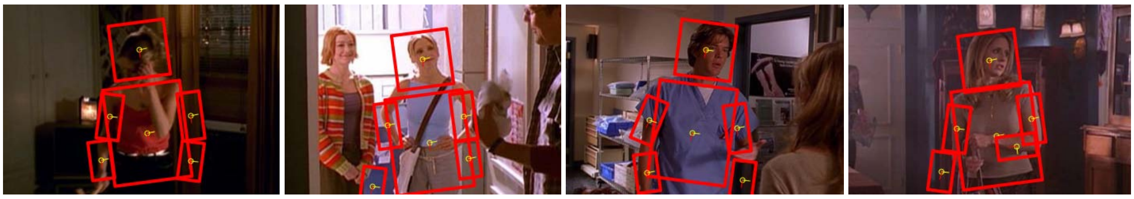
[5] `www.vision.ee.ethz.ch/~vferrari/datasets.html`



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

(e) (f) (g) (h)

Figure 4.12: **Upper body pose estimation:** Examples of pose estimation results of our method on the **"Buffy Stickmen"** dataset (see text for details). Note that color information is not used for detection.



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

Figure 4.13: **Upper body pose estimation:** Several examples of the typical failure cases of our method on the **"Buffy Stickmen"** dataset.

For both datasets the objective is to estimate configurations of torso, head, left and right forearm, and left and right upper arm. This is challenging due to the large variability of poses, varying, and loose fitting clothing, as well as strongly varying illumination. For the "ETHZ PASCAL Stickmen" dataset the task is further complicated by the varying quality of images, which are originating from diverse sources including community photo collections.

Due to the complexity of the task, the previously proposed approaches (Eichner and Ferrari, 2009; Ferrari et al., 2008, 2009*a*,*b*) use multiple stages to reduce the search space of admissible poses. In particular Ferrari et al. (2008, 2009*a*) propose to perform an additional automatic foreground/background separation step based on 'GrabCut' (Rother et al., 2004). This approach is further extend in (Eichner and Ferrari, 2009; Ferrari et al., 2009*b*) by introducing additional a-priori assumptions and learning dependencies between color distributions of body parts. In our approach we directly estimate the pose from images without any additional search space pruning. As pointed out before, one may think of the discriminative part models we use as a form of pruning.

Method	Torso	U. arm	Forearm	Head	Total
Ferrari et al. (2008)	–	–	–	–	61.4
Ferrari et al. (2009*b*)	–	–	–	–	74.5
Eichner and Ferrari (2009)	**98.7**	82.8	**59.8**	**97.9**	80.3
Our part detectors (generic)	28.1	11.1	6.2	56.2	19.8
Our part detectors (front/back)	56.2	20.9	8.6	67.7	30.4
Our model (generic)	97.5	83.8	55.8	96.2	78.8
Our model (front/back)	97.5	**92.7**	59.6	95.7	**83.1**

Table 4.7: **Upper body pose estimation:** Comparison of body part detection on the **"Buffy Stickmen"** dataset (numbers indicate the percentage of correctly localized body parts). Following the evaluation procedure only correctly localized people are taken into consideration. All experiments use the detector from (Ferrari et al., 2008), which can localize 85.1% of the annotated people.

Method	Torso	U. arm	Forearm	Head	Det. Rate	Total
Detector of Ferrari et al. (2008) and our model	97.5	**81.7**	**51.7**	96.6	**85.1**	**76.8**
Our model only	**99.5**	79.0	48.8	**99.9**	78.6	75.7

Table 4.8: **Upper body pose estimation:** Comparison of body part detection rates on the **"Buffy Stickmen"** dataset using the upper-body detector of Ferrari et al. (2008) for localization and our model only.

Method	Torso	U. arm	Forearm	Head	Total
Eichner and Ferrari (2009)	**97.2**	73.8	41.5	**88.1**	69.3
Our model (front/back)	96.4	**77.8**	**47.0**	85.0	**71.8**

Table 4.9: **Upper body pose estimation:** Comparison of body part detection on the **"ETHZ PASCAL Stickmen"** dataset (Eichner and Ferrari, 2009) (numbers indicate the percentage of correctly localized parts). According to the evaluation procedure only correctly localized people are taken into consideration. All experiments use the detector of (Ferrari et al., 2008), which can localize 65.6% of annotated people.

The authors of (Eichner and Ferrari, 2009; Ferrari et al., 2008, 2009*a*,*b*) report quantitative pose estimation results for the single-frame detector (no temporal coherency) only on a subset of people that have been correctly localized with a weak object detector used for prefiltering. This weak object detector is a HOG based upper body detector whose detections are made publicly available by the authors along with the “Buffy Stickmen” and “ETHZ PASCAL Stickmen” datasets. In order to facilitate comparison with previous work, our evaluation procedure on both datasets is to start with the provided upper-body detections and to only estimate the upper body pose in their neighborhood. Similarly to the authors of (Eichner and Ferrari, 2009; Ferrari et al., 2008, 2009*a*,*b*) we also assume that all people in the dataset have an upright pose and disregard all detections of head and upper body at other orientations.

We consider two versions of our model: the same generic model as used for full-body pose estimation in Sec. 4.4.3 for which we simply disregard lower-body parts, and a specialized model (front/back model) that is trained on the subset of images from episodes 3 and 4 of the Buffy dataset, which are not used for evaluation. Note that this front/back model only differs in terms of which training data was used; no other modifications were made. For the reference, we also report the performance of the part detectors of both models.

Table 4.7 shows localization results for all 6 upper body parts, along with the overall localization performance using the PCP measure, which was computed with the official evaluation code distributed along with the “Buffy Stickmen” dataset. Our method significantly outperforms the approach of (Ferrari et al., 2008) (78.8 vs. 61.4 PCP) and performs slightly better than (Ferrari et al., 2009*b*), even when using a fully generic model. Yet in contrast to (Ferrari et al., 2008), we do not require separate foreground/background segmentation or use color features. The best result is obtained with our model trained on front/back poses, which improves the performance by 5 PCP compared to the fully generic model (83.1 vs. 78.8 PCP).

It is also noteworthy that the part detectors alone, while powerful, do not perform well on this dataset, especially those of the arms. This is the case even if part detectors are trained specifically on front/back views, although in this case the performance improves significantly (19.8 PCP for generic detectors vs. 30.4 PCP for front/back detectors). On one hand, this illustrates the difficulty of the dataset, and on the other hand it demonstrates the importance of capturing the spatial relations between body parts, which for some of the body parts can improve localization by a factor of ten.

Tab. 4.8 compares part localization performance of our model with and without using the HOG based upper body detector of (Ferrari et al., 2008). In this experiment we are using our model with a generic prior and generic part detectors. In order to handle the case of multiple people being present in the image we proceed by first computing the marginal distribution of the torso and subsequently estimate the part configuration for each of its modes. The number of potential torso hypotheses is reduced by performing non-maximum suppression. The results show that our model can be successfully used in standalone mode, although the combination with the HOG detector leads to a slight improvement in part localization (76.8% for the combined approach vs. 75.7% for our model alone). The HOG detector manages to find the upper body in 85.1% of the frames, compared to 78.6% for our model. This

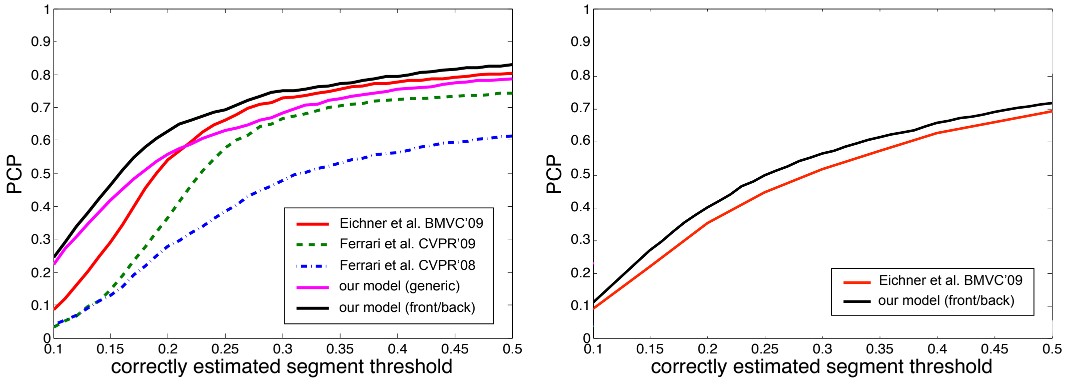


Figure 4.14: Comparison of part localization performance on **"Buffy Stickmen"** (top) and **"ETHZ PASCAL Stickmen"** (bottom) datasets.

is due to the fact that our model was trained on a more general training set. The main advantage of the combination with the HOG detector is certainly efficiency.

The official criteria for measuring PCP require the correct part localization threshold to be set to 0.5, which is rather forgiving to part localization errors. As has been recently proposed Eichner and Ferrari (2009), a better understanding of the performance of different methods is obtained by evaluating PCP for various localization thresholds and combining these results in a PCP-threshold plot. Fig. 4.14 shows the performance of different versions of our method, as well the PCP-threshold plots made publicly available by the authors of (Eichner and Ferrari, 2009; Ferrari et al., 2009*b*, 2008). Similarly to the numbers presented in Tab. 4.7, the curves are obtained using the official evaluation routine supplied with the "Buffy Stickmen" dataset. Note that while all considered methods except (Ferrari et al., 2008) perform comparably at the standard threshold of 0.5, our approach turns out to be more precise at pinpointing the exact configuration of body parts, which results in a much higher PCP for smaller thresholds.

Tab. 4.9 and Fig. 4.14 show the performance of our approach on the "ETHZ PASCAL Stickmen" dataset. The presented results are obtained with our model in which both prior and part detectors were trained on images from episodes 3 and 4 of "Buffy Stickmen". The results on the "ETHZ PASCAL Stickmen" and "Buffy Stickmen" datasets are consistent, with our model showing a slight improvement over the best published results of Eichner and Ferrari (2009). Interestingly, compared to this approach our model is able to better localize especially those body parts that are also hardest to find. For example on "ETHZ PASCAL Stickmen" we are able to localize the right forearm in 48% of the cases, which is in contrast to 41.5% for the approach of Eichner and Ferrari (2009). Note, that the differences in PCP between Tab. 4.9 and results published in (Eichner and Ferrari, 2009) are due to changes in the official evaluation procedure made after the publication of (Eichner and Ferrari, 2009).

Fig. 4.12 shows examples of estimated upper-body configurations, which demonstrate the effectiveness of our method even for difficult poses including self-occlusions (e.g., Fig. 4.12 (g, h)). On the Fig. 4.13 we also show some typical failure cases that are often due to incorrect scale estimation by the HOG upper body detector provided along with the dataset (c.f. Fig. 4.12 (b)) and failures of the part detector (Fig. 4.12 (a)). Since we assume a constant size of the object parts, our method is limited in how much foreshortening can be tolerated (Fig. 4.12 (c, d)).

4.4.3 Full body pose estimation

Finally, we evaluate our model on a full body pose estimation task and compare to the iterative image parsing (IIP) method of Ramanan (2006) and to the approach of Johnson and Everingham (2009). Both methods use a spatial model similar to ours, but approach appearance modeling somewhat differently. Note that IIP is the same algorithm that is used at the pose estimation stage in (Eichner and Ferrari, 2009; Ferrari et al., 2008, 2009*b*). In this comparison we use the publicly available multi-view and multi-articulation **People** dataset from (Ramanan, 2006), which contains people engaged in a wide variety of activities ranging from simple standing and walking to dancing and performing acrobatic exercises. The difficulty of the task is further increased by the limited training set of only 100 images, which only scarcely capture the variations in appearance and poses present in the test set. We evaluate the part localization performance using the same body part localization criteria as proposed in (Ferrari et al., 2008) and used in Sec. 4.4.2. The iterative image parsing results are obtained using the implementation of Ramanan (2006). Quantitative results are shown in Tab. 4.10.

Our findings show significant performance gains: The localization results of our full model surpass those of Ramanan (2006) by more than a factor of 2 (60.1% vs. 27.2% accuracy). The localization performance of all body parts is significantly improved, up to a factor of 3. Note that there is a large variability in localization performance across body parts. The part that is localized best is the torso (84.9 PCP), the most difficult parts are the forearms (35.2 PCP). The low performance on forearms is perhaps due to their small image extent and frequent occlusion of these body parts, but also due to the complex spatial distribution of these parts that is not captured well by the Gaussian part relationships.

It is interesting to note that our head detector alone (i.e., without any kinematic model) has a better localization performance for the head than the full model of Ramanan (2006). This clearly demonstrates the importance of powerful part detectors for obtaining good overall performance. The improvements in performance of our full model over the results published in Andriluka et al. (2009) are considerable, in particular for the lower leg parts, where the localization performance improved by more than 11%. On one hand, this is due to a finer discretization step for the part configurations; on the other hand, the model benefits from repulsive factors as described in Sec. 4.3.2.

Tab. 4.10 also shows the performance of our method when the boosted part detectors are replaced with discriminatively trained edge templates used in (Ramanan,

Method	Torso	U. leg	L. leg	U. arm	Forearm	Head	Total
IIP (Ramanan, 2006), 1st parse (edge features only)	39.5	20.7	20.7	12.6	11.6	21.4	19.2
IIP (Ramanan, 2006), 2nd parse (edge + color feat.)	52.1	30.9	29.0	17.5	13.6	37.5	27.2
HOG+Segmentation (Johnson and Everingham, 2009)	77.6	61.5	54.9	**53.2**	**39.2**	68.8	56.4
Our part detectors only	29.7	12.3	18.5	3.6	4.3	40.9	14.8
Our model, edge features from (Ramanan, 2006)	63.4	48.0	37.8	26.8	20.4	45.3	37.5
Andriluka et al. CVPR'09 (Andriluka et al., 2009)	81.4	63.1	55.1	47.5	31.6	75.6	55.2
Our model (1 px., 7.5°, repulsive fact.)	**84.9**	**69.8**	**62.9**	50.2	35.2	**79.5**	**60.1**

Table 4.10: **Full body pose estimation:** Comparison of body part detection rates and evaluation of different components of the model on the **"People"** dataset. Numbers indicate the percentage of correctly localized parts. The total number of part segments is 10 × 205 = 2050.

2006). For this experiment we extracted the responses of the part templates using the author's code (Ramanan, 2006) and fitted sigmoid functions to the foreground and background responses of each part template in order to make them comparable with one another. The average part localization rate in this experiment is 37.5%[6], which is significantly better than the results of iterative image parsing (IIP) (Ramanan, 2006), even though the same edge template features are used. We attribute this to the different representation of the part relationships in our kinematic model. The performance of the full model is still significantly better (60.1%), which again shows that the boosted part detectors contribute substantially to the overall performance.

Fig. 4.15 shows examples of estimated poses from the "People" dataset. For each image we also show the marginal posterior distributions for each part and give the number of correctly localized body parts. Note that the posteriors of our method often have strong peaks at the correct part configuration, while the posteriors of IIP

---

[6] Results without sigmoid fitting are considerably worse.

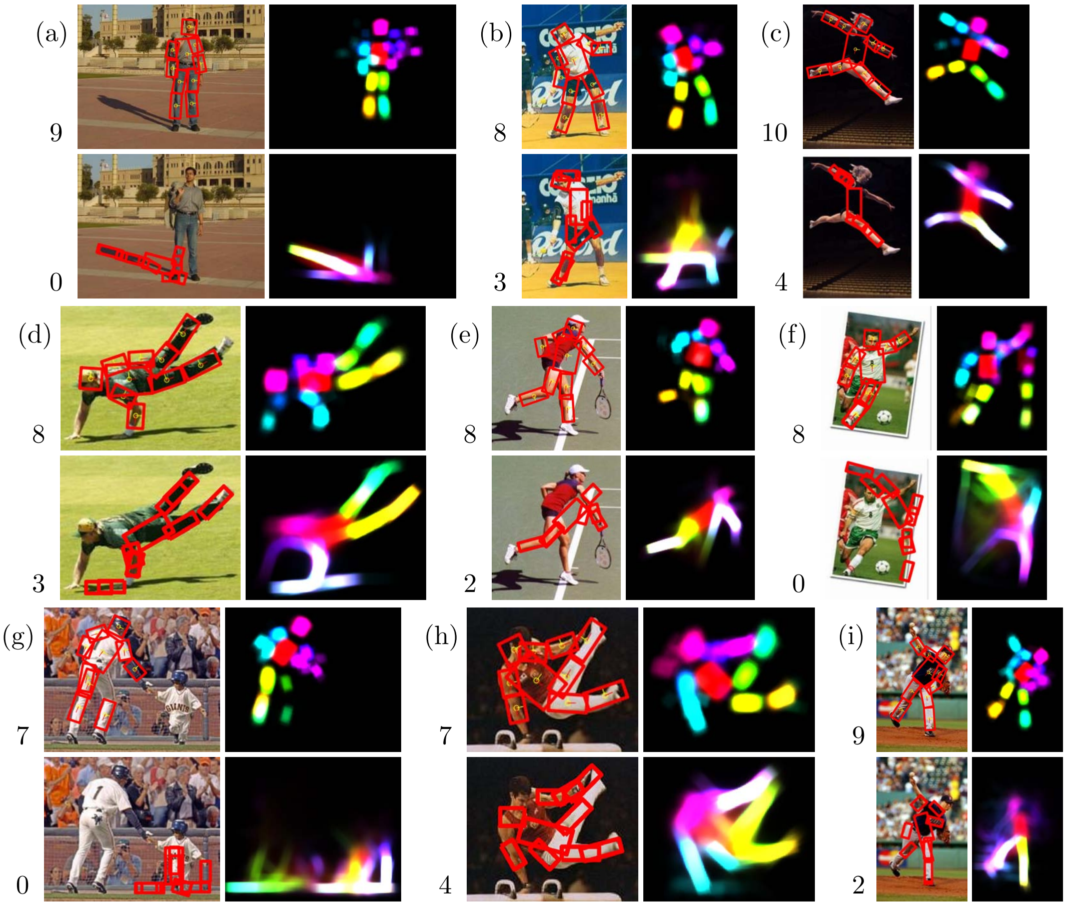


Figure 4.15: Comparison of full body pose estimation results between our approach (top) and (Ramanan, 2006) (bottom) on the **"People"** dataset. The numbers on the left of each image indicate how many of the 10 body parts were correctly localized.

appear to be flat along the limbs. It appears that the IIP method works well in relatively uncluttered images (e.g., Fig. 4.15 (c, h)); nonetheless, even in those scenes, we often localize more parts correctly. In strongly cluttered scenes (e.g., Fig. 4.15 (f, g)), our method seems to have a clear advantage in recovering the pose. This may be attributed to the fact that for such images the image parsing approach has difficulties building up appropriate appearance models during the initial parse. The line templates used in the initial parse may also be misguided by strong edges (Fig. 4.15 (a)), which our method handles more gracefully. On the Fig. 4.16 we show example of images for which both methods failed to estimate the poses correctly (i.e., less then 7 body parts where found). These failures are often due to complex poses that are not captured well by the Gaussian prior on part configurations (e.g., Fig. 4.16 (a, c, d, f)) and occlusion of body parts (e.g., Fig. 4.16 (c)). Also note that even in cases when several body parts were not correctly localized, the posterior distribution often still has a mode at the correct part location (e.g., Fig. 4.16 (b,e)).

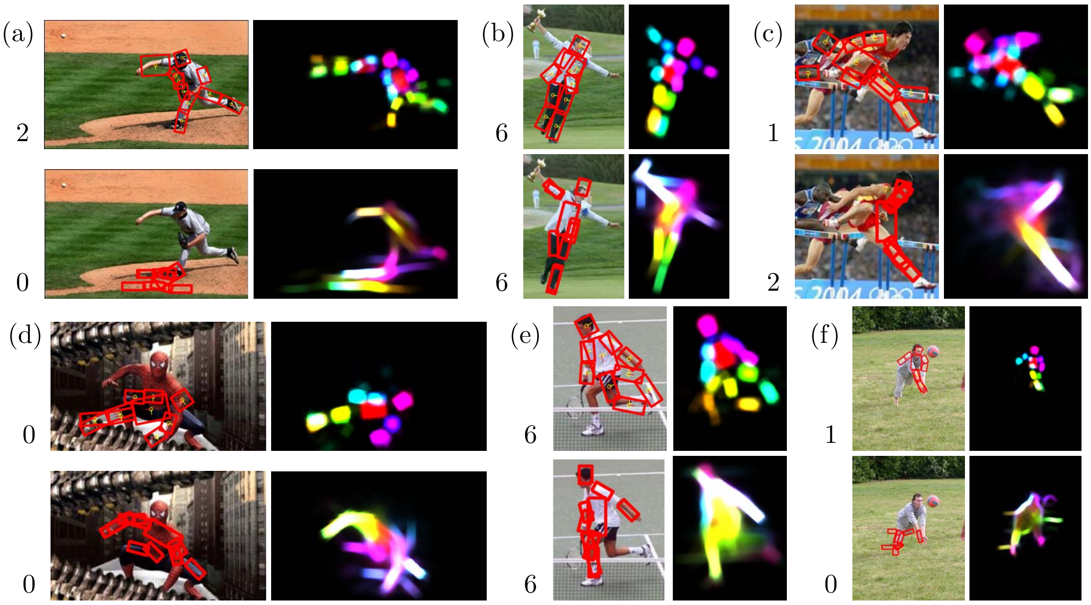


Figure 4.16: Several examples of difficult images from **"People"** dataset, on which neither our (top) nor the approach from (Ramanan, 2006) (bottom) was able to estimate more then 6 body parts correctly. The numbers on the left of each image indicate how many of the 10 body parts were correctly localized.

4.5 Conclusion

In this chapter we proposed a generic model for people detection and articulated pose estimation. We demonstrated the generality of our approach on several recently proposed datasets, where it consistently outperformed other approaches, which are often designed specifically for one of these tasks only. Despite that, our model is surprisingly simple. We attribute these excellent results to a powerful combination of two components: A strong discriminatively trained appearance model and a flexible kinematic tree prior on the configurations of body parts. In order to facilitate future comparison with our model we make the source code of our implementation available on our website[7].

Future work may consider our approach in a variety of ways. Currently, we do not make use of color information and do not model relationships between body parts beyond kinematic and repulsive constraints. We expect that the incorporation of additional constraints between body parts, extension of our approach with occlusion reasoning (c.f. (Sigal and Black, 2006*b*)), and modelling of body part foreshortening will further improve the performance.

The approach proposed in this chapter recovers 2D body configurations of people in the image coordinates, without reasoning about the position and orientation of parts in 3D, and relies on 2D body model. Such 2D estimates are often seen as prerequisite for recovery of 3D human poses and body shape (Guan et al., 2009; Lee

---

[7] `http://www.d2.mpi-inf.mpg.de/code`

and Chen, 1985; Taylor, 2000). These algorithms typically require very accurate estimation of 2D body poses, which are currently possible with our approach only for images with moderate amounts of clutter, occlusion and foreshortening. Furthermore, our approach does not make distinction between left and right limbs of the body, which are mapped to the same model parts.

Nonetheless, even such, at times ambiguous 2D estimates, contain much information about the underlying 3D body pose. As we show in Chapter 5, we can use them to define likelihood of 3D poses that when combined with dynamical priors on people motions allows us to estimate poses of people in 3D.

5

Monocular 3D Pose Estimation and Tracking by Detection

> *You will notice that in three dimensions the situation can be much more complicated than in two.*
>
> *Richard Feynman*, from *Six Not-So-Easy Pieces*

In Chapter 3 we introduced an approach to detection, tracking and pose estimation of people in crowded street scenes. In that approach, we relied on a part-based people detector in order obtain initial tracking hypotheses consistent over several subsequent frames. In the second step these hypotheses, denoted as *tracklets*, were refined using the learned model of body dynamics. Finally tracking over longer durations was performed by finding sequences of the refined hypotheses consistent in motion, appearance and people poses. In our approach, individual tracklets served as *data association hypotheses* linking detections in adjacent frames as potentially belonging to the same person. For each tracklet we could rely on evidence from multiple frames in order to to obtain more details about each hypothesis, such as body pose and appearance of the body parts. This detailed information was used to perform data association, which resulted in more reliable tracking despite long-term occlusions.

However, the approach in Chapter 3 was limited to 2D pose estimation, and relied on the assumption that people appear in the images in lateral pose. The applicability to other viewpoints was limited by our person detector, which was tuned to detect people in side views. In addition, our initialization and pose estimation procedure also relied on knowledge of the viewpoint and benefited from the fact that body poses of walking people seen from the side exhibit relatively little foreshortening of the body parts.

In this chapter we describe our approach to 3D human pose estimation and multi-person tracking. Here, in addition to data association, we also have to deal with depth ambiguities, which were avoided in Chapter 3. Here, we proceed by combining 2D evidence into tracklets, and rely on them to estimate viewpoints of people before actually performing 3D pose inference. The viewpoint estimation

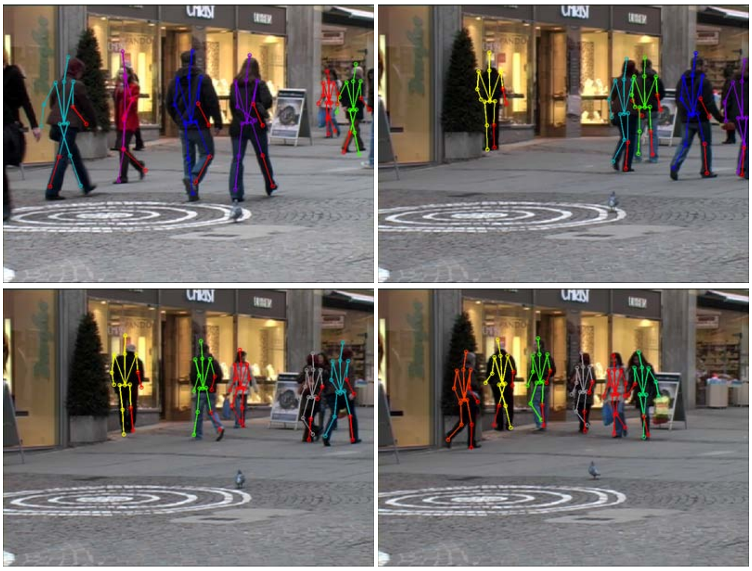


Figure 5.1: **Example results**: 3D tracking in a challenging scene.

is robust, because it combines evidence from all frames of the tracklet. Our 3D pose estimation procedure is different from many previous approaches in that we do not assume conditional independence of 2D image evidence given 3D poses in each frame. Rather, we use these dependencies to define a likelihood function for 3D poses, which has fewer local optima than a formulation in which single-frame evidence is assumed to be independent. In this chapter we also rely on our people detection and pose estimation techniques described in Chapter 4, which we use to detect people and estimate their viewpoints. Jointly these algorithmic and technical improvements allow us to extend the approach of Chapter 3 to tracking and pose estimation of people seen from arbitrary viewpoints in complex street scenes with multiple people, who often partially or fully occlude each other.

5.1 Introduction

In this chapter we addresses the challenging problem of 3D pose estimation and tracking of multiple people in cluttered scenes using a monocular, potentially moving camera. This is an important problem with many applications including video indexing, automotive safety, or surveillance. There are multiple challenges that contribute to the difficulty of this problem and need to be addressed simultaneously. Probably the most important challenge in articulated 3D tracking is the inherent ambiguity of 3D pose from monocular image evidence. This is particularly true for cluttered real-world scenes with multiple people that are often partially or even fully occluded for longer periods of time. Another important challenge, even for 2D pose recovery, is the complexity of human articulation and appearance. Additionally, complex and dynamically changing backgrounds of realistic scenes complicate data

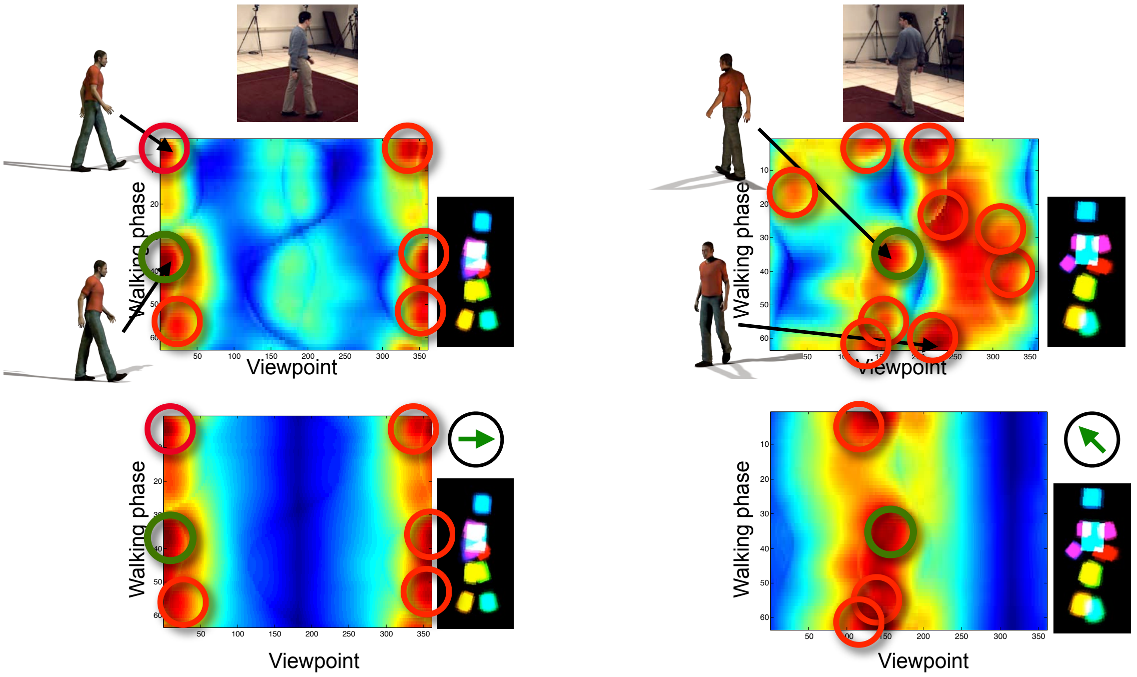


Figure 5.2: Visalization of the 3D pose likelihoods for two example images from the Human Eva II dataset. The x-axis of each plot corresponds to the viewpoint of the person with respect to the camera, and the y-axis corresponds to the position of the pose within the walking cycle. The top two plots correspond to the likelihood based only on the estimated 2D projections of the body parts, whereas the two plots along the bottom visualize the likelihood obtained by adding viewpoint estimates. In each case, the local maxima of the likelihood are denoted by circles, and the green circle denotes the local maxima corresponding to the correct solution. Note that the likelihood shown at the bottom, which combines both 2D projections and viewpoint estimates has significantly fewer strong local maxima corresponding to incorrect combination of pose and viewpoint, creating an easier search space.

association across multiple frames.

Our goal in this chapter is to contribute a sound Bayesian formulation to address this challenging problem. To that end we build on some of the most powerful approaches proposed for people detection and tracking in the literature. In three successive stages we accumulate the available 2D image evidence to enable robust 3D pose recovery.

Overview of the approach. Ultimately, our goal is to estimate the 3D pose Q_m of each person in all frames m of a sequence of length M, given the image evidence $\mathcal{E}_{1:M}$ in all frames. To that end, we define a posterior distribution over pose parameters given the evidence:

$$p(Q_{1:M}|\mathcal{E}_{1:M}) \propto p(\mathcal{E}_{1:M}|Q_{1:M})p(Q_{1:M}). \tag{5.1}$$

Here, $Q_{1:M}$ denotes the 3D pose parameters over the entire sequence. Clearly, a key difficulty is that the posterior in Eq. 5.1 has many local optima as the estimation of 3D poses is highly ambiguous given monocular images.

For example, the top row of Fig. 5.2 illustrates the likelihood of 3D pose parameters for the full range of viewpoints and phases of the walking cycle. Notice that for the image shown in the top-left, the 2D evidence provides sufficient constraints on pose and viewpoint, and the likelihood has only few local maxima. However, for the image in the top-right, a large portion of the parameter space has high likelihood, and there exist multiple local maxima which do not correspond to the correct pose. By extending 2D evidence with additional information, such as the viewpoint of the person, we can down-weight many of the incorrect local maxima, obtaining likelihood function shown in the bottom row of Fig. 5.2. However, reliable viewpoint estimates are difficult to obtain directly from single frames. In order to address this problem we propose a new three-stage approach, in which we sequentially reduce the ambiguity in 3D pose recovery.

In our approach the evidence in each frame is represented by the estimate of the person's 2D viewpoint w.r.t. the camera and the posterior distribution of the 2D positions and orientations of body parts. To estimate these from single frames, the *first stage* (Sec. 5.2) builds on a recently proposed people detection and pose estimation framework based on discriminative part detectors (Andriluka et al., 2009).

To accumulate further 2D image evidence, the *second stage* (Sec. 5.3) extracts people tracklets from a small number of consecutive frames using a 2D-tracking-by-detection approach. Here, the output of the first stage is refined in the sense that we obtain more reliable 2D detections of the people's body parts as well as more robust viewpoint estimates.

The *third stage* (Sec. 5.4) then uses the image evidence accumulated in the previous two stages to recover 3D pose. As described later, we model the temporal prior $p(Q_{1:M})$ over 3D poses as a hierarchical Gaussian process latent variable model (hGPLVM) (Lawrence and Moore, 2007). We combine this with a hidden Markov model (HMM) that allows to extend the people-tracklets, which cover only a small number of frames at a time, to possibly longer 3D people-tracks. Note that our 3D model is assumed to generate the bottom-up evidence from 2D body models and thus constitutes a hybrid generative/discriminative approach (c.f. (Tu et al., 2005)).

The main contribution of the work described in this chapter is a novel approach to human pose estimation, which combines 2D position, pose and viewpoint estimates into an evidence model for 3D tracking with a 3D motion prior, and is able to accurately estimate 3D poses of multiple walking people from monocular images in realistic street environments. The second contribution, which serves as a building block for 3D pose estimation, is a new pedestrian detection approach based on a combination of multiple part-based models. While the power of part-based models for people detection has already been demonstrated (e.g., Andriluka et al. (2009)), here we show that combining multiple part-based models leads to significant performance gains, and while improving over the state-of-the art in detection, also allows to estimate viewpoints of people in monocular images.

5.2 Multiview People Detection in Single Frames

2D people detection and pose estimation serves as one of our key building blocks for 3D pose estimation and tracking. Our approach is driven by three major goals: (1) We want to take advantage of the recent developments in 2D people detection and pose estimation to define robust appearance models for 3D pose estimation and tracking; (2) we aim to reduce the search space of possible 3D poses by taking advantage of inferred 2D poses; and (3) we want to extract the viewpoint from which people are visible to reduce the inherent 2D-to-3D ambiguity. To that end we build on and extend our 2D people detection and pose estimation approach from (Andriluka et al., 2009).

5.2.1 Basic pictorial structures model

Pictorial structures (Felzenszwalb and Huttenlocher, 2005) represent objects, such as people, as a flexible configuration of N different parts $L_m = \{\mathbf{l}_{m0}, \mathbf{l}_{m1}, \ldots, \mathbf{l}_{mN}\}$. m denotes the current frame of the sequence. The state of part i is given by $\mathbf{l}_{mi} = \{x_{mi}, y_{mi}, \theta_{mi}, s_{mi}\}$, where x_{mi} and y_{mi} denote its image position, θ_{mi} the absolute orientation, and s_{mi} the part scale. The posterior probability of the 2D part configuration L_m given the single frame image evidence D_m is given as

$$p(L_m|D_m) \propto p(D_m|L_m)p(L_m). \tag{5.2}$$

The prior on body configurations $p(L_m)$ has a tree structure and represents the kinematic dependencies between body parts. It factorizes into a unary term for the root part (here, the torso) and pairwise terms along the kinematic chains:

$$p(L_m) = p(\mathbf{l}_{m0}) \prod_{(i,j)\in K} p(\mathbf{l}_{mi}|\mathbf{l}_{mj}), \tag{5.3}$$

where K is the set of edges representing kinematic relationships between parts. $p(\mathbf{l}_{m0})$ is assumed to be uniform, and the pairwise terms are taken to be Gaussian in the transformed space of the joints between the adjacent parts (Andriluka et al., 2009; Felzenszwalb and Huttenlocher, 2005). The likelihood term is assumed to factorize into a product of individual part likelihoods

$$p(D_m|L_m) = \prod_{i=0}^{N} p(\mathbf{d}_{mi}|\mathbf{l}_{mi}). \tag{5.4}$$

To define the part likelihood, we rely on the boosted part detectors from (Andriluka et al., 2009), which use the truncated output of an AdaBoost classifier (Freund and Schapire, 1997) and a dense shape context representation (Belongie et al., 2000;

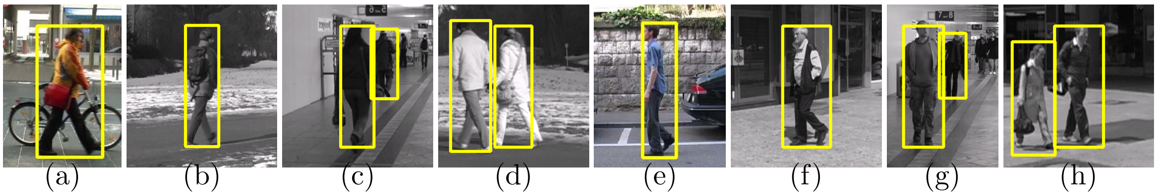


Figure 5.3: Training samples shown for each viewpoint: (a) right, (b) right-back, (c) back, (d) left-back, (e) left, (f) left-front, (g) front, (h) right-front.

Mikolajczyk and Schmid, 2005). Our model is composed of 8 body parts: left/right lower and upper legs, torso, head and left/right upper and lower arms (later sideview detectors also use left/right feet for better performance).

Apart from its excellent performance in complex real world scenes (Andriluka et al., 2009; Eichner and Ferrari, 2009), the pictorial structures model also has the advantage that inference is both optimal and efficient due to the tree structure of the model. We perform sum-product belief propagation to compute the marginal posteriors of individual body parts, which can be computed efficiently using convolutions (Felzenszwalb and Huttenlocher, 2005).

Data and evaluation. While the detector from (Andriluka et al., 2009) is in principle capable of detecting people from arbitrary views, its detection performance has only been evaluated on side views. To evaluate its suitability for our multiview setting, we collected a dataset of 1486 images for training, 248 for validation, and 248 for testing, which we carefully selected so that sufficiently many people are visible from all viewpoints. In addition to the persons' bounding boxes, we also annotated the viewpoint of all people in our dataset by assuming 8 evenly spaced viewpoints, each 45 degrees apart from each other (front/back, left/right, and diagonal front/back left/right). Fig. 5.3 shows example images from our training set, one for each viewpoint.

As expected and can be seen in Fig. 5.5(a), the detector trained on side-views as in (Andriluka et al., 2009) shows only modest performance levels on our multiview dataset. By retraining the model on our multiview training set we obtain a substantial performance gain, but still do not achieve the performance levels of monolithic, discriminative HOG-based detectors (Wojek et al., 2009) or HOG-based detectors with parts (Felzenszwalb et al., 2008) (see Fig. 5.5(b)). However, since we not only need to detect people, but also estimate their 2D pose, such monolithic or coarse part-based detectors are not appropriate for our task.

5.2.2 Multiview Extensions

To address this shortcoming, we develop an extended multiview detector that allows 2D pose estimation as well as viewpoint estimation. We train 8 viewpoint-specific detectors using our viewpoint-annotated multiview data. These viewpoint-specific

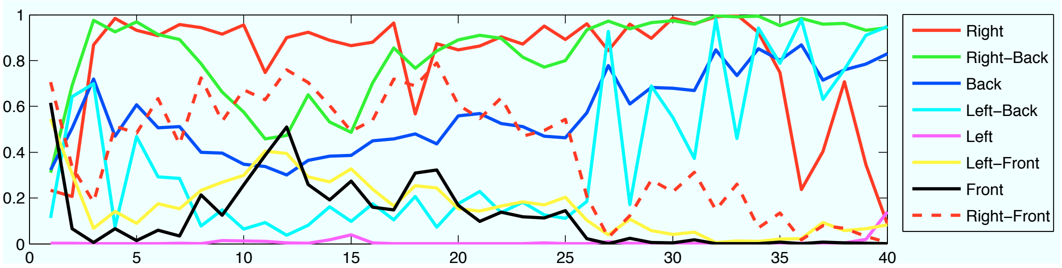


Figure 5.4: Calibrated output of the 8 viewpoint classifiers.

%	Right	Right-Back	Back	Left-Back	Left	Left-Front	Front	Right-Front	**Average**
Max	53.7	35.5	45.7	22.6	37.9	8.6	40.0	8.3	**31.1**
SVM	72.6	12.7	48.6	12.3	55.7	44.5	70.4	16.2	**42.2**
SVM-adj	71.4	22.3	29.5	18.0	84.7	18.1	50.7	29.2	**35.4**

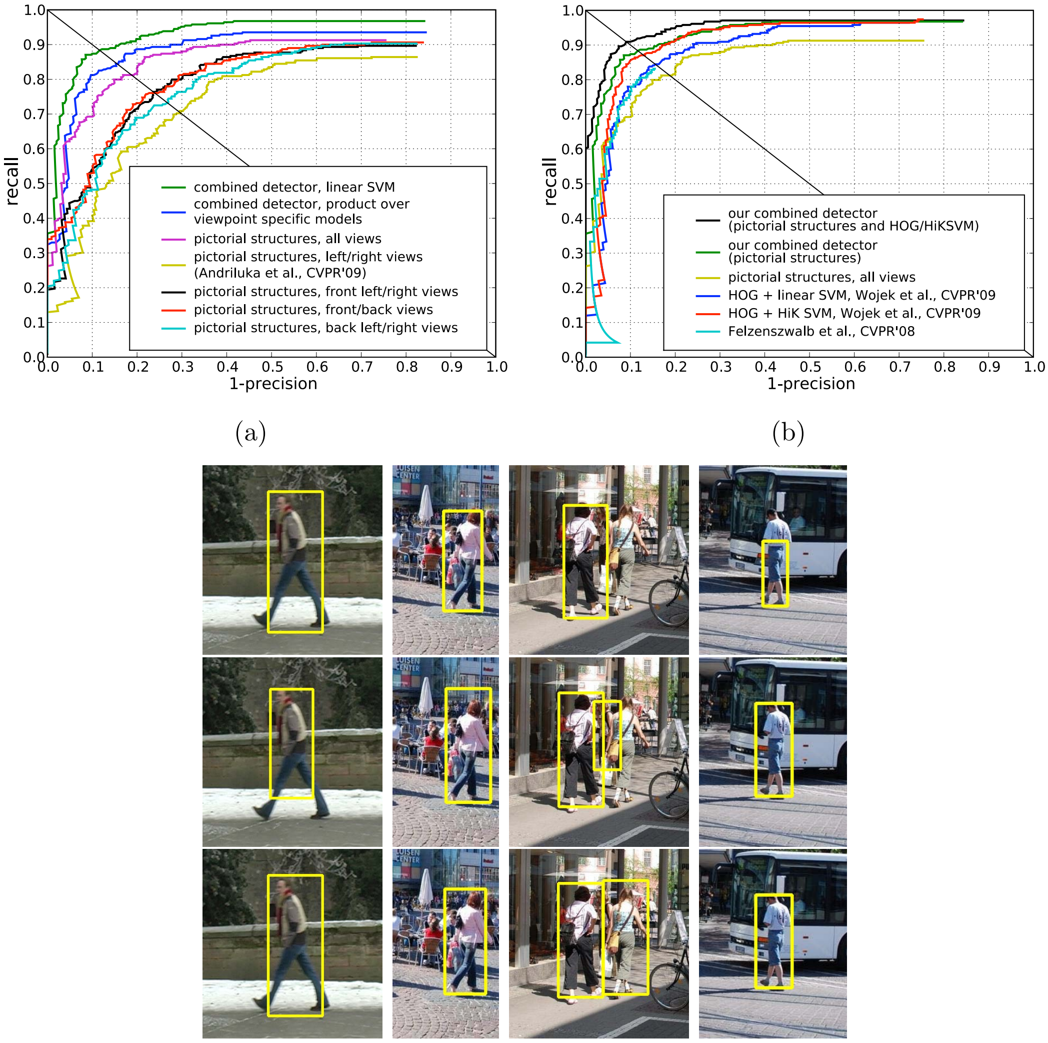
Table 5.1: Viewpoint estimation on the "MultiviewPeople" dataset. The task is to classify one of 8 viewpoints (chance level 12.5%).

detectors not only have the advantage that their kinematic prior is specific to each viewpoint, but also that part detectors are tuned to each view. We enrich this set of detectors with one generic detector trained on all views, as well as two side-view detectors as in (Andriluka et al., 2009) that additionally contain feet (which improves performance).

We explored two strategies for combining the output of this bank of detectors: (1) We simply add up the log-posterior of a person being at a particular image location as determined by the different detectors; and (2) we train a linear SVM using the 11-dimensional vector of the mean/variance-normalized detector outputs as features. The SVM detector was trained on the validation set of 248 images. Fig. 5.5(a) shows that the simple additive combination of viewpoint-specific detectors improved over the detection performance from each individual viewpoint-specific detector. It also outperforms the approach from (Andriluka et al., 2009).

Interestingly, the new SVM-based detector not only substantially improves performance, but also outperforms the current state-of-the-art in multiview people detection (Felzenszwalb et al., 2008; Wojek et al., 2009). As is shown in Fig. 5.5(b), the performance improves even further when we extend our bank of detectors with the HoG-based detector from (Wojek et al., 2009). While this is not the main focus of our work, this clearly shows the power of the first stage of our approach. Several example detections in Fig. 5.5(c) demonstrate the benefits of combining viewpoint-specific detectors.

Viewpoint estimation. Next we aim to estimate the person's viewpoints, since such viewpoint estimates allow to significantly reduce the ambiguity in 3D pose. To



(a) (b)

(c)

Figure 5.5: Comparison between (a) viewpoint-specific models and combined model, and (b) comparison to state-of-the-art on the "MultiviewPeople" dataset; (c) sample detections obtained with the side-view detector of Andriluka et al. (2009) (top), the generic detector trained on our multiview dataset (middle), and the proposed detector combining the output of viewpoint-specific detectors with a linear SVM (bottom).

that end, we rely on the bank of viewpoint-specific detectors from above, and train 8 viewpoint classifiers, linear SVMs, on the detector outputs of the validation set. We consider two training and evaluation strategies: (*SVM*) Only training examples from one of the viewpoints are used as positive examples, the remainder as negative ones; and (*SVM-adj*), where we group viewpoints into triplets of adjacent ones and train separate classifiers for each such triplet. As a baseline approach (*Max*), we estimate the viewpoint by taking the maximum over the outputs of the 8 viewpoint-

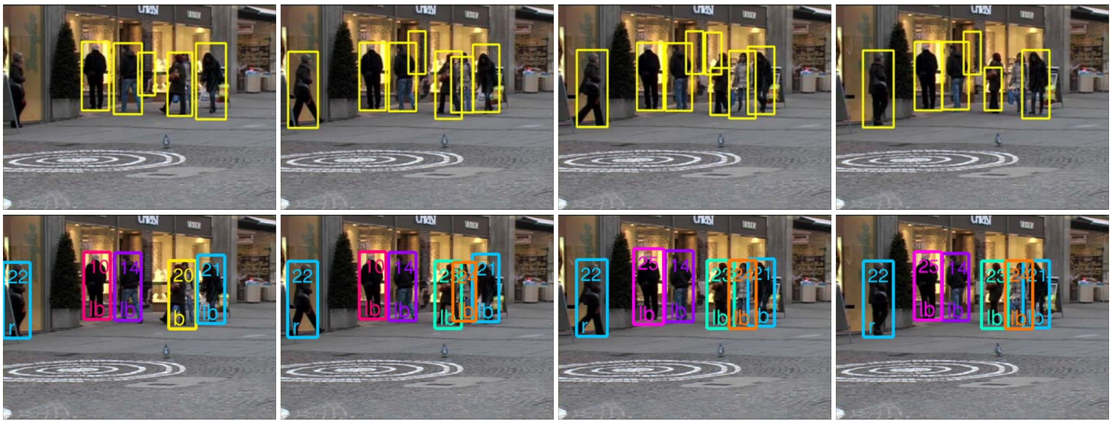


Figure 5.6: People detection based on single frames (top) and tracklets found by our 2D tracking algorithm; Different tracklets are identified by color and the estimated viewpoints are indicated with two letters (bottom). Note that several false positives in the top row are filtered out and additional – often partially occluded – detections are filled in (e.g., on the left side of the leftmost image).

specific detectors. Results are shown in Tab. 5.1. *SVM* improves over the baseline in case when we require exact recognition of viewpoint by approx. 11%, but *SVM-adj* also performs well. In addition, when we also consider the two adjacent viewpoints as being correct, *SVM* obtains an average performance of 70.0% and *SVM-adj* of 76.2%. This shows that *SVM-adj* more gracefully degrades across viewpoints, which is why we adopt it in the remainder.

Since the scores of the viewpoint classifiers are not directly comparable with each other, we calibrate them by computing the posterior of the correct label given the classifier score, which maps the scores to the unit interval. The posterior is computed via Bayes' rule from the distributions of classifier scores on the positive and negative examples. We assume these distributions to be Gaussian, and estimate their parameters from classifier scores on the validation set.

Fig. 5.4 shows calibrated outputs of all 8 classifiers computed for a sequence of 40 frames in which the person first appears from the "right" and then from the "right-back" viewpoint. The correct viewpoint is the most probable for most of the sequence, and failures in estimation often correspond to adjacent viewpoints.

5.3 2D Tracking and Viewpoint Estimation

As discussed in the introduction, our goal is to accumulate all available 2D image evidence prior to the third 3D tracking stage in order to reduce the ambiguity of 2D-to-3D lifting as much as possible. While the person detector described in the previous section is capable of estimating 2D positions of body parts and viewpoints of people from single frames, the second stage (described here) aims to improve these estimates by 2D-tracking-by-detection (Andriluka et al., 2008; Wu and Nevatia, 2007).

To exploit temporal coherency already in 2D, we extract short tracklets of people.

This, on the one hand, improves the robustness of estimates for 2D positions, scale and viewpoint of each person, since they are jointly estimated over an entire tracklet. Improved body localization in turn aids 2D pose estimation. On the other hand, it also allows to perform *early data association.* This is important for sequences with multiple people, where we can associate "anonymous" single frame hypotheses with the track of a specific person.

Tracklet extraction. From the first stage of our approach we obtain a set of N_m potentially overlapping bounding box hypotheses $\mathcal{H}_m = [\mathbf{h}_{m1}, \ldots, \mathbf{h}_{mN_m}]$ for each frame m of the sequence, where each hypothesis $\mathbf{h}_{mi} = \{h^x_{mi}, h^y_{mi}, h^s_{mi}\}$ corresponds to a bounding box at particular image position and scale. In order to obtain a set of tracklets, we follow the HMM-based tracking procedure introduced in Chapter 3.3.3. To that end we treat the person hypotheses in each frame as states and find state subsequences that are consistent in position, scale and appearance by iteratively applying Viterbi decoding. The emission probabilities for each state are derived from the detection score. The transition probabilities between states $\mathbf{h}_{mi}$ and $\mathbf{h}_{m-1,j}$ are modeled using first-order Gaussian dynamics and appearance compatibility:

$$\begin{aligned} p_{trans}(\mathbf{h}_{mi}, \mathbf{h}_{m-1,j}) &= \mathcal{N}(\mathbf{h}_{mi} | \mathbf{h}_{m-1,j}, \Sigma_{pos}) \cdot \\ &\quad \mathcal{N}(d_{app}(\mathbf{h}_{mi}, \mathbf{h}_{m-1,j}) | 0, \sigma^2_{app}). \end{aligned}$$

where $\Sigma_{pos} = \text{diag}(\sigma^2_x, \sigma^2_y, \sigma^2_s)$, and $d_{app}(\mathbf{h}_{mi}, \mathbf{h}_{m-1,j})$ is the Euclidean distance between RGB color histograms computed for the bounding rectangle of each hypothesis. We set $\sigma_x = \sigma_y = 5$, $\sigma_s = 0.1$ and $\sigma_{app} = 0.05$.

Viewpoint tracking. Finally, for each of the tracklets we estimate the viewpoint sequence $\omega_{1:N} = (\omega_1, \ldots, \omega_N)$, again using a simple HMM and Viterbi decoding. We consider the 8 discrete viewpoints as states, the viewpoint classifiers described in Sec. 5.2 as unary evidence, and Gaussian transition probabilities that enforce similar subsequent viewpoints to reflect that people tend to turn slowly.

Evaluation. Fig. 5.6 shows an example of a short subsequence in which we compare the detection results of the single-frame 2D detector with the extracted tracklets. Note how tracking helps to remove the spurious false positive detections in the background, and corrects failures in scale estimation, which would otherwise hinder correct 2D-to-3D lifting. In Fig. 5.4 we visualize the single frame prediction scores for each viewpoint for the tracklet corresponding to the person with index 22 on Fig. 5.6. Note that while viewpoint estimation from single frames is reasonably robust, it can still fail at times (the correct viewpoint is "right" for frames 4 to 30 and "right-back" for frames 31 to 40). The tracklet-based viewpoint estimation, in contrast, yields the correct viewpoint for the entire 40 frame sequence. Finally, as we demonstrate in Fig. 5.6, the tracklets also provide data association even in case of realistic sequences with frequent full and partial occlusions.

5.4 3D Pose Estimation

To estimate and track poses in 3D, we take the 2D tracklets extracted in the previous stage and lift the 2D pose estimated for each frame into 3D (c.f. Sigal and Black (2006*c*)), which is done with the help of a set of 3D exemplars (Mori and Malik, 2006; Shakhnarovich et al., 2003). Projections of the exemplars are first evaluated under the 2D body part posteriors, and the exemplar with the most probable projection is chosen as an initial 3D pose. This initial pose is propagated to all frames of the tracklet using the known temporal ordering on the exemplar set. Note that this yields multiple initializations for 3D pose sequences, one for each frame of the tracklet. This 2D-to-3D lifting procedure is robust, because it is based on reliable 2D pose posteriors, and detections and viewpoint estimates from the 2D tracklets. Starting from these initial pose sequences, the actual pose estimation and tracking is done in a Bayesian framework by maximizing the posterior defined in Eq. (5.1), for which they serve as powerful initializations. The 3D pose is parametrized as $Q_m = \{\mathbf{q}_m, \phi_m, \mathbf{h}_m\}$, where $\mathbf{q}_m$ denotes the parameters of body joints, ϕ_m the rotation of the body in world coordinates, and $\mathbf{h}_m = \{h_m^x, h_m^y, h_m^{scale}\}$ the position and scale of the person projected to the image. The 3D pose is represented using a kinematic tree with $P = 10$ flexible joints, in which each joint has 2 degrees of freedom. A configuration example is shown in Fig. 5.7(a).

The evidence at frame m is given by the $\mathcal{E}_m = \{D_m, \omega_m\}$, and consists of the single frame image evidence D_m and the 2D viewpoint estimate ω_m obtained from the entire tracklet.

We assume conditional independence of the evidence in each frame given the 3D pose parameters Q_m. The likelihood in Eq. 5.1 thus factorizes into single-frame likelihoods:

$$p(\mathcal{E}_{1:M}|Q_{1:M}) = \prod_{m=1}^{M} p(\mathcal{E}_m|Q_m). \tag{5.5}$$

We further assume conditional independence of 2D viewpoint and image evidence given the 3D pose, and define single frame likelihood as

$$p(\mathcal{E}_m|Q_m) = p(\omega_m|Q_m)p(D_m|Q_m). \tag{5.6}$$

Based on the estimated 2D viewpoint ω_m, we model the viewpoint likelihood of the 3D viewpoint ϕ_m as Gaussian centered at the rotation component of ϕ_m along the y-axis:

$$p(\omega_m|Q_m) = \mathcal{N}(\omega_m|\mathrm{proj}_y(\phi_m), \sigma_\omega^2). \tag{5.7}$$

We define the likelihood of the 3D pose $p(D_m|Q_m)$ with the help of the part posteriors given by the 2D body model

$$p(D_m|Q_m) = \prod_{n=1}^{N} p(\mathrm{proj}_n(Q_m)|D_m), \tag{5.8}$$

where $\mathrm{proj}_n(Q_m)$ denotes the projection of the n-th 3D body part into the image. While such a 3D likelihood is typically defined as the product of individual part likelihoods similarly to Eq. 5.4, this leads to highly multimodal posteriors and difficult inference. By relying on 2D part posteriors instead, the 3D model is focused on hypotheses for which there is sufficient 2D image evidence from previous stages.

To avoid expensive 3D likelihood computations, we represent each 2D part posterior using a non-parametric representation. In particular, for each body part n in frame m we find the J locations with the highest posterior probability $E_{mn} = \{(\mathbf{l}^j_{mn}, w^j_{mn}), j = 1, \ldots, J\}$, where $\mathbf{l}^j_{mn} \in \mathbb{R}^4$ corresponds to the 2D location (image position and orientation) and w^j_{mn} to the posterior density at this location. Given that, we approximate the 2D part posterior as a kernel density estimate with Gaussian kernel κ:

$$p(\mathrm{proj}_n(Q_m)|D_m) \approx \sum_j w^j_{mn} \kappa(\mathbf{l}^j_{mn}, \mathrm{proj}_n(Q_m)). \tag{5.9}$$

Dynamical model. In our approach we represent the temporal prior in Eq. 5.1 as the product of two terms:

$$p(Q_{1:M}) = p(\mathbf{q}_{1:M})p(\mathbf{h}_{1:M}), \tag{5.10}$$

which correspond to priors on the parameters of 3D pose as well as image position and scale. The prior on the person's position and scale $p(h_{1:M})$ is taken to be a broad Gaussian and models smooth changes of both the scale of the person and its position in the image.

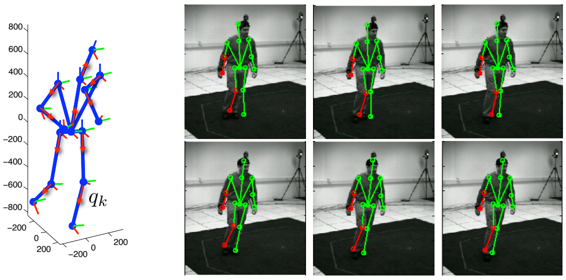
We model the prior over the parameters of the 3D pose $\mathbf{q}_{1:M}$ with a hierarchical Gaussian process latent variable model (hGPLVM) (Lawrence and Moore, 2007). We denote the M dimensional vector of the values of i-th pose parameter across all frames as $\mathbf{q}_{1:M,i}$. In the hGPLVM each dimension of the original high-dimensional pose is modelled as an independent Gaussian process defined over a shared low dimensional latent space $\mathbf{Z}_{1:M}$:

$$p(\mathbf{q}_{1:M}|\mathbf{Z}_{1:M}) = \prod_{i=1}^{P} \mathcal{N}(\mathbf{q}_{1:M,i}|0, \mathbf{K}_z), \tag{5.11}$$

where P is the number of parameters in our pose representation, $\mathbf{K}_z$ is a covariance matrix of the elements of the shared latent space $\mathbf{Z}_{1:M}$ defined by the output of the covariance function $\mathrm{k}(\mathbf{z}_i, \mathbf{z}_j)$, which in our case is taken to be squared exponential.

The values of the shared latent space $\mathbf{Z}_{1:M}$ are themselves treated as the outputs of Gaussian processes with a one dimensional input, the time $T_{1:M}$. Our implementation uses a $d_l = 2$ dimensional shared latent space. Such a hierarchy of Gaussian processes allows to effectively model both correlations between different dimensions of the original input space and their dynamics.

MAP estimation. The hGPLVM prior requires two sets of auxiliary variables $\mathbf{Z}_{1:M}$ and $T_{1:M}$, which need to be dealt with during maximum a-posteriori estimation.



(a) (b)

Figure 5.7: (a): Representation of the 3D pose in our model (parametrized joints are marked with arrows). (b): Initial pose sequence after 2D-to-3D lifting (top) and pose sequence after optimization of the 3D pose posterior (bottom).

Our strategy is to optimize $\mathbf{Z}$ only and keep the values of T fixed. This is possible since the values of T roughly correspond to the person's state within a walking cycle, which can be reliably estimated using the 2D tracklets. The full posterior over 3D pose parameters being maximized is given by:

$$p(Q_{1:M}, \mathbf{Z}_{1:M} | \mathcal{E}_{1:M}, T_{1:M}) \propto p(\mathcal{E}_{1:M} | Q_{1:M}) \cdot \\ p(\mathbf{q}_{1:M} | \mathbf{Z}_{1:M}) p(\mathbf{Z}_{1:M} | T_{1:M}) p(h_{1:M}). \quad (5.12)$$

We optimize the posterior using scaled conjugate gradients and initializing the optimization using the lifted 3D poses.

5.4.1 3D pose estimation for longer sequences

The MAP estimation approach just described is only tractable if the number of frames M is sufficiently small. In order to estimate 3D poses in longer sequences we apply a strategy similar to the one used in (Andriluka et al., 2008): First we estimate 3D poses in short ($M = 10$) overlapping subsequences of the longer sequence. Since for each subsequence we initialize and locally optimize the posterior multiple times, this leaves us with a large pool of 3D pose hypotheses for each of the frames, from which we find the optimal sequence using a hidden Markov model and Viterbi decoding. To that end we treat the 3D pose hypotheses in each frame as discrete states with emission probabilities given by Eq. 5.6 and define transition probabilities between states using the hGPLVM.

5.5 Experiments

We evaluate our model in two diverse scenarios. First, we show that our approach improves the state-of-the-art in monocular human pose estimation on the standard "HumanEva II" benchmark, for which ground truth poses are available. Additionally, we evaluate our approach on two cluttered and complex street sequences with multiple people including partial and full occlusions.

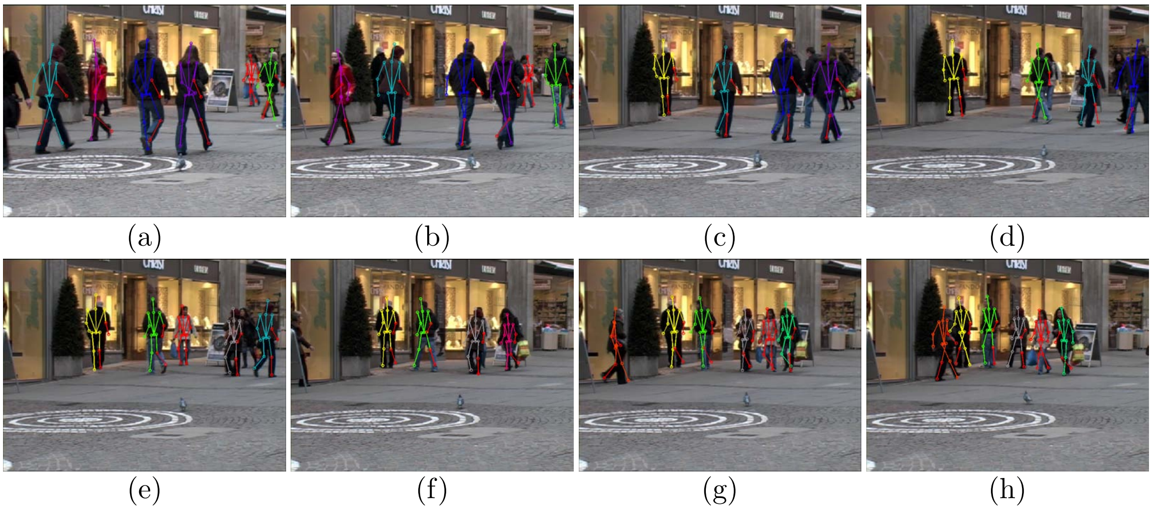

(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 5.8: 3D pose estimation on the "TUD Stadtmitte" dataset.

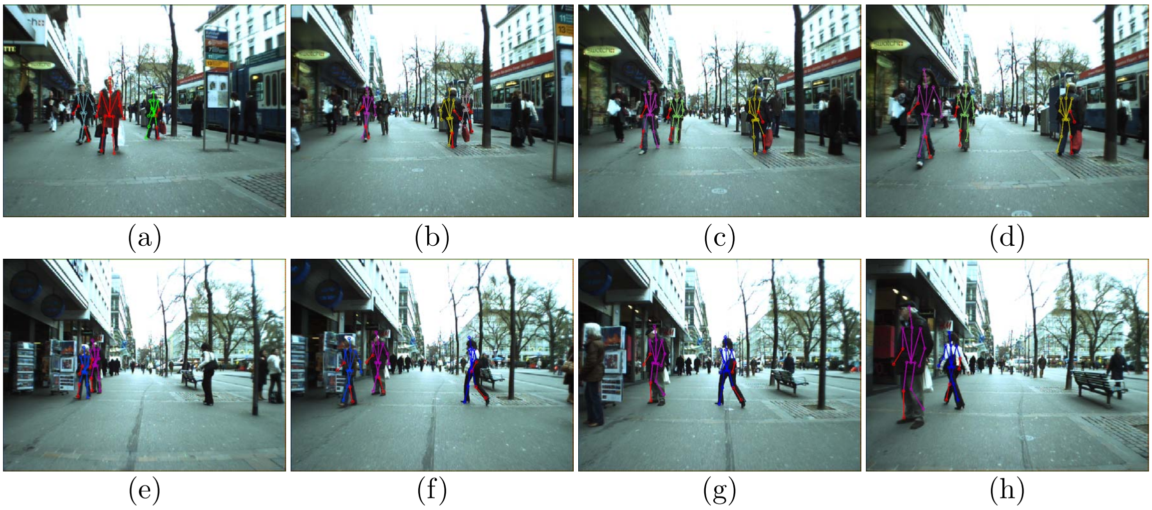

(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 5.9: 3D pose estimation on a sequence captured with a moving camera, previously used in (Gammeter et al., 2008).

5.5.1 Evaluation on the "HumanEva II" dataset

In order to quantitatively evaluate the performance of our 3D pose estimation method we use the "HumanEva II" dataset (Sigal and Black, 2006*a*), which provides synchronized images and motion capture data and is a standard evaluation benchmark for 2D and 3D human pose estimation. On this dataset we compare to (Rogez et al., 2008) as they obtain the best published results in a setting comparable to ours: They estimate poses in monocular image sequences without background subtraction, but rely on both appearance and temporal information.

For this experiment we train viewpoint specific models on the images of subjects "S1", "S2", "S3" from the "HumanEva" dataset. We found that adding more training data improves the performance of part detectors, especially for the lower and upper

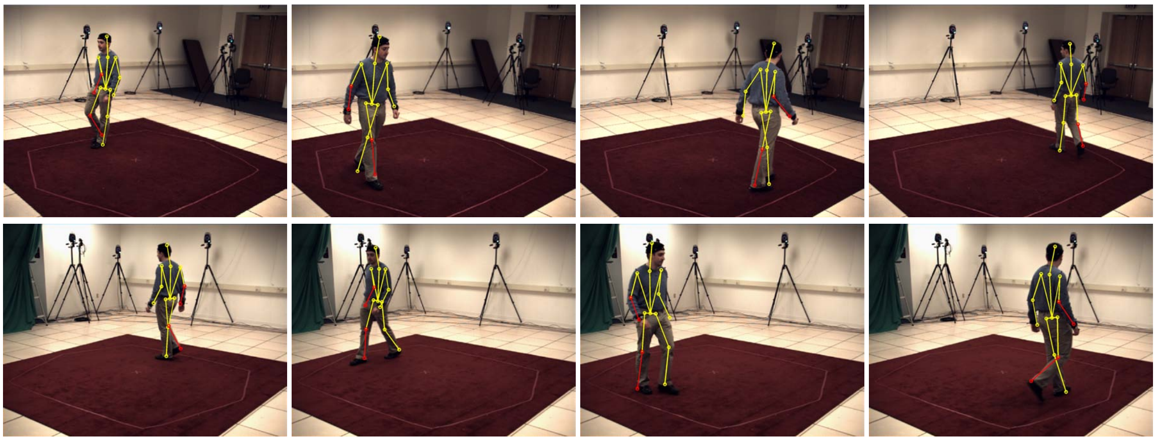


Figure 5.10: 3D pose estimation examples on HumanEva II for "Subject S2/Camera C1" (top) and "Subject S2/Camera C2" (bottom).

Subj./Cam.	2D Mean (Std) (Rogez et al., 2008)	2D Mean (Std)	3D Mean (Std)
S2/C1	12.98(3.5)	**10.49** (**2.70**)	107(15)
S2/C2	14.18(4.38)	**10.72** (**2.44**)	101(19)

Table 5.2: Quantitative evaluation on the HumanEva II dataset (frames 1-350). We report mean error and standard deviation of the relative 2D and 3D joint positions. 2D results are in pixels and 3D results are in millimeters.

arm body parts. Therefore, we extended the training data with the images from the "People" (Ramanan, 2006) and "Buffy" (Ferrari et al., 2008) datasets. The set of exemplars for 2D-to-3D lifting and the hGPLVM used to model the temporal dynamics on the pose sequences were obtained using training data for subject "S3" of the "HumanEva" dataset. As we show, despite the limited training data, this prior enables pose estimation on the "HumanEva II" dataset as well as in realistic street scenes.

Rogez et al. (2008) report pose estimation results for the first 350 frames of the sequence containing subject "S2", independently estimating poses for views obtained from cameras "C1" and "C2". Tab. 5.2 shows the mean error in the estimation of 2D and 3D joint locations, obtained using the official online evaluation tool. For both sequences our results improve substantially over those reported by Rogez et al. (2008). The improvement is especially large for the sequence taken with camera "C2", on which we obtain an average error of 10.72 pixels, compared to 14.18 pixels. We have also evaluated the 3D pose estimation performance of our approach, obtaining a mean error of 107 and 101 millimeters for cameras "C1" and "C2". Fig. 5.10 shows several examples of estimated poses obtained by our method on both sequences, visualizing every 100-th frame. We attribute the better localization accuracy of our method to the continuous optimization of 3D pose given the body part positions rather than selecting one from a discrete set of exemplars as in (Rogez et al., 2008).

5.5.2 3D pose estimation in street scenes

To evaluate our approach for a realistic street setting, we introduce the novel "TUD Stadtmitte"' dataset containing 200 consecutive frames taken in a typical pedestrian area. Over the 200 frames of this sequence our 2D tracking algorithm obtained 25 2D people-tracks, none of which contained false positive detections. Fig. 5.8 shows example images evenly spaced throughout the sequence. For every track we estimate the viewpoint of the person using exclusively our viewpoint classification algorithm. We could easily integrate the direction of movement of the person into the estimate, but this would limit the applicability of our method to the setting with static cameras. Note that we are able to estimate 3D poses correctly over a diverse range of viewpoints including those with significant depth ambiguity and difficult imaging conditions. Note, for example, the people on the right side of image Fig. 5.8(a) and the people in the middle of the image in Fig. 5.8(g).

Unfortunately we cannot report quantitative results on the 3D pose recovery for this sequence, as obtaining ground truth is very difficult for such realistic image sequences. However, the results are qualitatively close to those demonstrated by our method on the "HumanEva II" dataset, suggesting that the obtainable quantitative results would be comparable. Although our motion prior was trained on the "HumanEva" dataset, it generalized well to the street setting. Interestingly, we are also able to correctly estimate the pose of the person standing still, as is shown in Fig. 5.8(c,d,e). Looking at typical failure cases, several incorrectly estimated poses are due to incorrect scale estimation as in Fig. 5.8(a), partial occlusion Fig. 5.8(b,d), or failure in viewpoint estimation (e.g., the rightmost person in Fig. 5.8(g)).

We also evaluated our approach on a sequence recorded by a moving camera previously used in (Gammeter et al., 2008). Due to the high amount of background clutter, low frame-rate and many people in near frontal views, this sequence presents significant challenges for 3D pose estimation. Several examples of estimated 3D poses are shown in Fig. 5.9(right). Note that even under such challenging conditions our approach can track and estimate poses of people over a large number of frames, e.g., the rightmost person in Fig. 5.8(f,h). Also note that tracking and viewpoint estimation produced correct results even in the presence of strong background clutter, e.g., for the rightmost person in Fig. 5.8(b,c,d).

5.6 Conclusions

In this chapter we have presented a novel approach to monocular 3D human pose estimation and tracking, which is able to recover poses of people in realistic street conditions. The approach leverages recent advances in reliable 2D pose estimation from monocular images, tracking-by-detection, and powerful modeling of 3D dynamics based on hierarchical Gaussian process latent variable models. This allows to first accumulate the available 2D image evidence from which later 3D poses can be reliably recovered and tracked. The approach has been evaluated quantitatively on the “HumanEva II” benchmark and improves the state-of-the-art in this setting. We also showed excellent results on a challenging street sequence underlining the applicability of the approach for 3D pose estimation and tracking of multiple people in cluttered scenes using a monocular, potentially moving camera.

6

Vision Based Victim Detection from Unmanned Aerial Vehicles

In this chapter we consider the task of people detection in images taken from on-board camera of an unmanned aerial vehicle (UAV). This is an important task as its solution would allow to quickly survey a disaster site from the air and find humans needing help. This task is also extremely challenging as the large variability in clothing and possible poses of humans combined with possible partial occlusions create complex appearance patterns, which are hard to represent and detect well.

We proceed by evaluating various state-of-the-art visual people detection methods in the context of vision based victim detection from an UAV. The best performance in our comparison was shown by our approach presented in Chapter 4, and by the approach of Felzenszwalb et al. (2008), both of which rely on flexible part-based representations. We discuss their strengths and weaknesses, and demonstrate that by combining multiple detectors based on these approaches we can improve reliability of the overall system. We also demonstrate that the detection performance can be substantially improved by integrating the height and pitch information provided by on-board sensors. Jointly these improvements allow us to significantly boost the detection performance over the state of the art, which provides a substantial step towards making autonomous victim detection for UAVs practical.

6.1 Introduction

Finding human victims in post-disaster scenarios is one of the primary goals of any search and rescue (SAR) operation. Although significant progress has been made in developing ground robots for SAR applications, most of these robots still lack the mobility necessary for autonomous exploration of disaster sites. However, with the emergence of lightweight and inexpensive unmanned aerial vehicles (UAVs) it becomes possible to quickly survey a disaster site from the air in order to identify humans needing help (Cooper and Goodrich, 2008; Green et al., 2005; Nordberg et al., 2002).

In this chapter, we focus on victim detection from a UAV using an on-board daylight camera as our main sensing device. We envision that the development

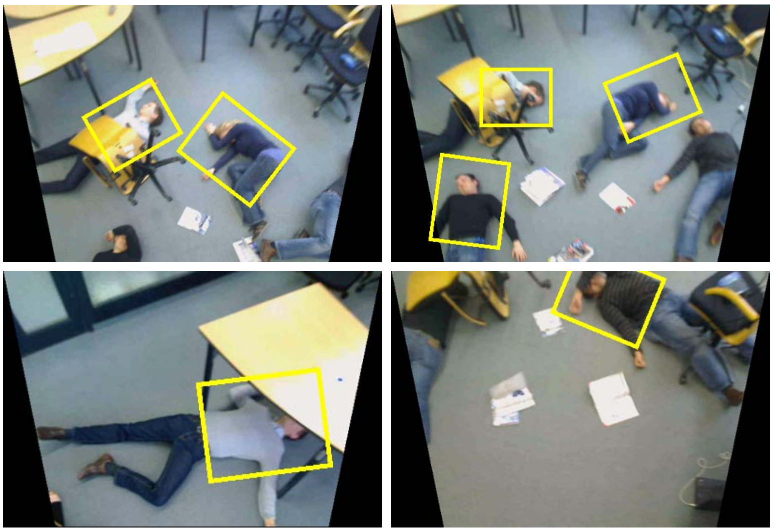


Figure 6.1: Several examples of people detection obtained with our approach on images captured from a quadrotor UAV.

of powerful vision-based victim detection methods will lead to a reduction in the number (and weight) of required on-board sensors and result in cheaper, smaller, and more power efficient UAVs. Additionally, robust vision-based detectors have proven to be an important building block when designing human detection systems based on multiple sensor modalities (Gate et al., 2009; Spinello et al., 2008).

Detection of people in images is a challenging problem. While significant progress has been made in specialized areas such as pedestrian detection (Dollár, Wojek, Schiele and Perona, 2009*b*), most approaches work best when people are fully visible and appear in a limited range of poses such as standing or walking. The best performing methods often use a monolithic representation of people, such as a HOG descriptor (Dalal and Triggs, 2005), and discriminative classifiers. Models of this type have recently been extended to incorporate motion (Dalal et al., 2006; Wojek et al., 2009) and color (Villamizar et al., 2009) features. They have also been applied to upper body detection (Ferrari et al., 2008), and have been integrated within larger systems to enable obstacle detection in mobile environments (Ess et al., 2009).

However, models based on monolithic representations are unlikely to generalize well to complex scenarios encountered in search and rescue applications (Murphy, 2004). In particular they are severely challenged by partial occlusions and high variabilities in poses of people, which frequently occur in such data. Fig. 6.1 shows several sample images acquired from the on-board camera of our UAV[1], which demonstrate the complexity of people detection in the scenario considered in this chapter: People are partially occluded and occur in a wide range of poses in an environment, which is typically highly cluttered. Using a monolithic person representation to detect characteristic full-body shapes is prone to fail in such a scenario.

A second type of people detection methods – called part-based models in the

---

[1] Shown detections correspond to the confidence level with equal precision and recall.

following – proceeds by decomposing complex appearances of humans into multiple components or parts (Andriluka et al., 2009; Bourdev and Malik, 2009; Felzenszwalb et al., 2008). In (Andriluka et al., 2009) we combine strong discriminatively trained body part detectors with a flexible body model based on the *pictorial structures* framework, which allows to detect highly articulated people. The approach of Felzenszwalb et al. (2008) also builds on the pictorial structures framework, but introduces a training procedure based on an unsupervised discovery of model parts, which are automatically chosen to optimize detection performance. Bourdev and Malik (2009) proposes to address the detection of articulated and partially occluded humans using a large number of specialized part detectors that are trained to detect body regions with characteristic appearances, which are also informative for the underlying 3D body configuration. In contrast to the monolithic representations mentioned above, these part-based models seem better suited to address the challenging problem of victim detection from a UAV.

In this chapter we make the following contributions. First, we discuss several state-of-the-art visual people detection methods, and evaluate them on a newly recorded dataset of images taken from a UAV. This evaluation includes both methods based on monolithic and on part-based representations. In particular we consider 2 people detectors (Dalal and Triggs, 2005; Ferrari et al., 2008) based on monolithic representations and 3 part-based detectors (Andriluka et al., 2009; Bourdev and Malik, 2009; Felzenszwalb et al., 2008). Second, after having identified the most suitable approaches to people detection from a UAV, we propose to augment these detectors with a prior distribution based on the pitch and height of the UAV measured by the on-board sensors. We demonstrate that this prior significantly improves the performance of people detectors and analyze reasons why not all detectors equally benefit from this. The third contribution is that we demonstrate that the considered people detectors are complementary, and that by combining them we can further improve the detection performance.

The rest of this chapter is organized as follows. Sec. 6.2 describes the quadrotor UAV used for data acquisition in our experiments. In Sec. 6.3 we introduce the approaches to people detection that we consider for comparison in this chapter, discuss their strengths and weaknesses, and describe our extensions. We present an experimental evaluation of these people detectors and our extensions in Sec. 6.4, and conclude and discuss future work in Sec. 6.5.

6.2 System Overview

The platform used for our experiments is a quadrotor helicopter developed at TU Darmstadt (Fig. 6.2). These kinds of vehicles are able to take off and land vertically and can hover at a fixed position, which motivates their application to search and rescue missions (Hoffmann et al., 2007). The propulsion system using four independently controlled motors and propellers allows the carriage of comparatively heavy payloads. Our quadrotor can carry up to 500g of cameras and other sensors

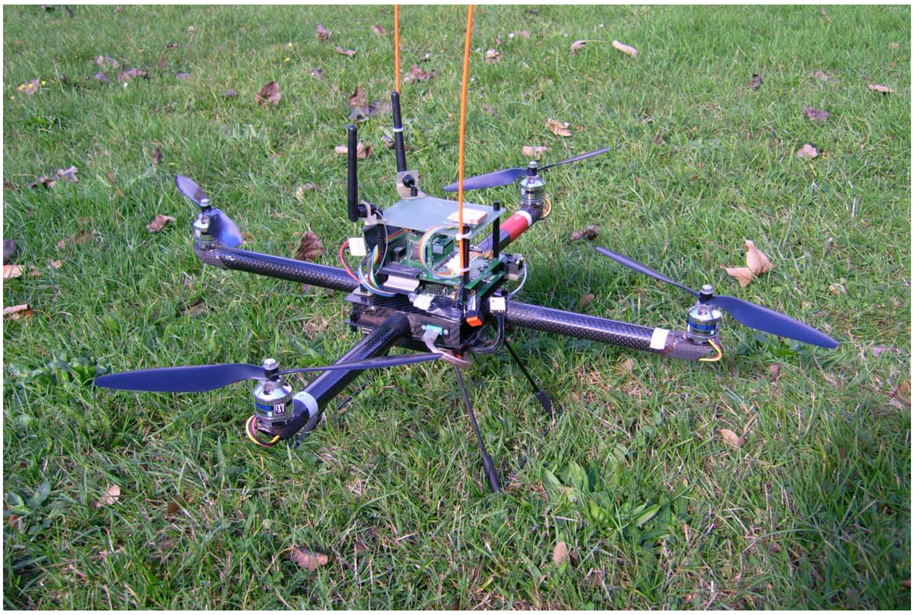


Figure 6.2: Quadrotor platform used for the experiments.

and weighs 1200g including the controller system and batteries for an endurance of approximately 20 minutes. With a diameter of 80cm it can be easily deployed in outdoor missions as well as for indoor scenarios.

Due to the instability of a quadrotor, the vehicle's attitude and velocity has to be controlled permanently. Therefore it is equipped with a 3-axis inertial sensor and magnetometer, a pressure sensor, a GPS receiver, and an ultrasonic ranger to measure the distance to the ground. The sensor information is fused using an extended Kalman filter running at 200 Hz deriving an integrated navigation solution using the algorithm of Titterton and Weston (2004). The outputs of the filter are fed to a cascaded PID controller to stabilize the attitude, height, velocity, and position of the quadrotor. As the rotational direction of two adjacent drives differ, the moments about all three axes and the total thrust can be controlled independently by simply varying the speed of the individual motors.

The control system is divided into two subparts: a microcontroller board interfacing the analog and digital sensors and motors; and a commercial embedded PC platform based on a current Intel Atom processor (Meyer and Strobel, 2010). The onboard computer executes the navigation, flight control, high-level mission control and communication tasks using the OROCOS Real-Time Toolkit (Bruyninckx, 2001). It interfaces the sensor board using a real-time enabled ethernet link.

For image acquisition a Logitech QuickCam Pro9000 camera is mounted to the quadrotor, which can transmit video images to the ground stations using the wireless network. Additionally, up to five frames per second are stored on an onboard flash media for after-mission analysis, including references to the available navigational data. The intrinsic camera parameters are calibrated using a publicly available calibration toolkit (Bouguet, 2010) and the extrinsic parameters relative to the ground plane are estimated using the height and attitude estimates provided by the UAV integrated navigation solution.

6.3 Vision-based People Detection

Detection of people in images is a challenging problem and many approaches to people detection have been proposed over the years. Approaches are often designed with a specific subproblem in mind, such as detection of pedestrians in street scenes (Dalal and Triggs, 2005), upper body detection (Ferrari et al., 2008), simultaneous detection and pose estimation (Andriluka et al., 2009), or generic people detection (Bourdev and Malik, 2009; Felzenszwalb et al., 2009).

One of the main contributions of this chapter is therefore to evaluate the applicability of these methods to search and rescue scenario and subsequently focus on improving performance of the best performing methods. While the evaluation is done in the context of victim detection from a UAV we believe that its results are applicable to people detection from mobile robots in general. In addition, we also demonstrate that in the case of a UAV we can further improve detection performance by using on-board sensor measurements in order to impose a prior on the scale of people in the image. In this section we briefly describe each of the considered approaches, and present an experimental comparison in Sec. 6.4.

Monolithic models. One of the most popular and effective models for people detection proposed to date is the histograms of oriented gradients (HOG) detector (Dalal and Triggs, 2005). In this model, histograms of image gradients are calculated and normalized in a local and overlapping block scheme and concatenated to a single descriptor of a detection window, which is densely scanned over all scales and locations in a test image. We consider HOG a monolithic model because the evidence of one detection window is encoded in a single descriptor, which is cast to a discriminative classifier (e.g., SVM), making a decision about presence or absence of the object of interest. This pairing with a powerful, discriminative classifier enables high levels of performance for object detection in cluttered scenes, e.g., pedestrian detection in street scenes (Dollár, Wojek, Schiele and Perona, 2009*b*). HOG was shown to learn a robust outer shape, which is shared by the positive training instances and delimits positive from negative samples. The local, overlapping normalization scheme enables robustness to illumination changes and to small variations in viewpoint. However, in the presence of high variability in articulation and partial occlusion HOG often fails because the model cannot recover from distorted monolithic descriptors. In our evaluation we consider two variants of HOG-based methods. i) The first variant is trained on full bodies of pedestrians. We make use of the implementation of our colleagues[2] (Wojek et al., 2009). ii) The second variant is trained on upper bodies of people[3] (Ferrari et al., 2008). Such monolithic approaches are a de-facto standard

---

[2] Project page: `www.mis.tu-darmstadt.de/tud-brussels/`

[3] Source code available at: `www.robots.ox.ac.uk/~vgg/software/UpperBody/`

when it comes to detection of people in settings with relatively little pose variation. However, it remains unclear, how these models generalize to the more challenging search and rescue scenario.

Part-based models. Part-based detection gives the flexibility necessary to deal with highly varying body poses. We consider three recently proposed part-based people detection methods: discriminatively trained part based models (DPM)[4] (Felzenszwalb et al., 2008), pictorial structures with discriminant part detectors (PS)[5] (Andriluka et al., 2009), and poselet based detection (PBD)[6] (Bourdev and Malik, 2009).

Our PS detector is built on the *pictorial structures* framework introduced in (Felzenszwalb and Huttenlocher, 2005). Here an object is represented as a flexible configuration of parts where one such configuration is denoted by $L = \{\mathbf{l}_0, \ldots, \mathbf{l}_N\}$, with $\mathbf{l}_i$ denoting the location of part i. In this generative formulation, the posterior over part configurations L given image evidence E is obtained via Bayes' rule: $p(L|E) \propto p(L)p(E|L)$. In order to enable efficient inference, PS employs a tree-structured Gaussian prior on L, and assumes that the overall likelihood can be decomposed into the product of individual part likelihoods. Under these assumptions the configuration posterior factorizes as:

$$p(L|E) \propto p(\mathbf{l}_0) \cdot \prod_{i=0}^{N} p(E|\mathbf{l}_i) \cdot \prod_{(i,j)\in G} p(\mathbf{l}_i|\mathbf{l}_j). \tag{6.1}$$

Sum-product belief propagation is applied in order to compute the marginal posterior of the torso, $p(\mathbf{l}_0|E)$, which is then used to delimit the detection bounding box. The detection results often account for multiple overlapping boxes that are post-processed with *non-maximum suppression* keeping only the hypothesis with the highest probability from significantly overlapping hypotheses.

The PS model employs body parts corresponding to upper and lower arms and legs, torso and head, and requires examples with labeled parts for training. Part likelihood terms $p(E|\mathbf{l}_i)$ are represented with discriminative part classifiers trained with AdaBoost. The pairwise terms $p(\mathbf{l}_i|\mathbf{l}_j)$ are estimated with maximum likelihood using the provided part labels.

The DPM model also relies on *pictorial structures* but differs in the prior imposed on the body parts. Here, a star shape prior is used, where all body parts are directly connected to the root part. Another difference is the interpretation of parts: While PS relies on manually labeled annotations, DPM automatically discovers the body parts that correspond to visually salient reoccurring structures in the training data. The configuration of body parts that maximizes Eq. (6.1) is found with max-product belief propagation. The entire model is trained in a purely discriminative fashion using the max-margin formalism. The appearance of each body part and the

---

[4] Source code available at: `people.cs.uchicago.edu/~pff/latent/`

[5] Source code available at: `www.d2.mpi-inf.mpg.de/code`

[6] Source code available at: `www.eecs.berkeley.edu/~lbourdev/poselets/`

root part are trained with SVMs, while a deformation cost of part constellations is obtained with gradient descent. DPM is specifically optimized for detection. Since no part annotations are required, it can be trained on a significantly larger training set than PS, which requires such part annotations.

The final approach to part-based people detection considered in our experiments is the poselet-based detector (PBD) recently proposed by Bourdev and Malik (2009). Instead of using the pictorial structures framework with a fixed number of parts, PBD relies on a large number of part detectors for diverse body regions denoted as “poselets”, which have consistent appearance and correspond to similar 3D body configurations. Detections of different poselets are integrated using a probabilistic voting procedure resembling the implicit shape model (Leibe et al., 2004) with weights learned using a max-margin framework (Maji and Malik, 2009*a*). Here every poselet votes for the location of the torso part, which in turn delimits the detection bounding box. Since the model is specifically designed to be robust to viewpoint and articulation changes, poselets often account for body regions that do not change significantly across articulation and viewpoint such as frontal faces, or correspond to frequently assumed body poses, such as legs of a standing person.

6.3.1 Proposed extensions

We propose two different kinds of extensions to the described vision based models: i) Since the different detectors focus on different aspects to be modeled, we propose to combine the complementary outputs of different detectors. ii) We introduce an extension that combines the vision based models with prior information obtained by the inertial sensors of the quadrotor.

Combining multiple models. The pictorial structures framework does not explicitly take the occlusion of body parts into account even though they frequently occur in our scenario. They happen due to complex poses in which some body parts are not visible, due to parts being outside of the view of the on-board camera, and due to miss-detections of some of the body parts from extreme foreshortening. In these cases the occluded body parts are fitted to spurious detections in the background, which results in a small probability of the overall configuration. In order to mitigate this problem we propose to combine the detection results of multiple models, each of which focuses on a different combination of body parts.

The DPM implementation used in our experiments is composed of two components, one upper-body and one full-body model. We extend the PS detector in a similar way and, complementary to a standard full-body detector, train an additional upper-body model, which is composed of torso, head, as well as upper and lower arms.

In order to fuse different models, we compute the posterior probability of each

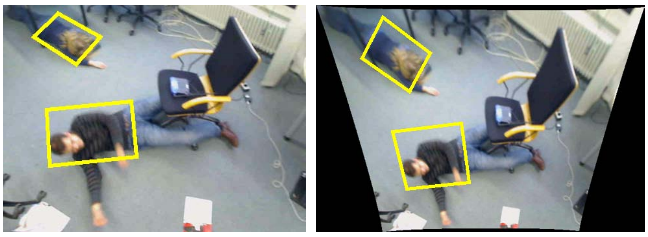


Figure 6.3: Original image taken by quadrotor at the height of 1.77 meters (left) and ground plane projection (right). The shown rectangles correspond to ground truth annotations.

hypothesis k given the detection score d_k of model $\mathcal{M}$ as:

$$p(h_k|d_k, \mathcal{M})) = \frac{p(d_k|h_k, \mathcal{M})}{p(d_k|h_k, \mathcal{M}) + p(d_k|\neg h_k, \mathcal{M})}, \tag{6.2}$$

where h_k is a Boolean variable corresponding to k-th hypothesis indicating whether it is correct or incorrect. $p(h_k|\mathcal{M})$ cancels as it is assumed to be uniform. The conditional distributions $p(d_k|h_k, \mathcal{M})$ and $p(d_k|\neg h_k, \mathcal{M})$ are assumed to be Gaussian, and fitted on a set of positive and negative detections.

The hypotheses of all models paired with the posterior probability are then cast forward to a joint non-maximum suppression step. Here, only the maximum scored detection is retained if several hypotheses overlap significantly. As the experiments demonstrate, this extension significantly improves the detection performance, especially on partially occluded people.

Scale prior based on UAV sensor measurements. The people detection methods discussed so far operate under the assumption that the camera position and depth for each image pixel are unknown. This implies that no prior information about the scale of the people in the image is available, and each model has to be exhaustively evaluated over all possible scales. However, in our scenario the quadrotor system is equipped with a calibrated camera and sensors capable of measuring the height and pitch angle of the vehicle. Combining these measurements allows to estimate the distance to the ground plane for each image pixel. Since our focus is on detecting people lying on the ground, this in turn provides an estimate of the scale of the person given an image position, subject to natural variation in people height and sensor noise. Similarly, knowing the position of the camera with respect to the ground plane we can back-project an image onto the ground plane taking both the homography transformation and image distortion into account (Heikkila and Silven, 1997). An example of this projection is shown in Fig. 6.3. Note that while the scale of people differs in the original image, after back-projection it becomes approximately the same. Additionally, the camera calibration and back-projection enables the relation of the height of the detection bounding boxes measured in pixels to the height of people measured in meters, which in turn allows to define a prior distribution on the bounding box height.

In the pictorial structures models, the posterior over configurations given by Eq. (6.1) contains the factor $p(\mathbf{l}_0)$ corresponding to the prior distribution on the position, scale and rotation of the root part $\mathbf{l}_0$, which is typically assumed to be uniform. When applying people detectors on back-projected images, we substitute this uniform prior with a Gaussian prior

$$p(\mathbf{l}_0) = \mathcal{N}(f(\mathbf{l}_0)|\mu_h, \sigma_h^2), \tag{6.3}$$

where $f(\mathbf{l}_0)$ is a linear transformation that converts the height of an hypothesis in pixels into metric units, $\mu_h = 0.8$ corresponds to an average upper body height of the person in meters, and $\sigma_h^2 = 0.1$. Note that this scale information is propagated to the other body parts through the body model.

As we show in the experimental section, not all models equally benefit from these priors. The PS model appears to be more precise in estimating the scale of people in an image, compared to DPM. While such precision is often not necessary when we are interested in detection only, it turned out to be beneficial when prior information about the expected scale of the person becomes available.

6.4 Experiments

6.4.1 Experimental setup

Dataset. The test set used in the experiments contains 220 images collected in an indoor office environment under uncontrolled daylight illumination conditions. During data collection our quadrotor was flying at a height between approximately 1.5 and 2.5 meters, capturing the images with interval of approximately 1 second. The captured dataset contains 285 ground truth annotations of people. Several sample images from the dataset are shown in Fig. 6.1.

When recording the test set we aimed to "simulate" difficulties typical for a search and rescue scenario: note the large diversity in poses of people present in the dataset; also note that many people are only partially visible, either because some parts of the body appear outside of the image, or due to self-occlusion, or due to occluding objects present in the scene. Obviously, in an ideal world, we would have access to a real and representative dataset from a real search and rescue situation. Besides the practical issues of obtaining such a dataset it is also unclear what such a "representative" dataset would be. So in order to increase the realism and difficulty of our evaluation, we decided not to train any of the evaluated people detection methods specifically for this scenario, but rather relied on the training sets provided with the respective method. Therefore, we explicitly evaluate the generalization performance of these methods to our test set, while simulating difficulties typical for search and rescue scenarios.

Evaluation methodology. In our dataset we annotated upper bodies of all people visible to at least 50%. For the evaluation we use the same criterion as in (Ferrari

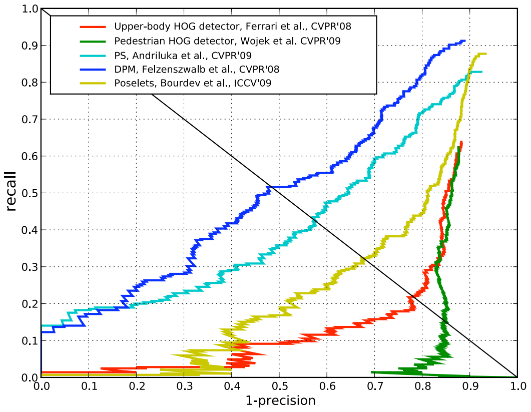


Figure 6.4: Comparison of people detection methods

et al., 2008), where a detection hypothesis is considered correct if the ratio of the intersection over the union of ground truth annotation and detection rectangles exceeded 0.25. We have chosen to define the detection task as detecting the upper body of the person, since all people detection methods considered in this paper are either designed to detect upper bodies (Bourdev and Malik, 2009; Ferrari et al., 2008) or can be easily adapted to do so. For each of the methods included in our comparison we are using the implementations and trained models made publicly available by the authors.

Our experiments consist of three parts: i) we compare different state-of-the-art methods of visual people detection, ii) we evaluate the importance of adding a scale prior to the model, and iii) we evaluate the performance of combinations of different detectors. For all of our experiments we report the equal error rate (EER) and show precision-recall curves.

6.4.2 Comparison of people detection methods

In our first experiment we evaluate the suitability of several recently proposed people detection methods for detecting articulated and partially occluded victims seen from our UAV.

The results are shown in Fig. 6.4 as recall-precision curves. Even though very common, the HOG based global template matching methods are not competitive in our setting. The HOG detector of Dalal and Triggs (2005) achieves 15.1% EER. The conceptually similar upper body detector “HOG-upper” of Ferrari et al. (2008) performs significantly better than the full body detector, but still achieves only 21.9% EER. Both HOG detectors do not perform well due to their monolithic structure, which does not take spatial variability in position of body parts into account. Several example detections of the full-body HOG detector are shown in Fig. 6.5 (first row). Note that while the HOG detector successfully copes with simple poses (e.g., bottom-left person in image (a)), it fails when body articulations vary significantly or when parts of the person become occluded as in images (b) and (c).

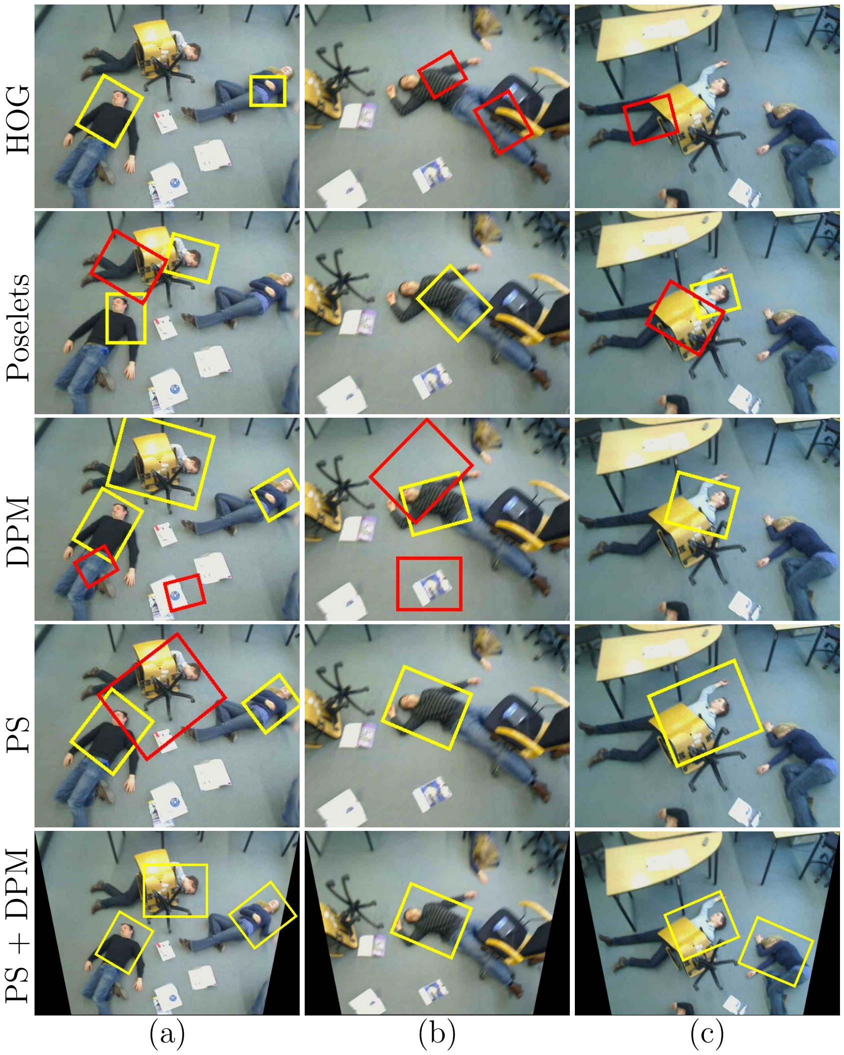
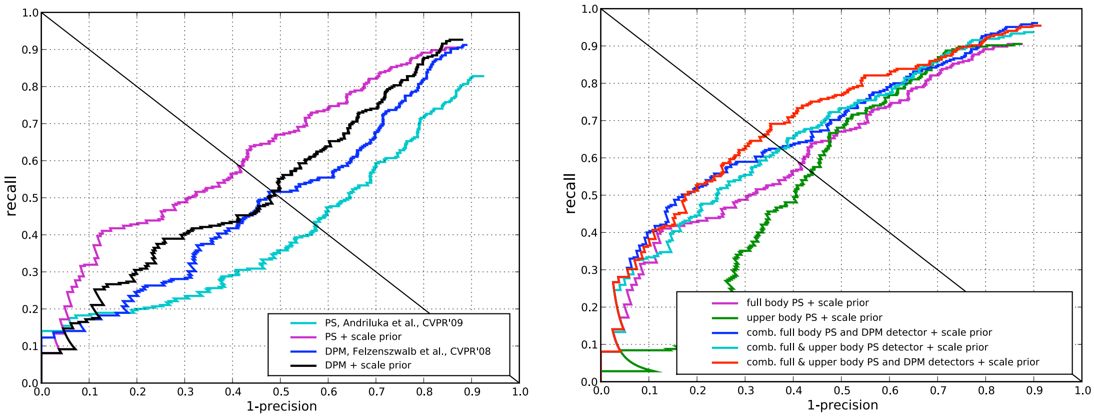


Figure 6.5: Several examples of detections at EER obtained with a **full-body HOG detector (Dalal and Triggs, 2005)** (1-st row), the **poselet detector (Bourdev and Malik, 2009)** (2-nd row), the **DPM detector (Felzenszwalb et al., 2009)** (3-rd row), the **full-body pictorial structures detector (Andriluka et al., 2009)** (4-th row) and the combined detector augmented with scale prior proposed in this paper (5-th row). True positive detections are plotted with yellow and false positives with red color.

The poselet detector (Bourdev and Malik, 2009) achieves 32.0% EER. Several example detections are shown in the second row of Fig. 6.5. Compared to the monolithic HOG detectors, the poselet detector appears to be more robust to partial occlusions. Note the correct localization of partially occluded people in images (a), (b), and (c). However, the poselet detector appears to be challenged by poses in which characteristic parts of the upper body are not visible, e.g., the rightmost person in Fig. 6.5(c). Additionally poselets seem to lack localization precision: Upper bodies are frequently localized with slight offsets from the correct position and scale,



(a) (b)

Figure 6.6: Comparison of performance with and without scale prior (a), and evaluation of different model combinations (b).

e.g. Fig. 6.5(b).

The two best performing detectors are both built on the pictorial structures framework. The PS detector (Andriluka et al., 2009) and the DPM detector (Felzenszwalb et al., 2008) achieve 42.5% and 51.5% EER respectively. Note that this corresponds to a performance improvement over the monolithic model (Dalal and Triggs, 2005) by 27.4% EER and 36.4% EER. The difference in performance between the PS and DPM detectors is most likely due to a significantly larger number of images used to train the DPM detector and the fact that the DPM model internally combines 2 models corresponding to a full-body and an upper-body configuration, while the PS detector uses a full-body model only. We have found that, although the DPM model yields better detection performances, it is often less precise in localizing people compared to the PS detector (see Fig. 6.5 images (b) and (c)). Such behavior might be due to the discriminative training procedure employed in the DPM model, which is specifically optimized for detection. This procedure does not reward improvement in localization beyond the limit set by the bounding box matching criterion. In contrast, the PS model is specifically designed for localization and body part detection and uses generative learning to estimate parameters of pairwise part relationships.

6.4.3 Integration of scale prior

Even the two best performing models (Andriluka et al., 2009; Felzenszwalb et al., 2008) frequently suffer from false positive detections. One source of such false positives are detections at incorrect scales. An example of such false positives produced by the PS model is shown in the fourth row of Fig. 6.5(a). Note that these false positives frequently correspond to unreasonable sizes of the human body when back-projected into world coordinates. For example the false positive detection in the fourth row of Fig. 6.5(a) would correspond to a person with a height of approxi-

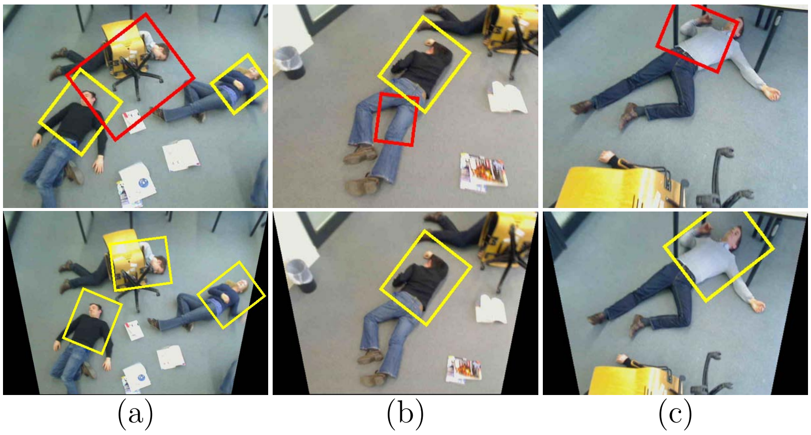

(a) (b) (c)

Figure 6.7: Examples of people detection at EER obtained with the **full-body pictorial structures detector** (**Andriluka et al., 2009**) without scale prior (first row) and with scale prior (second row).

mately 3.5 meters. As described in Sec. 6.3.1 we can reduce the influence of this kind of false positives by extending the detectors with a prior distribution on the height of detected people. This is accomplished by first projecting images onto the ground plane and subsequently introducing a Gaussian prior using known relations between pixels and metric units.

We compare the influence of the scale prior on the detection results for the two best performing methods in our evaluation. The results are shown in Fig. 6.6(a). While both models benefit from the scale prior, the improvement for the PS model is almost 16% EER, and is more significant than the improvement for the DPM model, which improves only slightly overall without improving the EER. An insight into these different behaviors can be gained by examining the distribution of scales of the true and false positive detections in the images projected onto the ground plane. These distributions are shown in Fig. 6.8 for the PS and DPM methods. Note that the PS model distribution of true positives has a clear peak around 120 pixels, which roughly corresponds to the upper-body height of 85 cm. False positives on the other hand occur mostly at small scales. For the DPM model the height of true positives is distributed almost uniformly in the range between 80 and 120 pixels. The DPM model appears to be less precise in scale estimation, which however is not reflected in the recall precision curve in Fig. 6.4 due to the rather loose bounding box matching criteria. However, this imprecision turns out to be a handicap when information about the detection scale is available from other sources. Fig. 6.7 shows several examples of detections of the original PS model, and the PS model augmented with the scale prior. Note that in addition to removing false positives as in images (a) and (b), the back-projection to the ground plane removes effects of perspective distortions, which also improves detection results, as for example in image (c).

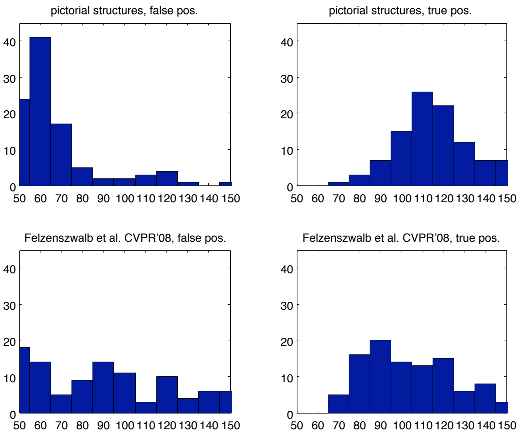


Figure 6.8: Distribution of the height of false positive (left) and true positive (right) detections for PS (Andriluka et al., 2009) (top row) and DPM (Felzenszwalb et al., 2008) (bottom row) detectors.

6.4.4 Combination of multiple detectors

Although both the PS and DPM detectors are built on the same pictorial structures framework, they differ significantly with respect to which parts are used in the model, which relationships between parts are considered and how the model parameters are learned from training data. The DPM model utilizes generic body parts that are automatically learned so that they are both discriminative and easy to localize in images. The DPM model appears to be more robust to occlusions since model parts are not fine tuned to detect particular parts of the body. This is in contrast to the PS model, which is designed to detect the actual body parts such as legs, torso, and head. As we found in our experiments, the PS model is superior to the DPM model in estimating the scale of a person, due to its more sophisticated body model enabling it to take advantage of a larger portion of the image evidence. In order to explore the complementarity of the PS and DPM detectors we derive a new detector based on their combination following the procedure described in Sec. 6.3.1. In addition to the original DPM and full-body PS detectors we also train an upper-body PS detector. The results of this experiment are shown in Fig. 6.6(b). In isolation, the upper-body PS detector did not perform nearly as well as the full-body PS detector, however the combination of these two detectors improves the EER from 58% for the full-body PS detector to 62%. A similar performance improvement is achieved when combining the full-body PS and DPM detectors. The best results are obtained by the detector combining full-body PS, upper-body PS and DPM detectors, which achieves 66% EER.

Several examples of correct detections and false positives are shown in Fig. 6.5 (bottom row) and Fig. 6.9. Note that compared to previously proposed detectors, our improved detector is able to find people occluded by the armchair in Fig. 6.5(a) and the strongly articulated person in Fig. 6.5(c). Top row of Fig. 6.10 shows several examples of people not detected by our system. Note, that such missing detections

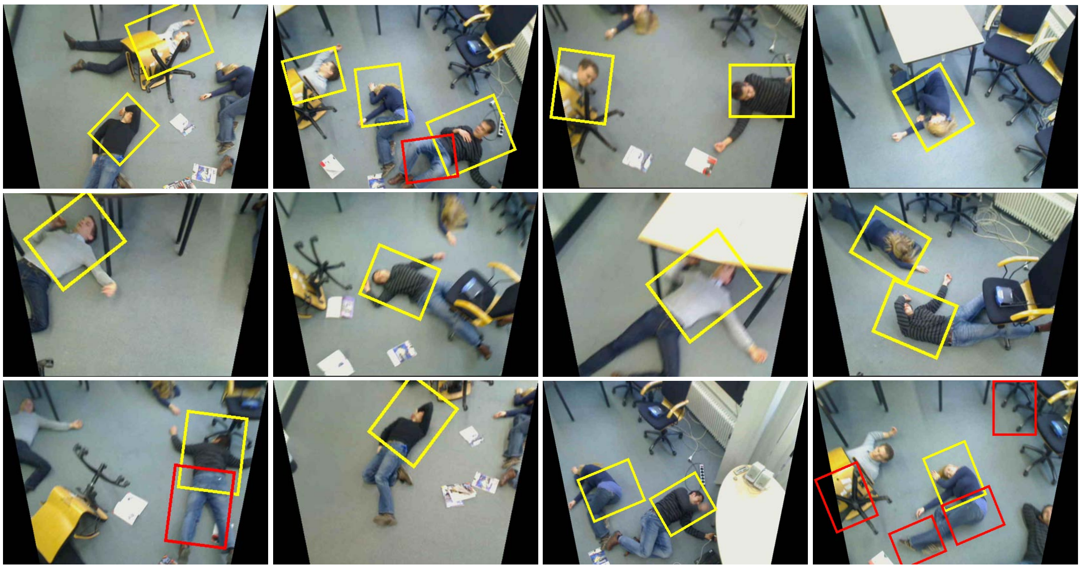


Figure 6.9: Examples of detections at EER obtained with the detector combining upper- and full-body PS, and DPM models, and scale prior.

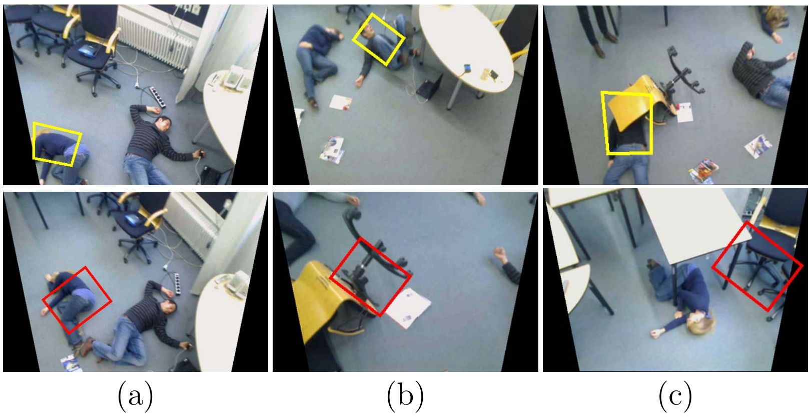


Figure 6.10: Missing recall (top row) and false positive (at ERR) detections (bottom row) of the detector combining upper- and full-body PS, and DPM models.

correspond to people with either especially severe occlusions as in images (a) and (c) or particularly complex articulations as in image (b). Such complex cases in which only few body parts are visible, appear to be beyond the capabilities of state-of-the-art detection methods. The bottom row of Fig. 6.10 also shows a couple of false positives obtained by our system at EER. While some of them correspond to nearly correct detections as in image (a), the detector also occasionally fires on background structures as in images (b) and (c).

6.5 Conclusion

In thic chapter we have evaluated the applicability of several state-of-the-art people detectors for victim detection from a UAV in a challenging search and rescue scenario. An important result of this comprehensive evaluation is that part-based models are better suited for victim detection than monolithic models, because they are able to represent variations in articulation and are robust to partial occlusions. As an extension to previous vision-based detectors we proposed to leverage complementary information of i) several detectors and ii) visual detectors and inertial sensor data of the UAV. Experimentally, we demonstrated that our extended framework substantially improved the detection performance, thus making a step towards autonomous victim detection in real world scenarios.

7

Conclusion and Future Work

Computer systems play an important role in our daily life, and there is little doubt that our reliance on them will only grow in the future. Increasingly often, the tasks delegated to computers go beyond simply performing computations and data processing and require active perception and interaction with the environment. In this thesis, we have addressed one particular challenge related to this new role of computers, focusing on algorithms that recognize people in static images and video. In the following, we briefly review the contributions of this thesis and discuss possible directions for future work.

7.1 Summary

One of the principal goals of our work was the development of algorithms for people detection that are applicable in realistic environments such as crowded street scenes. In Chapter 3, building on the approach of Seemann and Schiele (2006), we introduced a novel pedestrian detection system that relies on pictorial structures model. While this model has been previously used for pose estimation by Felzenszwalb and Huttenlocher (2005), we showed that it is also effective for people detection. In the same chapter, we also introduced a tracking-by-detection framework that combines output of the pedestrian detector with the dynamical body model in order to track and estimate poses of people in crowded street scenes. These results provided the foundation for the rest of the work presented in this thesis.

In Chapter 4, we introduced a generic model for people detection and pose estimation. One of the primary objectives of this work was to develop an approach with state-of-the-art performance in pedestrian detection that would also be applicable to other tasks, such as detection and pose estimation of actors in TV shows or victim detection in search-and-rescue scenarios. To that end, we switched from a star- to a tree-structured prior on part configurations and introduced a new discriminative appearance representation based on dense grids of local descriptors. We demonstrated that our new appearance model improves over the bottom-up appearance model of Seemann and Schiele (2006). We also showed that our approach achieves significant improvement in 2D pose estimation over the previous approaches of Ramanan (2006) and Ferrari et al. (2008).

In Chapter 5 we revisited the problem of pedestrian detection in conjunction with 3D pose estimation in crowded street scenes and demonstrated that we can further push the detection performance by combining output of multiple generic and viewpoint specific detectors . Compared to the pedestrian detector of Seemann and Schiele (2006), our approach from Chapter 5 achieves a significant improvement in performance and is capable of detecting pedestrians equally well for a full range of viewpoints and walking poses. As we have shown, our approach is also competitive with the best performing approaches to date. This is remarkable result considering a large effort undertaken in the computer vision community in the recent years towards solving pedestrian detection task. Compared to other approaches to pedestrian detection, our approach also provides more detailed description of the detected people in that it is capable of estimating their 3D body poses and orientations with respect to the camera.

Finally, in Chapter 6, we demonstrated an application of our approach to victim detection by unmanned aerial vehicles. We developed a solution which integrates the on-board sensor measurements into our detector and uses them in order to define informative prior on the scale of detections and remove the effects of perspective distortion.

In this thesis, we addressed a number of topics that are rarely handled within the same framework. In particular, people detection approaches often disregard explicit pose representation and are consequently unable to deliver any information beyond simple bounding box detections. At the same time, approaches for human pose estimation often employ simple appearance models based on image segmentation and require multiple cameras and clean static backgrounds, making them hard to apply in scenes of realistic complexity. Working on the intersection of detection, pose estimation and tracking allowed us to obtain impressive results that go beyond prior work in each of these areas. However, it has also limited the amount of time we could allot exclusively to each of these tasks. Consequently, some of the ideas that emerged during our work could not be fully explored. We would like to devote the rest of this chapter to discussing some of these ideas in the hope of coming back to them after completion of this thesis.

7.2 Perspectives for People Detection

A large portion of this thesis has been dedicated to pedestrian detection. Recent progress in this field suggests that we might see commercial pedestrian detection systems within the next few years. Therefore, in the future we would like to turn to more challenging problems such as generic people detection. In contrast to pedestrians, who appear in the images in walking poses and upright orientation, images of people exhibit much more complex patterns. These patterns arise due to interplay between body poses, appearance of body parts and partial occlusions, and they are very hard to represent and detect well. In order to address this challenge, we plan to focus on structured models, which represent the complex appearance of humans as a combination of local patterns and their spatial relationships. In order to properly model the self-occlusions, foreshortening of body parts and complex body articulations, we plan to rely on *3D body representations*, and combine them with appearance models of body parts that are capable of handling the multimodalities that result from changes in pose and viewpoint. One of our goals will be the development of local appearance models that are capable of estimating some of the 3D parameters of body parts, which could allow the system to iteratively aggregate the local evidence from individual parts into partial part assemblies that can be subsequently combined using 3D body representation.

7.3 Perspectives for 2D Pose Estimation

The approach to 2D human pose estimation introduced in (Andriluka et al., 2009) provides a solid baseline for further research in this area. The two main components of this work are the appearance representation of the body parts, and the articulated body model that defines constraints on the part configurations. We are certain that both of these components have a potential for future improvement. The appearance model used in our work relies on the shape context local features and AdaBoost training algorithm. A potential improvement to our appearance model could come from adding new types of appearance features, such as those capturing local self-similarity (Walk, Majer, Schindler and Schiele, 2010) or other types of image statistics (Dollár, Tu, Perona and Belongie, 2009). The improvement in performance could also be achieved by employing more elaborate learning algorithms that are better suited for capturing multimodalities in appearance of body parts. We would also like to explore more elaborate strategies for estimating model parameters. In our approach, we relied on a generative body model and interpreted the output of discriminative part classifiers as likelihood. Consequently, the spatial parameters of the model were estimated independently of each other using a maximum likelihood procedure. An interesting alternative would be to formulate the model as a conditional random field and rely on maximizing the conditional likelihood to learn the relative weighting of the unary appearance terms and the pairwise sparial terms.

However, some of the issues with the state-of-the-art approaches to 2D pose estimation (Andriluka et al., 2009; Eichner and Ferrari, 2009; Singh et al., 2010) are more profound and require fundamental changes which go beyond incremental improvements to appearance models or parameter learning algorithms. Currect approaches are limited in that that they rely on body models that do not take 3D effects such as foreshortening and self-occlusion into account and disregard some of the important correlations between positions of body parts (for example correlations between positions of the shoulder and hip joints that arise due to torso rotation). Essentially, these models did not get far from the "cardboard" people models that were proposed more than a decade ago (Ju et al., 1996). Therefore, in the future, we plan to focus on the approaches to 2D pose estimation that build on the more detailed body models and can properly represent foreshortening and self-occlusion. We expect that such body models will require more elaborate learning and inference techniques compared to those introduced in Chapter 4. In particular, with the increase in the number of parameters defining state of each body part it becomes prohibitively expensive to represent state-space using uniformly discretized grids. Standard message-passing algorithms also become prohibitively expensive, even when augmented with efficient message computation techniques. In the future in order to address these difficulties we plan to explore the adaptive discretizations (Isard et al., 2008) and other methods for approximate inference, such as Markov Chain Monte Carlo (Lee and Cohen, 2004).

Another important aspect that is not addressed by the current approaches is the handling of partial occlusions of body parts. Although several methods for explicit occlusion handling were proposed in the literature (Sudderth et al., 2004; Wang and Mori, 2008; Sigal and Black, 2006*b*), they typically rely on the availability of figure/ground segmentation and are applicable only to likelihood models that decompose into the product of pixel-wise terms. Such likelihoods are not directly applicable to cluttered, real-world scenes. Recently, it has been shown that output of part-based models can be used as a cue for figure/ground segmentation (Yang, Hallman, Ramanan and Fowlkes, 2010). In the spirit of this work, we can rely on our pose estimation approach to segment body parts of the person (see Fig. 7.1 for some preliminary results), and use these segmentations in conjunction with previously proposed occlusion handling methods (Sigal and Black, 2006*b*; Wang and Mori, 2008). In the future we would also like to extend the current evaluation proceedures to include evaluation of occlusion labeling and require disambiguation between left/right body parts. We feel that there is a clear need for such step, because the currently available datasets and metrics either ignore visibility of body parts (Ramanan, 2006; Ferrari et al., 2008; Eichner and Ferrari, 2009) or selectively exclude occluded parts from the evaluation (Yao and Fei-Fei, 2010).

Finally, in the future work we would like to explore the interplay between bottom-up and top-down approaches to appearance modelling. In this thesis, we explored both of them, first using a bottom-up appearance model based on the local features in Chapter 3, and then using a top-down sliding-window appearance model in Chapter 4. Generative bottom-up appearance models provide continuous estimates for the position, scale and orientation of body parts, and, in theory are able to operate accross wide range of spatial scales. However, these models are less robust to clutter and harsh imaging conditions relative to top-down discriminative appearance models. The latter operate by exhaustively searching over the discretized range of image positions, scales and orientations and consequently are able to estimate these variables only up to the discretization step. Recently, Levinshtein et al. (2009) have shown that bottom-up grouping procedures can be effective in finding salient image parts. In the future, we would like to explore the combination of bottom-up and top-down appearance models. To that end, bottom-up grouping can be employed to provide initial guesses for the locations of the body parts that is subsequently verified and refined in a top-down fashion.

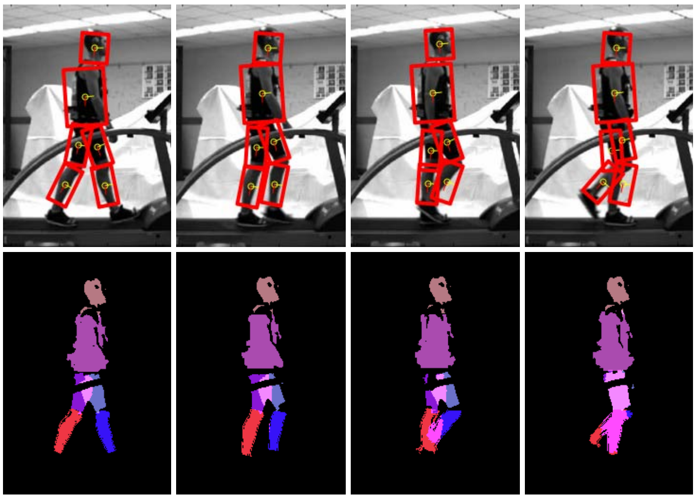


Figure 7.1: Estimated 2D body configurations (top row), and corresponding segmentations of the body parts (bottom row). The segmentations are obtained using color-based appearance models derived from the configuration estimates.

7.4 Perspectives for Monocular 3D Pose Estimation

In this thesis we introduced an approach for the monocular 3D pose estimation of walking people in crowded street scenes. An important objective for future work will be to generalize this approach to more complex settings with people involved in multiple, previously unseen activities. This is a significant challenge that requires further development of both appearance representation and dynamical body models.

Until now, low dimensional representations played an important role in 3D pose estimation. Models such as GPLVMs, Laplacian Eigenmaps, and PCA were used in (Urtasun et al., 2006; Fossati et al., 2007; Andriluka et al., 2008; Gammeter et al., 2008; Andriluka et al., 2010) in order to constrain the search to the relevant portion of the high dimensional pose space and capture the correlations between body joints. These models are well-suited for cyclic motions, such as walking or running, and were also shown to be applicable to other simple motions such as waving or swinging a golf club (Schwarz et al., 2010). However, it is not clear how well they generalize to unconstrained settings that feature multiple motions in which individual motion components often intervene with each other (e.g., walking and waving at the same time). Several authors have proposed to address this issue using switching activity priors (Schwarz et al., 2010) or hierarchical body models (Darby et al., 2009). However, the former method is limited to a few previously observed activities and is not able to represent transitions between them, while the latter has only been shown to work for simple activity combinations and has not been demonstrated to scale to larger volumes of training data. Promising results were shown by models that are designed to incorporate multiple motions within the same representation, such as Restricted Boltzmann Machines (Smolensky, 1986). Recently, these models were shown to be able to extract atomic motion primitives from motion capture data and could scale to much larger training sets containing thousands of examples (Taylor et al., 2010). In the future work we would like to extend our tracking-by-detection framework in order to incorporate these more generic motion models.

Multiple sources of constraints are necessary to resolve the depth and data association ambiguities that arise during 3D pose estimation. Clearly, some of these constraints can be provided by more elaborate dynamical body models and improved appearance models. Additional constraints can also be obtained by integration of image cues, such as shading and shadows (Balan, Black, Haussecker and Sigal, 2007). However, more substantial developments are necessary to scale from the handful of motions supported by the current models to the hundreds of motions routinely performed by humans in their daily life. We believe that one of the essential properties missing in the current approaches is tighter integration of the estimation of people's poses, their activities and the overall scene context. The missing sources of information include the interactions between multiple people present in the scene and the interactions of people with the environment and scene objects. These interactions might provide essential cues for pose estimation. At the same time, estimated poses themselves can be used as evidence for scene labeling and annotation. Our hypothesis is that tighter integration of models for scenes, activities, and poses of people is essential for understanding the images acquired in complex real-world conditions, and we will aim towards such integrated approaches in the future.

List of Figures

1.1 Several examples of images containing two people interacting with each other. Notice that the poses of the people often provide essential cues for determining the type of interaction. 4

1.2 Several examples of images with people that illustrate some of the challenges for people detection and articulated pose estimation. The images are ordered from left to right by their complexity for 2D pose estimation as measured by the percentage of body parts correctly localized by the approach of Andriluka et al. (2009). 4

2.1 Diagram showing spatial relations between parts of the face, and likelihood functions (or reference description) for the left edge of the face (a) and the eye (b). This figure is reproduced from the original pictorial structures work of Fischler and Elschlager (1973). 10

2.2 Diagram showing the main components of the pictorial structures based people detection and pose estimation system proposed here. . . 10

2.3 Several examples of detections obtained with the model of Felzenszwalb et al. (2009). 17

3.1 Examples of detection and tracking of *specific* persons in image sequences of crowded street scenes. 28

3.2 Graphical model structure describing the relation between articulation, parts, and features. 29

3.3 Examples of images from our training set. 30

3.4 The main components of the Implicit Shape Model (Leibe et al., 2008): appearance codebook (a) and learned spatial occurrence distribution corresponding to each the codebook entries (b). 31

3.5 (a) Test image, and (b) scale-invariant Hessian-Laplace interest points detected on the image. 33

3.6 (a) Several detected interest points, and conditional posterior distributions $p(\mathbf{x}^i|a, \mathbf{e}_k)$ of the foot (b), lower leg (c), and head (d) body parts given position $\mathbf{e}_k^{pos}$ and local image feature $\mathbf{e}_k^{app}$ corresponding to each of the interest points. Distributions corresponding to different interest points are visualized using different colors (see text for description). 33

3.7 Person hypothesis with corresponding probabilities of feet (b), lower legs (c), torso (d), and head (e). 34
3.8 Example detections at equal error rate of our detector (top), 4D-ISM (middle) and HOG (bottom) on the "TUD-Pedestrians" dataset. . . . 34
3.9 Representation of articulations in the latent space. 38
3.10 Comparison of our pedestrian detector with 4D-ISM (Seemann and Schiele, 2006) and HOG (Dalal and Triggs, 2005) on (a) the "TUD-Pedestrians" and (b) "TUD-Campus" datasets. Numbers in parenthesis indicate number of training images. 38
3.11 Examples of detector hypotheses (top row) and tracklet hypotheses (bottom row) at equal error rate on the "TUD-Campus" dataset. . . . 40
3.12 Detection and tracking on "TUD-Campus" dataset (see supplementary material). 41
3.13 Detection and tracking on "TUD-Crossing" dataset (see supplementary material). 42

4.1 **Sample results** (left to right): Full body pose estimation (1st and 2nd column), upper-body pose estimation, and pedestrian detection. *Bottom:* Part posteriors. 46
4.2 Left: **Kinematic prior** learned on a multi-view and multi-articulation dataset (Ramanan, 2006). Mean part positions (blue dots) and covariances of the part relations in the transformed space (red ellipses) are shown. Right: **Several independent samples** from the learned prior (to ease visualization visualized assuming a fixed torso position and orientation). 47
4.3 **Priors on the part configurations** (left to right): Pedestrian detection (star vs. tree model) and upper body detection. 50
4.4 Likelihood maps of the torso, head and upper arm visualized for the neighborhood around the ground-truth part location. 53
4.5 Visualization of the factors between body parts in our model. Solid lines correspond kinematic factors and dashed lines correspond to repulsive factors. 55
4.6 Example images from **"People"** dataset, with poses estimated by a tree model (top) and a non-tree model with repulsive factors (bottom). 60
4.7 Results obtained with sum-product (b, d) and max-product (a, c) algorithms for coarse 8 px. (top row) and fine 1 px. (bottom row) grid of evaluated part locations. 62
4.8 Several examples of detection at equal error rate obtained with our model (8 parts and tree prior, top) and partISM (bottom) on the **"TUD-Pedestrians"** dataset. 65
4.9 Several examples of detection and pose estimation obtained with our model (8 parts and tree prior) on the **"TUD-Campus"** sequence. . 65

4.10 **Pedestrian detection results:** Performance with varying number of parts on (a) **TUD-UprightPeople** and (b) **TUD-Pedestrians**. (c) Comparison of part detectors with varying step sizes between local descriptors on **TUD-UprightPeople**. 66

4.11 **Pedestrian detection results:** Comparison with previously proposed approaches partISM (Andriluka et al., 2008), HOG (Dalal and Triggs, 2005) on (a) **TUD-UprightPeople**, (b) **TUD-Pedestrians**, and (c) **TUD-Campus**. On the **TUD-Pedestrians** (b) we also compare with the “Hough Forests” approach of Gall and Lempitsky (2009). 67

4.12 **Upper body pose estimation:** Examples of pose estimation results of our method on the **“Buffy Stickmen”** dataset (see text for details). Note that color information is not used for detection. 69

4.13 **Upper body pose estimation:** Several examples of the typical failure cases of our method on the **“Buffy Stickmen”** dataset. . . . 69

4.14 Comparison of part localization performance on **“Buffy Stickmen”** (top) and **“ETHZ PASCAL Stickmen”** (bottom) datasets. 72

4.15 Comparison of full body pose estimation results between our approach (top) and (Ramanan, 2006) (bottom) on the **“People”** dataset. The numbers on the left of each image indicate how many of the 10 body parts were correctly localized. 75

4.16 Several examples of difficult images from **“People”** dataset, on which neither our (top) nor the approach from (Ramanan, 2006) (bottom) was able to estimate more then 6 body parts correctly. The numbers on the left of each image indicate how many of the 10 body parts were correctly localized. 76

5.1 **Example results**: 3D tracking in a challenging scene. 80

5.2 Visalization of the 3D pose likelihoods for two example images from the Human Eva II dataset. The x-axis of each plot corresponds to the viewpoint of the person with respect to the camera, and the y-axis corresponds to the position of the pose within the walking cycle. The top two plots correspond to the likelihood based only on the estimated 2D projections of the body parts, whereas the two plots along the bottom visualize the likelihood obtained by adding viewpoint estimates. In each case, the local maxima of the likelihood are denoted by circles, and the green circle denotes the local maxima corresponding to the correct solution. Note that the likelihood shown at the bottom, which combines both 2D projections and viewpoint estimates has significantly fewer strong local maxima corresponding to incorrect combination of pose and viewpoint, creating an easier search space. 81

5.3 Training samples shown for each viewpoint: (a) right, (b) right-back, (c) back, (d) left-back, (e) left, (f) left-front, (g) front, (h) right-front. 84
5.4 Calibrated output of the 8 viewpoint classifiers. 85
5.5 Comparison between (a) viewpoint-specific models and combined model, and (b) comparison to state-of-the-art on the "MultiviewPeople" dataset; (c) sample detections obtained with the side-view detector of Andriluka et al. (2009) (top), the generic detector trained on our multiview dataset (middle), and the proposed detector combining the output of viewpoint-specific detectors with a linear SVM (bottom). . 86
5.6 People detection based on single frames (top) and tracklets found by our 2D tracking algorithm; Different tracklets are identified by color and the estimated viewpoints are indicated with two letters (bottom). Note that several false positives in the top row are filtered out and additional – often partially occluded – detections are filled in (e.g., on the left side of the leftmost image). 87
5.7 (a): Representation of the 3D pose in our model (parametrized joints are marked with arrows). (b): Initial pose sequence after 2D-to-3D lifting (top) and pose sequence after optimization of the 3D pose posterior (bottom). 91
5.8 3D pose estimation on the "TUD Stadtmitte" dataset. 92
5.9 3D pose estimation on a sequence captured with a moving camera, previously used in (Gammeter et al., 2008). 92
5.10 3D pose estimation examples on HumanEva II for "Subject S2/Camera C1" (top) and "Subject S2/Camera C2" (bottom). 93

6.1 Several examples of people detection obtained with our approach on images captured from a quadrotor UAV. 98
6.2 Quadrotor platform used for the experiments. 100
6.3 Original image taken by quadrotor at the height of 1.77 meters (left) and ground plane projection (right). The shown rectangles correspond to ground truth annotations. 104
6.4 Comparison of people detection methods 106
6.5 Several examples of detections at EER obtained with a **full-body HOG detector** (**Dalal and Triggs, 2005**) (1-st row), the **poselet detector** (**Bourdev and Malik, 2009**) (2-nd row), the **DPM detector** (**Felzenszwalb et al., 2009**) (3-rd row), the **full-body pictorial structures detector** (**Andriluka et al., 2009**) (4-th row) and the combined detector augmented with scale prior proposed in this paper (5-th row). True positive detections are plotted with yellow and false positives with red color. 107
6.6 Comparison of performance with and without scale prior (a), and evaluation of different model combinations (b). 108

6.7 Examples of people detection at EER obtained with the **full-body pictorial structures detector (Andriluka et al., 2009)** without scale prior (first row) and with scale prior (second row). 109
6.8 Distribution of the height of false positive (left) and true positive (right) detections for PS (Andriluka et al., 2009) (top row) and DPM (Felzenszwalb et al., 2008) (bottom row) detectors. 110
6.9 Examples of detections at EER obtained with the detector combining upper- and full-body PS, and DPM models, and scale prior. 111
6.10 Missing recall (top row) and false positive (at ERR) detections (bottom row) of the detector combining upper- and full-body PS, and DPM models. 111

7.1 Estimated 2D body configurations (top row), and corresponding segmentations of the body parts (bottom row). The segmentations are obtained using color-based appearance models derived from the configuration estimates. 117

List of Tables

3.1 Recall of 4D-ISM, HOG, and our detectors at precision equal to 0.9 and at equal error rate on “TUD-Pedestrians” and “TUD-Campus” datasets. 40

4.1 **Evaluation of model parameters:** 2D pose estimation results on the **“People”** dataset for different image descriptors and discretizations of gradient orientation. 58

4.2 **Evaluation of model parameters:** 2D pose estimation results on the **“People”** dataset for varying numbers of boosting rounds. 59

4.3 **Evaluation of model parameters:** 2D pose estimation results on the **“People”** dataset for different step between evaluated part locations. 60

4.4 Comparison of models with and without repulsive factors on the **“TUD-UprightPeople”** dataset. 61

4.5 Comparison of models with and without repulsive factors on the **“People”** dataset. Numbers indicate the percentage of the correctly detected parts. The total number of part segments is $10 \times 205 = 2050$. Note that in this experiment part detectors are evaluated at the resolution of 1 px. in image and 7.5° in the orientation domain. 61

4.6 Comparison of sum-product and max-product belief propagation on **“People”** dataset. 62

4.7 **Upper body pose estimation:** Comparison of body part detection on the **“Buffy Stickmen”** dataset (numbers indicate the percentage of correctly localized body parts). Following the evaluation procedure only correctly localized people are taken into consideration. All experiments use the detector from (Ferrari et al., 2008), which can localize 85.1% of the annotated people. 70

4.8 **Upper body pose estimation:** Comparison of body part detection rates on the **“Buffy Stickmen”** dataset using the upper-body detector of Ferrari et al. (2008) for localization and our model only. . 70

4.9 **Upper body pose estimation:** Comparison of body part detection on the **"ETHZ PASCAL Stickmen"** dataset (Eichner and Ferrari, 2009) (numbers indicate the percentage of correctly localized parts). According to the evaluation procedure only correctly localized people are taken into consideration. All experiments use the detector of (Ferrari et al., 2008), which can localize 65.6% of annotated people. . 70

4.10 **Full body pose estimation:** Comparison of body part detection rates and evaluation of different components of the model on the **"People"** dataset. Numbers indicate the percentage of correctly localized parts. The total number of part segments is 10 × 205 = 2050. 74

5.1 Viewpoint estimation on the "MultiviewPeople" dataset. The task is to classify one of 8 viewpoints (chance level 12.5%). 85

5.2 Quantitative evaluation on the HumanEva II dataset (frames 1-350). We report mean error and standard deviation of the relative 2D and 3D joint positions. 2D results are in pixels and 3D results are in millimeters. 93

Bibliography

Agarwal, A. and Triggs, B. (2006), ‘Recovering 3d human pose from monocular images’, *IEEE T. Pattern Anal. Mach. Intell.* **28**(1), 44–58. 25

Andriluka, M., Roth, S. and Schiele, B. (2008), People-tracking-by-detection and people-detection-by-tracking, *in* ‘IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)’. 18, 20, 24, 25, 46, 48, 50, 51, 64, 65, 66, 67, 68, 87, 91, 117, 121

Andriluka, M., Roth, S. and Schiele, B. (2009), Pictorial structures revisited: People detection and articulated pose estimation, *in* ‘IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)’. 4, 15, 20, 24, 57, 58, 59, 61, 73, 74, 82, 83, 84, 85, 86, 99, 101, 102, 107, 108, 109, 110, 115, 119, 122, 123

Andriluka, M., Roth, S. and Schiele, B. (2010), Monocular 3d pose estimation and tracking by detection, *in* ‘IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)’. 23, 24, 25, 117

Arras, K., Grzonka, S., Luber, M. and Burgard, W. (2008), Eficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities, *in* ‘ICRA’. 26

Arras, K., Mozos, O. and Burgard, W. (2007), Using boosted features for the detection of people in 2d range data, *in* ‘ICRA’. 26

Balan, A. O., Black, M. J., Haussecker, H. and Sigal, L. (2007), Shining a light on human pose: On shadows, shading and the estimation of pose and shape, *in* ‘IEEE International Conference on Computer Vision (ICCV 2007)’. 118

Balan, A., Sigal, L., Black, M., Davis, J. and Haussecker, H. (2007), Detailed human shape and pose from images, *in* ‘IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)’. 19, 23

Barinova, O., Lempitsky, V. and Kohli, P. (2010), On detection of multiple object instances using hough transform, *in* ‘IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)’. 16

Belongie, S., Malik, J. and Puzicha, J. (2000), Shape context: A new descriptor for shape matching and object recognition, *in* 'Advances in Neural Information Processing Systems (NIPS*00)'. 20, 32, 51, 57, 83

Bergtholdt, M., Kappes, J., Schmidt, S. and Schnörr, C. (2009), 'A study of parts-based object class detection using complete graphs', *Int. J. Comput. Vision* **87**(1-2), 93–117. 22

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc. 12

Bouguet, J. Y. (2010), 'Camera calibration toolbox for Matlab', www.vision.caltech.edu/bouguetj. 100

Bourdev, L. and Malik, J. (2009), Poselets: Body part detectors trained using 3d human pose annotations, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 15, 17, 18, 99, 101, 102, 103, 106, 107, 122

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E. and Gool, L. V. (2009*a*), Markovian tracking-by-detection from a single, uncalibrated camera, *in* 'Proc. Int. IEEE CVPR Workshop on Performance Evaluation of Tracking and Surveillance (PETS'09)'. 19

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E. and Gool, L. V. (2009*b*), Robust tracking-by-detection using a detector confidence particle filter, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 19

Bruyninckx, H. (2001), Open robot control software: The OROCOS project, *in* 'ICRA', pp. 2523–2528. 100

Buehler, P., Everingham, M., Huttenlocher, D. P. and Zisserman, A. (2008), Long term arm and hand tracking for continuous sign language tv broadcasts, *in* 'British Machine Vision Conference (BMVC 2008)'. 47

Burl, M. C., Weber, M. and Perona, P. (1998), A probabilistic approach to object recognition using local photometry and global geometry, *in* 'European Conference on Computer Vision (ECCV 1998)', London, UK. 11

Carballo, A., Ohya, A. and Yuta, S. (2009), Multiple people detection from a mobile robot using double layered laser range finders, *in* 'ICRA Workshop on People Detection and Tracking'. 26

Cooper, J. and Goodrich, M. (2008), Towards combining UAV and sensor operator roles in UAV-enabled visual search, *in* 'International Conference on Human Robot Interaction', pp. 351–358. 97

Cootes, T. F., Taylor, C. J., Cooper, D. H. and Graham, J. (1995), 'Active shape models-their training and application', *Computer Vision and Image Understanding* **61**(1), 38–59. 14

Crandall, D., Felzenszwalb, P. F. and Huttenlocher, D. (2005), Spatial priors for part-based recognition using statistical models, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)'. 21, 54

Dalal, N. and Triggs, B. (2005), Histograms of oriented gradients for human detection, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)', San Diego, CA, USA. 5, 14, 15, 28, 30, 34, 38, 57, 66, 67, 98, 99, 101, 106, 107, 108, 120, 121, 122

Dalal, N., Triggs, B. and Schmid, C. (2006), Human detection using oriented histograms of flow and appearance, *in* 'ECCV'. 15, 98

Darby, J., Li, B., Costens, N., Fleet, D. and Lawrence, N. (2009), Backing off: Hierarchical decomposition of activity for 3d novel pose recovery, *in* 'British Machine Vision Conference (BMVC 2009)'. 117

Davis, J. and Sharma, V. (2004), Robust detection of people in thermal imagery, *in* 'ICPR'. 26

Demirdjian, D., Taycher, L., Shakhnarovich, G., Grauman, K. and Darrell, T. (2005), Avoiding the "streetlight effect": Tracking by exploring likelihood modes, *in* 'IEEE International Conference on Computer Vision (ICCV 2005)'. 29, 35

Deutscher, J. and Reid, I. (2005), 'Articulated body motion capture by stochastic search', *Int. J. Comput. Vision* **61**, 185–205. 23, 29, 35

Doherty, P. and Rudol, P. (2007), 'A UAV search and rescue scenario with human body detection and geolocalization', *Australian Joint Conference on Artificial Intelligence* **4830**, 1. 26

Dollár, P., Tu, Z., Perona, P. and Belongie, S. (2009), Integral channel features, *in* 'British Machine Vision Conference (BMVC 2009)'. 15, 115

Dollar, P., Wojek, C., Schiele, B. and Perona, P. (2009*a*), Pedestrian detection: A benchmark, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 15

Dollár, P., Wojek, C., Schiele, B. and Perona, P. (2009*b*), Pedestrian detection: A benchmark, *in* 'CVPR'. 98, 101

Eichner, M. and Ferrari, V. (2009), Better appearance models for pictorial structures, *in* 'British Machine Vision Conference (BMVC 2009)'. 19, 56, 68, 69, 70, 71, 72, 73, 84, 115, 116, 126

Eichner, M. and Ferrari, V. (2010), We are family: Joint pose estimation of multiple persons, *in* 'European Conference on Computer Vision (ECCV 2010)'. 5

Enzweiler, M. and Gavrila, D. M. (2009), 'Monocular pedestrian detection: Survey and experiments', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**, 2179–2195. 15

Ess, A., Leibe, B., Schindler, K. and Van Gool, L. (2009), Moving Obstacle Detection in Highly Dynamic Scenes, *in* 'ICRA'. 98

Ess, A., Leibe, B. and van Gool, L. (2007), Depth and appearance for mobile scene analysis, *in* 'IEEE International Conference on Computer Vision (ICCV 2007)'. 15

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. (n.d.), The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html. 68

Felzenszwalb, P. (2001), Learning models for object recognition, *in* 'CVPR', Kauai, Hawaii. 14

Felzenszwalb, P. F., Girshick, R., McAllester, D. and Ramanan, D. (2009), 'Object detection with discriminatively trained part based models', *IEEE T. Pattern Anal. Mach. Intell.* . 16, 17, 18, 101, 107, 119, 122

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004), Distance transforms of sampled functions, Technical Report TR2004-1963, Cornell University. 12

Felzenszwalb, P. F. and Huttenlocher, D. P. (2005), 'Pictorial structures for object recognition', *Int. J. Comput. Vision* **61**, 55–79. 5, 6, 11, 12, 13, 19, 21, 29, 30, 31, 32, 45, 46, 48, 49, 51, 54, 83, 84, 102, 113

Felzenszwalb, P. F., McAllester, D. and Ramanan, D. (2008), A discriminatively trained, multiscale, deformable part model, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)'. 15, 20, 48, 84, 85, 97, 99, 102, 108, 110, 123

Fergus, R., Perona, P. and Zisserman, A. (2003), Object class recognition by unsupervised scale-invariant learning, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)'. 11

Ferrari, V., Marin, M. and Zisserman, A. (2008), Progressive search space reduction for human pose estimation, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)'. 5, 19, 25, 46, 50, 56, 61, 68, 69, 70, 71, 72, 73, 93, 98, 99, 101, 105, 106, 113, 116, 125, 126

Ferrari, V., Marin, M. and Zisserman, A. (2009*a*), 2d human pose estimation in tv shows, *in* 'Proceedings of the Dagstuhl Seminar on Stastistical and Geometrical Approaches to Visual Motion Analysis'. 69, 70, 71

Ferrari, V., Marin, M. and Zisserman, A. (2009*b*), Pose search: retrieving people using their pose, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 19, 47, 55, 56, 69, 70, 71, 72, 73

Fischler, M. and Elschlager, R. (1973), 'The representation and matching of pictorial structures', *IEEE Transactions on Computers* **C-22**(1), 67–92. 5, 10, 11, 19, 30, 45, 46, 54, 119

Fod, A., Howard, A. and Mataric, M. (2002), Laser-based people tracking, *in* 'ICRA'. 26

Forsyth, D., Arikan, O., Ikemoto, L., OŠBrien, J. and Ramanan, D. (2006), *Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis*, Vol. 1 of *Foundations Trends in Computer Graphics and Vision*, Now Publishers Inc. 24

Fossati, A., Dimitrijevic, M., Lepetit, V. and Fua, P. (2007), Bridging the gap between detection and tracking for 3D monocular video-based motion capture, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)'. 23, 24, 25, 117

Freifeld, O., Weiss, A., Zuffi, S. and Black, M. (2010), Contour people: A parameterized model of 2d articulated human shape, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 5

Freund, Y. and Schapire, R. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *J. of Comp. and Sys. Sc.* **55**(1), 119–139. 14, 20, 52, 83

Frey, B. J. (1998), *Graphical Models for Machine Learning and Digital Communication*, MIT Press. 12

Gall, J. and Lempitsky, V. (2009), Class-specific hough forests for object detection, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 5, 16, 20, 66, 67, 121

Gall, J., Rosenhahn, B., Brox, T. and Seidel, H.-P. (2010), 'Optimization and filtering for human motion capture: A multi-layer framework', *Int. J. Comput. Vision* . 23

Gammeter, S., Ess, A., Jaeggli, T., Schindler, K., Leibe, B. and Gool, L. (2008), Articulated multi-body tracking under egomotion, *in* 'European Conference on Computer Vision (ECCV 2008)'. 23, 24, 92, 94, 117, 122

Gate, G., Breheret, A. and Nashashibi, F. (2009), Centralized fusion for fast people detection in dense environment, *in* 'ICRA'. 26, 98

Gavrila, D. (1999), Real-time object detection for smart vehicles, *in* 'IEEE International Conference on Computer Vision (ICCV 1999)'. 13

Gavrila, D. (2000), Pedestrian detection from a moving vehicle, *in* 'European Conference on Computer Vision (ECCV 2000)', Springer, pp. 37–49. 13

Grabner, H., Leistner, C. and Bischof, H. (2008), Semi-supervised on-line boosting for robust tracking, *in* 'European Conference on Computer Vision (ECCV 2008)'. 18

Green, W., Sevcik, K. and Oh, P. (2005), A competition to identify key challenges for unmanned aerial robots in near-earth environments, *in* 'ICAR'. 97

Grochow, K., Martin, S. L., Hertzmann, A. and Popovic, Z. (2004), Style-based inverse kinematics, *in* 'ACM SIGGRAPH'. 37

Guan, P., Weiss, A., Balan, A. and Black, M. J. (2009), Estimating human shape and pose from a single image, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 76

Hasler, N., Rosenhahn, B., Thormaehlen, T., Wand, M. and Seidel, H.-P. (2009), Markerless motion capture with unsynchronized moving cameras, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 23

Heikkila, J. and Silven, O. (1997), A four-step camera calibration procedure with implicit image correction, *in* 'CVPR', IEEE Computer Society, Washington, DC, USA, p. 1106. 104

Hoffmann, G., Huang, H., Waslander, S. and Tomlin, C. (2007), Quadrotor helicopter flight dynamics and control: Theory and experiment, *in* 'AIAA Guidance, Navigation and Control Conference'. 99

Ionescu, C., Bo, L. and Sminchisescu, C. (2009), Structural svm for visual localization and continuous state estimation, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 25

Isard, M. (2003), Pampas: Real-valued graphical models for computer vision, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)', pp. 613–620. 12

Isard, M. and MacCormick, J. (2001), BraMBLe: A Bayesian multiple-blob tracker, *in* 'IEEE International Conference on Computer Vision (ICCV 2001)'. 19

Isard, M., MacCormick, J. and Achan, K. (2008), Continuously-adaptive discretization for message-passing algorithms, *in* 'Advances in Neural Information Processing Systems (NIPS*08)'. 116

Jacobs, R., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991), 'Adaptive mixtures of local experts', *Neural Computation* **3**, 79–87. 21

Jiang, H. (2009), Human pose estimation using consistent max-covering, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 47

Jiang, H. and Martin, D. R. (2008), Global pose estimation using non-tree models, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)'. 22

Jie, L., Caputo, B. and Ferrari, V. (2009), WhoŠs doing what: Joint modeling of names and verbs for simultaneous face and pose annotation, *in* 'Advances in Neural Information Processing Systems (NIPS*09)'. 22

Johnson, S. and Everingham, M. (2009), Combining discriminative appearance and segmentation cues for articulated human pose estimation, *in* '2nd IEEE International Workshop on Machine Learning for Vision-based Motion Analysis, in conjunction with ICCV 2009'. 56, 57, 58, 73, 74

Johnson, S. and Everingham, M. (2010), Clustered pose and nonlinear appearance models for human pose estimation, *in* 'British Machine Vision Conference (BMVC 2010)'. 63

Jordan, M. I., Ghahramani, Z., Jaakkola, T. and Saul, L. K. (1999), 'An introduction to variational methods for graphical models', *Machine Learning* **37**(2), 183–233. 12

Ju, S. X., Black, M. J. and Yacoob, Y. (1996), Efficient inference with multiple heterogeneous part detectors for human pose estimation, *in* 'International Conference on Automatic Face- and Gesture-Recognition (FG'96)'. 115

Karlinsky, L., Dinerstein, M., Harari, D. and Ullman, S. (2010), The chains model for detecting parts by their context, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 5

Kleiner, A. and Kuemmerle, R. (2007), Genetic MRF model optimization for real-time victim detection in Search and Rescue, *in* 'IROS'. 26

Kschischang, F. R., Frey, B. J. and Loelinger, H.-A. (2001), 'Factor graphs and the sum-product algorithm', *IEEE T. Info. Theory* **47**(2), 498–519. 54

Kuo, C.-H., Huang, C. and Nevatia, R. (2010), Multi-target tracking by on-line learned discriminative appearance models, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 19

Lan, X. and Huttenlocher, D. P. (2005), Beyond trees: Common-factor models for 2d human pose recovery, *in* 'IEEE International Conference on Computer Vision (ICCV 2005)'. 20, 30, 49

Lawrence, N. D. (2005), 'Probabilistic non-linear principal component analysis with Gaussian process latent variable models', *J. Mach. Learn. Res.* **6**, 1783–1816. 36

Lawrence, N. D. and Moore, A. J. (2007), Hierarchical Gaussian process latent variable models, *in* 'International Conference on Machine Learning (ICML 2007)'. 5, 24, 28, 36, 82, 90

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989), 'Backpropagation applied to handwritten zip code recognition', *Neural Computation* **1**(4), 541–551. 14

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2324. 14

Lee, H.-J. and Chen, Z. (1985), 'Determination of 3d human body postures from a single view', *CVGIP* **30**, 148–168. 76

Lee, M. W. and Cohen, I. (2004), Proposal maps driven MCMC for estimating human body pose in static images, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)'. 21, 22, 116

Leibe, B., Leonardis, A. and Schiele., B. (2004), Combined object categorization and segmentation with an implicit shape model, *in* 'ECCV Workshop on statistical learning in computer vision', pp. 17–32. 5, 6, 16, 103

Leibe, B., Leonardis, A. and Schiele, B. (2008), 'Robust object detection with interleaved categorization and segmentation', *International Journal of Computer Vision* **77**(1), 259–289. 16, 31, 119

Leibe, B., Schindler, K. and Van Gool, L. (2007), Coupled detection and trajectory estimation for multi-object tracking, *in* 'IEEE International Conference on Computer Vision (ICCV 2007)'. 18, 19

Leibe, B., Seemann, E. and Schiele, B. (2005), Pedestrian detection in crowded scenes, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)'. 15, 16, 17, 18, 20, 28, 29, 46

Levinshtein, A., Dickinson, S. and Sminchisescu, C. (2009), Multiscale symmetric part detection and grouping, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 116

Lowe, D. G. (2004), 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vision* **60**(2), 91–110. 14, 20, 51, 57

Maji, S. and Malik, J. (2009*a*), Object detection using a max-margin hough tranform, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 103

Maji, S. and Malik, J. (2009*b*), Object detection using a max-margin hough transform, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 16

Meyer, J., Schnitzspan, P., Kohlbrecher, S., Petersen, K., Schwahn, O., Andriluka, M., Klingauf, U., Roth, S., Schiele, B. and von Stryk, O. (2010), A semantic world model for urban search and rescue based on heterogeneous sensors, *in* 'RoboCup Symposium, Singapore'. 26

Meyer, J. and Strobel, A. (2010), A flexible real-time control system for autonomous vehicles, *in* 'ISR / ROBOTIK'. 100

Mikolajczyk, K., Leibe, B. and Schiele, B. (2006), Multiple object class detection with a generative model, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)'. 20, 46, 51

Mikolajczyk, K. and Schmid, C. (2004), 'Scale and affine invariant interest point detectors', *Int. J. Comput. Vision* **60**, 63–86. 32

Mikolajczyk, K. and Schmid, C. (2005), 'A performance evaluation of local descriptors', *IEEE T. Pattern Anal. Mach. Intell.* **27**(10), 1615–1630. 14, 20, 57, 84

Mohan, A., Papageorgiu, C. and Poggio, T. (2001), 'Example-based object detection in images by componentes', *IEEE T. Pattern Anal. Mach. Intell.* **23**(4), 349–361. 15

Mooij, J. M. (2009), 'libDAI 0.2.2: A free/open source C++ library for Discrete Approximate Inference', http://www.libdai.org/. 55

Mori, G. and Malik, J. (2006), 'Recovering 3d human body configurations using shape contexts', *IEEE T. Pattern Anal. Mach. Intell.* **28**(7), 1052–1062. 25, 89

Murphy, K. (2005), Models for generic visual object detection, Technical report, University of British Columbia. 15

Murphy, R. (2004), 'Human-robot interaction in rescue robotics', *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **34**(2), 138–153. 98

Nordberg, K., Doherty, P., Farnebäck, G., Forssen, P.-E., Granlund, G., Moe, A. and Wiklund, J. (2002), Vision for a uav helicopter, *in* 'Proceedings of IROS'02, Workshop on aerial robotics'. 97

Okuma, K., Taleghani, A., De Freitas, N., Little, J. J. and Lowe, D. G. (2004), A boosted particle filter: Multitarget detection and tracking, *in* 'European Conference on Computer Vision (ECCV 2004)'. 18

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E. and Poggio, T. (1997), Pedestrian detection using wavelet templates, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 1997)'. 14, 15

Ormoneit, D., Sidenbladh, H., Black, M. J. and Hasti, T. (2001), Learning and tracking cyclic human motion, *in* 'Advances in Neural Information Processing Systems (NIPS*01)'. 6

Papageorgiou, C. and Poggio, T. (1999), Trainable pedestrian detection, *in* 'International Conference on Image Processing (ICIP-99)', pp. 35–39. 14

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 2nd edn, Morgan Kaufmann, San Francisco, California. 12, 54

Pham, Q., Gond, L., Begard, J., Allezard, N. and Sayd, P. (2007), Real-time posture analysis in a crowd using thermal imaging, *in* 'CVPR'. 26

Ramanan, D. (2006), Learning to parse images of articulated objects, *in* 'Advances in Neural Information Processing Systems (NIPS*06)'. 5, 6, 19, 46, 47, 48, 50, 53, 54, 56, 57, 58, 59, 61, 73, 74, 75, 76, 93, 113, 116, 120, 121

Ramanan, D., Forsyth, D. A. and Zisserman, A. (2007), 'Tracking people by learning their appearance', *IEEE T. Pattern Anal. Mach. Intell.* **29**, 65–81. 20, 35

Ramanan, D. and Sminchisescu, C. (2006), Training deformable models for localization, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)'. 22, 46, 49

Ren, X., Berg, A. C. and Malik, J. (2005), Recovering human body configurations using pairwise constraints between parts, *in* 'IEEE International Conference on Computer Vision (ICCV 2005)'. 22

Rogez, G., Rihan, J., Ramalingam, S., Orrite, C. and Torr, P. H. (2008), Randomized trees for human pose detection, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)'. 25, 92, 93

Ronfard, R., Schmid, C. and Triggs, B. (2002), Learning to parse pictures of people, *in* 'European Conference on Computer Vision (ECCV 2002)'. 19

Roth, S. and Black, M. J. (2009), 'Fields of experts', *Int. J. Comput. Vision* **82**(2), 205–229. 52

Rother, C., Kolmogorov, V. and Blake, A. (2004), '"grabcut": interactive foreground extraction using iterated graph cuts', *ACM Trans. Graph.* **23**, 309–314. 69

Sapp, B., Jordan, C. and Taskar, B. (2010), Adaptive pose priors for pictorial structures, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 5, 21, 53

Sapp, B., Toshev, A. and Taskar, B. (2010), Cascaded models for articulated pose estimation, *in* 'European Conference on Computer Vision (ECCV 2010)'. 5

Schiele, B. (2005), 'Model-free tracking of cars and people based on color regions', *Image and Vision Computing* **24**(11), 1172–1178. 18

Schindler, K., van Gool, L. and de Gelder, B. (2008), 'Recognizing emotions expressed by body pose: a biologically inspired neural model', *Neural Networks* **21**(9), 1238–1246. 22

Schulz, D., Burgard, W., Fox, D. and Cremers., A. (2003), 'People tracking with mobile robots using sample-based joint probabilistic data association', *IJRR* **22**(2). 26

Schwarz, L. A., Mateus, D. and Navab, N. (2010), Multiple-activity human body tracking in unconstrained environments, *in* 'AMDO'2010'. 117

Seemann, E. and Schiele, B. (2006), Cross-articulation learning for robust detection of pedestrians, *in* 'DAGM'. 5, 16, 30, 34, 38, 113, 114, 120

Shakhnarovich, G., Viola, P. A. and Darrell, T. (2003), Fast pose estimation with parameter-sensitive hashing, *in* 'IEEE International Conference on Computer Vision (ICCV 2003)'. 25, 89

Sigal, L., Bhatia, S., Roth, S., Black, M. J. and Isard, M. (2004), Tracking loose-limbed people, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)'. 6

Sigal, L. and Black, M. (2006*a*), Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Technical report, Brown University. 92

Sigal, L. and Black, M. J. (2006*b*), Measure locally, reason globally: Occlusion-sensitive articulated pose estimation, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)'. 22, 23, 35, 51, 76, 116

Sigal, L. and Black, M. J. (2006*c*), Predicting 3D people from 2D pictures, *in* 'AMDO 2006'. 21, 22, 24, 50, 89

Singh, V. K., Nevatia, R. and Huang, C. (2010), Efficient inference with multiple heterogeneous part detectors for human pose estimation, *in* 'European Conference on Computer Vision (ECCV 2010)'. 115

Sminchisescu, C., Kanaujia, A. and Metaxas, D. N. (2007), 'BM^{3}E: Discriminative density propagation for visual tracking', *IEEE T. Pattern Anal. Mach. Intell.* **29**, 2030–2044. 29, 35, 36

Smolensky, P. (1986), Information processing in dynamical systems: Foundations of harmony theory, *in* D. E. Rumelhart, J. L. McClelland and P. R. Group, eds, 'Parallel Distributed Processing: Volume 1: Foundations', MIT Press, Cambridge, pp. 194–281. 118

Spinello, L., Triebel, R. and Siegwart, R. (2008), Multimodal people detection and tracking in crowded scenes, *in* 'AAAI'. 98

Stark, M., Goesele, M. and Schiele, B. (2010*a*), Back to the future: Learning shape models from 3d cad data, *in* 'British Machine Vision Conference (BMVC 2010)'. 5

Stark, M., Goesele, M. and Schiele, B. (2010*b*), A shape-based object class model for knowledge transfer, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 12

Suard, F., Rakotomamonjy, A., Benshrair, A. and Broggi, A. (2006), Pedestrian detection using infrared images and histograms of oriented gradients, *in* 'IEEE Symposium on Intelligent Vehicle'. 26

Sudderth, E. B., Ihler, A. T., Freeman, W. T. and Willsky, A. S. (2003), Nonparametric belief propagation, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)', pp. 605–612. 12

Sudderth, E. B., Mandel, M. I., Freeman, W. T. and Willsky, A. S. (2004), Distributed occlusion reasoning for tracking with nonparametric belief propagation, *in* 'Advances in Neural Information Processing Systems (NIPS*04)'. 47, 116

Taylor, C. J. (2000), 'Reconstruction of articulated objects from point correspondences in a single uncalibrated image', *Comput. Vis. Image Und.* **80**, 349–363. 77

Taylor, G., Sigal, L., Fleet, D. and Hinton, G. (2010), Dynamical binary latent variable models for 3d human pose tracking, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 118

Tian, T.-P. and Sclaroff, S. (2010*a*), Fast globally optimal 2d human detection with loopy graph models, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 5

Tian, T.-P. and Sclaroff, S. (2010*b*), Fast multi-aspect 2d human detection, *in* 'European Conference on Computer Vision (ECCV 2010)'. 5

Titterton, D. and Weston, J. (2004), *Strapdown inertial navigation technology*, American Institute of Aeronautics, Washington, DC. 100

Tu, Z., Chen, X., Yuille, A. L. and Zhu, S.-C. (2005), 'Image parsing: Unifying segmentation, detection, and recognition', *Int. J. Comput. Vision* **63**(2), 113–140. 46, 50, 82

Urtasun, R. and Darrell, T. (2008), Sparse probabilistic regression for activity-independent human pose inference, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)'. 21, 25

Urtasun, R., Fleet, D. J. and Fua, P. (2006), 3D people tracking with Gaussian process dynamical models, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)'. 24, 35, 36, 49, 117

Vapnik, V. (1999), *The Nature of Statistical Learning Theory (Information Science and Statistics)*, Springer. 14

Villamizar, M., Scandaliaris, J., Sanfeliu, A. and Andrade-Cetto, J. (2009), Combining color-based invariant gradient detector with HoG descriptors for robust mage detection in scenes under cast shadows, *in* 'ICRA'. 98

Viola, P., Jones, M. and Snow, D. (2003), Detecting pedestrians using patterns of motion and appearance, *in* 'IEEE International Conference on Computer Vision (ICCV 2003)', pp. 734–741. 14, 20, 46

Vondrak, M., Sigal, L. and Jenkins, O. C. (2008), Physical simulation for probabilistic motion tracking, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)'. 23

Walk, S., Majer, N., Schindler, K. and Schiele, B. (2010), New features and insights for pedestrian detection, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 115

Walk, S., Schindler, K. and Schiele, B. (2010), Disparity statistics for pedestrian detection: combining appearance, motion, and stereo, *in* 'European Conference on Computer Vision (ECCV 2010)'. 15

Wang, J. M., Fleet, D. J. and Hertzmann, A. (2005), Gaussian process dynamical models, *in* 'Advances in Neural Information Processing Systems (NIPS*05)'. 36

Wang, Y. and Mori, G. (2008), Multiple tree models for occlusion and spatial constraints in human pose estimation, *in* 'European Conference on Computer Vision (ECCV 2008)'. 47, 116

Williams, C. K. I. and Allan, M. (2006), On a connection between object localization with a generative template of features and pose-space prediction methods, Technical Report EDI-INF-RR-0719, University of Edinburgh. 31

Wojek, C., Walk, S. and Schiele, B. (2009), Multi-cue onboard pedestrian detection, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)'. 15, 84, 85, 98, 101

Wu, B. and Nevatia, R. (2007), 'Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors', *Int. J. Comput. Vision* **75**, 247–266. 18, 87

Yang, W., Wang, Y. and Mori, G. (2010), Recognizing human actions from still images with latent poses, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 22

Yang, Y., Hallman, S., Ramanan, D. and Fowlkes, C. (2010), Layered object detection for multi-class segmentation, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)'. 116

Yao, B. and Fei-Fei, L. (2010), Modeling mutual context of object and human pose in human-object interaction activities, *in* 'European Conference on Computer Vision (ECCV 2010)'. 5, 22, 23, 116

Yedidia, J., Freeman, W. and Weiss, Y. (2000), Generalized belief propagation, *in* 'Advances in Neural Information Processing Systems (NIPS*00)'. 12

Yilmaz, A. and Shah, O. J. M. (2006), 'Object tracking: A survey', *ACM Computing Surveys* **38**(4). 18

Zhang, J., Luo, J., Collins, R. and Liu, Y. (2006), Body localization in still images using hierarchical models and hybrid search, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)'. 46

Zhang, X., Li, C., Tong, X., Hu, W., Maybank, S. and Zhang, Y. (2009), Efficient human pose estimation via parsing a tree structure based human model, *in* 'IEEE International Conference on Computer Vision (ICCV 2009)'. 47

Curriculum Vitae

	Mykhaylo Andriluka
Born in	Odessa, Ukraine
Citizenship:	ukrainian

Education:	2006–2010	**TU Darmstadt, Germany** PhD student at the Multimodal Interactive Systems Group of Prof. B. Schiele
	2000–2006	**TU Darmstadt, Germany** Studies of Mathematics with Computer Science, graduation with degree *Dipl.-Math.* Minor subjects: Machine Learning, Spline Approximation Diploma thesis: Multi-output Gaussian Process Regression, supervised by Lorenz Weizsaecker and Prof. Thomas Hofmann
Experience:	2006–2010	Research and teaching assistant, Multimodal Interactive Systems Group, TU Darmstadt, Germany.
	2001–2006	Software developer at “Advancis Software and Services”.
	2001	Teaching assistant at TU Darmstadt.
	2000–2001	Student research assistant at the Center of Computer Graphics (ZGDV) Darmstadt, division “Visual Computing”, supervised by Vitor Sa.

Publications

Monocular 3D Pose Estimation and Tracking by Detection
M. Andriluka, S. Roth and B. Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), San Francisco, USA, 2010

Vision Based Victim Detection from Unmanned Aerial Vehicles
M. Andriluka, P. Schnitzspan, J. Meyer, S. Kohlbrecher, K. Petersen, O. von Stryk, S. Roth and B. Schiele
In IEEE/RSJ In International Conference on Intelligent Robots and Systems (**IROS**), Taipei, Taiwan, 2010

Categorical Perception
M. Fritz, M. Andriluka, S. Fidler, M. Stark, A. Leonardis and B. Schiele
In Categorical Perception, Cognitive Systems Monographs (8), Springer, 2010

A Semantic World Model for Urban Search and Rescue Based on Heterogeneous Sensors
J. Meyer, P. Schnitzspan, S. Kohlbrecher, K. Petersen, O. Schwahn, M. Andriluka, U. Klingauf, S. Roth, B. Schiele and O. von Stryk
In **RoboCup Symposium**, Singapore, 2010

Pictorial Structures Revisited: People Detection and Articulated Pose Estimation
M. Andriluka, S. Roth and B. Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), Miami, USA, 2009

Visual People Detection: Different Models, Comparison and Discussion
B. Schiele, M. Andriluka, N. Majer, S. Roth and C. Wojek
In IEEE ICRA 2009 Workshop on People Detection and Tracking, Kobe, Japan, 2009

People-Tracking-by-Detection and People-Detection-by-Tracking
M. Andriluka, S. Roth and B. Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), Anchorage, USA, 2008

Multi-class Classification with Dependent Gaussian Processes
M. Andriluka, L. Weizsäcker and T. Hofmann
In International Conference on Applied Stochastic Models and Data Analysis (**ASMDA**), Crete, Greece, 2007