Articulated People Detection and Pose Estimation in Challenging Real World Environments

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

by
Leonid Pishchulin, M.Sc.

Saarbrücken
May 2016

Day of Colloquium	31st of May, 2016
Dean of the Faculty	Univ.-Prof. Dr. Frank-Olaf Schreyer Saarland University, Germany
Chair of the Committee	Prof. Dr. Matthias Hein Saarland University, Germany

Reporters

First Reviewer, Advisor	Prof. Dr. Bernt Schiele Max Planck Institute for Informatics, Germany Saarland University, Germany
Second Reviewer	Prof. Dr. Christian Theobalt Max Planck Institute for Informatics, Germany
Third Reviewer	Prof. Jitendra Malik, Ph.D. University of California at Berkeley
Academic Assistant	Dr. Björn Andres Max Planck Institute for Informatics, Germany

ABSTRACT

In this thesis we are interested in the problem of articulated people detection and pose estimation being key ingredients towards understanding visual scenes containing people. Although extensive efforts are being made to address these problems, we identify three promising directions that, we believe, didn't get sufficient attention recently.

First, we investigate how statistical 3D human shape models from computer graphics can be leveraged to ease training data generation. We propose a range of automatic data generation techniques that allow to directly represent relevant variations in the training data. Sampling from both the underlying human shape distribution and a large dataset of human poses allows to generate novel samples with controllable shape and pose variations that are relevant for the task at hand. Furthermore, we improve the state-of-the-art 3D human shape model itself by rebuilding it from a large commercially available dataset of 3D bodies.

Second, we develop expressive spatial and strong appearance models for 2D single- and multi-person pose estimation. We propose an expressive single person model that incorporates higher order part dependencies while remaining efficient. We augment this model with various types of strong appearance representations aiming to substantially improve the body part hypotheses. Finally, we propose an expressive model for joint pose estimation of multiple people. To that end, we develop strong deep learning based body part detectors and an expressive fully connected spatial model. The proposed approach treats multi-person pose estimation as a joint partitioning and labeling problem of a set of body part hypotheses: it infers the number of persons in a scene, identifies occluded body parts and disambiguates body parts between people in close proximity of each other.

Third, we perform thorough evaluation and performance analysis of leading human pose estimation and activity recognition methods. To that end we introduce a novel benchmark that makes a significant advance in terms of diversity and difficulty, compared to the previous datasets, and includes over $40,000$ annotated body poses and over $1.5M$ frames. Furthermore, we provide a rich set of labels which are used to perform a detailed analysis of competing approaches gaining insights into successes and failures of these methods.

In summary, this thesis presents a novel approach to articulated people detection and pose estimation. Thorough experimental evaluation on standard benchmarks demonstrates significant improvements due to the proposed data augmentation techniques and novel body models, while detailed performance analysis of competing approaches on our newly introduced large-scale benchmark allows to identify the most promising directions of improvement.

ZUSAMMENFASSUNG

In dieser Arbeit untersuchen wir das Problem der artikulierten Detektion und Posenschätzung von Personen als Schlüsselkomponenten des Verstehens von visuellen Szenen mit Personen. Obwohl es umfangreiche Bemühungen gibt, die Lösung dieser Probleme anzugehen, haben wir drei vielversprechende Herangehensweisen ermittelt, die unserer Meinung nach bisher nicht ausreichend beachtet wurden.

Erstens untersuchen wir, wie statistische 3D Modelle des menschlichen Umrisses, die aus der Computergrafik stammen, wirksam eingesetzt werden können, um die Generierung von Trainingsdaten zu erleichtern. Wir schlagen eine Reihe von Techniken zur automatischen Datengenerierung vor, die eine direkte Repräsentation relevanter Variationen in den Trainingsdaten erlauben. Indem wir Stichproben aus der zu Grunde liegenden Verteilung des menschlichen Umrisses und aus einem großen Datensatz von menschlichen Posen ziehen, erzeugen wir eine neue für unsere Aufgabe relevante Auswahl mit regulierbaren Variationen von Form und Posen. Darüber hinaus verbessern wir das neueste 3D Modell des menschlichen Umrisses selbst, indem wir es aus einem großen handelsüblichen Datensatz von 3D Körpern neu aufbauen.

Zweitens entwickeln wir ausdrucksstarke räumliche Modelle und Erscheinungsbild-Modelle für die 2D Posenschätzung einzelner und mehrerer Personen. Wir schlagen ein ausdrucksstarkes Einzelperson-Modell vor, das Teilabhängigkeiten höherer Ordnung einbezieht, aber dennoch effizient bleibt. Wir verstärken dieses Modell durch verschiedene Arten von starken Erscheinungsbild-Repräsentationen, um die Körperteilhypothesen erheblich zu verbessern. Schließlich schlagen wir ein ausdruckstarkes Modell zur gemeinsamen Posenschätzung mehrerer Personen vor. Dazu entwickeln wir starke Deep Learning-basierte Körperteildetektoren und ein ausdrucksstarkes voll verbundenes räumliches Modell. Der vorgeschlagene Ansatz behandelt die Posenschätzung mehrerer Personen als ein Problem der gemeinsamen Aufteilung und Annotierung eines Satzes von Körperteilhypothesen: er erschließt die Anzahl von Personen in einer Szene, identifiziert verdeckte Körperteile und unterscheidet eindeutig Körperteile von Personen, die sich nahe beieinander befinden.

Drittens führen wir eine gründliche Bewertung und Performanzanalyse führender Methoden der menschlichen Posenschätzung und Aktivitätserkennung durch. Dazu stellen wir einen neuen Benchmark vor, der einen bedeutenden Fortschritt bezüglich Diversität und Schwierigkeit im Vergleich zu bisherigen Datensätzen mit sich bringt und über 40.000 annotierte Körperposen und mehr als 1.5 Millionen Einzelbilder enthält. Darüber hinaus stellen wir einen reichhaltigen Satz an Annotierungen zur Verfügung, die zu einer detaillierten Analyse konkurrierender Herangehensweisen benutzt werden, wodurch wir Erkenntnisse zu Erfolg und Mißerfolg dieser Methoden erhalten.

Zusammengefasst präsentiert diese Arbeit einen neuen Ansatz zur artikulierten

Detektion und Posenschätzung von Personen. Eine gründliche experimentelle Evaluation auf Standard-Benchmarkdatensätzen zeigt signifikante Verbesserungen durch die vorgeschlagenen Datenverstärkungstechniken und neuen Körpermodelle, während eine detaillierte Performanzanalyse konkurrierender Herangehensweisen auf unserem neu vorgestellten großen Benchmark uns erlaubt, die vielversprechendsten Bereiche für Verbesserungen zu erkennen.

ACKNOWLEDGMENTS

This thesis would not be possible without the support of many people. First and foremost, I would like to thank my supervisor, Prof. Bernt Schiele, for his invaluable advice, and for being a constant source of motivation and inspiration throughout my PhD work. I'm grateful for his confidence in my abilities and potential from the very first days. He gave me an excellent PhD thesis topic which I'm very excited about, always provided timely and high quality feedback on my work. I'm grateful for his involvement and his help in spotting the weak points in my formulations and identifying the most promising directions of improvement. I'm thankful to him for creating a great research environment which I was lucky to be a part of.

I would like to thank Prof. Jitendra Malik for agreeing to serve as an external reviewer as part of the thesis committee. I would also like to thank Prof. Christian Theobalt for being part of the thesis committee.

I'm truly grateful to two colleagues of mine I was lucky to work with, Dr. Mykhaylo Andriluka and Dr. Peter Gehler. I would like to thank Micha for his energy and support. The ability to build on his great work helped to significantly speed up my learning curve. He was always there, ready to talk and help with advice, discuss new ideas and answer technical questions. I learned from him to never give up and fight until there is a small chance to improve the work, which was very valuable during the deadlines. I'm grateful to Peter for his fresh ideas and completely different perspectives on the problems we have been trying to address. From him I learned a lot about graphical models and powerful ways of leveraging them to address the challenges of human pose estimation. Numerous fruitful discussions during our meetings inspired many ideas of this thesis.

I'm grateful to further people I had a chance to collaborate with. I would like to thank Eldar Insafutdinov, for his work on developing fully-convolutional deep learning based body part detection models; I would like to thank Arjun Jain, for his help in leveraging computer graphics to make it serve people detection and pose estimation. I'm grateful to Dr. Stefanie Wuhrer, for her guidance and feedback on human body shape modeling. I would like to thank Dr. Christian Wojek from whom I learned a lot about people detection. I'm also grateful to Dr. Björn Andres and Siyu Tang for their help in marrying multicuts and human pose estimation. I would like to thank Ahmed Elhayek and Prof. Christian Theobalt who discovered the ways of leveraging strong 2D human pose estimation models to improve markerless 3D body motion capture.

Furthermore, I would like to thank my colleagues for sharing their code and having fruitful discussions on many topics discussed in this thesis: Dr. Michael Stark - for code and discussions on 3D rendering; Marcus Rohrbach - for code and advice on human activity recognition; Sikandar Amin - for sharing the code for body part tracking in videos and discussions on human pose estimation; Dr. Rodrigo Benenson

- for sharing his pedestrian detection code and many interesting discussions; Bojan Pepik - for many fruitful discussions on object detection; Mohamed Omran - for his advice and discussions regarding the Caffe deep learning framework. I would like to thank all my colleagues for their interest and active participation in discussions of my work at numerous group research seminars and group retreats, for their valuable feedback and many useful suggestions.

I would like to thank the students I was working with: Gautham Adithya who implemented 3D scan registration techniques; Mónica Vidriales who implemented statistical 3D shape model quality measures; Philipp Flotho who was working on people detection from reverse parking camera; Bahar Tarakameh who labeled thousands of human body poses.

I would like to thank many people whose everyday work targets all possible non-research related problems which every PhD student has to face. I'm grateful to our system administrators, Marcus Rohrbach and Jan Hosang, for doing a great job on organizing our working environment and managing the computation resources. I would like to thank all people in the help desk support team for maintaining our machines and computing cluster, and for always prompt responses to the hardware and software related issues. I'm also grateful to Uwe Brahm, Dag Kröper and Andreas Merkel for their continuous help regarding the web administration tasks. I would like to thank Connie Balzert for her excellent organizational work. Also, I would like to thank Jennifer Gerling and Michelle Cornell for their help and coordination of the graduate studies.

I'm also grateful to Mario Fritz and Andreas Bulling for our team work on screening and discussing new candidates as part of the admission committee for the graduate studies.

I would like to thank Dr. Zeynep Akata, Bojan Pepik, Mohamed Omran, Siyu Tang, Dr. Rodrigo Benenson and Mateusz Malinowski for their help in proofreading this thesis.

I would like to thank my office mates, Bojan Pepik and Wenbin Li, and all other colleagues, for creating the atmosphere of fun in our group.

Lastly, I'm very grateful to my fiancée Mónica for her continuous unconditional support and encouragement, and to my family who has been always believing in me and supporting me in all my undertakings.

CONTENTS

1 **Introduction** **1**
1.1 Challenges of articulated people detection and pose estimation 4
1.1.1 Obtaining representative training data with relevant variations 4
1.1.2 Building expressive models for human pose estimation 6
1.1.3 Benchmarking and analyzing the state of the art 8
1.2 Contributions of the thesis 9
1.2.1 Contributions to obtaining representative training data with relevant variations 9
1.2.2 Contributions to building expressive models for human pose estimation 11
1.2.3 Contributions to benchmarking and analyzing the state of the art 12
1.2.4 Other contributions 12
1.3 Outline of the thesis 13

2 **Related work** **17**
2.1 Obtaining representative training data 17
2.1.1 Training from synthetic data 18
2.1.2 Building statistical models of 3D human shape and pose 23
2.2 Building expressive models for human pose estimation 27
2.2.1 Basic pictorial structures based methods 27
2.2.2 Improving basic pictorial structures 29
2.2.3 Deep learning methods 32
2.2.4 Multi-person pose estimation 34
2.2.5 Relations to our work 35
2.3 Benchmarking and analyzing the state of the art 40
2.3.1 Benchmarking human pose estimation 40
2.3.2 Benchmarking human activity recognition 47

3 **Learning People Detection Models from Few Samples** **59**
3.1 Introduction 59
3.2 People detection models 61
3.3 MovieReshape: 3D human shape model 63
3.4 Experimental evaluation 64
3.4.1 Results using the Reshape data 66
3.4.2 Combining different datasets for training 68
3.4.3 Sliding-window detection using HOG 70
3.5 Conclusion 71

4 Robust People Detection based on Appearance and Shape 73
4.1 Introduction . . . 74
4.2 Pictorial structures model . . . 75
4.3 Synthetic data . . . 76
  4.3.1 3D model of human shape . . . 76
  4.3.2 Data generation . . . 76
4.4 Experimental evaluation . . . 79
  4.4.1 Results using rendered data alone and combined with image data 80
  4.4.2 Combination of people detection models . . . 81
4.5 Conclusion . . . 83

5 Articulated People Detection and Pose Estimation 85
5.1 Introduction . . . 85
5.2 Generation of novel training examples . . . 87
  5.2.1 Data annotation . . . 87
  5.2.2 3D human shape recovery and animation . . . 87
  5.2.3 Generation of novel images . . . 89
5.3 Articulated people detection . . . 90
5.4 Articulated pose estimation . . . 92
  5.4.1 Pictorial structures model . . . 93
  5.4.2 Joint PS-DPM model . . . 93
  5.4.3 Experimental evaluation . . . 94
5.5 Articulated pose estimation “in the wild” . . . 96
5.6 Conclusion . . . 98

6 Statistical Shape Spaces for 3D Human Modeling 101
6.1 Introduction . . . 102
6.2 Statistical modeling with SCAPE . . . 103
  6.2.1 Original SCAPE model . . . 103
  6.2.2 Simplified SCAPE space . . . 104
6.3 Data processing . . . 104
  6.3.1 Non-rigid template fitting . . . 104
  6.3.2 Initialization . . . 106
  6.3.3 Bootstrapping . . . 106
  6.3.4 Posture normalization . . . 106
6.4 Evaluation of template fitting . . . 107
  6.4.1 Implementation details . . . 107
  6.4.2 Quality measure . . . 108
  6.4.3 Initialization . . . 108
  6.4.4 *NRD* parameters . . . 109
  6.4.5 Bootstrapping . . . 110
6.5 Evaluation of statistical shape space . . . 110
  6.5.1 Quality measure . . . 110
  6.5.2 Bootstrapping . . . 111
  6.5.3 Number of training samples . . . 112

6.5.4 Posture normalization 112
6.6 Human body reconstruction 113
6.6.1 Dataset and experimental setup 113
6.6.2 Quantitative evaluation 113
6.6.3 Qualitative evaluation 115
6.7 Conclusion 117

7 Poselet Conditioned Pictorial Structures 119
7.1 Introduction 119
7.2 Review of pictorial structures 122
7.3 Poselet Conditioned Pictorial Structures 123
7.3.1 Poselet representation 124
7.3.2 Poselet dependent unary terms 124
7.3.3 Poselet dependent pairwise terms 125
7.4 Results 125
7.4.1 Results on LSP dataset 126
7.4.2 Results on Image Parse dataset. 129
7.4.3 Results on UIUC People dataset. 131
7.5 Conclusion 131

8 Expressive Models for Human Pose Estimation 133
8.1 Introduction 134
8.2 Pictorial structures model 136
8.2.1 Model formulation 136
8.2.2 Learning and inference 136
8.3 Better appearance representations 137
8.3.1 Body part detectors 137
8.3.2 Head and torso detectors (*spec-head, spec-torso*) 138
8.3.3 Implementation details 138
8.4 More flexible models 139
8.4.1 Body joints (*PS-flex*) 139
8.4.2 Mid-level representations (*mid-level*) 139
8.5 Results 140
8.5.1 Results on LSP dataset 140
8.5.2 Results on Image Parse dataset 144
8.6 Conclusion 146

9 Human Pose Estimation Benchmark and Analysis 149
9.1 Introduction 149
9.2 Dataset 151
9.3 Analysis of human pose estimation approaches 153
9.3.1 Analysis of pose estimation challenges 155
9.3.2 Evaluation of Deep Learning based Methods 163
9.4 Conclusion 166

10 Fine-grained Activity Recognition 167
10.1 Introduction 168
10.2 Dataset 169
10.3 Methods 170
10.3.1 Dense trajectories (DT) 170
10.3.2 Pose-based methods 170
10.3.3 Combinations of holistic and body based methods 171
10.4 Analysis of activity recognition performance 171
10.4.1 Overall performance 172
10.4.2 Analysis of activity recognition challenges 174
10.4.3 Detailed analysis on a subset of activities 175
10.5 Conclusion 177

11 Joint Multi Person Pose Estimation 179
11.1 Introduction 180
11.2 Problem Formulation 182
11.2.1 Feasible Solutions 182
11.2.2 Objective Function 183
11.2.3 Optimization 184
11.3 Pairwise Probabilities 184
11.3.1 Probability Estimation 185
11.4 Body Part Detectors 186
11.4.1 Adapted Fast R-CNN (*AFR-CNN*) 186
11.4.2 Dense architecture (*Dense-CNN*) 186
11.4.3 Evaluation of part detectors 188
11.4.4 Using detections in DeepCut models 189
11.5 DeepCut Results 190
11.5.1 Single person pose estimation 190
11.5.2 Multi-person pose estimation 192
11.6 Conclusion 201
11.7 Appendix: Additional Results on LSP dataset 201
11.7.1 LSP Person-Centric (PC) 201
11.7.2 LSP Observer-Centric (OC) 204

12 Conclusions and future perspectives 207
12.1 Discussion of contributions 208
12.2 Future perspectives 213
12.2.1 Obtaining representative training data with relevant variations 213
12.2.2 Building expressive models for human pose estimation 214
12.2.3 Benchmarking and analyzing the state of the art 216
12.2.4 A broader view on the topic 217

List of Figures 219

List of Tables 225

Bibliography **227**

Curriculum Vitae **245**

Publications **247**

1 INTRODUCTION

Contents

1.1 Challenges of articulated people detection and pose estimation . . . **4**
1.1.1 Obtaining representative training data with relevant variations 4
1.1.2 Building expressive models for human pose estimation . . . 6
1.1.3 Benchmarking and analyzing the state of the art 8
1.2 Contributions of the thesis **9**
1.2.1 Contributions to obtaining representative training data with relevant variations 9
1.2.2 Contributions to building expressive models for human pose estimation 11
1.2.3 Contributions to benchmarking and analyzing the state of the art 12
1.2.4 Other contributions 12
1.3 Outline of the thesis **13**

UNDERSTANDING visual scenes that contain people is one of the core research questions in computer vision. Over the course of the last decade plenty of articles have been published focusing on various facets of this complex problem, including face detection and recognition, people detection and tracking, human pose estimation and activity recognition, to name a few. Increased attention to this research topic can be accounted for several reasons.

First, numerous potential applications have motivated rapid development in this field. Intelligent humanoid robots performing a variety of tasks intend to help people in their households. Visual surveillance systems automatically detecting abnormal situations and suspicious behavior of humans increase our every day life security. Driver assistance systems become irreplaceable helpers that make our driving safer. Gaming and special effects industry benefit from human pose and motion analysis systems that create realistic and immersive entertainment experience. The success of these systems in achieving their goals depends on the extend they can perceive the human environment, how well they can interpret behavior and predict intentions of people, and how interactive their response is.

Second, the problem of understanding visual scenes containing people is an attractive research question from the scientific point of view. Advances towards solving this problem ultimately lead to the improved understanding of artificial intelligence in general. With human being one of the most complex object classes, the problem of understanding visual scenes containing people serves as a useful testbed to demonstrate the advances in computer vision and related fields. Depending on the level of detail the system should understand the scene, it requires solving a wide range of problems. For instance, at the lowest level of detail a system should provide information if a person is present in the scene, and if yes, determine his location. This requires solving an instance of a generic object detection task, namely people detection. However, due to the high degree of intra-class variability in appearance, pose and shape, and inherited generic object detection challenges, such as background clutter and imaging conditions, already this task of detecting arbitrary people in unconstrained real world environments is a hard problem. Inferring more details about people present in the scene by detecting their individual body parts is even harder. Body parts are small in size compared to the full body and miss characteristic appearance features due to their generic shapes and high intra-class appearance variability caused by changes in clothing, skin color and articulation, frequent occlusions and out of plane rotations. Thus the detection of individual body parts requires the development of methods that are invariant to pose, texture and lighting, while also being able to separate the parts from background. Further increasing the level of details that can be inferred about the scene, e.g., understanding human emotions and intentions, typically requires the knowledge about human behavioral and social interaction patterns, and using detailed information about human location and body pose as the key building blocks. Despite significant progress achieved for upright pedestrian detection (Benenson *et al.*, 2014) and human pose estimation in sports (Tompson *et al.*, 2014), articulated people detection and human pose estimation in unconstrained real world environments remains challenging.

In this thesis we are interested in articulated people detection and pose estimation as one of the key tasks towards understanding visual scenes containing people. Importantly, we focus on using a single monocular RGB image as input. This requirement makes both tasks more challenging compared to the case when using additional sensors, such as multi-view camera systems or depth sensors (Shotton *et al.*, 2011), where information about the scene depth allows to significantly constrain the search for human location and body pose. At the same time, the requirement of using a single monocular RGB image as input also makes this problem more general, as such methods can be applied outdoors and in unconstrained real world environments. Moreover, we do not make any assumption about appearance, viewpoint, shape and pose of individuals, imaging conditions, background, outdoor vs. indoor environment, etc. Thus the general setting we operate in makes our methods applicable in various scenarios, such as personal photo collections, movies and video sharing web sites.

Addressing articulated people detection and pose estimation in challenging real world scenarios requires methods that, on the one hand, are discriminative enough to separate humans from highly cluttered background, but, on the other hand, are also representative enough to capture relevant variations in human appearance, shapes and poses. In this thesis we investigate three directions towards building such methods. The first direction is *Obtaining representative training data with relevant variations*. Here we propose a range of automatic data generation techniques that directly encode relevant variations into the training data. As a core of our methods we use a state-of-the-art statistical 3D human shape model (Jain *et al.*, 2010) from computer graphics. Sampling from the underlying human shape distribution and a large set of human poses allows to generate novel samples with controllable shape and pose variations that are relevant for the task at hand. Our approaches are presented in Chapters 3, 4, and 5. In addition, in Chapter 6 we improve the 3D body shape model (Jain *et al.*, 2010) by building efficient and expressive shape spaces from a large commercially available 3D body shape dataset.

In the second direction, *Building expressive models for human pose estimation*, we explore ways of developing expressive spatial and strong appearance models for 2D single- and multi-person pose estimation. We observe that human motion and activities often simultaneously constrain the articulations of multiple body parts and propose an expressive image conditioned model that incorporates such higher order part dependencies while remaining efficient. Furthermore, we explore various types of appearance representations aiming to substantially improve the body part hypotheses. We draw on the best practices in human pose estimation and combine flexible spatial model with our expressive image conditioned model and strong appearance representations into a powerful human pose estimation model outperforming many competitors on prominent pose estimation benchmarks. Our novel image conditioned model is presented in Chapter 7 and the follow-up work analyzing and combining the best practices in human pose estimation is presented in Chapter 8. We further significantly improve appearance representations in Chapter 11 by developing strong deep learning based body part detection models. Building on strong part detectors, we propose an expressive spatial model for joint pose estimation of multiple people. Our model infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other.

Developing powerful pose estimation models requires deeper understanding of the limitations of current methods. Thus the third direction explored in this thesis, *Benchmarking and analyzing the state of the art*, is concerned with a thorough evaluation and performance analysis of leading human pose estimation and activity recognition methods. To that end, we introduce a novel benchmark that makes a significant advances in terms of diversity and difficulty, compared to the current datasets, and includes over $40,000$ images of people. Furthermore, we provide a rich set of labels which are used to perform a detailed analysis of the current approaches gaining insights into success and failures of these methods. We introduce the dataset and performance analysis of human pose estimation methods in Chapter 9, while analysis of human activity recognition approaches is presented in Chapter 10.

Given the three directions chosen above, one PhD thesis cannot cover all potential ways of advancing the state of the art in articulated people detection and pose estimation. In Chapter 12 we discuss potential directions for future work, such as using motion to improve detection of people and individual body parts and generating richer output in terms of 3D body pose or figure-ground segmentation. We believe though that the three directions chosen in this thesis address challenging topics that are highly relevant for the computer vision community. We show in Chapter 2 that the topics discussed in this thesis are timely, with many papers published during the course of this work.

In the following we analyze the challenges w.r.t. the three directions of this thesis, and how we approach these challenges. We then discuss the contributions of this work and conclude the chapter by providing the outline of this thesis and referring to respective publications resulting from this work.

1.1 CHALLENGES OF ARTICULATED PEOPLE DETECTION AND POSE ESTIMATION

Articulated people detection and pose estimation in monocular RGB images are two challenging and highly researched problems in computer vision. Addressing these problems in real-world scenarios is hard due to a number of factors. In Figure 1.1 we show sample images from the setting we are interested in. People dress in a large variety of different ways and have different body shapes. The same individual looks different depending on camera point of view and imaging conditions. At the same time, human body parts can move freely resulting in numerous body poses. Also appearance of the same body part changes significantly due to clothing, foreshortening and occlusion by other body parts, while appearance of different symmetric limbs, e.g., left and right arms, is similar. In addition, the spatial extent of majority of the body parts is rather small, and when taken independently all parts lack characteristic appearance features. Jointly these factors contribute to a high intra-class variability that makes part detection hard. Finally, detection of people and their individual body parts has to deal with classic problems of generic object detection, such as background clutter, varying lighting conditions and limited number of training samples. This problem becomes even harder when multiple humans are present in the image, as in this case one needs to correctly associate body part detections to the corresponding individuals and resolve ambiguities between limbs with similar appearance.

State-of-the-art articulated people detection and pose estimation methods aim to cope with these challenges using powerful supervised discriminative learning methods trained on large and representative datasets. The performance is then measured on public benchmarks using established evaluation measures enabling direct comparison between competing approaches. In the following subsections we discuss key challenges of obtaining large amounts of representative training data, building powerful pose estimation models and proper benchmarking of competing methods.

1.1.1 Obtaining representative training data with relevant variations

State-of-the-art methods for articulated people detection and pose estimation require large and representative training sets for best performance. In the following we focus on specific challenges of obtaining a large number of representative training samples that capture the essential variability of the complex people class distribution.

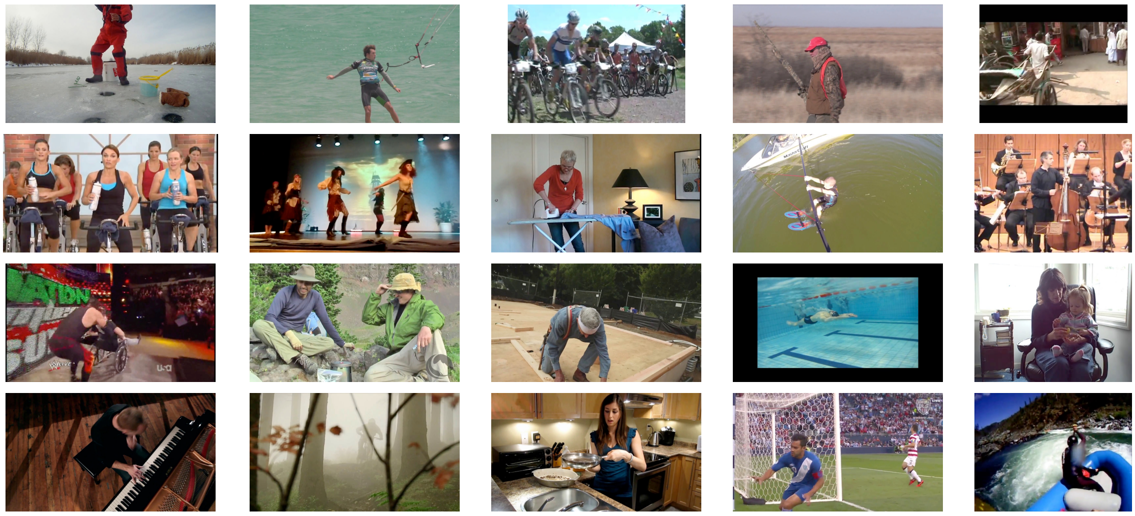


Figure 1.1: Example images showing challenges of articulated people detection and pose estimation.

Manual data collection and labeling. Recent progress in people detection and pose estimation has been possible mostly due to discriminative learning that allows to learn powerful models from a large training corpora. Large and representative training sets are essential for best performance and significant effort has been made in the computer vision community towards collecting large amounts of training data. Typically, images are extracted from public data sources (e.g. photo collections) and manually annotated, which is a tedious and time consuming task that often limits further improvements.

In this thesis we overcome the difficulty of manual data collection and labeling by automatically generating and annotating large amounts of synthetic data used to increase the available amounts of training data. We discuss different ways of generating synthetic data in Chapters 3, 4 and 5.

Representative training data. Collecting large amounts of representative training data is not only tedious but often an ill-defined process as it is unclear which part of the people class distribution is well-represented and which other parts of the distribution are still insufficiently sampled. A typical approach is to simply collect more training data without any guarantee that it will better cover the people class distribution. Current datasets are often limited to few thousand samples taken from consumer photographs without any statement which parts of the real distribution of human body shapes, poses and appearances are sampled.

In this thesis we address the challenge of collecting representative training data in two ways. First, in Chapters 3, 4 and 5 we use a 3D human shape model from computer graphics to produce a set of realistic shape deformations, and combine them with motion capture data to obtain feasible pose changes. This allows us to generate novel training samples with the full control over the shape and pose variations. Second, in Chapter 9 we use an established taxonomy of hundreds of human activities to collect a representative human pose estimation benchmark. Using the taxonomy during the manual data collection process allows to achieve a fair coverage of both common and rare poses that might be missed when simply collecting more images without aiming for good coverage.

Building efficient and expressive 3D human shape spaces. State-of-the-art 3D human shape models used for automatic data generation in Chapter 3, 4 and 5 are learned from a rather small publicly available collection of pre-processed human scans (Hasler *et al.*, 2009) with limited shape variations. Learning a model from much more representative datasets is challenging, as 1) such datasets are not freely available, and 2) it requires a significant engineering effort and know-how in order to implement a data pre-processing step that needs to bring the scans in correspondence.

In Chapter 6 we rebuild the widely used statistical body representation (Jain *et al.*, 2010) from a large commercially available scan database, by developing robust best practice solutions for scan alignment that quantitatively lead to best models. During the course of this work we also showed how to improve the 3D body shape and posture estimation under clothing (Wuhrer *et al.*, 2014).

1.1.2 Building expressive models for human pose estimation

After discussing challenges of obtaining large representative training sets including relevant variations, we switch our focus to discussing the major challenges when building expressive body models for the task of articulated human pose estimation in monocular images.

Efficient modeling of higher-order part dependencies. While tree-structured spatial models, commonly used for human pose estimation, allow for exact and efficient inference, they fail to capture important dependencies between *non-adjacent* body parts. Modeling such dependencies is important for effective pose estimation, but also challenging, as it is typically approached by adding cycles into the underlying graphical model, thus making exact and efficient inference infeasible.

In Chapter 7 we propose a novel model that incorporates higher order information between body parts by defining a conditional model in which all parts are a-priori connected, but which becomes a tractable pictorial structures model once the mid-level features are observed. This helps to effectively model dependencies between non-adjacent parts, while still allowing for exact and efficient inference in a tree-based model.

Mid-level pose and appearance representation. Building efficient body models based on image conditioned spatial and appearance terms requires robust mid-level image representation. This representation is typically used as an intermediate level between the observed visual information and the graphical model. On the one hand, mid-level representation has to be robust w.r.t. variations in people appearance, pose and imaging conditions. On the other hand, it has to be highly informative for the underlying human pose.

We analyze these requirements in Chapter 7 and use a non-parametric mid-level representation that jointly models appearance of multiple body parts in order to condition the spatial and appearance terms of our graphical model.

Obtaining strong body part detectors. Building strong detectors for all body parts is challenging due to several reasons. The appearance of body parts changes significantly due to clothing, foreshortening and occlusion. In addition, the spatial extent of majority of the body parts is rather small, and when taken independently each part lacks characteristic appearance features.

In Chapter 8 we argue that in order to obtain effective part detectors, it is necessary to leverage both the pose specific appearance of body parts, and the joint appearance of part constellations. In Chapter 11 we further significantly improve body part detectors by building on the recent advances of deep learning.

Analyzing various combinations of spatial and appearance representations. Most recent approaches to human pose estimation combine individual body parts with a set of pairwise part dependencies. Typically each work proposes a combination of a particular appearance model with a specific spatial model, and no study of different combinations has been performed so far. Performing a deep analysis of different combinations to discover highly complementary representations is challenging, as it requires thorough experimental evaluation in a single modeling framework.

In this thesis we combine and analyze several recently proposed powerful ideas, such as more flexible spatial models, as well as our image conditioned models. Starting with the basic tree-structured pictorial structures we perform a series of experiments incrementally adding various components corresponding to spatial and appearance representations and analyze the resulting performance gains in Chapter 8.

Multi-person pose estimation. Most recent work on articulated pose estimation considers a simplified problem by assuming that there is a single person in the image and that an approximate scale and location of the person is known. The proposed approaches typically output a single estimate of body configuration per image and do not provide any confidence score that the pose estimate is indeed correct. This ignores three important challenges which arise when applying these approaches on uncropped images. First, many images contain multiple people and thus in addition to estimating their poses one has to decide how many people are present. Second, for people in close proximity of each other it is necessary to reason about which body part detections belong to which individual. Third, for each person it requires searching over a wide range of possible positions and scales, and it is not clear how well current methods are able to deal with such increase in complexity.

We argue that in order to properly asses the state of the art in articulated people detection and pose estimation it is necessary to consider these problems jointly. To that end, in Chapter 5 we define a new dataset and evaluation criteria, and evaluate the performance of joint people detection and pose estimation in a more realistic scenario. Furthermore, in Chapter 11 we propose a novel multi-person pose estimation model which is intended to estimate the number of people in the image, correctly associate individual part detections to multiple individuals and resolve ambiguities between limbs with similar appearance. During the course of this thesis we further improved the proposed multi-person model by introducing novel image conditioned pairwise terms and incremental optimization strategies (Insafutdinov *et al.*, 2016), which allows to significantly push the state of the art in multi-person pose estimation while drastically reducing the run-time.

1.1.3 Benchmarking and analyzing the state of the art

Establishing representative benchmark. Current datasets for human pose estimation are limited in their coverage of the overall pose estimation challenges, as well as scope and variability of represented activities. Still these serve as the common sources to evaluate, train and compare different models on. Thus the key challenge when building a pose estimation benchmark is how to achieve a fair coverage of both common and rare pose of humans involved into various every day activities.

In Chapter 9 we introduce a novel benchmark collected using an established taxonomy of hundreds of human activities. The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational and householding activities, and capture people from a wider range of viewpoints.

Providing rich annotations. Current pose estimation benchmarks typically provide body joint annotations only. While 2D joint labels allow to evaluate pose estimation performance, this ground truth information is usually not enough to perform deeper performance analysis and evaluate the method's robustness to various pose estimation challenges. However, data annotation is a tedious and time consuming task. The richer the labeling is required, the more expensive it gets to label large amounts of data. Amazon Mechanical Turk (AMT) provides a possibility to scale up the annotation of large datasets, however easy-to-use annotation tools are required for efficient image labeling.

In Chapter 9 we present a novel benchmark that comes with a rich set of labels including 2D positions of body joints, full 3D torso and head orientation, visibility labels for joints and entire body parts, and activity labels. The annotations are obtained by in-house workers and via AMT. We developed a set of annotation tools that allow to efficiently obtain rich labelings for large datasets.

Building performance analysis tools. Given image labels, it is not obvious how to characterize the complexity of the image w.r.t. various pose estimation and activity recognition challenges. Quantitative complexity measures are required that map image annotations to real values which relate the complexity of the image w.r.t. each factor.

Given rich annotations of our novel dataset in Chapters 9 and 10 we define several quantitative complexity measures characterizing the scene complexity w.r.t. various human pose estimation and activity recognition challenges. This allows for detailed performance analysis of prominent pose estimation and activity recognition approaches demonstrating their successes and failures.

Establishing evaluation metrics. Defining a proper metric for evaluation of body part prediction is challenging, as this metric should be scale and articulation independent. The widely adopted "Percentage of Correct Parts (PCP)" and "Percentage of Correct Keypoints (PCK)" metrics evaluating pose estimation accuracy have drawbacks. The PCP metric uses part length as a threshold and thus requires that foreshortened body parts are localized with higher precision to be considered correct. The PCK metric defines a threshold as a fraction of the size of person bounding box including all body joints. This makes the PCK metric articulation dependent and thus unnecessarily loose in case of high degree of articulation, or too strict in case of compact body poses.

We analyze the drawbacks of existing metrics and propose improvements in Chapters 9 and 11.

1.2 CONTRIBUTIONS OF THE THESIS

After stating the individual challenges in the field and discussing how this thesis addresses these challenges, we now summarize the contributions w.r.t. the three directions of this thesis. In Chapter 2 we put our contributions in the context of related work. In Chapter 12 we discuss the contributions from the perspective of individual chapters, as a part of the concluding discussion.

1.2.1 Contributions to obtaining representative training data with relevant variations

In this thesis we investigate how 3D shape models from computer graphics can be leveraged to ease training data generation. The question we are asking is if the realism of today's computer graphic models can help computer vision to reduce the tedious task of data collection and at the same time improve the quality and the relevant variability of the training data. To that end, we propose several data generation methods based on a 3D human body shape model that represents pose and shape variations of human bodies.

Our first contribution is a novel data generation method that allows to generate thousands of photo-realistically looking synthetic training samples from only a few persons and views. Starting from an image sequence of an individual captured in a motion capture studio our method allows to generate large amounts of synthetic training data representing 3D shape and pose variations of the recorded individual. We show that surprisingly good results can be obtained from as few as one or two people. When training from eleven people we are able to match the performance of competing approaches learning from hundreds of individuals.

As a second contribution, we use a 3D human shape model to enrich an existing training data with additional non-photo-realistically looking training samples. We explore how complementary shape information sampled from the underlying 3D human shape distribution can be directly incorporated into the low level feature representation. We show that by careful design of the rendering procedure our feature representation can generalize well from synthetic training data to unseen real test data.

As a third contribution, we propose a novel method for automatic generation of multiple training examples from an arbitrary set of images with annotated human body poses. We use a 3D human shape model to produce a set of realistic shape deformations of person's appearance, and combine them with motion capture data to obtain feasible pose changes. This allows us to generate realistically looking training images of people while having full control over the shape and pose variations.

Our fourth contribution is an extensive evaluation of our data generation methods on the tasks of articulated people detection and human pose estimation. We explore how various parameters of the data generation process affect overall performance. We evaluate different strategies to combine novel synthetic and existing real data. We directly compare to other prominent methods trained on hundreds of manually labeled samples, as well on synthetic samples generated by competing method. On both tasks we can significantly improve performance when the training sets are extended with the automatically generated samples having relevant shape and pose variations.

Finally, our fifth contribution is concerned with improving the state-of-the-art 3D human shape model used in our data generation methods. This model was learned from the largest publicly available dataset consisting of rather small number of human scans lacking diversity in represented human shapes. We contribute by rebuilding the 3D human shape model from a large commercially available scan database (Robinette *et al.*, 1999), and making the resulting model available to the community. As preprocessing several thousand scans for learning the model is a challenge in itself, we contribute by developing robust best practice solutions for scan alignment that quantitatively lead to the best learned models. We also make the implementations of these pre-processing steps publicly available. We evaluate the improved accuracy and generality of our new model and show its improved performance for human body reconstruction from sparse input data.

1.2.2 Contributions to building expressive models for human pose estimation

In this thesis we propose several expressive body models for 2D single- and multi-person pose estimation.

First, we observe that despite high variability of body articulations, human motions and activities often simultaneously constrain the positions of multiple body parts. However, modeling such higher order part dependencies seemingly comes at a cost of more expensive inference, which resulted in their limited use in state-of-the-art methods. We thus propose a single person pose estimation model that incorporates higher order part dependencies while remaining efficient. We achieve this by defining a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structures model once the image observations are available. We evaluate different components of our method and show their contribution to the final performance. Furthermore, we demonstrate the high potential of the proposed approach by analyzing the performance in the ideal case. We demonstrate the effectiveness of our approach on three publicly available single person pose estimation benchmarks improving over or being on-par with the competitors in each case.

Second, we analyze and draw on several recently proposed powerful ideas such as strong local appearance models, flexible spatial models and our image conditioned method. We explore various types of appearance representations including rotation invariant or rotation specific appearance templates, mixtures of such local templates, specialized models tailored to appearance of salient body parts such as head and torso, and semi-global representations based on poselet features. In a series of experiments we draw several important conclusions: (1) we show that the proposed appearance representations are complementary; (2) we demonstrate that even a basic tree-structured spatial human body model achieves very good performance when augmented with the proper appearance representation; and (3) we show that the combination of the best performing appearance model with a flexible image conditioned spatial model achieves the best result, significantly improving over the competitors on prominent single person pose estimation benchmarks.

Third, we propose a novel multi-person human pose estimation model that is able to infer the number of people in a scene, identify occluded body parts and disambiguate body parts between people in close proximity of each other. To that end, we develop strong deep learning based body part detection models and an expressive fully connected spatial model. We treat multi-person pose estimation as a joint partitioning and labeling problem of a set of body part hypotheses. Our formulation implicitly performs non-maximum suppression on the set of part detections and groups them to form configurations of body parts that respect geometric and appearance constraints. The proposed formulation is an integer linear program and therefore allows the use of robust optimization techniques and facilitates the computation of bounds and feasible solutions with a certified optimality gap. We demonstrate significant improvements for single- and multi-person pose estimation over the state of the art on challenging and diverse public benchmarks.

1.2.3 Contributions to benchmarking and analyzing the state of the art

Based on our observation that current human pose estimation datasets are limited in their coverage of the overall pose estimation challenges, we contribute a novel benchmark that makes a significant advance in terms of diversity and difficulty. We collected this comprehensive dataset using an established taxonomy of several hundreds of human activities (Ainsworth *et al.*, 2011). The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational and householding activities, and capture people from a wider range of viewpoints. Furthermore, we contribute a rich set of labels including positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels. We release the dataset for public usage.

Our second contribution is a detailed analysis of prominent human pose estimation methods on our novel benchmark. We define a set of quantitative complexity measures that map rich body image annotations to a real value that relates the complexity of the image to human pose estimation challenges. Based on these complexity measures we contribute a set of performance analysis tools. We also establish novel evaluation measures intending to overcome the shortcomings of the current metrics. We complete a detailed performance analysis of prominent pose estimation methods, identify their strengths and drawbacks and propose the most promising future research directions.

Our third contribution is a thorough analysis of famous holistic and pose based activity recognition methods. Similar to human pose estimation, we define a set of complexity measures characterizing the scene difficulty w.r.t. activity recognition challenges. We contribute an extensive experimental evaluation of individual holistic and pose based methods and their combinations and discover a number of factors responsible for successes and failures of these methods.

1.2.4 Other contributions

In addition to the contributions presented in this thesis we contributed to the computer vision and computer graphics communities in the following ways. First, we prepared and released the source code of our image conditioned human pose estimation model for public usage thus allowing other researchers to build directly on the best practices in human pose estimation. Second, we released a novel 3D body shape model learned from a large commercially available dataset of human body shapes; we also released code to pre-process raw human scans and to fit model shape and pose to raw scans. Third, in order to facilitate the development and unify the performance comparison and analysis of human pose estimation and activity recognition methods, we collected, annotated and publicly released a novel benchmark; we contributed a set of performance analysis tools and created evaluation web pages with current best results on ours and other related benchmarks. Finally, we are currently working on releasing our state-of-the-art deep learning-based part detectors and our powerful multi-person pose estimation model.

1.3 OUTLINE OF THE THESIS

We now summarize the chapters of the thesis and put them in relation to each other. In addition, we refer to the respecting publications and collaborations with other researchers.

Chapter 2: Related work. In this chapter we provide an overview of the related work with a focus on three directions of this thesis, namely *Obtaining representative training data with relevant variations*, *Building expressive models for human pose estimation* and *Benchmarking and analyzing the state of the art*. We analyze the relations of other works to the methods and contributions presented in the thesis.

Chapter 3: Learning People Detection Models from Few Samples. In this chapter we introduce an approach to generate a large number of photo-realistically looking synthetic training samples from only a few persons and views. This data generation method is based on the method of (Jain *et al.*, 2010) that allows to reshape humans in videos. However, the focus of this chapter is on exploring the applicability of a state-of-the-art 3D body shape model to learn powerful people detection models from limited number of poses and appearances captured under controlled conditions.

The content of this chapter corresponds to the CVPR 2011 publication *Learning People Detection Models from Few Training Samples* (Pishchulin *et al.*, 2011a). Leonid Pishchulin was the lead author of this paper, while Arjun Jain contributed to data generation using his method (Jain *et al.*, 2010).

Chapter 4: Robust People Detection based on Appearance and Shape. This chapter explores the possibility of using the 3D body shape model directly to augment existing training data with complementary shape variations. In contrast to Chapter 3, this chapter does not aim to use visually appealing and photo-realistically rendered images but instead focuses on complementary and particularly important information for people detection, namely 3D human shape.

The content of this chapter corresponds to the BMVC 2011 publication *In Good Shape: Robust People Detection based on Appearance and Shape* (Pishchulin *et al.*, 2011b). Leonid Pishchulin was the lead author of this paper accepted as *Oral*. Arjun Jain contributed the code for 3D human shape modeling.

Chapter 5: Articulated People Detection and Pose Estimation. Based on experiences gained in Chapters 3 and 4, this chapter presents a method that uses the 3D body shape model to efficiently generate a large number of photo-realistically looking samples with controllable shape and pose variations from arbitrary monocular images. We demonstrate the applicability of this method to learn powerful detection and pose estimation models of highly articulated people.

The content of this chapter corresponds to the CVPR 2012 publication *Articulated People Detection and Pose Estimation: Reshaping the Future* (Pishchulin *et al.*, 2012). Leonid Pishchulin was the lead author of this paper. Arjun Jain implemented the appearance rendering part of the synthetic data generation pipeline.

Chapter 6: Statistical Shape Spaces for 3D Human Modeling. In this chapter we turn our attention to improving the state-of-the-art 3D human body shape model used in Chapters 3, 4 and 5. We specifically focus on how to build an efficient and expressive shape space from a large commercially available 3D body shape dataset (Robinette *et al.*, 1999). We evaluate different variants of state-of-the-art techniques for non-rigid template fitting and posture normalization to process the raw data. Thorough experimental evaluation shows several advantages of the learned shape models over the state of the art both in terms of statistical model quality and for the task of reconstructing 3D human body shapes from monocular depth images.

The content of this chapter corresponds to the Pattern Recognition 2016 submission that we additionally made publicly available on ArXiv (Pishchulin *et al.*, 2015). Leonid Pishchulin was the lead author of this paper.

Chapter 7: Poselet Conditioned Pictorial Structures. In this chapter we switch the focus from learning from synthetic data towards building expressive body models for human pose estimation. We propose a model that incorporates higher order part dependencies while remaining efficient. We achieve this by defining a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structures model once image observations are available.

The content of this chapter corresponds to the CVPR 2013 publication *Poselet Conditioned Pictorial Structures* (Pishchulin *et al.*, 2013a) accepted as *Oral*. Leonid Pishchulin was the lead author of this paper.

Chapter 8: Expressive Models for Human Pose Estimation. In this chapter we explore various types of appearance representations aiming to substantially improve the body part hypothesis. In addition, we draw on and combine several powerful ideas in human pose estimation, such as flexible spatial models as well as our image conditioned models from Chapter 7. As a result of our analysis we build a powerful human pose estimation model thereby outperforming the previous methods by a large margin on prominent pose estimation benchmarks.

The content of this chapter corresponds to the ICCV 2013 publication *Strong Appearance and Expressive Spatial Models for Human Pose Estimation* (Pishchulin *et al.*, 2013b). Leonid Pishchulin was the lead author of this paper.

Chapter 9: Human Pose Estimation Benchmark and Analysis. In this chapter we focus on benchmarking of prominent human pose estimation methods and thorough analysis of their performance. To that end, we introduce a novel large-scale benchmark that makes a significant advance in terms of diversity and difficulty, compared to other datasets. allows for detailed analysis of prominent human pose estimation approaches gaining insights into the successes and failures of these methods.

The content of this chapter corresponds to the CVPR 2014 publication *2D Human Pose Estimation: New Benchmark and State of the Art Analysis* (Andriluka *et al.*, 2014). Leonid Pishchulin and Mykhaylo Andriluka contributed equally, with Mykhaylo Andriluka being the initiator of this project.

Chapter 10: Fine-grained Activity Recognition. Based on the dataset (Chapter 9) including hundreds of everyday human activities, in this chapter we perform a thorough analysis of popular human activity recognition methods. In particular, we aim to clarify the underlying factors responsible for good performance of holistic methods based on dense trajectories (Wang *et al.*, 2013) and their counterparts using higher level encoding in terms of body pose and motion.

The content of this chapter corresponds to the GCPR 2014 publication *Fine-grained Activity Recognition with Holistic and Pose based Features* (Pishchulin *et al.*, 2014). Leonid Pishchulin was the lead author of this paper.

Chapter 11: Joint Multi Person Pose Estimation. In this chapter we switch our attention back to developing expressive body models for human pose estimation. In particular, we consider the task of multi-person pose estimation in real world images. We propose an approach that jointly solves the tasks of detection and pose estimation: it infers the number of people in a scene, identifies occluded body parts and disambiguates body parts between people in close proximity of each other. We achieve this by formulating the problem of pose estimation as partitioning and labeling of a set of body part hypotheses generated by the proposed strong deep learning based part detectors.

The content of this chapter corresponds to the CVPR 2016 publication *DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation* (Pishchulin *et al.*, 2016). Leonid Pishchulin was the lead author of this paper. Eldar Insafutdinov contributed to the implementation of fully-convolutional deep learning body part detectors. Bjoern Andres and Siyu Tang contributed to the implementation of the optimization solver.

Chapter 12: Conclusions and future perspectives. In this chapter we summarize this thesis' contributions and discuss current limitations, as well as possible directions to overcome the limitations. In addition, we provide an outlook on our ongoing and future work and discuss future directions for articulated people detection and pose estimation.

RELATED WORK 2

Contents

2.1 Obtaining representative training data **17**
  2.1.1 Training from synthetic data 18
  2.1.2 Building statistical models of 3D human shape and pose . . 23
2.2 Building expressive models for human pose estimation **27**
  2.2.1 Basic pictorial structures based methods 27
  2.2.2 Improving basic pictorial structures 29
  2.2.3 Deep learning methods 32
  2.2.4 Multi-person pose estimation 34
  2.2.5 Relations to our work 35
2.3 Benchmarking and analyzing the state of the art **40**
  2.3.1 Benchmarking human pose estimation 40
  2.3.2 Benchmarking human activity recognition 47

This thesis addresses the challenging interconnected problems of articulated people detection and human pose estimation in unconstrained real world environments. In this chapter we present related work while focusing on the directions explored in this thesis. We conclude each section by relating the works to the contributions of this thesis.

2.1 OBTAINING REPRESENTATIVE TRAINING DATA WITH RELEVANT VARIATIONS

The first direction explored in this thesis is concerned with obtaining training data that is representative of all relevant variations necessary to learn more powerful articulated people detection and human pose estimation models. In particular, we address the difficulties of manually collecting and annotating representative training sets. To that end, we leverage statistical models of 3D human shape and pose from computer graphics to automatically generate and annotate large amounts of synthetic training data with relevant shape and pose variations. Using computer graphics to support object modeling in general and human modeling in particular is obviously not a novel idea. First, in Section 2.1.1, we discuss the approaches leveraging computer graphics to generate synthetic training samples both with realistic and non-realistic appearance. Then, in Section 2.1.2, we present the methods to build computer graphics models of human body shape and pose that can directly be used for generation of synthetic training samples. We conclude each section by relating the discussed works to the contributions of this thesis.

2.1.1 Training from synthetic data

Here we discuss methods training from generated synthetic samples. We first discuss approaches that use synthetic training data in rather constrained scenarios where realistic appearance is not required. Then, we discuss methods training from synthetic data for people detection and pose estimation in more challenging monocular RGB images.

2.1.1.1 *Non-realistic appearance based methods*

We first briefly review human pose estimation and people detection methods based on non-realistic appearance and then discuss generic object detection methods using CAD models for training.

Human pose estimation and detection from silhouettes, depth and infrared data. One of the first to employ artificially created training data were (Grauman *et al.*, 2003), who used a commercially available tool to generate silhouettes for multi-view pose estimation with static cameras. They propose a probabilistic "shape+structure" model that represents the 3D shape of an object class by sets of contours from silhouette views observed from multiple calibrated cameras. Several 3D locations are used to define the corresponding object structure. Then, a prior density over the multi-view shape and structure is computed using a mixture of probabilistic principal components analyzers. At test time, given a novel set of contours, the unknown structure parameters are inferred from the probabilistic reconstruction of the new shape. In order to obtain sufficient amounts of training data to train their complex model they use commercially available 3D animation software to render thousands of pedestrian silhouette images. To that end, realistic humanoid models are manipulated, placed in the simulated scene and rendered as silhouettes from various viewpoints.

A more recent method that uses a commercial tool to generate a large number of synthetic training samples with non-realistic appearance has been proposed by (Shotton *et al.*, 2011). They re-target recorded motion capture data and automatically synthesize a large number of depth images for human body pose estimation with a depth sensor. Generated synthetic training samples closely resemble the depth images available at test time and contain good coverage of the variations due to changes in body pose, shape and viewpoint. These synthetic samples are used to train a classification forest for reliable per-pixel body part classification.

Training from synthetic data obtained from other visual sensors was explored as well. (Broggi *et al.*, 2005) use a simple 3D human shape template to generate synthetic data for a pattern matching approach applied to infrared based people detection. A grey scale 3D template is used to represent different poses and attributes of the human shape rendered from different viewpoints. First, this template is scaled to the person bounding box size and randomly translated within the bounding box to account for small localization errors. Then, matching is performed through a simple and fast cross-correlation function and a threshold is applied to obtain the final result.

The key advantages of these data generation methods is that they represent cheap ways of obtaining a large number of synthetic training samples. At the same time, they do not aim to produce realistically looking samples, but samples that resemble test time input. Constraining themselves to multi-view silhouettes or depth images, they avoid dealing with high appearance variation and background clutter that are typical for RGB images. This significantly simplifies generation of synthetic data and allows to generalize from synthetic samples to unseen testing examples.

Generic object detection based on CAD data. Another line of work explores how CAD models can be leveraged to generate training samples for better car and bicycle detection. (Liebelt *et al.*, 2008) build 3D representations of object classes, which allows to handle viewpoint changes and intra-class variability. To that end, a set of pose and class discriminant features is extracted from synthetic 3D object models and represented by their appearance and 3D position. Object recognition in real images is performed by matching synthetic to real descriptors in a 3D voting scheme. The follow-up work (Liebelt and Schmid, 2010) treats appearance and geometry as separate learning tasks with different training data. A local object appearance model is discriminatively trained from a database of real images, while 3D object geometry is captured using a generative representation built from a database of CAD models. As geometric information is linked to the 2D training data, the method allows to approximately perform 3D pose estimation for generic object classes.

The idea of using CAD models to learn 3D geometry is also explored by (Pepik *et al.*, 2012b). They design a detector that is tailored towards 3D geometric reasoning by extending the discriminatively trained deformable part models (DPM) (Felzenszwalb *et al.*, 2010) to include both estimates of viewpoint and 3D parts that are consistent across viewpoints. Consistency is achieved by imposing 3D geometric constraints on the latent positions of object parts, and the constraints are learned from a set of CAD models. In their follow-up work (Pepik *et al.*, 2012a) they extend the discrete 2D appearance representation to a continuous 3D appearance model by interpolating between discrete viewpoints based on CAD models.

(Stark *et al.*, 2010) also use CAD data to learn appearance, as well as spatial models from a small number of CAD examples and apply their models to car detection. In contrast to other methods they train appearance model from CAD renderings only and show that their method can generalize from synthetic training data to unseen real data.

2.1.1.2 *Realistic appearance based methods*

Here we discuss the methods that learn from synthetic training samples and use RGB images as input at test time. In order to generalize to unseen real examples, this group of methods aim to simulate realistic appearance by producing visually appealing synthetic training samples. This is in contrast to the methods presented above that train and test on non-realistic appearance. First, we present the methods that render realistically looking synthetic training samples. Then we discuss the approaches that apply transformations to real images while preserving their realism.

Rendering realistic appearance. One of first to leverage rendered synthetic training samples were (Shakhnarovich *et al.*, 2003). They present an exemplar based human pose estimation approach that learns a set of hashing functions to efficiently index examples relevant to estimate a particular pose. For each body pose their method finds approximate neighbors in sub-linear time in a large dataset of example synthetic samples generated using a commercially available rendering software.

Another method training from synthetic samples for human pose estimation in monocular RGB images has been proposed by (Okada and Soatto, 2008). They use a piece-wise linear regression method where multiple local linear regressors approximate the nonlinear mapping function from HOG-based feature vectors to 3D poses. Training human poses are randomly generated in a subspace constructed by PCA using the motion capture walking sequences. Human images corresponding to each pose are first rendered by a commercial rendering software and combined with cluttered background of natural images and with uniform background. Then, the poses are clustered using locations of body joints and a local liner regressor is trained from each cluster. This regressor implicitly selects image features that are discriminative for predicting 3D human poses. As this method employs a non-parametric learning approach, it critically depends on the ability to learn from a large set of synthetic training samples.

A more recent approach simulates realistic appearance by using a game engine to generate synthetic training samples from multiple viewpoints (Marin *et al.*, 2010). Generated training data is used to train pedestrian classification models applied to real images. First, an editor is used to create realistic, virtual cities with roads, streets, buildings, traffic signs, vehicles, pedestrians, different illumination conditions, etc., within a video game. Created virtual pedestrians can move through the virtual city respecting physical laws and following their artificial intelligence. Then, the authors play a game and record all encountered virtual pedestrians which are further used to train appearance-based pedestrian classifiers using HOG and linear SVM.

Transforming real images. An alternative way of obtaining synthetic training samples with realistic appearance is to apply transformations to real images preserving their realism. (Enzweiler and Gavrila, 2008) utilize synthesized virtual samples generated from a learned morphable 2D model to improve the classification performance of a discriminative model. A generative model captures prior knowledge about the pedestrian class in terms of several shape and texture models, each specific to a particular pedestrian pose. Active learning is performed to selectively sample generative models to obtain most informative samples for the discriminative learning procedure. Improved performance w.r.t. the original pool of training images has been obtained even though a significant part of the improvement can be achieved by simply adding spatial Gaussian noise (often called jittering) to the training data (Laptev, 2009). The reason for this is that the employed morphable model is still inherently 2D and thus limited in generating relevant shape and appearance variations.

Another line of research explores image-based rendering methods that create novel images by composing existing ones. (Tang *et al.*, 2012) explore this technique to learn occlusion specific people detection models. They observe that typical occlusions are caused by overlaps between people and propose a people detector that is tailored to various occlusion levels. They leverage the fact that person/person occlusions result in characteristic appearance patterns that can help to improve detection results. In order to obtain sufficient amounts of training data to learn a people detector for each occlusion pattern, they generate novel synthetic images. First, for each person, a silhouette is extracted based on the annotated foreground person map. Next, another single-person image is selected arbitrarily and combined with the extracted silhouettes. Such double person images are then used to train a people detector. In their follow-up work (Tang *et al.*, 2013), a similar procedure is used to improve people tracking. They propose a novel joint person detector that combines a single and a double person detector. Their detector explicitly exploits common patterns of person-person occlusions that are a frequent failure case for tracking in crowded scenes.

Recently (Ghiasi *et al.*, 2014) proposed to use an image based rendering method to model occlusions of individual body parts in the context of human pose estimation. They learn deformable models with many local part mixture templates using large quantities of synthetic training data. This allows to learn the appearance of different occlusion patterns, such as the shapes of occluding contours, as well as the co-occurrence statistics of occlusion between neighboring body parts. To train mixture components corresponding to different occlusion patterns, a large corpus of synthetic occlusion data is generated. This is achieved by compositing segmented objects over a base training data set that has been annotated with part locations and figure-ground masks. Occluders are scaled based on object annotations in the base image to produce realistic spatial distributions.

2.1.1.3 *Relations to our work*

Similar to a number of existing works, in Chapters 3, 4, and 5 we propose several methods leveraging synthetic training samples to learn more powerful people detection and pose estimation models. However, in contrast to existing methods, we use a statistical 3D model of human shape (Jain *et al.*, 2010) and a large set of 3D motion capture poses as core components of our approaches. Thus we have full control over pose and shape variations of generated training samples. Furthermore, in contrast to existing methods we develop approaches that are able to generalize well from synthetic training samples to unseen real samples with applications to articulated people detection and pose estimation in challenging monocular RGB scenes.

Similar to (Enzweiler and Gavrila, 2008) there is a generative model behind our synthetic data generation methods. However, in contrast to their method based on a 2D model of shape and texture, in our methods we use a generative 3D human shape model and 3D motion capture data to obtain realistic deformations of 2D data.

This makes our generative model more realistic and versatile.

In Chapter 5 we propose a method to train powerful articulated human pose estimation models from a large number of synthetic training samples, similar to (Shotton *et al.*, 2011). However, in contrast to their work relying on a depth sensor, we address the more challenging problem of human pose estimation in arbitrary RGB monocular images. This also makes our method more general and applicable in diverse unconstrained real world environments. The generality of our method comes with the need to cope with a high degree of intra-class appearance variation and background clutter. This has to be addressed when generating synthetic training samples thus making our data generation task more complex. Furthermore, our data generation methods rely on a generative 3D model of human shape, which allows for arbitrary shape variations, while the method of (Shotton *et al.*, 2011) employs a handful of different characters.

Similar to (Marin *et al.*, 2010), in Chapters 3 and 5 we learn appearance-based models from a large number of synthetic training samples. However, our synthetic training samples are obtained by reusing realistic appearance of the real world data, which is contrast to the game engine data of (Marin *et al.*, 2010) that lacks necessary realism. Direct comparison in Chapter 3 shows that training from synthetic samples with realistic appearances obtained from as many as eleven people performs much better than training from game engine data containing hundreds of distinctive virtual appearances.

In Chapter 4 we present a method to enrich existing training data with relevant shape variations. Contrary to existing works (Shakhnarovich *et al.*, 2003; Okada and Soatto, 2008; Marin *et al.*, 2010), we do not aim to use visually appealing and photo-realisticly rendered data but instead focus on complementary and particularly important information for people detection, namely 3D human shape. Direct comparison to (Marin *et al.*, 2010) shows that our data augmentation method is able to achieve significantly better detection results. On the other hand, we show that unlike other methods training *and* testing on non-realistic appearance (Shotton *et al.*, 2011; Grauman *et al.*, 2003; Broggi *et al.*, 2005), our feature representation can generalize from non-photo-realistic synthetic training data to unseen real test data.

(Liebelt *et al.*, 2008; Liebelt and Schmid, 2010; Stark *et al.*, 2010) propose interesting techniques to learn general object detectors from 3D CAD models. However, these works have not been shown to be applicable to articulated objects such as people. While our work in Chapter 4 uses similar features and rendering procedures, we additionally address the issue of intra-class variations in human shape and pose by automatically generating a large number of 3D models to detect people with large degrees of articulation.

Similar to (Tang *et al.*, 2012, 2013) and (Ghiasi *et al.*, 2014) in Chapters 3 and 5 we obtain synthetic training samples by applying transformations to existing real images preserving their realism. However, in contrast to these methods we use a morphable 3D body shape model to vary shape and pose parameters in 3D and use its projection to compute non-linear 2D image transformations corresponding to realistic 3D shape and pose deformations.

2.1.2 Building statistical models of 3D human shape and pose

We now switch our focus to building statistical models of 3D human shape and pose that can directly be used to generate synthetic training samples. Statistical human shape models represent variations in human physique and pose using low-dimensional parameter spaces and are valuable tools to solve difficult vision and graphics problems, e.g., in body pose tracking or animation. Typically these models learn a probability distribution over human shapes from a dataset of 3D human laser scans. Prior to statistical analysis, the human scans have to be processed and aligned to establish correspondence. In the following, we discuss the related work for each of the aspects of building statistical body models, such as establishing datasets of human scans, developing the scan alignment methods, and creating representations of 3D human shape and pose.

2.1.2.1 *Datasets*

Several datasets have been collected to analyze populations of 3D human bodies. Many publicly available research datasets allow for the analysis of shape and pose variations jointly. (Anguelov *et al.*, 2005) created two datasets: a pose dataset containing 70 scans of a particular person in a wide variety of poses, and a body shape dataset containing scans of 37 different people in a similar pose. For data acquisition they use a multi-view full body scanner that captures the scans with roughly 200, 000 3D points. (Hasler *et al.*, 2009) captured a dataset of 550 full body 3D scans of 114 individuals. 111 individuals were scanned in a standard pose that allows for learning a shape model. In addition, each subject was scanned in several randomly selected poses out of 34 poses, which allows to learn local shape deformations due to pose changes. The dataset provides registered meshes with resolution of 6449 3D points. Recently, FAUST dataset for benchmarking of 3D mesh registration methods has been proposed (Bogo *et al.*, 2014). It includes 300 triangulated meshes of 10 different subjects being professional models each scanned in 30 different poses. The average mesh resolution is 172, 000 vertices. Overall it can be concluded that the publicly available datasets contain data on the order of 100 individuals, which limits the range of shape variations. In contrast, commercially available CAESAR database (Robinette *et al.*, 1999) covers a wide variety of human body shapes. Being the largest dataset of 3D human scans to date, it contains the body shapes of 4, 400 American and European subjects in three different poses: standard standing pose, full coverage pose, and relaxed sitting pose. The database was collected in the course of two years. Selected subjects were solicited to ensure samples for weights, ethnic groups, gender, and geographic regions. The dataset contains high resolution human scans with more than 100, 000 vertices per scan. Each vertex has an assigned confidence score to signalize how reliably it was captured. Each scan has 74 manually placed landmarks and comes with a rich set of body measurements, such as height, waist girth, hip girth, weight, etc.

2.1.2.2 *Statistical shape spaces of 3D human bodies*

Building statistical shape spaces of human bodies is challenging, as there is strong and intertwined 3D shape and pose variability, yielding a complex function of multiple correlated shape and pose parameters. Methods to learn the shape spaces usually follow one of two routes. The first group of methods learn shape and pose related deformations separately and combine them afterwards (Anguelov *et al.*, 2005; Guan *et al.*, 2012; Chen *et al.*, 2013; Neophytou and Hilton, 2013; Jain *et al.*, 2010). These methods are inspired by the SCAPE model (Anguelov *et al.*, 2005) which couples a shape space capturing variation in body shape with a pose space learned from deformations of a single subject. In the original SCAPE model the transformation of each mesh triangle is modeled as combination of two kinds of linear transformations. The first kind represents the pose of the person as global rotation induced by the deformation of an underlying rigid skeleton, while the second kind encodes the individual deformations that originate from varying body shape or non-rigid pose dependent surface deformations, such as muscle bulging. As the model does not explicitly encode vertex position, one needs to solve a complex least squares problem to reconstruct the mesh surface. (Chen *et al.*, 2013) proposed a tensor-based method (TenBo) that jointly models shape and pose deformations using tensor decomposition technique. In contrast to SCAPE, their *TenBo* model effectively leverages training data from multiple people under multiple poses to improve accuracy. Both SCAPE and TenBo methods focus on human body shape only and do not account for any shape variations due to clothing. (Guan *et al.*, 2012) proposed a model for realistic animation of clothing that is learned from a physics-based simulation of clothing on bodies having different shapes and poses. SCAPE model is used to generate synthetic training samples with realistic shape and pose variations. Each synthetic body is manually dressed with each type of garment. Then, for each garment a factorized clothing model is learned that represents rigid rotation, pose independent variations of clothing shape, and pose dependent non-rigid deformations. At the test time the clothing is fit to the body by sequentially fitting each part of the factorized clothing model.

All SCAPE-like models use a set of transformations per triangle to encode shape variations in a shape space. Hence, to convert the vertex coordinates of a processed scan to its representation in shape space, a computationally demanding optimization problem needs to be solved. To overcome this difficulty, a simplified version of the SCAPE model has been proposed that models pose variation with an efficient skeleton-based surface skinning approach (Jain *et al.*, 2010). Laser scans in a standard pose are used to learn a PCA shape model. This shape space only covers variations in overall body shape and not in pose. An articulated skeleton is fitted to the average human shape, and linear blend skinning weights to attach the surface to the bones are computed. The skeleton scales in accordance to the body shape by expressing joint locations relative to nearby surface vertex locations. This representation allows for efficient shape and pose fitting at test time.

Learning shape and pose parameters separately requires training samples with factorized shape and pose variations. However, the poses of individual human scans used to learn statistical shape spaces typically contain slight variations caused by different identities. In this case, directly applying statistical analysis methods may lead to the learned shape spaces that explain body shape variations due to slight changes in pose. To account for these variations, several methods have been proposed. (Wuhrer *et al.*, 2012) factor out variations caused by pose changes by performing PCA on localized Laplacian coordinates. Their method was shown to perform well under significant pose variations of training samples. However, it requires an expensive non-linear optimization procedure. (Neophytou and Hilton, 2013) normalize the pose of each processed scan using a skeleton model and Laplacian surface deformation. While this type of normalization may introduce artifacts around joints when the pose is changed significantly, this approach was shown to work well for pose normalization of samples in a standard standing pose.

In contrast to SCAPE-like methods that learn shape and pose related deformations separately, another group of methods intends to perform simultaneous analysis of both types of variations (Allen *et al.*, 2006; Hasler *et al.*, 2009). These methods learn skinning weights for corrective enveloping of pose related shape variations, which allows to explore both shape and pose variations using a single shape space. Furthermore, it allows for realistic muscle bulging as shape and pose are correlated (Neumann *et al.*, 2013). (Allen *et al.*, 2006) proposed a model that captures both identity and pose dependent shape variations in a correlated fashion, which enables creation of a variety of virtual human characters with realistic and non-linear body deformations customized to the individual. To that end, they build a latent variable model that includes the full set of interpolation keys needed to generate human shape in any pose. This allows to encapsulate the correlation between pose and identity, while keeping these two modalities from being conflated. However, the method is based on a very expensive optimization procedure used to optimize a highly nonlinear function that simultaneously describes pose, skinning weights, bone parameters, and vertex positions. Computationally efficient method was proposed by (Hasler *et al.*, 2009) who analyze body shape and pose jointly by performing PCA on a rotation-invariant encoding of the model triangles. As a downside, their method operates on a low quality meshes and cannot represent high level of details. Thus additional step is required to add the high frequency information after model fitting. It has been shown, however, that for many applications in computer vision and graphics this level of detail is not required and simpler and computationally more efficient shape spaces can be used (Jain *et al.*, 2010; Helten *et al.*, 2013).

2.1.2.3 *Mesh registration*

Mesh registration is performed on the scans to bring them in correspondence for statistical analysis. Two surveys (van Kaick *et al.*, 2011; Tam *et al.*, 2013) review such techniques, and a full review is beyond the scope of this thesis. (Allen *et al.*, 2003) use non-rigid template fitting to compute correspondences between human body shapes in similar pose. The fitting is performed via non-rigid registration of the template to the human scan in the optimization framework by minimizing the combined fitting error. The error function consists of three terms: landmark term describing the goodness of fit of manually placed markers, data term computed as a distance between the vertices of the deformed template and scan, and smoothness term that requires neighboring vertices to move in a similar way. This technique has been extended to work for varying poses (Allen *et al.*, 2006; Hasler *et al.*, 2009), and in scenarios where no landmarks are available (Wuhrer *et al.*, 2011). (Allen *et al.*, 2006) first determine the pose of the scan using the marker positions and approximate pose enveloping weights. Next, the pose of the template is adjusted and body shape fitting is performed. (Hasler *et al.*, 2009) perform pose fitting based on the embedded skeleton. To that end, each revolute body joint is parameterized using a single rotation angle, while the complete body pose is additionally parameterized using global rotation and translation. Then, an ICP based optimization is performed to match the pose of the template model to the pose of the human scan by applying joint rotations in a kinematic chain. Their method requires manually placed landmarks to control the global stability of the pose fitting process. In contrast, (Wuhrer *et al.*, 2011) present a landmark free approach for fitting pose and shape. Their method learns locations of the anthropometric landmarks present in the database of human scans in various poses and automatically predicts locations of the landmarks on the unseen human scan. The predicted landmarks are then used to guide the body pose and shape fitting procedure.

2.1.2.4 *Relations to our work*

We now relate other works to the contributions of this thesis presented in Chapter 6 and concerned with building statistical spaces for 3D human modeling.

Our work is related to other works introducing datasets of 3D human scans (Robinette *et al.*, 1999; Hasler *et al.*, 2009). We systematically construct a model of 3D human shape and pose from the largest dataset of 3D laser scans to date (Robinette *et al.*, 1999) and use publicly available dataset (Hasler *et al.*, 2009) for learning pose and initial shape parameters of our model.

Our methods are similar to the approaches learning shape and pose related deformations separately and combining them afterwards (Anguelov *et al.*, 2005; Guan *et al.*, 2012; Chen *et al.*, 2013; Neophytou and Hilton, 2013; Jain *et al.*, 2010). In particular, we build on the simplified and efficient version of the SCAPE model (Anguelov *et al.*, 2005) proposed by (Jain *et al.*, 2010). In contrast to other SCAPE-like methods (Anguelov *et al.*, 2005; Guan *et al.*, 2012; Chen *et al.*, 2013; Neophytou and Hilton, 2013), our methods are much easier to train and more efficient to fit the shape and pose parameters at test time. Furthermore, we perform a series of experiments to directly compare our shape spaces learned from a large representative dataset of human shapes to the method (Jain *et al.*, 2010) trained from the largest publicly available dataset lacking shape variability. We show that our models significantly outperform their method in terms of statistical quality and fitting accuracy on the task of human body reconstruction from sparse input data.

Similar to other related methods we perform mesh registration of the training scans to bring them in correspondence for statistical analysis. We evaluate different variants of the state-of-the-art techniques for non-rigid template fitting and pose normalization to process the raw data (Allen *et al.*, 2003; Hasler *et al.*, 2009; Wuhrer *et al.*, 2012; Neophytou and Hilton, 2013). Our findings are not entirely new methods, but best practices and specific solutions for automatic pre-processing of large scan databases for learning statistical shape spaces in the best way. Our findings indicate that shape and pose fitting of an initial shape model to a raw scan prior to non-rigid deformation considerably improves the results. Furthermore, we discover that multiple passes over the dataset improve initialization and thus increase the overall fitting accuracy and statistical model qualities. Finally, we show that pose normalization prior to shape space learning leads to much better generalization and specificity and demonstrate that the models trained from the pose normalized samples are able to achieve higher accuracy of fitting to sparse input data.

2.2 BUILDING EXPRESSIVE MODELS FOR HUMAN POSE ESTIMATION

In this section we discuss the related works w.r.t. the second direction of this thesis concerned with building expressive spatial and appearance models for 2D human pose estimation. Compared to the previous section, we shift our focus from encoding relevant variations into the train data to building expressive human pose estimation models. These models, on the one hand, should be flexible enough to capture the complex and highly multi-variate distribution of the human class, and, on the other hand, should be sufficiently discriminative to separate the human body and individual body parts from highly cluttered backgrounds.

Human pose estimation methods typically assume that information about the person's location in the image is provided, either implicitly by using a person-centered image crop, or explicitly by providing a detection bounding box. While this restriction simplifies the task of human pose estimation by discarding large portions of background clutter and other distracting individuals present in the scene, it also allows the methods to focus on the essential task of detecting individual body parts. The majority of methods discussed below are single person pose estimation methods. However, we also briefly discuss methods addressing the more challenging multi-person pose estimation scenario where multiple potentially overlapping people are present in the image.

2.2.1 Basic pictorial structures based methods

Most recent methods for human pose estimation are based on the pictorial structures (PS) model (Fischler and Elschlager, 1973) that represents the body configuration as a collection of body parts and a set of pairwise part relations. Appearance of individual body parts is modeled by part detectors, while pairwise connections between the parts capture their preferred spatial arrangement. Finding the optimal solution in this model can be done in time quadratic in the number of parts.

This model has been made popular by (Felzenszwalb and Huttenlocher, 2005) who proposed an efficient linear time message passing algorithm based on distance transforms that made the inference in tree structured models tractable. This contribution resulted in a wide variety of methods building on and refining this basic PS model.

Several methods were proposed that focused on improving appearance modeling of individual body parts while using basic part connectivity. (Ramanan, 2006) improve the appearance model by extracting more powerful features tuned to a particular image using an iterative parsing approach: in each iteration the color distribution of each body part is estimated using the predicted part location which in turn is used to improve location prediction in the next iteration of parsing. (Ferrari *et al.*, 2008, 2009) further extend this model by integrating features from an automatic foreground segmentation step in order to obtain more powerful part detectors. They increase part detection performance by iteratively reducing the search space of valid articulations. (Eichner and Ferrari, 2009) improve the local appearance model by learning latent relationships between the appearance of different body parts. (Andriluka *et al.*, 2009, 2011) significantly improve performance by using a model that builds on strong discriminatively trained Adaboost part detectors. Their method allows for dense evaluation of part detectors at test time and requires neither iterative parsing, nor search space reduction techniques.

All of the methods discussed above have in common that they heavily rely on strong body part detectors modeled as *single "cardboard"* templates, while tree structured pairwise part interactions are based on simple *geometric* features only. The latter allows for exact and efficient inference. However, these requirements also result in several drawbacks shared across the methods. First, simple geometric features cannot capture richer interactions between connected body parts, such as color/texture similarity or compatibility of contours. Second, rigid "cardboard" part templates can barely capture variations in body part geometry due to out of plane rotations. Moreover, appearance of body parts can be quite generic and rigid part templates often learn characteristic appearance cues, such as two parallel edges, that can easily be confused with background clutter. Third, tree structured connections model dependencies between adjacent body parts only. Hence, important dependencies between non-adjacent body parts (e.g. skin color similarity between lower arms) are not captured. Fourth, higher order part dependencies are not captured by simple pairwise connectivity. Despite high variability of body articulations, human motions and activities often simultaneously constrain the positions of multiple body parts, thus modeling such higher order dependencies is important for effective pose estimation. Fifth, a single body model with fixed parameters for part appearance and geometry can barely capture the high variability in human appearance and pose observed in natural images. Single rigid part templates cannot capture all appearance variations caused by clothing, imaging conditions, part size and articulation. Neither can single pairwise terms account for the highly multi-modal nature of part-part interactions.

The rest of the methods discussed in this section aim to address at least one of these limitations.

2.2.2 Improving basic pictorial structures

We now briefly discuss the improvements addressing limitations of the basic pictorial structures model.

2.2.2.1 *Image dependent pairwise terms*

In order to address the first limitation of the basic PS model, several methods integrating richer part-part interactions have been proposed. (Sapp *et al.*, 2010a) defines a PS model where pairwise terms are image conditioned. Their model relies on silhouette based similarity cues that capture local pairwise part interactions. An effective shape based kernel is used to express pairwise model parameters as kernel regression estimates from a learned sparse set of exemplars. Using image conditioned pairwise parameters in the tree structured models prevents the use of efficient distance transforms and requires more general inference techniques with complexity quadratic in the number of parts. In order to include richer part-part interactions while keeping the inference tractable, (Sapp *et al.*, 2010b) proposed a coarse-to-fine cascade based technique. They learn a sequence of structured models at different pose resolution levels where coarser models prune the state space of the next level by using their marginals. The final level contains much fewer states per part and hence a more general inference method that allows for richer pairwise interactions can be applied. (Karlinsky and Ullman, 2012) proposed another approach that incorporates image information into the pairwise terms. They learn specific linking features that support particular pairwise part configurations. In contrast to other methods, pairwise connectivity structure is not fixed but discovered automatically during training. At test time an approximate inference technique is used.

2.2.2.2 *Flexible body models*

Addressing the second limitation of the basic PS model requires more flexible body models that can account for variations in body part geometry, but also capture more distinctive appearance cues. Modeling appearance of body joints instead of body parts between the joints fulfills these requirements. The appearance of body joints is far less affected by articulations and changes in scale, and exhibits more characteristic traits compared to that of body parts. One of the first to model the appearance of body joints were (Yang and Ramanan, 2011). They replace a rectangular fixed size body part template with two squared templates placed on body joints and an additional squared template between the joints. This allows for a better fit to the ground truth during max-margin learning and increased flexibility of test time part matching. (Sapp *et al.*, 2011) also proposed a flexible body model for pose estimation in videos. They showed that modeling the appearance of body joints is necessary to improve the pose estimation of highly articulated lower arms. Both methods inspired other researchers to use flexible body models. (Rohrbach *et al.*, 2012) use a flexible body model for pose estimation of individuals in a kitchen environment. (Karlinsky and Ullman, 2012) model body joints in a fully connected graphical model. (Dantone *et al.*, 2013) proposed non-linear joint regressors by employing a two-layered random forest. All of these methods conclude that flexible body models improve the pose estimation of highly articulated lower limbs.

2.2.2.3 *Loopy models*

The third group of methods improves over the basic tree structured PS model by adding more pairwise connections between non-adjacent body parts, which leads to a loopy part graph. Time complexity for exact inference in such graphs grows exponentially with the size of the largest clique in the graph. Thus approximate methods are typically used for inference. (Tran and Forsyth, 2010) construct a fully connected graphical model and use approximate search for inference. They show that the full relational model performs better than the tree model, despite using approximate inference. (Yao and Fei-Fei, 2010) models mutual context of object and human pose in human-object interaction activities. They also use a loopy graph and approximate inference, but learn the structural connectivity between the object, overall human pose and different body parts in a max-margin framework. In contrast to the last two works that rely on approximate inference, (Tian and Sclaroff, 2010) employs Branch and Bound (BB) to search for a globally optimal solution in a loopy graphical model. To that end, they re-use the dynamic programming tables computed by the tree model for efficient lower bound look-up. The method is shown to run fast empirically while also slightly outperforming the basic tree structured model. (Sun *et al.*, 2012) also use BB for exact and efficient inference in a fully connected graph. They obtain bounds by relaxing the loopy model into a mixture of star-models and use a specialized data structure along with an efficient search routine.

2.2.2.4 *Mixtures of trees*

Another group of methods improves over basic PS by using mixtures of tree structured models. This accounts for the high multi-modality of human pose and appearance distributions observed in natural images. (Johnson and Everingham, 2010) proposed to learn separate pairwise and appearance terms for different global body configurations. To that end, they cluster poses based on joint annotations and learn a separate tree structured PS model for each cluster. At test time each model is applied to an image and the highest scoring pose configuration is selected. In order to make the scores of different PS models comparable, a separate set of balancing weights is learned. (Dantone *et al.*, 2014) follow a similar strategy to improve pose estimation results while learning much stronger body part regressors using regression forest and auto context. One limitation of these methods is that many global body models are needed to cover the set of possible articulations and much training data is required to train the models, as training data is divided across them. (Yang and Ramanan, 2011, 2013) aim to overcome these shortcomings by proposing a flexible mixture of parts model. Instead of learning global body configurations they employ local mixtures of appearance templates and pairwise terms while also preserving the tree structure. This allows for modeling exponentially many trees using only a few mixture components per body part and part-part connection. This model was further extended in several ways. (Eichner and Ferrari, 2012a) incorporate color as additional feature into the flexible mixtures of parts model and use foreground and background appearance sharing among multiple images at test time. (Desai and Ramanan, 2012) extend the mixtures by additional components learned from the images where a body part is occluded. They show that explicitly modeling the appearance of the occluder helps to improve pose estimation performance in case of partially or even fully occluded body parts. (Wang and Li, 2013a) proposed to learn the tree structure instead of using kinematic tree connections.

Recently, (Kiefel and Gehler, 2014) proposed a Fields of Parts (FoP) formulation that is conceptually different from other mixture models. The key difference is in the underlying graph structure: the presence and absence of a body part at every possible position, orientation, and scale in an image is modeled with a binary random variable. This results in a large number of binary random variables, which is in contrast to the traditional pictorial structures formulation with few random variables, one for each body part, and a large state space. The advantage of this novel formulation is its ability to explain the entire image, i.e. foreground *and* background, while classic PS models explain foreground only. In contrast, the FoP model allows for simultaneous image segmentation and pose estimation. Variables in the model are densely connected across the fields of parts in a kinematic tree structure and within the fields requiring approximate inference. To that end, an efficient marginal inference procedure was proposed. Authors demonstrate significant performance improvements compared to (Yang and Ramanan, 2013) when using the same number of parameters.

2.2.2.5 *Hierarchical models*

Hierarchical methods model higher-order part relations thus addressing one of the shortcomings of the basic PS model relying on simple pairwise part connectivity. Hierarchical methods introduce semi-global parts at different levels of body pose abstraction and connect such parts across the hierarchy while preserving geometric consistency constraints. At the highest level of abstraction higher-order relations between all body parts are simultaneously captured by a global template. This template is then composed of semi-global parts that capture the relations between subsets of body parts at a lower level of the hierarchy. At the lowest level the hierarchical model decomposes into atomic body parts. In order to capture all appearance variations of compositional body parts, mixtures of appearance templates are typically used. (Wang *et al.*, 2011) manually pre-define several levels of body pose abstraction and learn mixtures of appearance templates for each level by clustering corresponding body joints and learning a single template per cluster. Dense connections between compositional parts and their components at lower levels result into a cyclic graph that requires approximate inference with loopy belief propagation. (Sun and Savarese, 2011) follow a similar decomposition strategy. However, their model recursively represents an object as a collection of parts without introducing unnecessary cycles, thus allowing for tractable exact inference. (Duan *et al.*, 2012) propose a hierarchical model where each sub-model is a separate tree based model. This design choice allows for efficient inference based on dual-decomposition. (Sapp and Taskar, 2013) propose a method that explicitly captures a variety of global pose modes and uses a convex objective and joint training for mode selection and pose estimation of atomic body parts.

2.2.2.6 *Pose estimation by detection*

The methods presented so far heavily rely on spatial connectivity models to filter out many false body part detections. Another line of research argued that for the tasks involving complex combinatorial optimization strong detectors are especially important, as they effectively narrow down the search to the relevant part of the search space (Tu *et al.*, 2005). Pose estimation by detection has recently received more attention. This requires models which, in contrast to more traditional approaches, focus on part detection and rely either on loose geometric features (Sapp *et al.*, 2011; Dantone *et al.*, 2013) or ignore them altogether (Shotton *et al.*, 2011; Mittal *et al.*, 2012; Gkioxari *et al.*, 2013). A structured output ranking method is proposed in (Mittal *et al.*, 2012). Their method first generates detection candidates using individual part detectors, combines the candidates and jointly optimizes and ranks the output space using a structured output ranking approach. (Gkioxari *et al.*, 2013) train highly discriminative classifiers that differentiate between multiple arm configurations. The classifiers are discovered in a non-parametric way and are based on rich representations integrating several strong appearance cues. (Ramakrishna *et al.*, 2014) proposed a method to jointly train strong body part detectors while encoding rich spatial interactions among multiple body parts into detectors themselves. In order to deal with the highly multi-modal appearance of individual body parts, they employ a modular architecture based on high capacity predictors.

2.2.3 Deep learning methods

Appearance modeling of all human pose estimation methods discussed so far is based on hand crafted low level appearance feature descriptors, e.g., HOG (Dalal and Triggs, 2005) or Shape Context (Mikolajczyk and Schmid, 2005), that encode edge orientation statistics in local image patches. These features are typically pooled over local spatial regions and sometimes even across scales in order to make the descriptors robust to slight changes in translation and scale, and also to reduce the descriptor size. All descriptor parameters are manually selected and validated on a particular task, and then used unchanged for various applications at hand.

Recently, an alternative approach to manually defined features, representation learning, has shown significant performance improvements for object classification (Krizhevsky *et al.*, 2012), object detection (Girshick *et al.*, 2014; Sermanet *et al.*,

2014) and human pose estimation (Jain *et al.*, 2014; Toshev and Szegedy, 2014; Tompson *et al.*, 2014). This line of methods employs deep learning with convolutional neural networks, where multiple layers of representation invariant to various factors are learned directly from the data. In contrast to body part detection approaches discussed earlier in this section, (Toshev and Szegedy, 2014) formulate pose estimation as a regression problem where a cascade of deep learning regressors is used to directly regress body joint locations from image pixels. To that end, a generic convolutional neural network with five convolutional and two fully connected layers is used. In the first stage, this network is used to jointly regress rough locations of all body joints given a full image as input. Simultaneous regression of multiple body joints avoids the necessity to explicitly design the model topology and interactions between the joints. In order to refine initial joint predictions, further cascades are applied to higher resolution sub-images cropped around initial estimates. Another holistic deep learning-based approach was recently proposed in (Carreira *et al.*, 2016). The proposed deep learning framework operates both on input and output spaces by extracting representations from image pixels and body part location information. Rather than directly predicting the outputs in a single pass, an iterative self-correcting approach progressively refines initially predicted locations of individual body parts by feeding back error predictions. (Ouyang *et al.*, 2014) use a deep learning network to re-score the unary and pairwise features learned using the model of (Yang and Ramanan, 2013). The proposed method can be viewed as a post-processing pose estimation technique that extracts non-linear representations from multiple information sources. Experimental results show significant performance gains due to deep learning based re-scoring. (Jain *et al.*, 2014) proposed a neural network architecture consisting of three convolutional and three fully connected layers to learn individual body part detectors. First, a contrast normalized RGB image patch is used as input. This patch is then processed by three convolutional layers and two max-pooling layers intended to reduce computational complexity and increase robustness to small translations in the input image. As the deeply learned part detectors still produce many false positives, a simple hand designed spatial connectivity model is applied to filter out false detections. These deeply learned part detectors were shown to outperform the part detectors based on mixtures of HOG templates. (Tompson *et al.*, 2014) proposed another deep learning part detection method that uses a sliding-window network architecture to jointly train part detectors with a simple spatial model in a unified learning framework. The model incorporates a multi-resolution input with overlapping receptive fields, which allows the network to see a larger portion of the image with only a slight increase in the number of learned parameters. Dense feature maps are efficiently obtained in a sliding window fashion by performing convolutions on the full image, which avoids redundant evaluations of the learned filters on the overlapping image regions. Experimental evaluation shows significant performance improvements compared to other methods, and most of the performance can be attributed to the part detectors even without using the spatial connectivity model. The follow-up work (Tompson *et al.*, 2015) further improves the part detection performance by using an additional refinement stage that increases the localization accuracy. As in their previous work, a very simple spatial model is used. In contrast, (Chen and Yuille, 2014) proposed a deep learning method based on a much stronger image conditioned spatial model. They extended the model of (Yang and Ramanan, 2013) by conditioning the pairwise terms on deeply learned image representations while also using strong part detectors trained in the same deep learning framework. They demonstrate significant performance gains when using the image conditioned spatial model thereby matching the performance of (Tompson *et al.*, 2014) and outperforming other competing methods. Recently, (Hu and Ramanan, 2016) proposed a bidirectional architectures that combines bottom-up reasoning with top-down feedback, where neural units are influenced by both lower and higher-level units. (Wei *et al.*, 2016) proposed a Convolutional Pose Machines approach being an extension of their prior work (Ramakrishna *et al.*, 2014). They incorporate a convolutional network architecture into the pose machine framework allowing the learning of representations for both image and spatial context directly from the data. In order to address the problem of vanishing gradients when training deep architectures, additional supervision is used in the intermediate network layers. All deep learning approaches discussed so far operate on single monocular images. (Pfister *et al.*, 2015) extended single frame-based human pose estimation model to video domain by combining the outputs of deep learning part detectors with deep optical flow to improve predictions in each frame. The idea is further extended in the follow-up work (Charles *et al.*, 2016) where a personalized body part detection approach automatically adapts to the uniqueness of a person's appearance to improve pose estimation in long videos.

2.2.4 Multi-person pose estimation

Most of the human pose estimation methods discussed above are specifically designed for pose estimation of single isolated individuals that are typically pre-localized using a detection bounding box or a crop around a person. These methods assume that each body part belongs to the same single individual and thus inevitably fail in the presence of multiple interacting people. Here we discuss methods that tackle such challenging cases.

Surprisingly few approaches exist for joint multi-person pose estimation in monocular images. Typically these methods explicitly model interactions between different people by reasoning about occlusions generated by the overlapping body parts of different individuals. (Eichner and Ferrari, 2010) proposed an occlusion probability predictor based on person detection bounding boxes and integrate the occlusion predictions into a multi-person PS model. Additionally, an inter-person exclusion penalty preventing body parts of different people from occupying the same image region has been proposed. Both innovations lead to better pose estimates in group photos, where several persons stand nearby and partially occlude each other. (Ladicky *et al.*, 2013) present a model combining pictorial structures based human pose estimation with per pixel body part labeling in the Markov random field framework. This joint formulation models multiple persons in the image, estimates their body joint locations and additionally infers a pixel-wise body part labeling. This method was shown to work particularly well in cases where some limbs are occluded or one person is partially occluding another. However, their method requires expensive manual labeling of body part segmentations at training time and initial set candidate poses at test time. In contrast to the latter two methods that explicitly reason about occlusions, the approach of (Yang *et al.*, 2012) intended to model multiple interaction patterns between the touching body parts of two individuals. They automatically discover six different patterns, such as, e.g., "hand touches hand", "hand touches shoulder", or "shoulder touches shoulder". For each interaction pattern a flexible mixture of trees model is learned. At test time each model is fitted to the image and the interaction is classified in one of the six classes based on the fitting error. Importantly, this method estimates the locations of interacting body parts only. Recently, (Chen and Yuille, 2015) proposed a single person pose estimation model that performs explicit reasoning about occlusions of individual body parts. To that end, they build on their prior work (Chen and Yuille, 2014) and introduce additional mixture components to handle person-person body part occlusions, and similarly condition component selection based on the CNN features. Their approach primarily focuses on the single-person case and handles multi-person scenes similar to (Yang and Ramanan, 2013), while achieving state-of-the-art performance on the public multi-person pose estimation benchmark (Eichner and Ferrari, 2010).

2.2.5 Relations to our work

In this section we relate the prior work to our contributions w.r.t. building expressive spatial and strong appearance models for single- and multi-person pose estimation in monocular images (Chapters 7, 8 and 11).

Similar to most methods presented above our single person pose estimation approaches in Chapters 7 and 8 are based on tree structured pictorial structures that allow for exact and efficient inference. However, in contrast to the tree based methods that model dependencies between *adjacent* body parts only, our approaches are able to capture important dependencies between *non-adjacent* body parts while still allowing for efficient inference.

Several approaches also model non-adjacent body part dependencies by considered non-tree loopy models (Tian and Sclaroff, 2010; Tran and Forsyth, 2010; Yao and Fei-Fei, 2010; Sapp *et al.*, 2011; Sun *et al.*, 2012). With a few exceptions none of these models consider interactions between body parts that go beyond simple pairwise relationships. In contrast, in Chapter 7 we propose a model that encodes higher-order part dependencies based on a mid-level image representation. To that end, we define a conditional model in which all parts are a-priori connected, but which becomes a tractable PS model once the mid-level features are observed. Our method allows for exact and efficient inference, in contrast to loopy models that require approximate inference.

Our single person pose estimation methods proposed in Chapters 7 and 8 are related to work aiming to increase the flexibility of the PS approach by jointly training a mixture of tree-structured PS models (Johnson and Everingham, 2010; Yang and Ramanan, 2011; Desai and Ramanan, 2012; Eichner and Ferrari, 2012a). Similar to the flexible mixtures of parts methods (Yang and Ramanan, 2011; Desai and Ramanan, 2012; Eichner and Ferrari, 2012a) and in contrast to global body pose mixtures (Johnson and Everingham, 2010) our models rely on local mixture components. In particular, our model can be seen as an exponentially large collection of PS models with a selection function that chooses a suitable model based on the observed image features. Using image conditioned component predictors in our approaches is in contrast to (Yang and Ramanan, 2011) and other flexible mixtures of parts models where selection of unary and pairwise mixture components is performed during inference. Similar to mixture models, our approach allows for efficient inference at test time, yet we are also able to incorporate dependencies between parts that go beyond pairwise interactions. Those are not captured in the model structure but in the conditioning step.

Similar to the hierarchical models (Wang *et al.*, 2011; Sun and Savarese, 2011; Duan *et al.*, 2012; Sapp and Taskar, 2013) our approaches in Chapters 7 and 8 capture higher-order relations between individual body parts. To that end, similar to (Wang *et al.*, 2011), we rely on semi-global poselet detectors of body part configurations. However, in contrast to (Wang *et al.*, 2011) who incorporates poselet detectors directly into the hierarchical graphical model thereby introducing loops and requiring approximate inference, we use poselet detectors as mid-level representation to predict pairwise and unary parameters in our tree structured graphical model. This allows our method to capture higher-order part relations while still enabling exact and efficient inference in the tree structured model. Although (Sun and Savarese, 2011; Sapp and Taskar, 2013) also perform efficient inference, our mid-level representation captures part relations on different levels of body pose abstraction, which makes our methods more powerful.

Similar to methods introducing richer pairwise dependencies between body parts by conditioning the pairwise potentials on the image (Sapp *et al.*, 2010a,b; Karlinsky and Ullman, 2012), in Chapter 7 we define a PS model where unary and pairwise terms are image conditioned. However, our method is more general as it implicitly models dependencies between multiple parts by using an intermediate poselet based feature representation. In contrast, the methods (Sapp *et al.*, 2010a,b; Karlinsky and Ullman, 2012) rely on edge based similarity cues that are ineffective in the presence of background clutter. These cues are extracted from small local patches and thus capture mostly local pairwise part interactions. Capturing the dependencies simultaneously between multiple body parts makes our method applicable in challenging real world scenarios where highly articulated humans are seen from different viewpoints (see results in Chapter 9), while the methods of (Sapp *et al.*, 2010a,b; Karlinsky and Ullman, 2012) were applied to frontal poses only with a relatively small degree of articulation.

Our methods in Chapter 8 are related to (Yang and Ramanan, 2011; Desai and Ramanan, 2012; Eichner and Ferrari, 2012a), as our methods explore strong local appearance representations modeled as mixtures of rotation specific and rotation invariant templates. We directly compare both types of detectors and show the superior performance of rotation specific mixtures, which underlines the consideration that the appearance of individual body parts has rotation dependent characteristic traits. Our approaches are also related to the pose estimation by detection methods which learn specialized detectors tailored to the appearance of salient body parts, such as hands, head and torso (Mittal *et al.*, 2012; Gkioxari *et al.*, 2013). We show that strong head and torso detectors are important in the presence of significant background clutter. Furthermore, our methods are related to (Bourdev *et al.*, 2010; Wang *et al.*, 2011) that model the appearance of part configurations using semi-global representations based on poselet features, yet these models have not been shown to lead to the state-of-the-art performance in human pose estimation. In contrast to other methods typically relying on a single kind of detectors, in Chapter 8 we combine different appearance representations and show that strong representations operating at different levels of granularity (mixtures of local templates vs. semi-global poselets) are complementary. We demonstrate that when augmented with the complementary appearance models, the basic tree-structured pictorial structures models perform better than other methods based on mixtures of trees (Yang and Ramanan, 2011; Eichner and Ferrari, 2012a; Dantone *et al.*, 2014). Finally, we show that combining our complementary appearance representations with the image conditioned spatial model leads to the best pose estimation performance.

Similar to (Yang and Ramanan, 2011; Sapp *et al.*, 2011; Karlinsky and Ullman, 2012; Dantone *et al.*, 2013), our approach in Chapter 8 relies on a flexible body model and models the appearance of body joints. However, in contrast to (Sapp *et al.*, 2011; Karlinsky and Ullman, 2012; Dantone *et al.*, 2013) we include additional cardboard body parts in between the joints, which makes our appearance modeling stronger and allows us to improve overall pose estimation performance.

Our methods in Chapters 7 and 8 are related to deep learning pose estimation approaches (Chen and Yuille, 2014; Jain *et al.*, 2014; Tompson *et al.*, 2014; Toshev and Szegedy, 2014; Carreira *et al.*, 2016; Wei *et al.*, 2016). Similar to these methods, our approaches rely on strong part detectors that, in contrast, are based on hand crafted image representations, such as HOG or shape context. As we show in Chapter 11, deep learning significantly improves part detection performance. We envision that replacing hand crafted appearance representations by deep learning part detectors will significantly improve the performance of our Poselet Conditioned Pictorial Structures approach. Similar to (Chen and Yuille, 2014), we rely on mid-level representation to predict the pairwise potentials in the graphical model. However, our mid-level representation is based on poselet detectors obtained by independently training shape context templates with AdaBoost. This is in contrast to more powerful deep learning features used as a mid-level representation by (Chen and Yuille, 2014). We expect that using deep learning features to condition pairwise and unary potentials of our model will further increase pose estimation performance. Prediction of pairwise potentials in our case is done globally for the whole image, which makes our inference efficient, while (Chen and Yuille, 2014) predict locally using the features extracted from a local image patch. Similar to (Toshev and Szegedy, 2014; Carreira *et al.*, 2016) who use holistic approaches to directly regress the positions of body joints given deep learning features, we use semi-holistic poselet detectors to predict the unary and pairwise parameters in our model. In contrast to their work, we perform inference on the graphical model given the predicted model parameters and complementary local part appearance modeling obtained by sliding window part detectors, which allows our method to recover from wrong predictions and perform well even if image evidence for certain body parts is weak. Our methods are similar to (Jain *et al.*, 2014; Tompson *et al.*, 2014, 2015) who rely on strong body part detectors applied in a sliding window fashion. In contrast to their methods using a weak pairwise part connectivity, our methods are based on a more expressive image conditioned spatial model that captures the higher-order dependencies between multiple body parts. We believe that these deep learning methods will also profit when combined with our spatial model.

Our multi-person pose estimation method in Chapter 11 is related to (Eichner and Ferrari, 2010; Yang *et al.*, 2012; Sun and Savarese, 2011; Ladicky *et al.*, 2013; Chen and Yuille, 2015). Similar to these approaches we perform pose estimation of multiple people. However, in contrast to these approaches that focus on the upper body with little variability in pose and consider a simplified case of people seen from the front, we address the more difficult problem of full body articulated pose estimation in challenging real world scenarios. To that end, similar to (Chen and Yuille, 2015) and in contrast to (Eichner and Ferrari, 2010; Yang *et al.*, 2012; Sun and Savarese, 2011; Ladicky *et al.*, 2013) building on hand-crafted HOG image features, we rely on deep learning representations to model strong body part detectors. Similarly to (Chen and Yuille, 2015), we aim to distinguish between visible and occluded body parts. However, in contrast to their approach that primarily focuses on the single-person case and handles multi-person scenes akin to (Yang and Ramanan, 2013), our approach jointly reasons about body part interactions across several people. Furthermore, unlike (Chen and Yuille, 2015), our approach is not limited by the number of possible occlusion patterns and cover person-person occlusions and other types as truncation and occlusion by objects in one formulation. Direct comparison to (Chen and Yuille, 2015) on a prominent multi-person pose estimation benchmark (Eichner and Ferrari, 2010) shows significantly better performance of our joint multi-person model. In contrast to (Yang *et al.*, 2012) who model interactions between two individuals, our method works for an arbitrary number of people. Unlike their method which is restricted to modeling a few interaction patterns in a chain structure between touching body parts of two people, our approach can model arbitrary interactions between individuals by using a fully connected graphical model. In contrast to (Eichner and Ferrari, 2010; Ladicky *et al.*, 2013) relying on a person detector to generate initial hypotheses for the joint model, our approach does not require a person detector and jointly solves the tasks of detection and pose estimation: it infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other. Unlike (Eichner and Ferrari, 2010), our multi-person approach is not limited by a number of occlusion states among people. We directly compare to (Eichner and Ferrari, 2010) on their multi-person benchmark and demonstrate significant performance improvements by our models. Unlike (Ladicky *et al.*, 2013) who resorts to a greedy approach by adding one person hypothesis at a time until the joint objective can be reduced, our formulation can be solved with a certified optimality gap. In addition, our method does not require expensive manual labeling of body part segmentation at training time, unlike the model of (Ladicky *et al.*, 2013). In contrast to (Sun and Savarese, 2011) who perform independent pose estimation for each individual, we use an expressive fully connected spatial model that allows for joint partition and body part labeling of all individuals present in the image.

Our multi-person models in Chapter 11 are also related to deep learning approaches. In particular, we propose a strong part detection model by building on the state-of-the-art generic object detector based on convolutional neural networks (Girshick *et al.*, 2014; Girshick, 2015). In contrast to their methods using selective search (Uijlings *et al.*, 2013) to generate object proposals, we rely on our strong DPM based body part detectors presented in Chapter 8 to generate reliable detection proposals for body parts. We carefully evaluate multiple design choices required for the best body part detection performance and empirically demonstrate significant improvements of our detection model over (Girshick *et al.*, 2014; Girshick, 2015) on the task of human pose estimation. Furthermore, we also propose a strong fully-convolutional body part detection model that outputs dense scoremaps, similar to (Jain *et al.*, 2014; Chen and Yuille, 2014; Tompson *et al.*, 2014, 2015; Wei *et al.*, 2016). However, unlike (Tompson *et al.*, 2014, 2015; Wei *et al.*, 2016), we use single size receptive field and do not include multi-resolution context information. In contrast to (Carreira *et al.*, 2016; Wei *et al.*, 2016) we developed a single stage single level fully-convolutional part detection architecture that is much easier to train. Appearance and spatial components of our approaches are trained piece-wise, which is in contrast to (Tompson *et al.*, 2014). We envision that using multi-resolution filters and joint training should improve the performance of our models as well. Similar to (Tompson *et al.*, 2015), we improve body part localization by using location refinement. However, in contrast to their method that trains an additional CNN for location refinement, we employ a much simpler strategy by optimizing a joint classification and location refinement objective function during CNN training. While the approaches (Jain *et al.*, 2014; Chen and Yuille, 2014; Tompson *et al.*, 2014, 2015; Wei *et al.*, 2016; Carreira *et al.*, 2016) target the problem of pose estimation of single isolated individuals, we address the much more challenging task of joint pose estimation of multiple interacting people. Most importantly, we propose a novel powerful spatial model that allows us to correctly assign the body part detections to the corresponding individuals. As our model can handle single person cases as well, we directly compare to prior single person pose estimation methods (Chen and Yuille, 2014; Tompson *et al.*, 2014, 2015) on challenging single person benchmarks and demonstrate significant performance improvements.

2.3 BENCHMARKING AND ANALYZING THE STATE OF THE ART

After discussing the prior works on human pose estimation and their relation to our approaches we change our focus on discussing the works concerned with the benchmarking and performance analysis of human pose estimation methods (Section 2.3.1). As the third direction explored in this thesis is also concerned with performance analysis of popular methods in human activity recognition, we additionally review the advances in this area and discuss current benchmarks in Section 2.3.2.

2.3.1 Benchmarking human pose estimation

Here we discuss the datasets created for benchmarking human pose estimation methods. As we are interested into 2D human pose estimation in monocular images, we concentrate on the datasets established for this setting. Typically these datasets come with labeled 2D positions of individual body joints or stick annotations of body parts (also known as *stickmen*), while some of the datasets additionally provide body joint occlusion labels. Two different body joint annotation settings defining the complexity of pose estimation were proposed. In the first setting body joints are annotated in the *observer centric (OC)* fashion meaning that the left body joints are the ones located on the left w.r.t. the line connecting the mean of shoulder annotations and mean of hip annotations, and the right body joints are the ones laying on the right side w.r.t. this line. This setting has been proposed in order to simplify the pose estimation task and minimize the influence of prediction errors due to incorrect estimation of the front/back person viewpoint. *Person centric (PC)* annotations label left/right body joints according to the *true* left and right of the shown person. PC body joint annotations correspond to the harder pose estimation task requiring the correct estimation of the person's viewpoint. The datasets provided with the stickmen annotations follow the OC approach. In the following we discuss related pose estimation benchmarks and point out which of the two annotation settings is chosen in each case. First, we discuss the datasets created for full body human pose estimation. Then, we present the datasets focusing on the task of upper body pose estimation. Finally, we discuss various evaluation metrics used to measure the performance of human pose estimation methods.

2.3.1.1 *Full body human pose estimation*

Image Parsing (IP) (Ramanan, 2006). The IP dataset was one of the first datasets proposed for benchmarking 2D human pose estimation methods. It consists of 100 training and 205 testing images of fully visible people performing a diverse set of activities such as sports, dancing and acrobatics. The images were collected from previous sport datasets and personal photos. All images were rescaled such that the height of the individuals, if they were standing upright, is roughly 150 pixels. The dataset is provided with OC annotations of all body joints.

Leeds Sports Poses (LSP) (Johnson and Everingham, 2010). The LSP dataset is one of the most widely used human pose estimation benchmark to date. It consists of images collected from Flickr and showing people involved in eight different sports, namely athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball. The images exhibit strong variations in articulation and viewpoint. The dataset includes 1,000 people for training and 1,000 for testing. Both subsets are provided with ground truth body joint annotations including position and occlusion labels. Each annotated individual was cropped from the original full size image around the labeled body joints. Then, each crop was rescaled such that the person is roughly 150 pixels high, similar to IP dataset. Originally this dataset was provided with the PC body joint annotations. However, (Eichner and Ferrari, 2012a) proposed a simplified pose estimation task and released OC annotations widely used for evaluation.

Leeds Sports Poses Extended (LSPE) (Johnson and Everingham, 2011). The follow up work (Johnson and Everingham, 2011) addresses the limitation that the training set of LSP is rather small and extends the training set by 10,000 additional images collected from Flickr. Most of the collected training images show people performing gymnastics, parkour and athletics. The PC body joint labels were obtained using Amazon Mechanical Turk (AMT) and include a significant portion of noise due to imprecise localization and errors in body structure. Similar to the original LSP dataset, each annotated individual was cropped and rescaled to be roughly 150 pixels high.

UIUC People (Tran and Forsyth, 2010). The UIUC People dataset contains 593 images of people in variable body poses playing different sports. The majority of images show people playing badminton. 346 images are used for training and 247 are used for testing. In contrast to the IP and LSP datasets, no scale normalization is performed. This dataset is provided with PC annotations of all body joints.

UIUC Sport (Wang *et al.*, 2011). The UIUC Sport dataset intends to overcome the limitation of the UIUC People dataset that focuses mostly on people playing badminton. It includes a more diverse set of about 20 sport categories, such as acrobatics, American football, croquet, cycling, hockey, figure skating, soccer, golf, and horseback riding. There are in total 1,299 images, out of which 649 are randomly chosen for training and the rest for testing. Similar to UIUC People, this dataset is provided with PC body joint annotations.

FashionPose (Dantone *et al.*, 2013). The FashionPose dataset includes 7,543 images of people downloaded from a variety of fashion blogs. This dataset focuses on pose estimation challenges due to highly variable appearance. That said, pose variability in this dataset is not high and mostly limited to upright standing people. Each image contains a single person with the entire body visible. The dataset comes with PC body joint annotations augmented with visibility flags.

PASCAL Person Layout Challenge (Everingham *et al.*, 2010). This dataset and the corresponding evaluation strategy are different from the above conventional pose estimation benchmarks. This challenge requires correct prediction of the presence/absence of the head, hands and feet *and* correct prediction of the corresponding bounding boxes of visible body parts. This is in contrast to other benchmarks that typically require correct location prediction of individual body joints and assume that all body joints are always present in the image. The Person Layout dataset contains 850 images for training and 849 images for testing with annotated bounding boxes around the head, hands and feet. Similar to other full body benchmarks, images were obtained from Flickr, but, in contrast, are not restricted to sport activities and fully visible persons and often contain people with some of the body parts truncated or occluded.

2.3.1.2 *Upper body human pose estimation*

Buffy Stickmen (Ferrari *et al.*, 2008). The Buffy Stickmen dataset was the first benchmark for upper body human pose estimation. This dataset consists of frames extracted from several episodes of the popular TV show "Buffy the Vampire Slayer". It contains 472 training and 276 testing images of people labeled with stickmen annotations. Each frame may contain multiple people. In order to correctly associate the detections to multiple persons at test time, an upper body people detector is used, and only the highest scoring detection per person matching the ground truth bounding box is considered. While being a popular upper body pose estimation benchmark until recently, it lacks the variability in body poses and suffers from low contrast in many frames, as noted by (Tran and Forsyth, 2010).

ETHZ PASCAL Stickmen (Eichner and Ferrari, 2009). This dataset is a subset of the PASCAL VOC 2008 dataset (Everingham *et al.*, 2008) containing 549 annotated test images of roughly upright frontal people with at least the upper body visible. The dataset consists mainly of amateur photographs with difficult illumination and low image quality, which makes it more challenging compared to the Buffy Stickmen dataset. At the same time, the pose variability is not high either. Body parts are labeled using stickmen annotations.

Video Pose 2.0 (Sapp *et al.*, 2011). The Video Pose 2.0 consists of video clips extracted from the TV shows "Friends" and "Lost". Clips were hand picked to represent a wide range of lower/upper arm articulations. The dataset consists of 1,286 frames distributed between 44 short clips each 2-3 seconds long. 26 clips are used for training and 18 for testing. In contrast to other pose estimation benchmarks containing still frames, this dataset encourages the usage of motion information. It focuses on pose estimation of upper and lower arms only annotated with OC body joint labels. The global scale and translation of the person is fixed to avoid the influence of people detection errors on pose estimation accuracy. Similar to the Buffy Stickmen and ETHZ PASCAL Stickmen datasets, this dataset is limited to roughly upright frontal poses.

Frames Labeled In Cinema (FLIC) (Sapp and Taskar, 2013). The FLIC dataset contains 5,003 still frames (1,016 used for training) automatically extracted from popular Hollywood movies. To that end, first a people detector was applied on every tenth frame of 30 movies. Then, the upper body joints of people detected with high confidence were annotated with OC labels using AMT. Images containing occluded or severely non-frontal people were manually rejected, which shifted the focus of the dataset towards frontal upright people, similar to previous upper body pose estimation benchmarks.

Armlets (Gkioxari *et al.*, 2013). The Armlets dataset consists of 9,593 training and 2,996 testing people with PC annotated body joints of upper/lower arms. The training set of the Armlets dataset uses the images from the PASCAL VOC 2011 (Everingham *et al.*, 2011a) (person category and action recognition challenge) and H3D (Bourdev and Malik, 2009) datasets. Testing images are taken from the validation set of PASCAL VOC 2009 (Everingham *et al.*, 2009). The dataset contains 2 people per image on average. The majority of people in the test set are frontal upright.

We Are Family Stickmen (Eichner and Ferrari, 2010). In contrast to single person pose estimation benchmarks presented above, this dataset is intended to serve as a test bed for multi-person pose estimation approaches. The dataset consists of group photos, such as, e.g., classmates, music bands, or sport teams, collected by querying Google image search and Flickr with words like "family", "band", "team", etc. In total 525 image were collected with 6 people each on average. 350 images are used for training and 175 for testing. The images contain frontal upright people that often occlude each other. Upper body stickmen annotations are used for evaluation.

Synchronic Activities Stickmen (Eichner and Ferrari, 2012b). This dataset was introduced for human pose co-estimation of multiple individuals. It was collected from the Internet and consists of 357 images with a total of 1,128 persons all used for evaluation. Each image contains multiple people performing cheer-leading, aerobic and dancing activities and being in roughly synchronized poses. The dataset is provided with OC stickmen upper body part annotations used for evaluation. The images cover frontal upright people with a high variability of arm articulations.

Proxemics (Yang *et al.*, 2012). The Proxemics dataset is intended to cover multiple interaction patterns between two people. The dataset was collected from personal photos of family and friends queried from Flickr, Getty Images, and using Google and Bing image searches. For querying the images, abstract concepts such as "arm around shoulder", "hugging", "holding hands", etc., were used. The images show mostly frontal upright people. Each pose was labeled with OC body joint annotations. Furthermore, for each pair of people, one of six pre-defined touch codes between pairs of body parts was labeled as well.

2.3.1.3 *Evaluation measures*

We now briefly describe the measures used for evaluation of 2D human pose estimation methods.

Percentage of Correct Parts (PCP) (Ferrari *et al.*, 2008). According to the definition of the PCP, a body part estimation is considered correct if *both* of its endpoints are closer to their ground truth positions than a threshold. As a distance threshold 0.5 of the ground truth segment length is used. PCP is then computed as a percentage of correctly predicted body segments out of the total number of segments. This evaluation metric requires correct prediction of both body joints belonging to the same limb, thus enforcing consistent predictions for body limbs. However, the PCP metric also has a drawback that foreshortened body parts should be localized with higher precision to be considered correct. This makes the task of pose estimation of foreshortened forearms and lower legs harder compared to other body parts.

Percentage of Correct Keypoints (PCK) (Sapp *et al.*, 2011). In contrast to the PCP metric used for evaluation of body limb matching, PCK measures the accuracy of localization of individual body joints. The body joint prediction is considered to be correct, if predicted location falls within a matching distance from the ground truth. The accuracy is plotted as a function of the whole range of matching thresholds. The matching distance depends on the image scale, and several works proposed to relate the matching distance to image annotations. (Yang and Ramanan, 2013) use the portion of the height of the person bounding box enclosing all body joints as a matching distance. The disadvantage of this definition is that it makes the matching distance dependent on body articulation. For instance, for the upright person with hands up the matching distance computed in this way is higher compared to the person with hands down. In order to overcome this limitation, (Sapp and Taskar, 2013) proposed to use the portion of the ground truth torso height as matching distance. As the torso is less deformable compared to the whole human body, the definition of matching distance based on torso height is more robust and not affected by body part articulations. However, the matching distance defined in this way is strongly affected by torso foreshortening due to out-of-plane rotations.

Average Precision of Keypoints (APK) (Yang and Ramanan, 2013). The APK metric has been proposed in order to evaluate the accuracy of body joint detection using precision-recall curves. Similar to PCK, it counts the joint to be correctly detected if the predicted location falls within a matching distance from the ground truth joint location. However, unlike PCK and PCP that consider only a single highest scoring detection per part, APK considers all detections while using the detection scores as confidence. Thus APK correctly penalizes both missing detections *and* false positives. This measure can be used directly to evaluate detection accuracy of body parts of multiple people present in the image.

Multi-person PCP (Ferrari *et al.*, 2008; Eichner and Ferrari, 2010). As PCP always considers a single detection per body part, it cannot directly deal with multiple people present in the image. In order to address this scenario, first, a person detector is used to associate body part detections to corresponding persons, and then the PCP evaluation is performed for each person independently. Typically only the people are considered whose detections match the ground truth bounding boxes. This biases evaluation towards people in more simple and common poses that are easier to detect, while ignoring harder and thus more interesting cases.

2.3.1.4 *Relations to our work*

We now relate the prior work to our contributions presented in Chapter 9 w.r.t. benchmarking and analysis of the state-of-the-art in 2D human pose estimation.

Existing benchmarks focus on individual aspects of the human pose estimation task, such as sport scenes (Johnson and Everingham, 2010; Tran and Forsyth, 2010; Wang *et al.*, 2011), frontal-facing people (Ferrari *et al.*, 2008; Eichner and Ferrari, 2009; Sapp *et al.*, 2011; Sapp and Taskar, 2013; Dantone *et al.*, 2013), pose estimation in group photos (Eichner and Ferrari, 2010; Yang *et al.*, 2012), and pose estimation of people performing synchronized activities (Eichner and Ferrari, 2012b). Thus these benchmarks are still limited in their scope and variability of considered activities. In contrast, our dataset covers a much wider variety of human activities including various recreational, occupational and householding activities, and captures people from a wider range of viewpoints. The key rationale behind our data collection strategy is that we want to represent both common and rare human poses that might be missed when simply collecting more images without aiming for good coverage. To that end, we specifically use an established taxonomy of over 800 human activities (Ainsworth *et al.*, 2011) to guide the data collection process. This results in a diverse set of images covering not only different activities, but also indoor and outdoor scenes, a variety of imaging conditions, as well as both amateur and professional recordings.

Our dataset is related to the PASCAL Person Layout Challenge (Everingham *et al.*, 2010), as it contains a significant number of people with occluded and truncated body parts. However, in contrast to (Everingham *et al.*, 2010), we require correct estimation of *all* body parts, not only head and lower arms/legs. In contrast to other full body and upper body pose estimation benchmarks, we do not make any assumption that all body parts or at least all upper body parts are present in the image. This makes the task of pose estimation harder, as it requires dealing with the variable number of body parts.

Our work is related to earlier datasets, such as Image Parsing (Ramanan, 2006), Buffy Stickmen (Ferrari *et al.*, 2008) and ETHZ PASCAL Stickmen (Eichner and Ferrari, 2009). However, the small training sets of few hundreds of annotated people included in these datasets make them unsuitable for training models with complex appearance representations and multiple components (Johnson and Everingham, 2011; Dantone *et al.*, 2014; Tompson *et al.*, 2014), which have been shown to perform best. In contrast, our dataset includes a large training set with over 28, 000 individuals with annotated body joints, which allows for training rich pose estimation models. Although datasets with large training sets exist (Johnson and Everingham, 2011; Gkioxari *et al.*, 2013; Sapp and Taskar, 2013), they cover individual aspects, such as particular sports and frontal upright people. In contrast, our dataset covers hundreds of everyday human activities, which results in a wide variety of poses, appearances and viewpoints, while being 3-5 times larger. In contrast to other benchmarks, also the test set is much larger, thus enabling thorough performance analysis on subsets of data and allowing for more solid conclusions.

Similar to other pose estimation benchmarks, we provide body joint annotations. We choose to use the more challenging PC labeling setting that requires both correct localization of the body parts along with the correct match w.r.t. the left/right assignment. Similar to LSP (Johnson and Everingham, 2010) and LSPE (Johnson and Everingham, 2011), we also annotate visibility of body joints. However, in contrast to all related 2D human pose estimation benchmarks, we additionally provide a rich set of labels, such as visibility of limbs, 3D orientation of head and torso, locations of eyes and nose, and activity labels. These rich annotations allow for a detailed analysis of various factors influencing the performance of the state-of-the-art approaches, such as foreshortening, occlusion, viewpoint and activity, previously not possible in this level of detail. We thoroughly analyze the performance and evaluate the robustness of current approaches to various challenges of articulated pose estimation and identify the limitations of current methods.

Similar to (Johnson and Everingham, 2011), we use AMT to obtain body joint labels. However, in contrast to their work, we carefully pre-select the qualified workforce based on a qualification task, and then maintain label quality by manually inspecting the annotated data. This results in reliable body labels that can be directly used to train supervised learning methods, which is in contrast to (Johnson and Everingham, 2011) who need to take the label noise into account.

Our work is also related to multi-person pose estimation datasets (Eichner and Ferrari, 2010, 2012b; Yang *et al.*, 2012), as a substantial number of people (over 30%) are in groups of two or more individuals. However, in contrast to other datasets restricted to frontal upright people in group pictures (Eichner and Ferrari, 2010), personal photo collections (Yang *et al.*, 2012) or synchronized sport-related activities (Eichner and Ferrari, 2012b), our dataset covers a wider variety of poses, viewpoints and clothing types of people interacting with each other in real world environments. We establish a multi-person pose estimation task in Chapter 11.

Similar to the Video Pose 2.0 dataset (Sapp *et al.*, 2011), for each image in our benchmark we provide a short video clip containing preceding and following frames. This facilitates the usage of motion information and makes our dataset a valid benchmark for human pose estimation methods in videos. In contrast to the dataset (Sapp *et al.*, 2011) restricted to frontal upright people from popular TV shows, we used YouTube that allows to access a rich collection of videos originating from various sources, including amateur and professional recordings and capturing a variety of public events and performances. Using YouTube videos gives us access to real world data with uncontrolled variations in pose, clothing, viewpoint, scale and imaging conditions, data where people are interacting with other people and objects in unconstrained real world environments.

Our work is also related to other works establishing evaluation measures for single person 2D pose estimation. We address the main issues of PCP (Ferrari *et al.*, 2008) and PCK (Sapp *et al.*, 2011) by improving both metrics. We address the limitation of PCP that matches distance depending on ground truth limb length thus making the task of estimating the foreshortened limbs harder. To that end, we define a new metric that uses matching distance depending on the mean ground-truth segment length over the entire test set, but otherwise follows the definition of PCP. We improve the PCK measure that uses a fraction of the enclosing person bounding box height as matching threshold (Yang and Ramanan, 2013) and thus strongly depends on body articulations. We propose a slight modification of PCK and define the matching threshold as a portion of the head segment length. We choose to use head size to make the metric articulation independent. At the same time, the head is less frequently foreshortened compared to the torso, which makes our definition also more robust than the one proposed by (Sapp and Taskar, 2013).

Similar to (Eichner and Ferrari, 2010; Sun and Savarese, 2011), we introduce an evaluation metric suited for multi-person pose estimation. In contrast to (Eichner and Ferrari, 2010) who use upper body detector and perform PCP evaluation only for detections matching the ground truth bounding boxes, our evaluation metric considers all predictions in the image. This addresses a harder problem, as it is not restricted to the correct upper body detections only and takes into account false positive predictions. At the same time, this makes evaluation independent from a choice of people detector. Similar to (Sun and Savarese, 2011; Yang and Ramanan, 2013), we use APK as evaluation measure. However, in contrast to their works that aim to evaluate the detection of *any* instance of the body part class in the image, we evaluate consistent body part configurations by assigning entire pose prediction to the ground truth based on the highest PCK score. Our metric requires that only single pose prediction can be assigned to the ground truth pose and counts unassigned predictions as false positives. We compute AP for each body part prediction and report APK as total performance over all body parts.

2.3.2 Benchmarking human activity recognition

We now switch our focus to benchmarking and performance analysis of current action and activity recognition methods in videos. In general, one should differentiate between recognition of short atomic *actions*, such as "catch the ball" or "hit the ball", from recognition of human *activities* consisting of multiple atomic actions, such as "playing soccer". However, with slight abuse of notation, for the rest of this section we will use the word "activity" for both human actions and activities. First, we briefly summarize the advances in human activity recognition. Then, we discuss the most related activity recognition benchmarks.

2.3.2.1 *Advances in activity recognition*

Here we review the most relevant methods for recognition of human activities in video. We specifically focus on how to represent activities for the purpose of recognition. We distinguish three groups of methods: holistic representations characterizing human activities in terms of low level features, pose based representations using higher level encoding of activities in terms of body pose and motion, and deep learning methods where multiple layers of representation invariant to various factors are learned directly from the data. Given these representations, a codebook based quantization is typically performed, followed by learning a linear classifier that tells apart samples belonging to different activity classes. In the following, we focus on different representations extracted for recognition of human activities.

Holistic representations for activity recognition. Holistic appearance based features combined with the Bag-of-Words representations (Laptev *et al.*, 2008; Duchenne *et al.*, 2009; Wang *et al.*, 2013; Wang and Schmid, 2013; Jhuang *et al.*, 2013) have been considered the de facto standard for human activity recognition in video. Many methods create discriminative feature representations of a video by first detecting spatio-temporal interest points (STIP) (Chakraborty *et al.*, 2011; Laptev, 2005), or sampling them densely (Wang *et al.*, 2009), and then extracting various feature descriptors in the space-time volume. (Laptev, 2005) builds on the ideas of Harris and Förstner of using interest point operators and detects local spatio-temporal structures where the image values have significant local variations in both space and time. The spatio-temporal extent of the detected events is estimated by maximizing a normalized spatio-temporal Laplacian operator over spatial and temporal scales. (Chakraborty *et al.*, 2011) improve over other STIP detectors by detecting more repeatable, stable and distinctive points, while suppressing unwanted detections on the background. To that end, surround suppression combined with local and temporal constraints is performed. (Wang *et al.*, 2009) perform an evaluation of several STIP detectors in a common experimental setup and compare their performances to dense sampling of space-time features. They demonstrate that dense sampling on a regular space and time grid consistently outperforms all tested STIP detectors for human activities in realistic settings.

Arguably the most popular holistic approach to date is the Dense Trajectories (DT) method (Wang *et al.*, 2013) that tracks dense feature points and extracts strong appearance based features along the trajectories. In particular, DT computes histograms of oriented gradients (HOG) (Dalal and Triggs, 2005), histograms of flow (HOF) (Laptev *et al.*, 2008), and motion boundary histograms (MBH) (Dalal *et al.*, 2006) around densely sampled points that are tracked for 15 frames using median filtering in a dense optical flow field. In addition, the method extracts geometric features, such as x and y displacements between neighboring frames in a trajectory. The follow up work (Wang and Schmid, 2013) improves the DT method by removing trajectories consistent with camera motion. To that end, camera motion is estimated by matching feature points between frames using SURF (Willems *et al.*, 2008) descriptors and dense optical flow. This method achieves state-of-the-art results on several human activity recognition benchmarks.

Several other holistic approaches have been proposed. (Rodriguez *et al.*, 2008) introduce a template-based method for recognizing human actions in video. Their method is based on a Maximum Average Correlation Height (MACH) filter that aims to capture intra-class variability by synthesizing a single spatio-temporal MACH filter for a given action class. In order to make the method efficient, they analyze the response of the filter in the frequency domain, thus avoiding the high computational cost commonly incurred in template-based approaches. (Brendel and Todorovic, 2011) proposed a method that automatically learns spatio-temporal relations and activity parts that are most relevant for representing human activities. To that end, they proposed a novel representation of videos based on spatio-temporal graphs, where nodes correspond to multi-scale video segments and edges capture their hierarchical, temporal, and spatial relationships. Inference and learning are performed within the same framework based on robust least-squares optimization. At test time the model is applied to detect and localize relevant parts of activities in videos. (Pirsiavash and Ramanan, 2014) proposed to use simple grammars that capture hierarchical temporal structure of human activities while allowing for efficient inference. To that end, they develop grammar based approaches that decompose videos into action segments, and recursively decompose actions to sub-actions. They proposed Specialized grammars that can be efficiently parsed in an online fashion and learned from partially labeled data. Training videos assumed to be provided with action labels, and the method is able to infer latent sub-action structure during the max-margin learning.

Pose based representations for activity recognition. Another line of research explores ways of higher level video encoding in terms of body pose and motion (Singh and Nevatia, 2011; Rohrbach *et al.*, 2012; Jhuang *et al.*, 2013). The intuition there is that many activities exhibit characteristic body motions and thus can be reliably described using human body pose based features. (Singh and Nevatia, 2011) adopt a joint tracking and recognition approach to track the actor's pose by sampling from 3D action models. The action models are obtained by lifting the manually annotated 2D key poses to 3D and then computing the transformation matrices between the 3D key pose figures. In order to better fit the poses sampled from coarse action models to the observations, pose-specific part models capturing appropriate kinematic and occlusion constraints in a tree-structure are used. Their approach has been shown to work particularly well in images with little clutter and fully visible people.

(Rohrbach *et al.*, 2012) proposed a human pose tracking based method for activity recognition in more challenging kitchen scenarios with frequent occlusions, truncations and complex poses. In their work, first, they estimate upper body poses in every tenth frame by adapting the method of (Andriluka *et al.*, 2009) and track the locations of body joints over a fixed temporal neighborhood forward and backward. Tracking is performed based on SIFT features that are extracted and matched for each joint separately across consecutive frames. Then, given the body joint trajectories, two different feature representations are computed: manually defined statistics over the body model trajectories and Fourier transform features (FFT) (Zinnen *et al.*, 2009). Experimental evaluation in the challenging kitchen scenarios shows that proposed pose based representations are less effective than holistic appearance based representations.

(Jhuang *et al.*, 2013) proposed an activity recognition method that also relies on automatically estimated body joint locations. They apply the pose estimation model of (Yang and Ramanan, 2013) to each individual frame and compute a much richer set of within and across frame geometric features, compared to (Rohrbach *et al.*, 2012). They show that body features significantly outperform holistic methods. However, these conclusions are made on a subset of HMDB (Kuehne *et al.*, 2011) where actions with global body motion are performed by isolated and fully visible individuals – a setting that seems to be well suited for pose estimation methods.

(Raptis and Sigal, 2013) chose a different strategy for pose based activity recognition. Instead of detecting individual body joints, they use spatially-localizable poselet (Bourdev and Malik, 2009) representations to encode key frames. This allows to model the whole video sequence as a sparse set of temporally discriminative keyframes. A max-margin learning framework is used to learn a set of most discriminative keyframes that are treated as latent variables, and the local temporal connections between them. They show competitive performance when detecting different interactions between people.

Deep learning representations for activity recognition. Both holistic and pose based activity recognition methods rely on manually designed representations based on either hand crafted low level appearance and motion feature descriptors, or higher level but still pre-defined geometric body pose and motion features. Another line of research follows the recent trend and develops deep learning methods based on convolutional neural networks (CNNs) that learn multiple representation layers directly from the video pixels. In contrast to CNNs applied to images, the key challenge of video based deep learning is in extending the connectivity of CNNs to the time domain to take advantage of local spatio-temporal information while still allowing for efficient training.

One of the first to extend existing single frame based CNNs to the third, temporal, dimension for recognition of human activities in videos were (Karpathy *et al.*, 2014). To that end, they suggest a multi-resolution spatial foveated architecture that allows to significantly speed up training. The low resolution context stream receives the down-sampled frames at half the original spatial resolution, while the fovea stream receives the center of the down-sampled region at the original resolution. This allows to halve the total input dimensionality while also taking advantage of the camera bias present in many online videos. They explore different strategies to fuse temporal information and found slow fusion, where higher layers get access to progressively more global information in both spatial and temporal dimensions, to work best. However, experimental evaluation shows only a modest improvement of the best spatio-temporal CNN over single frame models.

(Simonyan and Zisserman, 2014a) also proposed a two stream architecture that is significantly different from (Karpathy *et al.*, 2014). The architecture is based on two separate recognition streams, spatial and temporal, implemented as CNNs and combined by late fusion. The spatial stream is used for action recognition from still video frames, while the temporal stream is trained to recognize actions from motion described by dense optical flow. Decoupling the spatial and temporal CNNs allows for separate pre-training of both nets. Thus, the spatial CNN can exploit the large amounts of available annotated image data, and the temporal CNN can be pre-trained from pre-computed multi-frame dense optical flow. Experimental results show that using optical flow for training of the temporal CNN allows to achieve very good performance in spite of limited training data.

Another group of action recognition methods builds on the state-of-the-art region based generic object detector (RCNN) (Girshick *et al.*, 2014). (Gkioxari and Malik, 2015) proposed a deep learning based action detection approach that combines appearance and motion cues. First, in each frame, motion salient proposal regions are selected from the entire set of region proposals, which leads to a significant reduction in the number of regions. Second, spatio-temporal feature representations are extracted to build strong classifiers using CNNs. In the follow-up work (Gkioxari *et al.*, 2015b) extend their previous approach by incorporating contextual cues. To that end, RCNN is adapted to use more than one region for classification while still maintaining the ability to localize the action. Joint training of action specific models and image representations in a common deep learning framework leads to rich action specific representations. Experimental evaluation demonstrates significant improvements over the state of the art. Recently, (Gkioxari *et al.*, 2015a) proposed a part-based approach to action recognition, where detectors are a deep version of poselets (Bourdev and Malik, 2009) that capture parts of the human body under characteristic poses. They show that adding parts to holistic CNNs achieves the best performance both for action and attribute recognition. However, they also observe that using parts in deeper networks leads to less significant improvements.

2.3.2.2 *Datasets*

We now discuss datasets created for benchmarking human activity recognition in video. Even when not considering benchmarks for single image based action recognition, such as (Everingham *et al.*, 2011b), the number of existing activity recognition datasets is still large (e.g. over 30 datasets are listed in (Ahad *et al.*, 2011)). We thus focus on the most relevant ones w.r.t. the dataset used for benchmarking human activity recognition methods in this thesis. We distinguish four large groups of datasets: full body pose, movie and YouTube video, surveillance, and assisted daily living datasets. Different evaluation modes have been proposed, with the most popular ones being activity classification and detection. The former requires classification of a *pre-localized* activity into one of the classes. The latter corresponds to a more challenging setting, where an activity is not pre-localized temporally and thus first has to be detected. In this case, a single video may contain multiple instances of the same activity class, or even multiple activities belonging to different classes. In the following, we discuss the most related activity recognition benchmarks and their evaluation settings.

KTH Action (Schuldt *et al.*, 2004). The KTH Action dataset contains 2,391 sequences of six full body human actions, namely "walking", "jogging", "running", "boxing", "hand waving", and "hand clapping", performed by 25 people in four different scenarios (outdoors, outdoors with scale variation, outdoors with different clothes, and indoors). All sequences were taken in front of homogeneous backgrounds with a static camera. The task is to classify each action into one of six classes.

MSR Action (Yuan *et al.*, 2009). The MSR Action is another full body pose action dataset. It consists of 16 video sequences containing in total 63 actions belonging to three action categories, namely "hand clapping", "hand waving", and "boxing". The actions are performed by 10 subjects. Each sequence contains multiple types of actions and some sequences contain actions performed by different people. The actions are performed indoors and outdoors. In contrast to the KTH Action dataset, the video sequences of the MSR Action benchmark have more realistic cluttered and moving backgrounds. Furthermore, the task is more challenging, as it requires multi-class action detection in longer sequences. Later, a second version of the MSR Action dataset was released[1]. It contains the same number of action classes, but a much larger number of video sequences (54) and action instances (203).

Hollywood Human Actions (Laptev *et al.*, 2008). This dataset was collected by extracting 663 short video clips from 32 Hollywood movies (12 movies were used to extract training video clips, 20 movies to extract testing video clips). Each video clip contains one of eight actions: "answer phone", "get out of the car", "hand shake", "hug person", "kiss", "sit down", "sit up", and "stand up". Annotations for half of the training set were obtained by automatically aligning the movie scripts and subtitles to videos. This dataset has been further extended to include four additional action classes, namely "drive car", "eat", "fight person", and "run", and a larger number of video clips (1,709) extracted from 69 movies (Marszałek *et al.*, 2009). The task on both datasets is action classification.

Coffee and Cigarettes (Laptev and Perez, 2007). The Coffee and Cigarettes dataset contains 264 short video clips of two action classes, namely "drinking" and "smoking". The video clips were extracted from several Hollywood movies and contain the actions appearing in different scenes performed by different people. In contrast to the Hollywood Human Actions datasets, this benchmark promotes action detection rather than action classification.

High Five (Patron *et al.*, 2010). High Five is another action detection benchmark. It was compiled of 300 video clips extracted from 23 different TV shows. Each video

---

[1] http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/

clip from the positive set of 200 videos contains one of four people interactions, namely "hand shake", "high five", "hug", and "kiss", and each interaction appears in 50 videos. The remaining 100 videos constitute the negative set that does not contain any interactions. The interactions are not temporally aligned and have a high degree of intra-class variability due to variations in the number of actors, their scales and viewpoints. Evaluation is performed by two fold cross validation.

HMDB51 (Kuehne *et al.*, 2011). HMDB51 is one of the most challenging action recognition benchmarks to date. The dataset was collected from a variety of sources ranging from digitized movies to YouTube and Google videos. It consists of 51 action categories distributed among 6,766 video clips, and each category has at least 101 associated clips. Action categories are grouped into five types: general facial actions, facial actions with object manipulation, general body movements, body movements with object interaction, and body movements for human interaction. For each action category and split, 70 training and 30 testing videos were selected, such that the training and testing clips are not originating from the same video. Performance is reported as average action classification accuracy over three dataset splits of increased difficulty.

Joint-annotated HMDB (J-HMDB) (Jhuang *et al.*, 2013) J-HMDB dataset was collected by extracting 928 clips with a total of 31,838 frames comprising 21 action categories of the challenging HMDB51 dataset. The selected action categories correspond to distinctive global body motions, such as "jump", "kick ball", "golf", "run", "sit", etc. Each frame was annotated on AMT using a tool based on the 2D puppet model (Zuffi *et al.*, 2012). The annotation requires adjusting the joint positions of the puppet so that its contours align with the contours of the shown person. This dataset was specifically created to analyze the performance of holistic and pose based methods for activity recognition. A subset of this dataset with 12 actions distributed over 316 clips is used as a pose estimation benchmark. Similar to other full body pose estimation datasets, this subset contains isolated individuals with all body parts present in the image.

UCF50 and UCF101[2]. The UCF50 dataset contains 50 action categories distributed over 6,618 video clips collected from YouTube. Most of the action categories are sport related, such as, e.g., "baseball pitch", "biking", "pull ups", or "rowing", whereas the rest of the categories correspond to playing an instrument (e.g., "playing violin", "playing piano"), dancing (e.g., "swing", "salsa spins") and walking (e.g., "walking with a dog", "military parade"). For all 50 categories, the videos are grouped into 25 groups and each group consists of more than 4 action clips. The video clips in the same group may share some common features, such as the same person, similar background, similar viewpoint, etc. Evaluation is done using leave-one-group-out cross validation and performance is reported as mean classification accuracy over all action categories. UCF50 has been further extended to UCF101 to double the

---

[2] http://crcv.ucf.edu/data/index.php

number of action categories (101) and the number of available video clips (13, 320). However, the focus of the extended dataset remains on sports-centric activities.

Sports-1M (Karpathy *et al.*, 2014). The Sports-1M dataset is the largest activity classification dataset to date. It consists of $1M$ YouTube videos annotated with 487 classes. The classes are arranged into a manually defined hierarchy that contains internal nodes, such as "aquatic sports", "team sports", and "sports with animals". The hierarchy is fine grained at the lowest level and contains, e.g., 6 different types of bowling and 23 types of billiards, among other activity categories. Each class has 1, 000-3, 000 videos assigned automatically by analyzing the text meta data surrounding the videos. 70% of all videos is used for training, 10% for validation and 20% for testing. Performance is reported as mean classification accuracy using top one and top five predictions per video. The large-scale training set allows for training expressive video classification models based on convolutional neural networks (CNNs).

PETS 2007[3]**.** In contrast to the previously discussed datasets addressing action recognition in movies and Internet videos, the PETS workshop dataset is targeted towards surveillance. The dataset contains real world situations captured by surveillance cameras in shops, airports, or subway stations. Its videos typically contain multiple people with high degrees of occlusion. Video sequences are grouped into three scenarios with increasing scene complexity: loitering, attended luggage removal (theft), and unattended luggage. The goal is to detect one or more of these security/criminal events within a real-world environment.

VIRAT (Oh *et al.*, 2011). VIRAT is another surveillance benchmark with focus on continuous visual event recognition in outdoor areas with wide coverage. It is a large-scale dataset including diverse events that involve interactions between multiple actors, vehicles and facilities. In total it contains 23 events, such as "walking", "standing", "opening/closing the trunk", "loading/unloading", "entering/exiting facility", etc., distributed over 29 hours of video. The dataset is provided with detailed annotations including both moving object tracks and event labels. Additionally, the dataset defines novel evaluation metrics and different types of evaluation modes for visual recognition tasks.

University of Rochester Activities of Daily Living (URADL) (Messing *et al.*, 2009). This dataset falls into the fourth category of the benchmarks discussed in this section, namely assisted daily living datasets. The dataset consists of 150 high-resolution videos clips of 10 activities useful for an assisted cognition task. Some sample activities are "answering the phone", "drinking a glass of water", or "peeling a banana". Each activity was performed three times by five different people. Performance is reported as mean activity classification accuracy.

---

[3] http://www.cvg.reading.ac.uk/PETS2007/data.html

CMU-Multimodal Activity Database (CMU-MMAC) (De la Torre *et al.*, 2008). The CMU-MMAC is another assisted daily living dataset. It contains several hours of detailed multi-modal sensor data capturing people while cooking several dishes in the kitchen scenarios. The sensing modalities include video from several external and one wearable camera, multi-channel audio, inertial measurement units and marker based body motion capture performed with multi-view cameras. Although this dataset provides rich spatial and temporal data, multiple sensors and markers attached to the body make the videos look unrealistic.

MPII Cooking Activities (Rohrbach *et al.*, 2012). The MPII Cooking Activities dataset is also concerned with capturing human activities in kitchen scenarios. However, in contrast to CMU-MMAC, the dataset contains much more realistic videos of 12 participants performing 65 different cooking activities, such as "cut slices", "pour", or "spice". In order to record realistic behavior and not isolated individual activities, participants were asked to prepare one to six of a total of 14 dishes, such as, e.g., *fruit salad*, or *cake*. Preparation of each dish contains several atomic cooking activities. The dataset consists of 44 videos with a total length of over 8 hours, and over $800K$ frames. The dataset supports both activity detection and activity classification evaluation modes.

2.3.2.3 *Relations to our work*

We now relate the other works to the contributions of this thesis concerned with establishing a new activity recognition benchmark (Chapter 9) and using the benchmark for thorough performance analysis of popular human activity recognition methods (Chapter 10).

Our benchmark in Chapter 9 is provided with activity labels and thus can be used for training and evaluation of current activity recognition methods, similar to the activity datasets presented above. However, related benchmarks typically focus on a single aspect of human activity recognition challenges, such as full body pose actions (Schuldt *et al.*, 2004; Yuan *et al.*, 2009; Jhuang *et al.*, 2013), sports (Karpathy *et al.*, 2014, UCF data), actions in movies (Laptev *et al.*, 2008; Laptev and Perez, 2007; Patron *et al.*, 2010; Kuehne *et al.*, 2011), event recognition (Oh *et al.*, 2011, PETS), or daily assisted living activities (Messing *et al.*, 2009; De la Torre *et al.*, 2008; Rohrbach *et al.*, 2012). In contrast, our dataset is not restricted to single atomic actions, but aims for much broader scale and coverage of activity classes. To that end, the dataset was collected from YouTube videos using an established taxonomy of over 800 every day human activities. This allows to represent both common and rare human activities that might be missed when relying on the ad-hoc selection of activity classes. In contrast to other benchmarks containing a handful to several dozens of action classes, our dataset includes several hundreds of human activities thus being more representative. Compared to the Sports-1M dataset (Karpathy *et al.*, 2014) that also includes multiple hundreds of activities, the focus of our benchmark is on much broader set of every day human activities, with sports related activities representing only a portion of the entire set.

Compared to the related activity recognition benchmarks containing from few hundreds up to few thousands of video clips, our benchmark consists of over $40,000$ video clips with over $1.5M$ frames, and thus is one to two orders of magnitude larger. The only dataset containing a larger number of training videos is the sports related benchmark of (Karpathy *et al.*, 2014). However, in contrast to their weakly annotated dataset containing much label noise and unrelated activities, our benchmark was subject to a manual quality control at different stages of data collection and annotation. Additionally, by providing the URLs to YouTube videos and time intervals when the videos were extracted, we provide a possibility to extend the number of frames for each video clip.

Compared to other activity recognition datasets restricted to a few actions performed at front of simple background (Schuldt *et al.*, 2004; Yuan *et al.*, 2009) and in Hollywood movies (Laptev *et al.*, 2008; Laptev and Perez, 2007; Patron *et al.*, 2010), or focusing on activities in improvised kitchen scenarios (De la Torre *et al.*, 2008; Rohrbach *et al.*, 2012), our dataset is more realistic. Using YouTube allows us to access a rich collection of videos originating from various sources, including amateur and professional recordings and capturing a variety of public events and performances.

Similar to other activity recognition benchmarks, our dataset also comes with activity labels. However, in contrast to many other benchmarks we also provide a much richer set of labels including location and visibility annotations of individual body joints, visibility annotations for body parts, full 3D torso and head orientation. The J-HMDB dataset (Jhuang *et al.*, 2013) is closest to ours in terms of richness of labels. However, compared to 21 activity classes considered in (Jhuang *et al.*, 2013) our dataset includes 410 activities and more than an order of magnitude more images ($\sim$ 32K in J-HMDB vs. over $1.5M$ images in our dataset).

Rich annotations and a large representative dataset allow for large-scale comparison of the popular holistic method based on dense trajectories (Wang *et al.*, 2013; Wang and Schmid, 2013) and pose based activity recognition methods (Chapter 10). Our results complement the findings in (Jhuang *et al.*, 2013), indicating that pose based features indeed outperform holistic features for certain cases. However, we also find that both types of features are complementary and their combination performs best, which is in contrast to (Jhuang *et al.*, 2013) who concluded that pose based methods outperform holistic methods and did not show any improvement by combining both types of representations. In contrast to other works typically providing the final activity recognition performance of multiple methods without analyzing where the performance differences come from (Wang *et al.*, 2013; Jhuang *et al.*, 2013; Karpathy *et al.*, 2014, e.g.), we aim for an in depth performance analysis and reveal numerous cases where the methods are complementary. To that end, we use rich annotations and analyze the factors responsible for the success and failure of holistic and pose based methods. For instance, we found that holistic methods are mostly affected by the number and speed of trajectories, whereas pose-based methods are mostly influenced by viewpoint of the person. We observe striking performance differences across activities. For some activities pose based features are more than twice as accurate compared to holistic features, and vice versa. The best performing approach in our comparison is based on the combination of holistic and pose based approaches, which again underlines their complementarity.

LEARNING PEOPLE DETECTION MODELS FROM FEW TRAINING SAMPLES

3

Contents

3.1 Introduction **59**
3.2 People detection models **61**
3.3 MovieReshape: 3D human shape model **63**
3.4 Experimental evaluation **64**
  3.4.1 Results using the Reshape data 66
  3.4.2 Combining different datasets for training 68
  3.4.3 Sliding-window detection using HOG 70
3.5 Conclusion **71**

In this chapter we introduce the approach to generate a large number of photo-realistically looking synthetic training samples from only a few persons and views and demonstrate that this automatically generated synthetic training data can be successfully used to learn better people detection models. State-of-the-art methods learn appearance based models requiring tedious collection and annotation of large data corpora. Also, obtaining data sets representing all relevant variations with sufficient accuracy for the intended application domain at hand is often a non-trivial task. Therefore this chapter investigates how 3D shape models from computer graphics can be leveraged to ease training data generation. In particular we employ a rendering-based reshaping method in order to generate thousands of synthetic training samples from only a few persons and views. We evaluate our data generation method for two different people detection models. Our experiments on a challenging multi-view dataset indicate that the data from as few as eleven persons suffices to achieve good performance. When we additionally combine our synthetic training samples with real data we improve the performance even further.

3.1 INTRODUCTION

People detection has been actively researched over the years due to its importance for applications such as mobile robotics, image indexing and surveillance. Prominent methods for people detection rely on appearance-based features paired with supervised learning techniques. This is true for full-body models such as (Dalal and Triggs, 2005) as well as part-based models such as (Andriluka *et al.*, 2009; Felzenszwalb *et al.*, 2010; Tompson *et al.*, 2014). Key to very good performance for these methods is to collect representative and substantial amounts of training data which is a tedious

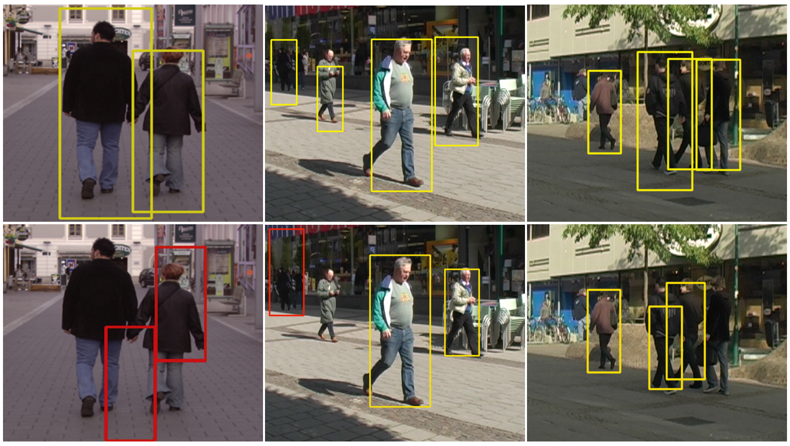


Figure 3.1: Sample detections at the equal error rate by the model trained on synthetic data generated from 6 people (top row) and on the game engine data of (Marin *et al.*, 2010) (bottom row). Even training on a subset of data obtained from only 6 different people, we are able to outperform the detector trained on much more variable game engine data. (see Sec. 3.4.1 for more details)

and time-consuming task and often limits further improvements.

The question we are asking in this chapter is if the realism of computer graphic models such as (Anguelov *et al.*, 2005; Balan *et al.*, 2007; Jain *et al.*, 2010) can help computer vision to reduce the tedious task of data collection and at the same time improve the quality and the relevant variability of the training data. Even in the early days of computer vision, computer graphics has been seen as a rich source for object models (Brooks *et al.*, 1979; Lowe, 1987; Marr and Nishihara, 1978). While these early models lacked realism in appearance more recent rendering techniques have indeed allowed to learn models for objects such as cars using computer graphics models alone (Liebelt *et al.*, 2008; Stark *et al.*, 2010). Also in the context of people detection computer graphics models have been used to generate training data. (Marin *et al.*, 2010), e.g., reports promising results using a game engine to produce training data. While game engines have improved dramatically over the years they are still not as realistic as more elaborate 3D human models such as (Anguelov *et al.*, 2005; Balan *et al.*, 2007; Jain *et al.*, 2010).

The first major contribution is to explore the applicability of a state-of-the-art 3D person model from the computer graphics community to learn powerful people detection models. We directly compare to people detection systems based on the well-known pictorial structures model (Andriluka *et al.*, 2009) as well as the Histogram of oriented gradients (HOG) model (Dalal and Triggs, 2005) learned from hundreds of manually labeled training data. Our findings indicate that surprisingly good results can be obtained training from as few as 1 or 2 people only and that comparable results can be obtained already with 11 people. The second main contribution is to compare these results to prior work such as (Marin *et al.*, 2010). The third contribution is to analyze different combinations of real and synthetic training data thereby outperforming methods using standard training data only. These results are obtained for two famous people detection methods, namely the pictorial structures model and the HOG model.

3.2 PEOPLE DETECTION MODELS

In this section we briefly recapitulate the two prominent people detection models used as the basis for our study. We will start with the pictorial structures model (Fischler and Elschlager, 1973) which has been made popular by (Andriluka *et al.*, 2009; Felzenszwalb and Huttenlocher, 2005) and then briefly introduce the sliding-window detection model with HOG features (Dalal and Triggs, 2005).

Pictorial structures model. In this model the human body is represented by a flexible configuration $L = \{l_0, l_1, ..., l_N\}$ of N body parts. The state of part i is given by $l_i = (x_i, y_i, \theta_i, s_i)$, where (x_i, y_i) denotes the part position in image coordinates, θ_i the absolute part orientation, and s_i denotes the part scale relative to the part size in the scale normalized training set. Given image evidence E, the posterior of the part configuration L is given by

$$p(L|E) \propto p(E|L)p(L) \tag{3.1}$$

where $p(L)$ is the kinematic tree prior and $p(E|L)$ corresponds to the likelihood of image evidence E under the particular body part configuration L. The tree prior expresses the dependencies between parts and can be factorized as

$$p(L) = p(l_0) \prod_{(i,j) \in G} p(l_i|l_j) \tag{3.2}$$

where G is the set of all directed edges in the kinematic tree, l_0 is assigned to the root node (torso) and $p(l_i|l_j)$ are pairwise terms along the kinematic chains. $p(l_0)$ is assumed to be uniform, and pairwise terms are modeled to be Gaussians in the transformed space of part joints (Andriluka *et al.*, 2009; Felzenszwalb and Huttenlocher, 2005).

The likelihood term is decomposed into the product of individual part likelihoods:

$$p(E|L) = p(l_0) \prod_{i=0}^{N} p(e_i(l_i)) \tag{3.3}$$

where $e_i(l_i)$ is the evidence for part i at image location l_i.

As we use the publicly available implementation provided by (Andriluka *et al.*, 2009), part likelihoods are computed by boosted part detectors, which use the output of an AdaBoost classifier (Freund and Schapire, 1997) computed from dense shape context descriptor (Belongie *et al.*, 2002). Inference is performed by means of sum-product belief propagation to compute marginal posteriors of individual body parts. For pedestrian detection, the marginal distribution of the torso location is used to predict the bounding box, similar to the work of (Andriluka *et al.*, 2009).

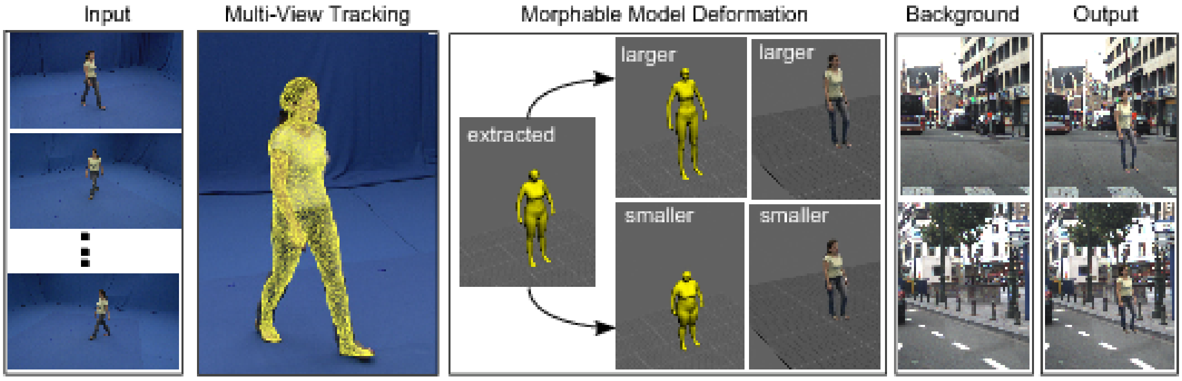


Figure 3.2: Overview of the approach to generate training data from real examples using a morphable 3D body model that drives a 2D image deformation.

We slightly adapt the pictorial structures model of (Andriluka *et al.*, 2009) to use 6 body parts which are relevant for pedestrian detection: left/right lower and upper legs, torso and head. Also, we use a star prior on the part configuration, as it was shown to perform on par with a tree prior (Andriluka *et al.*, 2009) while making the inference much simpler.

In the experiments reported below the part likelihoods as well as the star prior are learned on different training sets ranging from real images as used by (Andriluka *et al.*, 2010), over game-engine produced data as used by (Marin *et al.*, 2010) to images produced from a state-of-the-art 3D human shape model introduced in the section 3.3.

Sliding-window detection with HOG features. In the sliding-window detection framework the image is scanned over all positions and scales and each window is represented by a feature and classified independently to contain a pedestrian or not. Contrary to the pictorial structures model pedestrians are often represented by a monolithic template without the notion of body parts. In this work we employ HOG features (Dalal and Triggs, 2005). This feature has been shown to yield good performance for pedestrian detection in several benchmarks (Dollár *et al.*, 2009). For a 128×64 detection window HOG features vote the gradient orientation into 8×8 pixel large cell histograms weighted by the gradient's magnitude. To tolerate slight variations in position and scale the responses are interpolated with respect to orientation and location and distributed into neighboring bins and cells. More robustness with respect to lighting conditions is achieved by normalization over 2×2 groups of cells. As a classifier we employ a histogram intersection kernel SVM which can be computed efficiently at test time (Maji *et al.*, 2008). To merge nearby detections on the same object we employ mean-shift mode search as a non-maximum suppression step.

3.3 MOVIERESHAPE: 3D HUMAN SHAPE MODEL

In order to generate synthetic training data for the people detection models, we adopt an approach to reshape humans in videos (Jain *et al.*, 2010). The core component of this work is a morphable 3D body model that represents pose and shape variations of human bodies. Starting from an image sequence of an individual this model allows to generate large amounts of synthetic training data representing 3D shape and pose variations of the recorded individual. Fig. 3.2 shows an overview of the approach. To generate the required input data, we ask subjects to perform movements in front of a uniformly colored background in our motion capture studio. Each person is captured with 8 HD cameras with a resolution of 1296×972 pixels. First, the subject is segmented from the background and the extracted silhouettes are used to automatically fit the morphable 3D body model to the input sequences. We then randomly sample from the space of possible 3D shape variations that is defined by the morphable body model. These shape parameters drive a 2D deformation of the image of the subject. In the last step, an arbitrary background is selected and is composited with the image of the deformed subject. To generate large amounts of training data for each subject, the random selection of the 3D shape parameters and the background is repeated several times resulting in an arbitrary number of composited training images with different body shapes for all subjects, all performed poses, and all camera views.

Morphable 3D body model. The morphable body model is generated from a database of 3D laser scans of humans (114 subjects in a subset of 35 poses). Additionally, body weight, gender, age, and several other biometric measures of the subjects are recorded (Hasler *et al.*, 2009). From this data a morphable 3D body model is built, similar to the well known SCAPE model (Anguelov *et al.*, 2005). This morphable model is capable of representing almost all 3D pose and shape variations available in the database. The pose variations are driven by a skeleton in combination with linear blend skinning that is defined once manually for the template mesh fitted to all 3D scans in the database. The shape variations across individuals are analyzed and represented via principal component analysis (PCA). The first 20 PCA components are used capturing 97% of the variations in the observed body shapes.

Markerless motion capture. Given the segmented input images, we employ a particle filter-based estimator (Jain *et al.*, 2010) to fit the parameters of the morphable body model to the extracted silhouettes. The estimated parameters are the 28 joint angles of the skeleton and the 20 PCA coefficients. The approach selects those particles whose parameters produce the lowest silhouette error in all camera views.

Image deformation. Once we know the parameters of the subject in the video this defines our deformation source. The corresponding deformation target is defined by randomly selecting different shape parameters from our database. Thereby, we allow samples from 3 times the standard deviations that was observed in the 3D

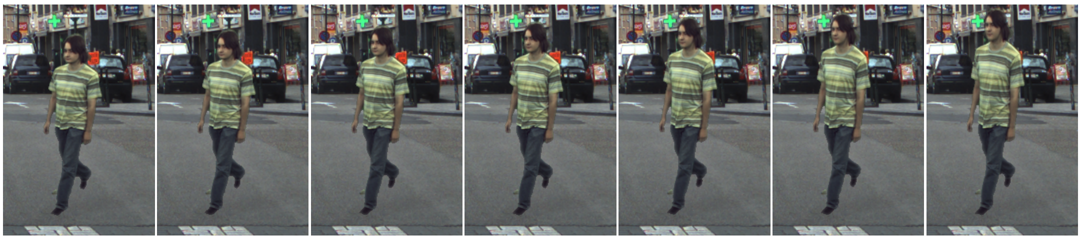


Figure 3.3: Sample Reshape images of a person with modified height. The leftmost and the rightmost images represent extreme deviations and the middle image corresponds to the original height; the 2nd and 6th images show deviations of 2σ, while the 3rd and 5th images correspond to the deviations of 1σ from the original height.

shape database of scanned subjects (corresponding to a 99% confidence interval). The difference between the 3D source and target model defines 3D offset vectors for all the vertices of the morphable model template mesh. As detailed in (Jain *et al.*, 2010), a subset of these 3D offset vectors can be used to drive a 2D deformation in the image plane. This 2D deformation is consequently motivated by the knowledge about the shape variations of subjects in the database and the results are different from simple image transformations (like non-uniform scaling or shearing). It is, e.g., possible that the depicted subject only becomes bigger at the belly, or gets shorter legs, or enjoys more muscular arms. The image deformation is repeated multiple times with randomly sampled body shapes.

Background compositing. In the final step, we sample randomly from a database of backgrounds containing images of urban scenes without pedestrians. We blend the segmentation masks of persons with a Gaussian with σ of 2 pixels. Then, we composite the background with the deformed images of subjects by adding weighted background and foreground pixel values together. See Fig. 3.3 for sample outputs of the system varying the height of the person.

3.4 EXPERIMENTAL EVALUATION

This section experimentally evaluates the applicability of training data obtained by the 3D human shape model described in section 3.3. These results are compared to training data obtained from real images (Andriluka *et al.*, 2010) as well as from a game engine (Marin *et al.*, 2010). First, we briefly introduce the different datasets used for training and evaluation. Then, we show that already a small number of people in our training dataset allows to achieve performance almost on par with the detector trained on real data containing hundreds of different people. We also show that combining detectors trained on real and synthetic data allows to outperform the detectors trained only on real data.

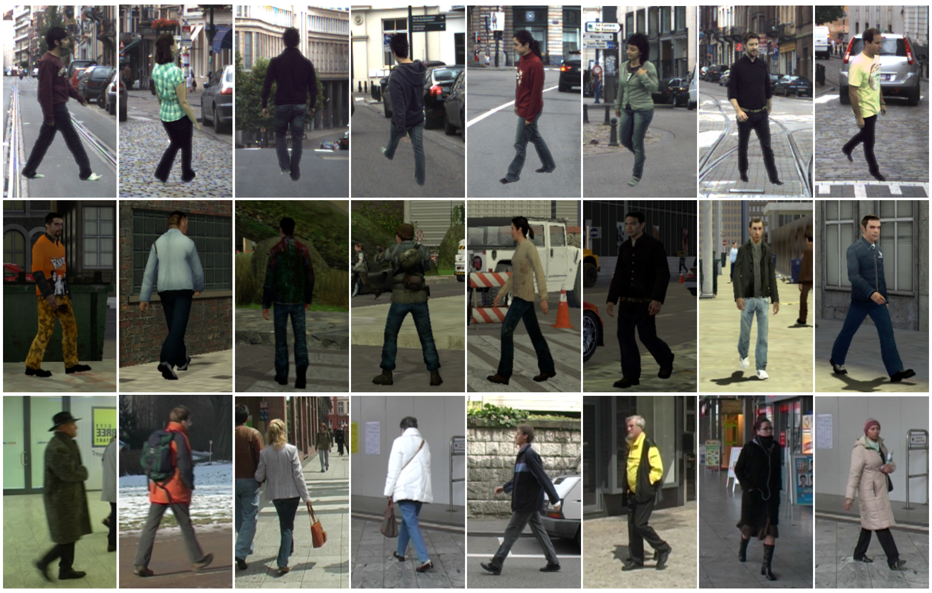


Figure 3.4: Samples from the training data used in our experiments: Reshape images (top row), CVC virtual pedestrians (middle row) and multi-viewpoint dataset (bottom row). Synthetic Reshape images look similar to the real ones while being much more realistic than CVC pedestrians. Real images often contain persons wearing long or wide clothes and caring a bag, which does not occur in the synthetic data.

Reshape training dataset. In order to obtain synthetic training data, we collected a dataset of 11 subjects each depicted in 6–9 different poses corresponding to a walking cycle. Each pose is seen from 8 different viewpoints separated by 45 degrees apart from each other. Synthetic images were obtained as described in section 3.3. For each original image we generated 30 gradual changes of height: 15 modifications making a person shorter and 15 making a person taller, which results in almost 2000 images per person and 20400 positive training samples in total (see for samples Fig. 3.3). We note that the applied transformation is non-linear and therefore different from simply scaling the original image. The MovieReshape model also allows to automatically obtain bounding box as well as body part annotations which are required for the pictorial structures model. The annotations for the unmodified image are obtained by backprojecting the morphable 3D model to the image plane. For the reshaped images we apply the same inverse mapping to these annotations which is used to morph appearance. This is one of the key advantages which facilitates the generation of large amounts of data without the need to manually annotate each image. All persons are rescaled to 200 pixel in height and embedded in background images of driving sequences containing no pedestrians. To record the background sequences a calibrated camera has been used and thus synthetically generated pedestrians can be easily embedded at geometrically plausible positions on the ground plane. Some sample images are shown in Fig. 3.4 (top row). We additionally perform smoothing along the shape boundaries separating persons from background in order to get more realistic gradients for the shape context descriptor. Finally, we adjust the luminance of the embedded pedestrians such that their mean approximately matches the backgrounds' mean luminance.

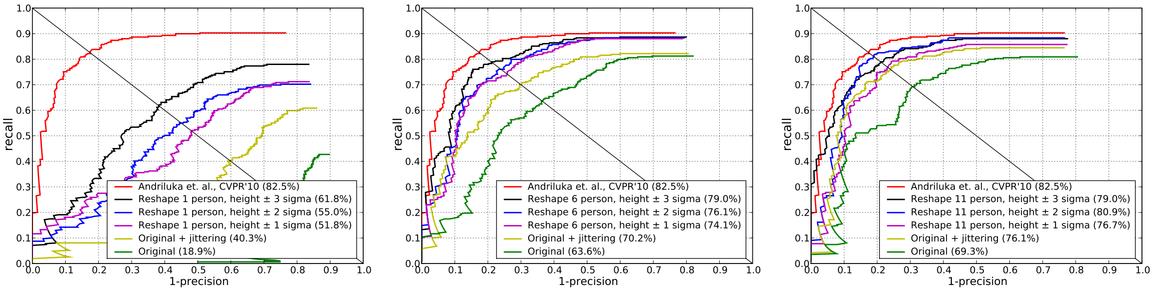
CVC training dataset. The second dataset contains synthetic images produced by a game engine which were kindly provided by the authors of (Marin *et al.*, 2010). These images of virtual pedestrians are generated by driving through virtual cities in the computer game Half-Life 2. The CVC dataset which we were provided with consists of 1716 pedestrians shown from arbitrary views with annotated bounding boxes. In comparison to our Reshape dataset, the appearance variability of the CVC dataset is significantly larger (cf. Fig. 3.4, middle row). We manually annotated the body parts of people and also rescaled the images so that all subjects have the same height of 200 pixels. Finally, we mirrored all images in order to obtain more training data, resulting in 3432 images in total. This data is complemented by a negative set of 2047 images of the same virtual urban scene environment without pedestrians.

Multi-viewpoint dataset. The third dataset we used in our experiments is the challenging multi-viewpoint dataset (Andriluka *et al.*, 2010) consisting of real images of hundreds of pedestrians shown from arbitrary views. The dataset comes with 1486 part-annotated pedestrians for training, 248 for testing and 248 for validation. The images from the training set were mirrored in order to increase the amount of training data. Sample images for different viewpoints can be seen in Fig. 3.4 (bottom row).

Experimental setup. To evaluate all trained models we use the multi-viewpoint dataset's test data and are thus directly comparable to the method of (Andriluka *et al.*, 2010) on this dataset. Thus, whenever we use the multi-viewpoint training data we refer to the experiment as *Andriluka*. For our experiments which use the Reshape dataset we use the multi-viewpoint dataset's negative training data. Experiments on the CVC data showed minor performance differences between using the negative data provided with the CVC data or the negative data provided by the multi-viewpoint data. In the following we thus only report results obtained with the CVC negative dataset. All results are provided as precision vs. recall curves and throughout this chapter we use the equal error rate (EER) to compare results. EERs for each experiment are also reported in the respective plots' legend. To match ground truth annotations to objects detections we use the PASCAL criterion (Everingham *et al.*, 2010), which demands at least 50% overlap of ground truth bounding box and detection.

3.4.1 Results using the Reshape data

We start by evaluating the pictorial structures model's performance when it is trained on the Reshape data and compare its performance to training on the multi-viewpoint training dataset and the CVC training dataset.



(a) 1 person (b) 6 people (c) 11 people

Figure 3.5: Results using Reshape data. Shown are results using 1 (a), 6 (b) and 11 (c) people to train a generic pictorial structures model. Each plot show results obtained by (Andriluka *et al.*, 2009) on real data (red), training on the unmodified training data (green), the reshape model with different variations of σ (violet, blue, black) and results using jittering (yellow)

Fig 3.5(a)-(c) show the results obtained using one, six, and eleven people to train a generic pictorial structures model. To understand the influence of different parameters of the model we vary the employed subset of the Reshape data. The green lines in figure Fig 3.5(a)-(c) show the results obtained using the original training sequences acquired from one, six and eleven people without applying the human reshape model of section 3.3. While the performance increases with more people the maximum performance obtained with eleven people is only 69.2% EER (equal error rate).

Although the wide range of height modifications allows to cover 99% of data variability spanned in this direction, having extremely short and tall pedestrians in the training set can be unnecessary, since they are quite rare in real world data. This consideration motivates to subsample the Reshape data w.r.t. maximal and minimal height of subjects. For that purpose we train pictorial structures model on subsets of images corresponding to no modification, $\pm 1, 2$ and 3σ (standard deviation) from the original mean height of people. The results for 1, 6 and 11 persons are again shown in Fig. 3.5. It can be observed that in all cases including the images with increasing number of height modifications helps to improve performance.

In order to understand whether the improvement comes from the increased variability of data rather than from the increased amount of positive samples, we also train the model on the set of original images enriched by jittering (Laptev, 2009). The results are shown in Fig. 3.5(a)-(c) in yellow. As expected, the performance of the model in the latter case is worse, as nonlinear data transformation due to height modifications allows to capture more realistic variability of the data than simple 2D jittering. The largest difference can be observed when using a single person only for training (cf. Fig. 3.5(a)). In this case, jittering helps to improve the performance from 18.7% to 40.1% EER, while training on the data with height modification $\pm 1\sigma$ results in an EER of 51.8%. When using six and eleven people the difference between using ± 2 or $\pm 3\sigma$ becomes less pronounced. For instance, for six people this difference

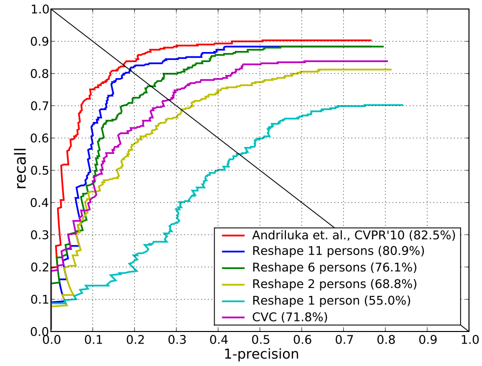


Figure 3.6: Detection rate w.r.t. the number of different persons represented in the training data. Already one person is enough to provide reasonable variability in the Reshape data. The increasing number of persons results in significant improvement which allows to achieve performance almost on par with the detector trained on the real data. The model trained on CVC data performs well, but noticeably worse than ours.

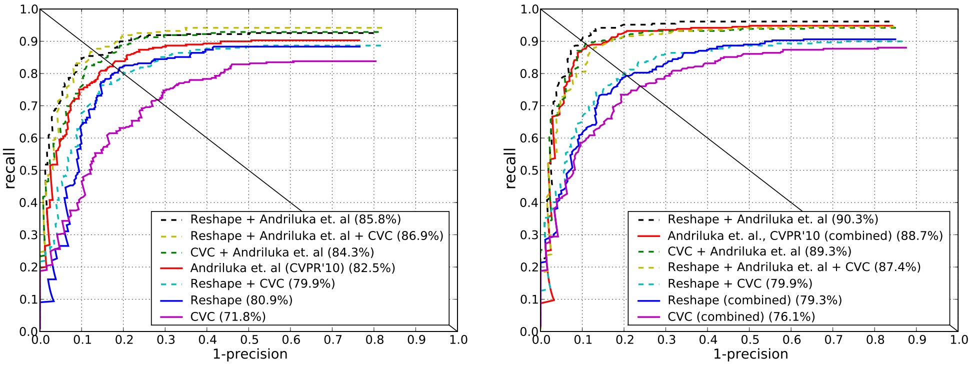
constitutes 3.9% on EER. As $\pm 2\sigma$ corresponds to faster training times due to less data we use this setting for the remainder of the paper.

Fig. 3.6 summarizes how the number of different persons contained in the Reshape dataset affects performance. Surprisingly, already training data from a single person obtains an EER of 55.0% suggesting that this data already covers a reasonable variability (this performance can be further improved using $\pm 3\sigma$ as shown in Fig 3.5(a)). Not surprisingly, increasing the number of people improves performance considerably. More interesting however is the fact that with as few as 11 people we are able to achieve performance of 80.9% EER, which is almost on par with the model trained on the real multi-viewpoint data (red curve, 82.5% EER) containing hundreds of different people.

Fig. 3.6 also contains a curve (in violet) for the model trained on the CVC dataset. As expected, the model trained on the virtual people and thus less realistic data performs worse achieving 71.8% EER, despite much larger number of different appearances contained in the dataset. For comparison, the model trained on a subset of our data from just six persons achieves 76.1% on EER. We also provide some sample detections obtained in this case which are shown in Fig. 3.1. These results clearly show the advantage of using our Reshape data for training.

3.4.2 Combining different datasets for training

In the previous section good results have been obtained using the Reshape data from as few as eleven different people as well as using training data from real images. Therefore this section explores the possibility to combine models trained on different types of data in order to boost performance further. In order to combine detectors, we follow a detector stacking strategy (also used in (Andriluka *et al.*,



(a) Combination of generic detectors

(b) Combination of viewpoint specific detectors

Figure 3.7: Combination of generic detectors (a) and viewpoint specific detectors (b). In both cases, the combination of our detector with the one trained on the real data helps to improve detection performance.

2010)). More precisely we train detectors on different datasets first and then combine them by an another SVM that is trained using the vectors of detector outputs as features (normalized by mean/variance). For SVM training, we use the validation set provided with the multi-viewpoint dataset.

We consider two different settings. First, we consider the combinations of the models trained on all viewpoints of the corresponding data, as it is done in the previous section. The results are shown in Fig. 3.7(a), where single detectors are denoted by solid lines, and combined ones are marked by dotted lines. The combination *Andriluka+CVC* (84.1% EER) improves performance slightly over *Andriluka* alone (82.5% EER) whereas the combination *Reshape+CVC* (79.9%) does not improve performance w.r.t. *Reshape* (80.9%). The combination *Reshape+Andriluka* (85.8%) does improve both over *Andriluka* alone as well as *Reshape* alone. Further adding CVC (*Reshape+Andriluka+CVC*) slightly improves the performance achieving 87% EER. Overall this combination obtains the best performance reported in the literature for this setting (multi-viewpoint pictorial structures model). The combination *Reshape+CVC* performs similarly to *Reshape* data alone. This might be due to the fact that in both types of data subjects wear tight clothes such as trousers, jackets and T-shirts, but no coats or dresses which sometimes occur in the test data. Additionally this combination suffers from less realistic appearance of the virtual pedestrians. Hence, the additional *CVC* samples are not complementary to the Reshape samples. This intuition is also confirmed by a noticeable improvement obtained by combining the detector trained on Reshape data with the one of (Andriluka *et al.*, 2010) trained on real multi-viewpoint data. As quite a few images in the real multi-viewpoint training set contain persons wearing long clothes the training data and thus the detectors are more complementary. For the same reason, the combination of the CVC detector and the Andriluka detector performs better than Andriluka's detector, though the combination's performance is slightly worse than the combination with the Reshape data.

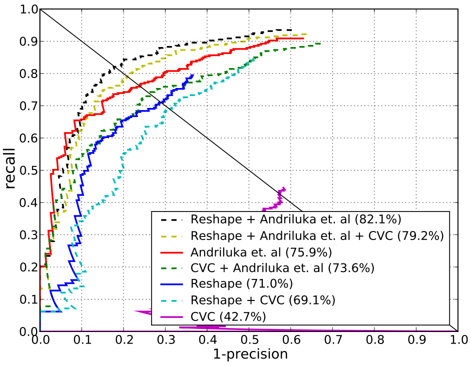


Figure 3.8: Performance of the sliding window detector of on different types of data. Similar to pictorial structures model, the best performance achieved when trained on human reshape together with real data. Detector trained on CVC pedestrians again performs worst.

The second setting explored in this section is to combine not only one detector trained on each dataset but to first train viewpoint-specific detectors on appropriate subsets of the different data and then train a stacked classifier on combinations thereof. The main advantage is that the part detectors as well as the kinematic tree prior are more specific for each view and thus more discriminative. The results are shown in Fig. 3.7(b). First, the combination of 8 viewpoint-specific detectors trained on Reshape data clearly outperforms those trained on CVC virtual pedestrians (79.3% against 76.1% EER) which again shows the advantage of training on our synthetic Reshape data. However, the performance achieved is still below the results provided by (Andriluka *et al.*, 2010) (red curve) who combined 8 viewpoint-specific detectors, 2 side-view detectors that contain feet and one generic detector trained on all views. By enriching this set of detectors by 8 viewpoint-specific and one generic detector trained on human Reshape data, we are able to outperform the results of (Andriluka *et al.*, 2010) increasing the detection rate from 88.6% to 90.3% EER. The 8 CVC viewpoint-specific detectors are not complementary enough to further boost performance w.r.t. the combinations mentioned above.

3.4.3 Sliding-window detection using HOG

For the combination of different datasets we additionally verified our findings for a sliding-window detector framework (see Fig. 3.8). For this experiment we trained a generic detector for all viewpoints consisting of a monolithic HOG feature representation (Dalal and Triggs, 2005) combined with a fast histogram intersection kernel as classifier (Maji *et al.*, 2008). We used the exact same training data as for the experiments reported above. Overall the results obtained are slightly below the pictorial structure model's results in Fig. 3.7(a). This may be explained by the test set's difficulty, which contains people seen from all viewpoints and under all poses for which a part-based representation is favorable. As for the pictorial structures model the combination of the Reshape data with the multi-viewpoint data provided by Andriluka obtains best performance with an EER of 82.1%. When we additionally add the CVC virtual samples the performance drops to an EER of 79.2% which can be explained by the less realistic appearance of these samples. However, both combinations outperform the detector which is only trained on data by (Andriluka *et al.*, 2010) (EER 75.9%). Consistent with our finding for the pictorial structures model, the performance drops to an EER of 73.6% when the CVC data is added. Also the detector trained only on the Reshape data (EER 71.0%) performs worse than the detector trained on real data. Similarly to the real multi-viewpoint data, the combination of Reshape with CVC data decreases performance (EER 69.1%). The performance with the detector only trained on the CVC virtual samples is substantially worse. Interestingly Marin et al. (Marin *et al.*, 2010) have reported equal performance of virtual samples and their real data when a sliding-window size of 48×96 pixels is used. This might be explained by the fact that real data and virtual data appear more similar on the lower resolution DaimlerDB (Enzweiler and Gavrila, 2009) automotive test data (pedestrian median height is 47 pixels), while for higher resolution (median pedestrian height on the multi-viewpoint test data is 184 pixel) the classifier might loose performance due to unrealistic appearance. Overall we find that the results for the sliding-window detector framework to be consistent with the results obtained by the pictorial structures model leading to the same conclusions. We would also like to highlight that a detector trained on the combination of multi-viewpoint and Reshape data clearly outperforms a detector which is only trained on real multi-viewpoint data.

3.5 CONCLUSION

This chapter explored the possibility to generate synthetic training data from a state-of-the-art computer graphics 3D human body model (called Reshape data in the chapter). Learning people detection models from as few as 11 people enabled to achieve performance on par with competing systems trained on hundreds of manually labeled images. This result has been obtained for two of the well known people detection models, namely the pictorial structures model (in two different settings) as well as the HOG-detector. Using less realistic training data generated from a game engine (Marin *et al.*, 2010) has led to far less compelling results. Combining the detectors trained on the Reshape data with detectors trained on the manually labeled data has allowed to outperform the competitors for challenging multi-viewpoint data introduced by (Andriluka *et al.*, 2010).

Considering the fact that only 11 people have been recorded and used to train the respective appearance models the results reported in this chapter are indeed promising. In fact, using recordings from several hundreds of people should allow to reach performance levels that are beyond what can be reached with manually and tediously labeled data. In order to overcome the expensive data acquisition step limiting the total number of training appearances available for data generation, in the next chapter we explore the possibility of using the 3D body shape model directly to augment existing training data with complementary shape and pose variations.

4

IN GOOD SHAPE: ROBUST PEOPLE DETECTION BASED ON APPEARANCE AND SHAPE

Contents

4.1 Introduction **74**
4.2 Pictorial structures model **75**
4.3 Synthetic data **76**
  4.3.1 3D model of human shape 76
  4.3.2 Data generation 76
4.4 Experimental evaluation **79**
  4.4.1 Results using rendered data alone and combined with image data 80
  4.4.2 Combination of people detection models 81
4.5 Conclusion **83**

In this chapter we further develop the ideas for learning from automatically generated synthetic data. Rather than generating visually appealing photo-realisticly rendered synthetic training samples, which requires an expensive data acquisition step described in Chapter 3, this chapter explores the possibility to use a 3D human shape and pose model directly to add relevant shape information to learn more powerful people detection models. By sampling from the space of 3D shapes we are able to control data variability while covering the major shape variations of humans which are often difficult to capture when collecting real-world training images. We evaluate our data generation method for a people detection model based on pictorial structures. As we show on a challenging multi-viewpoint dataset, the additional information contained in the 3D shape model helps to outperform models trained on image data alone (see e.g. Fig. 4.1), and models trained on photo-realistic synthetic data obtained by our method in Chapter 3.

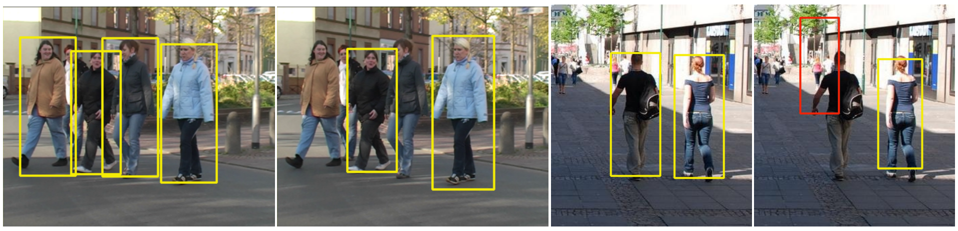

(a) (b) (c) (d)

Figure 4.1: Performance of our detector without (b,d) and with (a,c) complementary shape information

4.1 INTRODUCTION

People detection is a challenging task in computer vision with applications to surveillance, automotive safety and human-computer interaction. Prominent challenges for this task are the high intra-class variability, the high degree of articulation and the varying shape from different viewpoints. As state-of-the-art approaches heavily rely on supervised learning methods it is important to collect sufficient amounts of training data that capture the essential variability of the complex people class distribution. However, collecting large amounts of relevant training data is not only tedious but often an ill-defined process as it is unclear which part of the people class distributions is well-represented and which other parts of the distribution are still insufficiently sampled. A typical approach is to simply collect more training data without any guarantee that it will cover the people class distribution any better. Current datasets are often limited to few thousand samples taken from consumer photographs without any statement which parts of the real distribution of human shapes are sampled.

(Liebelt *et al.*, 2008; Marin *et al.*, 2010; Stark *et al.*, 2010) has proposed to use synthetic data to increase the available amount of training data. Similar to these works we use a graphics based 3D human model (Jain *et al.*, 2010) to enrich an existing training dataset with additional training samples. Contrary to existing works however, we do not aim to use visually appealing and photo-realisticly rendered data but instead focus on complementary and particularly important information for people detection, namely 3D human shape. The main intuition is that it is important to enrich image-based training data with the data that contains *complementary shape information* and that this *data is sampled from the underlying human 3D shape distribution*. Specifically, we sample the 3D human shape space to produce several thousand synthetic instances that cover a wide range of human poses and shapes. Next, we render non-photorealistic 2D edge images from these samples seen from a large range of viewpoints and compute the low-level edge-based feature representation (Belongie *et al.*, 2002) from these. It is important to point out, that in no stage of this pipeline photo-realistic images are produced. Instead, we show that by careful design of the rendering procedure our feature representation can generalize from synthetic training data to unseen real test data. Our experiments on people detection show that the combination of real and large amounts of synthetic data sampled from a previously learned 3D human shape distribution allows to train a detector which outperforms models trained from real data only (see e.g. Fig. 4.1).

4.2 PICTORIAL STRUCTURES MODEL

Here we briefly recapitulate the pictorial structures model (Fischler and Elschlager, 1973) made popular by (Felzenszwalb and Huttenlocher, 2005; Andriluka *et al.*, 2009).

Pictorial structures model. This model represents the human body as a flexible configuration $L = \{l_0, l_1, ..., l_N\}$ of N different parts. The state of each part i is provided by $l_i = (x_i, y_i, \theta_i, s_i)$, with (x_i, y_i) denoting its image position, θ_i the absolute part orientation, and s_i the part scale which is relative to the part size in the scale normalized training set.

Given image evidence E, the posterior probability of the part configuration L is provided by $p(L|E) \propto p(E|L)p(L)$, where $p(E|L)$ is the likelihood of image evidence E given a part configuration L and $p(L)$ is the kinematic tree prior describing dependencies between parts.

The prior can be factorized as $p(L) = p(l_0) \prod_{(i,j) \in G} p(l_i|l_j)$, with G denoting the set of edges representing kinematic dependencies between the parts, l_0 assigned to the root node (torso) and $p(l_i|l_j)$ are pairwise terms along the kinematic chains. $p(l_0)$ is assumed to be uniform, while pairwise terms are modeled to be Gaussians in the transformed space of joints between the adjacent parts (Felzenszwalb and Huttenlocher, 2005; Andriluka *et al.*, 2009).

The likelihood term can be factorized into the product of individual part likelihoods $p(E|L) = p(l_0) \prod_{i=0}^{N} p(e_i(l_i))$, where $e_i(l_i)$ represents the evidence for part i at location l_i in the image. In the publicly available implementation provided by (Andriluka *et al.*, 2009) part likelihoods are computed by boosted part detectors which use the output of AdaBoost classifiers (Freund and Schapire, 1997) computed from dense shape context descriptors (Belongie *et al.*, 2002). Marginal posteriors of individual body parts are found by sum-product belief propagation. Similar to the work of (Andriluka *et al.*, 2009, 2010), we use the marginal distribution of the torso location to predict the bounding box around pedestrian making a direct and fair comparison with this approaches feasible.

For our experiments we adjust the pictorial structures model of (Andriluka *et al.*, 2009) to use 6 body parts which are relevant for pedestrian detection, namely right/left upper and lower legs, head and torso. We use a star prior on the part configuration, as it was shown to perform similar to the tree prior while making the inference more efficient. Individual part likelihoods along with the star prior are learned from real images, as it was done by (Andriluka *et al.*, 2010), as well as from the rendered images produced by the 3D human shape model which we describe in Sec. 4.3.1.

4.3 SYNTHETIC DATA

In this section we introduce our novel approach to generate large amounts of synthetic data with a realistic shape distribution from a 3D human shape model. First we briefly describe the statistical model of human shape (Jain *et al.*, 2010). Then we present our data generation method and evaluate its parameters on a challenging multi-viewpoint dataset of real pedestrians (Andriluka *et al.*, 2010).

4.3.1 3D model of human shape

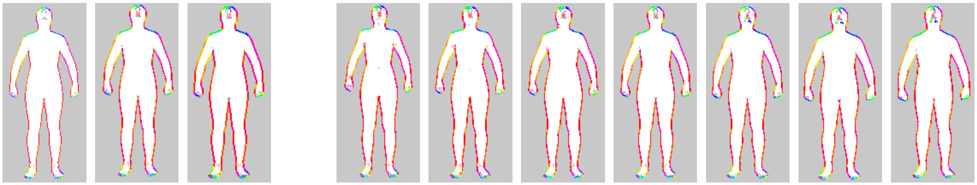
In order to generate non-photorealistic 2D edge images of pedestrians we employ a statistical model of 3D human shape (Jain *et al.*, 2010) which is a variant of SCAPE model (Anguelov *et al.*, 2005). The model is learned from a publicly available database of 3D laser scans of humans and describes plausible shape and pose variations of human bodies. A total of 550 full body scans (with roughly 50% male and 50% female subjects, aged 17 to 61) in varied poses were used to generate the model.

The shape variation across individuals is represented via principal component analysis (PCA). We use the first 20 PCA components capturing 97% of the body shape variation. Unlike the SCAPE model, no per triangle deformation is learned to represent pose-specific surface deformations as for the rendering of edge images such level of detail is not necessary. Instead, the linear blend skinning is used to model shape motion. To this end, a kinematic skeleton was rigged into the average human shape model by a professional animation artist.

The motion of the model is represented by a kinematic skeleton comprising 15 joints, with a total of 22 degrees of freedom (DoF) plus 6 DoF for the whole body. The surface of the model consists of a triangle mesh with 6450 3D vertices and 12894 faces. In total the model has 28 pose parameters and 20 parameters representing the body shape variations. Following the approach of (Jain *et al.*, 2010) we remap the 20 PCA shape parameters onto six meaningful semantic attributes, such as persons' height, weight, legs length, as well as waist, hips and breast girth, and directly change their values in our experiments.

4.3.2 Data generation

As argued before we propose to obtain synthetic training data by non-photorealistic rendering of the 3D human shape model described in Sec. 4.3.1. First, we uniformly sample shape parameters of six semantic attributes. Then we sample 3D joint angles from a set of 181 different walking poses. After shape and pose changes are applied, the 3D model is rendered from an arbitrary viewpoint into a 2D edge image by using an adapted rendering software kindly provided by the authors of (Stark *et al.*, 2010). Specifically, we render an edge created by two facets of the 3D model with respect to the normals of both facets. The closer the angle between camera view



(a) Varying edge thickness

(b) Varying waist

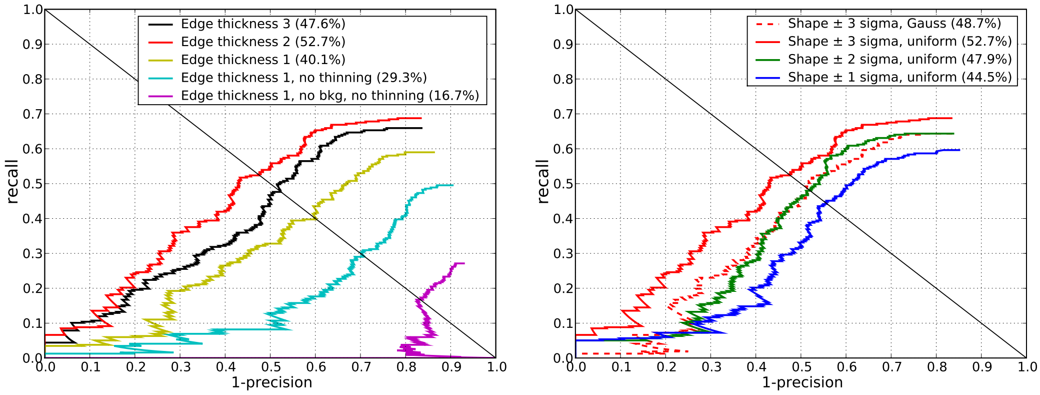
Figure 4.2: Examples of rendered images with (a) 1, 2 and 3 pixel thick edges and (b) deviations from the mean waist girth (middle). Shown are gradual shape changes in the range ±3 standard deviations from the mean shape.

and normals to the right angle, the stronger the edge. Samples of rendered images are shown in Fig. 4.4(a). Prior to training of individual part detectors we combine the rendered edges with the background edges obtained by applying Canny edge detector (Canny, 1986) to an image without pedestrians. Finally, we perform edge thinning in the composed edge image, exactly as it is done in the edge detector, and use the obtained results as a direct input to shape context descriptors. All part annotations needed for training of the pictorial structures model are produced automatically from known 3D joint positions of the human shape model.

The thickness of rendered edges, composition with the background and the magnitude of shape variability may impact the detection performance. Therefore, we experimentally study the choice of these parameters. For each set of parameters we generate 6,000 (6K) synthetic images and train the pictorial structures model which we evaluate on the whole test set of the multi-view dataset (Andriluka *et al.*, 2010) (see Sec. 4.4 for the dataset description and experimental setup).

Composition with background. First, we show the necessity of composing the rendered edge images with the background edges to achieve reasonable detection performance. The results are shown in Fig. 4.3(a). As it can be seen from the picture, combining rendered and background edges is absolutely essential as it helps to noticeably improve the results. Initial performance of the detector trained on the rendered edge images alone (purple curve, 16.6% EER – equal error rate) is significantly improved in case of composed images (cyan curve, 29.3% EER). This is due to the fact that dense part description covers a part of background in addition to the foreground structure, and thus the classifier learns that some edges come from the background. If these edges are completely missing in the training data, the boosted part detectors do not expect them during testing, which leads to frequent misclassifications.

Although the rendered edges are one pixel thick, their composition with the background edges can result in cluttered regions and thus edge thinning is needed. Fig. 4.3(a) shows that edge thinning in a Canny-detector-like manner helps to increase the performance from 29.3% EER to 40.1% EER. Thus, in the following experiments,



(a) Edge thickness and background composition

(b) Sigma shape

Figure 4.3: Effects of (a) rendered edge thickness and composition with the background, and (b) allowed deviations from the mean shape. Composition of rendered and background edges following by edge thinning is essential for reasonable performance. Maximum range ±3 sigma of realistic shape variations leads to the best detection results.

we always perform edge thinning after composing rendered and background edge images.

Thickness of rendered edges. We observe that during rescaling and rotation of individual parts to the canonical orientation prior to learning of individual part detectors (Andriluka *et al.*, 2009) some fragments of edges can be lost. We thus increase the thickness of rendered edges to two and three pixels and compare the results. As it can be seen from Fig. 4.3(a), rendering of two pixel edges helps to increase the performance from 40.1% to 52.7% EER. This result is also better than for three pixel edges. Therefore we render images with two pixel edges in the rest of our experiments. Examples of rendered images with various edges are shown in Fig. 4.2(a).

Shape variations. Another important parameter is the range of shape variations in the 3D human shape model. We study the optimal variations by uniformly sampling shape parameters from the range of ± 1, 2 and 3 sigma (= standard deviation) from the mean shape. Fig. 4.3(b) shows detection performances for those cases. It can be seen that the best result is achieved for the range ± 3 sigma (red curve, 52.7% EER). This is due to the fact that the data covers more variability then in case of ± 2 and ± 1 sigma (47.9% and 44.5% EER, respectively). We also compare this results to Gauss-sampling from the interval ± 3 sigma. As it can be seen, uniform sampling outperforms Gauss-sampling for the same interval (52.7% EER vs. 48.7% EER). Here the intuition is that the shape distribution learned from the database of human scans is narrower than the distribution in the test set of real pedestrians, and thus the tails of the shape distribution are essential and can be better represented by uniform sampling. This underlines the argument that it is important to vary the sample distribution systematically rather than blindly adding more training samples in order to cover essential parts of the distribution that improve detection performance. Fig. 4.2(b) shows sample images with various deviations from the mean waist girth.

(a) Rendered edge images (b) Real pedestrians

Figure 4.4: Sample images of (a) rendered humans with shape, pose and viewpoint variations and (b) real pedestrians from the test set. Edge orientation is encoded as hue and edge strength as saturation in the HSV color system. Notice significant variability in clothing, shape, poses and viewpoints from which real pedestrians are seen.

4.4 EXPERIMENTAL EVALUATION

In this section we experimentally evaluate our approach of synthetic data generation described in Sec. 4.3.2 in various settings and compare it to the best known results from the literature obtained when using real training images only. First, we briefly introduce the datasets used for training and evaluation. Then we show the results when the detector is trained on synthetic data alone. Importantly and as argued before, joint training on real data with complementary shape information coming from the rendered samples allows to increase the performance over real data alone. Finally, we combine detectors trained on different datasets which allows us to achieve the best people detection results on a challenging multi-viewpoint dataset.

Rendered dataset. We generate 15,000 non-photorealisticly rendered edge images, as described in section 4.3.2. We uniformly sample shape parameters of six semantic attributes within ±3 standard deviation from the mean shape and set the edge width to two pixels as these parameters led to the best detection performance (c.f. Sec. 4.3.2). Each training sample is rendered with 200 pixels height. As background dataset we use a set of 350 images without pedestrians. See for samples without background Fig. 4.4(a) where edge orientation is color-coded.

Multi-viewpoint dataset. The second dataset used in our experiments is the multiview data proposed by (Andriluka *et al.*, 2010). The dataset contains images of real

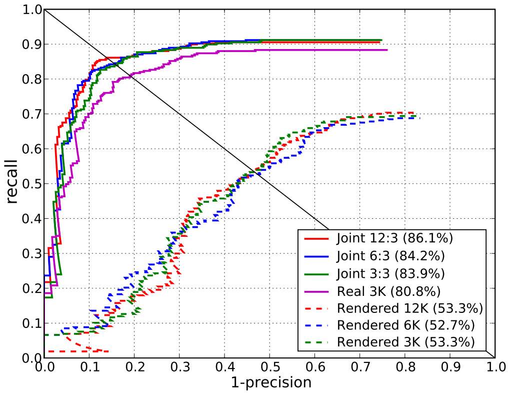


Figure 4.5: Results using rendered data alone and jointly with real data. Our detector trained on joint data outperforms the one trained on real data alone.

pedestrians from arbitrary viewpoints. It includes 1486 images of part-annotated pedestrians for training, 248 for testing and 248 for validation. In order to increase the amount of training data the images from the training set were mirrored. Samples from the multi-viewpoint dataset are shown in Fig. 4.4(b).

Experimental setup. For evaluation of trained detection models we use the test set of the multi-view dataset and thus are able to directly compare the obtained results to the performance of (Andriluka *et al.*, 2010) which we denote as *Andriluka* in the following, and to our method described in Chapter 3. In our experiments with rendered data we use the negative set of the multi-viewpoint dataset. We provide all results as precision-recall curves and use equal error rates (EER) to compare the performance. To match ground truth annotations to object detections we use the PASCAL criterion (Everingham *et al.*, 2010) which implies at least 50% overlap between ground truth and detection. We consider only detections which are at least 100 pixels high and additionally annotate all smaller pedestrians which were not annotated in the test set and thus when detected counted as false positives. For fairer comparison, we re-evaluate detector of (Andriluka *et al.*, 2010) and our detector from Chapter 3 for this setting.

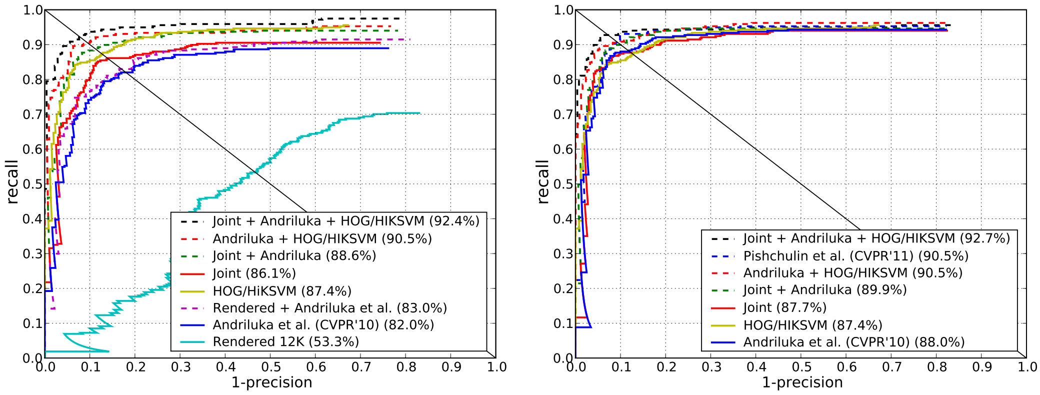
4.4.1 Results using rendered data alone and combined with image data

First we train the pictorial structures model on non-photorealisticly rendered edge images alone. We also compare the results to the performance obtained while training on real and joint real/ rendered data. Fig. 4.5 shows the results when using 3K, 6K and 12K synthetic images. It can be seen that training on the rendered data alone achieves reasonable performance of 53.3% EER (dotted green and red curves), especially when taking into account that internal edges are missing in the synthetic data. However, the difference in performance due to various amounts of synthetic data is insignificant. In contrast, by adding different numbers of synthetic samples to the real ones and thus bringing various portions of complementary shape information into the training data we are able to noticeably improve detection performance over real data alone. The best result is achieved when jointly trained on 3K real and 12K synthetic samples, improving the detection rate from 80.8% (real data alone) to a remarkable 86.1% EER (combined synthetic with real data). This improvement clearly shows the power of using our synthetic data in addition to real images as it helps to increase the variability of training data and thus allows for better generalization of people detection models. The obtained result corresponds to the best ratio (12:3) between synthetic and real data. Increasing the amount of synthetic data leads to worse performance (84.9% EER) due to overfitting of the detection model to synthetic samples while decreasing reduces the shape variability gained through additional samples (c.f. Fig. 4.5 blue and green curves).

4.4.2 Combination of people detection models

Good results for joint training on synthetic and real data motivate experiments to combine different and potentially complementary people detectors trained on both datasets. In this section we explore various combinations of people detection models by following a detector stacking strategy. For that purpose we train the pictorial structures model on different datasets and then combine detectors by a linear SVM trained on vectors of mean-variance normalized detector outputs. For combination we use dense detection score maps which contain position and scale of detection hypotheses produced by individual detectors. For each detection score of each detector we build a feature vector by concatenating this score with the scores of all other detectors at current scale and position respecting the overlap between detections' bounding boxes. We use the validation set of multi-view data for SVM training.

To combine people detection models we consider two settings. In the first setting we train generic detectors – i.e. single detectors that are trained simultaneously on all viewpoints – and combine them with the best generic detector of (Andriluka *et al.*, 2010) which we refer to as *Andriluka*. In the following discussion *Andriluka* corresponds to the detector trained on real data only, *Rendered* to the detector trained on synthetic data only and *Joint* trained on the combination of both data. The results are shown in Fig. 4.6(a). As it can be seen from the plot the combination *Andriluka* and *Rendered* improves the performance over *Andriluka* alone (83.0 vs. 82.0% EER) which means that the *Rendered* is complementary to the *Andriluka* despite missing internal edges and clothes. However, combination of *Joint* and *Andriluka* helps to increase the detection rate from 86.1% EER to 88.6% EER. This improvement means that indeed the detector trained on joint data is complementary to the one trained on real data alone. We compare the obtained result to the monolithic sliding window-based detector computed from histograms of oriented gradients (HOG) (Dalal and Triggs, 2005) and used for SVM training with the histogram intersection



(a) Generic detectors (b) Multi-viewpoint detectors

Figure 4.6: Combination of (a) generic and (b) multi-viewpoint detectors. Combination of our detector jointly trained on the real and synthetic data with the one trained on the real data helps to improve detection performance.

kernel (HIKSVM) (Maji *et al.*, 2008). The *HOG/HIKSVM* (yellow line, 87.1% EER) performs worse than our *Joint+Andriluka* detector (88.6% EER). The performance improves in case of *HOG/HIKSVM+Andriluka* (90.5% EER), while the combination of *HOG/HIKSVM* with *Joint* and *Andriluka* allows to achieve the best detection results (black dotted curve, 92.4% EER) outperforming *Joint+Andriluka* and *Andriluka* alone. This can be explained by more accurate scale estimation by *HOG/HIKSVM* detector which resolves the first false positives by the pictorial structures model such as pedestrian detections on the wrong scale.

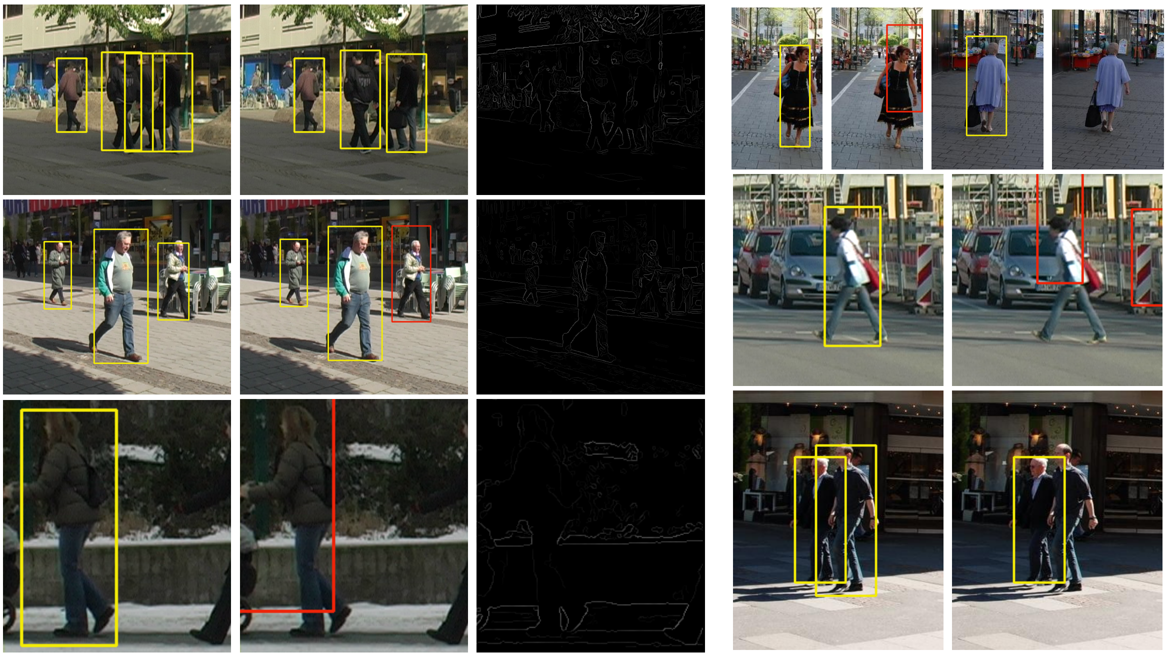
In the second setting we first train individual viewpoint-specific detectors on appropriate subsets of the respective real and synthetic data and combine them again by means of an SVM classifier. In contrast to the previous case, where only one generic detector was trained on each subset, training the pictorial structures model separately for each single view allows for better adjustment of the individual part detectors and the tree prior for each viewpoint, making the detection model more discriminative. The results are shown in Fig. 4.6(b). First, the combination of 8 viewpoint-specific and one generic detector all trained on joint real and synthetic data outperforms the generic detector alone (87.7% EER vs. 86.1% EER, c.f. Fig. 4.6(a)), which supports the intuition that viewpoint-specific part detectors are more discriminative. The obtained results are on par with (Andriluka *et al.*, 2010) (88.0% ERR) who not only combined 8 viewpoint-specific and one generic detector, but also used 2 side-view detectors additionally containing feet. By enriching this set of detectors by 8 viewpoint-specific detectors and one generic detector trained on joint data, we outperform the results of (Andriluka *et al.*, 2010) achieving 89.9% EER. Moreover, adding a *HOG/HIKSVM* to the bank of these detectors helps to further boost the performance up to 92.7% ERR outperforming *HOG/HIKSVM+Andriluka* (90.5% EER). We compare the results to our previous method from Chapter 3 which we denote as *Pishchulin et al., (CVPR'11)* (blue dotted curve) where we also use morphable 3d body model for generating photorealistic synthetic training samples and combine the detectors in the same way to boost the performance. It can be seen that we outperform the results of our previous method (92.7% vs 90.5 %ERR) despite using non-photorealistically rendered images of pedestrians with missing clothes and internal edges. It can be explained by the fact that the approach introduced in this chapter is more flexible and allows to easily produce lots of training data with much variation in shape and pose whereas the method from the previous chapter is unable to represent pedestrians in unseen poses and appearances. method achieves the best known detection

Discussion. In order to understand the reasons for better performance of joint training on synthetic and real data compared to the training on real data alone, we analyze typical failures of *Andriluka* and compare those to the detections by our *Joint* detector. Fig. 4.7(a) shows sample detections of both models at the EER, as well as edge images obtained by Canny edge detector (Canny, 1986). It can be seen that *Andriluka* (middle column) fails when edge evidence for several body parts is missing or only partial e.g. due to occlusion (top row) or poor contrast (middle and bottom row). However, our detector (left column) is able to cope with such hard cases by focusing on shape evidence taken from external (human shape) edges. This underlines the argument that the synthetic data does indeed contain complementary information, namely the additional shape information, w.r.t. the real data alone. This clearly shows the advantage of using our approach for rapid generation of synthetic data with relevant shape distribution to increase the variability covered by the limited number of real samples.

We also analyze the advantages of using the combination of generic detectors *Joint + Andriluka + HOG/HIKSVM* over *Andriluka* alone. Sample detections at the EER in both cases are shown in Fig. 4.7(b). It can clearly be seen that our stacked detector is able to more accurately detect pedestrians when *Andriluka* fails. This suggests that our stacked detector generalizes better and can even detect pedestrians having unusual shapes (top row). This again means that both *HOG/HIKSVM* and our *Joint* detector are complementary to *Andriluka* and provide additional information helping to improve detection performance.

4.5 CONCLUSION

In this chapter we explored the potential of synthetically generated data from a 3D human shape model in order to enrich image-based data with complementary shape information. As we use a 3D human shape and pose model we are able to generate thousands of synthetic instances having a relevant and complementary shape distribution, covering a wide range of human shapes and poses. We show that careful design of the rendering procedure where model instances are rendered from various viewpoints into non-photorealistic edge images allows to compute low level feature representations which generalize to unseen real test data. Experimental



(a) Joint vs. Andriluka

(b) Our combined detector vs. Andriluka

Figure 4.7: Typical failures of Andriluka's detector at EER and its comparison to (a) generic detector trained on joint rendered and real data and (b) combination of generic detectors. Andriluka's detector (every second picture) fails due to missing edge evidence while our detector trained on joint data relies on partial shape evidence taken from external edges. Edge images better seen on the screen when zoomed. Combination of our detector with *Andriluka* and *HOG/HIKSVM* helps to detect pedestrians despite unusual shapes, clutter, image blur and poor contrast.

results revealed significant improvements in detection performance when training a detector on joint synthetic and real data over training on real data alone, supporting our claim about complementarity of our synthetically generated data to the real data alone. Finally, we show in the generic as well as in the multi-viewpoint setting that the combination of our detector trained on the joint data with other detectors trained on real data alone allows to improve the performance on a challenging multi-viewpoint dataset.

We observed though that the performance of synthetic data only is below the performance of real data, as appearance distribution of the synthetic training data is different from the test time real data due to missing internal edges and clothing information. We address this shortcoming in the next chapter.

ARTICULATED PEOPLE DETECTION AND POSE ESTIMATION: RESHAPING THE FUTURE

5

Contents

5.1 Introduction **85**
5.2 Generation of novel training examples **87**
  5.2.1 Data annotation 87
  5.2.2 3D human shape recovery and animation 87
  5.2.3 Generation of novel images 89
5.3 Articulated people detection **90**
5.4 Articulated pose estimation **92**
  5.4.1 Pictorial structures model 93
  5.4.2 Joint PS-DPM model 93
  5.4.3 Experimental evaluation 94
5.5 Articulated pose estimation "in the wild" **96**
5.6 Conclusion **98**

In this chapter we propose a method that allows to generate a large number of *photo-realistically* looking synthetic training samples with controllable shape and pose variations from *arbitrary monocular* images, thus overcoming the main limitations of the methods presented in Chapters 3 and 4. To that end we build on recent advances in computer graphics to generate samples with realistic appearance and background while modifying body shape and pose. We validate the effectiveness of our approach on the task of articulated human detection *and* articulated pose estimation. We report the improvements of pose estimation results on the popular Image Parsing (Ramanan, 2006) human pose estimation benchmark and demonstrate superior performance for articulated human detection. In addition we define a new challenge of combined articulated human detection and pose estimation in real world scenes.

5.1 INTRODUCTION

Recent progress in people detection and articulated pose estimation may be contributed to two key factors. First, discriminative learning allows to learn powerful models on a large training corpora (Bourdev and Malik, 2009; Felzenszwalb *et al.*, 2010; Yang and Ramanan, 2011; Tompson *et al.*, 2014). Second, robust image representations enable to deal with image clutter, occlusions and appearance variation (Dalal and Triggs, 2005; Mikolajczyk and Schmid, 2005; Krizhevsky *et al.*, 2012). Large and

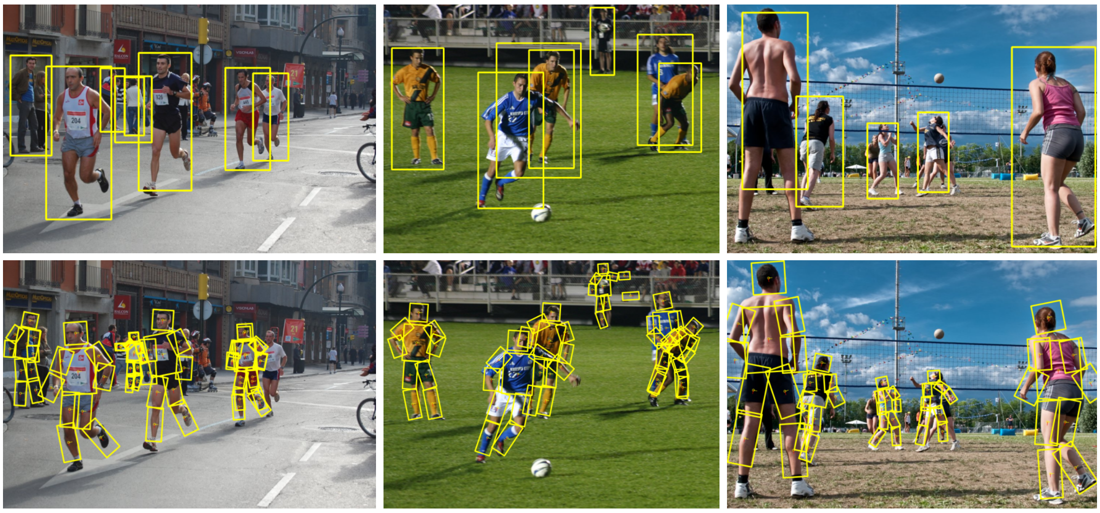


Figure 5.1: Sample detections (top) and pose estimates (bottom) of multiple articulated people obtained with our model trained on images from our new data generation method.

representative training sets are essential for best performance and significant effort has been made collecting them (Bourdev and Malik, 2009; Johnson and Everingham, 2011; Everingham *et al.*, 2010; Lin *et al.*, 2014). Typically, images are extracted from public data sources (e.g. photo collections) and manually annotated. However, even for large datasets it remains a challenge to ensure that they adequately cover the space of possible body poses, shapes and appearances. Even more importantly, depending on the task (e.g. detecting people in basketball vs. golf vs. street-scenes) the *relevant* distribution of shape, body pose and appearance varies greatly and cannot be easily controlled using manually collected datasets.

In this chapter we are interested in the challenging problem of articulated people detection *and* pose estimation in challenging real world scenes. In order to achieve this goal (e.g. illustrated in Fig. 5.1), we make several contributions. As a first contribution, we propose a novel method for automatic generation of multiple training examples from an arbitrary set of images with annotated human body poses. We use a 3D human shape model (Jain *et al.*, 2010) to produce a set of realistic shape deformations of person's appearance, and combine them with motion capture data to produce a set of feasible pose changes. This allows us to generate realistically looking training images of people where we have full control over the shape and pose variations. As a second contribution, we evaluate our data generation method on the task of articulated human detection and on the task of human pose estimation. We explore how various parameters of the data generation process affect overall performance. On both tasks we can significantly improve performance when the training sets are extended with the automatically generated images. As a third contribution, we propose a joint model that directly integrates evidence from an appropriately trained deformable part model (DPM) (Felzenszwalb *et al.*, 2010) into a pictorial structures framework and demonstrate that this joint model further improves performance. Last, as fourth contribution, we define a new challenge of joint detection and pose estimation of multiple articulated people in challenging real world scenes.

5.2 GENERATION OF NOVEL TRAINING EXAMPLES

To improve both articulated people detection and pose estimation we aim to generate training images with full control over pose and shape variations. Fig. 5.2 gives an overview of our novel data generation process consisting of three stages. Starting from approximate 3D pose annotations we first recover the parameters of the 3D human shape model (Jain *et al.*, 2010). The body shape is then modified by *reshaping* and *animating*. Reshaping changes the shape parameters according to the learned generative 3D human shape model and animating changes the underlying body skeleton. Given the new reshaped and/or animated 3D body shape we back-project it into the image and morph the segmentation of the person. To that end we employ the linear blend skinning procedure with bounded biharmonic weights described in (Jacobson *et al.*, 2011). The following describes these steps in more detail.

5.2.1 Data annotation

For each subject in the training set we manually provide a 3D pose and a semi-automatic segmentation of the person. The 3D pose is obtained using the annotation tool introduced in (Bourdev and Malik, 2009). The pose is used later to resolve the depth ambiguities which otherwise arise when fitting the 3D human shape model to 2D observations. The initial segmentation is obtained with GrabCut (Rother *et al.*, 2004) which we automatically initialize using annotated 2D joint positions and projected 3D shape from the fitted shape model (see below). While this procedure already produces reasonable results, segmented images often require user interaction to refine the segmentation due to low resolution, poor contrast and bad lighting. We use the segmentation to compute a 2D image mesh which is then deformed to change human shape and pose.

5.2.2 3D human shape recovery and animation

3D human shape model. In order to generate photorealistic synthetic images of people in different poses we employ a statistical model of 3D human shape and pose (Jain *et al.*, 2010) which is a variant of the SCAPE model (Anguelov *et al.*, 2005). The model is learned from a public database of 3D laser scans of humans and thus represents the available shape and pose variations in the population. The shape variation across individuals is expressed via principal component analysis (PCA). We use the first 20 PCA components capturing 97% of the body shape variation. Linear blend skinning is used to perform pose changes. To this end, a kinematic skeleton was rigged into the average human shape model. The 3D model pose is represented by a kinematic skeleton with 15 joints having a total of 24 degrees of

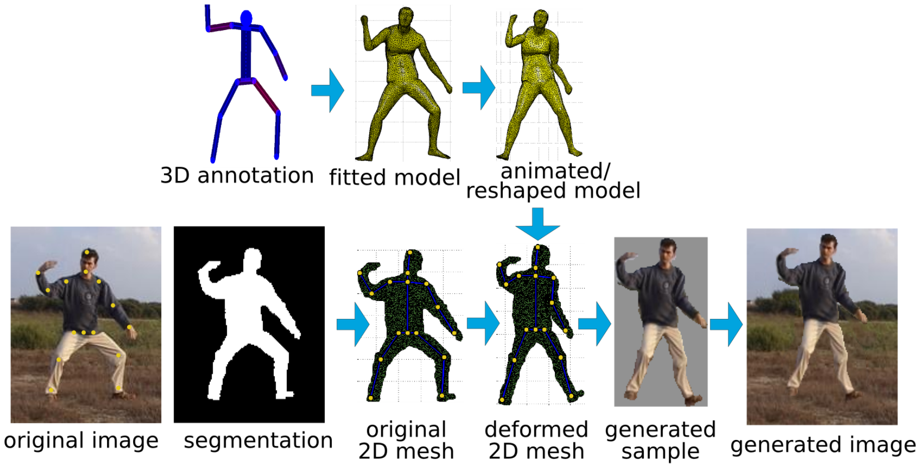


Figure 5.2: Overview of our novel data generation method.

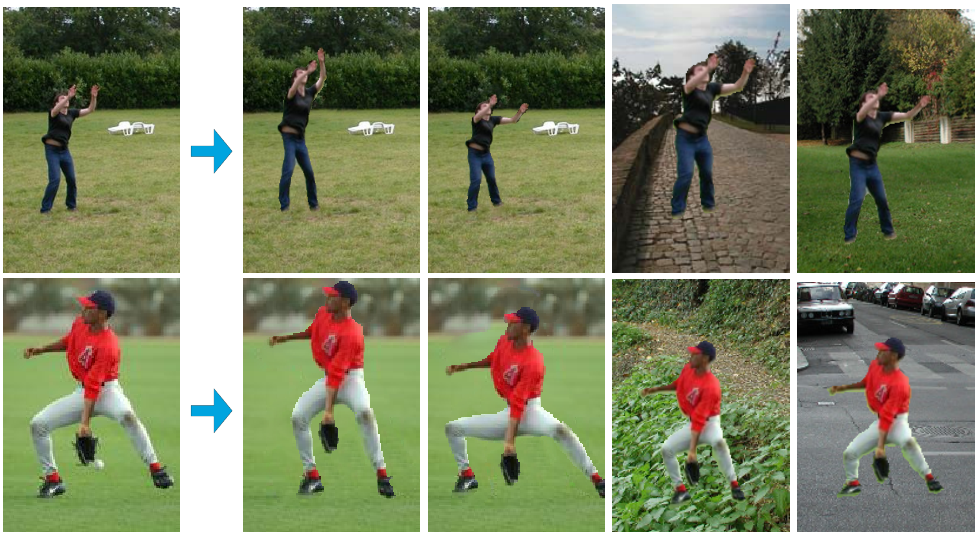
freedom (DoF) plus 6 DoF for global body position and orientation. The model surface consists of a triangle mesh with 6450 vertices and 12894 faces.

Model fitting. Having an annotated 3D pose allows to resolve the depth ambiguity while fitting the 3D shape model's kinematic skeleton to a 2D image. We retarget the skeleton to an annotated 3D pose by computing inverse kinematics through minimizing the Euclidean distance between a set of corresponding 3D joint positions, namely left/right ankles, knees, hips, wrists and elbows, upper neck and head. We use a constrained optimization based on the iterative interior point method. Optimization is done in shape and pose parameters space. Obtaining a good fit of the skeleton is essential for the rest of our data generation process and can significantly influence the realism of generated images. The fitting dependents on the flexibility of the kinematic skeleton and also on how well the corresponding 3D joint positions match. We thus do not include shoulders, pelvis and thorax joints into the objective function as these tend to have different positions in the annotated 3D pose and the 3D model's kinematic skeleton.

Varying model shape and pose. After fitting the skeleton we vary the 3D shape and pose parameters. To change the shape we randomly sample from the underlying 3D human shape distribution. For 3D shape animation we require a database of poses. To that end we retargeted the shape model's kinematic skeleton to over 280,000 of highly articulated poses from freely available mocap data[4]. To do so, we fix the bone lengths of the mocap skeleton to be the same as for the shape model's skeleton and compute inverse kinematics by optimizing over global rotation, translation and pose parameters only, which reduces the search space and produces better results. To animate the fitted skeleton we use the nearest retargeted poses with an average joint distance of less than 90 mm. Informal experiments showed

---

[4] CMU MoCap Database http://mocap.cs.cmu.edu/



(a) (b)

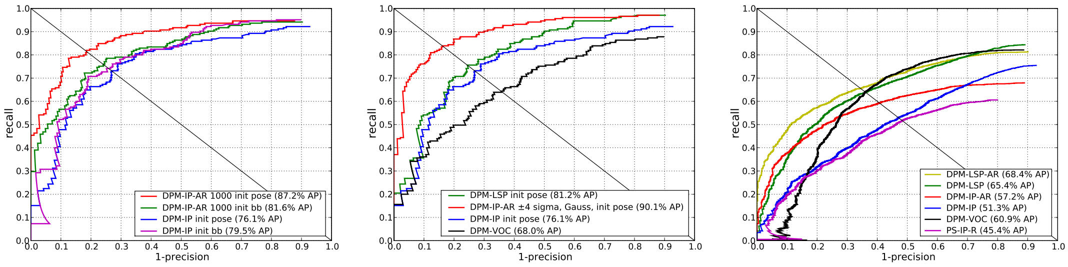
Figure 5.3: Examples of automatically generated novel images: (a) original image and (b) animated and reshaped synthetic samples with different backgrounds. Note the realism of the generated samples.

that going further away from the fitted pose may result in unrealistically looking generated images.

5.2.3 Generation of novel images

After shape and pose changes are applied to the fitted 3D shape model, we project its 3D joint positions into the image and move 2D annotated joints towards corresponding projected joints. This results in a smooth 2D mesh deformations described by linear blend skinning (Jacobson *et al.*, 2011). We only animate "dangling" arms and legs, and do not deform occluded or occluding limbs as this leads to unrealistic deformations.

To obtain a final training sample we render the deformed 2D mesh into a photorealistically looking individual by reusing the original appearance of the person. Finally we combine the rendered subject with the background. We either replace the original person with the generated one by first removing the original person from the image using a commercial implementation of (Barnes *et al.*, 2009), or embed the generated sample at a random place of a new people-free image. Fig. 5.3 shows original images from the "Image Parsing" set and automatically generated novel images with animated and reshaped humans and different types of backgrounds.



(a) (b) (c)

Figure 5.4: Comparison of different initializations for DPM components (a). Comparison of detection results of the DPM model on (b) "Image Parsing" and (c) multiscale "Leeds Sports Poses" datasets.

5.3 ARTICULATED PEOPLE DETECTION

This section evaluates our data generation method for articulated people detection. For this we use the deformable part model (DPM) (Felzenszwalb *et al.*, 2010) and evaluate its performance on the "Image Parsing" dataset (Ramanan, 2006). For training we use training sets from the publicly available datasets: PASCAL VOC 2009 (VOC) (Everingham *et al.*, 2010) consists of 2, 819 images of people captured over a wide range of imaging conditions; "Image Parsing" (IP) (Ramanan, 2006) consists of 100 images of fully visible people in a diverse set of activities such as sports, dancing, and acrobatics; the recently proposed "Leeds Sports Poses" (LSP) dataset (Johnson and Everingham, 2010) that includes 1, 000 images of people involved in various sports. We denote the models trained on these sets as DPM-VOC, DPM-IP and DPM-LSP. We introduce two new training sets obtained from IP by reshaping (R) and the combination of animating and reshaping (AR) training examples[5]. The models trained on this data *together* with the IP data are denoted DPM-IP-R and DPM-IP-AR accordingly. Average precision (AP) is used to compare performance and the PASCAL criterion (Everingham *et al.*, 2010) is used for matching.

DPM training. Training of DPM proceeds as usual (Felzenszwalb *et al.*, 2010). However, we found that the initialization of DPM components significantly influences detection performance. I.e. the standard way to initialize the components based on the bounding box (BB) aspect ratio does not appear to be well suited for our task, as people with different poses often have similar BBs. We explore an alternative initialization strategy, where we cluster the images according to the relative displacement of the 2D joint locations w.r.t. the fixed body joint (neck joint in our case). The comparison of detection performance is presented in Fig. 5.4(a). DPM-IP-AR outperforms DPM-IP (81.6% vs. 79.5% AP) even when initialized by BB aspect ratios. Initializing DPM by pose clustering leads to an unequal distribution of training samples among different components and thus some components suffer from the lack of training data. This explains the performance decrease for DPM-IP (76.1% AP). However pose clustering accounts for a significant improvement for

---

[5] The data is available for research purposes on our web page.

real/synthetic	AP, [%]
100 IP/0	76.1
100 IP/400 R	83.9
100 IP/400 AR	87.2
100 IP/900 AR	**88.6**
100 IP/1900 AR	88.1

Table 5.1: Results using "Image Parsing" (IP) data alone and jointly with Reshape (R) or Animate-Reshape (AR).

range	sampling	
	uniform	Gauss
$\pm 4\sigma$	85.4	**90.1**
$\pm 3\sigma$	**88.6**	85.1
$\pm 2\sigma$	88.0	83.2
$\pm 1\sigma$	85.6	87.3

Table 5.2: Results for different samplings of shape parameters in Animate-Reshape data.

DPM-IP-AR (87.2% AP), as each component gets enough training data. This underlines the argument that our data generation method does indeed help to cover more shape and pose variations compared to the real data alone.

Data ratio. We study the influence of increasing shape and pose variations in the training data by changing the ratio between AR and IP data (results in Tab. 5.1). Clearly, performance is worst when training on IP data alone (76.1% AP). Adding 400 of R samples (increasing only shape variations) noticeably improves performance (83.9% AP). However adding the same number of AR samples (increasing both shape and pose variations) accounts for further improvements (87.2% AP). This supports the intuition that a global articulated people detector requires training data with large shape and pose variations and thus can significantly profit from our data generation method. Increasing the amount of AR data further improves the performance to 88.6% AP. Adding even more AR samples leads to a slight decrease in performance due to overfitting.

Shape variations. The ability to sample from the underlying 3D human shape distribution provides a direct control over generated data variability. Thus it is important to evaluate various ranges of shape changes and different sampling strategies. We sample shape parameters within ± 1, 2, 3 and 4σ (standard deviation) from the mean shape using uniform and Gauss-sampling and report the results in Tab. 5.2 for 100 IP/900 AR data. For both uniform and Gauss strategies sampling from $\pm 3\sigma$ outperforms $\pm 2\sigma$ as it better covers the space of possible shapes. Interestingly, by Gauss-sampling from $\pm 4\sigma$ and thus oversampling the tails of possible shape variations represented by our 3D human shape model we are able to improve the performance to 90.1% AP. Intuitively, the tails of the data distribution are important for learning powerful detectors. Increasing the sampling range increases the likelihood to sample unlikely but possible shape variations, which is far more difficult to achieve when using manually collected datasets only.

Summary of detection results. In Fig. 5.4(b) we summarize our findings and compare the obtained results to both DPM-VOC and DPM-LSP. DPM-VOC performs

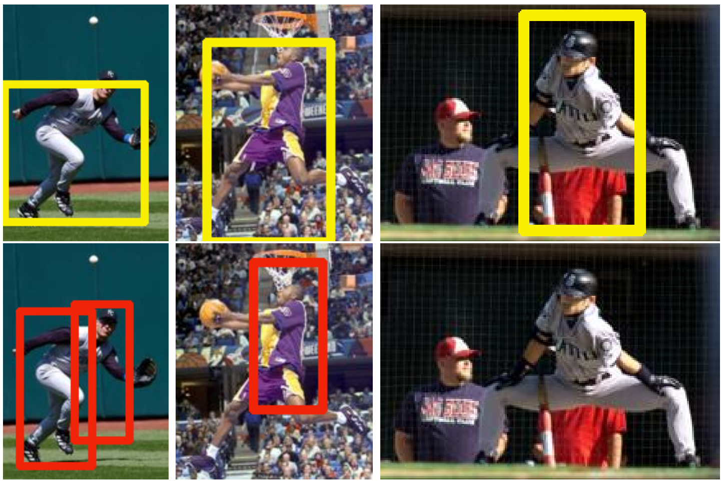


Figure 5.5: Examples of articulated people detections at EER by DPM trained on our joint synthetic and real "Image Parsing" (IP) data (top) and IP data alone (bottom). DPM trained on VOC2009 failed to detect people in these images.

the worst (68.0% AP) trained on mostly upright people without strong articulations. This intuition is also supported by a better performance of DPM-IP (76.1% AP) trained from a much smaller set of images containing highly articulated people. Although training from a larger number of real samples (DPM-LSP) increases the detection rate (81.2% AP) this improvement is less pronounced compared to DPM-IP-AR (90.1% AP). This is due to the fact that the data variability is uncontrolled in LSP, as thus by adding more real samples we do not necessarily increase the variability. Training on our data generated from only 100 real images and having controllable pose and shape variations outperforms other models by a large margin achieving a remarkable 90.1% AP. We also show example detections at the equal error rate for DPM-IP-AR and DPM-IP in Fig. 5.5. Both qualitative and quantitative results clearly show the advantage of our method to increase the shape and pose variability of training data by sampling from the underlying 3D human shape distribution and changing human poses.

5.4 ARTICULATED POSE ESTIMATION

Motivated by the success of our data generation method to enable articulated people detection, this section proposes a new joint model for body pose estimation combining our pictorial structures model with DPM. We first briefly describe the Pictorial Structures (PS) model (Fischler and Elschlager, 1973; Felzenszwalb and Huttenlocher, 2005) and then introduce our novel Joint PS-DPM model. We evaluate both models on the challenging "Image Parsing" dataset and show that pose estimation can directly profit from our strong articulated people detector. We use the percentage of correct parts (PCP) (Ferrari *et al.*, 2008) measure for performance comparison.

5.4.1 Pictorial structures model

Pictorial structures (PS) (Felzenszwalb and Huttenlocher, 2005; Fischler and Elschlager, 1973) represent the human body as a flexible configuration $L = \{l_0, l_1, ..., l_N\}$ of body parts. The state of each part i is denoted by $l_i = (x_i, y_i, \theta_i, s_i)$, where (x_i, y_i) gives the part position in image coordinates, θ_i the absolute part orientation, and s_i indicates the part scale relative to the part size in the scale normalized training set. Given image evidence E, the posterior of the part configuration L is described by $p(L|E) \propto p(E|L)p(L)$, where $p(L)$ is the kinematic tree prior and $p(E|L)$ is the likelihood of image evidence E for the body part configuration L. The tree prior describes dependencies between model parts and can be factorized as $p(L) = p(l_0) \prod_{(i,j) \in G} p(l_i|l_j)$, where G is the set of all directed edges in the kinematic tree, l_0 is assigned to the root node (torso) and $p(l_i|l_j)$ are pairwise terms along the kinematic chains. Pairwise terms are modeled by Gaussians in the transformed space of part joints while $p(l_0)$ is assumed to be uniform. The likelihood term is decomposed into the product of single part likelihoods $p(E|L) = \prod_{i=0}^{N} p(E|l_i)$,

We use our publicly available implementation (Andriluka *et al.*, 2011). In this implementation part likelihoods are modeled with AdaBoost classifiers (Freund and Schapire, 1997) and image evidence is represented by a grid of shape context descriptors (Belongie *et al.*, 2002). Inference is performed by sum-product belief propagation, which allows to compute marginal posteriors of each body part.

5.4.2 Joint PS-DPM model

While being conceptually similar, the DPM model and the PS model are designed for different tasks. The DPM model is designed for object detection and its parts are optimized to localize a bounding box of the person only. In particular it is non-trivial to map these parts to the locations of the anatomical body parts as is necessary for human pose estimation. On the contrary the PS model is defined directly in terms of anatomical parts and explicitly models their mutual positions and orientations. However anatomical body parts are not necessary optimal for detection, as they might be non-discriminative with respect to background.

To benefit from the complementary properties of PS and DPM models we define a joint model by embedding the evidence provided by DPM model into the PS framework. In the joint model we define the likelihood of the torso part as a product of two likelihood terms $p(E|l_i) = p_{ps}(E|l_i)p_{dpm}(E|l_i)$, where the first term is the original PS torso likelihood, and the second term is given by the torso prediction from the DPM. We adapt the DPM model to estimate the torso location by training linear regression model that predicts torso endpoints from the positions of the DPM model parts. These estimates are robust since the torso is typically associated with multiple parts of the DPM, which reduces uncertainty in the prediction. For each predicted torso location l_i we define $p_{dpm}(E|l_i) = \sigma(\mathrm{m}(l_i))$, where $\mathrm{m}(l_i)$ is the confidence score of the DPM detection, and $\sigma(\cdot)$ is a sigmoid function that calibrates the DPM score with respect to the PS likelihood. For all locations that did not have

Setting	Torso	Upper legs	Lower legs	Upper arms	Fore-arms	Head	Total
Image Parsing (IP)	84.9	71.5	61.5	50.2	36.6	71.2	59.6
+ Reshape (R)	87.8	75.1	65.9	52.4	36.1	71.7	61.9
+ Joint PS+DPM	**88.8**	**77.3**	**67.1**	53.7	36.1	73.7	63.1
(Andriluka *et al.*, 2011) *	83.9	70.5	63.4	50.5	35.1	70.7	59.4
(Yang and Ramanan, 2011) *	82.9	69.0	63.9	55.1	35.4	**77.6**	60.7
(Johnson and Everingham, 2011)	87.6	74.7	**67.1**	**67.3**	**45.8**	76.8	**67.4**

* evaluated using our implementation of PCP criteria introduced in (Ferrari *et al.*, 2008)

Table 5.3: Pose estimation results (PCP) on the "Image Parsing" (IP) dataset.

torso predictions we set the likelihood to $\varepsilon = 10^{-3}$.

5.4.3 Experimental evaluation

Here we evaluate both original PS and the proposed joint PS-DPM model on the task of pose estimation. In the following experiments the spatial and the part likelihoods of both models are learned on different training data, namely real "Image Parsing" (IP) data alone and together with the Reshape (R) data produced by our data generation method.

Training on IP data alone. First we report the best results obtained by training the PS model on IP data only. Similar to (Andriluka *et al.*, 2011) we train part detectors on the training set augmented with the slightly rotated, translated and scaled versions of the original training samples. As in (Andriluka *et al.*, 2011), we use a repulsive factor for lower and upper legs and perform inference by loopy belief propagation on the reduced state space of samples from part posteriors. Using IP data only we achieve 59.6% PCP. The results are shown in Tab. 5.3.

Training on IP and Reshape data. Our findings indicate that by jointly training on IP and Reshape data we improve over IP data alone. The best result is achieved by adding 1200 synthetic samples to the training data (61.9% PCP). Further increasing the proportion of Reshape samples leads to worse performance due to overfitting to synthetic samples, while decreasing the number of Reshape samples reduces variability and leads to worse performance. Training on IP and Animate-Reshape (AR) data (60.0% PCP) performs slightly worse than the Reshape data. The PS does not benefit from the animated training data as it can already model such transformations via a flexible pose prior. The best performance is achieved by uniformly sampling from $\pm 1\sigma$ (61.9% PCP) while Gauss-sampling performs slightly worse ($\pm 2\sigma$, 61.2% PCP).

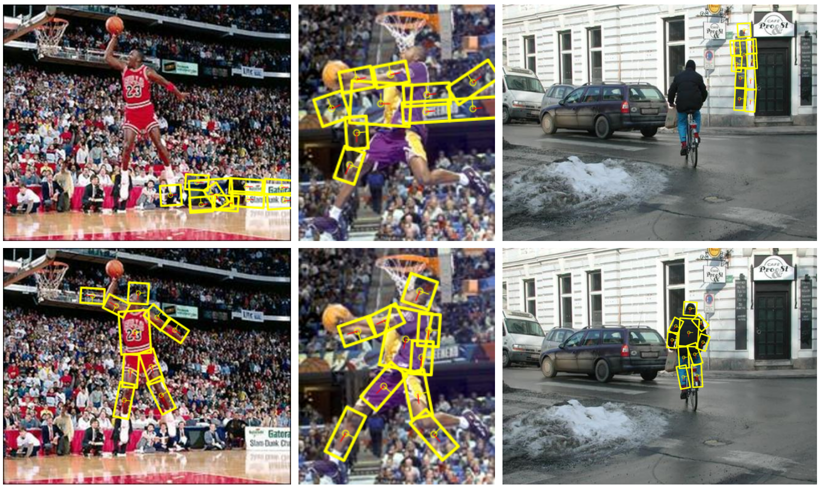


Figure 5.6: Comparison of body pose estimation results between the PS trained on IP (top) and our Joint PS+DPM model trained on IP + Reshape data (bottom).

Training Joint PS-DPM. Results of training our Joint PS-DPM model on IP and AR data are shown in Tab. 5.3 (row 3). The Joint PS-DPM model outperforms the PS model alone (63.1% vs. 61.9% PCP). Expectedly, the localization of torso improved (87.8% vs. 88.8% PCP) which is explained by the increased confidence of torso estimation in the Joint PS-DPM model. Clear improvement is achieved for all body parts apart from forearms, while the limbs directly connected to the torso profit at most.

Comparison to competitors. We compare our results to other results from the literature in Tab. 5.3. We outperform the method (Andriluka *et al.*, 2011) and more complex discriminatively trained mixtures of parts model (Yang and Ramanan, 2011). The achieved performance is slightly below (Johnson and Everingham, 2011) who use far more training data and learn *multiple* PS models after clustering similar poses. We envision that their clustering scheme could be effective in our case as well, in particular since we could generate sufficient amounts of training data even for clusters with rare poses. We leave this extension to future work.

Note that the results of (Yang and Ramanan, 2011) presented in Tab. 5.3 differ from those found in the original publication. The difference is due to the use of evaluation toolkit provided with the “Buffy” dataset (Ferrari *et al.*, 2008), which deviates from the PCP criteria introduced in (Ferrari *et al.*, 2008) in several ways leading to higher PCP scores[6]. For the sake of comparison we re-evaluate our method

---

[6] According to the definition of PCP from (Ferrari *et al.*, 2008) the body part is considered correct if *both* of its endpoints are closer to their ground truth positions than a threshold. The code in “Buffy” toolkit requires that the *average* over endpoint distances is smaller than the threshold. Such loose matching allows a segment to be accepted even if it is far from the ground-truth, because small distance of one endpoint can compensate for a large distance of the other endpoint. Another difference is that the code accepts multiple pose hypotheses as input, and evaluates the PCP score *only* for the hypothesis matching the ground-truth upper body bounding box. This is the “best case”

Setting	Torso	Upper legs	Lower legs	Upper arms	Fore-arms	Head	Total
ours, our evaluation	88.8	77.3	67.1	53.7	36.1	73.7	63.1
ours, loose matching	92.7	84.1	74.4	62.2	44.1	81.0	70.3
ours, evaluation of (Yang and Ramanan, 2011)	**98.9**	**90.1**	**79.6**	68.8	48.1	92.5	**76.5**
(Yang and Ramanan, 2011), our evaluation	82.9	69.0	63.9	55.1	35.4	77.6	60.7
(Yang and Ramanan, 2011), loose matching	88.8	78.5	71.7	70.7	41.7	81.5	69.6
(Yang and Ramanan, 2011)	97.6	83.9	75.1	**72.0**	**48.3**	**93.2**	74.9

Table 5.4: Pose estimation results (PCP) on the "Image Parsing" (IP) when using our evaluation and evaluation of (Yang and Ramanan, 2011).

using the publicly available toolkit (Ferrari *et al.*, 2008). The results are shown in Tab. 5.4. Clearly, both peculiarities of evaluation procedure employed by (Yang and Ramanan, 2011) contribute to significantly higher PCP results.

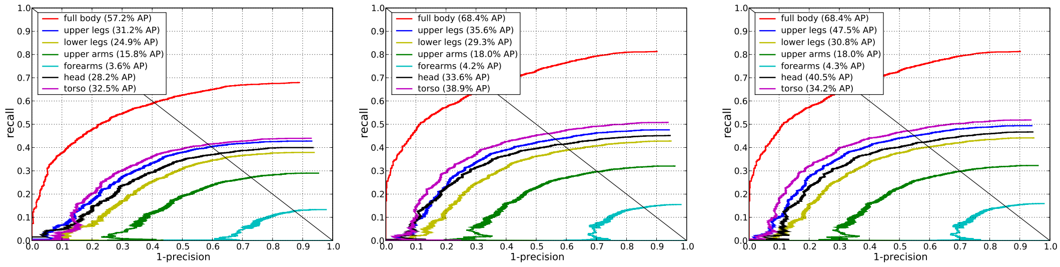
In Fig. 5.6 we show examples of pose estimation results by our joint PS+DPM model trained on Reshape data and PS model trained on IP data alone. Note that the PS fails due to background clutter (left and middle) and presence of human-like structures (right). The Joint model uses additional information from the DPM torso prediction and thus is more robust. Clearly, correct estimation of torso position is the key to correct estimation of the rest parts.

5.5 ARTICULATED POSE ESTIMATION "IN THE WILD"

Most recent work on articulated pose estimation considers a simplified problem by assuming that there is a single person in the image and that an approximate scale and position of the person is known (Ramanan, 2006; Johnson and Everingham, 2010; Sapp *et al.*, 2010a; Wang *et al.*, 2011). The proposed approaches typically output a single estimate of body configuration per image and do not provide any confidence score that the pose estimate is indeed correct. This ignores two important issues which arise when applying these approaches on real images. First, many images contain multiple people and so in addition to estimating poses of people it is also necessary to decide how many people are present. Second, for each person it becomes necessary to search over a wide range of possible positions and scales, and it is not clear how well the proposed methods are able to deal with such increase in complexity. We argue that in order to properly asses the state-of-the-art in articulated people detection and pose estimation it is necessary to consider these problems jointly. To that end we define a new dataset and evaluation criteria, and use them to validate the results obtained in Sec. 5.3 and Sec. 5.4 in a more realistic setting.

---

evaluation that relies on the ground-truth annotation. In contrast, the PCP criteria (Ferrari *et al.*, 2008) assumes there is one hypothesis for each part per image.



(a) DPM-IP-AR + PS (b) DPM-LSP-AR + PS (c) DPM-LSP-AR + Joint PS+DPM

Figure 5.7: Detection results of different body parts on the multi-scale "Leeds Sport Poses" dataset.

Dataset and evaluation criteria. The "Leeds Sport Poses" (LSP) dataset (Johnson and Everingham, 2010) contains images of people rescaled to the same scale and cropped around the person bounding box. We define a new dataset based on the LSP by using the publicly available original non-cropped images. This dataset, in the following denoted as "multi-scale LSP", contains 1000 images depicting multiple people in different poses and at various scales. We extended the annotations on the new dataset to include ground truth body configurations and bounding boxes of all people taller than 150 pixels resulting in $2,551$ annotated people total. To jointly asses the performance of detection and articulated pose estimation we evaluate the pose estimation in terms of recall and precision curves (RPC) and use AP to compare the performance. The PASCAL criterion (Everingham *et al.*, 2010) is used for matching people detections to the ground truth. For part matching to the ground truth we employ the PCP measure (Sec. 5.4) and use the people detector score as a confidence score of the hypothesis of each part. In addition to already mentioned training data we animate and reshape original LSP (Johnson and Everingham, 2010) training images (LSP-AR) and use them to train a DPM.

Results. Similar to (Ferrari *et al.*, 2008) we use pre-filtering by running an articulated people detector. We collect all detections at the highest recall, and estimate poses independently for each of the detections matching the ground truth. All misdetections are considered when computing an RPC curve for each part.

We first evaluate the performance of DPM trained on different types of data. Results are shown in Fig. 5.4(c). Again DPM-IP-AR is much better than DPM-IP (57.2% vs. 51.3% AP), while DPM-LSP-AR outperforms DPM-LSP (68.4% vs. 65.4%) achieving the best result. These results show that the detectors trained on data augmented with reshaped and animated examples are more robust to strong pose variations. All DPM models outperform the PS model that is not trained discriminatively and is therefore more prone to failures in the presence of background clutter.

Fig. 5.7 shows RPC curves for individual body parts corresponding to different combinations of detection and pose estimation models. The best result is achieved by combining the DPM-LSP-AR detector with our Joint PS+DPM model (Fig. 5.7(c)). The performance varies greatly across parts. The localization is especially difficult

Setting	Torso	Upper legs	Lower legs	Upper arms	Fore-arms	Head	Total
DPM-LSP-AR + Joint PS+DPM	**40.5**	**37.5**	**30.8**	**18.0**	**4.3**	**34.2**	**25.6**
DPM-LSP-AR	38.9	35.6	29.3	18.0	4.2	33.6	24.7
DPM-IP-AR + PS	32.5	31.2	24.9	15.8	3.6	28.2	21.2
DPM-VOC + PS	29.9	25.2	20.0	14.2	3.6	27.4	18.3
PS-IP-R + PS	29.1	28.7	23.5	14.7	4.0	24.5	19.5

Table 5.5: Average precision (AP) of part estimations by different methods on multi-scale "Leeds Sport Poses" dataset.

for small parts such as forearms that are frequently occluded and foreshortened. To compare part detection performance across different models, we summarize the results in Tab. 5.5. Using DPM-VOC for pre-filtering achieves 18.2% AP, which is below PS-IP-R + PS (19.2% AP) performing better at high precision (cf. Fig 5.4(c)). DPM-IP-AR + PS achieves 21.2% AP. By using DPM-LSP-AR which is a better people detector we significantly improve the performance to 24.7% AP: localization of torso and head improves by more than 5% AP, while upper and lower legs improve by 4.4%. This clearly shows the importance of using a robust people detector to improve pose estimation of highly articulated people on multiple scales. Finally, DPM-LSP-AR + Joint PS+DPM achieves the best result (25.6% AP) outperforming other models for all parts. Torso, head and upper legs benefit most from better torso detection, as our joint model is able to detect the torso with higher confidence. The somewhat low overall results are due to a large number of partially occluded and strongly articulated people seen from untypical viewpoints.

5.6 CONCLUSION

In this chapter we propose a novel method for automatic generation of training examples from an arbitrary set of images. By using a 3D human shape model we generate realistic shape deformations of peoples' appearance. In addition, we animate reshaped samples by using a large set of motion capture data to generate plausible pose variations. We evaluate our data generation method for articulated people detection and pose estimation and show that for both tasks we significantly improve the performance when augmenting existing training data with our automatically generated images. In particular, we achieve very good results on the challenging "Image Parsing" benchmark using just 100 real training images and a basic pictorial structures model. We also propose a joint model which integrates the evidence provided by DPM into the pictorial structures framework and experimentally show that the new model allows to further increase the performance. Finally we propose a new challenge of joint detection and pose estimation of multiple articulated people in cluttered sport scenes.

Similar to Chapters 3 and 4 our data generation method introduced in this chapter is based on the statistical 3D human shape model learned from the largest publicly available dataset that contains human body scans of hundred individuals and thus is limited in the range of captured shape variations. In Chapter 6 we systematically construct the statistical shape model from a large and much more representative commercially available dataset of human shapes and demonstrate superior performance of the newly build model.

BUILDING STATISTICAL SHAPE SPACES FOR 3D HUMAN MODELING

6

Contents

6.1 Introduction **102**
6.2 Statistical modeling with SCAPE **103**
  6.2.1 Original SCAPE model 103
  6.2.2 Simplified SCAPE space 104
6.3 Data processing **104**
  6.3.1 Non-rigid template fitting 104
  6.3.2 Initialization 106
  6.3.3 Bootstrapping 106
  6.3.4 Posture normalization 106
6.4 Evaluation of template fitting **107**
  6.4.1 Implementation details 107
  6.4.2 Quality measure 108
  6.4.3 Initialization 108
  6.4.4 *NRD* parameters 109
  6.4.5 Bootstrapping 110
6.5 Evaluation of statistical shape space **110**
  6.5.1 Quality measure 110
  6.5.2 Bootstrapping 111
  6.5.3 Number of training samples 112
  6.5.4 Posture normalization 112
6.6 Human body reconstruction **113**
  6.6.1 Dataset and experimental setup 113
  6.6.2 Quantitative evaluation 113
  6.6.3 Qualitative evaluation 115
6.7 Conclusion **117**

STATISTICAL models of 3D human shape and pose learned from scan databases have developed into valuable tools to solve a variety of vision and graphics problems. Unfortunately, most publicly available models are of limited expressiveness as they were learned on very small databases that hardly reflect the true variety in human body shapes. In this chapter, we contribute by rebuilding a popular statistical body representation used in Chapters 3, 4 and 5 from a large commercially available scan database, and making the resulting model available to the community (visit *humanshape.mpi-inf.mpg.de*). As preprocessing several thousand scans for learning the model is a challenge in itself, we contribute by developing robust best practice solutions for scan alignment that quantitatively lead to the best learned models. We make implementations of these preprocessing steps also publicly available. We extensively evaluate the improved accuracy and generality of our new model, and show its improved performance for human body reconstruction from sparse input data.

6.1 INTRODUCTION

Statistical human shape models represent variations in human physique and pose using low-dimensional parameter spaces, and are valuable tools to solve difficult vision and graphics problems, e.g. in pose tracking or animation. Despite significant progress in modeling the statistics of the complete 3D human shape and pose (Allen *et al.*, 2003; Anguelov *et al.*, 2005; Guan *et al.*, 2012; Chen *et al.*, 2013; Neophytou and Hilton, 2013; Hasler *et al.*, 2009; Jain *et al.*, 2010) only few publicly available statistical 3D body shape spaces exist (Hasler *et al.*, 2009; Jain *et al.*, 2010). Further on, the public models are often learned on only small datasets with limited shape variations (Hasler *et al.*, 2009). The reason is a lack of large representative *public* datasets and the significant effort required to process and align raw data scans prior to learning a statistical shape space.

This chapter contributes by systematically constructing a model of 3D human shape and pose from a large *commercially* available dataset of 3D laser scans (Robinette *et al.*, 1999) and making it publicly available to the research community (Section 6.2). Our model is based on a simplified and efficient variant of the SCAPE model (Anguelov *et al.*, 2005) (henceforth termed *simplified SCAPE space*), that was described by Jain et al. (Jain *et al.*, 2010), and was used for different applications in computer vision and graphics (Jain *et al.*, 2010; Helten *et al.*, 2013; Mündermann *et al.*, 2007, Chapters 3, 4 and 5), but was never learned from such a complete dataset. This compact shape space learns a probability distribution from a dataset of 3D human laser scans. It models variations due to changes in identity using a principal component analysis (PCA) space, and variations due to pose using a skeleton-based surface skinning approach. This representation makes the model versatile and computationally efficient compared to SCAPE.

Prior to statistical analysis, the human scans have to be processed and aligned to establish correspondence. We contribute by evaluating different variants of state-of-the-art techniques for non-rigid template fitting and posture normalization to process the raw data (Allen *et al.*, 2003; Hasler *et al.*, 2009; Wuhrer *et al.*, 2012; Neophytou and Hilton, 2013). Our findings are not entirely new methods, but best practices and specific solutions for automatic preprocessing of large scan databases for learning the simplified SCAPE model in the best way (Section 6.3). First, shape and posture fitting of an initial shape model to a raw scan prior to non-rigid deformation considerably improves the results. Second, multiple passes over the dataset improve initialization and thus increase the overall fitting accuracy and statistical model qualities. Third, posture normalization prior to shape space learning leads to much better generalization and specificity.

The main contribution of our work is a set of simplified SCAPE spaces learned from the largest database that is commercially available (Robinette *et al.*, 1999). The differences in our simplified SCAPE spaces stem from differences in the registration and pre-alignment of the human body scans. We evaluate different data processing techniques in Section 6.4 and the resulting shape spaces in Section 6.5. Finally, we compare our simplified SCAPE spaces to the state-of-the-art, which is a simplified SCAPE space learned from a publicly available database (Hasler *et al.*, 2009) for the application of reconstructing full 3D body models from monocular depth images in Section 6.6. Our experimental evaluation clearly demonstrates the advantage of our more expressive shape models in terms of shape space quality and for the task of reconstructing 3D human body shapes from monocular depth images (Section 6.6).

We release the newly built shape spaces with code to (1) pre-process raw scans and (2) fit a shape space to a raw scan for public usage. We believe this contribution is required for future development in human modeling.

6.2 STATISTICAL MODELING WITH SCAPE

We briefly recap the efficient simplified representation of the SCAPE model (Jain *et al.*, 2010) which we use and discuss its differences to the original SCAPE model (Anguelov *et al.*, 2005) in more detail. For learning the model, both methods assume that a template mesh **T** has been deformed to each raw scan in a database. All scans of the database are assumed to be rigidly aligned, e.g. by Procrustes Analysis (Goodall, 1991).

6.2.1 Original SCAPE model

In the original SCAPE model, the transformation of each triangle of **T** is modeled as combination of two linear transformations $\mathbf{R}_{m,i} \in SO(3)$ and $\mathbf{Q}_{m,i} \in \mathbb{R}^{3\times3}$. Index i indicates one particular scan **T** is fitted to and we refer to the fitting result after rigid alignment with **T** as instance mesh $\mathbf{M}_i$. While $\mathbf{R}_{m,i}$ represents the posture of the person as global rotation induced by the deformation of an underlying rigid skeleton, $\mathbf{Q}_{m,i}$ encodes the individual deformations of each triangle that originates from varying body shape or non-rigid posture dependent surface deformations such as muscle bulging. Computing $\mathbf{Q}_{m,i}$ for each vertex separately is an under-constrained problem. Therefore, smoothing is applied so that $\mathbf{Q}_{m,i}$ of neighboring vertices are dependent. Finally, by applying dimensionality reduction techniques to the transformations $\mathbf{R}_{m,i}$ and $\mathbf{Q}_{m,i}$, one obtains a flexible model that covers a wide range of possible surface deformations. However, as the model does not explicitly encode vertex position, one needs to solve a complex least squares problem to reconstruct the mesh surface.

6.2.2 Simplified SCAPE space

The aforementioned computational overhead is often prohibitive in applications where speed is more important than the overall reconstruction quality, or when many samples need to be drawn from the shape space. Our simplified SCAPE space (Jain *et al.*, 2010) reconstructs vertex positions in a given posture and shape without needing to solve a Poisson system. To learn the model, only laser scans in a standard posture χ_0 are used. A PCA model of the meshes $\mathbf{M}_i$ is learned, which represents each shape using a parameter vector φ. This shape space only covers variations in overall body shape and not posture. An articulated skeleton is fitted to the average human shape, and linear blend skinning weights to attach the surface to the bones are computed. The skeleton scales in accordance to the body shape by expressing joint locations relative to nearby surface vertex locations.

For reconstructing a model of shape φ in skeleton pose χ (joint angle parameters), the method first calculates a personalized mesh $\mathbf{M}_{\varphi,\chi_0}$ using φ. Then a linear blend skinning is applied to the personalized mesh to obtain the final mesh $\mathbf{M}_{\varphi,\chi}$ in pose χ. While such a simplified SCAPE approach shows much faster reconstruction speed, especially when the personalized mesh and skeleton can be precomputed, its reconstruction quality is inferior to the original SCAPE approach. In the rest of this chapter, we use this simple and efficient shape space.

6.3 DATA PROCESSING

This section describes how to pre-process a set of raw body scans to establish correspondence and pre-align the models. We show best-practice ways how non-rigid template fitting can be used to register raw scans, how to initialize the template fitting, how bootstrapping can help to improve the correspondence, and how the postures of registered scans can be normalized. Tools to reproduce these steps will be made publicly available.

6.3.1 Non-rigid template fitting

Our method to fit a human shape template $\mathbf{T}$ to a human scan $\mathbf{S}$ is inspired by Allen et al. (Allen *et al.*, 2003). In non-rigid template fitting (henceforth abbreviated *NRD*), each vertex $\mathbf{p}_i$ of $\mathbf{T}$ is transformed by a 4×4 affine matrix $\mathbf{A}_i$, which allows for twelve degrees of freedom during the transformation. The aim is to find a set of matrices $\mathbf{A}_i$ that define vertex positions in a deformed template matching well with $\mathbf{S}$. The fitting is done by minimizing a combination of data, smoothness and landmark errors.

Data term. The data term requires each vertex of the transformed template to be as close as possible to its corresponding vertex of $\mathbf{S}$, and takes the form

$$E_d = \sum_{i=1}^{N} w_i ||\mathbf{A}_i \mathbf{p}_i - NN_i(\mathbf{S})||_F^2, \tag{6.1}$$

where N is the number of vertices in $\mathbf{T}$, w_i weights the error contribution of each vertex, $||.||_F$ denotes the Frobenius norm, and NN_i is a closest compatible point in $\mathbf{S}$. If surface normals of closest points are less than 60° apart and the distance between the points is less than 20 mm, we set w_i to 1, otherwise to 0.

Smoothness term. Fitting using E_d only may lead to situations where neighboring vertices of $\mathbf{T}$ match to disparate vertices in $\mathbf{S}$. To enforce smooth surface deformations we use a smoothness term E_s that requires affine transformations applied to connected vertices to be similar, i.e.

$$E_s = \sum_{\{i,j | (\mathbf{p}_i, \mathbf{p}_j) \in edges(\mathbf{T})\}} ||\mathbf{A}_i - \mathbf{A}_j||_F^2. \tag{6.2}$$

Landmark term. Although using E_d and E_s would suffice to fit two surfaces that are close to each other, the optimization gets stuck in a local minimum when $\mathbf{T}$ and $\mathbf{S}$ are far apart. A remedy is to identify a set of points on $\mathbf{T}$ corresponding to known anthropometric landmarks on $\mathbf{S}$. In each CAESAR scan these are obtained by placing markers on each subject prior to scanning. Our landmark term penalizes misalignments between landmark locations

$$E_l = \sum_{i=1}^{M} ||\mathbf{A}_{k_i} \mathbf{p}_{k_i} - \mathbf{l}_i||_F^2, \tag{6.3}$$

where k_i is the landmark index on $\mathbf{T}$, and $\mathbf{l}_i$ is the landmark point on $\mathbf{S}$. Although there are only 64 landmarks compared to the total number of 6449 vertices, good landmark fitting is enough to get the deformed surface of $\mathbf{T}$ close to $\mathbf{S}$, and avoid local convergence.

Combined energy. The three terms are combined into a single objective

$$E = \alpha E_d + \beta E_s + \gamma E_l. \tag{6.4}$$

For optimization we use L-BFGS-B (Zhu *et al.*, 1997). We vary the weights α, β and γ according to the following empirically found schedule. We first perform a single iteration of optimization without data term by setting $\alpha = 0$, $\beta = 10^6$, $\gamma = 10^{-3}$, which allows to bring the surfaces into a rough correspondence. We then allow the data term to contribute by setting $\alpha = 1$, $\beta = 10^6$, $\gamma = 10^{-3}$. In addition, we relax smoothness and landmark weights after each iteration of fitting to $\beta := 0.25\beta$ and $\gamma := 0.25\gamma$, thus allowing the data term to dominate. This is repeated until $\beta \leq 10^3$. Reducing β increases the flexibility of deformation and allows $\mathbf{T}$ to better reproduce fine details, while reducing γ is necessary due to unreliable placement of landmarks in some scans.

6.3.2 Initialization

For non-rigid template fitting to succeed, **T** should be pre-aligned to **S**. We explore two initialization strategies.

A first standard way to initialize *NRD* is to use a static template with annotated landmarks. Corresponding landmarks are then used to rigidly align **S** to **T**.

A second way to initialize the fitting is to start with a simplified SCAPE space that was learned from a small registered dataset. Fitting the shape space to a scan is achieved in three steps. First, **S** is rigidly aligned to **T**. Second, the shape parameters of the simplified SCAPE space are fixed while the posture parameters are adjusted to minimize the landmark term given in Eq. 6.3. Third, the shape and posture parameters of the simplified SCAPE space are optimized iteratively to minimize the data term in Eq. 6.1. In each iteration, the posture is fitted in a first step and the shape is fitted in a second step. After each fitting step, the set $NN_i(\mathbf{S})$ is recomputed. This iterative procedure is repeated until E_d does not change significantly. For optimization, we use the iterative interior point method.

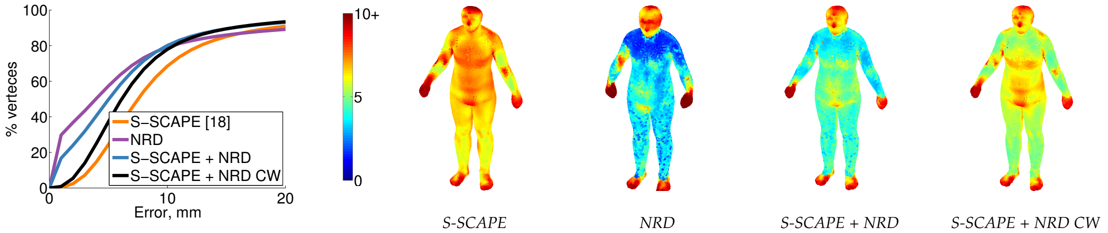
6.3.3 Bootstrapping

In many cases, even after *NRD*, **T** is far from **S**. Using registered scans with a high fitting error for shape space learning may lead to unrealistic shape deformations in the learned space. A remedy is to visually examine each fitting, discard fittings of low quality, and learn a simplified SCAPE space using the samples that passed the visual inspection. This simplified SCAPE space is then used as initialization to perform a fitting during the next pass. This bootstrapping process is performed until nearly all registered scans pass the visual inspection. Note that visual inspection is required, as low average fitting errors do not always correspond to good results, since the fitting of localized areas may be inaccurate.

6.3.4 Posture normalization

The simplified SCAPE space used in this study aims to represent shape and posture variations independently. However, by performing PCA over the vertex coordinates of processed scans captured in a standard posture, the shape space capturing variations caused by different identities is not normalized for posture. This may cause problems because the scans in standing posture present in the CAESAR and MPI Human Shape databases inevitably contain slight posture variations, mostly in the areas of the arms. To account for these variations, we compare the statistical shape space learned on the registered data directly to the one learned on data that was modified to remove posture variations. We consider two recent posture normalization approaches.

Wuhrer et al. (Wuhrer *et al.*, 2012) factor out variations caused by posture changes by performing PCA on localized Laplacian coordinates. While this approach leads to



(a) Total fitting error (b) Mean fitting error

Figure 6.1: Fitting error on the CAESAR dataset when using the *S-SCAPE* space (Jain *et al.*, 2010) alone, *NRD* alone, and initializing using *S-SCAPE* prior to *NRD* with different weighting schedules (*S-SCAPE + NRD*, *S-SCAPE + NRD CW*). Shown are (a) the proportion of vertices [%] with fitting error below a threshold and (b) the average fitting error per vertex.

better shape spaces than performing PCA on the vertex coordinates, it is difficult to compare this shape space to the simplified SCAPE space. We therefore modify each model $\mathbf{M}_i$ obtained by fitting $\mathbf{T}$ to $\mathbf{S}_i$ by initializing each shape to the mean shape $\overline{\mathbf{M}}$ and by optimizing the localized Laplacian coordinates to be as close as possible to the ones computed on $\mathbf{M}_i$. This leads to models that have the body shape of $\mathbf{M}_i$ in the posture of $\overline{\mathbf{M}}$.

Neophytou and Hilton (Neophytou and Hilton, 2013) normalize the posture of each processed scan using a skeleton model and Laplacian surface deformation. While this type of normalization may introduce artifacts around joints when the posture is changed significantly, this approach is suitable to normalize the posture of models of the CAESAR database as the posture variations are minor. We use this method to modify the posture of each $\mathbf{M}_i$.

6.4 EVALUATION OF TEMPLATE FITTING

We now evaluate the different components of our registration procedure on the CAESAR dataset (Robinette *et al.*, 1999). Each CAESAR scan contains 73 manually placed landmarks. We exclude several landmarks located on open hands, as those are missing for our template, resulting in 64 landmarks used for registration. Furthermore, we remove all laser scans without landmarks and corrupted scans, resulting in 4308 scans.

6.4.1 Implementation details

Non-rigid template fitting requires a human shape template as input, and the initialization procedure requires an initial shape space. We use registered scans of 111 individuals in neutral posture of the MPI Human Shape dataset to compute these initializations.

However, these data have artifacts in non-smooth areas at the head and neck. We smooth these areas by identifying problematic vertices and by iteratively recomputing their positions as an average position of direct neighbors. Furthermore, due to privacy reasons, head vertices of each human scan were replaced by the same dummy head, which is not representative and of low quality at the backside. We adjust the vertex compatibility criteria to compute nearest neighbors during *NRD* by allowing 30° deviation of the head face normals while increasing the distance threshold to 50 mm.

We employ the algorithm from Section 6.3.1 to compute correspondences for the CAESAR dataset. One inconsistency between the datasets is that the hands in the MPI Human Shape dataset are closed, while they are open in the CAESAR dataset. As remedy, we set α and γ to zero for hand vertices in Eq. 6.4, thus only allowing E_s to contribute. Prior to fitting, we sub-sample each CAESAR scan to have a total number of vertices that exceeds the number of vertices of $\mathbf{T}$ by a factor of three (6449 vertices in $\mathbf{T}$ vs. 19347 vertices in $\mathbf{S}$). This gives a good trade-off between fitting quality and computational efficiency.

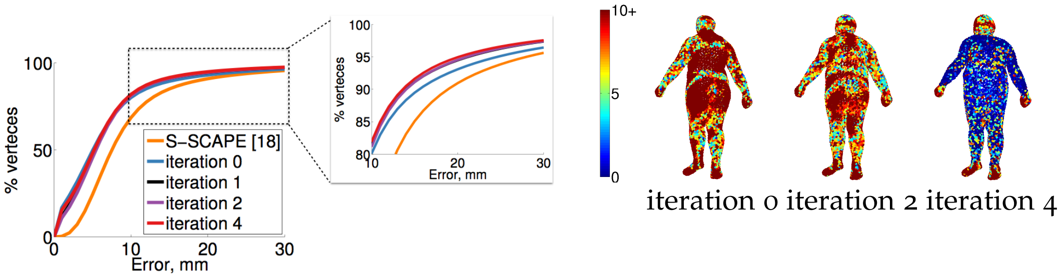
6.4.2 Quality measure

Measuring the accuracy of surface fitting is not straightforward, as no ground truth correspondence between $\mathbf{S}$ and $\mathbf{T}$ is available. We evaluate the fitting accuracy by finding the nearest neighbor in $\mathbf{S}$ for each fitted template vertex. If this neighbor not further from its correspondence in $\mathbf{T}$ than 50 mm and its face normals do not deviate more than 60°, the Euclidean vertex-to-vertex distance is computed. In our experiments we report both the proportion of vertices falling below a certain threshold and the distance per vertex averaged over all fitted templates. In the following, we first show the effects of various types of initialization and weighting schemes in the *NRD* procedure on the fitting error. Second, we show the effect of performing multiple bootstrapping rounds.

6.4.3 Initialization

First, we evaluate two different initialization strategies used in our fitting procedure. We compare the results when using an average human template (*NRD*) to the result when using the simplified SCAPE space learned on the MPI Human Shape dataset (*S-SCAPE + NRD*) for initialization. We compare the results by both non-rigid deformation schemes to the fitting accuracy when fitting the publicly available simplified SCAPE space by Jain et al. (Jain *et al.*, 2010) without any non-rigid deformation (*S-SCAPE*).

The results are shown in Fig. 6.1. The total fitting error in Fig. 6.1(a) shows that *NRD* achieves good fitting results in the low error range of 0 – 10 mm, as it can produce good template fits for the areas where $\mathbf{T}$ is close to $\mathbf{S}$. However, as *NRD* is a model-free method, the smooth topology of $\mathbf{T}$ may not be preserved during the deformation, e.g. convex surfaces of $\mathbf{T}$ may be deformed into non-convex



(a) Total fitting error (b) Mean fitting error on a sample scan

Figure 6.2: Fitting error after up to four bootstrapping rounds over the CEASAR database when using the model of Jain et al. (Jain *et al.*, 2010) as initialization for iteration 0. Shown is the proportion of vertices [%] with fitting error below a threshold.

surfaces after *NRD*. This leads to large fitting errors for areas of **T** that are far from **S**. *S-SCAPE* + *NRD* uses a shape space fitting prior to NRD, which allows for a better initial alignment of **T** to **S**. Note that *S-SCAPE* + *NRD* results in a better fitting accuracy in the high error range of $10 - 20$ mm. The fitting result by *S-SCAPE* + *NRD* favorably compares against using *S-SCAPE* alone. Although *S-SCAPE* results in deformations preserving the human body shape topology, the shape space is learned from the relatively specialized MPI Human Shape dataset containing mostly young adults and thus cannot represent all shape variations.

We also analyze the differences in the mean fitting errors per vertex in Fig. 6.1(b). *NRD* achieves good fitting results for most of the vertices. However, the arms are not fitted well due to differences in body posture of **T** and **S**. Furthermore, the average fitting error is not smooth, which shows that despite using E_s, *NRD* may produce non-smooth deformations. In contrast, the result of *S-SCAPE* + *NRD* is smoother and has a lower fitting error for the arms. Clearly, the average fitting error of *S-SCAPE* is much higher, with notably worse fitting results for arms, belly and chest.

6.4.4 *NRD* parameters

Second, we evaluate the influence of the weight relaxation during *NRD* on the fitting accuracy. Specifically, we compare the standard weighting scheme where weights are relaxed in each iteration (*S-SCAPE* + *NRD*) to the case where the weights stay constant (*S-SCAPE* + *NRD CW*). Fig. 6.1(a) shows that the total fitting error of *S-SCAPE* + *NRD* is lower than *S-SCAPE* + *NRD CW*. This is because *S-SCAPE* + *NRD CW* enforces higher localized rigidity by keeping weights constantly high, while *S-SCAPE* + *NRD* relaxes the weights so that **T** can fit more accurately to **S**. This explanation is supported by consistently higher per-vertex mean fitting errors in case of *S-SCAPE* + *NRD CW* compared to *S-SCAPE* + *NRD*, as shown in Fig. 6.1(b). The highest differences are in the areas of high body shape variability, such as belly and chest. Different weight reduction schemes such as $\beta := 0.5\beta$, $\gamma := 0.5\gamma$ and $\beta := 0.25\beta$, $\gamma := 0.25\gamma$ lead to better fitting accuracy compared to constant weights, with the latter scheme achieving slightly better results and faster convergence rates. We thus use the proposed weight reduction scheme in the following.

6.4.5 Bootstrapping

Third, we evaluate the fitting accuracy before and after performing multiple rounds of bootstrapping. To that end, we use the output of *S-SCAPE + NRD* (iteration 0) to learn a new statistical shape space, which is in turn used to initialize *NRD* during the second pass over the data (iteration 1). This process is repeated for five passes. The number of registered scans that survived the visual inspection after each round is 1771, 3253, 3641, 4237 and 4301, respectively. This results show that bootstrapping allows to register and thus to learn from an increasing number of scans. Fitting results are shown in Fig. 6.2. The close-up shows that although the overall fitting accuracy before and after bootstrapping is similar, bootstrapping allows to slightly improve the fitting accuracy in the range of $10 - 30\ mm$. Fitting results after three passes over the dataset (iteration 2) are slightly better compared to the initial fitting (iteration 0), and the accuracy is further increased after five passes (iteration 4). Fig. 6.2 (b) shows sample fitting results before and after several bootstrapping rounds. Largest improvements are achieved for the belly and chest; these are areas with large variability. The fitting improves with an increasing number of bootstrapping rounds. We use the fitting results after five passes (iteration 4) to learn the simplified SCAPE space used in the following.

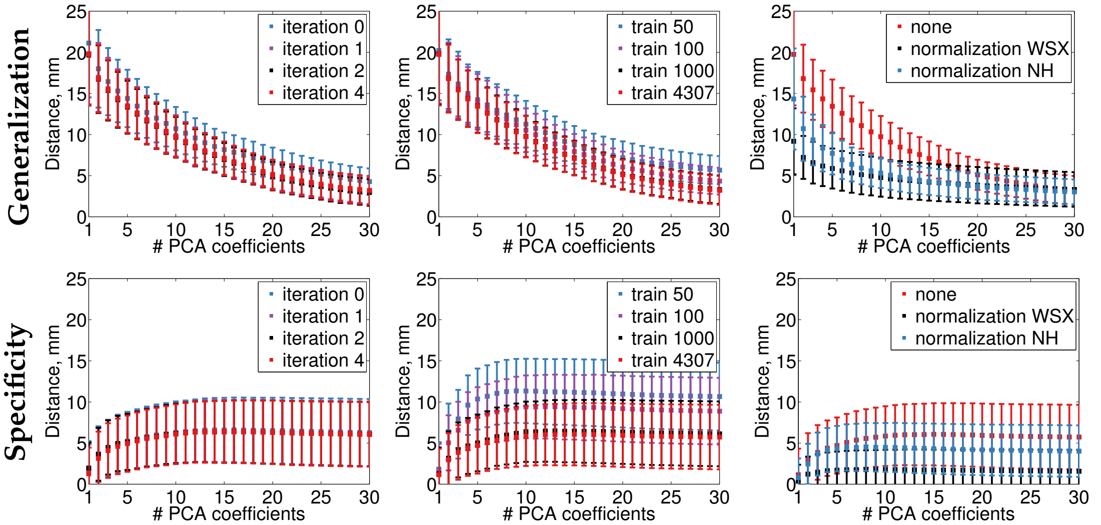
6.5 EVALUATION OF STATISTICAL SHAPE SPACE

In this section, we evaluate the simplified SCAPE space using the statistical quality measures generalization and specificity (Styner *et al.*, 2003).

6.5.1 Quality measure

We use two complementary measures of shape statistics. Generalization evaluates the ability of a shape space to represent unseen instances of the object class. Good generalization means the shape space is capable of learning the characteristics of an object class from a limited number of training samples, poor generalization indicates overfitting of the training set. Generalization is measured using leave-one-out cross reconstruction of training samples, i.e. the shape space is learned using all but one training samples and the resulting shape space is fitted to the excluded sample. The fitting error is measured using the mean vertex-to-vertex Euclidean distance. Generalization is reported as mean fitting error averaged over the complete set of trials, and plotted as a function of the number of shape space parameters. It is expected that the mean error decreases until convergence as the number of shape space parameters increases.

Specificity measures the ability of a shape space to generate instances of the object



(a) bootstrapping (b) # training samples (c) posture normalization

Figure 6.3: Influence of different design choices on statistical quality measures. Shown are influence of (a) bootstrapping, (b) number of training samples and (c) posture normalization on generalization (top row) and specificity (bottom row).

class that are similar to the training samples. The specificity test is performed by generating a set of instances randomly drawn from the learned shape space and by comparing them to the training samples. The error is measured as average distance of the generated instances to their nearest neighbors in the training set. It is expected that the mean distance increases until convergence with increasing number of shape space parameters. We follow Styner et al. (Styner *et al.*, 2003) and generate 10 000 random samples.

6.5.2 Bootstrapping

We evaluate the influence of the bootstrapping on the quality of the statistical shape space by comparing models obtained after zero, one, two and four iterations of bootstrapping. The geometry of the training samples changes in each bootstrapping round, which makes the generalization and specificity results incomparable across different shape spaces. We thus use the training samples obtained after four iterations of bootstrapping as "ground truth", i.e., the reconstruction error of generalization and the nearest neighbor distance of specificity for each shape space is computed w.r.t. fitting results after four bootstrapping rounds. This allows for a fair comparison across different statistical shape spaces.

The results are shown in Fig. 6.3(a). Generalization is already reduced after a single iteration of bootstrapping because after one iteration, the shape space is learned from a significantly larger number of training samples, thereby using samples with higher shape variation that were discarded in the 0^{th} iteration. The following rounds of bootstrapping have little influence on generalization and specificity, with the shape space after four iterations resulting in a slightly lower specificity error than for previous iterations for a small number of shape parameters.

6.5.3 Number of training samples

To evaluate the influence of the number of training samples, we vary the number of samples obtained after four bootstrapping iterations. Specifically, we consider subsets of 50, 100, 1000 and 4307 ($all - 1$) training samples. To compute a shape space, the desired number of training shapes are sampled from all training samples according to a learned PCA space. For generalization, we cross-evaluate on all 4308 training samples by leaving one sample out and by sampling the desired number of training shapes from the remaining samples. For specificity, we compute the nearest-neighbor distances to all 4308 training samples to find the closest sample.

The results are shown in Fig. 6.3(b). The shape space learned from the smallest number of samples performs worst. Increasing the number of samples consistently improves the performance with the best results achieved when using the maximum number. Both generalization and specificity error reduction is most pronounced when increasing the number of samples from 50 to 100. Further increasing the number of samples to 1000 affects specificity much stronger than generalization. This shows that the shape space learned from only 100 samples generalizes well, while its generative qualities are poor. Increasing the number of samples from 1000 to 4307 only slightly reduces both generalization and specificity errors, which shows that a high-quality statistical shape space can be learned from 1000 samples.

6.5.4 Posture normalization

Finally, we evaluate the generalization and specificity of the shape space obtained when performing posture normalization using the methods of Wuhrer et al. (Wuhrer *et al.*, 2012) (*WSX*) and Neophytou and Hilton (Neophytou and Hilton, 2013) (*NH*). The results are shown in Fig. 6.3 (c). Posture normalization significantly improves generalization and specificity, with *WSX* achieving the best result. The reduction of the average fitting error in case of generalization is highest for a low number of shape parameters. This is because both *WSX* and *NH* lead to shape spaces that are more compact compared to the shape space obtained with non-normalized training shapes. Additionally, both posture-normalized shape spaces exhibit much better specificity. Compared to the shape space trained before posture normalization, randomly generated samples from the both shape spaces trained after *WSX* and *NH* exhibit less variation in posture and are thus more similar to their corresponding posture-normalized training samples.

Finally, we qualitatively examine the first five PCA components learned by the following simplified SCAPE spaces: the current state-of-the-art shape space *S-SCAPE* (Jain *et al.*, 2010), our shape space without posture normalization and with posture normalization using *WSX* and *NH*. The results are shown in Fig. 6.4. Many of the major modes of shape variation by *S-SCAPE* (row 1) are affected by global and local posture-related deformations, such as moving of arms or tilting the body. In contrast, the principal components of variation by our shape space (row 2) are mostly due to shape changes thanks to a better template fitting procedure and a more representative training set. However, little posture variations are still part of the learned shape space. Performing posture normalization of the training samples prior to learning the shape space completely factors out changes due to posture, as can be seen in the shape spaces learned using *WSX* (row 3) and *NH* (row 4).

6.6 HUMAN BODY RECONSTRUCTION

Finally, we evaluate our improved simplified SCAPE spaces in the specific application of estimating human body shape from sparse visual input. We follow the approach by Helten et al. (Helten *et al.*, 2013) to estimate the body shape of a person from two sequentially taken front and back depth images. First, body shape and posture are fitted independently to each depth image. Second, the obtained results are used as initialization of a method that jointly optimizes over shape and independently optimizes over posture parameters. This optimization strategy is used because the shape in both depth scans is of the same person, but the pose may differ.

6.6.1 Dataset and experimental setup

We use a publicly available dataset (Helten *et al.*, 2013) containing Kinect body scans of three males and three females. Examples of the Kinect scans are shown in Fig. 6.6(a). For each subject, a high-resolution laser scan was captured, which is used to determine "ground truth" body shape by fitting a simplified SCAPE space to the data. We follow the evaluation protocol of Helten et al. (Helten *et al.*, 2013), which computes the fitting error as a difference between the results of fitting a simplified SCAPE space to the depth data and the "ground truth" computed as a vertex-to-vertex Euclidean distance. As the required landmarks are not available for this dataset, we manually placed 14 landmarks on each depth and laser scan.

6.6.2 Quantitative evaluation

For quantitative evaluation, we compare the following four shape models presented above: the current state-of-the-art shape space (Jain *et al.*, 2010), our shape space without posture normalization and with posture normalization using *WSX* and *NH*. In our experiments, we vary the number of shape space parameters and the number of training samples the simplified SCAPE models are learned from. To evaluate the fitting accuracy, we report the proportion of vertices below a certain threshold.

The results are shown in Fig. 6.5, where the number of shape space parameters varies in the columns and the number of training samples varies in the rows. In all cases our simplified SCAPE spaces learned from the CAESAR dataset significantly

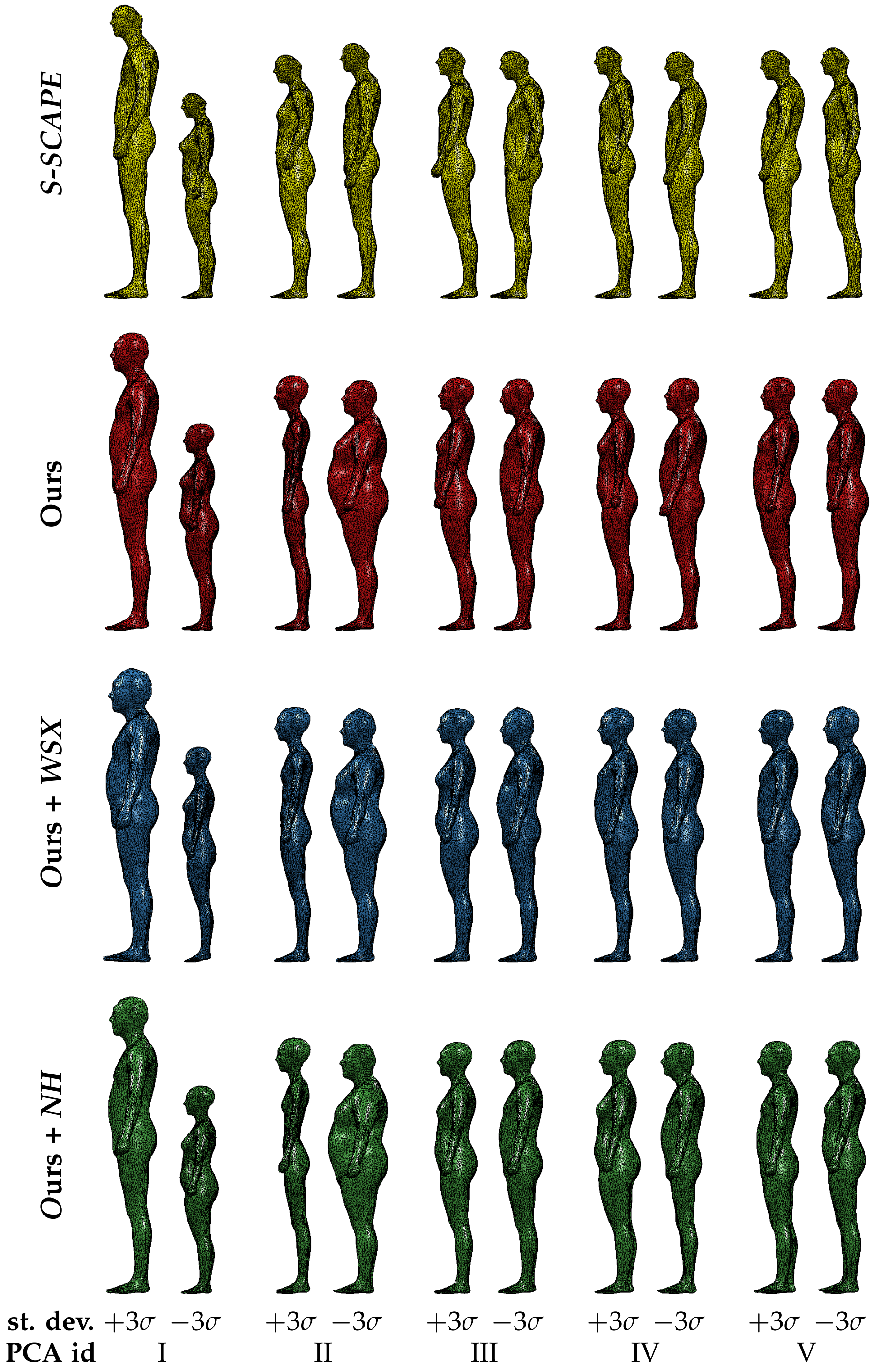

S-SCAPE
Ours
Ours + WSX
Ours + NH
st. dev. $+3\sigma$ -3σ $+3\sigma$ -3σ $+3\sigma$ -3σ $+3\sigma$ -3σ $+3\sigma$ -3σ
PCA id I II III IV V

Figure 6.4: Visualization of the first five PCA eigenvectors scaled by $\pm 3\sigma$ (standard deviation). Shown are eigenvectors of the simplified SCAPE space (Jain *et al.*, 2010) (row 1) and the simplified SCAPE spaces trained using our processed data without (row 2) and with posture normalization using *WSX* (Wuhrer *et al.*, 2012) (row 3) and *NH* (Neophytou and Hilton, 2013) (row 4).

outperform the shape space by Jain et al., which is learned from the far less representative MPI Human Shape dataset. Our models achieve good fitting accuracy when using as few as 20 shape parameters, and the performance stays stable when increasing the number of shape parameters up to 50 (first row). In contrast, the performance of the shape space by Jain et al. drops, possibly due to overfitting to unrealistic shape deformations in noisy depth data. Interestingly, better performance by our models is evident even in the case when all models are learned from the same number of training samples (third and fourth rows). This shows that the CAESAR data has higher shape variability than the MPI Human Shape data. In the majority of cases, the shape space learned from the posture-normalized samples with *NH* outperforms the shape space learned from samples without posture normalization. This shows that the posture normalization method of Neophytou and Hilton (Neophytou and Hilton, 2013) helps to improve the accuracy of fitting to noisy depth data. Surprisingly, the shape space learned from samples without posture normalization outperforms the shape space learned from the posture-normalized samples with *WSX* in most cases. Overall, the quantitative results show the advantages of our approach of building simplified SCAPE spaces learned from a large representative set of training samples with additional posture normalization.

6.6.3 Qualitative evaluation

To qualitatively evaluate the fitting, we visualize the per-vertex fitting errors. We consider the simplified SCAPE spaces learned from all available training samples and use 20 shape space parameters. For visualization we choose two subjects, male and female, where the differences among the shape spaces are most pronounced.

Results are shown in Fig. 6.6. Our shape spaces better fit the data, in particular in the areas of belly and chest. This is to be expected, as we learn from the larger and more representative CAESAR dataset. Both shape spaces trained from posture normalized models can better fit the arms compared to non-normalized models.

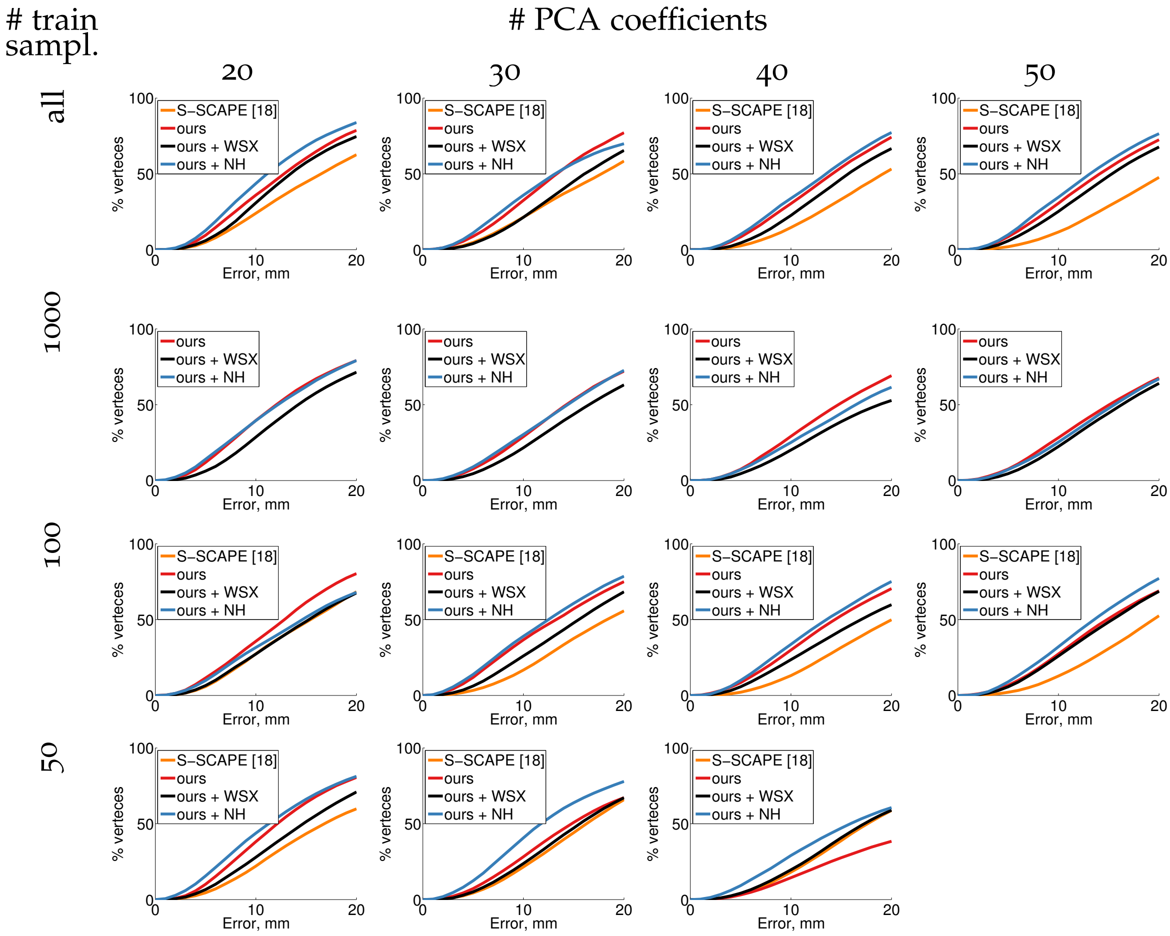


Figure 6.5: Fitting error on dataset of depth scans (Helten *et al.*, 2013) of simplified SCAPE spaces by Jain et al. (Jain *et al.*, 2010) and simplified SCAPE spaces trained using our processed data without posture normalization and with posture normalization using WSX and NH. Shown is the proportion of vertices [%], for which the fitting error falls below a threshold.

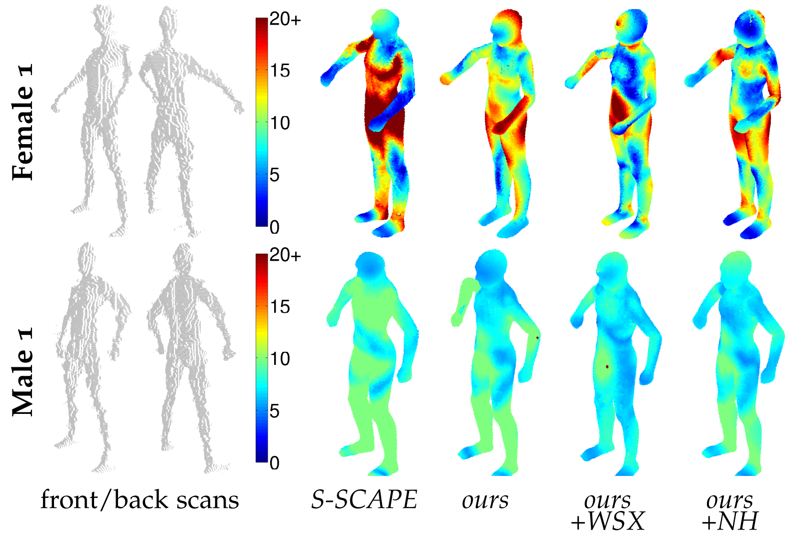


Figure 6.6: Per-vertex shape fitting error (mm) of multiple methods on sample individuals of Helten et al. (Helten *et al.*, 2013).

6.7 CONCLUSION

In this chapter we address the challenging problem of building an efficient and expressive 3D body shape space from the largest commercially available 3D body scan dataset (Robinette *et al.*, 1999). We carefully design and evaluate different data preprocessing steps required to obtain high-quality body shape models. To that end, we evaluate different template fitting procedures. We observe that shape and posture fitting of an initial shape space to a scan prior to non-rigid deformation considerably improves the fitting results. Our findings indicate that multiple passes over the dataset improve initialization and thus increase the overall fitting accuracy and statistical shape space qualities. Furthermore, we show that posture normalization prior to learning a shape space leads to significantly better generalization and specificity of the simplified SCAPE spaces. Finally, we demonstrate the advantages of our learned shape spaces over the state-of-the-art shape space of Jain et al. (Jain *et al.*, 2010) learned on largest publicly available dataset (Hasler *et al.*, 2009) for the task of human body tracking from monocular depth images.

We release the simplified SCAPE spaces, raw scan preprocessing code, code to fit a simplified SCAPE space to a scan and evaluation code for public usage[7]. We believe that this contribution is required for future development in human body modeling.

---

[7] Available at humanshape.mpi-inf.mpg.de.

POSELET CONDITIONED PICTORIAL STRUCTURES 7

Contents

7.1 Introduction **119**
7.2 Review of pictorial structures **122**
7.3 Poselet Conditioned Pictorial Structures **123**
  7.3.1 Poselet representation 124
  7.3.2 Poselet dependent unary terms 124
  7.3.3 Poselet dependent pairwise terms 125
7.4 Results **125**
  7.4.1 Results on LSP dataset 126
  7.4.2 Results on Image Parse dataset. 129
  7.4.3 Results on UIUC People dataset. 131
7.5 Conclusion **131**

In this chapter we consider the challenging problem of articulated human pose estimation in still images. We observe that despite high variability of the body articulations, human motions and activities often simultaneously constrain the positions of multiple body parts. Modeling such higher order part dependencies seemingly comes at a cost of more expensive inference, which resulted in their limited use in state-of-the-art methods. In this paper we propose a model that incorporates higher order part dependencies while remaining efficient. We achieve this by defining a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structures model once the image observations are available. In order to derive a set of conditioning variables we rely on the poselet-based features that have been shown to be effective for people detection but have so far found limited application for articulated human pose estimation. We demonstrate the effectiveness of our approach on three publicly available pose estimation benchmarks.

7.1 INTRODUCTION

In this chapter we consider the challenging task of articulated human pose estimation in monocular images. Prominent approaches in this area (Andriluka *et al.*, 2011; Johnson and Everingham, 2011; Yang and Ramanan, 2011; Sapp and Taskar, 2013; Tompson *et al.*, 2014; Chen and Yuille, 2014) are based on the pictorial structures model (PS) and are composed of unary terms modeling body part appearance and pairwise terms between *adjacent* body parts and/or joints capturing their preferred spatial arrangement. While this approach leads to tree-based models and thus efficient and exact inference, it fails to capture important dependencies between *non-adjacent* body parts. That modeling such dependencies is important for effective pose estimation can be seen e.g. in Fig. 7.1: activities of people like playing soccer, tennis or volleyball results in strong dependencies between many if not all body parts; this can not be modeled with the above approach.

This well known problem has so far been addressed in two ways. The first simply uses a mixture of tree models thus learning separate pairwise terms for different global body configurations e.g. (Johnson and Everingham, 2010, 2011; Dantone *et al.*, 2014). The second approach is to add more pairwise terms including non-adjacent body parts leading to a loopy part graph that requires approximate inference (Andriluka *et al.*, 2011; Tran and Forsyth, 2010; Sun and Savarese, 2011; Wang *et al.*, 2011). A key challenge in designing models for pose estimation is thus to encode the higher-order part dependencies while still allowing efficient inference. In this chapter we propose a novel model that incorporates higher order information between body parts by defining a conditional model in which all parts are a-priori connected, but which becomes a tractable PS model once the mid-level features are observed. This allows to effectively model dependencies between non-adjacent parts while still allowing for exact and efficient inference in a tree-based model.

Clearly, the choice of the particular mid-level image representation used for conditioning our model is crucial for good performance of the overall approach. On the one hand, this representation has to be robust with respect to variations in people appearance, pose and imaging conditions. On the other hand, it has to be highly informative for the underlying human pose. In order to satisfy these requirements we rely on the non-parametric poselet representation introduced in (Bourdev *et al.*, 2010). Note that for the task of people detection until recently the most popular approaches were those which rely on a representation that jointly models appearance of multiple body parts (Bourdev *et al.*, 2010; Felzenszwalb *et al.*, 2010). Yet these models have not been shown to lead to state-of-the-art performance in human pose estimation, likely because they rely on a pose representation that is not fine-grained enough to enable localization of all body joints.

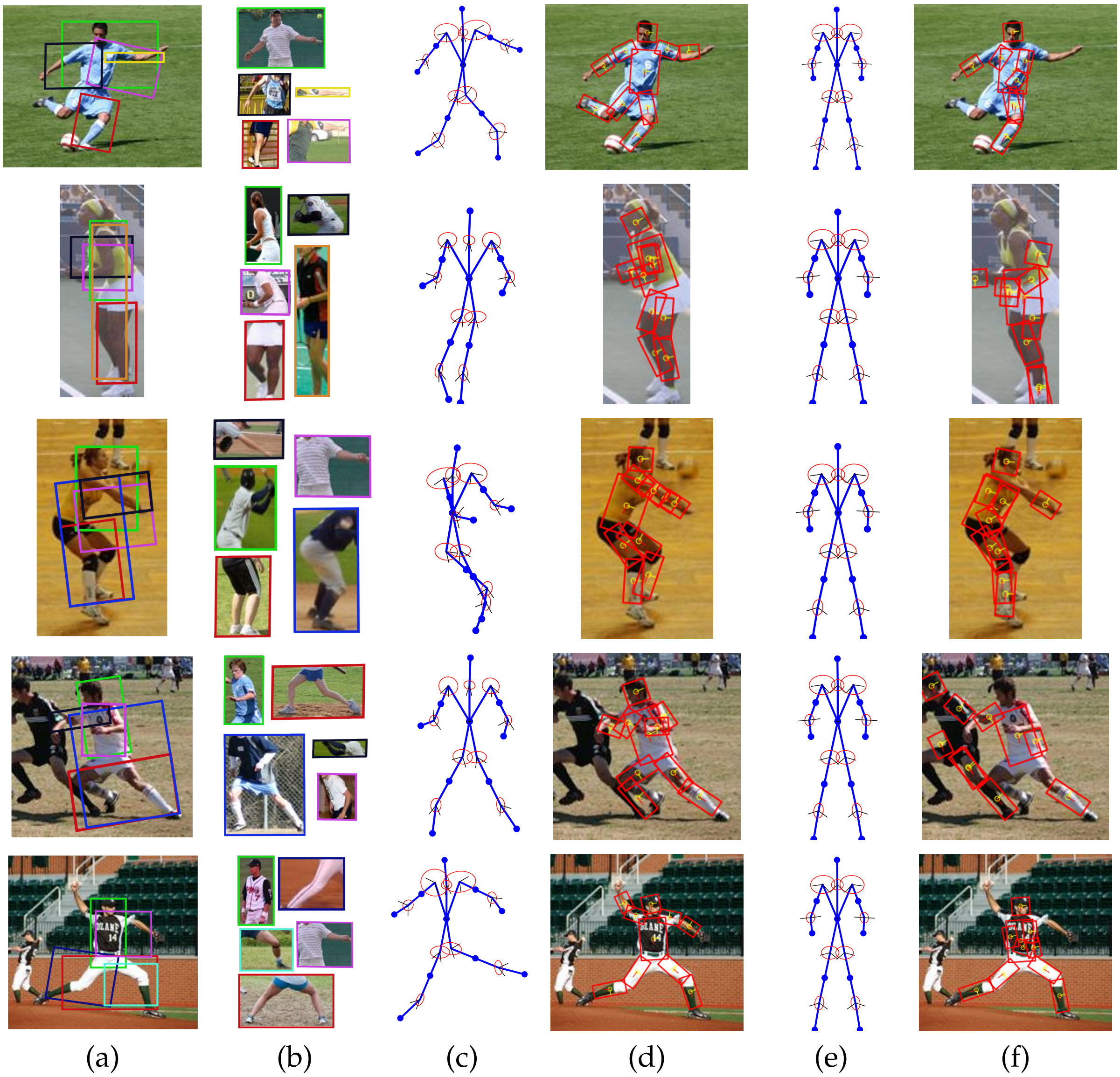


Figure 7.1: Visualization of our approach. (a) shows the top scoring poselet detections with the corresponding poselet cluster medoids (b). It is visible that the poselets capture the anatomical configuration of the human in the input image. All poselet detections contribute to a prediction of the deformable pairwise terms, the outcome of which is shown in (c). Using the PS model with these pair-wise terms achieves the detection outcome (d). In contrast we show the generic prior (Andriluka *et al.*, 2009) (e) and the corresponding pose prediction (f).

7.2 REVIEW OF PICTORIAL STRUCTURES

In this section we introduce the Pictorial Structures (PS) version (Andriluka *et al.*, 2011, 2009) that we are building on and that will serve as a baseline in the experiments. This implementation has been found to be competitive across a range of datasets. Although we focus on this particular incarnation of the PS model, we believe the extensions are applicable to other models, such as the one from (Yang and Ramanan, 2011). The extension of this model will then be the topic of the next section.

We phrase the PS model as a conditional random field (CRF), modelling the conditional probability of a body pose configuration given image evidence. We denote by $L = (l_1, \ldots, l_M)$ a full body pose, consisting of M parts. A part $l_m = (x_m, y_m, \theta_m, s_m)^\top$ is parameterized by its x, y center position, rotation $\theta \in [0, 360)$, and scale $s \in \mathbb{R}_+$. With D we denote any form of image evidence and with β the vector of model parameters. For convenience we distinguish between parameters for unary β^u and pairwise β^p factors. The PS model then takes the form

$$E(L; D, \beta) = \sum_{m=1}^{M} E^u(l_m; D, \beta^u) + \sum_{n \sim m} E^p(l_n, l_m; \beta^p). \tag{7.1}$$

With $n \sim m$ we denote the neighborhood relationship between the body parts. This typically is restricted to form a tree in order to enable exact and efficient inference.

Unary potentials We use the following unary potential functions

$$E^u(l_m; D, \beta^u) = \log \phi^u(l_m; D), \quad \forall m = 1, \ldots, M, \tag{7.2}$$

with pre-trained AdaBoost classifiers as the feature functions

$$\phi^u(l_m; D) = \max\left(\frac{\sum_t \alpha_i^t h_t(l_m, D)}{\sum_t \alpha_i^t}, \epsilon_0 \right). \tag{7.3}$$

A decision stump h_t in Eq.(7.3) is of the following form

$$h_t(l_m, D) = \operatorname{sign}(\xi_t(\mathbf{x}_{n(t)} - \varphi_t)), \tag{7.4}$$

where $\mathbf{x}$ is a feature vector, $\varphi_t \in \mathbb{R}$ a threshold, $\xi_t \in \{-1, 1\}$, and $n(t)$ is a feature index. The feature vector is obtained by concatenating the shape context descriptors computed on a regular grid inside the part bounding box. We refer the reader to (Andriluka *et al.*, 2011, 2009) for details on training and descriptors.

Pairwise potentials Pairwise potential functions take the form

$$E^p(l_n, l_m; \beta^p) = \left\langle \beta^p_{n,m}, \phi^p_{n,m}(l_n, l_m) \right\rangle, \quad \forall n \sim m. \tag{7.5}$$

The features for the potential $\phi^p_{n,m}$ acting on n and m are computed as follows. First both parts are transformed into a common reference space, that is the location of the joint between these parts. We use the transformation

$$T_{mn}(l_n) = \begin{pmatrix} x_n + s_n\mu_x^{mn}\cos\theta_n - s_n\mu_y^{mn}\sin\theta_n \\ y_n + s_n\mu_x^{mn}\cos\theta_n - s_n\mu_y^{mn}\sin\theta_n \\ \theta_n + \tilde{\theta}_{mn} \\ s_n \end{pmatrix}, \tag{7.6}$$

where $\mu^{mn} = (\mu_x^{mn}, \mu_y^{mn})^T$ is the mean relative position of the joint between parts m and n in the coordinate system of part n; $\tilde{\theta}_{mn}$ is the relative angle between parts. The pairwise term is then a Gaussian on the difference vector between the two transformations $T_{mn}(l_n) - T_{nm}(l_m)$, as is standard practice in all PS works (Andriluka *et al.*, 2011, 2009; Yang and Ramanan, 2011; Felzenszwalb and Huttenlocher, 2005). We derive a linear form for the pairwise term in Eq. 7.5 using the natural parameterization of the Gaussian as in (Felzenszwalb *et al.*, 2010; Yang and Ramanan, 2011), and place positivity constraints on those parameters in β^p that correspond to variances.

We learn unary and pairwise terms in a piecewise strategy, unary potentials using AdaBoost and the pairwise terms using a Maximum-Likelihood estimate.

7.3 POSELET CONDITIONED PICTORIAL STRUCTURES

Our approach is based on the following idea: we use a mid-level representation that captures possible anatomical configurations of a human pose to predict an image-specific pictorial structures (PS) model that in turn is applied to the image. The representation we are using is inspired by the work (Bourdev and Malik, 2009; Wang *et al.*, 2011) which is why we refer to it as *poselets*. Poselets go beyond standard pairwise part-part configurations and capture the configuration of multiple body parts jointly. As we still predict a tree connected PS model we retain efficient and tractable inference.

The idea of our model is visualized in Fig. 7.1. On the input images we compute poselet responses that capture different portions of the person's body configuration. Highest scoring poselet detections are shown in Fig. 7.1(a), together with representative examples for them in Fig. 7.1(b). This information is then used to augment both unary and pairwise terms of the PS model. In Fig. 7.1(c) we show the deformation terms of the resulting PS model that we are able to predict. Pose of the person estimated with our poselet-conditioned model is shown in Fig. 7.1(d). For comparison we show the deformation model of (Andriluka *et al.*, 2009) (a generic pose prior being the same for all images) along with the corresponding pose estimate in the last two columns.

The idea of having multiple deformation models is similar to the idea of encoding body pose configurations through different mixture components as in (Yang and Ramanan, 2011). However, in their work the pairwise mixture components are – in contrast to our model – not dependent on the image but estimated during inference. We experimentally compare to this approach.

This section first describes the feature representation used to capture human poses. We then present the extension of the standard PS model outlined in the previous section and show how both unaries (sec. 7.3.2) and pairwise terms (sec. 7.3.3) can be enhanced using poselet information.

7.3.1 Poselet representation

The goal of the mid-level representation is to capture common dependencies of multiple body parts. We implemented the following strategy to train a set of poselet detectors and compute a feature based on their responses.

For a reference body part, we cluster the relative positions of a subset of related body parts. For example, when picking the 'neck' part we cluster relative offsets of all upper body parts using Euclidean distance and K-means. We prune clusters that have less than 10 examples and use the remaining ones as poselets. In this chapter we run this process multiple times, picking different reference points and multiple subsets of related parts to obtain a total of P clusters. Together with every poselet p we store its mean offset from the torso annotation μ_p.

The next step is to learn a detector for each poselet. Following (Andriluka *et al.*, 2011, 2009), we train AdaBoost detectors on dense shape context features. A separate detector is trained for every poselet cluster using all training images that fall within this cluster. Example outcomes can be seen in Figure 7.1(a+b) showing the highest scoring poselets for some sample images and their medoids.

To form a feature vector $f \in \mathbb{R}^P$ we first predict the torso position μ_{torso} in the test image. Given a torso prediction and the relative offset μ_p of the poselet p, we compute the maximum poselet response in a small region[8] around $\mu_{torso} + \mu_p$. This corresponds to a max-pooling step in a local region for every poselet p. Then we aggregate the maximum scores for all $p = 1, \ldots, P$ poselets to form a feature vector $f \in \mathbb{R}^P$. Similar to (Wang *et al.*, 2011), we define 11 body part configurations, namely full body, upper body with arms, torso and head, right arm and torso, left arm and torso, right arm alone, left arm alone, torso with legs, legs, right leg alone, and left leg alone. For each of these configurations we cluster the data as described above and learn poselet detectors. During test time we additionally run each detector for +/-7.5 degrees to compensate for slight rotations. Torso prediction is done using the detector from Chapter 5 that we augment with a spatial prior learned on the training set.

Next we present two different ways how the features f can be used to obtain image conditioned PS models.

7.3.2 Poselet dependent unary terms

We first use the poselet features to obtain a location and rotation prediction for each body part separately.

---

[8] The size of the region is set to 20×20 pixels in our experiments.

Let us describe the location preference for a single part m only. During training, for part m, we cluster the relative distance between the torso and the part into $k = 1, \dots, K$ clusters. For each cluster k we compute its mean offset from the torso μ_k and the variance of the differences Σ_k. This now forms a classification problem, from the poselet response f into the set of K clusters. To this end we train a classifier using sparse linear discriminant analysis (SLDA) (Clemmensen *et al.*, 2011) on the training set. We chose a sparse method since we expect a different set of poselets to be predictive for different body parts.

During test time we apply the learned classifier to predict from f the mean μ_k, and variance Σ_k that are subsequently used as a Gaussian unary potential for the part. We proceed analogously for rotation, that is we learn a classifier that predicts the absolute rotation of the body part based on poselet responses. Both unary parts together form a Gaussian potential $E^{u,poselet}$, and the complete set of unary terms of our model then reads

$$E^u(l_m; D) = E^{u,boost}(l_m; D) + w_p E^{u,poselet}(l_m; D), \tag{7.7}$$

where $E^{u,boost}$ is the original term given by Eq. 7.2 and w_p is the weighting parameter estimated on the validation set.

7.3.3 Poselet dependent pairwise terms

To extend the pairwise terms we make them image dependent. For each pair of parts l_n, l_m we cluster their relative rotations into K clusters and obtain the parameters $\beta^{p,k}$ independently for each cluster using a maximum likelihood estimate. Similar to unary terms, we learn a SLDA classifier that predicts, given the feature f, into the set of clusters. This in turn yields the parameters β^p to be used for the image in question. The new pairwise potential that replaces E^p from Eq. 7.5 reads

$$E^{p,poselet}(l_n, l_m; D) = \langle \beta^p_{n,m}(f; D), \phi^p_{n,m}(l_n, l_m) \rangle. \tag{7.8}$$

We wrote $\beta(f)$ to make explicit its dependency on the poselet responses and that this parameter is being predicted.

7.4 RESULTS

In this section we evaluate the proposed poselet-conditioned PS model on three well-known pose estimation benchmarks. We demonstrate that our new model achieves a significant improvement compared to the original PS model, while performing on par or even outperforming other competing approaches.

Datasets. For evaluation we use the following publicly available pose estimation benchmarks exhibiting strong variations in articulation and viewpoint: the recently proposed "Leeds Sports Poses" (LSP) dataset (Johnson and Everingham, 2010) that includes 1000 images for training and 1000 for testing showing people involved in various sports; the "Image Parsing" (IP) (Ramanan, 2006) dataset consisting of 100 train images and 205 test images of fully visible people performing various activities such as sports, dancing and acrobatics; the "UIUC People" dataset (Tran and Forsyth, 2010) consisting of 346 training and 247 test images of people in highly variable body poses playing different sports such as Frisbee or badminton. For each dataset we increase the training set size by adding the mirrored versions of the training images.

7.4.1 Results on LSP dataset

As in (Johnson and Everingham, 2010) we allocate 500 training images for the validation set and use it to estimate the weighting parameter in Eq. 7.7 and the number K of unary and pairwise clusters via grid search. The estimated values are $w_p = 0.05$ and $K = 12$. The poselets are trained as described in Section 7.3.1, which results in $P = 1036$ poselets. We follow (Eichner and Ferrari, 2012a) and use the observer-centric annotations provided by the authors of (Eichner and Ferrari, 2012a), which allows us to directly compare to their work. In the following we evaluate different model components and compare our approach to the competitors.

Using an oracle to select components. First we show the performance of our model assuming that the correct component for every potential is chosen by an oracle. This is the best case scenario that provides an upper bound on the performance our proposed model can achieve. We experimented with the number of components and found that 12 components per potential perform best. Increasing the number of components did not lead to improved results because of the limited number of training images available for parameter estimation for each component.

Results are shown in Tab. 7.1. It can be seen that adding poselet dependent terms improves the performance w.r.t. the baseline PS model (Andriluka *et al.*, 2009). Large improvements are consistently observed for all body parts. Correct predictions of unary rotation components improve the localization of lower arms and legs most. This is explained by the fact that the rotation of these body parts is far less constrained compared to the rest of the limbs. Constraining part rotations to small ranges around the correct rotations reduces the uncertainty and steers the pose estimation towards the correct body pose. Similar effects can be seen when constraining positions of the unary potentials and learning the pairwise parameters from correct components, as this further constrains the predicted pose. The results show that using the parameters from correctly predicted components dramatically improves the localization of all body parts in each particular setting. At the same time, the combination of all settings produces the best results which indicates that the constraints coming from different settings are complementary to each other. Note that even the model with oracle component prediction does not achieve values close to 100% because of test examples with extremely foreshortened or occluded body parts.

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
(Andriluka *et al.*, 2009)	80.9	67.1	60.7	46.5	26.4	74.9	55.7
+ predict unary rotation (ur)	96.4	91.1	86.1	76.6	60.2	88.5	81.3
+ predict unary position (up)	97.1	91.4	80.7	80.2	49.5	90.1	79.1
+ predict pairwise (p/wise)	93.2	88.5	81.6	73.6	58.0	87.6	78.4
+ ur + up + p/wise	**98.3**	**96.0**	**89.4**	**87.0**	**71.8**	**94.0**	**88.1**

Table 7.1: Pose estimation results (PCP) on the "Leeds Sport Poses" (LSP) dataset by our method *when using an oracle* to choose the correct component for every potential out of 12 possible values. This confirms the intuition that predicting the correct PS model directly translates to better PCP performance.

Evaluation of poselet-conditioned potentials. We evaluate each of the poselet-conditioned potentials described in Sec. 7.3 by plugging them one by one into our model. As each potential includes a classifier that maps poselet features to one of the components, we also evaluate the performance of these classifiers. The results are shown in Tab. 7.2. It can be seen that using PS + torso prediction improves the results compared to PS alone (56.2% vs. 55.7% PCP). Interestingly, when predicting the unary position parameters even despite the somewhat low component prediction accuracy of 43.9% we are able to improve the pose estimation result from 56.2% to 59.3% PCP. Similar results are obtained when predicting the unary rotation parameters (60.3% PCP). Combination of both further improves the performance to 60.8% PCP, as both potentials are complementary to each other.

We also analyze how prediction of pairwise parameters affects pose estimation. The prediction scores of pairwise components are generally lower than the absolute unary ones. A possible explanation is that the classification problem becomes harder because several rather different poselets might still correspond to the same relative angle between the two body parts. However, the final pose estimation result is again improved (60.9% PCP). The combination of all three types of poselet-dependent potentials leads to further improvement and achieves 62.9% PCP. This indicates that the information provided by each type of potentials is complementary. Overall, our method achieves an improvement of 7.2% PCP over the original PS model that uses a generic pose prior. It shows that incorporating long range dependencies via mid-level feature representation can significantly boost the performance while keeping the inference efficient.

Comparison to competing approaches. We compare our method to competing approaches in Tab. 7.3. Interestingly, our method outperforms not only the baseline PS model (62.9% vs. 55.7% PCP), but also the popular pose estimation model (Yang and Ramanan, 2011) which we downloaded from the authors' web page and retrained on the LSP dataset for fair comparison (62.9% vs. 60.8% PCP). The improvement is most prominent in case of localizing upper legs (+6.2% PCP) whose configurations

Setting	Avg. prediction accuracy, [%]	PCP, [%]
(Andriluka *et al.*, 2009)	-	55.7
+ torso prediction	-	56.2
+ predict unary position (up)	43.0	59.3
+ predict unary rotation (ur)	37.4	60.3
+ ur + up	-	60.8
+ predict pairwise (p/wise)	30.8	60.9
+ up + ur + p/wise	-	**62.9**

Table 7.2: Accuracy of predicting a correct component for each unary and pairwise potential and corresponding pose estimation results (PCP) on the "Leeds Sport Poses" (LSP) dataset.

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
ours	**87.5**	**75.7**	68.0	54.2	33.9	78.1	62.9
(Andriluka *et al.*, 2009)	80.9	67.1	60.7	46.5	26.4	74.9	55.7
(Yang and Ramanan, 2011)	84.1	69.5	65.6	52.5	35.9	77.1	60.8
(Eichner and Ferrari, 2012a)	86.2	74.3	**69.3**	**56.5**	**37.4**	**80.1**	**64.3**

Table 7.3: Pose estimation results on the "Leeds Sport Poses" (LSP) dataset with observer-centric annotations.

can be reliably captured by the legs- and torso-legs-poselets. The improvement is also pronounced for the lower legs which profit a lot from the improved upper legs localization and for the upper arms (both +2.4% PCP). This result is very interesting since the method of (Yang and Ramanan, 2011) is a mixture of parts model that is quite different from ours, as it uses multiple unary templates for every part and image-independent pairwise potentials that do not allow to model long range part dependencies. In contrast, our model uses generic templates for each part, but incorporates a wide range of part unary terms by conditioning on poselet-representation. We also compare our method to the work (Eichner and Ferrari, 2012a), that extends the model of Yang&Ramanan by using additional background/foreground color information across images of the same dataset and modify the hard negative mining procedure. Therefore, when comparing the numbers one has to bear in mind that the reported numbers of (Eichner and Ferrari, 2012a) are based on additional information about the dataset statistics. Compared to our method the difference is most pronounced in case of forearms where the skin color information could be particularly helpful. Overall we conclude that both competing methods are orthogonal to our approach and are likely to improve when using multiple specific part templates and incorporating a color model. In Fig. 7.2 we show example pose estimation results using our method (row 1) and comparison to both (Andriluka

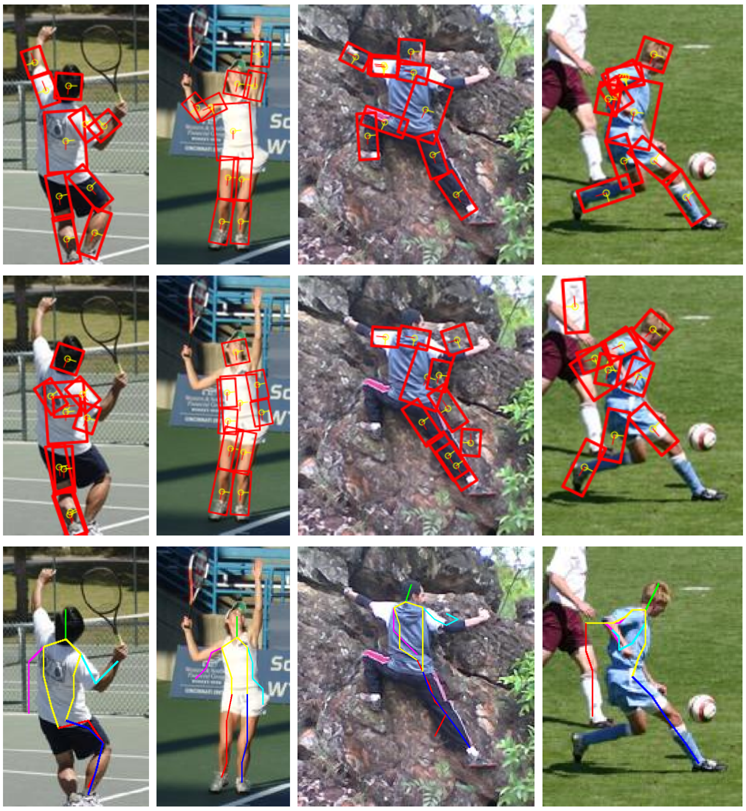


Figure 7.2: Sample pose estimation results on the LSP dataset obtained by our method (row 1), PS (Andriluka *et al.*, 2009) and the method of (Yang and Ramanan, 2011) (row 3). Modeling long-range part dependencies by our method results in better performance on highly articulated people.

et al., 2009) (row 2) and (Yang and Ramanan, 2011) (row 3). Our method is able to exploit long-range dependencies between parts across a variety of activities such as tennis serve (columns 1 and 2), climbing (column 3) and running (column 4). In Fig. 7.3 (top row) we also show several examples of failure cases. The failure cases often correspond to images of people in poses that are underrepresented in the training set, and for which the prediction of unary and pairwise components is not accurate enough.

7.4.2 Results on Image Parse dataset.

We now show the performance of our method on the "Image Parse" (IP) dataset. For evaluation we reuse the model learned on the LSP train set, but estimate the parameters w_p and K on the training set of the IP dataset. The estimated values are $w_p = 0.1$ and $K = 12$. Note that the value of w_p increased with respect to the LSP dataset, which results in a stronger influence of the poselet features on the final solution. This could be due to a larger variability of people poses on the LSP dataset compared to IP (see (Johnson and Everingham, 2011) for the discussion and

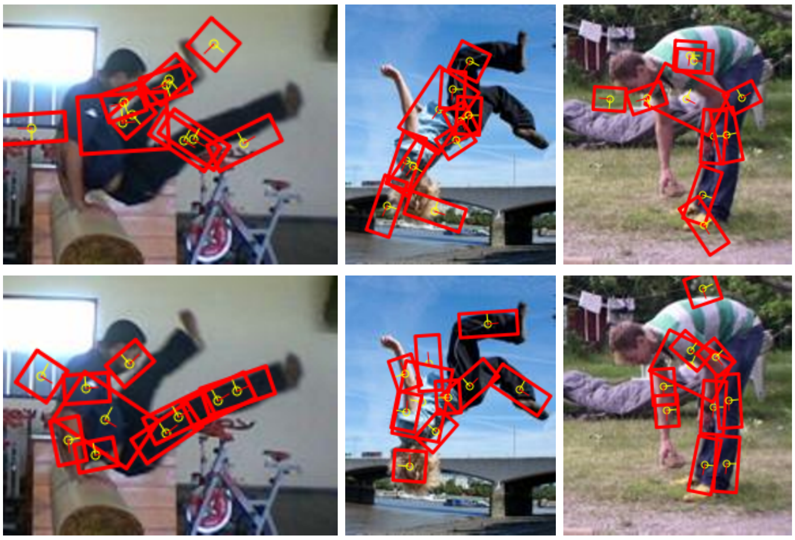


Figure 7.3: Typical failure cases on the LSP dataset. Shown are the results by our method (row 1) and PS (Andriluka *et al.*, 2009) (row 2).

comparison of the two datasets).

The results are shown in Tab. 7.4. It can be seen that our method outperforms the baseline PS model (62.9% vs. 59.2% PCP), which is in line with the results on the larger LSP dataset. Our approach favorably compares to (Yang and Ramanan, 2011), outperforming it on all body parts apart from the lower arms. The most prominent improvement is observed for the torso, but the improvement for upper/lower legs is also pronounced. Our method is slightly better than the multi-layer composite model of (Duan *et al.*, 2012). Their approach aims to capture non-tree dependencies between the parts by decomposing the model into multiple layers and performing dual decomposition to cope with cycles in the part graph. In contrast to their method, which incorporates multiple layers directly into the inference procedure making it infeasible without relaxations, our method implicitly models long-range dependencies between the parts and allows exact and efficient inference.

Our approach introduced in this chapter performs slightly worse compared to the approach presented in Chapter 5, where we extended the tree-structured pictorial structures model with additional repulsive factors between non-adjacent parts and a stronger torso detector. We extend the poselet conditioned pictorial structures approach with the repulsive factors and employ the same two-stage inference procedure. The results are shown in Tab. 7.4. The extended model corresponds to "ours + repulsive" and achieves 66.1% PCP, improving over other models trained on the LSP dataset only. Our result is only slightly worse than the result of the model from (Johnson and Everingham, 2011) that was trained on a significantly larger training set of 10000 images.

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
ours	**92.2**	74.6	63.7	54.9	39.8	70.7	62.9
ours + repulsive	90.7	**80.0**	**70.0**	59.3	37.1	77.6	66.1
(Andriluka *et al.*, 2009)	86.3	66.3	60.0	54.6	35.6	72.7	59.2
(Yang and Ramanan, 2011)	82.9	69.0	63.9	55.1	35.4	77.6	60.7
(Duan *et al.*, 2012)	85.6	71.7	65.6	57.1	36.6	**80.4**	62.8
ours (Chapter 5)	88.8	77.3	67.1	53.7	36.1	73.7	63.1
(Johnson and Everingham, 2011)	87.6	74.7	67.1	**67.3**	**45.8**	76.8	**67.4**

Table 7.4: Pose estimation results (PCP) on "Image Parse" (IP).

Method	Torso	Upper arm	Lower arm	Upper arm	Fore arm	Head	Total
ours	**91.5**	**66.8**	**54.7**	38.3	**23.9**	**85.0**	**54.4**
(Andriluka *et al.*, 2009)	88.3	64.0	50.6	**42.3**	21.3	81.8	52.6
(Wang *et al.*, 2011)	86.6	56.3	50.2	30.8	20.3	68.8	47.0

Table 7.5: Pose estimation results (PCP) on the "UIUC People".

7.4.3 Results on UIUC People dataset.

For complete evaluation of our method we finally present results on the "UIUC People" dataset. We reuse the setting from the LSP dataset. We cluster the data into 20 clusters, again preserving only those containing at least 10 examples and learn poselet detectors on both UIUC+LSP data. The results are shown in Tab. 7.5. It can be seen that using only dataset-specific poselets already improves the results over the baseline PS model. This finding is consistent for all three datasets, we always improved when using poselet conditioned features. Interestingly, our method performs better than the approach of (Wang *et al.*, 2011) that also falls behind the baseline PS model. This method is based on hierarchical poselets which intend to capture the non-tree dependencies between the parts via multiple layers. Such a model structure inevitably introduces cycles and requires an approximate inference.

7.5 CONCLUSION

Pose estimation is often addressed with pictorial structures (PS) models based on a tree-structured graph leading to efficient and exact inference. However, tree-structured models fail to capture important dependencies between non-connected body parts leading to estimation failures. This work proposes to capture such dependencies using poselets that serve as a mid-level representation that jointly encodes articulation of several body parts. We show how an existing PS model for human pose estimation can be improved using a poselet representation. The resulting model is as efficient as the original tree-structured PS model, and is at the same time capable of representing complex dependencies between multiple parts. Experimental results show that a better prediction of human body layout using poselets improves body part estimation. We observe a consistent improvement on all of the considered datasets.

We believe that the components of our model could be further improved. In particular, local appearance of individual body parts is modeled using a single shape context template that cannot account for multi-modal part appearance distribution due to clothing, imaging conditions, part size and articulation. Furthermore, current body model is a "cardboard" model that can barely capture variations in body part geometry due to out of plane rotations. Therefore, in the next chapter we address these shortcomings by analyzing and building on the successful ideas from the human pose estimation literature.

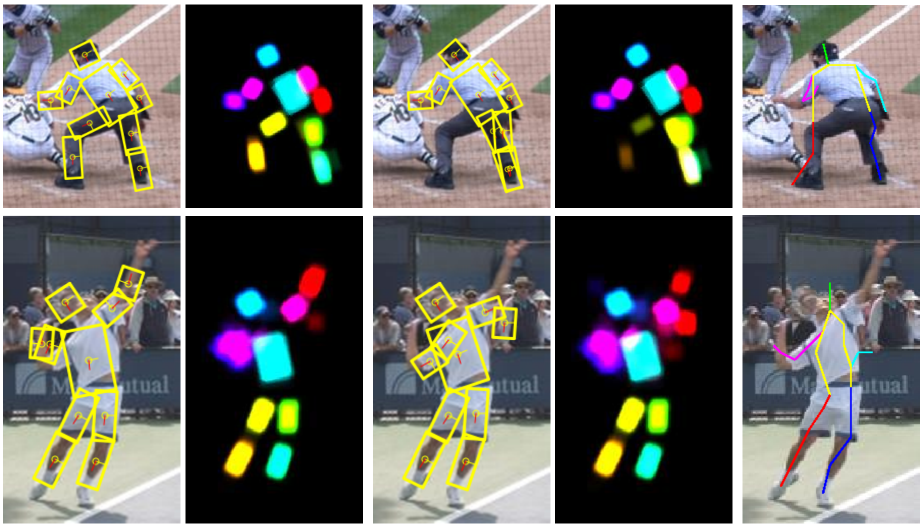
8

STRONG APPEARANCE AND EXPRESSIVE SPATIAL MODELS FOR HUMAN POSE ESTIMATION

Contents

8.1 Introduction **134**
8.2 Pictorial structures model **136**
  8.2.1 Model formulation 136
  8.2.2 Learning and inference 136
8.3 Better appearance representations **137**
  8.3.1 Body part detectors 137
  8.3.2 Head and torso detectors (*spec-head*, *spec-torso*) 138
  8.3.3 Implementation details 138
8.4 More flexible models **139**
  8.4.1 Body joints (*PS-flex*) 139
  8.4.2 Mid-level representations (*mid-level*) 139
8.5 Results **140**
  8.5.1 Results on LSP dataset 140
  8.5.2 Results on Image Parse dataset 144
8.6 Conclusion **146**

Typical approaches to articulated pose estimation combine spatial modeling of the human body with appearance modeling of body parts. This chapter aims to advance articulated pose estimation in two ways. First we explore various types of appearance representations aiming to substantially improve the body part hypotheses. And second, we draw on and combine several powerful ideas such as more flexible spatial models as well as our image-conditioned spatial models proposed in Chapter 7. In a series of experiments we draw several important conclusions: (1) we show that the proposed appearance representations are complementary; (2) we demonstrate that even a basic tree-structure spatial human body model achieves very good performance when augmented with the proper appearance representation; and (3) we show that the combination of the best performing appearance model with a flexible image-conditioned spatial model achieves very good results, significantly improving over competitors, on the "Leeds Sports Poses" and "Image Parsing" benchmarks.



(a) (b) (c)

Figure 8.1: Example pose estimation results and corresponding part marginal maps obtained by (a) our full model combining local appearance and mid-level representation, (b) our best local appearance model and (c) results by Yang&Ramanan (Yang and Ramanan, 2011).

8.1 INTRODUCTION

Prominent approaches to human pose estimation rely on the pictorial structures model representing the human body as a collection of rigid parts and a set of pairwise part dependencies. The appearance of the parts is often assumed to be mutually independent. Part detectors are either trained independently (Johnson and Everingham, 2011; Andriluka *et al.*, 2011) or jointly with the rest of the model (Yang and Ramanan, 2011; Desai and Ramanan, 2012). While effective detectors have been proposed for specific body parts with characteristic appearance such as heads and hands (Mittal *et al.*, 2012; Gkioxari *et al.*, 2013), detectors for other body parts are typically weak. Obtaining strong detectors for all body parts is challenging for a number of reasons. The appearance of body parts changes significantly due to clothing, foreshortening and occlusion by other body parts. In addition, the spatial extent of the majority of the body parts is rather small, and when taken independently each of the parts lacks characteristic appearance features. For example lower legs often appear as a pair or parallel edges.

We argue that in order to obtain effective part detectors it is necessary to leverage both the pose specific appearance of body parts, and the joint appearance of part constellations. Pose specific person and body part detectors have appeared in various forms in the literature. For example, people tracking approaches (Ramanan *et al.*, 2005; Fossati *et al.*, 2007) rely on specialized detectors tailored to specific people poses that are easy to detect. Similarly, prominent approaches to people detection (Bourdev *et al.*, 2010) build on a large collection of pose specific poselet detectors. Local (Yang

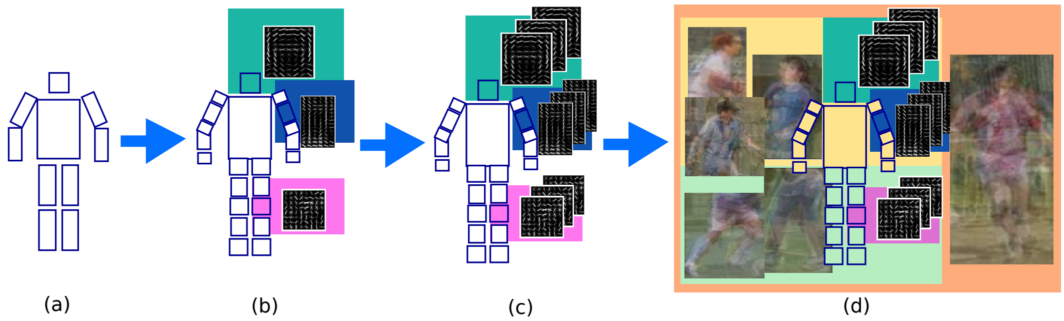


Figure 8.2: Overview of our method. We extend basic *PS* model (Andriluka *et al.*, 2009) (a) to more flexible structure with stronger local appearance representations including single component part detectors (b) and mixtures of part detectors (c). Then we combine local appearance model with mid-level representation based on semi-global poselets which capture configurations of multiple parts (d). Shown are the means of sample poselet clusters. Color coding shows different levels of granularity of our appearance and spatial models.

and Ramanan, 2011) and global (Johnson and Everingham, 2011) mixture models that capture pose specific appearance of individual body parts and joints have shown to be effective for pose estimation. These approaches capture appearance at different levels of granularity: full person vs. subset of parts vs. individual parts and differ in the way they represent the appearance.

This chapter builds on findings from the literature and follows two complementary routes to a more powerful pose model: improving the appearance representation and increasing the expressiveness of the joint body part model (see Fig. 8.1 and 8.3 for samples). Specifically, we consider local appearance representations based on rotation invariant or rotation specific appearance templates, mixtures of such local templates, specialized models tailored to appearance of salient body parts such as head and torso, and semi-global representations based on poselet features (Sec. 8.3). The second main contribution of the chapter is to combine the improved appearance model with more expressive body representations. These include the flexible models of (Sapp *et al.*, 2011; Yang and Ramanan, 2011) and our image conditioned spatial model proposed in Chapter 7. This contribution is presented in Sec. 8.4.

Starting with the basic tree-structured pictorial structures we perform a series of experiments incrementally adding various components and analyzing the resulting performance gains (Fig. 8.2). Our analysis reveals several surprising facts (Sec. 10.4). The performance of the best appearance model for individual body parts is surprisingly high and can even compete with some approaches using weaker appearance terms but a full spatial model (Tab. 8.4). When augmented with the best appearance model, the basic tree-structured pictorial structures model perform superior to competing models (Eichner and Ferrari, 2012a; Yang and Ramanan, 2011) (Tab. 8.3). We show that strong appearance representations operating at different levels of granularity (mixtures of local templates vs. semi-global poselets) are complementary.

Finally, we report very good results on the "Image Parsing" and "Leeds Sports Poses" benchmarks, which are obtained by combining the best appearance model with our image conditioned pictorial structures spatial model (Tabs. 8.5 and 8.6).

8.2 PICTORIAL STRUCTURES MODEL

In the following we briefly summarize the basic tree-structured pictorial structures model (Fischler and Elschlager, 1973; Felzenszwalb and Huttenlocher, 2005), that will serve as the baseline model for our analysis. In Sec. 8.3 and 8.4 we describe several extensions.

8.2.1 Model formulation

The pictorial structures model represents the human body as a collection of rigid parts $L = \{l_1, \ldots, l_N\}$ and a set of pairwise part relationships. The state of each part is denoted by $l_n = (x_n, y_n, \theta_n, s_n)$, where (x_n, y_n) is the image position of the part, θ_n is the absolute orientation, and s_n is the part scale relative to the part size in the scale-normalized training set. Denoting the image observations by D, the energy of the body part configuration L defined by the pictorial structures model is given by

$$E(L; D) = \sum_{n=1}^{N} E^u(l_n; D) + \sum_{n \sim m} E^p(l_m, l_n). \tag{8.1}$$

The pairwise relationships between body parts are denoted by $n \sim m$. They follow the kinematic chain and thus result in a tree structured model.

We use the pictorial structures model introduced in (Felzenszwalb and Huttenlocher, 2005) as our baseline model, and refer to it as *PS* in the remainder. This model is composed of $N = 10$ body parts: head, torso, and left and right upper arms, forearms, upper legs and lower legs. The parts are pairwise connected to form a tree corresponding to the kinematic chain, see Fig. 8.2(a). The pairwise terms E^p encode the kinematic dependencies and are represented with Gaussians in the transformed space of joints between parts. We refer to the original chapter (Felzenszwalb and Huttenlocher, 2005) for the details on the pairwise terms. Note that in the basic model the spatial extent of each part, and in particular the distance between part center and position of its joints is fixed, which potentially restricts the model to the configurations with relatively little foreshortening.

8.2.2 Learning and inference

In this chapter we use the publicly available implementation of the pictorial structures approach (Andriluka *et al.*, 2011). The parameters of unary and pairwise factors are learned using piecewise training. The pairwise term is set using a Maximum-Likelihood estimate that is available in closed form. The unary terms are described in Sec. 8.3.

Inference in the model is performed with sum-product belief propagation. Due to the tree structure this is an exact inference procedure yielding the marginal distributions for each body part. Predictions are then obtained by taking the maximum marginal state for each part. Some PS model variants that we will describe include auxiliary (latent) variables, this procedure thus marginalizes them out.

8.3 BETTER APPEARANCE REPRESENTATIONS

We now turn our attention to improving the appearance representations for body parts. These correspond to the unary terms E^u in Eq. 8.1.

As the baseline model we consider the appearance representation introduced in (Andriluka *et al.*, 2011). These factors use boosted part detectors over shape context features, one detector per body part. This appearance representation is made independent to the part rotation, by normalizing the training examples with respect to part rotation prior to learning. At test time, the detector is evaluated for each of 48 rotations in the discretized state-space of the PS model. The model that uses only this unary factor will be denoted as *PS*. [9]

8.3.1 Body part detectors

The rotation independent representation from (Andriluka *et al.*, 2011) is based on a simplifying assumption, namely that the appearance of model parts does not change with part rotation. This typically is not true. For example the upper arms raised above the head and the ones held in front of the torso look quite different because of the overlap with other parts and change in the contours of the shoulders. This motivated rotation dependent detectors as in (Yang and Ramanan, 2011; Gkioxari *et al.*, 2013).

We augment *PS* with two types of such local representations: 1) a rotation dependent detector tailored to the absolute orientation of the part (*rot-dep mix*) and 2) a rotation invariant representation tailored to a particular body pose (*pose-dep mix*). As an implementation we choose the deformable part model (DPM) (Felzenszwalb *et al.*, 2010) that has proven to be very reliable for detection purposes.

Absolute rotation. Rotation dependent part detectors are obtained in the following way. We discretize the rotation space in $N = 16$ different bins, corresponding to a span of 22.5 degrees. All training data is assigned to the corresponding rotation bin based on the annotation. We then train a 16 component model, one component for each bin. As these models do capture rotation dependent appearance changes, we refer to this variant as *rot-dep mix*. A simpler baseline is a single component model trained for all rotations together. We include this model in the comparison under the name *rot-dep single*.

---

[9] Please see (Andriluka *et al.*, 2011) for further implementation details.

Relative rotation. Rotation of the body parts is related to the orientation of the entire body, not necessarily to the absolute value in the image plane. We model this using a part detector that depends on the body pose. For this we normalize the part to a common rotation but rotate the entire body along with it. Then a binning in again 16 clusters is obtained by using the visibility features proposed in (Desai and Ramanan, 2012). This clustering results in components that are compact w.r.t. the body pose in the proximity of the body part. The resulting detector is referred as *pose-dep mix*. Since this is "rotation invariant", in the sense the absolute rotation is irrelevant, during test time we evaluate this detector for all rotations in the state space of *PS*. We also include a simpler baseline which is a single component model trained from rotation-normalized body parts and then again evaluated for all rotations. We refer to it as *rot-inv single*.

8.3.2 Head and torso detectors (*spec-head*, *spec-torso*)

We consider two types of specialized part detectors proposed in the literature. Our torso detector from Chapter 5 and the head detector from (Marin-Jimenez *et al.*, 2011). The main rationale behind using such specialized detectors is that body parts such as head and torso have rather specific appearance that calls for specialized part models.

Specifically, the torso detector introduced in Chapter 5 is directly adapted from the articulated person detector based on a DPM. A torso prediction is obtained by regression using the positions of the latent DPM parts as features. This specialized torso detector benefits from evidence from the entire person and captures the pose. This is in contrast to the previous local torso model as it is not bound to evidence within the torso bounding box only. We refer to the specialized torso detector as *spec-torso*.

The head detector of (Marin-Jimenez *et al.*, 2011) uses the observation that the main source of variability for the head is due to the viewpoint of the head w.r.t. the camera, e.g. front and profile views have a different but rather distinctive appearance. Following (Marin-Jimenez *et al.*, 2011) we train a DPM detector for the head with 8 components corresponding to a set of viewpoints discretized with a step of 45 degrees. Note that the particular set of components is not available for the local detectors of the head that are either grouped by the in plane rotation or by the pose of the surrounding parts. We refer to specialized head detector as *spec-head*.

8.3.3 Implementation details

All detectors outlined above are based on the DPM v4.0 framework and we utilize the publicly available software (Felzenszwalb *et al.*, 2010). To turn a set of DPM detections after non-maximum suppression into a dense score for every pixel we apply a kernel density estimate (KDE). From the set $\{(d_k, s_k)\}, k = 1, \ldots, K$ with d_k denoting the detection position and s_k the detection score we define the score for part l_n as the value of the KDE $E^u(l_n; D) = \log \sum_k w_k \exp(-\|l_k - d_k\|^2/\sigma^2)$, where $w_k = s_k + m$, and m is a minimal detection score produced by the detector, which is set to -3.0 in our experiments. We then add the normalized DPM scores to the boosted part detector (Andriluka *et al.*, 2009) scores at every position of the dense scoregrid and use these summed scores in the inference.

8.4 MORE FLEXIBLE MODELS

Besides improving the pure appearance representations several works suggested to alter the model representation to make it more flexible. We incorporate their findings and include two modifications to the standard PS model.

8.4.1 Body joints (*PS-flex*)

The original PS model represents body parts as variables, which in turn make appearance changes such as foreshortening very drastic. Follow-up work has suggested to build appearance representation for more local parts while allowing more flexibility in their composition (Sapp *et al.*, 2011; Yang and Ramanan, 2011). We incorporate this by including an additional 12 variables that represent location of the joints in the human body. These parts correspond to the left and right shoulder, elbow, wrist, hip, knee and ankle. In order to retain deterministic inference we incorporate these parts such that the resulting model is still tree-structured, as illustrated in Fig. 8.2(b). The additional pairwise terms between joint parts and body parts are modeled as a Gaussian factor w.r.t. their position. Since some body and joint parts are restricted to have the same absolute rotation, such as lower arm and wrist, we add a constraint on their rotation and scale to be identical. We refer to our flexible model as *PS-flex*.

8.4.2 Mid-level representations (*mid-level*)

Poselet conditioned deformation terms. The basic *PS* model has a limitation that the spatial distribution of the body parts is modeled as a Gaussian and can not properly represent the multi-modalities of human poses. We therefore take advantage of our image conditioned model introduced in Chapter 7 and substitute the unimodal image independent spatial factors in Eq. 8.1 with image conditioned factors. We define multiple pairwise terms for each joint by clustering the training data w.r.t. relative part rotation, and then predict the type of the pairwise term at test time based on the image features. To do so we train part configuration detectors called poselets and then use their responses during test time as mid-level feature representation (c.f. Fig. 8.2(d)). Prediction is treated as a multi-class classification problem where we use a classifier based on sparse linear discriminant analysis (sLDA) (Clemmensen *et al.*, 2011). We denote this image conditioned flexible configuration as *mid-level p/wise*.

Poselet conditioned appearance. The local appearance models introduced in Sec. 8.3 are designed to capture pose dependent appearance of individual parts and pairs of adjacent parts. In order to capture appearance of the person at a higher level of granularity we extend our model with a mid-level poselet based representation and use poselet features described above to obtain rotation and position prediction of each body part separately. For instance, to predict part positions, we cluster the training data for each part based on part relative offset w.r.t. torso center into set of clusters. Then for each cluster its mean offset from the torso and the variance are computed. We then train a sLDA classifier to predict from the poselet features the mean and variance of the relative offset for every part and use these values as a Gaussian unary potential, which we add to other unary potentials introduced in Sec. 8.3. Prediction of absolute part orientation is done in a similar way. We call these representations in the following experiments as *mid-level rot* and *mid-level pos*, respectively and refer to Chapter 7 for further details on the implementation of these terms.

8.5 RESULTS

In this section we evaluate the proposed extensions on two well-known pose estimation benchmarks and compare to other competitors. As a performance measure we use the common PCP loss (Ferrari *et al.*, 2008).

Datasets. For evaluation we use the publicly available pose estimation benchmarks exhibiting strong variations in articulation and viewpoint: "Leeds Sports Poses" (LSP) dataset (Johnson and Everingham, 2010) that includes 1000 images for training and 1000 for testing showing people involved in various sports; the "Image Parsing" (IP) (Ramanan, 2006) dataset consisting of 100 train images and 205 test images of fully visible people in diverse set of activities such as sports, dancing and acrobatics.

8.5.1 Results on LSP dataset

In this section we report on the results obtained using the various extensions outlined in the last two sections. We follow (Eichner and Ferrari, 2012a) and use *observer-centric (OC)* annotations provided by the authors for evaluation. We train all the representations using the training set of LSP dataset.

Flexible Model We start with a comparison of models using body part appearance alone (*PS*) with the flexible model *PS-flex* that includes both joint and body part appearance. The results are shown in Tab. 8.1. We observe an improvement (+2.4%) due to better localization of lower legs and arms. This reinforces the findings of (Yang and Ramanan, 2011): a flexible model of joints copes better with foreshortening. When removing the body parts for arms and legs and use only body joints (joints only) the performance drops. We attribute this to the easier confusion of joint

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore- arm	Head	Total
PS (Andriluka *et al.*, 2009)	**80.9**	67.1	60.7	46.5	26.4	**74.9**	55.7
PS-flex (joints only)	80.1	69.0	64.7	43.6	27.3	70.5	56.0
PS-flex	80.5	**70.2**	**66.5**	**46.7**	**32.0**	70.2	**58.1**

Table 8.1: Results on LSP when varying number of parts in PS.

detectors to background clutter. We conclude that the *PS-flex* model should benefit from better appearance representations which we will evaluate next.

Single component detectors. Performance of rotation dependent (*rot-dep single*) and rotation invariant (*rot-inv single*) single component detectors is reported in Tab. 8.2. Surprisingly, adding *rot-dep single* already improves the overall result (+2.7%), mostly due to better head localization (+8.1%). The majority of the poses in the dataset are upright, thus much of head appearance change is captured by the *rot-dep single* detector. As expected, the result is further improved by *rot-inv single*, and the improvement is most prominent for lower arms (+7.8%). This clearly shows that rotation invariance of a single component detector is key to cope with the high degree of articulation by training and testing samples.

Mixtures of part detectors. Rotation dependent mixture of detectors (*rot-dep mix*) accounts for the characteristic appearance changes of body parts under rotation. These types of detectors indeed improve the results, see line 4 in Tab. 8.2. When compared with the single counterparts we observe significant performance gain for all body parts.

While the former detectors are (in)variant to local rotations, they do not take the pose-specific appearance into account. The detectors *pose-dep mix* do. However, we do not observe any performance increase over *rot-dep mix*. We believe this is due to more compact cluster representations of the *rot-dep mix*, which makes them more discriminative. In summary, the best local mixture appearance representation improves over best single component detector by 2.8%, improving results for all parts. This indicates that mixtures better handle the highly multi-modal local appearance of body parts.

Specialized detectors. We discussed the possibility for designing specialized body part detectors in Section 8.3.2. We add those detectors to the *pose-dep mix* model, also including a Gaussian term on the torso location estimated via Maximum Likelihood on the training annotations. The results can be found in the last two lines of Tab. 8.2. Both the specialized torso and head detector improve the performance of torso and head localization, and via the connected model also improve the performance of other body parts. Even though the better torso prediction improves head localization (+0.3%), a specialized head detector still improves the performance (+1.1%). Since the parts are connected to the head via the torso, the influence of the *spec-head* detector on

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
PS-flex	80.5	70.2	66.5	46.7	32.0	70.2	58.1
+ rot-dep single	82.2	72.5	67.9	51.6	31.6	78.3	60.8
+ rot-inv single	83.6	73.6	69.8	52.4	39.4	78.1	63.2
+ rot-dep mix	87.2	76.0	72.2	55.9	40.5	83.3	66.0
+ pose-dep mix	84.5	75.4	70.3	53.4	40.5	78.0	64.2
+ spec torso	88.4	76.5	72.6	56.5	41.1	83.6	66.6
+ spec head	**89.2**	**76.7**	**72.8**	**56.9**	**41.2**	**84.7**	**66.9**

Table 8.2: Results on LSP using local appearance models.

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
local appearance	89.2	76.7	72.8	56.9	41.2	84.7	66.9
+ mid-level rot	89.0	77.6	73.2	58.1	42.5	85.3	67.7
+ pos	**89.4**	78.7	**74.0**	59.7	43.9	**86.0**	68.8
+ p/wise	88.7	**78.8**	73.4	**61.5**	**44.9**	85.6	**69.2**

Table 8.3: Results on LSP using mid-level representations.

other body parts is found to be smaller. In summary, specialized detectors improve estimation results for all body parts, and give a +0.9% better results in terms of PCP. We expect this result would carry over to other models from the literature.

Mid-level representations. Now we combine the best performing local appearance representation with the mid-level representation introduced in Chapter 7. We use the same parameters as reported by the authors. Results are shown in Tab. 8.3. Predicting absolute orientation of parts based on mid-level representation (*mid-level rot*) noticeably improves results (+1.2%). Consistent improvement is achieved for each limb with forearms improving the most (+1.3%). Adding prediction of part positions based on mid-level features (*mid-level pos*) leads to further improvements (+1.1%). Again upper/lower arms profit the most from semi-global poselet detectors. They exhibit higher degree of articulation compared to other parts and thus are more difficult to detect using local detectors. Finally, adding prediction of pairwise terms (*mid-level p/wise*) improves the total performance, achieving an outstanding 69.2%. Overall, adding mid-level representations to the best performing local appearance model improves the results by 2.3%, giving improved results for all body parts. These results demonstrate the complementary effect of local appearance models and mid-level representations. Mid-level representation based on semi-global poselets models long range part dependencies, while local appearance model concentrate on local changes in the appearance of body parts.

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
PS-flex	36.2	20.1	27.1	6.8	5.5	40.2	19.5
+ local appearance	67.1	36.2	35.2	18.6	10.6	63.0	33.1
+ mid-level	**79.5**	**65.5**	**63.5**	**46.9**	**26.9**	**77.1**	**56.2**

Table 8.4: Performance on LSP using part appearance only.

Performance using unaries only. Finally we evaluate how much the appearance representation alone contributes to the final performance. To do so we remove all connections between the parts and evaluate part detectors only. Results are shown in Tab 8.4. As expected, boosted detectors of *PS-flex* perform worst. Adding our best local appearance model significantly improves the results (+13.6%), which demonstrates the strengths of the local appearance models compared to the original boosted detectors. Local mixtures of part detectors allow to model pose-dependent appearance of limbs while strong specific head and torso detectors push the performance of both most salient body parts (67.1 vs 36.2% for torso and 63.0 vs. 40.2% for head). Including the mid-level representation significantly improves the result further (+23.1%). So, upper/lower arms which are difficult to detect by local detectors profit a lot from semi-global poselets (+28.3 and +16.3%). A similar trend can be observed for upper/lower legs. This again demonstrates the strengths of mid-level representation and its complementary w.r.t. the local appearance models.

Comparison to competitors. We compare our approach to competitors in Tab. 8.5. Interestingly, our full model including local appearance and mid-level representations outperforms not only the baseline *PS* (Andriluka *et al.*, 2009) (69.2 vs 55.7%), but also other competitors by quite a margin, improving 4.9% over the next best performing method (Eichner and Ferrari, 2012a). The results also improve over our Poselet Conditioned PS proposed in Chapter 7 (69.2 vs. 62.9%) where we use similar mid-level representations but have a more simplistic local appearance model based on (Andriluka *et al.*, 2009). This is consistent for all body parts: torso +1.7%, upper legs +3.1%, lower leg +5.4%, upper arm +7.3%, forearm +11.0%, head +7.5%. We found this result interesting, as it clearly shows how much performance gain can be achieved by improving local part appearance while preserving the mid-level representation. We also compare our method to the popular pose estimation model (Yang and Ramanan, 2011) which we downloaded from the authors' web page and retrained on LSP dataset for fair comparison. Interestingly, our local appearance model combined with basic Gaussian pairwise terms already outperforms their method (66.9% vs. 60.8%). This demonstrates the strengths of the proposed local appearance model based on mixtures of pose-dependent detectors and specific torso and head detectors. When using our full model we outperform (Yang and Ramanan, 2011) by 8.4%. Finally, we compare our method to recent work (Eichner and Ferrari, 2012a), that extends the model (Yang and Ramanan, 2011) using additional back-

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
Our local appearance	**89.2**	76.7	72.8	56.9	41.2	84.7	66.9
Our full model	88.7	**78.8**	**73.4**	**61.5**	**44.9**	**85.6**	**69.2**
(Andriluka *et al.*, 2009)	80.9	67.1	60.7	46.5	26.4	74.9	55.7
(Yang and Ramanan, 2011)	84.1	69.5	65.6	52.5	35.9	77.1	60.8
Poselet Conditioned PS (Chapter 7)	87.5	75.7	68.0	54.2	33.9	78.1	62.9
(Eichner and Ferrari, 2012a)	86.2	74.3	69.3	56.5	37.4	80.1	64.3

Table 8.5: Comparison of pose estimation results (PCP) on LSP dataset to current methods using observer-centric (OC) annotations. Results using person-centric (PC) annotations available on our evaluation web page human-pose.mpi-inf.mpg.de under “Related Benchmarks”)

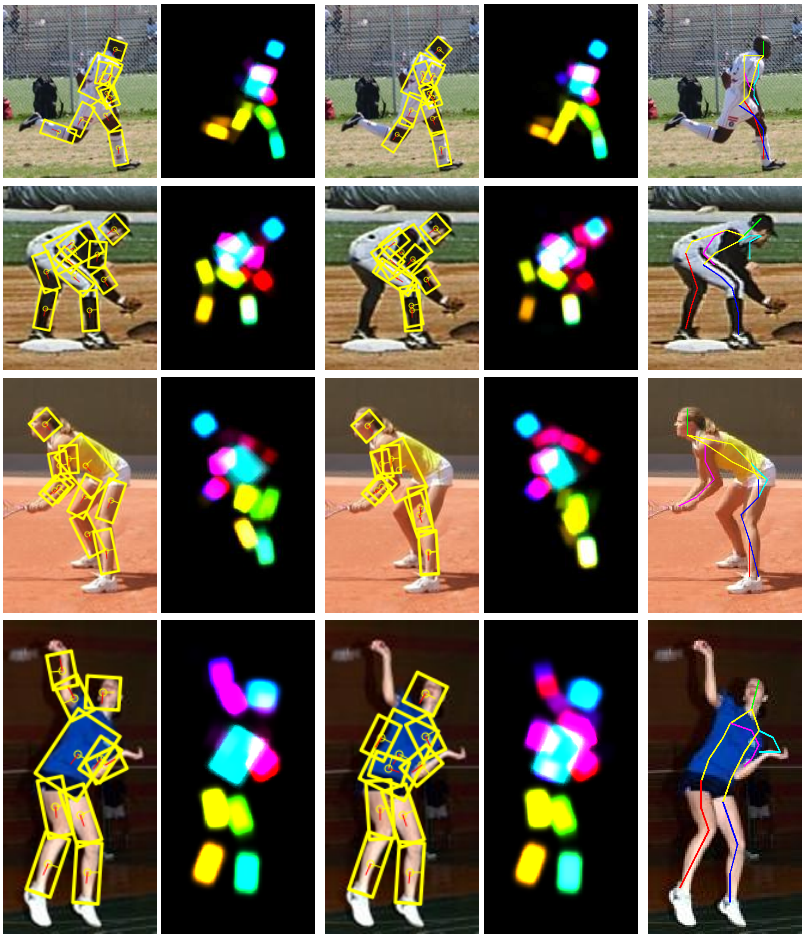
ground/foreground color information across images of the same dataset and modify the hard negative mining procedure. Thus when comparing to (Eichner and Ferrari, 2012a) one should bear in mind that the reported numbers are based on additional information about the dataset statistics. Again, our local appearance model already performs better (66.9 vs. 64.3%). Comparing our full model, we observe an improvement of striking 4.9% over the best performing competitor (Eichner and Ferrari, 2012a). This demonstrates the strength of combining local appearance modeling with flexible mid-level representations.

Qualitative evaluation. Successful results of our model are shown in Fig. 8.3 (rows 1-4). Our local appearance model already achieves good results (Fig. 8.3(b)), as it is able to cope with highly variable part appearance. Our full model which also includes mid-level representations further improves the results (Fig. 8.3(a)), as it captures the entire pose of the body and models other part dependencies. This is in contrast to Yang&Ramanan (Yang and Ramanan, 2011) (Fig. 8.3(c)) who rely only on local image evidence. Typical failure cases of our model include large variations in scale between body parts (Fig. 8.3 (line 5)), untypical appearance and poses (line 6) and massive self-occlusion (line 7).

8.5.2 Results on Image Parse dataset

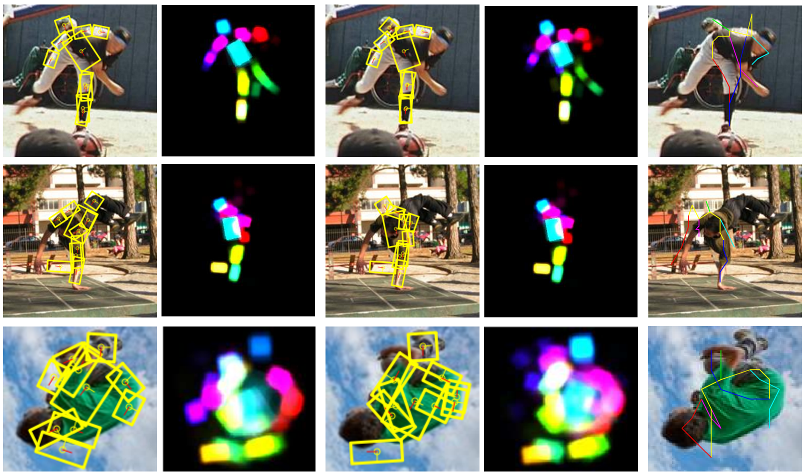
In the experiments on the Image Parse dataset (Ramanan, 2006) we use our full model trained on the LSP dataset and set the parameters of the mid-level representation as reported in Chapter 8. In Tab. 8.6 we compare our full model with a number of pose estimation approaches from the literature. Our method improves over the best performing competitor by 2.0%.

The approach proposed in this chapter outperforms our own Poselet Conditioned PS proposed in Chapter 7 (+6.5%). This result is in line with the findings on LSP, and shows the importance of better appearance models. Our method consistently



(a) (b) (c)

successful cases



(a) (b) (c)

failure cases

Figure 8.3: Qualitative results: estimated poses and corresponding part marginal maps obtained by (a) our full model combining local appearance and flexible mid-level representation, (b) our local appearance model and (c) results by Yang&Ramanan (Yang and Ramanan, 2011).

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
Our full model	**93.2**	77.1	68.0	63.4	**48.8**	86.3	**69.4**
(Andriluka *et al.*, 2011)	86.3	66.3	60.0	54.6	35.6	72.7	59.2
(Yang and Ramanan, 2011)	82.9	69.0	63.9	55.1	35.4	77.6	60.7
(Duan *et al.*, 2012)	85.6	71.7	65.6	57.1	36.6	80.4	62.8
Joint PS (Chapter 5)	88.8	**77.3**	67.1	53.7	36.1	73.7	63.1
Poselet Conditioned PS (Chapter 7)	92.2	74.6	63.7	54.9	39.8	70.7	62.9
(Yang and Ramanan, 2013)	85.9	74.9	**68.3**	63.4	42.7	**86.8**	67.1
(Johnson and Everingham, 2011)	87.6	74.7	67.1	**67.3**	45.8	76.8	67.4

Table 8.6: Comparison of pose estimation results (PCP) on "Image Parse" dataset to current methods.

improves over the pose estimation model of Yang&Ramanan (Yang and Ramanan, 2011) (+8.7%) and the over the newer version of this model from (Yang and Ramanan, 2013) (+2.3%). The improvement is achieved for all body parts apart from head and lower legs. In particular, we improve on highly articulated forearms (+6.1%) and upper legs (+2.2%). This demonstrates that much improvement can be gained from the complementary mid-level representation. Our result is also significantly better than the multi-layer composite model of (Duan *et al.*, 2012) (+6.6%), who captures non-tree part dependencies by decomposing the model into several layers and using dual decomposition to cope with the resulting loopy graph. In contrast, our method implicitly models long-range dependencies between the parts by using mid-level representation while allowing exact and efficient inference. The proposed approach outperforms our Joint PS method proposed in Chapter 5 (+6.3%), where we also integrate the evidence from a people detector into the PS framework to improve torso localization. Joint PS introduces loops between the corresponding upper/lower legs to prevent over-counting, again yielding more expensive inference. Finally, our method outperforms (Johnson and Everingham, 2011) (+2.0%). Note that their model also uses strong local appearance models and is trained on an additional dataset of 10000 images.

8.6 conclusion

In this chapter we investigated the use of 1) stronger appearance models and 2) more flexible spatial models. We observe that better local appearance representations directly result in better performance and even a basic tree-structured human body model achieves very good performance when augmented with the proper appearance representation. The second route explored in this chapter are more flexible spatial body models with image conditioned terms based on mid-level representations, implemented as poselets. We find significant improvement using this information, both when using a connected and even a disconnected body model. The effects of the terms studied are found to be additive, the combination significantly improves the performance as demonstrated on two benchmark datasets. The source code of our approach was made publicly available[10]. Note that all representations considered in this chapter rely on the image gradient information only. In Chapter 11 we introduce a novel pose estimation model based on deep learning part detectors and demonstrate significantly improved human pose estimation performance.

---

[10] www.d2.mpi-inf.mpg.de/poselet-conditioned-ps

9 2D HUMAN POSE ESTIMATION: NEW BENCHMARK AND STATE OF THE ART ANALYSIS

Contents

9.1 Introduction **149**
9.2 Dataset **151**
9.3 Analysis of human pose estimation approaches **153**
  9.3.1 Analysis of pose estimation challenges 155
  9.3.2 Evaluation of Deep Learning based Methods 163
9.4 Conclusion **166**

As discussed in Chapters 7 and 8 human pose estimation has made significant progress during the last years. However current datasets are limited in their coverage of the overall pose estimation challenges. Still these serve as the common sources to evaluate, train and compare different models on. In this chapter we introduce a novel benchmark "MPII Human Pose"[11] that makes a significant advance in terms of diversity and difficulty, a contribution that we feel is required for future developments in human body models. This comprehensive dataset was collected using an established taxonomy of over 800 human activities Ainsworth *et al.* (2011). The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational and householding activities, and capture people from a wider range of viewpoints. We provide a rich set of labels including positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels. For each image we provide adjacent video frames to facilitate the use of motion information. Given these rich annotations we perform a detailed analysis of several human pose estimation approaches gaining insights for the success and failures of these methods.

9.1 INTRODUCTION

Recent pose estimation methods employ complex appearance models (Dantone *et al.*, 2013; Gkioxari *et al.*, 2013; Tompson *et al.*, 2014; Chen and Yuille, 2014) and rely on learning algorithms to estimate model parameters from the training data. The performance of these approaches crucially depends on the availability of annotated training images that are representative for the appearance of people clothing, strong articulation, partial (self-)occlusions and truncation at image borders. Although

---

[11] Available at human-pose.mpi-inf.mpg.de.

bicycling
bicycling, BMX

conditioning exercise
ski machine

dancing
ballroom

fishing and hunting
fish. from river bank

home activities
tanning hides

home repair
carpentry

inactivity quiet
sitting quietly

lawn and garden
driving tractor

miscellaneous
standing

music playing
violin, sitting

occupation
horse grooming

religious activities
sit., playing instrum.

running
running, stairs, up

self care
taking medication

sports
soccer

transportation
riding in a bus

volunteer activities
playing with children

walking
bird watching

water activities
snorkeling

winter activities
skating, ice dancing

Figure 9.1: Randomly chosen images from each of 20 activity categories of the proposed "MPII Human Pose" dataset. Image captions indicate activity category (1st row) and activity (2nd row). To view the full dataset visit human-pose.mpi-inf.mpg.de.

there exists training sets for special scenarios such as sport scenes (Johnson and Everingham, 2010, 2011) and upright people (Sapp and Taskar, 2013; Dantone *et al.*, 2013), these benchmarks are still limited in their scope and variability of represented activities. Sport scene datasets typically include highly articulated poses, but are limited with respect to variability of appearance since people are typically wearing tight sports outfits. In turn, datasets such as "FashionPose" (Dantone *et al.*, 2013) and "Armlets" (Gkioxari *et al.*, 2013) aim to collect images of people wearing a variety of different clothing types, and include occlusions and truncation but are dominated by people in simple upright standing poses.

To the best of our knowledge no attempt has been made to establish a more representative benchmark aiming to cover a wide pallet of challenges for human pose estimation. We believe that this hinders further development on this topic and propose a new benchmark "MPII Human Pose". Our benchmark significantly advances state of the art in terms of appearance variability and complexity, and includes more than 40,000 images of people. We used YouTube as a data source and collected images and image sequences using queries based on the descriptions of more than 800 activities. This results in a diverse set of images covering not only different activities, but indoor and outdoor scenes, a variety of imaging conditions, as well as both amateur and professional recordings (cf. Fig. 1). This allows us to study existing body pose estimation techniques and identify their individual failure modes.

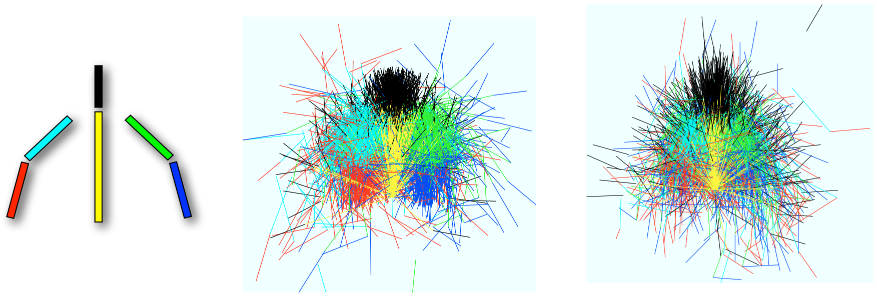
9.2 DATASET

In this chapter we introduce a large dataset of images that covers a wide variety of human poses and clothing types and includes people interacting with various objects and environments. The key rationale behind our data collection strategy is that we want to represent both common and rare human poses that might be missed when simply collecting more images without aiming for good coverage. To this end, we use a two-level hierarchy of human activities proposed in (Ainsworth *et al.*, 2011) to guide the collection process. This hierarchy was developed for the assignment of standardized energy levels during physical activity surveys and includes 823 activities in total of 21 different activity categories. The activities at the first level of the hierarchy correspond to thematically related groups of activities such as "Home Activities", "Lawn and Garden" or "Sports". The activities at the second level then correspond to individual activities such as "Washing windows", "Picking fruit" or "Rock climbing". Note that using the activity hierarchy for collection has an additional advantage that all images have an associated activity label. As a result one can assess and analyze any performance measure also on subsets of activities or activity categories.

Due to the coverage of the hierarchy the images in our dataset are representative of the diversity of human poses, overcoming one of the main limitations of previous collections. In Fig. 9.2 we visualize this diversity by comparing upper body annotations of the "Armlets" dataset Fig. 9.2(b) and our proposed dataset (c). Note that although "Armlets" contain about 13,500 images, the annotations resemble a person with arms down along the torso (distribution of red, cyan, green, and blue sticks).

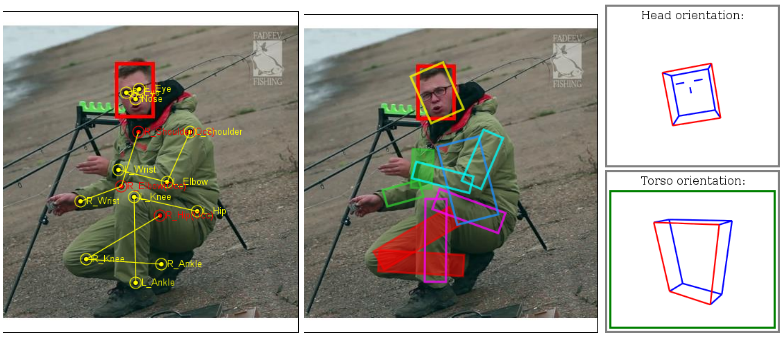
We collect images from YouTube using queries based on the activity descriptions. Using YouTube allows us to access a rich collection of videos originating from various sources, including amateur and professional recordings and capturing a variety of public events and performances. In Fig. 9.2 (c) we show the distribution of upper body poses on our dataset. Note the variability in the location of hands and the absence of distinctive peaks for the upper and lower arms that are present in the case of the "Armlets" dataset.

Data collection. As a first step of the data collection we manually query YouTube using descriptions of activities from (Ainsworth *et al.*, 2011). We select up to 10 videos for each activity filtering out videos of low quality and those that do not include people. This resulted in $3,913$ videos spanning 410 different activities. Note that we merged a number of the original 823 activities due to high similarity between them, such as cycling at different speeds. In the second step we manually pick several frames with people from each video. As the focus of our benchmark is pose



(a) (b) (c)

Figure 9.2: Visualization of upper body pose variability. From left to right we show, (a) color coding of the body parts (b) annotations of the "Armlets" dataset (Gkioxari *et al.*, 2013), and (c) annotations of this dataset.



(a) (b) (c)

Figure 9.3: Example of the provided annotations. Annotated are (a) positions and visibility of the main body joints, locations of the eyes and nose and the head bounding box (occluded joints are shown in red), (b) occlusion of the main body parts (occluded parts are shown with filled rectangles), and (c) 3D viewpoints of the head and torso. On the illustration the viewpoint is shown using a simplified body model, the front face of the model is shown in red.

estimation we do not include video frames in which people are severely truncated or in which pose is not recognizable due to poor image quality or small scale. We aim to select frames that either depict different people present in the video or the same person in a substantially different pose. In addition we restrict the selected frames to be at least 5 seconds apart. This step resulted to a total of 24,920 extracted frames from all collected videos. Next, we annotate all people present in the collected images, but ignore dense people crowds in which significant number of people are almost fully occluded. Following this procedure we collect images of 40,522 people. We allocate roughly tree quarters of the collected images for training and use the rest for testing. Images from the same video are either all in the training or all in the test set. This results in a training/test set split of 28,821 to 11,701.

Data annotation. We provide rich annotations for the collected images, an example can be seen in Fig. 9.3. Annotated are the body joints, 3D viewpoint of the head and torso, and position of the eyes and nose. Additionally for all body joints and parts visibility is annotated. Following (Johnson and Everingham, 2011; Gkioxari *et al.*, 2013) we annotate joints in a "person centric" way, meaning that the left/right joints refer to the left/right limbs of the person. At test time this requires pose estimation with both a correct localization of the limbs of a person along with the correct match to the left/right limb. The annotations are performed by in-house workers and via Amazon Mechanical Turk (AMT). In our annotation process we build and extend the annotation tools described in (Maji, 2011). Similarly to (Johnson and Everingham, 2011; Vondrick *et al.*, 2012) we found that effective use of AMT requires careful selection of qualified workforce. We pre-select AMT workers based on a qualification task, and then maintain data quality by manually inspecting the annotated data.

Experimental protocol and evaluation metrics. We define the baseline evaluation protocol on our dataset following the current practices in the literature (Johnson and Everingham, 2011; Gkioxari *et al.*, 2013; Sapp and Taskar, 2013). We assume that at test time the rough location and scale of a person are known, and we exclude the cases with multiple people in close proximity to each other from the evaluation. We feel that these simplifications are necessary for the rapid adoption of the dataset as the majority of the current approaches does not address multiple people pose estimation and does not search over people positions and scales.

We consider three metrics as indicators for the pose estimation performance. The widely adopted "PCP" metric (Ferrari *et al.*, 2008) that considers a body part to be localized correctly if the estimated body segment endpoints are within 50% of the ground-truth segment length from their true locations. The "PCP" metric has a drawback that foreshortened body parts should be localized with higher precision to be considered correct. We define a new metric denoted as "PCPm" that uses 50% of the mean ground-truth segment length over the entire test set as a matching threshold, but otherwise follows the definition of "PCP". Finally, we consider the "PCK" metric from (Yang and Ramanan, 2013) that measures accuracy of the localization of the body joints. In (Yang and Ramanan, 2013) the threshold for matching of the joint position to the ground-truth is defined as a fraction of the person bounding box size. We use a slight modification of the "PCK" and define the matching threshold as 50% of the head segment length. We denote this metric as "PCKh". We choose to use head size because we would like to make the metric articulation independent.

9.3 ANALYSIS OF HUMAN POSE ESTIMATION APPROACHES

In this section we analyze the performance of prominent human pose estimation approaches on our benchmark. We take advantage of our rich annotations and conduct a detailed analysis of various factors influencing the results, such as foreshortening,

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Upper body	Full body
(Gkioxari *et al.*, 2013)	51.3	-	-	28.0	12.4	-	26.4	-
(Sapp and Taskar, 2013)	51.3	-	-	27.4	16.3	-	27.8	-
(Yang and Ramanan, 2013)	61.0	36.6	36.5	34.8	17.4	70.2	33.1	38.3
Pishchulin et al. (Chapter 8)	63.8	39.6	37.3	39.0	26.8	70.7	39.1	42.3
(Gkioxari *et al.*, 2013) + loc	65.1	-	-	33.7	14.9	-	32.4	-
(Sapp and Taskar, 2013) + loc	65.1	-	-	32.6	19.2	-	33.7	-
(Yang and Ramanan, 2013) + loc	**67.2**	39.7	39.4	37.4	18.6	**75.7**	35.8	41.4
Pishchulin et al. (Chapter 8) + loc	66.6	**40.5**	**38.2**	**40.4**	**27.7**	74.5	**40.6**	**43.9**

Table 9.1: Pose estimation results (PCPm) on the proposed dataset without and with using rough body location ("+ loc" in the table).

activity and viewpoint, previously not possible in this detail. The goal of this analysis is to evaluate the robustness of prominent approaches in various challenges for articulated pose estimation, identify the existing limitations and stimulate further research advances.

In our analysis we consider two full body and two upper body pose estimation approaches. The full body approaches are the version 1.3 of the *Flexible Mixture of Parts* (FMP) approach of Yang and Ramanan (Yang and Ramanan, 2013) and our best *Poselet Conditioned Pictorial Structures* model proposed in Chapter 7 and extended in Chapter 8 (we refer to our method as *Pishchulin et al.* in the rest of this chapter). The upper body pose estimation approaches are the *Multimodal Decomposable Models* (MODEC) approach of Sapp et al. (Sapp and Taskar, 2013) and the *Armlets* approach of Gkioxari et al. (Gkioxari *et al.*, 2013). In case of FMP and MODEC we use publicly available code and pre-trained models. In case of the Armlets model, the code and pre-trained model provided by the authors correspond to the version from (Gkioxari *et al.*, 2013) that includes the HOG features only. The performance of our version of Armlets on the "Armlets" dataset is 3.3 PCP lower than the version based on combination of all features.[12] In the following experiments we use "PCPm" as our working metric, while also providing overview results for "PCP" and "PCPh". While we observe little performance differences when using each metric, all conclusions obtained during "PCPm"-based evaluation are valid for "PCP" and "PCKh"-based evaluations as well.

Overall performance evaluation. We begin our analysis by reporting the overall pose estimation performance of each approach and summarize the results in Tab. 9.1. We include both upper- and full body results to enable comparison across different models. The PS approach achieves the best result of 42.3% PCPm, followed by the FMP approach with 38.3% PCPm. On the upper body evaluation, PS performs best with 39.1%, while both MODEC (27.8% PCPm) and Armlets (26.4% PCPm) perform significantly worse.

---

[12] See Tab.1 in (Gkioxari *et al.*, 2013) for the comparison.

Setting	PCP		PCPm		PCKh	
	Upper body	Full body	Upper body	Full body	Upper body	Full body
(Gkioxari *et al.*, 2013)	26.2	-	26.4	-	25.9	-
(Sapp and Taskar, 2013)	27.5	-	27.8	-	27.9	-
(Yang and Ramanan, 2013)	32.1	37.8	33.1	38.3	34.9	37.7
Pishchulin et al. (Chapter 8)	**38.3**	**42.2**	**39.1**	**42.3**	**38.6**	**41.2**

Table 9.2: Summary of pose estimation results using various metrics.

The interesting outcome of this comparison is that both upper body approaches MODEC and Armlets are outperformed by the full body approaches PS and FMP evaluated on upper body only. This is interesting because significant portion of the dataset (15 %) includes people that have only upper body visible. It appears that the PS and FMP approaches are sufficiently robust to missing parts to produce reliable estimates even in the case of lower body occlusion.

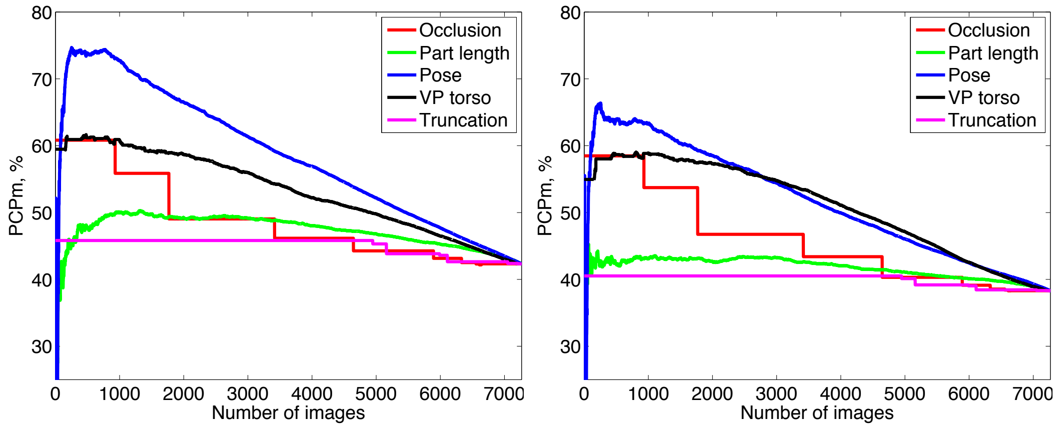
Lower part of Tab. 9.1 shows the results when using provided rough location of person during test time inference. We observe, that while the performance increases for all methods, upper body approaches profit at most, as they heavily depend on correct torso localization. For the sake of fair comparison among the methods, we *do not* use the rough location in the following experiments. Another interesting outcome is that the achieved performance is substantially lower than the performance on the sports-centric LSP dataset, but comparable to results on the "Armlets" dataset (42.2% PCP on our benchmark vs. 69.2% on LSP (Chapter 8) vs. 36.2% PCP on "Armlets"). This suggests that sport activities are not necessary the most difficult cases for pose estimation; challenges such as appearance variability, occlusion and truncation apparently deserve more attention in the future.

Overview of pose estimation results using "PCP" and "PCKh" metrics and comparison to "PCPm" metric is shown in Tab. 9.2. We observe slight performance differences when using various evaluation metrics. However, it can be seen that the relating ranking of methods is preserved.

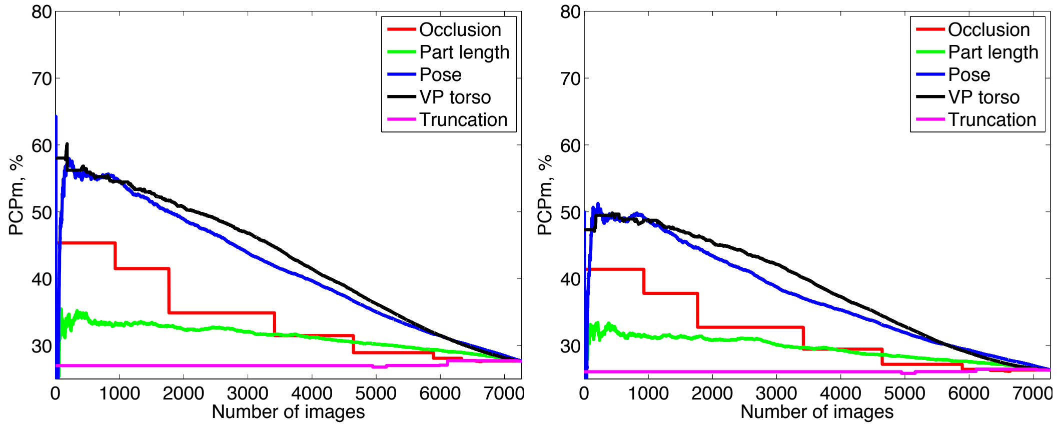
9.3.1 Analysis of pose estimation challenges

We now analyze the performance of each approach with respect to the following five factors: part occlusion, foreshortening, body pose, viewpoint, and activity of the person. For the purpose of this analysis we define quantitative complexity measures that map body image annotations to a real value that relates to the complexity of the image with respect to each factor.

Let us denote the annotation of the person by $L = \{L^{pose}, L^{view}, L^{vis}\}$, where $L^{pose} = \{l_i, i = 1, \ldots, N\}$ corresponds to the positions of body parts, $L^{view} = \{\alpha_1, \alpha_2, \alpha_3\}$ are the Euler angles representation of the torso rotation, and $L^{vis} = \{(\rho_i, \theta_i), i = 1, \ldots, N\}$ encodes body part visibility via a set of occlusion labels



Pishchulin et al. (Chapter 8) (Yang and Ramanan, 2013)



(Sapp and Taskar, 2013) (Gkioxari *et al.*, 2013)

Figure 9.4: Performance (PCPm) as a function of the five complexity measures.

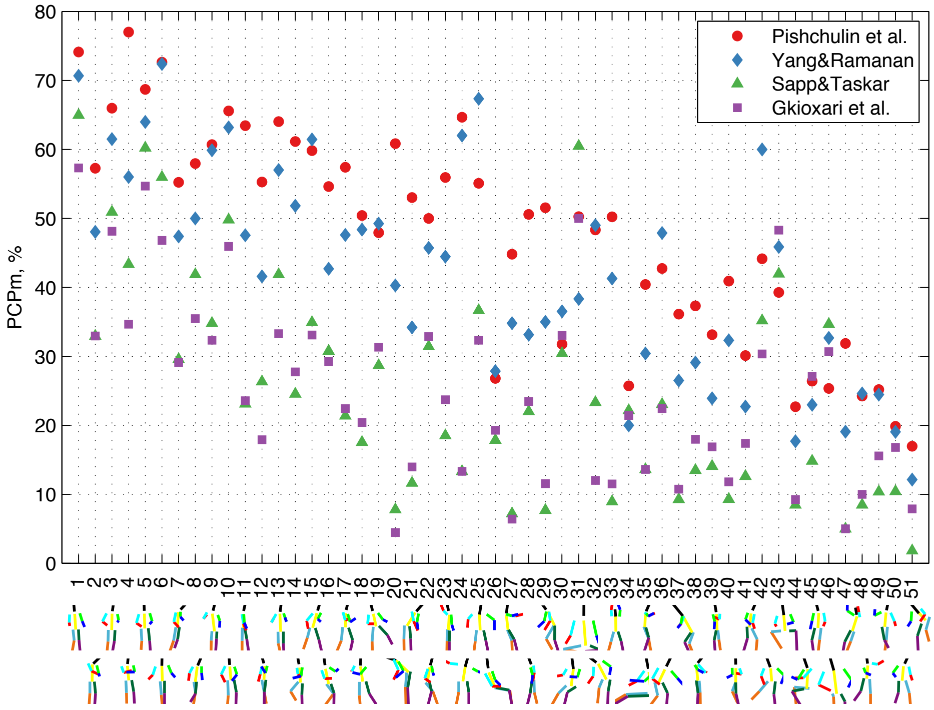
$\rho_i \in \{0, 1\}$ and truncation labels $\theta_i \in \{0, 1\}$.

We define the following complexity measures. Pose complexity is measured as the deviation from the mean pose on the entire dataset. We define $\mathrm{m}_{pose}(L) = \prod_{(i,j)\in E} p_{ps}(l_i|l_j)$, where E is a set of body joints and $p_{ps}(l_i|l_j)$ is a Gaussian distribution measuring relative position of the two adjacent body parts using the transformed state-space representation introduced in (Felzenszwalb and Huttenlocher, 2005). Note that $\mathrm{m}_{pose}(L)$ corresponds to the likelihood of the pose under the tree structured pictorial structures model (Felzenszwalb and Huttenlocher, 2005). The amount of foreshortening is measured by $\mathrm{m}_f(L) = \sum_{i=1}^{N} |d(l_i) - m_i|/m_i$, where $d(l_i)$ is the length of the body part i, and m_i is the mean length over the entire dataset. The viewpoint complexity is measured by the deviation from the frontal viewpoint: $\mathrm{m}_v(L) = \sum_{i=1}^{3} \alpha_i$. Finally, the amount of occlusion and truncation correspond to the number of occluded and truncated body parts: $\mathrm{m}_{occ} = \sum_{i=1}^{N} \rho_i$, and $\mathrm{m}_t = \sum_{i=1}^{N} \tau_i$.

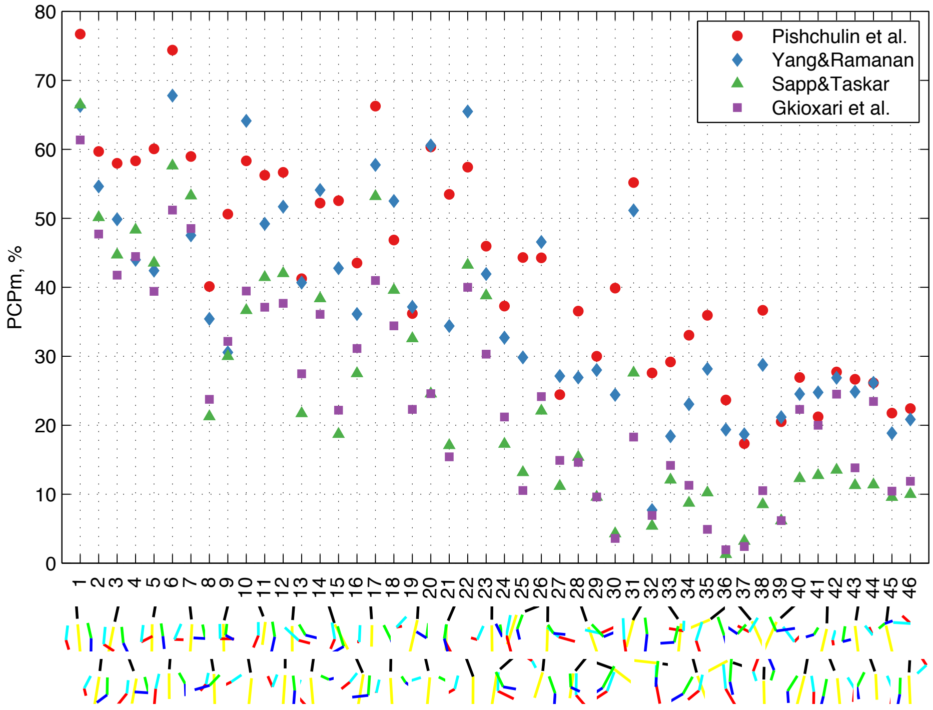
Performance as a function of the complexity measures. To visualize the influence of the various factors on pose estimation performance we plot PCPm scores for the images sorted in the order of increasing complexity (see Fig. 9.4). In general and as expected, the performance drops for all measures as the complexity increases. There are interesting differences however. Body pose complexity clearly influences the performance of all approaches the most. The second most influential factor is the viewpoint of the torso. For upper body pose estimation approaches this factor is equally influential as body pose. The third most influential factors is occlusion while for the full body estimation approaches this is equally influential as the torso orientation. Contrary to our expectation we found that the part length is less influential. Part length and in particular foreshortening effects are considered to be the key difficulties for both pose estimation. Based on this analysis the above mentioned factors have a higher influence on the performance. The least influential factor is truncation having the smallest effect. In the case of upper body estimation the performance even slightly increases as the amount of truncation increases due to two factors. As truncation if more likely for the lower body these approaches suffer less from truncation and also truncated poses are biased towards frontal views for which the methods are more suited. We now discuss and analyze each factor in more detail.

Body pose performance. As stated above the complexity of the pose is a dominating factor for the performance of all considered approaches. For example the PS approach achieves 72.8% PCPm on the 1000 images with lowest pose complexity, compared to 42.3% for the entire dataset. The same is true for the FMP model, 63.4% PCPm on 1000 least pose complex images vs. 38.3% overall.

To highlight variations in performance across different body configurations we cluster the test images according to the body pose and measure performance for each cluster. We repeat this two times, clustering all body joints and the upper body joints only. In the latter case we measure performance on the upper body parts only. These two clusterings correspond both to different types of challenges as well as applications. Furthermore, this allows to directly compare full vs. upper body techniques. We show the average PCPm for all pose clusters with more than 25 examples in Fig. 9.5 ordering the results from left to right by increasing mean pose complexity. We now analyze the results for full body clusters (Fig. 9.5 (a)), while also providing the performance for upper body clusters in Fig. 9.5 (b). Note the significant variations in performance across different full body clusters. For example, results on full body clusters vary between 77% and 2% PCPm. The best performance is achieved on full body clusters with poses similar to the mean pose e.g. clusters 1 and 5 (see Figure 9.5 (a)). Examining clusters with poor performance we immediately discover several failure modes of PS and FMP approaches. Consider the clusters 42 and 43 that correspond to people with slightly foreshortened torso. FMP improves over PS by 14% PCPm on cluster 25 (54% PCPm for PS vs. 68% PCPm for FMP) and by 16% PCMm on cluster 42 (44% PCPm for PS vs. 60% PCPm for FMP), as it can better model torso foreshortening by representing torso as configuration of



(a) full body



(b) upper body

Figure 9.5: Performance (PCPm) on images clustered by (a) full body and (b) upper body pose. Clusters are ordered by increasing mean pose complexity and representatives are shown beneath.

multiple flexible parts, whereas PS models torso as a single rigid part. Also, the flexibility of FMP model accounts for its better performance on frontal sitting people (cluster 43) where FMP improves over PS by 7% PCPm (46% PCPm for FMP vs. 39% PCPm for PS), mainly due to better modeling of the foreshortened upper legs. However, performance on the side-view sitting people (e.g. clusters 26, 30, 34, 44) is poor for all methods. Another prominent failure mode for all approaches are people facing away from the camera, e.g. cluster 50. Such part configurations are commonly mistaken for the frontal view which leads to a mismatch between left and right body parts resulting in incorrect estimation. These findings demonstrate inability of current methods to reliably discriminate between frontal and backward views of people. Interestingly, upper body approaches outperform full body methods on the full body cluster 31. This is an easy case for the former group of methods due to frontal upright upper body, but is a challenging task for the full body approaches as legs are hard to estimate in this case. However, both MODEC and Armlets fail on examples when torso start deviating from canonical orientation (e.g. clusters 20, 27, 37). At the same time both full body methods perform better, as they are more robust to the viewpoint changes. Surprisingly, full body methods outperform upper body approaches on "easy" examples (c.f. cluster 1, 3 and 5). We attribute this effect to the correct integration of signals from the legs into a more reliable upper body estimate.

Occlusion and truncation performance. In Fig. 9.4 we clearly see difference in how occlusion and truncation influences the results. As expected we observe that the performance is best for fully visible people, but full visibility does not result in success rate similar to the one we observed for the images with simple poses, e.g. PS approach achieves 72.8% PCPm for 1000 most simple poses vs. 60% PCPm for same amount of people with least occlusion. We observe that occlusion results in significant performance drop on the order of 10% PCPm, e.g. in the case of PS approach 19.3% vs. 31.2% PCPm for the forearm with and without occlusion.

As mentioned above, truncation showed the least influence overall among the discussed factors. There are at least two reasons. First, the number of images with truncation is limited in our dataset (about 30% of the test data contain truncated people). Second, and more importantly, for truncation one cannot annotate positions of body parts outside of the image. Therefore the standard procedure is to exclude truncated body parts from the evaluation. In that sense approaches that wrongly estimate the position of a truncated body part are not punished for that. This limitation could be addressed by requiring that models have to also report which parts of the body are truncated.

Viewpoint performance. We evaluate the pose estimation for various torso viewpoints in two ways. In Fig. 9.4 we show results using our standard analysis method based on images ordered by deviation from the frontal viewpoint. For a more detailed analysis we quantize the space of viewpoints by clustering training examples according to their 3D torso orientations. We show results for the viewpoint clusters in Fig. 9.6 ordering them by the number of examples corresponding to each cluster. The number of examples per cluster ranges between 1453 examples for the largest cluster corresponding to the frontal viewpoint, and 53 examples for the viewpoint with extreme torso tilt.

We observe that in contrast to the full body approaches, viewpoint has profound influence on the performance of the upper body approaches considered in our evaluation. The performance of both Armlets and MODEC approaches drops significantly for non-frontal views.

A per viewpoint evaluation reveals significant performance differences across viewpoints. In Fig. 9.6 we show the results for the "person centric" annotations that we use throughout experiments in this chapter and in addition for the "observer centric" (OC) annotations, in which body limbs are labeled as left/right based on their image location with respect to the torso. Frontal and near-frontal viewpoints are performing best. We observe a large drop in performance for backward facing people when performance is measured in "person centric" manner, which suggests that large portion of incorrect pose estimates for backward views is due to incorrect matching of left/right limbs.

We observe that all approaches handle extreme viewpoints poorly. PS approach is the only one in our evaluation that gracefully handles in-plane rotations (cluster 12), whereas performance of other approaches significantly degrades in that case. Also, PS outperforms other methods in case of extreme torso tilts (e.g. cluster 14). The performance for clusters with extreme torso rotation is on the level of 20 - 30% PCPm for the best method, corresponding to only 2 - 3 out of 10 body parts being localized correctly for such viewpoints.

Part length performance. Fig. 9.4 also shows the influence of part length on the performance of each approach. In this context, foreshortening is the most influential aspect and considered an important challenge for articulated pose estimation. The key observation is that the presence or absence of foreshortening has relatively little influence on the result compared to the other factors such as pose and occlusion. The best performing PS model is the most robust to foreshortening compared to other three approaches. For example the performance for the first 4000 images ordered by increasing foreshortening remains nearly constant.

Activity performance. Finally, we evaluate pose estimation performance as a function of the person activity. To that end we group test images by the activity categories in the hierarchy used for the image collection (Ainsworth *et al.*, 2011) and compute PCPm for each category. The results are shown in Fig. 9.7, where we order categories from left to right according to the number of test examples.

We observe strong variation of performance for different activity types. Best results are obtained on the sports- and dancing-centric activities (e.g. "Sports", "Running", "Winter Activities" and "Dancing"). Most difficult turn out to be activities that are performed in bulky clothing and involve use of tools (e.g. "Home Repair") and activities performed in cluttered scenes (e.g. "Fishing and Hunting"). MODEC

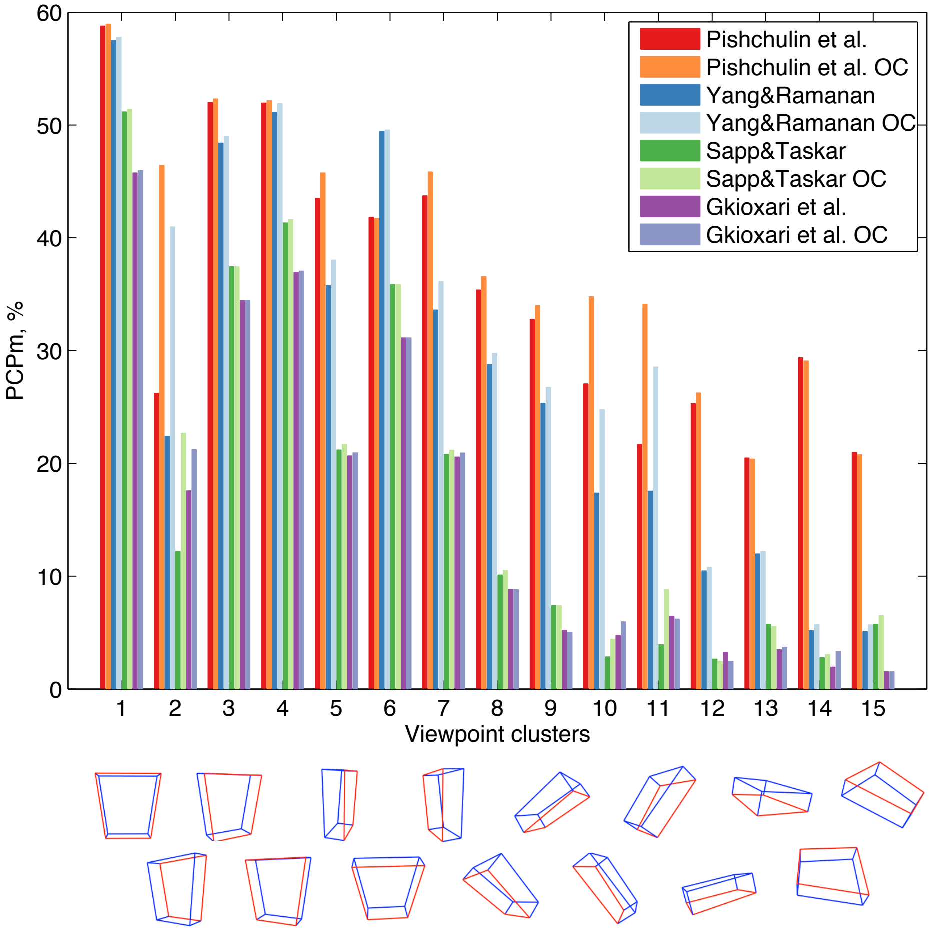


Figure 9.6: Pose estimation results (PCPm) grouped by viewpoint. Viewpoint clusters ordered decreasingly w.r.t. number of images. Each cluster is visualized in bottom row using 3D model of the torso corresponding to the cluster medoid.

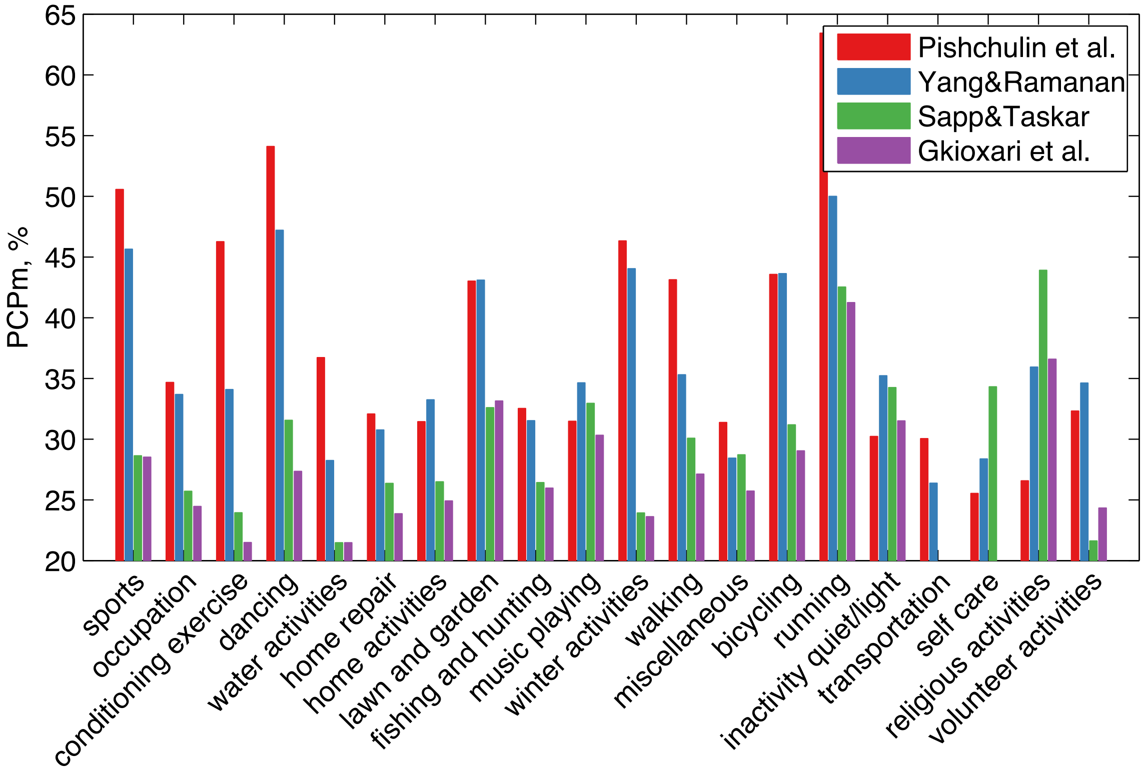


Figure 9.7: Pose estimation results (PCPm) grouped by activity categories shown in decreasing order w.r.t. number of images.

Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Upper body	Full body
(Yang and Ramanan, 2013)	61.0	36.6	36.5	34.8	17.4	70.2	33.1	38.3
(Yang and Ramanan, 2013) retrained	**69.3**	39.5	38.8	**43.4**	27.7	74.6	**42.3**	44.7
Pishchulin et al. (Chapter 8)	63.8	39.6	37.3	39.0	26.8	70.7	39.1	42.3
Pishchulin et al. (Chapter 8) retrained	68.4	**42.7**	**42.8**	42.0	**29.2**	**76.3**	42.1	**46.1**

Table 9.3: Comparison of performance (PCPm) before and after retraining. For PCKh results see supplementary material.

outperforms all other approaches on the "Self care" activities (examples of activities from this category are "Eating, sitting", "Hairstyling", "Grooming" etc. with "Eating, sitting" containing by far the largest number of images.)

Retrained models. To showcase the usefulness of the benchmark as an analysis tool we retrain the PS and FMP models on the training set from our benchmark. To speed up training we consider a subset of 4000 images, which is 4 times as many images as in the LSP and 40 times as many as in the Image Parsing datasets used by the publicly available PS and FMP models. The results are shown in Tab. 9.3. FMP significantly benefits from retraining (44.7% PCPm for retrained vs. 38.3% for original). PS achieves slightly better result, although overall improvement due to retraining is smaller (46.1% PCPm for retrained vs. 42.3% PCPm the original).

Although performances for FMP and PS are close overall, we observe interesting differences when examining performance at the level of individual activities and viewpoints (thereby exploiting the rich annotations of our benchmark). Results are shown in Fig. 9.8. We observe that our publicly available PS model is winning by a large margin on the highly articulated categories, such as "Dancing" and "Running". Retraining the model boosts performance on activities with less articulation but more complex appearance (e.g. "Home Activities", "Lawn and Garden", "Bicycling", and "Occupation"). Our results show that training on the larger amount of more variable data significantly improved robustness of FMP to viewpoint changes. Performance of FMP improves on the difficult viewpoints by a large margin (e.g. for viewpoint cluster 10 improvement is from 17% to 31% PCPm). Retraining improves the performance of PS model on difficult viewpoints as well, although not as dramatically as for FMP, likely because PS already models in-plane rotations explicitly.

9.3.2 Evaluation of Deep Learning based Methods

We perform evaluation of the recent deep learning based methods using the proposed PCKh evaluation measure. In particular, we consider following approaches from the literature: fully-convolutional part detection approach of (Tompson *et al.*, 2014) who jointly train part detectors with a simple spatial model; their follow-up method (Tompson *et al.*, 2015) that uses an additional refinement stage increasing the localization accuracy; holistic deep learning based approach (Carreira *et al.*, 2016)

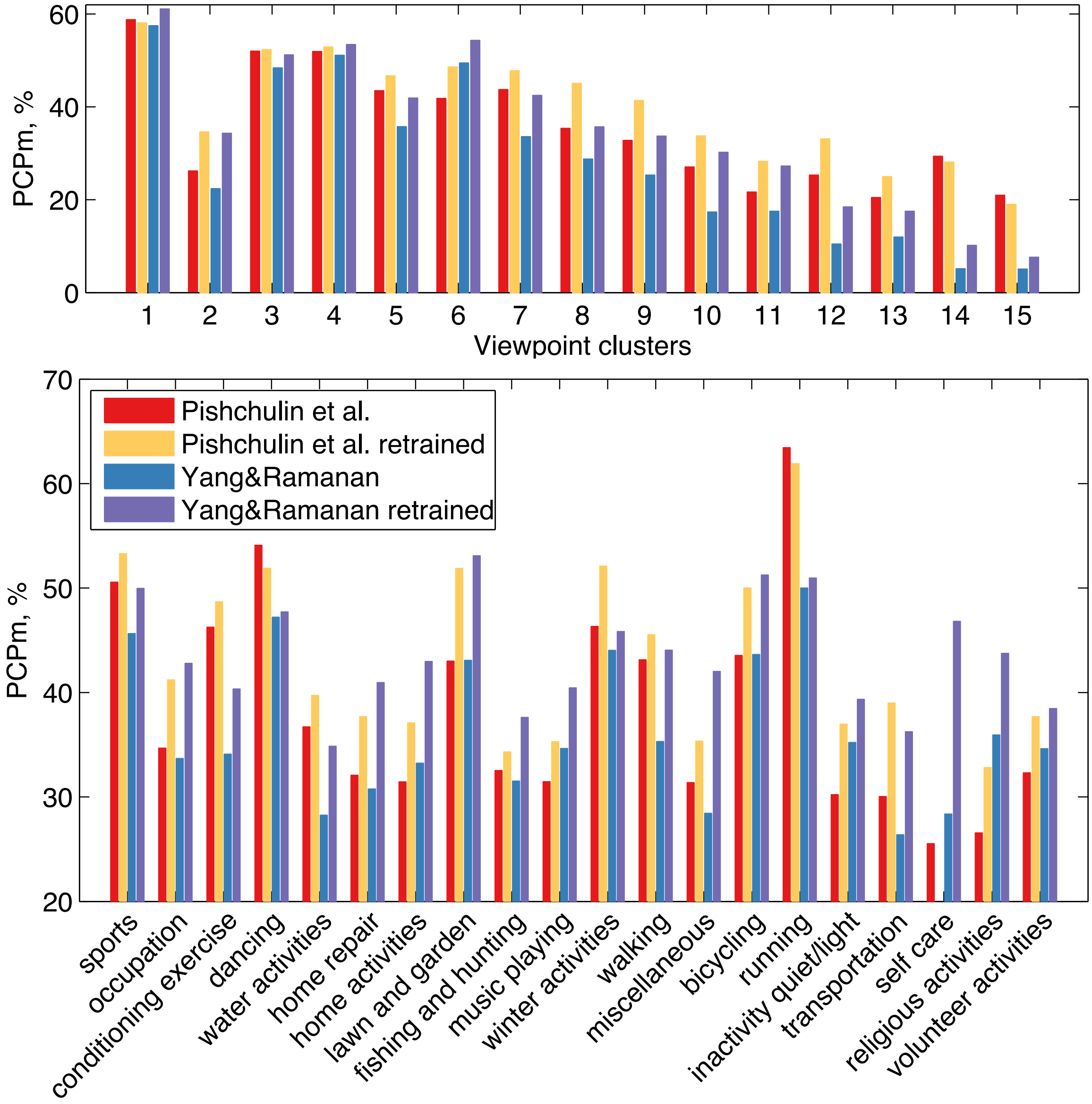


Figure 9.8: Comparison of performance (PCPm) on viewpoint (top) and activity category clusters (bottom) before and after retraining. See Fig. 9.6 for visualization of the viewpoint clusters.

Setting	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Full Body
Pishchulin et al. (Chapter 8) retrain.	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
DeepCut (Chapter 11)	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
DeeperCut (Insafutdinov *et al.*, 2016)	96.6	94.6	88.5	**84.4**	87.6	**83.9**	**79.4**	88.3
(Tompson *et al.*, 2014)	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
(Carreira *et al.*, 2016)	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
(Tompson *et al.*, 2015)	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
(Hu and Ramanan, 2016)	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
(Wei *et al.*, 2016)	**97.8**	**95.0**	**88.7**	84.0	**88.4**	82.8	**79.4**	**88.5**

Table 9.4: Pose estimation performance (PCKh) by deep learning based approaches.

that uses iterative error feedback strategy to progressively refine initially predicted locations of individual body parts; recent bidirectional architecture (Hu and Ramanan, 2016) that combines bottom-up reasoning with top-down feedback, where neural units are influenced by both lower and higher-level units; and recent Convolutional Pose Machines approach (Wei *et al.*, 2016) that incorporates a convolutional network architecture into the pose machine framework. In addition, we provide the performance by our *DeepCut* approach introduced in Chapter 11 and by our follow-up *DeeperCut* method (Insafutdinov *et al.*, 2016), and compare the results to our *PS* approach (Chapter 8) relying on hand-crafted image representations. We report per-part PCKh results in Tab. 9.4 and results for the entire range of distance thresholds in Fig. 9.9. All deep learning based approaches significantly outperform *PS*, as they are able to learn much richer image representations. (Tompson *et al.*, 2015) outperforms their previous method (Tompson *et al.*, 2014) (82.0% vs. 79.6% PCKh) due to additional body part location refinement that increases localization accuracy. The difference is more prominent for smaller distance thresholds (Fig. 9.9). (Tompson *et al.*, 2015) also outperforms holistic approach (Carreira *et al.*, 2016). This is due to the fact that fully-convolutional multi-scale architecture of the former allows for more precise body part localization compared to the holistic part location prediction of the latter. Performance difference becomes much clear when analyzing PCKh curves for smaller distance thresholds in Fig. 9.9. For instance, for PCKh @ 0.2 the difference is over 12% PCKh. Our *DeepCut* approach proposed in Chapter 11 slightly outperforms (Tompson *et al.*, 2015) (82.4 vs. 82.0% PCKh), while the differences are more pronounced for smaller thresholds due to better part localization. This is interesting, as in contrast to their method that relies on multi-resolution filter banks and trains an additional CNN for location refinement, *DeepCut* operates on single resolution and employs a much simpler strategy by optimizing a joint classification and location refinement objective function during CNN training. *DeepCut* also performs on par with the recent bidirectional approach (Hu and Ramanan, 2016) at maximum PCKh threshold, but significantly outperforms at lower distance thresholds. (Wei *et al.*, 2016) outperforms *DeepCut* (88.5 vs. 82.4% PCKh): it encodes more contextual information into the part detectors via the multi-stage training

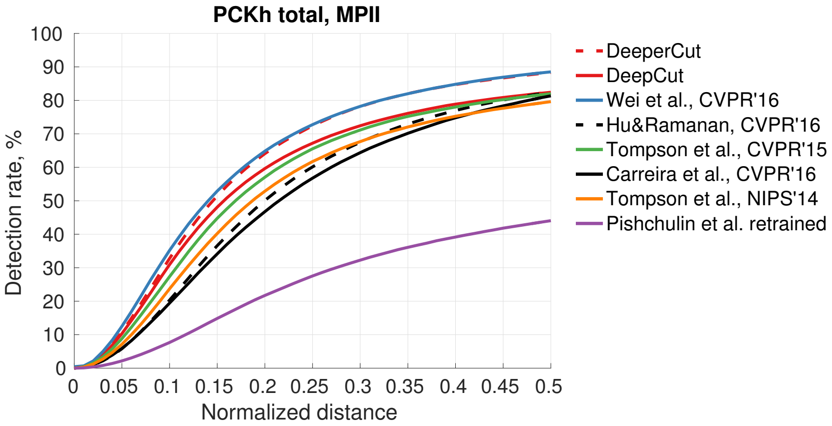


Figure 9.9: Pose estimation results over all PCKh distance thresholds.

procedure. Relying on deeper part detectors and expressive image conditioned pairwise terms in *DeeperCut* allows for on-par performance with (Wei *et al.*, 2016) while using much simpler sinle-stage single-scale architecture.

9.4 CONCLUSION

In this chapter we advanced the state of the art in human pose estimation by establishing new qualitatively higher standards for evaluation and analysis of pose estimation methods and demonstrate the most promising research directions for the next years. To that end, we proposed a novel "MPII Human Pose" benchmark that we collected by leveraging a taxonomy of literature. Compared to other datasets our benchmark covers significantly wider range of human poses spanning from householding to recreational activities and sports. Rich labeling of the collected data and a set of developed evaluation tools enable comprehensive analysis which we perform to demonstrate the strengths and weaknesses of the competing approaches.

Our findings indicate that human pose estimation methods are significantly challenged by cases outside their comfort zone, such as large torso rotation and loose clothing. From all other factors, pose complexity has the most profound effect on the pose estimation performance. Current methods perform best on activities with simple tight clothing (e.g. in sport scenes), and are challenged by images with complex clothing and background clutter that are typical for many occupational and outdoor activities.

We made the data, annotations for training images and evaluation tools publicly available in order to enable detailed analysis of future pose estimation methods. To prevent accidentally tuning on the test set, we withheld the annotations for the test images and perform evaluation on demand using the developed performance analysis tools.

In the next chapter we use this comprehensive benchmark to analyze the performance of prominent human activity recognition methods on the task of fine-graned activity recognition with hundreds of everyday activities.

10

FINE-GRAINED ACTIVITY RECOGNITION WITH HOLISTIC AND POSE BASED FEATURES

Contents

10.1 Introduction **168**
10.2 Dataset **169**
10.3 Methods **170**
    10.3.1 Dense trajectories (DT) 170
    10.3.2 Pose-based methods 170
    10.3.3 Combinations of holistic and body based methods 171
10.4 Analysis of activity recognition performance **171**
    10.4.1 Overall performance 172
    10.4.2 Analysis of activity recognition challenges 174
    10.4.3 Detailed analysis on a subset of activities 175
10.5 Conclusion **177**

In this chapter we switch our focus to analyzing the performance of popular human activity recognition methods. Holistic methods based on dense trajectories (Wang *et al.*, 2013; Wang and Schmid, 2013) have been de facto standard for recognition of human activities in video. On the other hand, at the high level human activities can often be accurately characterized in terms of body pose, motion, and interaction with scene objects (Jhuang *et al.*, 2013). In this chapter we aim to clarify the underlying factors responsible for good performance of holistic and pose based representations. To that end, we build on our comprehensive dataset introduced in Chapter 9 leveraging the existing taxonomy of human activities. This dataset includes 24, 920 video snippets covering 410 human activities in total. Our analysis reveals that holistic and pose based methods are highly complementary, and their performance varies significantly depending on the activity. We find that holistic methods are mostly affected by the number and speed of trajectories, whereas pose based methods are mostly influenced by viewpoint of the person. We observe striking performance differences across activities: for certain activities results with pose based features are more than twice as accurate compared to holistic features, and vice versa. The best performing approach in our comparison is based on the combination of holistic and pose based approaches, which again underlines their complementarity.

10.1 INTRODUCTION

In this chapter we consider the task of human activity recognition in realistic videos such as feature movies or videos from YouTube. We specifically focus on how to represent activities for the purpose of recognition. Various representations were proposed in the literature, ranging from low level encoding using point trajectories (Wang *et al.*, 2013; Wang and Schmid, 2013) to higher level representations using body pose trajectories (Jhuang *et al.*, 2013) and collection of action detectors (Sadanand and J., 2012). At the high level human activities can often be accurately characterized in terms of body pose, motion, and interaction with scene objects. Representations based on such high level attributes are appealing as they allow to abstract the recognition process from nuisances such as camera viewpoint or person clothing, and facilitate sharing of training data across activities. However, articulated pose estimation is a challenging and non-trivial task that is subject of ongoing research (Yang and Ramanan, 2013; Dantone *et al.*, 2013; Sapp and Taskar, 2013; Tompson *et al.*, 2014; Wei *et al.*, 2016). Therefore, most popular methods in activity recognition have been ralying on holistic representations (Laptev *et al.*, 2008; Duchenne *et al.*, 2009; Wang *et al.*, 2013; Wang and Schmid, 2013) that extract appearance and motion features from the entire video and leverage discriminative learning techniques to identify information relevant for the task.

Results on the JHMDB dataset (Jhuang *et al.*, 2013) suggest that recent pose estimation methods might have reached sufficient accuracy to be effective for activity recognition. Motivated by these results, we further explore holistic and pose based representations aiming for much broader scale and coverage of activity classes. To that end we employ our recent "MPI Human Pose" dataset introduced in Chapter 9. Compared to 21 activity classes considered in (Jhuang *et al.*, 2013) the "MPI Human Pose" dataset includes 410 activities and more than an order of magnitude more images ($\sim$ 32K in JHMDB vs. over 1.5M images in "MPI Human Pose"). "MPI Human Pose" aims to systematically cover a range of activities using an existing taxonomy (Ainsworth *et al.*, 2011). This is in contrast to existing datasets (Kuehne *et al.*, 2011; Soomro *et al.*, 2012) that typically include ad-hoc selections of activity classes. Using the rich labeling of people provided with "MPI Human Pose" we evaluate the robustness of holistic and pose based representations to factors such as body pose, viewpoint, and body-part occlusion, as well as to the number and speed of dense trajectories covering the person.

This chapter makes the following contributions. First, we perform a large-scale comparison of holistic and pose based features on the "MPI Human Pose" dataset. Our results complement the findings in (Jhuang *et al.*, 2013), indicating that pose based features indeed outperform holistic features for certain cases. However we also find that both types of features are complementary and their combination performs best. Second, we analyze factors responsible for success and failure, including number and speed of trajectories, occlusion, viewpoint and pose complexity.

10.2 DATASET

In order to analyze the challenges for fine-grained human activity recognition, we build on our recent publicly available "MPI Human Pose" dataset (Chapter 9). The dataset was collected from YouTube videos using an established two-level hierarchy of over 800 every day human activities. The activities at the first level of the hierarchy correspond to thematic categories, such as "Home repair", "Occupation", "Music playing", etc., while the activities at the second level correspond to individual activities, e.g. "Painting inside the house", "Hairstylist" and "Playing woodwind". In total the dataset contains 20 categories and 410 individual activities covering a wider variety of activities than other datasets, while its systematic data collection aims for a fair activity coverage. Overall the dataset contains $24,920$ video snippets and each snippet is at least 41 frames long. Altogether the dataset contains over a 1M frames. Each video snippet has a key frame containing at least one person with a sufficient portion of the body visible and annotated body joints. There are $40,522$ annotated people in total. In addition, for a subset of key frames richer labels are available, including full 3D torso and head orientation and occlusion labels for joints and body parts.

Static pose estimation complexity measures. In addition to the dataset, in Chapter 9 a set of quantitative complexity measures aiming to asses the difficulty of pose estimation in each particular image was proposed. These measures map body image annotations to a real value that relates the complexity of each image w.r.t. each factor. These complexity measures are listed below.

1. *Pose*: deviation from the mean pose on the entire dataset.
2. *Occlusion*: number of occluded body parts.
3. *Viewpoint*: deviation of 3D torso rotation from the frontal viewpoint.
4. *Part length*: deviation of body part lengths from the mean part lengths.
5. *Truncation*: number of truncated body parts.

Novel motion specific complexity measures. We augment the above set with the measures assessing the amount of motion present in the scene.

1. *# dense trajectories (# DT)*: total number of DT computed by (Wang and Schmid, 2013).
2. *# dense trajectories on body (# DT body)*: number of DT on body mask.
3. *Motion speed (MS)*: mean over all trajectory displacements in the video.
4. *Motion speed on body (MS body)*: *MS* extracted on body mask.
5. *# people*: number of people in the video.

10.3 METHODS

In order to analyze the performance on the challenging task of fine-grained human activity recognition, we explore two lines of methods that extract relevant features. The first line of methods extracts holistic appearance based features and is represented by the "Dense Trajectories" method (Wang *et al.*, 2013) which achieves very good performance on several datasets. The second line of methods computes features from locations of human body joints following the intuition that body part configurations and motion should provide strong cues for activity recognition. We now describe both types of methods and their combinations.

10.3.1 Dense trajectories (DT)

DT computes histograms of oriented gradients (HOG) (Dalal and Triggs, 2005), histograms of flow (HOF) (Laptev *et al.*, 2008), and motion boundary histograms (MBH) (Dalal *et al.*, 2006) around densely sampled points that are tracked for 15 frames using median filtering in a dense optical flow field. In addition, x and y displacements in a trajectory are used as a fourth feature. We use a publicly available implementation of the improved DT method (Wang and Schmid, 2013), where additional estimation removes some of the trajectories consistent with camera motion. Following (Wang and Schmid, 2013) we extract all features on our data and generate a codebook for each of the four features of 4K words using k-means from a million of sampled features, and stack L_2-normalized histograms for learning.

10.3.2 Pose-based methods

It has been shown that body features provide a strong signal for recognition of human activities on a rather limited set of 21 distinctive full body actions in monocular RGB video sequences (Jhuang *et al.*, 2013). We thus investigate the usefulness of body features for our task where the variability of poses and granularity of activities is much higher. We explore different ways of obtaining body joint locations and extract the body features using the code kindly provided by (Jhuang *et al.*, 2013). We use the same trajectory length of 7 frames with a step size of 3, generate a codebook of 20 words for each descriptor type and finally stack the L_2-normalized histograms for learning. We now present different ways of obtaining body joint locations.

GT single pose (GT). We directly use the ground truth locations of body joints in the key frame to compute single pose based features. As some of the body parts may be truncated, we compute features only for the present parts.

GT single pose + track (GT-T). As the ground truth information is not available for the rest of the frames in a sequence, we approximate the positions of body joints in the neighboring frames by tracking the joints using sift-based tracker (Rohrbach *et al.*, 2012). The tracker is initialized with correct positions of body joints, and thus provides reliable tracks of joints in the local temporal neighborhood.

PS single pose + track (PS-T). It is not realistic to expect the ground truth information to be available at test time in real world scenarios. We thus replace the given body joint locations by automatically estimated ones using the publicly available implementation (Yang and Ramanan, 2013). As we show in Chapter 9, pose estimation performance of their method on the "MPI Human Pose" is slightly below the performance by our best model introduced in Chapter 7 and extended in Chapter 8. However, we use the method of (Yang and Ramanan, 2013) due to efficiency reasons.

PS multi-pose (PS-M). Using pose estimation method also allows to obtain joint locations independently in each frame of a sequence without using the sift tracker. Notably, the same method was shown by (Jhuang *et al.*, 2013) to outperform the holistic approach.

10.3.3 Combinations of holistic and body based methods

As the holistic *DT* approach does not extract any pose information, and pose based methods do not compute any appearance features, both are potentially complementary. Thus we expect that an activity recognition system will profit from their combinations. We investigate two ways of combining the methods.

PS-M + DT (features). We perform a *feature* level fusion of both *DT* and *PS-M* by matching both types of features independently to the respective codebooks and then stacking the normalized histograms into a single representation.

PS-M + DT (classifiers). We also investigate a *classifier* level fusion. To do so we first run pre-trained *DT* and *PS-M* classifiers (see Sec. 10.4) independently on each sequence and stack the scores together into a single feature vector.

PS-M filter DT. Another way of combining both types of methods is using estimated joint locations to filter the trajectories computed by *DT*. We first estimate poses in all video frames and generate a binary mask using the union of rectangles around detected body parts for all single top detections per frame. We then only preserve the trajectories overlapping with the mask in all frames.

10.4 ANALYSIS OF ACTIVITY RECOGNITION PERFORMANCE

In this section we analyze the performance of holistic and pose based methods and their combinations on the challenging task of fine-grained human activity recognition with hundreds of activity classes.

Data splits. As main test bed for our analysis, a split with videos containing sufficiently separated individuals (Chapter 9) is used. This restriction is necessarily for using the pose estimation method. This *Separate people* split contains 15244 training and 5699 testing video snippets. Fig. 10.1(a) shows statistics of the training and testing videos per activity. Notably, the videos may still contain multiple people and some body parts may be truncated by a frame border. To rule out the confusion potentially caused by the presence of multiple truncated people, we define a subset of the test set from *Separate people*. This subset contains 2622 videos with exactly one fully visible person per video. This *Single fully visible people* setup is inspired by (Jhuang *et al.*, 2013) and is favorable for the pose estimation method designed to predict body joints of fully visible people.

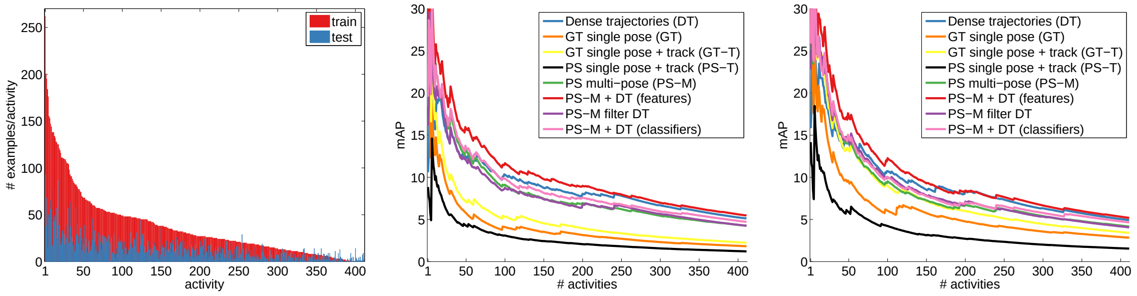
Training and evaluation. We train activity classifiers using feature representations described in Sec. 10.3 and ground truth activity labels. In particular, we train one-vs-all SVMs using mean stochastic gradient descent (SGD) (Rohrbach *et al.*, 2011) with a χ^2 kernel approximation (Vedaldi and Zisserman, 2010). At test time we perform one-vs-all prediction per each class independently and report the results using mean Average Precision (AP) (Everingham *et al.*, 2007). When evaluating on a subset, we always report the results on the top N activity classes arranged w.r.t training set sizes.

10.4.1 Overall performance

We start the evaluation by analyzing the performance on all activity classes.

Separate people. It can be seen from Fig. 10.1(b) that performance is reasonable for a relatively small number of classes (the typical case for many activity recognition datasets), but quickly degrades for a large number of classes, clearly leaving room for improvement of activity recognition methods.

We observe that *Dense trajectories (DT)* alone outperforms all pose based methods achieving 5.1% mAP. Expectedly, *GT single pose (GT)* performs worst (1.8% mAP). Although *GT* uses ground truth joint positions to extract body features, they are computed in a single key frame, thus ignoring motion. Expectedly, adding motion via sift tracking (*GT single pose + track (GT-T)*) improves the results to 2.2% mAP. Replacing ground truth by predicted joint locations (*PS single pose + track (PS-T)*) results in a performance drop (1.2% vs 2.2% mAP) due to unreliable initialization of the tracker by imperfect pose estimation. Surprisingly, *PS multi-pose (PS-M)* significantly improves the results, achieving 4.2% mAP. It shows that performing body joint predictions in each individual frame is more reliable than simple tracking. Interestingly, the feature level fusion *PS-M + DT (features)* noticeably improves over *DT* alone and classifier level fusion *PS-M + DT (classifiers)*, achieving 5.5% mAP. This shows that both holistic *DT* and pose based *PS-M* methods are complementary. We analyze whether the complementarity of *DT* comes from the holistic features extracted on the person or elsewhere in the scene. By restricting the extraction to the



(a) # examples/activity on *Separate people*

(b) Performance (mAP) on *Separate people*

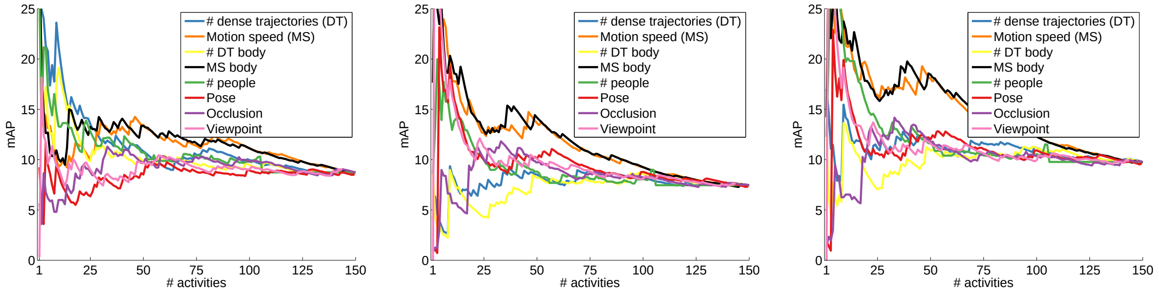
(c) Performance (mAP) on *Single fully visible people*

Figure 10.1: Dataset statistics and performance (mAP) as a function training set size. Shown are (a) number of training/testing examples/activity in *Separate people* subset; performance on (b) *Separate people* and (c) *Single fully visible people*. Best viewed in color.

body mask (*PS-M filter DT*), we observe a drop of performance w.r.t. *DT* (4.3% mAP vs. 5.1% mAP). It shows that the features extracted outside of the body mask do contain additional information which helps to better discriminate between activities in a fine-grained recognition setting. This intuition is additionally supported by the similar performance of *PS-M filter DT* w.r.t. *PS-M*. Overall, we conclude that holistic and pose based methods are complementary and should be used in a combination for better activity recognition.

Single fully visible people. We now analyze the results in Fig. 10.1(c). Although the absolute performances are higher, which is explained by an easier setting, the ranking is similar to Fig. 10.1(b). Two differences are: 1) *GT-T* achieves similar performance to *PS-M* on many activity classes, but looses in total (3.4% mAP vs. 4.2% mAP); and 2) *PS-M filter DT* is better than both *DT* and *PS-M* on a small set of classes, probably because the trajectory features on the background mostly contribute to confusion on this set of activities.

Differences to (Jhuang *et al.*, 2013). Our analysis in a fine-grained activity recognition setting on hundreds of classes leads to conclusions which go beyond the results of (Jhuang *et al.*, 2013) obtained from much smaller number of classes from HMDB dataset (Kuehne *et al.*, 2011). First, we compare the performance of *DT* to a larger number of pose based methods and show the superior performance of *DT*, when evaluated on hundreds of activities. This is in contrast to (Jhuang *et al.*, 2013) showing that the pose based *PS-M* is better. Second, we discover that holistic *DT* and pose based *PS-M* are complementary and their combination outperforms each of the approaches alone. This contradicts the conclusions of (Jhuang *et al.*, 2013) which does not show any improvement when combining *DT* and *PS-M*. Finally, we showed that using the trajectories restricted to body only degrades the performance, which



(a) *DT*

(b) *PS multi-pose (PS-M)*

(c) *PS-M + DT (features)*

Figure 10.2: Performance (mAP) on a subset of 150 activities from *Separate people* as a function of the complexity measures. Best viewed in color and with additional zooming.

suggests that the context adds to the discrimination between activity classes.

10.4.2 Analysis of activity recognition challenges

After analyzing the overall recognition performance on all classes, we explore which factors affect the performance of best performing holistic *DT*, pose based *PS-M* and combination *PS-M + DT (features)* of both methods. We use the complexity measures $1-3$ specific for static pose estimation and our novel $1-5$ motion specific complexity measures described in Sec. 10.2. To make the evaluation consistent with the rest of the experiments, we compute the average complexities for the whole activity class and use the obtained real values to rank the classes. This is in contrast to the evaluation of pose estimation in Chapter 9 where we compute the measures per single pose and thus operate on individual instance level. To visualize the performance, we sort the activities using the pose related complexity measures in *increased* complexity order, and motion related complexity measures in the *decreased* order. As performance may still be dominated by the training set size when only few examples are available, we restrict the evaluation to the 150 largest activity classes. This corresponds to a slice at 150 in Fig. 10.1(b). The results are shown in Fig. 10.2.

Dense trajectories (DT). Analyzing the results in Fig. 10.2(a) we observe that a high number of dense trajectories everywhere in the video (*# DT*) and on human body (*# DT body*) leads to the best performance of the *DT* method. Also, we notice that high motion speed (*MS*, *MS body*) is an indicative factor for good recognition results. Surprisingly, *DT* performs better on activities with a high number of people (*# people*). This is explained by the fact that more people potentially produce more motion, which is a positive factor for *DT*. On the other hand, being close to the average pose (*Pose*) and having little occlusion (*Occlusion*) hurts performance. The former is not very surprising, as the average pose is common to many activities, which makes it more difficult for *DT* to capture distinctive features. We discover that activities with little occlusion often contain either little motion (e.g. "sitting, talking in person") or fine-grained motion (e.g. "wash dishes") and thus are hard for *DT*.

PS multi-pose (PS-M). We now analyze Fig. 10.2(b) and observe several distinctive differences w.r.t. which factors mostly affect the performance of *PS-M*. It can be seen that *MS* and *MS body* have stronger effect on *PS-M* compared to *DT*, and the higher the speed, the better the performance. In order to better understand this nontrivial trend, we analyze which activities happen to produce highest *MS body*. We note that those are sports, dancing and running related activities, for which the pose estimation performance of (Yang and Ramanan, 2013) is above average (cf. Fig. 9.8). Also, these activities exhibit characteristic body part motions and can successfully be encoded using body features. At the other end of the *MS body* scale are the activities with low fine-grained motion, related to home repair, self care and occupation, for which the pose estimation performance is much worse. *Pose* and *Viewpoint* strongly affect the performance as well, as frontal upright people whose pose is close to the mean pose are easier for pose estimation. This is again in contrast to *DT*, where the performance is not noticeably affected by *Viewpoint* and even drops in case of low *Pose*. Surprisingly, high *# people* positively affects *PS-M*. Looking at top ranked activities, we notice that many of them are related to active group exercises or team sports, such as "aerobic" and "frisbee", or to simple standing postures, such as "standing, talking in person". Body features can again be successfully used to encode the corresponding motions. On the other hand, we observe that the high *# DT* and *# DT body* hurts performance, which is in contrast to the *DT* method. We observe that for high *# DT body* many activities correspond to water related activities, such as "fishing in stream", "swimming, general", "canoeing, kayaking". Interestingly, the presence of water leads to high *# DT* and characteristic motions captured by *DT*. At the same time *PS-M* fails due to unreliable pose estimation caused by complex appearance and occlusions.

PS-M + DT (features). The differences for *DT* and *PS-M* methods suggest that both methods are complementary. We analyze in Fig. 10.2(c) which factors affect the performance of *PS-M + DT (features)*. It can be seen that positively affecting factors are either positive for both *DT* and *PS-M* (*MS*, *MS body*, *# people*), or positive for *PS-M* only (*Pose*, *Viewpoint*). In contrast to *PS-M* the high *# DT* slightly improves the performance, while high *# DT body* does not hurt as much. Expectedly, *Viewpoint* hurts performance as it does for both *DT* and *PS-M*. This shows the complementarity of both *DT* and *PS-M*and leaves room for improvement in finding better ways of combining both methods.

10.4.3 Detailed analysis on a subset of activities

After analyzing the factors affecting the results by different methods, we conduct a detailed analysis on a smaller set of the top 15 activities from *Separate people*.

	yoga, power	bicyc., moun.	skiing, down.	cooking or food	skate-board.	rope skip.	softb., gener.	forest.
Dense trajectories (DT)	10.6	14.5	51.9	0.5	11.4	36.0	**12.7**	8.4
GT single pose (GT)	22.3	26.5	7.5	1.8	3.4	51.2	2.2	1.4
GT single pose + track (GT-T)	**37.0**	28.0	10.9	**2.6**	4.6	69.2	3.6	1.2
PS single pose + track (PS-T)	8.8	6.6	6.0	1.3	1.7	63.1	1.6	1.8
PS multi-pose (PS-M)	18.3	34.0	27.3	2.6	17.2	**90.5**	3.0	5.2
PS-M + DT (features)	19.6	**40.7**	32.9	2.2	**19.5**	88.7	3.9	7.2
PS-M filter DT	16.1	20.4	**52.2**	0.8	13.5	55.7	4.2	**10.6**

	carpen., gener.	bicyc., racing	golf	rock climb.	ballet, modern	aerobic step	resist. train.	total
Dense trajectories (DT)	5.5	5.5	33.0	**41.5**	12.7	24.5	**16.5**	19.0
GT single pose (GT)	2.7	7.1	36.1	2.3	1.0	1.1	1.4	11.2
GT single pose + track (GT-T)	2.8	8.7	25.3	8.9	1.7	3.3	1.3	13.9
PS single pose + track (PS-T)	5.3	0.5	14.7	1.2	2.8	11.1	1.6	8.5
PS multi-pose (PS-M)	3.4	8.6	47.9	4.7	22.9	10.4	7.2	20.2
PS-M + DT (features)	5.0	12.1	**51.9**	14.4	**23.7**	17.1	14.4	**23.5**
PS-M filter DT	**6.1**	**15.5**	15.9	38.6	7.1	**25.8**	9.6	19.5

Table 10.1: Activity recognition results (mAP) on 15 largest classes from *Separate people*.

The results are shown in Tab. 10.1. None of the methods outperforms all others on all activities and different approaches are better on different activities. On average methods perform well on activities with simple poses and motions e.g. "rope skipping", "skiing, downhill" and "golf" - typical cases in most of the current activity recognition benchmarks. However, the performance of all methods is low for activities with more variability in motion and poses, e.g. "cooking", "carpentry, general" and "forestry". This leaves room for improvement of all competing methods. Analyzing the performance on individual activities, we observe that for "yoga, power" activity *GT* outperforms holistic *DT* and *PS-M filter DT* methods (22.3% vs. 10.6% and 16.1% mAP, respectively) and is better than the pose based *PS-M* (22.3% vs. 18.3% mAP). It is interesting, as *GT* does not use any motion and relies on static body features only. The explanation is that the "yoga, power" activity contains distinctive body poses and thus can be reliably captured by *GT*, while *PS-M* fails due to unreliable pose estimations. It can be seen that in many cases the combination *PS-M + DT (features)* noticeably outperforms both *PS-M* and *DT* alone. The differences are most pronounced for "bicycling, mountain", "bicycling, racing", "skateboarding" exhibiting characteristic motions, and "golf" activity having distinctive body motion and poses. Overall *PS-M + DT (features)* achieves the best performance of 23.5% mAP. We visualize several successful and failure cases of the methods in Fig. 10.3.

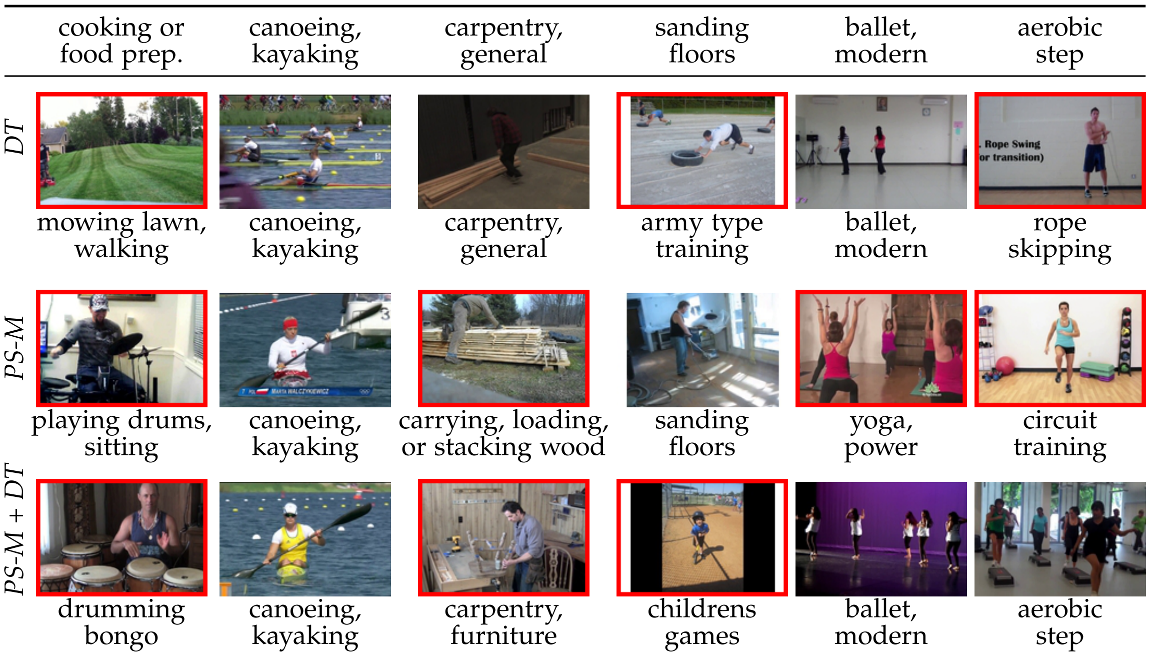


Figure 10.3: Successful and failure cases on several activity classes. Shown are the most confident prediction per class. False positives are highlighted in red.

10.5 CONCLUSION

In this work we address the challenging task of fine-grained human activity recognition on a recent comprehensive dataset with hundreds of activity classes. We study holistic and pose based representations and analyze the factors responsible for their performance. We reveal that holistic and pose based methods are complementary, and their performance varies significantly depending on the activity. We found that both methods are strongly affected by the speed of trajectories. While the holistic method is also strongly influenced by the number of trajectories, pose based methods are strongly affected by human pose and viewpoint. We observe striking performance differences across activities and experimentally show that the combination of both methods performs best.

11

DEEPCUT: JOINT SUBSET PARTITION AND LABELING FOR MULTI PERSON POSE ESTIMATION

Contents

11.1 Introduction **180**
11.2 Problem Formulation **182**
  11.2.1 Feasible Solutions 182
  11.2.2 Objective Function 183
  11.2.3 Optimization 184
11.3 Pairwise Probabilities **184**
  11.3.1 Probability Estimation 185
11.4 Body Part Detectors **186**
  11.4.1 Adapted Fast R-CNN (*AFR-CNN*) 186
  11.4.2 Dense architecture (*Dense-CNN*) 186
  11.4.3 Evaluation of part detectors 188
  11.4.4 Using detections in DeepCut models 189
11.5 DeepCut Results **190**
  11.5.1 Single person pose estimation 190
  11.5.2 Multi-person pose estimation 192
11.6 Conclusion **201**
11.7 Appendix: Additional Results on LSP dataset **201**
  11.7.1 LSP Person-Centric (PC) 201
  11.7.2 LSP Observer-Centric (OC) 204

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

11.1 INTRODUCTION

Human body pose estimation methods have become increasingly reliable. Powerful body part detectors (Tompson *et al.*, 2015) in combination with tree-structured body models (Tompson *et al.*, 2014; Chen and Yuille, 2014) show impressive results on diverse datasets (Johnson and Everingham, 2011; Andriluka *et al.*, 2014; Sapp and Taskar, 2013). These benchmarks promote pose estimation of single pre-localized persons but exclude scenes with multiple persons. This problem definition has been a driver for progress, but also falls short on representing a realistic sample of real-world images. Many photographs contain multiple people of interest (see Fig 11.1) and it is unclear whether single pose approaches generalize directly. We argue that the multi person case deserves more attention since it is an important real-world task.

Key challenges inherent to multi person pose estimation are the partial visibility of some people, significant overlap of bounding box regions of people, and the a-priori unknown number of people in an image. The problem thus is to infer the number of persons, assign part detections to person instances while respecting geometric and appearance constraints. Most strategies use a two-stage inference process (Gkioxari *et al.*, 2014; Sun and Savarese, 2011, Chapter 5) to first detect and then independently estimate poses. This is unsuited for cases when people are in close proximity since they permit simultaneous assignment of the same body-part candidates to multiple people hypotheses.

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

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

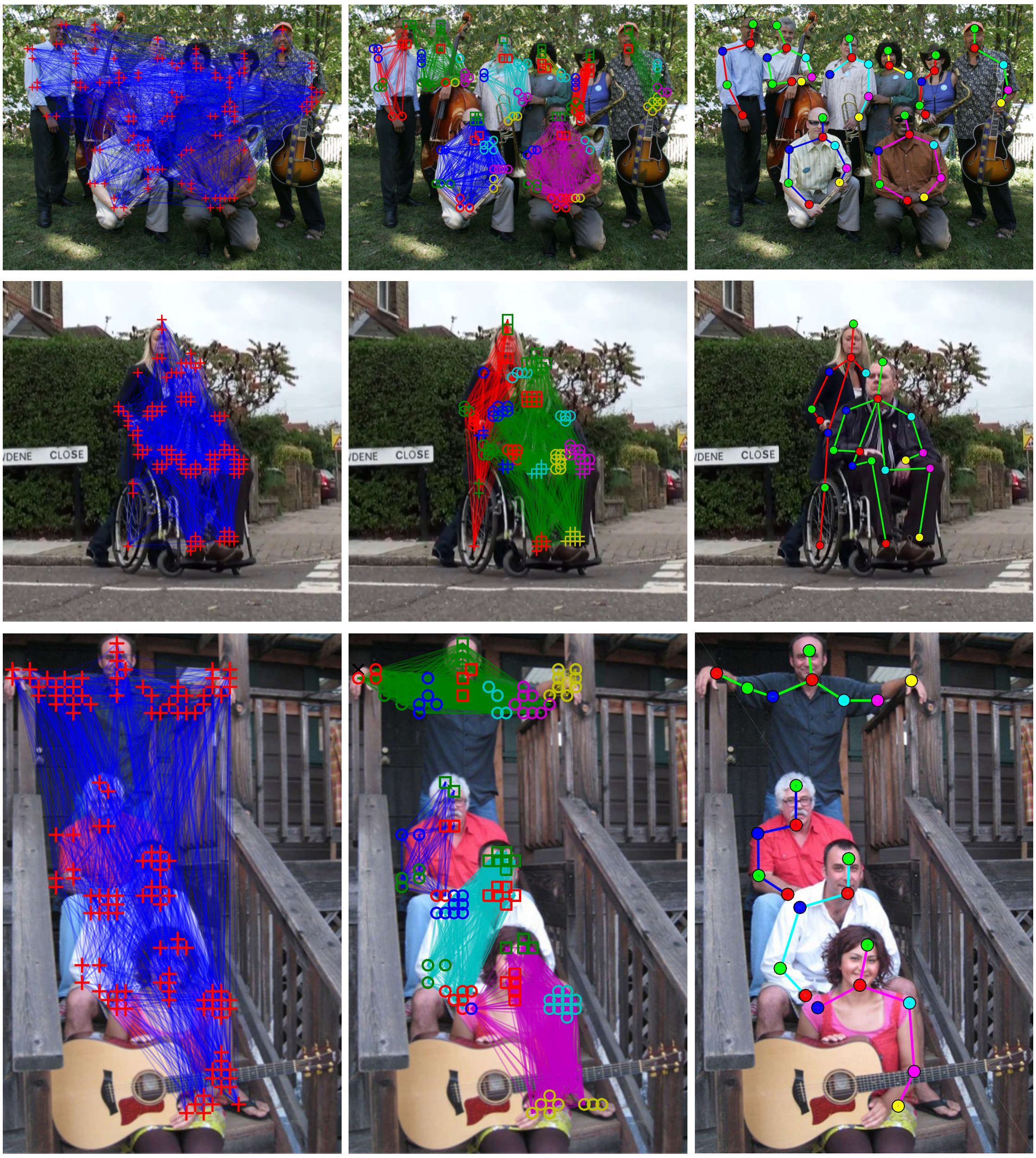

(a) (b) (c)

Figure 11.1: Method overview: (a) initial detections (= part candidates) and pairwise terms (graph) between all detections that (b) are jointly clustered belonging to one person (one colored subgraph = one person) and each part is labeled corresponding to its part class (different colors and symbols correspond to different body parts); (c) shows the predicted pose sticks.

11.2 PROBLEM FORMULATION

In this section, the problem of estimating articulated poses of an unknown number of people in an image is cast as an optimization problem. The goal of this formulation is to state three problems jointly: 1. The selection of a subset of body parts from a set D of *body part candidates*, estimated from an image as described in Section 11.4 and depicted as nodes of a graph in Fig. 11.1(a). 2. The *labeling* of each selected body part with one of C *body part classes*, e.g., "arm", "leg", "torso", as depicted in Fig. 11.1(c). 3. The *partitioning* of body parts that belong to the same person, as depicted in Fig. 11.1(b).

11.2.1 Feasible Solutions

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

In order to constrain the 01-labelings (x, y, z) to well-defined articulated poses of one or more people, we impose the linear inequalities (11.1)–(11.3) stated below. Here, the inequalities (11.1) guarantee that every body part is labeled with at most one body part class. (If it is labeled with no body part class, it is suppressed). The inequalities (11.2) guarantee that distinct body parts d and d' belong to the same person only if neither d nor d' is suppressed. The inequalities (11.3) guarantee, for any three pairwise distinct body parts, d, d' and d'', if d and d' are the same person (as indicated by $y_{dd'} = 1$) and d' and d'' are the same person (as indicated by $y_{d'd''} = 1$), then also d and d'' are the same person ($y_{dd''} = 1$), that is, transitivity, cf. (Chopra and Rao, 1993). Finally, the inequalities (11.4) guarantee, for any $dd' \in \binom{D}{2}$ and any $cc' \in C^2$ that $z_{dd'cc'} = x_{dc} x_{d'c'} y_{dd'}$. These constraints allow us to write an objective function as a linear form in z that would otherwise be written as a cubic form in x and y. We denote by X_{DC} the set of all (x, y, z) that satisfy all inequalities, i.e., the set of feasible solutions.

$$\begin{aligned}
\forall d \in D \forall cc' \in \tbinom{C}{2}: \quad & x_{dc} + x_{dc'} \leq 1 && (11.1)\\
\forall dd' \in \tbinom{D}{2}: \quad & y_{dd'} \leq \sum_{c \in C} x_{dc} &&\\
& y_{dd'} \leq \sum_{c \in C} x_{d'c} && (11.2)\\
\forall dd'd'' \in \tbinom{D}{3}: \quad & y_{dd'} + y_{d'd''} - 1 \leq y_{dd''} && (11.3)\\
\forall dd' \in \tbinom{D}{2} \forall cc' \in C^2: \quad & x_{dc} + x_{d'c'} + y_{dd'} - 2 \leq z_{dd'cc'} &&\\
& z_{dd'cc'} \leq x_{dc} &&\\
& z_{dd'cc'} \leq x_{d'c'} &&\\
& z_{dd'cc'} \leq y_{dd'} && (11.4)
\end{aligned}$$

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

$$\forall dd' \in \tbinom{D}{2} \forall cc' \in C^2: \quad x_{dc} + x_{d'c'} - 1 \leq y_{dd'} \tag{11.5}$$

11.2.2 Objective Function

For every pair $(d, c) \in D \times C$, we will estimate a probability $p_{dc} \in [0, 1]$ of the body part d being of class c. In the context of CRFs, these probabilities are called *part unaries* and we will detail their estimation in Section 11.4.

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

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

$$\min_{(x,y,z) \in X_{DC}} \quad \langle \alpha, x \rangle + \langle \beta, z \rangle, \tag{11.6}$$

where we used the short-hand notation

$$\alpha_{dc} := \log \frac{1 - p_{dc}}{p_{dc}} \tag{11.7}$$

$$\beta_{dd'cc'} := \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} \tag{11.8}$$

$$\langle \alpha, x \rangle := \sum_{d \in D} \sum_{c \in C} \alpha_{dc} \, x_{dc} \tag{11.9}$$

$$\langle \beta, z \rangle := \sum_{dd' \in \binom{D}{2}} \sum_{c,c' \in C} \beta_{dd'cc'} \, z_{dd'cc'} \quad . \tag{11.10}$$

11.2.3 Optimization

In order to obtain feasible solutions of the ILP (11.6) with guaranteed bounds, we separate the inequalities (11.1)–(11.5) in the branch-and-cut loop of the state-of-the-art ILP solver Gurobi. More precisely, we solve a sequence of relaxations of the problem (11.6), starting with the (trivial) unconstrained problem. Each problem is solved using the cuts proposed by Gurobi. Once an integer feasible solution is found, we identify violated inequalities (11.1)–(11.5), if any, by breadth-first-search, add these to the constraint pool and re-solve the tightened relaxation. Once an integer solution satisfying all inequalities is found, together with a lower bound that certifies an optimality gap below 1%, we terminate.

11.3 PAIRWISE PROBABILITIES

Here we describe the estimation of the pairwise terms. We define pairwise features $f_{dd'}$ for the variable $z_{dd'cc'}$ (Sec. 11.2). Each part detection d includes the probabilities $f_{p_{dc}}$ (Sec. 11.4.4), its location (x_d, y_d), scale h_d and bounding box B_d coordinates. Given two detections d and d', and the corresponding features $(f_{p_{dc}}, x_d, y_d, h_d, B_d)$ and $(f_{p_{d'c}}, x_{d'}, y_{d'}, h_{d'}, B_{d'})$, we define two sets of auxiliary variables for $z_{dd'cc'}$, one set for $c = c'$ (same body part class clustering) and one for $c \neq c'$ (across two body part classes labeling). These features capture the proximity, kinematic relation and appearance similarity between body parts.

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

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

$\dots,(IOMax)^2, \exp(-\Delta x),\dots,\exp(-IOMax))$.

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

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

where $p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1)$ is obtained by conducting a normalized 2D histogram of $S_{dd'}$ and $R_{dd'}$ from positive training examples, analogous to the negative likelihood $p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 0)$. In Sec. 11.5.1 we also experiment with encoding the appearance into the pairwise feature by concatenating the feature $f_{p_{dc}}$ from d and $f_{p_{d'c}}$ from d', as $f_{p_{dc}}$ is the output of the CNN-based part detectors. The final pairwise feature is $(p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'}), f_{p_{dc}}, f_{p_{d'c}})$.

11.3.1 Probability Estimation

The coefficients α and β of the objective function (Eq. 11.6) are defined by the probability ratio in the log space (Eq. 11.7 and Eq. 11.8). Here we describe the estimation of the corresponding probability density: *(1)* For every pair of detection and part classes, namely for any $(d, c) \in D \times C$, we estimate a probability $p_{dc} \in (0, 1)$ of the detection d being a body part of class c. *(2)* For every combination of two distinct detections and two body part classes, namely for any $dd' \in \binom{D}{2}$ and any $cc' \in C^2$, we estimate a probability $p_{dd'cc'} \in (0, 1)$ of d and d' belonging to the same person, meanwhile d and d' are body parts of classes c and c', respectively.

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

$$p(z_{dd'cc'} = 1|f_{dd'}, \theta_{cc'}) = \frac{1}{1 + \exp(-\langle \theta_{cc'}, f_{dd'} \rangle)}. \tag{11.11}$$

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

Inference Given two detections d and d', the coefficients α_{dc} for x_{dc} and $\alpha_{d'c}$ for $x_{d'c}$ are obtained by Eq. 11.7, the coefficient $\beta_{dd'cc'}$ for $z_{dd'cc'}$ has the form

$$\beta_{dd'cc'} = \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} = -\langle f_{dd'}, \theta_{cc'} \rangle. \tag{11.12}$$

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

11.4 BODY PART DETECTORS

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

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

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

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

Larger context. Increasing the size of DPM detections by up-scaling every bounding box by a fixed factor allows to capture more context around each part. In Sec. 11.4.3 we evaluate the influence of up-scaling and show that using larger context around parts is crucial for best performance.

Details. Following standard FR-CNN training procedure ImageNet models are finetuned on pose estimation task. Center of a predicted bounding box is used for body part location prediction. See Sec. 11.7 for detailed parameter analysis.

11.4.2 Dense architecture (*Dense-CNN*)

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

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

Scale. Selecting the scale at which CNN is applied is crucial. We empirically found that scaling an image such that an upright standing person is 340 px high leads to best results. This way 224×224 VGG receptive field sees sufficiently large portion of human to disambiguate body parts.

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

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

Regression to other parts. Similar to location refinement we add an extra term to the objective function where for each part we regress onto all other part locations. We empirically found this auxiliary task to improve the unary performance (c.f. Sec. 11.4.3). We envision these predictions to improve the spatial model as well and leave this for the future work.

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

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

Table 11.1: Unary only performance (PCK) of *AFR-CNN* on the LSP (Person-Centric) dataset. *AFR-CNN* is finetuned from ImageNet to MPII (lines 3-6), and then finetuned to LSP (line 7).

11.4.3 Evaluation of part detectors

Datasets. For training and evaluation we use three public benchmarks: "Leeds Sports Poses" (LSP) (Johnson and Everingham, 2010) (person-centric (PC) annotations) including 1000 training and 1000 testing images of people doing sports; "LSP Extended" (LSPET) (Johnson and Everingham, 2011) consisting of 10000 training images; "MPII Human Pose" ("Single Person") (Andriluka *et al.*, 2014) consisting of 19185 training and 7247 testing people in every day activities. The MPII training set is used as default. In some cases LSP training *and* LSPET is included, this is denoted as MPII+LSPET in the experiments. As LSPET has severe labeling noise, all original high-resolution images were re-annotated.[13]

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

AFR-CNN. Evaluation of *AFR-CNN* on LSP is shown in Tab. 11.1. Oracle selecting per part the closest from 2000 proposals achieves 97.8% PCK, as proposals cover majority of the ground truth locations. Choosing a single proposal per part using DPM score achieves 23% PCK – not surprising given the difficulty of the body part detection problem. Re-scoring the proposals using *AFR-CNN* with AlexNet (Krizhevsky *et al.*, 2012) dramatically improves the performance to 56.9% PCK, as CNN learns richer image representations. Extending the regions by 4x (1x ≈ head size) achieves 65.1% PCK, as it incorporates more context including the information about symmetric parts and allows to implicitly encode higher-order part relations. Using data augmentation and slightly tuning training parameters improves the performance

---

[13] Data will be made publicly available.

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

Table 11.2: Unary only performance (PCK) of *Dense-CNN* VGG on LSP (PC) dataset. *Dense-CNN* is finetuned from ImageNet to MPII (line 1), to MPII+LSPET (lines 2-5), and finally to LSP (line 6).

to 72.4% PCK. We refer to Sec. 11.7 for detailed analysis. Deeper VGG architecture improves over smaller AlexNet reaching 77.9% PCK. All results so far are achieved by finetuning the ImageNet models on MPII. Further finetuning to LSP leads to remarkable 82.8% PCK: network learns LSP-specific image representations. Strong increase in AUC (57.0 vs. 50%) is due to improvements for smaller PCK thresholds. No bounding box regression leads to performance drop (81.3% PCK, 53.2% AUC): location refinement is crucial for better part localization. Overall *AFR-CNN* obtains very good results on LSP by far outperforming the state of the art (c.f. Tab. 11.3, rows $7-9$). Evaluation on MPII Single Person shows competitive performance (Tab. 11.4, row 1).

Dense-CNN. The results are in Tab. 11.2. Training with VGG on MPII with softmax loss achieves 80.8% PCK thereby outperforming *AFR-CNN* (c.f. Tab. 11.1, row 6). This shows the advantages of fully convolutional training and evaluation. Expectedly, training on larger MPII+LSPET dataset improves the results (83.0 vs. 80.8% PCK). Using cross-entropy loss with sigmoid activations improves the results to 83.8% PCK, as it better models the appearance of close-by parts. Location refinement improves localization accuracy (84.8% PCK), which becomes more clear when analyzing AUC (61.5 vs. 55.6%). Interestingly, regressing to other parts further improves PCK to 86.1% showing a value of training with the auxiliary task. Finally, finetuning to LSP achieves the best result of 87.1% PCK, which is significantly higher than the best published results (c.f. Tab. 11.3, rows $7-9$). Unary-only evaluation on MPII reveals slightly higher AUC results compared to the state of the art (Tab. 11.4, row $3-4$).

11.4.4 Using detections in DeepCut models

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

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AFR-CNN (unary)	95.4	86.5	77.8	74.0	84.5	82.6	78.8	82.8	57.0
+ *DeepCut SP*	95.4	86.7	78.3	74.0	84.3	82.9	79.2	83.0	58.4
+ appearance pairwise	95.4	87.2	78.6	73.7	84.7	82.8	78.8	83.0	58.5
+ *DeepCut MP*	95.2	86.7	78.2	73.5	84.6	82.8	79.0	82.9	58.0
Dense-CNN (unary)	97.2	90.8	83.0	79.3	90.6	85.6	83.1	87.1	63.6
+ *DeepCut SP*	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1	63.5
+ *DeepCut MP*	96.2	**91.2**	83.3	77.6	**91.3**	**87.0**	80.4	86.7	62.6
PS (Chapter 8)	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1	35.8
(Tompson *et al.*, 2014)	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3
(Carreira *et al.*, 2016)	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1	41.5
(Chen and Yuille, 2014)	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	40.1
(Fan *et al.*, 2015)*	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2
(Wei *et al.*, 2016)	**97.8**	**92.5**	**87.0**	**83.9**	**91.5**	**90.8**	**89.9**	**90.5**	**65.4**

* re-evaluated using the standard protocol, for details see project page of (Fan *et al.*, 2015)

Table 11.3: Pose estimation results (PCK) on LSP (PC) dataset.

11.5 DEEPCUT RESULTS

The aim of this Chapter is to tackle the multi-person case. To that end, we evaluate the proposed *DeepCut* models on four diverse benchmarks. We confirm that both single person (*SP*) and multi-person (*MP*) variants (Sec. 11.2) are effective on standard *SP* pose estimation datasets (Johnson and Everingham, 2010; Andriluka *et al.*, 2014). Then, we demonstrate superior performance of *DeepCut MP* on the multi-person pose estimation task.

11.5.1 Single person pose estimation

We now evaluate single person (*SP*) and more general multi-person (*MP*) *DeepCut* models on LSP and MPII *SP* benchmarks described in Sec. 11.4. Since this evaluation setting implicitly relies on the knowledge that all parts are present in the image we always output the full number of parts.

Results on LSP. We report per-part PCK results (Tab. 11.3) and results for a variable distance threshold (Fig. 11.2 (a)). *DeepCut SP AFR-CNN* model using 100 detections improves over unary only (83.0 vs. 82.8% PCK, 58.4 vs. 57% AUC), as pairwise connections filter out some of the high-scoring detections on the background. The improvement is clear in Fig. 11.2 (a) for smaller thresholds. Using part appearance scores in addition to geometrical features in $c \neq c'$ pairwise terms only slightly improves AUC, as the appearance of neighboring parts is mostly captured by a relatively large region centered at each part. As geometrical only pairwise lead to faster experiments. The performance of *DeepCut MP AFR-CNN* matches the *SP* and improves over *AFR-CNN* alone: *DeepCut MP* correctly handles the *SP* case. Performance of *DeepCut SP Dense-CNN* is almost identical to unary only, unlike the results for *AFR-CNN*. *Dense-CNN* performance is noticeably higher compared to *AFR-CNN*, and "easy" cases that could have been corrected by a spatial model are resolved by stronger part detectors alone.

Comparison to the state of the art (LSP). Tab. 11.3 compares results of *DeepCut* models to our single person *PS* approach (Chapter 8) and deep learning methods specifically designed for single person pose estimation. All *DeepCuts* significantly outperform *PS*, as they are able to learn much richer representations compared to hand-crafted image features used in *PS*. This clearly shows the advantages of using deep learning to build much stronger part detection models. All *DeepCuts* significantly outperform prior work (Tompson *et al.*, 2014; Chen and Yuille, 2014; Fan *et al.*, 2015), with *DeepCut SP Dense-CNN* model improving by 13.7% PCK over (Chen and Yuille, 2014). The improvement is even more dramatic for lower thresholds (Fig. 11.2 (a)): for PCK @ 0.1 the best model improves by 19.9% over (Tompson *et al.*, 2014), by 26.7% over (Fan *et al.*, 2015), and by 32.4% PCK over (Chen and Yuille, 2014). The latter is interesting, as (Chen and Yuille, 2014) use a stronger spatial model that predicts the pairwise conditioned on the CNN features, whereas *DeepCuts* use geometric-only pairwise connectivity. Including body part orientation information into *DeepCuts* should further improve the results. *DeepCut* also significantly outperforms concurrent approach (Carreira *et al.*, 2016) (87.1 vs. 73.1% PCK), as fully-convolutional part detectors of *DeepCut* can better localize individual body parts compared to holistic regressions used by (Carreira *et al.*, 2016). Another concurrent approach (Wei *et al.*, 2016) slightly outperforms *DeepCut* (90.5 vs. 87.1% PCK), as it encodes more contextual information into the part detectors via the multi-stage multi-scale training procedure. Difference is visible for larger distance thresholds, while performance in higher precision regime is identical. We envision that extending the proposed approach to incorporate multiple scales should improve the performance.

Results on MPII Single Person. Results are shown in Tab. 11.4 and Fig. 11.2 (b). *DeepCut SP AFR-CNN* noticeably improves over *AFR-CNN* alone (79.8 vs. 78.8% PCK, 51.1 vs. 49.0% AUC). The improvement is stronger for smaller thresholds (c.f. Fig. 11.2), as spatial model improves part localization. *Dense-CNN* alone trained on MPII outperforms *AFR-CNN* (81.6 vs. 78.8% PCK), which shows the advantages of dense training and evaluation. As expected, *Dense-CNN* performs slightly better when trained on the larger MPII+LSPET. Finally, *DeepCut Dense-CNN SP* is slightly better than *Dense-CNN* alone leading to the best result on MPII dataset (82.4% PCK).

Comparison to the state of the art (MPII). We compare the performance of *DeepCut* models to our *PS* approach (Chapter 8), the best previous deep learning approaches (Tompson *et al.*, 2014, 2015)[14], and to concurrent approaches (Carreira

---

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

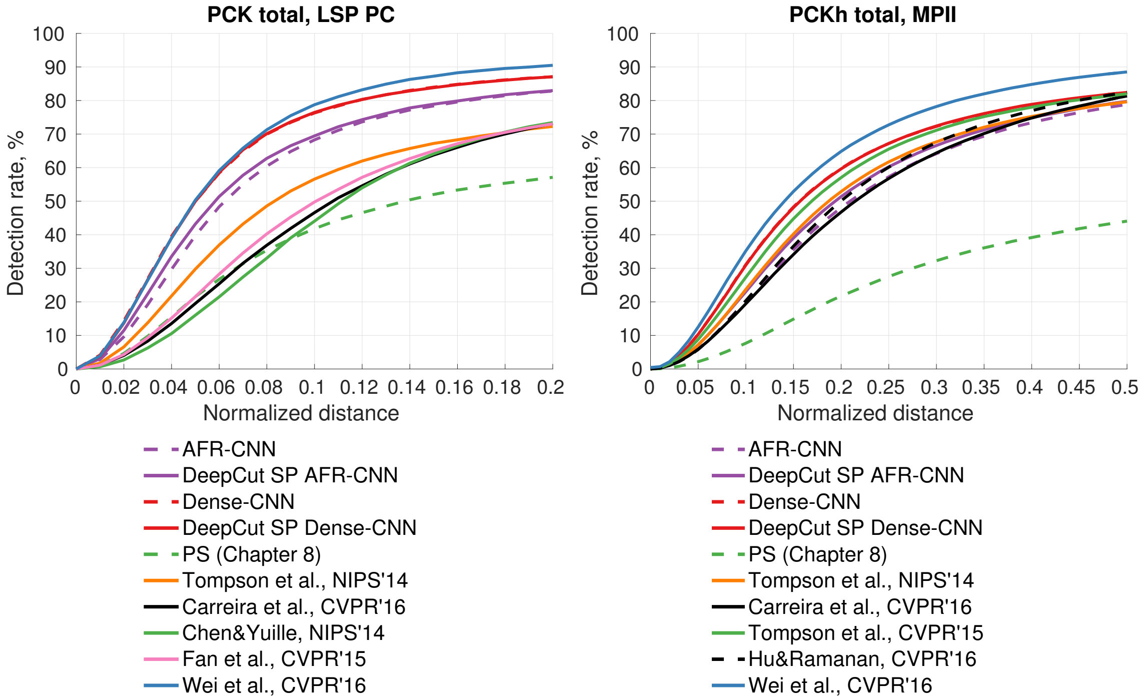
Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK_h	AUC
AFR-CNN (unary)	91.5	89.7	80.5	74.4	76.9	69.6	63.1	78.8	49.0
+ *DeepCut SP*	92.3	90.6	81.7	74.9	79.2	70.4	63.0	79.8	51.1
Dense-CNN (unary)	93.5	88.6	82.2	77.1	81.7	74.4	68.9	81.6	56.0
+LSPET	94.0	89.4	82.3	77.5	82.0	74.4	68.7	81.9	56.5
+*DeepCut SP*	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	56.5
PS (Chapter 8)	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1	24.5
(Tompson *et al.*, 2014)	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6	51.8
(Carreira *et al.*, 2016)	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3	49.1
(Tompson *et al.*, 2015)	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0	54.9
(Hu and Ramanan, 2016)	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4	51.1
(Wei *et al.*, 2016)	**97.8**	**95.0**	**88.7**	**84.0**	**88.4**	**82.8**	**79.4**	**88.5**	**61.4**

Table 11.4: Pose estimation results (PCK_h) on MPII Single Person.

et al., 2016; Hu and Ramanan, 2016; Wei *et al.*, 2016). All deep learning based *DeepCut* models significantly outperform *PS* approach that relies on hand-crafted image representations. *DeepCut SP Dense-CNN* outperforms both (Tompson *et al.*, 2014, 2015) (82.4 vs 79.6 and 82.0% PCK, respectively). Similar to them *DeepCuts* rely on dense training and evaluation of part detectors, but unlike them use single size receptive field and do not include multi-resolution context information. Also, appearance and spatial components of *DeepCuts* are trained piece-wise, unlike (Tompson *et al.*, 2014). We observe that performance differences are higher for smaller thresholds (c.f. Fig. 11.2 (b)). This is remarkable, as a much simpler strategy for location refinement is used compared to (Tompson *et al.*, 2015). Using multi-resolution filters and joint training should improve the performance. *DeepCut* also outperforms concurrent approach (Carreira *et al.*, 2016) (82.4 vs. 81.3% PCK, 56.5 vs. 49.1% AUC). Largest differences are observed for smaller distance thresholds: fully-convolutional part detection architecture with location refinement implemented in *DeepCut* allows for much more precise body part localization compared to the holistic predictions used in (Carreira *et al.*, 2016). *DeepCut* performs on par with another concurrent approach (Hu and Ramanan, 2016) for maximum distance threshold, but significantly outperforms when taking the entire PCK curve into account (56.5 vs. 51.1% AUC), which shows the advantages of the proposed fully-convolutional part detectors. Concurrent approach (Wei *et al.*, 2016) outperforms *DeepCut* (88.5 vs. 82.4% PCK): it incorporates intermediate supervision and uses multi-stage training procedure that allows to increase the size of context seen by the part detectors. We envision that using larger context and intermediate supervision will likely increase the performance of our method as well.

11.5.2 Multi-person pose estimation

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



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

Figure 11.2: Pose estimation results over all PCK thresholds.

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

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

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

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

Table 11.5: Pose estimation results (*m*PCP) on WAF dataset.

the highest PCK$_h$ (Andriluka *et al.*, 2014). Only single pose can be assigned to GT. Unassigned predictions are counted as false positives. Finally, AP for each body part is computed and mAP is reported.

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

Results on WAF. Results are shown in Tab. 11.5. *det ROI* is obtained by extending provided upper body detection boxes. *AFR-CNN det ROI* achieves 57.6% *m*PCP and 73.9% AOP. *DeepCut MP AFR-CNN* significantly improves over *AFR-CNN det ROI* achieving 82.2% *m*PCP. This improvement is stronger compared to LSP and MPII due to several reasons. First, *m*PCP requires consistent prediction of body sticks as opposite to body joints, and including spatial model enforces consistency. Second, *m*PCP metric is occlusion-aware. *DeepCuts* can deactivate detections for the occluded parts thus effectively reasoning about occlusion. This is supported by strong increase in AOP (85.6 vs. 73.9%). Results by *DeepCut MP Dense-CNN* follow the same tendency achieving the best performance of 84.7% *m*PCP and 86.5% AOP. Both increase in *m*PCP and AOP show the advantages of *DeepCuts* over traditional *det ROI* approaches.

Tab. 11.5 shows that *DeepCuts* outperform all prior methods. Deep learning method (Chen and Yuille, 2015) is outperformed both for *m*PCP (84.7 vs. 80.7%) and AOP (86.5 vs. 84.9%) measures. This is remarkable, as *DeepCuts* reason about part interactions across several people, whereas (Chen and Yuille, 2015) primarily focuses on the single-person case and handles multi-person scenes akin to (Yang and Ramanan, 2013). In contrast to (Chen and Yuille, 2015), *DeepCuts* are not limited by the number of possible occlusion patterns and cover person-person occlusions and other types as truncation and occlusion by objects in one formulation. *DeepCuts* significantly outperform (Eichner and Ferrari, 2010) while being more general: unlike (Eichner and Ferrari, 2010) *DeepCuts* do not require person detector and not limited by a number of occlusion states among people.

Qualitative evaluation on WAF dataset. Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* to the traditional two-stage approach *Dense-CNN det ROI* relying on person detector, and to the approach (Chen and Yuille, 2015) on WAF dataset is shown in Fig. 11.3. *det ROI* does not reason about occlusion and often predicts inconsistent body part configurations by linking the parts across the nearby staying people (image 9, right shoulder and wrist of person 2 are linked to the right elbow of person 3; image 10, left elbow of person 4 is linked to the left wrist of person 3). In contrast, *DeepCut MP* predicts body part occlusions, disambiguates multiple and potentially overlapping people and correctly assembles independent detections into plausible body part configurations (image 9, left arms of people 1-3 are correctly predicted to be occluded; image 10, linking of body parts across people 3 and 4 is corrected; image 12, occlusion of body parts is correctly predicted and visible parts are accurately estimated). In contrast to (Chen and Yuille, 2015), *DeepCut MP* better predicts occlusions of person's body parts by the nearby staying people (images 2, 4-6, 8-14), but also by other objects (image 7, left arm of person 1 is occluded by the chair). Furthermore, *DeepCut MP* is able to better cope with strong articulations and foreshortenings (image 1, person 1, 3; image 2 person 1 bottom row; image 3, person 1-2; image 6, person 6; image 8, person 2; image 10, person 4; image 12, person 4; image 13, person 1). Typical *DeepCut MP* failure case is shown in image 15: the right upper arm of person 3 and both arms of person 4 are not estimated due to missing part detection candidates.

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

Results are shown in Tab. 11.6. *DeepCut MP AFR-CNN* improves over *AFR-CNN det ROI* by 4.3% achieving 51.4% AP. The largest differences are observed for the ankle, knee, elbow and wrist, as those parts benefit more from the connections to other parts. *DeepCut MP UB AFR-CNN* using upper body parts only slightly improves over the full body model when compared on common parts (60.5 vs 58.2% AP). Similar tendencies are observed for *Dense-CNN*s, though improvements of *MP UB* over *MP* are more significant.

All *DeepCuts* outperform *Chen&Yuille SP GT ROI*, partially due to stronger part detectors compared to (Chen and Yuille, 2014) (c.f. Tab. 11.3). Another reason is that *Chen&Yuille SP GT ROI* does not model body part occlusion and truncation always predicting the full set of parts, which is penalized by the AP measure. In contrast, our formulation allows to deactivate the part hypothesis in the initial set of part candidates thus effectively performing non-maximum suppression. In *DeepCuts* part hypotheses are suppressed based on the evidence from all other body parts making this process more reliable.

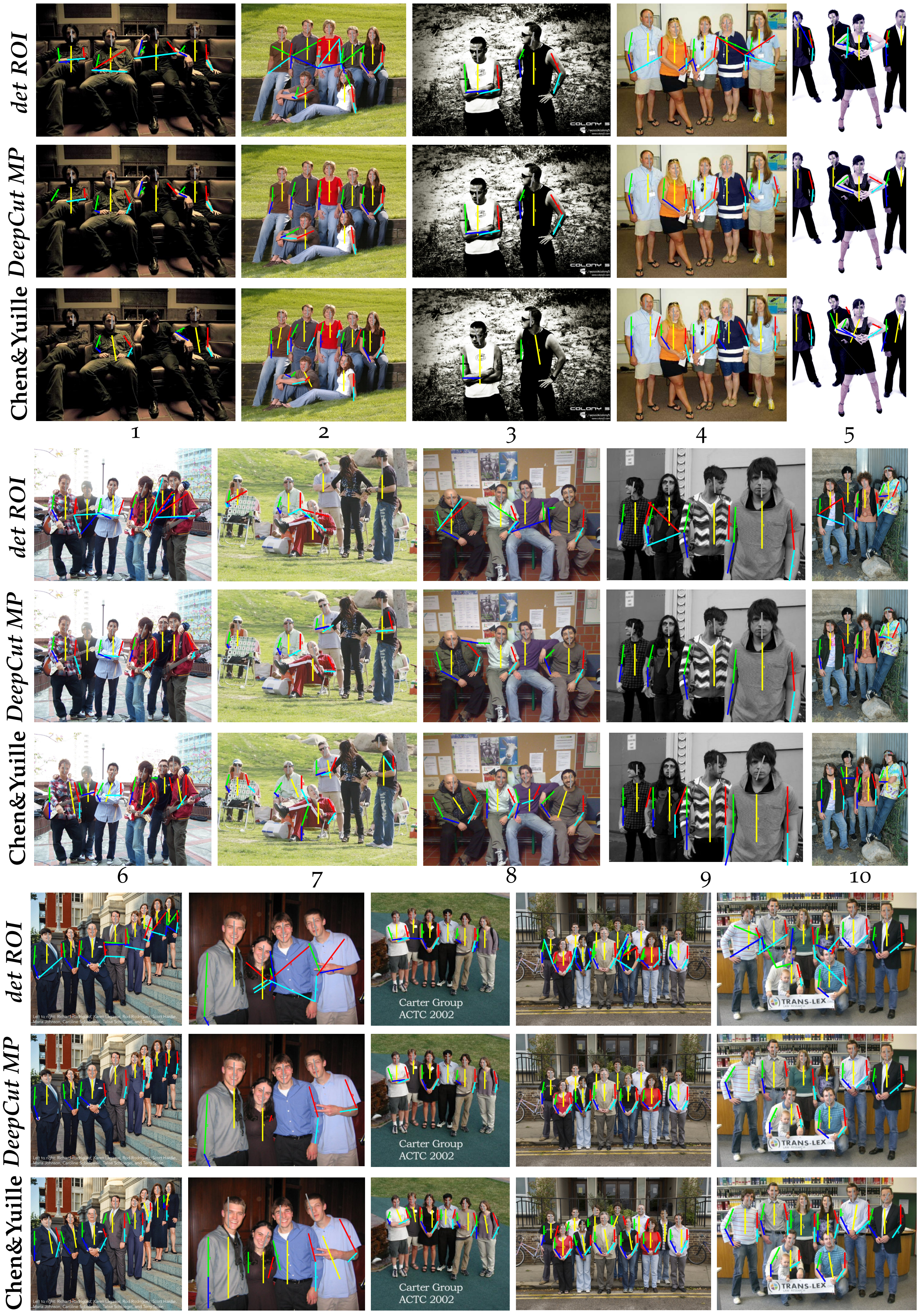


Figure 11.3: Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (rows 2, 5, 8) to the traditional two-stage approach *Dense-CNN det ROI* (rows 1, 4, 7) and to the approach of (Chen and Yuille, 2015) (rows 3, 6, 9) on WAF dataset.

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

Table 11.6: Pose estimation results (AP) on MPII Multi-Person.

Qualitative evaluation on MPII Multi-Person. Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* to the traditional two-stage approach *Dense-CNN det ROI* on MPII Multi-Person dataset is shown in Fig. 11.4 and 11.5. *Dense-CNN det ROI* works well when multiple fully visible individuals are sufficiently separated and thus their body parts can be partitioned based on the person detection bounding box. In this case the strong *Dense-CNN* body part detection model can correctly estimate most of the visible body parts (image 16, 17, 19). However, *Dense-CNN det ROI* cannot tell apart the body parts of multiple individuals located next to each other and possibly occluding each other, and often links the body parts across the individuals (images 1-16, 19-20). In addition, *Dense-CNN det ROI* cannot reason about occlusions and truncations always providing a prediction for each body part (image 4, 6, 10). In contrast, *DeepCut MP Dense-CNN* is able to correctly partition and label an initial pool of body part candidates (each image, top row) into subsets that correspond to sets of mutually consistent body part candidates and abide to mutual consistency and exclusion constraints (each image, row 2), thereby outputting consistent body pose predictions (each image, row 3). $c \neq c'$ pairwise terms allow to partition the initial set of part detection candidates into valid pose configurations (each image, row 2: person-clusters highlighted by dense colored connections). $c = c'$ pairwise terms facilitate clustering of multiple body part candidates of the same body part of the same person (each image, row 2: markers of the same type and color). In addition, $c = c'$ pairwise terms facilitate a repulsive property that prevents nearby part candidates of the same type to be associated to different people (image 1: detections of the left shoulder are assigned to the front person only). Furthermore, *DeepCut MP Dense-CNN* allows to either merge or deactivate part hypotheses thus effectively performing non-maximum suppression and reasoning about body part occlusions and truncations (image 3, row 2: body part hypotheses on the background are deactivated (black crosses); image 6, row 2: body part hypotheses for the truncated body parts are deactivated (black crosses); image 1-6, 8-9, 13-14, row 3: only visible body parts of the partially occluded people are estimated, while non-visible body parts are correctly predicted to be occluded). These qualitative examples show that *DeepCuts MP* can successfully deal with the unknown number of people per image and the unknown number of visible body parts per person.

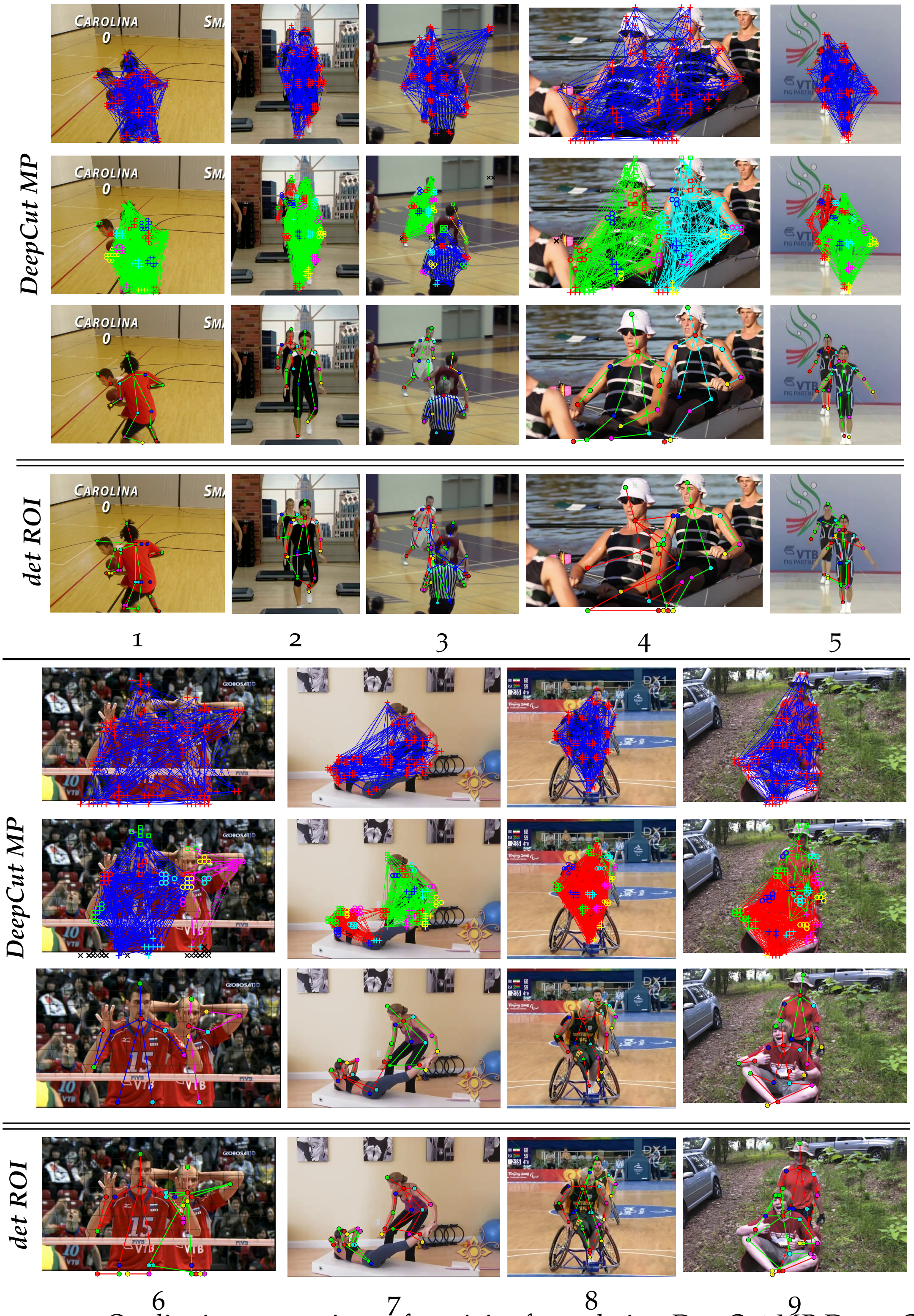


Figure 11.4: Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (rows 1-3, 5-7) to the traditional two-stage approach *Dense-CNN det ROI* (rows 4, 8) on MPII Multi-Person dataset. See Fig. 11.1 for the explanation of color-coding.

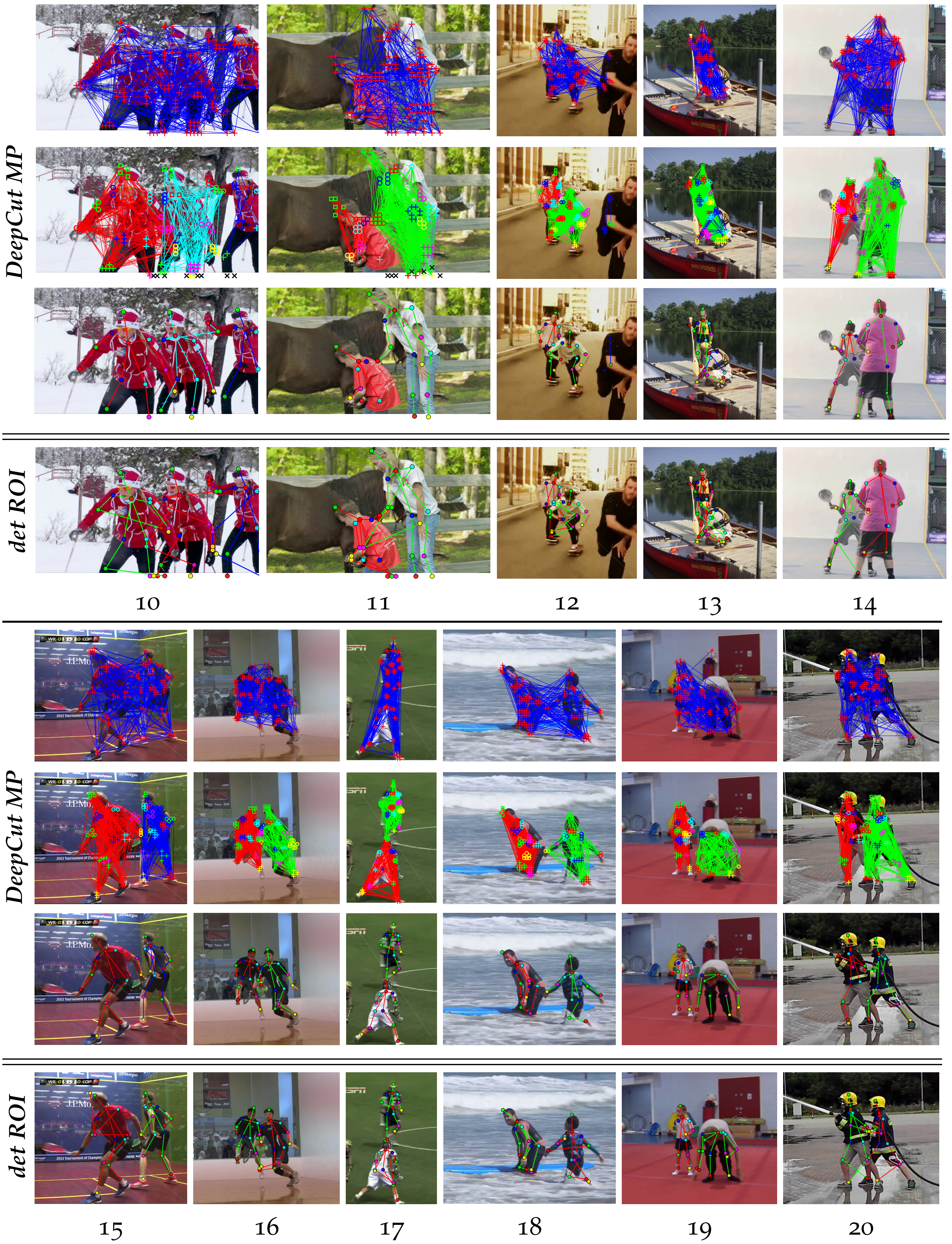


Figure 11.5: Qualitative comparison (contd.) of our joint formulation *DeepCut MP Dense-CNN* (rows 1-3, 5-7) to the traditional two-stage approach *Dense-CNN det ROI* (rows 4, 8) on MPII Multi-Person dataset. See Fig. 11.1 for the explanation of color-coding.

11.6 CONCLUSION

Articulated pose estimation of multiple people in uncontrolled real world images is challenging but of real world interest. In this work, we proposed a new formulation as a joint subset partitioning and labeling problem (SPLP). Different to previous two-stage strategies that separate the detection and pose estimation steps, the SPLP model jointly infers the number of people, their poses, spatial proximity, and part level occlusions. Empirical results on four diverse and challenging datasets show significant improvements over all previous methods not only for the multi-person, but also for the single-person pose estimation problem. On multi-person WAF dataset we improve by 30% PCP over the traditional two-stage approach. This shows that a joint formulation is crucial to disambiguate multiple and potentially overlapping persons. Models and code will be made publicly available.

11.7 APPENDIX: ADDITIONAL RESULTS ON LSP DATASET

We provide additional quantitative results on LSP dataset using person-centric (PC) and observer-centric (OC) evaluation settings.

11.7.1 LSP Person-Centric (PC)

First, detailed performance analysis is performed when evaluating various parameters of *AFR-CNN* and results are reported using PCK (Sapp and Taskar, 2013) evaluation measure. Then, performance of the proposed *AFR-CNN* and *Dense-CNN* part detection models is evaluated using strict PCP (Ferrari *et al.*, 2008) measure.

Detailed *AFR-CNN* performance analysis (PCK). Detailed parameter analysis of *AFR-CNN* is provided in Tab. 11.7 and results are reported using PCK evaluation measure. Respecting parameters for each experiment are shown in the first column and parameter differences between the neighboring rows in the table are highlighted in bold. Re-scoring the 2000 DPM proposals using *AFR-CNN* with AlexNet (Krizhevsky *et al.*, 2012) leads to 56.9% PCK. This is achieved using basis scale 1 ($\approx$ head size) of proposals and training with initial learning rate (lr) of 0.001 for 80k iterations, after which lr is reduced by 0.1, for a total number of 140k SGD iterations. In addition, bounding box regression and default IoU threshold of 0.5 for positive/negative label assignment (Girshick, 2015) have been used. Extending the regions by 4x increases the performance to 65.1% PCK, as it incorporates more context including the information about symmetric body parts and allows to implicitly encode higher-order body part relations into the part detector. No improvements observed for larger scales. Increasing lr to 0.003, lr reduction step to 160k and training for a larger number of iterations (240k) improves the results to 67.4, as higher lr allows for for more significant updates of model parameters when finetuned on the task of human body part detection. Increasing the number of

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AlexNet scale 1, lr 0.001, lr step 80k, # iter 140k, IoU 0.5	82.2	67.0	49.6	45.4	53.1	52.9	48.2	56.9	35.9
AlexNet **scale 4**, lr 0.001, lr step 80k, # iter 140k, IoU 0.5	85.7	74.4	61.3	53.2	64.1	63.1	53.8	65.1	39.0
AlexNet scale 4, **lr 0.003, lr step 160k, # iter 240k**, IoU 0.5	87.0	75.1	63.0	56.3	67.0	65.7	58.0	67.4	40.8
AlexNet scale 4, lr 0.003, lr step 160k, # iter 240k, **IoU 0.4**	87.5	76.7	64.8	56.0	68.2	68.7	59.6	68.8	40.9
AlexNet scale 4, lr 0.003, lr step 160k, # iter 240k, IoU 0.4, **data augment**	87.8	77.8	66.0	58.1	70.9	66.9	59.8	69.6	42.3
AlexNet scale 4, **lr 0.004, lr step 320k, # iter 1M**, IoU 0.4, data augment	88.1	79.3	68.9	62.6	73.5	69.3	64.7	72.4	44.6
+ finetune LSP, lr 0.0005, lr step 10k, # iter 40k	92.9	81.0	72.1	66.4	80.6	77.6	75.0	77.9	51.6
VGG scale 4, lr 0.003, lr step 160k, # iter 320k, IoU 0.4, data augment	91.0	84.2	74.6	67.7	77.4	77.3	72.8	77.9	50.0
+ finetune LSP lr 0.0005, lr step 10k, # iter 40k	**95.4**	**86.5**	**77.8**	**74.0**	**84.5**	**78.8**	**82.6**	**82.8**	**57.0**

Table 11.7: PCK performance of *AFR-CNN* (unary) on LSP (PC) dataset. *AFR-CNN* is finetuned from ImageNet on MPII (lines 1-6, 8), and then finetuned on LSP (lines 7, 9).

training examples by reducing the training IoU threshold to 0.4 results into slight performance improvement (68.8 vs. 67.4% PCK). Further increasing the number of training samples by horizontally flipping each image and performing translation and scale jittering of the ground truth training samples improves the performance to 69.6% PCK and 42.3% AUC. The improvement is more pronounced for smaller distance thresholds (42.3 vs. 40.9% AUC): localization of body parts is improved due to the increased number of jittered samples that significantly overlap with the ground truth. Further increasing the lr, lr reduction step and total number of iterations altogether improves the performance to 72.4% PCK, and very minor improvements are observed when training longer. All results above are achieved by finetuning the AlexNet architecture from the ImageNet model on the MPII training set. Further finetuning the MPII-finetuned model on the LSP training set increases the performance to 77.9% PCK, as the network learns LSP-specific image representations. Using the deeper VGG (Simonyan and Zisserman, 2014b) architecture improves over more shallow AlexNet (77.9 vs. 72.4% PCK, 50.0 vs. 44.6% AUC). Finetuning VGG on LSP achieves remarkable 82.8% PCK and 57.0% AUC. Strong increase in AUC (57.0 vs. 50%) characterizes the improvement for smaller PCK evaluation thresholds. Switching off bounding box regression results into performance drop (81.3% PCK, 53.2% AUC) thus showing the importance of the bounding box regression for better part localization. Overall, we demonstrate that proper adaptation and tweaking of the state-of-the-art generic object detector FR-CNN (Girshick, 2015) leads to a strong body part detection model that dramatically improves over the vanilla FR-CNN (82.8 vs. 56.9% PCK, 57.8 vs. 35.9% AUC) and significantly outperforms the state of the art (+9.4% PCK over the best known PCK result (Chen and Yuille, 2014) and +9.7% AUC over the best known AUC result (Tompson *et al.*, 2014).

Overall performance using PCP evaluation measure. Performance when using the strict "Percentage of Correct Parts (PCP)" (Ferrari *et al.*, 2008) measure is reported in Tab. 11.8. In contrast to PCK measure evaluating the accuracy of predicting body joints, PCP evaluation metric measures the accuracy of predicting body part sticks. *AFR-CNN* achieves 78.3% PCP. Similar to PCK results, *DeepCut SP AFR-CNN* slightly improves over unary alone, as it enforces more consistent predictions of

	Torso	Upper Leg	Lower Leg	Upper Arm	Fore-arm	Head	PCP
AFR-CNN (unary)	93.2	82.7	77.7	75.5	63.5	91.2	78.3
+ *DeepCut SP*	93.3	83.2	77.8	76.3	63.7	91.5	78.7
+ appearance pairwise	93.4	83.5	77.8	76.6	63.8	91.8	78.9
+ *DeepCut MP*	93.6	83.3	77.6	76.3	63.5	91.2	78.6
Dense-CNN (unary)	96.2	87.8	81.8	81.6	**72.3**	95.6	83.9
+ *DeepCut SP*	**97.0**	**88.8**	**82.0**	**82.4**	71.8	**95.8**	**84.3**
+ *DeepCut MP*	96.4	**88.8**	80.9	**82.4**	71.3	94.9	83.8
PS (Chapter 8)	88.7	63.6	58.4	46.0	35.2	85.1	58.0
(Tompson *et al.*, 2014)	90.3	70.4	61.1	63.0	51.2	83.7	66.6
(Chen and Yuille, 2014)	96.0	77.2	72.2	69.7	58.1	85.6	73.6
(Fan *et al.*, 2015)*	95.4	77.7	69.8	62.8	49.1	86.6	70.1
PS (Chapter 8)	88.7	63.6	58.4	46.0	35.2	85.1	58.0
(Wang and Li, 2013b)	87.5	56.0	55.8	43.1	32.1	79.1	54.1

* re-evaluated using the standard protocol, for details see project page of (Fan *et al.*, 2015)

Table 11.8: Pose estimation results (PCP) on LSP (PC) dataset.

body part sticks. Using more general multi-person *DeepCut MP AFR-CNN* model results into similar performance, which shows the generality of *DeepCut MP* method. *DeepCut SP Dense-CNN* slightly improves over *Dense-CNN* alone (84.3 vs. 83.9% PCP) achieving the best PCP result on LSP dataset using PC annotations. This is in contrast to PCK results where performance differences *DeepCut SP Dense-CNN* vs. *Dense-CNN* alone are minor.

We now compare the PCP results to the state of the art. The *DeepCut* models outperform all other methods by a large margin. The best known PCP result by Chen&Yuille (Chen and Yuille, 2014) is outperformed by 10.7% PCP. This is interesting, as their deep learning based method relies on the image conditioned pairwise terms while our approach uses more simple geometric only connectivity. Interestingly, *AFR-CNN* alone outperforms the approach of Fan et al. (Fan *et al.*, 2015) (78.3 vs. 70.1% PCP), who build on the previous version of the R-CNN detector (Girshick *et al.*, 2014). At the same time, the best performing dense architecture *DeepCut SP Dense-CNN* outperforms (Fan *et al.*, 2015) by +14.2% PCP. Surprisingly, *DeepCut SP Dense-CNN* dramatically outperforms the method of Tompson et al. (Tompson *et al.*, 2014) (+17.7% PCP) that also produces dense score maps, but additionally includes multi-scale receptive fields and jointly trains appearance and spatial models in a single deep learning framework. We envision that both advances can further improve the performance of *DeepCut* models. Finally, all proposed approaches significantly outperform earlier non-deep learning based methods (Wang and Li, 2013b, Chapter 8) relying on hand-crafted image features.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AFR-CNN (unary)	95.3	88.3	78.5	74.2	87.3	84.2	81.2	84.2	58.1
Dense-CNN (unary)	**97.4**	**92.0**	**83.8**	**79.0**	**93.1**	**88.3**	**83.7**	**88.2**	**65.0**
PS (Chapter 8)	87.5	77.6	61.4	47.6	79.0	75.2	68.4	71.0	45.0
(Chen and Yuille, 2014)	91.5	84.7	70.3	63.2	82.7	78.1	72.0	77.5	44.8
(Ouyang *et al.*, 2014)	86.5	78.2	61.7	49.3	76.9	70.0	67.6	70.0	43.1
(Kiefel and Gehler, 2014)	83.5	73.7	55.9	36.2	73.7	70.5	66.9	65.8	38.6
(Ramakrishna *et al.*, 2014)	84.9	77.8	61.4	47.2	73.6	69.1	68.8	69.0	35.2

Table 11.9: Pose estimation results (PCK) on LSP (OC) dataset.

11.7.2 LSP Observer-Centric (OC)

We now evaluate the performance of the proposed part detection models on LSP dataset using the observer-centric (OC) annotations (Eichner and Ferrari, 2012a). In contrast to the person-centric (PC) annotations used in all previous experiments, OC annotations do not penalize for the right/left body part prediction flips and count a body part to be the right body part, if it is on the right side of the line connecting pelvis and neck, and a body part to be the left body part otherwise.

Evaluation is performed using the official OC annotations provided by (Eichner and Ferrari, 2012a). Prior to evaluation, we first finetune the *AFR-CNN* and *Dense-CNN* part detection models from ImageNet on MPII and MPII+LSPET training sets, respectively, (same as for PC evaluation), and then further finetuned the models on LSP OC training set.

PCK evaluation measure. Results using OC annotations and PCK evaluation measure are shown in Tab. 11.9 and in Fig. 11.6. *AFR-CNN* achieves 84.2% PCK and 58.1% AUC. This result is only slightly better compared to *AFR-CNN* evaluated using PC annotations (84.2 vs 82.8% PCK, 58.1 vs. 57.0% AUC). Although PC annotations correspond to a harder task, only small drop in performance when using PC annotations shows that the network can learn to accurately predict person's viewpoint and correctly label left/right limbs in most cases. This is contrast to earlier approaches based on hand-crafted features whose performance drops much stronger when evaluated in PC evaluation setting (e.g. Chapter 8) drops from 71.0% PCK when using OC annotations to 58.0% PCK when using PC annotations). Similar to PC case, *Dense-CNN* detection model outperforms *AFR-CNN* (88.2 vs. 84.2% PCK and 65.0 vs. 58.1% AUC). The differences are more pronounced when examining the entire PCK curve for smaller distance thresholds (c.f. Fig. 11.6).

Comparing the performance by *AFR-CNN* and *Dense-CNN* to the state of the art, we observe that both proposed approaches significantly outperform other methods. Both deep learning based approaches of Chen&Yuille (Chen and Yuille, 2014) and Ouyang et al. (Ouyang *et al.*, 2014) are outperformed by $+10.7$ and $+18.2\%$ PCK when compared to the best performing *Dense-CNN*. Analysis of PCK curve for the

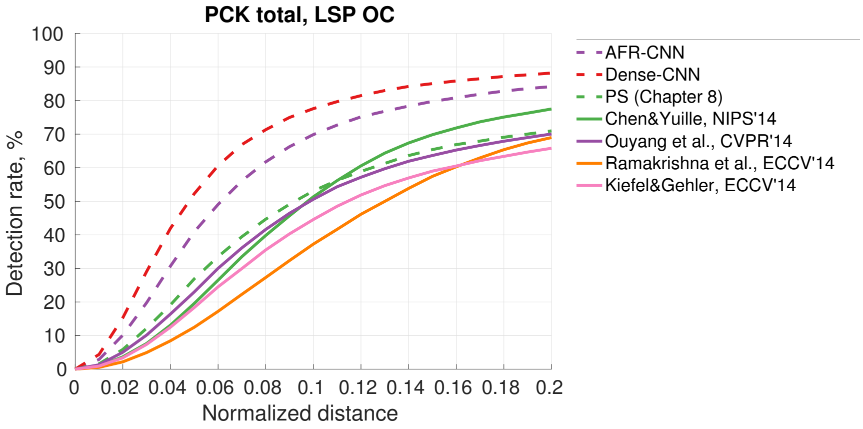


Figure 11.6: Pose estimation results over all PCK thresholds on LSP (OC) dataset.

	Torso	Upper Leg	Lower Leg	Upper Arm	Fore-arm	Head	PCP
AFR-CNN (unary)	92.9	86.3	79.8	77.0	64.2	91.8	79.9
Dense-CNN (unary)	**96.0**	**91.0**	**83.5**	**82.8**	**71.8**	**96.2**	**85.0**
PS (Chapter 8)	88.7	78.9	73.2	61.8	45.0	85.1	69.2
(Chen and Yuille, 2014)	92.7	82.9	77.0	69.2	55.4	87.8	75.0
(Ouyang *et al.*, 2014)	88.6	77.8	71.9	61.9	45.4	84.3	68.7
(Kiefel and Gehler, 2014)	84.3	74.5	67.6	54.1	28.3	78.3	61.2
(Ramakrishna *et al.*, 2014)	88.1	79.0	73.6	62.8	39.5	80.4	67.8

Table 11.10: Pose estimation results (PCP) on LSP (OC) dataset.

entire range of PCK distance thresholds reveals even larger performance differences (c.f. Fig. 11.6). The results using OC annotations confirm our findings from PC evaluation and clearly show the advantages of the proposed part detection models over the state-of-the-art deep learning methods (Chen and Yuille, 2014; Ouyang *et al.*, 2014), as well as over earlier pose estimation methods based on hand-crafted image features (Kiefel and Gehler, 2014; Ramakrishna *et al.*, 2014, Chapter 8).

PCP evaluation measure. Results using OC annotations and PCP evaluation measure are shown in Tab. 11.10. Overall, the trend is similar to PC evaluation: both proposed approaches significantly outperform the state-of-the-art methods with *Dense-CNN* achieving the best result of 85.0% PCP thereby improving by +10% PCP over the best published result (Chen and Yuille, 2014).

CONCLUSIONS AND FUTURE PERSPECTIVES 12

Contents

12.1 Discussion of contributions **208**
12.2 Future perspectives **213**
  12.2.1 Obtaining representative training data with relevant variations 213
  12.2.2 Building expressive models for human pose estimation . . . 214
  12.2.3 Benchmarking and analyzing the state of the art 216
  12.2.4 A broader view on the topic 217

ARTICULATED people detection and human pose estimation have been significantly advanced over the last years. To a large extend the success of the current methods can be accounted for strong appearance models relying either on hand-crafted image features boosted with non-parametric decision forests (Benenson *et al.*, 2014; Zhang *et al.*, 2015), or multi-layer image representations completely learned from data using deep learning methods (Toshev and Szegedy, 2014; Chen and Yuille, 2014; Tompson *et al.*, 2014; Wei *et al.*, 2016). This observation leads to three implications. First, as the complexity of methods increases and so the number of model parameters that have to be estimated from the data, large representative training sets are crucial for the best performance. Second, developing expressive spatial models for human pose estimation has enjoyed less attention in the literature, but becomes crucial when multiple detections of body parts have to be grouped into valid pose configurations and correctly assigned among potentially multiple individuals present in the image. Third, with the rapid progress of human pose estimation over the last years, comprehensive benchmarks are required for fair comparison and thorough performance analysis of highly competing approaches. Thus, in this thesis we investigated three directions towards advancing articulated people detection and pose estimation: (i) *obtaining representative training data with relevant variations*, (ii) *building expressive models for human pose estimation*, and (iii) *benchmarking and analyzing the state of the art*. In the following, we briefly summarize the thesis w.r.t. the three directions and discuss our contributions and future perspectives.

First, we examined multiple ways of *obtaining representative training data with relevant variations*. More specifically, we proposed a range of automatic data generation methods that allow to directly encode relevant variations into the training data. At the core of our methods we used a state-of-the-art statistical 3D human shape model (Jain *et al.*, 2010) from computer graphics. Sampling from the underlying human shape distribution and a large set of human poses allowed us to generate novel samples with controllable shape and pose variations that are relevant for the task at hand. Furthermore, we improved the 3D human body shape model by building efficient and expressive shape spaces from a large commercially available 3D body shape dataset (Robinette *et al.*, 1999).

The second direction of this thesis, *building expressive models for human pose estimation*, was concerned with exploring ways of developing expressive spatial and strong appearance models for 2D single- and multi-person pose estimation. We proposed an expressive single person pose estimation model that incorporates higher order part dependencies while remaining efficient, and explored various types of appearance representations aiming to substantially improve the body part hypotheses. Furthermore, we proposed an expressive model for joint pose estimation of multiple people. To that end, we develop strong deep learning based body part detectors and an expressive fully connected spatial model. The proposed approach infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other. We demonstrated significant improvements over the state of the art on single person and multi-person pose estimation tasks.

In the third direction explored in this thesis, *benchmarking and analyzing the state of the art*, we performed a thorough evaluation and performance analysis of prominent human pose estimation and activity recognition methods. To that end, we introduced a novel benchmark that makes a significant advance in terms of diversity and difficulty, compared to the current datasets, and includes over $40,000$ annotated people. We provided a rich set of labels that allowed for detailed performance analysis of prominent approaches gaining insights into successes and failures of these methods.

Furthermore, we advanced the field by making the source code and data freely available to the community. First, we released the source code and learned models for our image conditioned human pose estimation approach thus allowing other researchers to build directly on the best practices in human pose estimation. Second, we made an effort to collect, annotate and release for public usage a new comprehensive large scale benchmark that aims to unify the work in 2D human pose estimation. Third, we released a state-of-the-art human body shape model learned from a large commercially available dataset, as well as the pre-processed data used for learning; in addition, we made the code for data pre-processing, model building and fitting publicly available. Finally, we are currently working on releasing our state-of-the-art deep learning based part detectors and powerful multi-person pose estimation model. We believe that the deliverables of this thesis in terms of code and data is a valuable contribution accelerating the dynamic development of the field.

12.1 DISCUSSION OF CONTRIBUTIONS

The overall goal of this thesis was to advance the articulated people detection and pose estimation in challenging real world scenarios. Towards this goal we investigated three orthogonal directions by looking at the training data, expressive models, and proper ways of benchmarking the competing approaches. In the following we would like to discuss the steps we performed towards these goals and contributions of this thesis with respect to the individual chapters.

First, in Chapter 3 we demonstrated that a state-of-the-art 3D human shape model from the computer graphics can be successfully used to learn powerful people detection models. We contributed a novel training data generation method that relies on the parametric body shape model to generate thousands of photo-realistically looking synthetic training samples from only a few persons and views. We demonstrated that surprisingly good results can be obtained from as few as one or two people only and that comparable results can be obtained already with eleven people. We directly compared to people detection systems based on the well-known pictorial structures model (Andriluka *et al.*, 2009), as well as the Histogram of oriented gradients (HOG) model (Dalal and Triggs, 2005) trained using the standard real data and in both settings showed the improvements when using our automatically generated synthetic training data. Furthermore, we directly compared to the main competitor (Marin *et al.*, 2010) that uses game engine data to train people detectors. We retrained both people detection systems using their game engine data and in both cases demonstrated significant performance improvements when using our photo-realistically rendered synthetic training samples over the computer game engine data.

Second, in Chapter 4 we showed that a 3D human shape model can be used directly to enrich an existing training data with relevant shape information in order to learn more powerful people detection models. We demonstrated that complementary shape information sampled from the underlying 3D human shape distribution can be directly incorporated into the low level feature representation via non-photo-realistically rendered training examples. We showed that although in no stage of this pipeline photo-realistic images are produced, by careful design of the rendering procedure our feature representation can generalize well from synthetic training data to unseen real test data. We contributed a thorough experimental analysis of different parameters of the data generation pipeline and analyzed different combinations of real and synthetic training data. Our experiments on people detection showed that the combination of real and large amounts of synthetic data sampled from a previously learned 3D human shape distribution allows to train a detector which outperforms models trained from real data only, synthetic training data generated from computer games (Marin *et al.*, 2010), and their combinations.

Third, we analyzed the advantages and limitations of both initially proposed data generation methods and developed a novel approach that enables automatic generation of numerous *photo-realistically* looking synthetic training examples from *arbitrary monocular* images with annotated human body poses (Chapter 5). We used a 3D human shape model to produce a set of realistic shape deformations of person's appearance, and combined them with motion capture data to produce a set of feasible pose changes. This allowed us to generate realistically looking training images of people where we have full control over the shape and pose variations. We evaluated our data generation method on the task of articulated human detection and on the task of human pose estimation. On both tasks we could significantly improve performance when the training sets are extended with the automatically generated images. Motivated by the very good articulated people detection performance, we proposed a joint model that directly integrates evidence from an appropriately trained deformable part model (DPM, (Felzenszwalb *et al.*, 2010)) into a pictorial structures framework and demonstrated that this joint model further improves performance. Furthermore, we advanced the field by contributing a new challenge of joint detection and pose estimation of multiple articulated people in challenging real world scenes.

Fourth, in Chapter 6 we improved the state-of-the-art 3D human shape model (Jain *et al.*, 2010) used in our data generation methods. This model was learned from the largest publicly available dataset consisting of rather small number of human scans lacking diversity in represented human shapes. We contributed by rebuilding the 3D human shape model from a large commercially available scan database (Robinette *et al.*, 1999), and making the resulting model available to the community. As preprocessing several thousand scans for learning the model is a challenge in itself, we contributed by developing robust best practice solutions for scan alignment that quantitatively lead to the best learned models and made the implementations of these pre-processing steps also publicly available. We performed extensive experimental evaluation and demonstrated the improved accuracy and generality of the shape model. Furthermore, we experimentally showed its improved performance for human body reconstruction from sparse input data. The published code and models have been downloaded numerous times which demonstrates the impact of this work in the field. Furthermore, in (Wuhrer *et al.*, 2014) we explored ways of improving the 3D body shape and posture estimation under clothing and proposed a novel method that uses a posture-invariant shape space to model body shape variation combined with a skeleton-based deformation to model changes in pose. We demonstrated that using the proposed posture-invariant shape space allows to achieve higher accuracy when fitting the shape model to a dressed individual, compared to a canonical shape model.

Fifth, we contributed an expressive human pose estimation model in Chapter 7. We observed that despite high variability of body articulations, human motions and activities often simultaneously constrain the positions of multiple body parts and proposed a model that incorporates higher order part dependencies while remaining efficient. To that end, we defined a conditional model in which all body parts are connected a-priori, but which becomes a tractable tree-structured pictorial structures model once the mid-level features are available. We analyzed different choices of particular mid-level image representations used for conditioning our model and choose the non-parametric poselet representation introduced in (Bourdev *et al.*, 2010). We showed that this representation works particularly well as it jointly models appearance of multiple body parts. We performed a thorough evaluation of different model's components and showed their contribution to the final performance. Furthermore, we demonstrated the potential of the proposed approach by analyzing the performance in the ideal case. We showed the effectiveness of our model on three publicly available pose estimation benchmarks improving or being on-par with the competing approaches in each case.

Sixth, in Chapter 8, we analyzed and drew on several recently proposed powerful ideas such as strong local appearance models, flexible spatial models and our image conditioned method. We explored various types of appearance representations including rotation invariant or rotation specific appearance templates, mixtures of such local templates, specialized models tailored to appearance of salient body parts such as head and torso, and semi-global representations based on poselet features. Then we combined the improved appearance model with more expressive body representations including the flexible models of (Sapp *et al.*, 2011; Yang and Ramanan, 2011) and our image conditioned spatial model. Starting with the basic tree-structured pictorial structures we perform a series of experiments incrementally adding various components and analyzing the resulting performance gains. Our analysis resulted in several important conclusions: (1) we showed that the proposed appearance representations operating at different levels of granularity (mixtures of local templates vs. semi-global poselets) are complementary; (2) we demonstrated that even a basic tree-structure spatial human body model achieves very good performance when augmented with the proper appearance representation; and (3) we showed that the combination of the best performing appearance model with a flexible image-conditioned spatial model achieves the best result, significantly improving over many competing methods on prominent pose estimation benchmarks. We made the implementation of our best performing model publicly available. The source code was downloaded numerous times which shows a high interest of the community.

Seventh, we collected, labeled and released for public usage a novel comprehensive benchmark for 2D human pose estimation and established a set of performance analysis tools (Chapter 9). Compared to current human pose estimation datasets limited in their coverage of the overall pose estimation challenges, our benchmark made significant advance in terms of diversity and difficulty. We collected this comprehensive dataset from YouTube videos using an established taxonomy of several hundreds of everyday human activities (Ainsworth *et al.*, 2011). The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational and householding activities, and capture people from a wider range of viewpoints. Furthermore, we contributed a rich set of labels including positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels. The dataset contains over $40,000$ annotated poses and over 1.5M frames. We released the dataset for public usage. In addition to the dataset we contributed a detailed analysis of several prominent human pose estimation methods on our novel benchmark. We defined a set of quantitative complexity measures that map rich body image annotations to a real value that relates the complexity of the image to human pose estimation challenges. Based on these complexity measures we contributed a set of performance analysis tools. In addition, we also established novel evaluation measures intending to overcome the shortcomings of the current metrics. We completed a detailed performance analysis of prominent pose estimation methods, identified their strengths and drawbacks and proposed the most promising future research directions.

Eighths, using our comprehensive benchmark we contributed a thorough performance analysis of popular holistic and pose based activity recognition methods in Chapter 10. Similar to human pose estimation, we defined a set of activity recognition specific complexity measures characterizing the scene difficulty w.r.t. activity recognition challenges. We contributed an extensive experimental evaluation of individual activity recognition methods and their combinations and discovered a number of factors responsible for successes and failures of holistic and pose based methods. In a series of experiments we discovered that holistic and pose based methods are complementary, and their performance varies significantly depending on the activity. We demonstrated that both methods are strongly affected by the speed of trajectories. While the holistic method is also strongly influenced by the number of trajectories, pose based methods are strongly affected by human pose and viewpoint. Motivated by our experimental analysis showing that the holistic and pose estimation methods are highly complementary, we proposed a novel activity recognition approach based on the combination of both methods and showed empirically that the proposed approach achieves the best performance.

Finally, in Chapter 11 we contributed a novel multi-person pose estimation model. We proposed strong deep learning based appearance representations for body part detection. Building on strong appearance models we proposed an expressive spatial model for joint pose estimation of multiple people. This was achieved by treating the multi-person pose estimation as a joint partitioning and labeling problem of a set of body part hypotheses. Our formulation implicitly performs non-maximum suppression on the set of part detections and groups them to form configurations of body parts that respect geometric and appearance constraints. Our model is able to infer the number of persons in a scene, identify occluded body parts, and disambiguate body parts between people in close proximity of each other. The proposed formulation is an integer linear program and therefore allows the use of robust optimization techniques and allows for the computation of bounds and feasible solutions with a certified optimality gap. We demonstrated significant improvements over the state of the art for single-, as well as multi-person pose estimation on challenging public benchmarks. Furthermore, in (Insafutdinov *et al.*, 2016) we further improved the proposed multi-person model by re-visiting and significantly improving each of its key ingredients. In particular, we (1) significantly improved part detectors by building on extremely deep architectures; (2) introduced novel image conditioned pairwise terms by learning to regress from each body part location onto locations of all other parts during CNN training; and (3) proposed incremental optimization strategies that significantly reduce run-time while improving pose estimation accuracy. Proposed improvements allowed to significantly boost multi-person and single-person pose estimation performance while dramatically reducing run-time. Finally, in (Elhayek *et al.*, 2015) we also addressed a problem of multi-person 3D pose estimation. To that end, we developed a novel approach for accurate marker-less capture of 3D articulated skeleton motion of several subjects in general scenes. Combining a discriminative image based joint detection method based on deep learning with a model based generative motion tracking algorithm allowed us to track full articulated joint angles at the state-of-the-art accuracy and temporal stability with as few as two cameras.

12.2 FUTURE PERSPECTIVES

We now first discuss future work w.r.t. the different directions of this thesis. Then, we conclude this section with giving a broader view on the topic.

12.2.1 Obtaining representative training data with relevant variations

Generative 3D human pose model. Synthetic data generation methods proposed in this thesis allow to generate novel training samples by sampling 3D body shape and pose parameters. While shape parameters are sampled from the underlying continuous 3D shape space learned from 3D body scans using PCA, pose parameters are sampled from a large database of discrete 3D human poses. The main draw back of the pose sampling step is that the poses of novel training samples are restricted to the exemplars present in the database. On the other hand, it has been shown that 3D human pose can be successfully represented as a linear combination of a sparse set of bases learned from 3D motion capture data (Ramakrishna *et al.*, 2012; Wang *et al.*, 2014). We believe that combining our generative human shape spaces with a generative 3D human pose model is a promising direction to improve the synthetic data generation methods proposed in thesis.

Generative human appearance model. We have shown in Chapter 5 that 3D body shape model can be used to generate a large number of photo-realistically looking samples with controllable shape and pose variations from arbitrary monocular images by reshaping and animating original images. We also showed that training from these samples allows to learn powerful detection and pose estimation models of highly articulated people. However, image morphing requires an expensive manual foreground segmentation for training samples. Furthermore, it may introduce morphing artifacts that make the generated image look unrealistic. Therefore, it might be beneficial to learn a full 3D appearance model of dressed individuals and use this model directly to generate photo-realistically looking synthetic training samples. While learning such a model is a challenge on its own, first steps have already been made by other researchers in the field, e.g. (Guan *et al.*, 2012).

Learning from the generative body model directly. In this thesis we proposed several approaches to learn better people detection and human pose estimation models from the 3D human body model. In all cases, the relevant information is transferred from the 3D body model via an intermediate step of rendering synthetic training samples. However, the better solution would be to learn directly from the 3D shape model without intermediate rendering of training examples. This would make the learning process more efficient and would allow to avoid the information loss due to rendering and image morphing. We thus leave this promising research direction for the future work.

Active and weakly-supervised learning. In this thesis we contributed a comprehensive benchmark for 2D human pose estimation and human activity recognition with 40,000 labeled poses. Availability of this dataset to the research community has become one the key factors to success of recent strongly supervised deep learning approaches to human pose estimation (Tompson *et al.*, 2014; Wei *et al.*, 2016, Chapter 11). However, while collecting and labeling the data, we realized that manual collection and annotation of large amounts of data is a tedious process. Although using Amazon Mechanical Turk (AMT) allowed to scale the data labeling process to tens of thousands of human poses, it required significant effort of establishing annotation infrastructure, managing the turkers and performing a thorough quality assurance of the manually labeled images. At the same time, large amounts of weakly labeled data are freely available from various Internet sources, such as YouTube or Flickr. Weak labels are typically provided in form of user tags describing objects and activities. These labels can potentially be used to facilitate the mining of novel training samples for the task of human pose estimation. While a substantial body of the active and weakly-supervised learning literature exist for general object recognition (e.g., (Ebert *et al.*, 2012; Liang and Grauman, 2014; Mac Aodha *et al.*, 2014)), active and weakly-supervised learning in the context of human pose estimation remains largely unaddressed. Thus, in the future we would like to explore the ways of automatically selecting novel training samples from large amounts of the weakly labeled data.

12.2.2 Building expressive models for human pose estimation

Robust and versatile mid-level features. In this thesis we showed that mid-level representations based on semi-global poselet detectors can be successfully used to condition spatial and appearance models. However, poselet detectors rely on hand-crafted HOG feature based image representations. (Chen and Yuille, 2014) has shown that using deep learning CNN features as a mid-level representation allows to significantly improve pose estimation performance thereby outperforming our Poselet Conditioned PS on LSP dataset (Johnson and Everingham, 2010) (75.0 vs. 69.2% PCP). We envision that using deep learning features as a mid-level representation should significantly improve our image conditioned method as well.

Conditioning on local information. Further analyzing possible improvements of our Poselet Conditioned PS (Chapter 7 and Chapter 8) we note that currently conditioning is performed globally using the content of the entire image. This implies that during the inference a fixed set of globally predicted pairwise and unary mixture components is used, and selection of each mixture component is not affected by any particular location in the image. Conditioning on local information, similar as it is done by (Chen and Yuille, 2014), should add the flexibility to the model and allow to recover from the wrong predictions of some of the components. We thus would like to investigation this direction in the future.

Applications beyond pose estimation. In this thesis we considered human pose estimation as a major application of image conditioned approaches. However, we envision that image conditioned spatial and appearance models can be successfully applied for other computer vision tasks as well. (Tang *et al.*, 2012, 2013) showed that modeling occlusion patterns by multiple specifically tackled components is advantageous in the context of people detection in crowded scenes and (Pepik *et al.*, 2013) used similar ideas for detection of densely parked cars. In both lines of research DPM (Felzenszwalb *et al.*, 2010) was used for component modeling, and component selector was modeled as a latent variable. Conditioning component selection based on mid-level representation that, e.g. could provide a rough information on the number of objects in the scene, is a possible way of improving both methods. Another task that could potentially profit from the image conditioned models is human activity recognition in unconstrained environments. We have shown in Chapter 10 that this is a hard problem due to high intra- and low inter-class variability that is hard to capture by a single model. Using image conditioned approach that allows to selection a specific activity model trained to tackle particular set of activities is a potential way of improving the performance of fine grained human activity recognition.

Joint training. In Chapter 11 we proposed an expressive multi-person pose estimation model that achieves state-of-the-art results on several public benchmarks. The latter is very encouraging when taking into account that appearance and spatial components of the model are trained piece-wise. We envision that joint training of both components will further improve the performance. We thus would like to explore ways of combining the spatial and appearance components in a single deep learning framework that can be trained end-to-end from the raw image data to output the unary and pairwise probabilities. Joint training will not only balance the unary and pairwise potentials in a better way, but will also allow for richer interactions between the potentials.

Richer outputs. In this thesis we focused on developing human pose estimation models that produce reliable estimations of 2D body joint locations of single (Chapters 7 and 8) and multiple (Chapter 11) people. However, other works have shown that single image based human pose estimation models can be successfully extended to produce richer outputs, such as foreground/background segmentations of body parts (Ladicky *et al.*, 2013), or 3D body part hypotheses (Andriluka *et al.*, 2010; Wang *et al.*, 2014). We envision that our models can be extended as well to output richer hypotheses beyond 2D joint locations. One possible direction of research is to extend our dense fully-convolutional body part detection model proposed in Chapter 11 to output the per-part semantic segmentation. Since the model has been developed to train from dense ground truth scoremaps and output dense score predictions at test time, it can easily be adapted for body part segmentation. Another possible directions of research is to combine the 3D body shape model with the 2D human pose estimation model. 3D body shape information and 3D kinematic and anthropometric constraints encoded into the 3D body shape model can significantly constrain the search of plausible 3D human pose configurations. At the same time, in order to achieve a reliable lifting from 2D joint coordinates to 3D part hypotheses, the 2D human pose estimation model can be extended to infer intermediate 2.5D representations, such as depth ordering of body parts. These intermediate representations should be much easier to annotate in the training data than full 3D body poses and also easier to infer at test time given the image observations.

Explore temporal information. Our work on human pose estimation presented in this thesis assumes that a static monocular image is available at the test time. However, as increasingly more visual information is produced in form of video content, the assumption that the entire sequence of images is available may be often fulfilled nowadays. It has been shown that using the motion information between the neighboring frames and extending the inference in the temporal domain allows to improve the pose estimation in each individual frame (Sapp *et al.*, 2011; Weiss and Taskar, 2013; Tokola *et al.*, 2013; Pfister *et al.*, 2015). On the other hand, it has been shown that a multi-cut formulation can be successfully used for joint non-maximum suppression and tracking of multiple people (Tang *et al.*, 2015). We envision that our multi-cut based method presented in Chapter 11 can be extended for tracking of individual body parts of multiple people over time. This extension should allow in improve labeling and grouping of body parts in each frame and jointly reason about part occlusions over the entire video sequence.

12.2.3 Benchmarking and analyzing the state of the art

Richer annotations. Our comprehensive dataset introduced in Chapter 9 includes not only 2D body joint location and visibility labels, but also richer annotations, such as visibility of body part sticks, continuous 3D orientation of head and torso, and activity labels. In the future we plan to extend the set of labels in order to facilitate developing of methods outputting a richer set of hypothesis, such as 3D body poses and body part segmentations. First, we would like to annotate the depth ordering of body parts. Labeling depth layers is much easier compared to the full 3D body pose annotation. At the same time it enables learning of intermediate representations that allow to resolve depth ambiguities and bridge the gap between 2D body part predictions and corresponding 3D poses. Second, we would like to obtain a per-pixel body part segmentation. This would allow for training and benchmarking of semantic per-pixel body part labeling methods. The problem of semantic body part segmentation is highly overlooked in the community, mostly due to unavailability of the large datasets providing per-pixel body part segmentation labels. Thus, extending our benchmark with corresponding labels should facilitate the development of this research direction.

Online evaluation tool. In order facilitate the fair comparison of pose estimation methods on our benchmark and to prevent accidentally tuning on the test set, we withheld the annotations for the test images and perform evaluation on demand using the developed performance analysis tools. However, as more and more people are interested into the dataset, we would like to make the evaluation procedure more convenient by establishing an online evaluation tool following the tools currently available for evaluation of optical flow (Baker *et al.*, 2011; Butler *et al.*, 2012) and computer vision tasks in automotive setting (Geiger *et al.*, 2012). User interface will allow to upload the predictions and automatically re-compute all results using the performance analysis tools we created. We believe that this will further increase the interest to the dataset and make it a golden standard for evaluation of 2D human pose estimation methods.

12.2.4 A broader view on the topic

After discussing concrete ideas to address the limitations and provide different future directions w.r.t. contributions of this thesis, we now outline broader directions to advance the state of the art in articulated people detection and pose estimation.

Larger annotated datasets. Recent advances in articulated people detection and human pose estimation are mostly due to the development of strongly supervised discriminative learning techniques based on deep learning (Toshev and Szegedy, 2014; Zhang *et al.*, 2014; Girshick *et al.*, 2014; Tompson *et al.*, 2014; Chen and Yuille, 2014; Wei *et al.*, 2016). These approaches represent high capacity multi-layer classifiers with a large number of parameters that have to be estimated from the data, and thus require large strongly annotated datasets for the best performance. It has been shown that performance of deep learning body part detection methods increases with the increasing amounts of training data. Thus, obtaining large representative training sets with body part annotations is one of the most promising future directions. Limitations of current deep learning methods will become much clear when they are provided with enough training to rich the performance plateau, i.e. when adding more training samples does not result in noticeable increase of performance.

Tighter connections between different aspects of understanding humans. Current human pose estimation and articulated people detection methods typically consider these problems in isolation from other higher-level aspects of understanding humans, such as human activity recognition, human behavior analysis, understanding body language, social roles and interactions. While human pose estimation often is used as a building block in a feed forward architecture to model these aspects, improving pose estimation itself based on such higher-level cues remains largely unexplored. For instance, there is a high correlation between human activities and poses that human body assumes to perform the activity. Thus, knowing the activity label should to constrain the human pose estimation towards most likely poses for this activity. We believe that in the future more work should be focused on closing the loop between the higher-level aspects of understanding humans and human pose estimation.

More analysis. In the course of this thesis many related works have been published, which shows increasing interest in the community to the problem of articulated people detection and pose estimation. However, for the most of new methods, the emphasis is on model novelty and top performance, while typically little analysis is performed to showcase the advantages and limitations of the proposed method. Better understanding the drawbacks and failures of current methods rather than stating that a method X is better on dataset Y is a key to advancing the state of the art. To that end, a fair comparison among competing methods and thorough performance analysis is required on dedicated benchmarks. In this thesis we have made an initial step in this direction by establishing a comprehensive benchmark and a set of performance analysis tools. However, much more work is to be done to establish the culture of fair comparison and thorough performance analysis in the human pose estimation community.

More effort in data and code sharing. Articulated people detection and human pose estimation are highly competitive areas of computer vision with many research groups working in parallel to address common challenges. However, despite the competitive nature, collaboration in terms of code and data sharing among the research groups should be strongly encouraged. The progress made by a group in advancing the state of the art should be available to the others in order to facilitate even faster development of this highly dynamic research area. Being able to build on the best performing methods instead of re-implementing them will allow the researchers to focus on the not-yet-solved problems rather than trying first to rich the performance of already published approaches.

LIST OF FIGURES

1.1 Example images showing challenges of articulated people detection and pose estimation. 5
3.1 Sample detections at the equal error rate by the model trained on synthetic data generated from 6 people (top row) and on the game engine data of (Marin *et al.*, 2010) (bottom row). Even training on a subset of data obtained from only 6 different people, we are able to outperform the detector trained on much more variable game engine data. (see Sec. 3.4.1 for more details) 60
3.2 Overview of the approach to generate training data from real examples using a morphable 3D body model that drives a 2D image deformation. 62
3.3 Sample Reshape images of a person with modified height. The leftmost and the rightmost images represent extreme deviations and the middle image corresponds to the original height; the 2nd and 6th images show deviations of 2σ, while the 3rd and 5th images correspond to the deviations of 1σ from the original height. 64
3.4 Samples from the training data used in our experiments: Reshape images (top row), CVC virtual pedestrians (middle row) and multi-viewpoint dataset (bottom row). Synthetic Reshape images look similar to the real ones while being much more realistic than CVC pedestrians. Real images often contain persons wearing long or wide clothes and caring a bag, which does not occur in the synthetic data. 65
3.5 Results using Reshape data. Shown are results using 1 (a), 6 (b) and 11 (c) people to train a generic pictorial structures model. Each plot show results obtained by (Andriluka *et al.*, 2009) on real data (red), training on the unmodified training data (green), the reshape model with different variations of σ (violet, blue, black) and results using jittering (yellow) 67
  (a) 1 person 67
  (b) 6 people 67
  (c) 11 people 67
3.6 Detection rate w.r.t. the number of different persons represented in the training data. Already one person is enough to provide reasonable variability in the Reshape data. The increasing number of persons results in significant improvement which allows to achieve performance almost on par with the detector trained on the real data. The model trained on CVC data performs well, but noticeably worse than ours. . 68
3.7 Combination of generic detectors (a) and viewpoint specific detectors (b). In both cases, the combination of our detector with the one trained on the real data helps to improve detection performance. 69

  - (a) Combination of generic detectors 69
  - (b) Combination of viewpoint specific detectors 69

3.8 Performance of the sliding window detector of on different types of data. Similar to pictorial structures model, the best performance achieved when trained on human reshape together with real data. Detector trained on CVC pedestrians again performs worst. 70

4.1 Performance of our detector without (b,d) and with (a,c) complementary shape information 74
  - (a) 74
  - (b) 74
  - (c) 74
  - (d) 74

4.2 Examples of rendered images with (a) 1, 2 and 3 pixel thick edges and (b) deviations from the mean waist girth (middle). Shown are gradual shape changes in the range ±3 standard deviations from the mean shape. 77
  - (a) Varying edge thickness 77
  - (b) Varying waist 77

4.3 Effects of (a) rendered edge thickness and composition with the background, and (b) allowed deviations from the mean shape. Composition of rendered and background edges following by edge thinning is essential for reasonable performance. Maximum range ±3 sigma of realistic shape variations leads to the best detection results. 78
  - (a) Edge thickness and background composition 78
  - (b) Sigma shape 78

4.4 Sample images of (a) rendered humans with shape, pose and viewpoint variations and (b) real pedestrians from the test set. Edge orientation is encoded as hue and edge strength as saturation in the HSV color system. Notice significant variability in clothing, shape, poses and viewpoints from which real pedestrians are seen. 79
  - (a) Rendered edge images 79
  - (b) Real pedestrians 79

4.5 Results using rendered data alone and jointly with real data. Our detector trained on joint data outperforms the one trained on real data alone. 80

4.6 Combination of (a) generic and (b) multi-viewpoint detectors. Combination of our detector jointly trained on the real and synthetic data with the one trained on the real data helps to improve detection performance. 82
  - (a) Generic detectors 82
  - (b) Multi-viewpoint detectors 82

4.7 Typical failures of Andriluka's detector at EER and its comparison to (a) generic detector trained on joint rendered and real data and (b) combination of generic detectors. Andriluka's detector (every second picture) fails due to missing edge evidence while our detector trained on joint data relies on partial shape evidence taken from external edges. Edge images better seen on the screen when zoomed. Combination of our detector with *Andriluka* and *HOG/HIKSVM* helps to detect pedestrians despite unusual shapes, clutter, image blur and poor contrast. 84
  (a) Joint vs. Andriluka 84
  (b) Our combined detector vs. Andriluka 84
5.1 Sample detections (top) and pose estimates (bottom) of multiple articulated people obtained with our model trained on images from our new data generation method. 86
5.2 Overview of our novel data generation method. 88
5.3 Examples of automatically generated novel images: (a) original image and (b) animated and reshaped synthetic samples with different backgrounds. Note the realism of the generated samples. 89
5.4 Comparison of different initializations for DPM components (a). Comparison of detection results of the DPM model on (b) "Image Parsing" and (c) multiscale "Leeds Sports Poses" datasets. 90
  (a) 90
  (b) 90
  (c) 90
5.5 Examples of articulated people detections at EER by DPM trained on our joint synthetic and real "Image Parsing" (IP) data (top) and IP data alone (bottom). DPM trained on VOC2009 failed to detect people in these images. 92
5.6 Comparison of body pose estimation results between the PS trained on IP (top) and our Joint PS+DPM model trained on IP + Reshape data (bottom). 95
5.7 Detection results of different body parts on the multi-scale "Leeds Sport Poses" dataset. 97
  (a) DPM-IP-AR + PS 97
  (b) DPM-LSP-AR + PS 97
  (c) DPM-LSP-AR + Joint PS+DPM 97
6.1 Fitting error on the CAESAR dataset when using the *S-SCAPE* space (Jain *et al.*, 2010) alone, *NRD* alone, and initializing using *S-SCAPE* prior to *NRD* with different weighting schedules (*S-SCAPE* + *NRD*, *S-SCAPE* + *NRD CW*). Shown are (a) the proportion of vertices [%] with fitting error below a threshold and (b) the average fitting error per vertex. . . 107
  (a) Total fitting error 107
  (b) Mean fitting error 107

6.2 Fitting error after up to four bootstrapping rounds over the CEASAR database when using the model of Jain et al. (Jain *et al.*, 2010) as initialization for iteration 0. Shown is the proportion of vertices [%] with fitting error below a threshold. 109
    (a) Total fitting error 109
    (b) Mean fitting error on a sample scan 109
6.3 Influence of different design choices on statistical quality measures. Shown are influence of (a) bootstrapping, (b) number of training samples and (c) posture normalization on generalization (top row) and specificity (bottom row). 111
6.4 Visualization of the first five PCA eigenvectors scaled by $\pm 3\sigma$ (standard deviation). Shown are eigenvectors of the simplified SCAPE space (Jain *et al.*, 2010) (row 1) and the simplified SCAPE spaces trained using our processed data without (row 2) and with posture normalization using *WSX* (Wuhrer *et al.*, 2012) (row 3) and *NH* (Neophytou and Hilton, 2013) (row 4). 114
6.5 Fitting error on dataset of depth scans (Helten *et al.*, 2013) of simplified SCAPE spaces by Jain et al. (Jain *et al.*, 2010) and simplified SCAPE spaces trained using our processed data without posture normalization and with posture normalization using WSX and NH. Shown is the proportion of vertices [%], for which the fitting error falls below a threshold. 116
6.6 Per-vertex shape fitting error (mm) of multiple methods on sample individuals of Helten et al. (Helten *et al.*, 2013). 117
7.1 Visualization of our approach. (a) shows the top scoring poselet detections with the corresponding poselet cluster medoids (b). It is visible that the poselets capture the anatomical configuration of the human in the input image. All poselet detections contribute to a prediction of the deformable pairwise terms, the outcome of which is shown in (c). Using the PS model with these pair-wise terms achieves the detection outcome (d). In contrast we show the generic prior (Andriluka *et al.*, 2009) (e) and the corresponding pose prediction (f). 121
7.2 Sample pose estimation results on the LSP dataset obtained by our method (row 1), PS (Andriluka *et al.*, 2009) and the method of (Yang and Ramanan, 2011) (row 3). Modeling long-range part dependencies by our method results in better performance on highly articulated people. 129
7.3 Typical failure cases on the LSP dataset. Shown are the results by our method (row 1) and PS (Andriluka *et al.*, 2009) (row 2). 130
8.1 Example pose estimation results and corresponding part marginal maps obtained by (a) our full model combining local appearance and mid-level representation, (b) our best local appearance model and (c) results by Yang&Ramanan (Yang and Ramanan, 2011). 134

8.2 Overview of our method. We extend basic *PS* model (Andriluka *et al.*, 2009) (a) to more flexible structure with stronger local appearance representations including single component part detectors (b) and mixtures of part detectors (c). Then we combine local appearance model with mid-level representation based on semi-global poselets which capture configurations of multiple parts (d). Shown are the means of sample poselet clusters. Color coding shows different levels of granularity of our appearance and spatial models. 135

8.3 Qualitative results: estimated poses and corresponding part marginal maps obtained by (a) our full model combining local appearance and flexible mid-level representation, (b) our local appearance model and (c) results by Yang&Ramanan (Yang and Ramanan, 2011). 145

9.1 Randomly chosen images from each of 20 activity categories of the proposed "MPII Human Pose" dataset. Image captions indicate activity category (1st row) and activity (2nd row). To view the full dataset visit human-pose.mpi-inf.mpg.de. 150

9.2 Visualization of upper body pose variability. From left to right we show, (a) color coding of the body parts (b) annotations of the "Armlets" dataset (Gkioxari *et al.*, 2013), and (c) annotations of this dataset. 152

9.3 Example of the provided annotations. Annotated are (a) positions and visibility of the main body joints, locations of the eyes and nose and the head bounding box (occluded joints are shown in red), (b) occlusion of the main body parts (occluded parts are shown with filled rectangles), and (c) 3D viewpoints of the head and torso. On the illustration the viewpoint is shown using a simplified body model, the front face of the model is shown in red. 152

9.4 Performance (PCPm) as a function of the five complexity measures. . 156

9.5 Performance (PCPm) on images clustered by (a) full body and (b) upper body pose. Clusters are ordered by increasing mean pose complexity and representatives are shown beneath. 158

(a) full body 158

(b) upper body 158

9.6 Pose estimation results (PCPm) grouped by viewpoint. Viewpoint clusters ordered decreasingly w.r.t. number of images. Each cluster is visualized in bottom row using 3D model of the torso corresponding to the cluster medoid. 161

9.7 Pose estimation results (PCPm) grouped by activity categories shown in decreasing order w.r.t. number of images. 162

9.8 Comparison of performance (PCPm) on viewpoint (top) and activity category clusters (bottom) before and after retraining. See Fig. 9.6 for visualization of the viewpoint clusters. 164

9.9 Pose estimation results over all PCKh distance thresholds. 166

10.1 Dataset statistics and performance (mAP) as a function training set size. Shown are (a) number of training/testing examples/activity in *Separate people* subset; performance on (b) *Separate people* and (c) *Single fully visible people*. Best viewed in color. 173
  (a) # examples/activity on *Separate people* 173
  (b) Performance (mAP) on *Separate people* 173
  (c) Performance (mAP) on *Single fully visible people* 173
10.2 Performance (mAP) on a subset of 150 activities from *Separate people* as a function of the complexity measures. Best viewed in color and with additional zooming. 174
  (a) *DT* 174
  (b) *PS multi-pose (PS-M)* 174
  (c) *PS-M + DT (features)* 174
10.3 Successful and failure cases on several activity classes. Shown are the most confident prediction per class. False positives are highlighted in red. 177
11.1 Method overview: (a) initial detections (= part candidates) and pairwise terms (graph) between all detections that (b) are jointly clustered belonging to one person (one colored subgraph = one person) and each part is labeled corresponding to its part class (different colors and symbols correspond to different body parts); (c) shows the predicted pose sticks. 181
11.2 Pose estimation results over all PCK thresholds. 193
11.3 Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (rows 2, 5, 8) to the traditional two-stage approach *Dense-CNN det ROI* (rows 1, 4, 7) and to the approach of (Chen and Yuille, 2015) (rows 3, 6, 9) on WAF dataset. 196
11.4 Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (rows 1-3, 5-7) to the traditional two-stage approach *Dense-CNN det ROI* (rows 4, 8) on MPII Multi-Person dataset. See Fig. 11.1 for the explanation of color-coding. 199
11.5 Qualitative comparison (contd.) of our joint formulation *DeepCut MP Dense-CNN* (rows 1-3, 5-7) to the traditional two-stage approach *Dense-CNN det ROI* (rows 4, 8) on MPII Multi-Person dataset. See Fig. 11.1 for the explanation of color-coding. 200
11.6 Pose estimation results over all PCK thresholds on LSP (OC) dataset. . 205

LIST OF TABLES

Tab. 5.1 Results using "Image Parsing" (IP) data alone and jointly with Reshape (R) or Animate-Reshape (AR). 91
Tab. 5.2 Results for different samplings of shape parameters in Animate-Reshape data. 91
Tab. 5.3 Pose estimation results (PCP) on the "Image Parsing" (IP) dataset. 94
Tab. 5.4 Pose estimation results (PCP) on the "Image Parsing" (IP) when using our evaluation and evaluation of (Yang and Ramanan, 2011). 96
Tab. 5.5 Average precision (AP) of part estimations by different methods on multi-scale "Leeds Sport Poses" dataset. 98
Tab. 7.1 Pose estimation results (PCP) on the "Leeds Sport Poses" (LSP) dataset by our method *when using an oracle* to choose the correct component for every potential out of 12 possible values. This confirms the intuition that predicting the correct PS model directly translates to better PCP performance. 127
Tab. 7.2 Accuracy of predicting a correct component for each unary and pairwise potential and corresponding pose estimation results (PCP) on the "Leeds Sport Poses" (LSP) dataset. 128
Tab. 7.3 Pose estimation results on the "Leeds Sport Poses" (LSP) dataset with observer-centric annotations. 128
Tab. 7.4 Pose estimation results (PCP) on "Image Parse" (IP). 131
Tab. 7.5 Pose estimation results (PCP) on the "UIUC People". 131
Tab. 8.1 Results on LSP when varying number of parts in PS. 141
Tab. 8.2 Results on LSP using local appearance models. 142
Tab. 8.3 Results on LSP using mid-level representations. 142
Tab. 8.4 Performance on LSP using part appearance only. 143
Tab. 8.5 144
Tab. 8.6 Comparison of pose estimation results (PCP) on "Image Parse" dataset to current methods. 146
Tab. 9.1 154
Tab. 9.2 155
Tab. 9.3 163
Tab. 9.4 165
Tab. 10.1 176
Tab. 11.1 188
Tab. 11.2 189
Tab. 11.3 190
Tab. 11.4 192
Tab. 11.5 194
Tab. 11.6 197

Tab. 11.7 202
Tab. 11.8 203
Tab. 11.9 204
Tab. 11.10 205

BIBLIOGRAPHY

M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa (2011). Action dataset - A survey, in *SICE Annual Conference (SICE) 2011*. Cited on page 51.

B. Ainsworth, W. Haskell, S. Herrmann, N. Meckes, D. Bassett, C. Tudor-Locke, J. Greer, J. Vezina, M. Whitt-Glover, and A. Leon (2011). 2011 Compendium of Physical Activities: a second update of codes and MET values, *Medicine & Science in Sports & Exercise*, vol. 43(8), pp. 1575–1581. Cited on pages 12, 45, 149, 151, 160, 168, and 211.

B. Allen, B. Curless, and Z. Popović (2003). The space of human body shapes: reconstruction and parameterization from range scans, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 22(3), pp. 587–594. Cited on pages 25, 27, 102, and 104.

B. Allen, B. Curless, Z. Popović, and A. Hertzmann (2006). Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis, in *Symposium on Computer Animation 2006*. Cited on pages 25 and 26.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 15, 180, 188, 190, 193, 194, and 195.

M. Andriluka, S. Roth, and B. Schiele (2009). Pictorial Structures Revisited: People Detection and Articulated Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 28, 49, 59, 60, 61, 62, 67, 75, 78, 121, 122, 123, 124, 126, 127, 128, 129, 130, 131, 135, 139, 141, 143, 144, 209, 219, 222, and 223.

M. Andriluka, S. Roth, and B. Schiele (2010). Monocular 3D Pose Estimation and Tracking by Detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 62, 64, 66, 68, 69, 70, 71, 75, 76, 77, 79, 80, 81, 82, and 215.

M. Andriluka, S. Roth, and B. Schiele (2011). Discriminative Appearance Models for Pictorial Structures, *International Journal of Computer Vision (IJCV)*, vol. 99(3), pp. 259–280. Cited on pages 28, 93, 94, 95, 119, 120, 122, 123, 124, 134, 136, 137, and 146.

D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis (2005). SCAPE: Shape Completion and Animation of People, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 24(3), pp. 408–416. Cited on pages 23, 24, 26, 60, 63, 76, 87, 102, and 103.

S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski (2011). A Database and Evaluation Methodology for Optical Flow, *International Journal of Computer Vision (IJCV)*, vol. 92, pp. 1–31. Cited on page 217.

A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker (2007). Detailed human shape and pose from images., in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on page 60.

C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman (2009). PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 28(3), pp. 24:1–24:11. Cited on page 89.

S. Belongie, J. Malik, and J. Puzicha (2002). Shape Matching and Object Recognition Using Shape Contexts, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24(4), pp. 509–522. Cited on pages 61, 74, 75, and 93.

R. Benenson, M. Omran, J. Hosang, and B. Schiele (2014). Ten years of pedestrian detection, what have we learned?, in *2nd Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving (CVRSUAD) in conjunction with the European Conf. on Computer Vision (ECCV) 2014*. Cited on pages 2 and 207.

F. Bogo, J. Romero, M. Loper, and M. J. Black (2014). FAUST: Dataset and evaluation for 3D mesh registration, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 23.

L. Bourdev, S. Maji, T. Brox, and J. Malik (2010). Detecting People Using Mutually Consistent Poselet Activations, in *Proc. of the European Conf. on Computer Vision (ECCV) 2010*. Cited on pages 37, 120, 134, and 210.

L. Bourdev and J. Malik (2009). Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2009*. Cited on pages 43, 50, 51, 85, 86, 87, and 123.

W. Brendel and S. Todorovic (2011). Learning spatiotemporal graphs of human activities, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2011*. Cited on page 49.

A. Broggi, A. Fascioli, P. Grisleri, T. Graf, and M. Meinecke (2005). Model-Based Validation Approaches and Matching Techniques for Automotive Vision Based Pedestrian Detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on pages 18 and 22.

R. Brooks, R. Creiner, and T. Binford (1979). The ACRONYM Model-Based Vision System, in *Intern. Joint Conference on Artificial Intelligence 1979*. Cited on page 60.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black (2012). A naturalistic open source movie for optical flow evaluation, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on page 217.

J. Canny (1986). A computational approach to edge detection, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 8, pp. 679–698. Cited on pages 77 and 83.

J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik (2016). Human Pose Estimation with Iterative Error Feedback, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on pages 33, 37, 38, 39, 163, 165, 190, 191, and 192.

B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzalez, and F. Xavier Roca (2011). A Selective Spatio-temporal Interest Point Detector for Human Action Recognition in Complex Scenes, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2011*. Cited on page 48.

J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman (2016). Personalizing Video Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 34.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2015). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, in *International Conf. on Learning Representations (ICLR) 2015*. Cited on page 187.

X. Chen and A. Yuille (2014). Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations, in *Advances in Neural Information Processing Systems (NIPS) 2014*. Cited on pages 34, 35, 37, 39, 119, 149, 180, 190, 191, 195, 202, 203, 204, 205, 207, 214, 215, and 217.

X. Chen and A. Yuille (2015). Parsing Occluded People by Flexible Compositions, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 35, 38, 193, 194, 195, 196, and 224.

Y. Chen, Z. Liu, and Z. Zhang (2013). Tensor-Based Human Body Modeling, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 24, 26, and 102.

S. Chopra and M. Rao (1993). The partition problem, *Mathematical Programming*, vol. 59(1–3), pp. 87–115. Cited on page 182.

L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll (2011). Sparse Discriminant Analysis, *Technometrics*, vol. 53(4), pp. 406–413. Cited on pages 125 and 139.

N. Dalal and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on pages 32, 48, 59, 60, 61, 62, 70, 81, 85, 170, and 209.

N. Dalal, B. Triggs, and C. Schmid (2006). Human Detection Using Oriented Histograms of Flow and Appearance, in *Proc. of the European Conf. on Computer Vision (ECCV) 2006*. Cited on pages 48 and 170.

M. Dantone, J. Gall, C. Leistner, and L. J. V. Gool (2014). Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36(11), pp. 2131–2143. Cited on pages 30, 37, 45, and 120.

M. Dantone, J. Gall, C. Leistner, and L. V. Gool. (2013). Human Pose Estimation using Body Parts Dependent Joint Regressors, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 30, 32, 37, 41, 45, 149, 150, and 168.

F. De la Torre, J. K. Hodgins, J. Montano, and S. Valcarcel (2008). Detailed Human Data Acquisition of Kitchen Activities: the CMU-Multimodal Activity Database (CMU-MMAC), Technical report, CMU. Cited on pages 55 and 56.

C. Desai and D. Ramanan (2012). Detecting Actions, Poses, and Objects with Relational Phraselets, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on pages 31, 36, 134, and 138.

P. Dollár, C. Wojek, B. Schiele, and P. Perona (2009). Pedestrian Detection: A Benchmark, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on page 62.

K. Duan, D. Batra, and D. Crandall (2012). A Multi-layer Composite Model for Human Pose Estimation, in *Proc. of the British Machine Vision Conf. (BMVC) 2012*. Cited on pages 32, 36, 130, 131, and 146.

O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce (2009). Automatic annotation of human actions in video, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2009*. Cited on pages 48 and 168.

S. Ebert, M. Fritz, and B. Schiele (2012). RALF: A Reinforced Active Learning Formulation for Object Class Recognition, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 214.

M. Eichner and V. Ferrari (2009). Better Appearance Models For Pictorial Structures, in *Proc. of the British Machine Vision Conf. (BMVC) 2009*. Cited on pages 28, 42, and 45.

M. Eichner and V. Ferrari (2010). We Are Family: Joint Pose Estimation of Multiple Persons, in *Proc. of the European Conf. on Computer Vision (ECCV) 2010*. Cited on pages 34, 35, 38, 39, 43, 45, 46, 47, 193, and 194.

M. Eichner and V. Ferrari (2012a). Appearance Sharing for Collective Human Pose Estimation, in *Proc. of the Asian Conf. on Computer Vision (ACCV) 2012*. Cited on pages 31, 36, 37, 41, 126, 128, 135, 140, 143, 144, and 204.

M. Eichner and V. Ferrari (2012b). Human Pose Co-Estimation and Applications, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34(11), pp. 2282–2288. Cited on pages 43, 45, and 46.

A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt (2015). Efficient ConvNet-Based Marker-Less Motion Capture in General Scenes With a Low Number of Cameras, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 212.

M. Enzweiler and D. M. Gavrila (2008). A mixed generative-discriminative framework for pedestrian classification, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 20 and 21.

M. Enzweiler and D. M. Gavrila (2009). Monocular Pedestrian Detection: Survey and Experiments, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31(12), pp. 2179–2195. Cited on page 71.

M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2008). *The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html*. Cited on page 42.

M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2011a). *The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results, http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html*. Cited on page 43.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html*. Cited on page 172.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2009). *The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results, http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html*. Cited on page 43.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2010). The PASCAL Visual Object Classes (VOC) Challenge, *International Journal of Computer Vision (IJCV)*, vol. 88(2), pp. 303–338. Cited on pages 42, 45, 66, 80, 86, 90, and 97.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2011b). *The PASCAL Action Classification Taster Competition*. Cited on page 51.

X. Fan, K. Zheng, Y. Lin, and S. Wang (2015). Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 190, 191, and 203.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan (2010). Object Detection with Discriminatively Trained Part-Based Models, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32(9), pp. 1627–1645. Cited on pages 19, 59, 85, 86, 90, 120, 123, 137, 138, 210, and 215.

P. F. Felzenszwalb and D. P. Huttenlocher (2005). Pictorial Structures for Object Recognition, *International Journal of Computer Vision (IJCV)*, vol. 61(1), pp. 55–79. Cited on pages 28, 61, 75, 92, 93, 123, 136, and 156.

V. Ferrari, M. Marin, and A. Zisserman (2008). Progressive Search Space Reduction for Human Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 28, 42, 44, 45, 47, 92, 94, 95, 96, 97, 140, 153, 201, and 202.

V. Ferrari, M. Marin, and A. Zisserman (2009). Pose search: Retrieving people using their pose, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on page 28.

M. A. Fischler and R. A. Elschlager (1973). The Representation and Matching of Pictorial Structures, *IEEE Trans. Comput*, vol. 22(1), pp. 67–92. Cited on pages 27, 61, 75, 92, 93, and 136.

A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua (2007). Bridging the Gap between Detection and Tracking for 3D Monocular Video-Based Motion Capture, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on page 134.

Y. Freund and R. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences (JCSS)*, vol. 55(1), pp. 119–139. Cited on pages 61, 75, and 93.

A. Geiger, P. Lenz, and R. Urtasun (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 217.

G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes (2014). Parsing Occluded People, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 21, 22, and 194.

R. Girshick (2015). Fast R-CNN, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2015*. Cited on pages 39, 186, 187, 201, and 202.

R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014). Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 32, 39, 51, 203, and 217.

G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik (2013). Articulated Pose Estimation using Discriminative Armlet Classifiers, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 32, 37, 43, 46, 134, 137, 149, 150, 152, 153, 154, 155, 156, and 223.

G. Gkioxari, R. Girshick, and J. Malik (2015a). Actions and Attributes from Wholes and Parts. Cited on page 51.

G. Gkioxari, R. Girshick, and J. Malik (2015b). Contextual Action Recognition with R*CNN. Cited on page 51.

G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik (2014). Using k-poselets for detecting people and localizing their keypoints, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 180.

G. Gkioxari and J. Malik (2015). Finding Action Tubes, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 51.

C. Goodall (1991). Procrustes Methods in the Statistical Analysis of Shape, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53(2), pp. 285–339. Cited on page 103.

K. Grauman, G. Shakhnarovich, and T. Darrell (2003). Inferring 3D Structure with a Statistical Image-Based Shape Model, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2003*. Cited on pages 18 and 22.

P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. Black (2012). DRAPE: DRessing Any PErson, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 31(4), pp. 35:1–10. Cited on pages 24, 26, 102, and 213.

N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel (2009). A Statistical Model of Human Pose and Body Shape, in *Comput. Graph. Forum (Proc. of Eurographics) 2009*. Cited on pages 6, 23, 25, 26, 27, 63, 102, 103, and 117.

T. Helten, A. Baak, G. Bharai, M. Müller, H.-P. Seidel, and C. Theobalt (2013). Personalization and Evaluation of a Real-time Depth-based Full Body Scanner, in *3D Vision 2013*. Cited on pages 25, 102, 113, 116, 117, and 222.

P. Hu and D. Ramanan (2016). Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on pages 34, 165, and 192.

E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model, *ArXiv*. Cited on pages 8, 165, and 212.

A. Jacobson, I. Baran, J. Popović, and O. Sorkine (2011). Bounded Biharmonic Weights for Real-Time Deformation, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 30(4), pp. 78:1–78:8. Cited on pages 87 and 89.

A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt (2010). MovieReshape: Tracking and Reshaping of Humans in Videos, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH Asia)*, vol. 29(5), pp. 148:1–148:10. Cited on pages 2, 3, 6, 13, 21, 24, 25, 26, 60, 63, 64, 74, 76, 86, 87, 102, 103, 104, 107, 108, 109, 112, 113, 114, 116, 117, 207, 210, 221, and 222.

A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler (2014). Learning Human Pose Estimation Features with Convolutional Networks, in *International Conf. on Learning Representations (ICLR) 2014*. Cited on pages 33, 37, 38, and 39.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black (2013). Towards understanding action recognition, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2013*. Cited on pages 48, 49, 50, 53, 55, 56, 167, 168, 170, 171, 172, and 173.

S. Johnson and M. Everingham (2010). Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, in *Proc. of the British Machine Vision Conf. (BMVC) 2010*. Cited on pages 30, 36, 41, 45, 46, 90, 96, 97, 120, 125, 126, 140, 150, 188, 190, and 214.

S. Johnson and M. Everingham (2011). Learning Effective Human Pose Estimation from Inaccurate Annotation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 41, 45, 46, 86, 94, 95, 119, 120, 129, 130, 131, 134, 135, 146, 150, 153, 180, and 188.

L. Karlinsky and S. Ullman (2012). Using Linking Features in Learning Non-parametric Part Models, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on pages 29, 36, and 37.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei (2014). Large-scale Video Classification with Convolutional Neural Networks, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 50, 51, 54, 55, and 56.

M. Kiefel and P. Gehler (2014). Human Pose Estimation with Fields of Parts, in *Proc. of the European Conf. on Computer Vision (ECCV) 2014*. Cited on pages 31, 204, and 205.

A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems (NIPS) 2012*. Cited on pages 32, 85, 188, and 201.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre (2011). HMDB: a large video database for human motion recognition, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2011*. Cited on pages 50, 53, 55, 168, and 173.

L. Ladicky, P. H. S. Torr, and A. Zisserman (2013). Human Pose Estimation using a Joint Pixel-wise and Part-wise Formulation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 34, 38, 39, and 215.

I. Laptev (2005). On Space-Time Interest Points, *International Journal of Computer Vision (IJCV)*, vol. 64(2-3), pp. 107–123. Cited on page 48.

I. Laptev (2009). Improving object detection with boosted histograms, *Image Vision Comput.*, vol. 27(5), pp. 535–544. Cited on pages 20 and 67.

I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld (2008). Learning Realistic Human Actions from Movies, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 48, 52, 55, 56, 168, and 170.

I. Laptev and P. Perez (2007). Retrieving actions in movies, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2007*. Cited on pages 52, 55, and 56.

L. Liang and K. Grauman (2014). Beyond Comparing Image Pairs: Setwise Active Learning for Relative Attributes, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 214.

J. Liebelt and C. Schmid (2010). Multi-view object class detection with a 3D geometric model, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 19 and 22.

J. Liebelt, C. Schmid, and K. Schertler (2008). Viewpoint-independent object class detection using 3D Feature Maps, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 19, 22, 60, and 74.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick (2014). Microsoft COCO: Common Objects in Context, in *Proc. of the European Conf. on Computer Vision (ECCV) 2014*. Cited on page 86.

D. Lowe (1987). Three-Dimensional Object Recognition from Single Two-Dimensional Images, *Artificial Intelligence*, vol. 31, pp. 355–395. Cited on page 60.

O. Mac Aodha, N. D. Campbell, J. Kautz, and G. J. Brostow (2014). Hierarchical Subquery Evaluation for Active Learning on a Graph, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 214.

S. Maji (2011). Large Scale Image Annotations on Amazon Mechanical Turk, Technical report, EECS UC Berkeley. Cited on page 153.

S. Maji, A. Berg, and J. Malik (2008). Classification using intersection kernel SVMs is efficient, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 62, 70, and 82.

J. Marin, D. Vazquez, D. Geronimo, and A. Lopez (2010). Learning appearance in virtual scenarios for pedestrian detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 20, 22, 60, 62, 64, 66, 71, 74, 209, and 219.

M. Marin-Jimenez, A. Zisserman, and V. Ferrari (2011). "Here's looking at you, kid." Detecting people looking at each other in videos, in *Proc. of the British Machine Vision Conf. (BMVC) 2011*. Cited on page 138.

D. Marr and H. Nishihara (1978). Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. of the Royal Society of London B: Biological Sciences*, vol. 200(1140), pp. 269–194. Cited on page 60.

M. Marszałek, I. Laptev, and C. Schmid (2009). Actions in Context, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on page 52.

R. Messing, C. Pal, and H. Kautz (2009). Activity recognition using the velocity histories of tracked keypoints, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2009*. Cited on pages 54 and 55.

K. Mikolajczyk and C. Schmid (2005). A Performance Evaluation of Local Descriptors, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27(10), pp. 1615–1630. Cited on pages 32 and 85.

A. Mittal, M. Blaschko, A. Zisserman, and P. Torr (2012). Taxonomic Multi-class Prediction and Person Layout using Efficient Structured Ranking, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on pages 32, 37, and 134.

L. Mündermann, S. Corazza, and T. P. Andriacchi (2007). Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models., in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on page 102.

A. Neophytou and A. Hilton (2013). Shape and Pose Space Deformation for Subject Specific Animation, in *3D Vision 2013*. Cited on pages 24, 25, 26, 27, 102, 107, 112, 114, 115, and 222.

T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and C. Theobalt (2013). Sparse localized deformation components, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH Asia)*, vol. 32(6), pp. 179:1–179:10. Cited on page 25.

S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai (2011). A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 54 and 55.

R. Okada and S. Soatto (2008). Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images, in *Proc. of the European Conf. on Computer Vision (ECCV) 2008*. Cited on pages 20 and 22.

W. Ouyang, X. Chu, and X. Wang (2014). Multi-source Deep Learning for Human Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 33, 204, and 205.

A. Patron, M. Marszalek, A. Zisserman, and I. Reid (2010). High Five: Recognising human interactions in TV shows, in *Proc. of the British Machine Vision Conf. (BMVC) 2010*. Cited on pages 52, 55, and 56.

B. Pepik, P. Gehler, M. Stark, and B. Schiele (2012a). $3D^2$PM - 3D Deformable Part Models, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on page 19.

B. Pepik, M. Stark, P. Gehler, and B. Schiele (2012b). Teaching 3D Geometry to Deformable Part Models, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 19.

B. Pepik, M. Stark, P. Gehler, and B. Schiele (2013). Occlusion Patterns for Object Class Detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 215.

T. Pfister, J. Charles, and A. Zisserman (2015). Flowing ConvNets for Human Pose Estimation in Videos, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2015*. Cited on pages 34 and 216.

H. Pirsiavash and D. Ramanan (2014). Parsing Videos of Actions with Segmental Grammars, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 49.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele (2013a). Poselet Conditioned Pictorial Structures, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 14.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele (2013b). Strong Appearance and Expressive Spatial Models for Human Pose Estimation, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2013*. Cited on pages 15 and 186.

L. Pishchulin, M. Andriluka, and B. Schiele (2014). Fine-grained Activity Recognition with Holistic and Pose based Features, in *Proc. of German Conference on Pattern Recognition (GCPR) 2014*. Cited on page 15.

L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele (2016). DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on page 15.

L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele (2012). Articulated People Detection and Pose Estimation: Reshaping the Future, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 14.

L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele (2011a). Learning People Detection Models from Few Training Samples, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 13.

L. Pishchulin, A. Jain, C. Wojek, T. Thormaehlen, and B. Schiele (2011b). In Good Shape: Robust People Detection based on Appearance and Shape, in *Proc. of the British Machine Vision Conf. (BMVC) 2011*. Cited on page 13.

L. Pishchulin, S. Wuhrer, H. Thomas, C. Theobalt, and B. Schiele (2015). Building Statistical Shape Spaces for 3D Human Modeling, *ArXiv 1503.05860*. Cited on page 14.

V. Ramakrishna, T. Kanade, and Y. A. Sheikh (2012). Reconstructing 3D Human Pose from 2D Image Landmarks, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on page 213.

V. Ramakrishna, D. Munoz, M. Hebert, A. J. Bagnell, and Y. Sheikh (2014). Pose Machines: Articulated Pose Estimation via Inference Machines, in *Proc. of the European Conf. on Computer Vision (ECCV) 2014*. Cited on pages 32, 34, 204, and 205.

D. Ramanan (2006). Learning to Parse Images of Articulated Objects, in *Advances in Neural Information Processing Systems (NIPS) 2006*. Cited on pages 28, 40, 45, 85, 90, 96, 126, 140, and 144.

D. Ramanan, D. A. Forsyth, and A. Zisserman (2005). Strike a Pose: Tracking People by Finding Stylized Poses, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on page 134.

M. Raptis and L. Sigal (2013). Poselet Key-framing: A Model for Human Activity Recognition, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 50.

K. Robinette, H. Daanen, and E. Paquet (1999). The CAESAR Project: A 3-D Surface Anthropometry Survey, in *Proc. of Conf. on 3D Digital Imaging and Modeling 1999*. Cited on pages 10, 14, 23, 26, 102, 103, 107, 117, 208, and 210.

M. D. Rodriguez, J. Ahmed, and M. Shah (2008). Action MACH: a spatio-temporal maximum average correlation height filter for action recognition, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on page 49.

M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele (2012). A Database for Fine Grained Activity Detection of Cooking Activities, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 29, 49, 50, 55, 56, and 170.

M. Rohrbach, M. Stark, and B. Schiele (2011). Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 172.

C. Rother, V. Kolmogorov, and A. Blake (2004). "GrabCut": interactive foreground extraction using iterated graph cuts, *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 23(3), pp. 309–314. Cited on page 87.

S. Sadanand and C. J. J. (2012). Action Bank: A High-Level Representation of Activity in Video, in *Proc. of the European Conf. on Computer Vision (ECCV) 2012*. Cited on page 168.

B. Sapp, C. Jordan, and B. Taskar (2010a). Adaptive Pose Priors for Pictorial Structures, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 29, 36, and 96.

B. Sapp and B. Taskar (2013). Multimodal Decomposable Models for Human Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 32, 36, 43, 44, 45, 46, 47, 119, 150, 153, 154, 155, 156, 168, 180, 188, and 201.

B. Sapp, A. Toshev, and B. Taskar (2010b). Cascaded Models for Articulated Pose Estimation, in *Proc. of the European Conf. on Computer Vision (ECCV) 2010*. Cited on pages 29 and 36.

B. Sapp, D. Weiss, and B. Taskar (2011). Parsing human motion with stretchable models, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 29, 32, 35, 37, 42, 44, 45, 46, 47, 135, 139, 211, and 216.

C. Schuldt, I. Laptev, and B. Caputo (2004). Recognizing human actions: a local SVM approach, in *Proc. of the International Conf. on Pattern Recognition (ICPR) 2004*. Cited on pages 52, 55, and 56.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun (2014). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, in *International Conf. on Learning Representations (ICLR) 2014*. Cited on page 32.

G. Shakhnarovich, P. Viola, and T. Darrell (2003). Fast pose estimation with parameter sensitive hashing, in *Proc. of the European Conf. on Computer Vision (ECCV) 2003*. Cited on pages 20 and 22.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake (2011). Real-Time Human Pose Recognition in Parts from a Single Depth Image, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 2, 18, 22, and 32.

K. Simonyan and A. Zisserman (2014a). Two-Stream Convolutional Networks for Action Recognition in Videos, in *Advances in Neural Information Processing Systems (NIPS) 2014*. Cited on page 50.

K. Simonyan and A. Zisserman (2014b). Very Deep Convolutional Networks for Large-Scale Image Recognition, in *International Conf. on Learning Representations (ICLR) 2014*. Cited on pages 187 and 202.

V. K. Singh and R. Nevatia (2011). Action recognition in cluttered dynamic scenes using Pose-Specific Part Models, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2011*. Cited on page 49.

K. Soomro, A. R. Zamir, and M. Shah (2012). UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, Technical report, UCF. Cited on page 168.

M. Stark, M. Goesele, and B. Schiele (2010). Back to the Future: Learning Shape Models from 3D CAD Data, in *Proc. of the British Machine Vision Conf. (BMVC) 2010*. Cited on pages 19, 22, 60, 74, and 76.

M. Styner, K. Rajamani, L.-P. Nolte, G. Zsemlye, G. Székely, C. Taylor, and R. Davies (2003). Evaluation of 3D Correspondence Methods for Model Building, in *Information Processing in Medical Imaging 2003*. Cited on pages 110 and 111.

M. Sun and S. Savarese (2011). Articulated Part-based Model for Joint Object Detection and Pose Estimation, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2011*. Cited on pages 32, 36, 38, 39, 47, 120, 180, and 193.

M. Sun, M. Telaprolu, H. Lee, and S. Savarese (2012). An Efficient Branch-and-Bound Algorithm for Optimal Human Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 30 and 35.

G. Tam, Z.-Q. Cheng, Y.-K. Lai, F. Langbein, Y. Liu, D. Marshall, R. Martin, X.-F. Sun, and P. Rosin (2013). Registration of 3D Point Clouds and Meshes: A Survey From Rigid to Non-Rigid, *Trans. on Visualization and Computer Graphics*, vol. 19(7), pp. 1199–1217. Cited on page 25.

S. Tang, B. Andres, M. Andriluka, and B. Schiele (2015). Subgraph Decomposition for Multi-Target Tracking, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 216.

S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele (2013). Learning People Detectors for Tracking in Crowded Scenes, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2013*. Cited on pages 21, 22, and 215.

S. Tang, M. Andriluka, and B. Schiele (2012). Detection and Tracking of Occluded People, in *Proc. of the British Machine Vision Conf. (BMVC) 2012*. Cited on pages 21, 22, and 215.

T.-P. Tian and S. Sclaroff (2010). Fast Globally Optimal 2D Human Detection with Loopy Graph Models, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 30 and 35.

R. Tokola, W. Choi, and S. Savarese (2013). Breaking the Chain: Liberation from the Temporal Markov Assumption for Tracking Human Poses, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2013*. Cited on page 216.

J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler (2015). Efficient Object Localization Using Convolutional Networks, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 33, 38, 39, 163, 165, 180, 187, 191, and 192.

J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler (2014). Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation, in *Advances in*

Neural Information Processing Systems (NIPS) 2014. Cited on pages 2, 33, 34, 37, 38, 39, 45, 59, 85, 119, 149, 163, 165, 168, 180, 187, 188, 190, 191, 192, 202, 203, 207, 214, and 217.

A. Toshev and C. Szegedy (2014). DeepPose: Human Pose Estimation via Deep Neural Networks, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 33, 37, 38, 188, 207, and 217.

D. Tran and D. A. Forsyth (2010). Improved Human Parsing with a Full Relational Model, in *Proc. of the European Conf. on Computer Vision (ECCV) 2010*. Cited on pages 30, 35, 41, 42, 45, 120, and 126.

Z. Tu, X. Chen, A. L. Yuille, and S. chun Zhu (2005). Image parsing: Unifying segmentation, detection, and recognition, *International Journal of Computer Vision (IJCV)*, vol. 63(2), pp. 113–140. Cited on page 32.

J. Uijlings, K. van de Sande, T. Gevers, and A.W.M. (2013). Selective Search for Object Recognition, *International Journal of Computer Vision (IJCV)*, vol. 104(2), pp. 154–171. Cited on pages 39 and 186.

O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or (2011). A survey on shape correspondence, *Computer Graphics Forum*, vol. 3(6), pp. 1681–1707. Cited on page 25.

A. Vedaldi and A. Zisserman (2010). Efficient additive kernels via explicit feature maps, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 172.

C. Vondrick, D. Patterson, and D. Ramanan (2012). Efficiently Scaling Up Crowdsourced Video Annotation, *International Journal of Computer Vision (IJCV)*, vol. 101(1), pp. 184–204. Cited on page 153.

C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao (2014). Robust Estimation of 3D Human Poses from a Single Image, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 213 and 215.

F. Wang and Y. Li (2013a). Beyond Physical Connections: Tree Models in Human Pose Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 31.

F. Wang and Y. Li (2013b). Beyond Physical Connections: Tree Models in Human Pose Estimation., in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 203.

H. Wang, A. Kläser, C. Schmid, and C.-L. Liu (2013). Dense trajectories and motion boundary descriptors for action recognition, *International Journal of Computer Vision (IJCV)*, vol. 103(1), pp. 60–79. Cited on pages 15, 48, 56, 167, 168, and 170.

H. Wang and C. Schmid (2013). Action Recognition with Improved Trajectories, in *Proc. IEEE International Conf. on Computer Vision (ICCV) 2013*. Cited on pages 48, 56, 167, 168, 169, and 170.

H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid (2009). Evaluation of local spatio-temporal features for action recognition, in *Proc. of the British Machine Vision Conf. (BMVC) 2009*. Cited on page 48.

Y. Wang, D. Tran, and Z. Liao (2011). Learning Hierarchical Poselets for Human Parsing, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 31, 36, 37, 41, 45, 96, 120, 123, 124, and 131.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh (2016). Convolutional Pose Machines, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on pages 34, 37, 39, 165, 166, 168, 190, 191, 192, 207, 214, and 217.

D. J. Weiss and B. Taskar (2013). Learning Adaptive Value of Information for Structured Prediction, in *Advances in Neural Information Processing Systems (NIPS) 2013*. Cited on page 216.

G. Willems, T. Tuytelaars, and L. Van Gool (2008). An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector, in *Proc. of the European Conf. on Computer Vision (ECCV) 2008*. Cited on page 48.

S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang (2014). Estimation of Human Body Shape and Posture Under Clothing, *Computer Vision and Image Understanding (CVIU)*, vol. 127, pp. 31–42. Cited on pages 6 and 210.

S. Wuhrer, C. Shu, and P. Xi (2011). Landmark-Free Posture Invariant Human Shape Correspondence, *The Visual Computer*, vol. 27(9), pp. 843–852. Cited on page 26.

S. Wuhrer, C. Shu, and P. Xi (2012). Posture-Invariant Statistical Shape Analysis Using Laplace Operator, *Computers & Graphics*, vol. 36(5), pp. 410–416. Cited on pages 25, 27, 102, 106, 112, 114, and 222.

Y. Yang, S. Baker, A. Kannan, and D. Ramanan (2012). Recognizing proxemics in personal photos, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 35, 38, 43, 45, and 46.

Y. Yang and D. Ramanan (2011). Articulated pose estimation with flexible mixtures-of-parts, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 29, 30, 36, 37, 85, 94, 95, 96, 119, 122, 123, 127, 128, 129, 130, 131, 134, 135, 137, 139, 140, 143, 144, 145, 146, 211, 222, 223, and 225.

Y. Yang and D. Ramanan (2013). Articulated Human Detection with Flexible Mixtures of Parts, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35(12), pp. 2878–2890. Cited on pages 30, 31, 33, 34, 35, 38, 44, 47, 50, 146, 153, 154, 155, 156, 163, 168, 171, 175, 193, and 194.

B. Yao and L. Fei-Fei (2010). Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 30 and 35.

J. Yuan, Z. Liu, and Y. Wu (2009). Discriminative subvolume search for efficient action detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 52, 55, and 56.

N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev (2014). PANDA: Pose Aligned Networks for Deep Attribute Modeling, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on page 217.

S. Zhang, R. Benenson, and B. Schiele (2015). Filtered Feature Channels for Pedestrian Detection, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 207.

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal (1997). Algorithm 778: L-BFGS-B Fortran subroutines for large-scale bound-constrained optimization., *ACM Trans. Math. Softw.*, vol. 23(4), pp. 550–560. Cited on page 105.

A. Zinnen, U. Blanke, and B. Schiele (2009). An Analysis of Sensor-Oriented vs. Model-Based Activity Recognition, in *IEEE International Symposium on Wearable Computers (ISWC) 2009*. Cited on page 50.

S. Zuffi, O. Freifeld, and M. J. Black (2012). From pictorial structures to deformable structures, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 53.

CURRICULUM VITAE

Leonid Pishchulin

Born: 03/10/1986 in Ryazan, Russia
Citizenship: Russian

Education:	2010–2015	**Max Planck Institute for Informatics, Germany**
		Ph.D. student at the Computer Vision and Multimodal Computing Group of Prof. Bernt Schiele
	2007–2010	**RWTH Aachen University, Germany**
		M.Sc. student, Computer Science (very good)
	2003–2007	**National University of Science and Technology "MISIS", Russia**
		Diploma student, Computer Science (with distinction)
Experience:	2015–now	**Research Scientist**, Amazon.com, USA
	2010–2010	**Research Assistant** with Prof. Hermann Ney, computer vision, RWTH Aachen University, Germany
	2006-2007	**Software Developer** at FarmCom.ru
Reviewer:		IEEE Trans. Pattern Anal. Mach. Intell. (2015)
		International Journal of Computer Vision (2012, 2015)
		Computer Vision and Image Understanding (2013–2016)
		ECCV 2014, CVPR 2015 (both Outstanding Reviewer Award); ICCV 2015; CVPR 2016, ECCV 2016
Awards:	2010–2012	International Max Planck Research School Scholarship
	2007–2009	Doctor Carl-Arthur Pastor-Foundation Scholarship

PUBLICATIONS

[15] *"DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model"*
Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka,
Bernt Schiele
ArXiv, 2016

[14] *"DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation"*
Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres,
Mykhaylo Andriluka, Peter Gehler, Bernt Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2016

[13] *"Building Statistical Shape Spaces for 3D Human Modeling"*
Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt,
Bernt Schiele
ArXiv 1503.05860, 2015

[12] *"Efficient ConvNet-Based Marker-Less Motion Capture in General Scenes With a Low Number of Cameras"*
Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin,
Mykhaylo Andriluka, Chris Bregler, Bernt Schiele, Christian Theobalt
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2015

[11] *"Fine-grained Activity Recognition with Holistic and Pose based Features"*
Leonid Pishchulin, Mykhaylo Andriluka, Bernt Schiele
In German Conference on Pattern Recognition (**GCPR**), 2014

[10] *"Estimation of Human Body Shape and Posture Under Clothing"*
Stefanie Wuhrer, Leonid Pishchulin, Alan Brunton, Chang Shu, Jochen Lang
Computer Vision and Image Understanding (**CVIU**), 2014

[9] *"Human Pose Estimation: New Benchmark and State of the Art Analysis"*
Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2014

[8] *"Strong Appearance and Expressive Spatial Models for Human Pose Estimation"*
Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele
In IEEE International Conference on Computer Vision (**ICCV**), 2013

[7] *"Poselet Conditioned Pictorial Structures"*
Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele

In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2013

[6] *"Articulated People Detection and Pose Estimation: Reshaping the Future"*
Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, Bernt Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2012

[5] *"Image warping for face recognition: From local optimality towards global optimization"*
Leonid Pishchulin, Tobias Gass, Philippe Dreuw, Hermann Ney
Pattern Recognition (**PR**), 45:3131–3140, 2012

[4] *"In Good Shape: Robust People Detection based on Appearance and Shape"*
Leonid Pishchulin, Arjun Jain, Christian Wojek, Thorsten Thormählen, Bernt Schiele
In British Machine Vision Conference (**BMVC**), 2011

[3] *"Learning People Detection Models from Few Training Samples"*
Leonid Pishchulin, Arjun Jain, Christian Wojek, Mykhaylo Andriluka, Thorsten Thormählen, Bernt Schiele
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2011

[2] *"The Fast and the Flexible: Extended Pseudo Two-Dimensional Warping for Face Recognition"*
Leonid Pishchulin, Tobias Gass, Philippe Dreuw, Hermann Ney
In Iberian Conference on Pattern Recognition and Image Analysis (**IbPRIA**), 2011

[1] *"Warp that Smile on your Face: Optimal and Smooth Deformations for Face Recognition"*
Tobias Gass, Leonid Pishchulin, Philippe Dreuw, Hermann Ney
IEEE International Conference on Automatic Face and Gesture Recognition (**FG**), 2011