Technische Universität München

Fakultät für Informatik

Multi-view Part-based Models for 3D Human Pose Estimation in Real-World Scenes

Sikandar Amin

Vollständiger Abdruck der von der Fakultät für Informatik
der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende/-r:		Prof. Dr. Rüdiger Westermann
Prüfende/-r der Dissertation:	1.	Prof. Dr. Bernd Radig
	2.	Prof. Dr.-Ing. Darius Burschka

Die Dissertation wurde am 27.03.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 14.07.2018 angenommen.

Abstract

We address the task of articulated 3D human pose estimation in real-world scenes with multiple humans and dynamic backgrounds. A number of factors make this task challenging including, appearance-shift between train and test sets, high dimensional state-space of 3D body configurations, person-person occlusions, across-view ambiguities, and inter-mixing of body parts of multiple humans. In addition to these challenges, we are working in a domain with limited amount of available training data. Recent approaches in this domain rely on activity specific motion models and hence require complex inference procedures for body pose estimation. In practice, such approaches end up relying on stochastic search methods to work. On top of all that, these approaches are mainly evaluated in controlled laboratory settings with static backgrounds, and therefore their ability to cope with real-world settings is unknown.

In this work, we take inspiration from the recent success of pictorial structures based approaches for 2D human pose estimation. First, we propose several improvements to the basic 2D pictorial structure model, and later extend our 2D model for human pose estimation jointly in multiple views. We name it *Multi-view Pictorial Structures* approach. We show that our approach achieves on par or better results on standard 3D pose estimation dataset, *i.e*. HumanEva-I. Notably, our approach significantly outperforms the state-of-the-art for activities with more complex motions. Moreover, we also introduce two challenging multi-view datasets for single and multiple human 3D pose estimation in the ”wild”. Our evaluations demonstrate that the proposed multi-view pictorial structures approach is also applicable to such challenging settings without bells and whistles.

Finally, we propose a model adaptation approach which capitalizes on the 3D pose reconstructions in the previous images of the testing scenario, and thus progressively adapts the model to the scene at hand. Our approach for model adaptation can operate in both *offline* and *online* modes, without any modifications. We demonstrate the advantages of both modes of operation in dynamic multi-view settings.

Zusammenfassung

Wir beschäftigten uns mit der Aufgabe der artikulierten Schätzung der menschlichen Körperhaltung in 3D in Echtweltsituationen mit mehreren Menschen und dynamischen Hintergründen. Eine Vielzahl von Faktoren macht diese Aufgabe zu einer Herausforderung, darunter die Blickverschiebung zwischen Train- und Test-Sets, den hohen dimensionalen Zustandsraum von 3D-Körperkonfigurationen, Okklusionen zwischen Menschen, blickübergreifende Zweideutigkeiten sowie die Vermischung der Körperteile mehrerer Menschen. Zusätzlich zu diesen Herausforderungen arbeiten wir in einem Bereich mit einer begrenzten Menge an verfügbaren Trainingsdaten. Neuere Annährungen in diesem Bereich vertrauen auf aktivitätsspezifische Bewegungsmodelle und benötigen daher komplexe Inferenzverfahren für diese Schätzung der Körperhaltung. In der Praxis vertrauen solche Annäherungen schließlich auf stochastische Suchmethoden, um zu funktionieren. Hinzu kommt, dass diese Annäherungen hauptsächlich in kontrollierten Laborumgebungen mit statischen Hintergründen ausgewertet werden, weswegen ihre Fähigkeit, mit Echtweltsituationen zurecht zu kommen, unbekannt ist.

In dieser Dissertation nehmen wir unsere Inspiration von den neueren Erfolgen der Annäherungen für die Schätzung menschlicher Körperhaltung in 2D auf der Grundlage von Bildstrukturen (pictorial structures). Zunächst schlagen wir mehrere Verbesserungen des grundlegenden 2D-Modells der Bildstrukturen vor und erweitern unser 2D-Modell später für die Schätzung der Körperhaltung in mehreren Blickwinkeln zusammen. Wir nennen es die "Multi-View Pictorial Structures Approach". Wir zeigen, dass unsere Annäherungen gleichwertige oder bessere Ergebnisse bei einem Standarddatensatz für 3D-Körperhaltung, d.h. HumanEva-I, erzielt. Es ist bemerkenswert, dass unsere Annäherung den Stand der Technik für Aktivitäten mit komplexeren Bewegungen deutlich übertrifft. Zusätzlich führen wir noch zwei herausfordernde Datensätze aus mehreren Blickwinkeln für eine 3D-Schätzung der Körperhaltung einzelner und mehrerer Menschen "in the wild" ein. Unsere Auswertungen zeigen, dass die vorgeschlagene Annäherung mit pictorial structures aus mehreren Blickwinkeln auch für solche herausfordernde Situationen ohne irgendwelche extras anwendbar ist.

Schließlich schlagen wir eine Annäherung der Modellanpassung vor, das von den 3D Körperhaltungsnachbildungen in den vorangegangenen Bildern des Testszenarios profitiert und daher das Modell zunehmend an das vorliegende Modell anpasst. Unsere Annäherung für die Modellanpassung kann sowohl im Online- als auch im Offline-Modus ohne Modifizierungen funktionieren. Wir zeigen die Vorteile beider Verfahrensweisen in Einstellungen mit mehreren Blickwinkeln.

Acknowledgements

First of all, I am especially grateful to my advisor Prof. Bernd Radig for his continuous support and encouragement for my Ph.D research, for his patience, motivation, and immense knowledge that helped me grow as a research scientist.

Next, I would like to express my sincere gratitude to Dr. Mykhaylo Andriluka for his hands-on support throughout my Ph.D work, and enabling collaboration with Max Planck Institute. Moreover, his guidance helped me in writing of this thesis.

My sincere thanks also goes to Prof. Bernt Schiele, who provided me a wonderful opportunity to join his research group at Max Planck Institute for Informatics as a visiting researcher, and supported me for attending the computer vision conferences. Without this precious support it would not be possible to conduct this research.

Furthermore, I would like to thank my colleague Dr. Marcus Rohrbach for the research collaboration on a number of exciting topics.

I would also like to mention my friends Abbas and Zeeshan for their moral support throughout my Ph.D research.

Last but not the least, I would like to thank my parents, my brothers Khizer & Haris, and my good friend Anne Steffen for supporting me spiritually, keeping my morale high throughout the writing of this thesis, and their wonderful impact on my life in general.

Contents

Dedicated to my parents,
Muhammad Amin and Rukhsana Amin

Chapter 1

Introduction

Object pose estimation is one of the most active and challenging domains in computer vision. It refers to the study of algorithms to recover the object pose. Pose of an object means the parameters which define the object *configuration* or *shape layout*. Estimating this object information automatically from images or videos would help an autonomous system to intelligently interact with the object and reason about the scene more convincingly. Object grasping and manipulation are typical examples of such interactions with the objects. Recently, pose estimation of rigid objects has seen significant improvement for both monocular [47, 51, 117, 121] and multi-view settings [122]. Naturally, the advancement in local feature representations and statistical learning techniques have significantly contributed to the progress on this task. However, the excellent results in this particular domain can be strongly attributed to the object rigidity which allows to construct detailed 3D models and therefore the task of 3D pose estimation for such objects reduces to matching of 2D images to the projections of 3D models, over a small set of parameters *i.e. scale, translation, and rotation*. For many such object categories the detailed 3D models of the objects are available online *e.g*. Google 3D warehouse, KIT object models database [5], NYC3DCars [75], and PASCAL3D+ [155], etc.

On the other hand, the 3D pose estimation performance of deformable or articulated objects has not yet achieved the acceptable accuracy. Human body is both articulated and deformable, which makes it much more challenging to formulate its body layout in vision research. The human pose estimation approaches which are based on intertial sensors [99, 111] or retro-reflective markers [101, 148] are quite precise and reliable. But they restrict the pose estimation and motion capture to controlled laboratory settings, and limit the human's ability to display natural body postures. On the other hand, computer vision based solutions for markerless pose estimation enable human motion analysis in natural settings, making it currently one of the most sought-after domains for many real-world applications. *Surveillance*, *bio-mechanical analysis*, *man-machine communication*, *automatic annotation* and *automatic driving assistance systems*

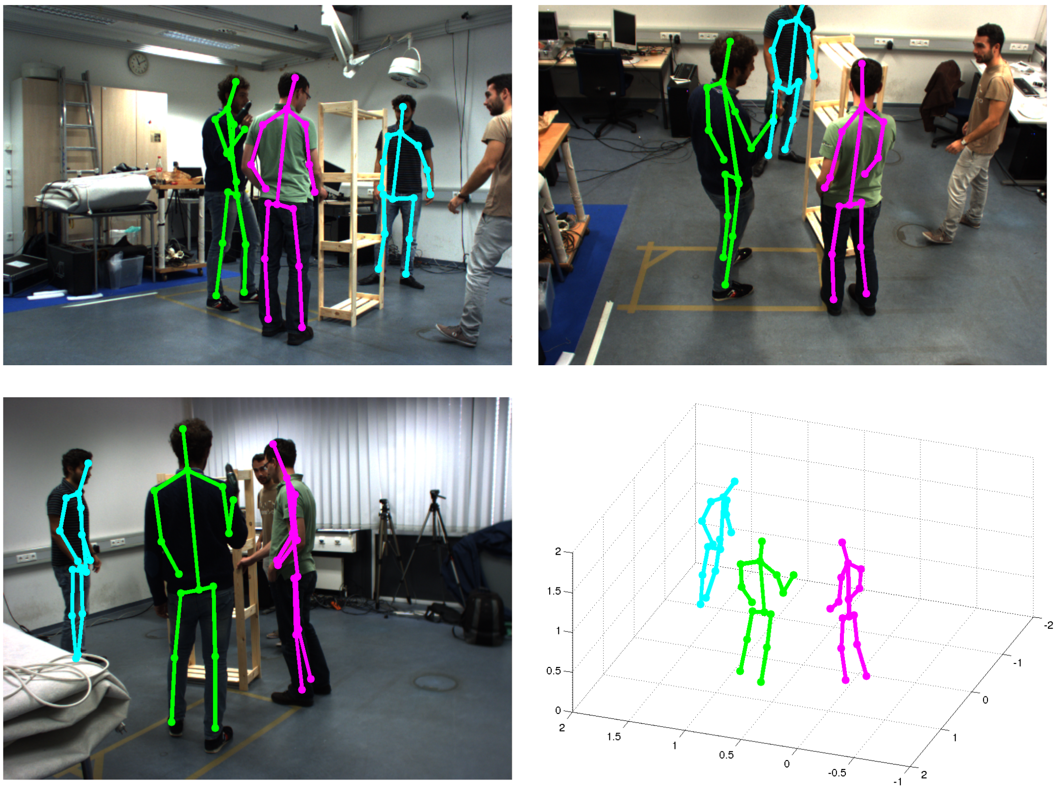


Figure 1.1: Output of our proposed 3D human pose estimation approach and their projections in 2D.

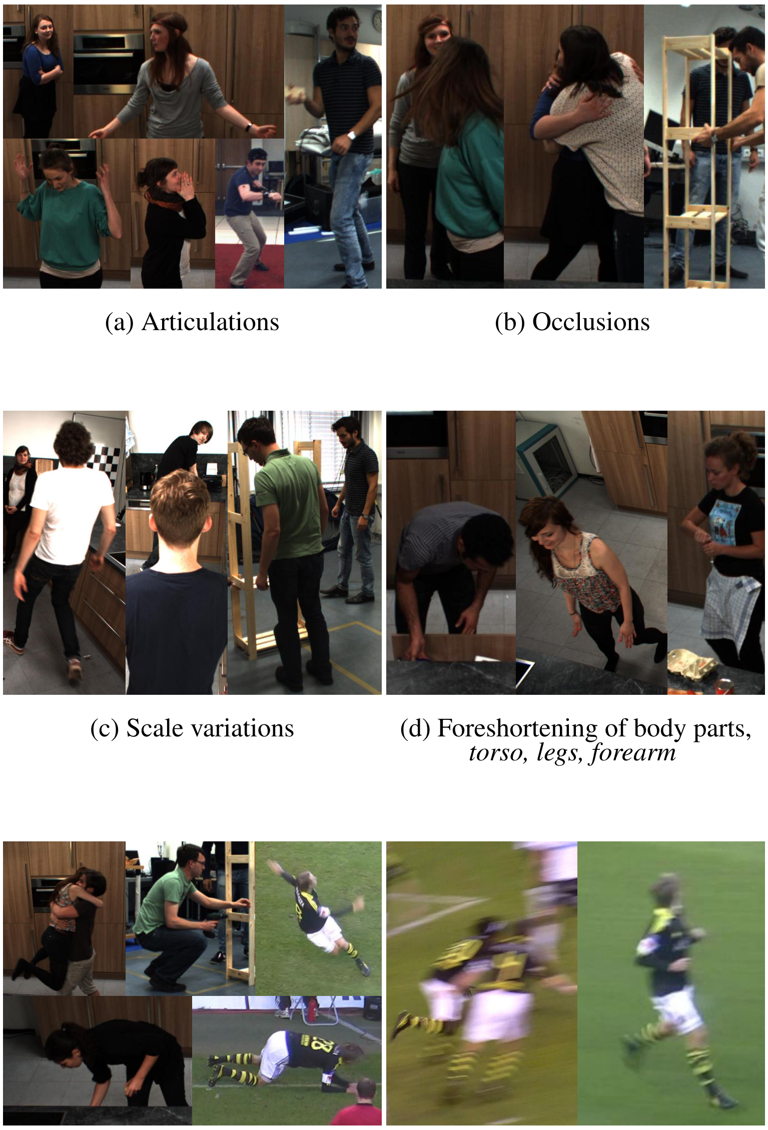
among several other applications can greatly benefit from such solutions. Moreover, vision-based solutions for pose estimation are relatively much more affordable compared to special body suits that are being used in typical motion capture scenarios. The fundamental step for these applications is to accurately estimate human body poses, atomic gestures or sometimes even higher-level (compositional) activities. However, recovering the 2D/3D body layout of a person using only camera images is not straight-forward and suffers from a number of challenges such as, projection ambiguities, high degrees of freedom, occlusions, background clutter, high diversity in clothing appearance, sub-optimal imaging conditions, and sensor noise. These challenges often cause human pose estimation algorithms to fail. We illustrate some of the challenges in the next section 1.1.

There is a variety of settings where articulated 3D human pose estimation has been considered. From *static image* pose estimations to motion capture scenarios in *videos*, *2D or 3D*, *monocular or multiview*, etc. Also different applications require different level of body pose details, from *stick figures* [10] to *detailed mesh models* [13]. In this work, we consider different settings from monocular for 2D to multi-view for 3D human pose estimation, and provide robust and efficient solutions to enable the use of pose estimation for higher level recognition tasks, *e.g*. activity recognition and body-language understanding. Unlike typical approaches, we address the 3D human pose estimation problem in complex settings with multiple humans in the scene with both person-person and person-object interactions. As well we consider the more general pose estimation task for people across a variety of activities and do not rely on activity specific motion priors. We construct the foundation of our work on part-based models due to their success on the task of 2D human pose estimation, in the presence of real-world challenges. The extent of variability in the appearance of human body images in the realistic scenes is quite high. Besides, the human body is deformable and has a large degree of freedom. Hence, in natural settings a human pose can appear in a huge range of configurations. Therefore, a machine learning based approach to recognize the pose of the person in an image require great amount of labeled training data, covering all possible poses and appearances. However, this is not always practical and thus generally leads to pose estimation failures due to blind spots in the training data.

On the other hand, the part-based models such as pictorial structures have shown their ability to cope well in the face of limited amount of training images since each part appearance is modeled separately and hence require way less number of labeled examples. We propose several improvements and refinements to basic 2D pictorial structures model [10], to push their expressive power. Later in Chapter 3, we extend our improved 2D model to enable joint estimation of 2D human body poses in multiple camera views. It is important to note that, we do not perform inference directly in the high-dimensional space of 3D pose configurations to avoid the search overhead in the continuous 3D space. Rather our joint inference allows for geometrically consistent 2D poses across views, which could then be directly triangulated to recover the 3D pose. We demonstrate state-of-the-art results on several 2D and 3D pose estimation benchmark datasets. Moreover, to address the challenges of pose estimation in dynamic environment we propose to adapt the model to the scene (background and humans) at test-time. Our motivation to pursue this task, is the natural intuition to expect that the pose estimation performance should improve if we continuously observe the same scene with the same human subjects for a longer period of time. We study different strategies to integrate the evidence from the test-domain to improve the overall pose estimation model for the scene at hand, without requiring any test-domain labeled data. Figure 1.1 demonstrates output of our 2D and 3D human pose estimation methods.

1.1 Challenges

Reasoning about the 3D scene from 2D images is the general challenge for computer vision tasks. Human pose estimation in particular have to deal with a wide range of difficulties that typically occur due to tremendous variability exhibited by the human body pose in appearance and structure during different activities, and photometric and geometric conditions. These challenges include, projection ambiguities, multiple body scales, multiple viewpoints, huge range of articulations, shape deformation of body parts, illumination changes, background clutter,



(a) Articulations

(b) Occlusions

(c) Scale variations

(d) Foreshortening of body parts, *torso, legs, forearm*

(e) Awkward poses

(f) Motion blur

Figure 1.2: Examples of typical challenges from our evaluation scenarios.(a) huge pose space due to large degrees of freedom, (b) different types of occlusions generally observed, (c) different levels of scale variations, (d) foreshortening of body parts due to out-of-plane rotations and projection ambiguities, (e) awkward poses that are rarely observed at the training time, (f) motion blur.

different types of occlusions (person-person, person-object, or self), high diversity in clothing appearance, sub-optimal imaging conditions, and sensor noise. In Figure 1.2, we illustrate some of the challenges typically observed in our evaluation scenarios. The detailed descriptions about the evaluation scenarios are given in the corresponding chapters in this thesis.

Besides, there are additional challenges specifically for pose estimation in 3D. The degrees of freedom per body part increases from typically four in 2D (x and y translation, rotation, and scale) to seven in 3D, as we have three parameters for translation, three more for rotation, and one for the scale. Therefore, the major difficulty is the optimization or inference in a high dimensional continuous space of 3D human body configurations. Discretization of such a high dimensional space in order to reduce the computational cost severely impacts the pose estimation performance. The state-of-the-art approaches rely on techniques like background subtraction, and/or stochastic search to cut down computational requirements. In contrast to the related work, we deal with the complexity of 3D space by proposing a framework for joint inference of 2D body poses in all views. This is how we avoid the challenges of the 3D space altogether (see Chapter 3). However, this leads to another challenge in the multi-view setup *i.e.* association of the image evidence across different camera views. While one can estimate the homography (linear transformation) between projections of static planar objects in different views for across-view correspondence, however that is certainly not the case for typical human body parts.

1.2 Related Work

Human pose estimation is one of the most studied problems in Computer Vision. Reliable and accurate parsing of human poses in realistic scenes is key to a number of applications such as action recognition [64, 114, 158, 159], human computer interaction [91], and motion capture [80]. Therefore, this task has drawn much research attention during the past 30 years and there exists a huge body of literature and several classes of approaches to recover the human pose from images and videos. In section 1.5 we will survey in more detail the variety of applications that can benefit from human pose estimation. Depending on the application requirement, human pose can either be estimated in 2D image coordinates or real-world 3D coordinates. We consider both domains in this work, but mainly focusing on 3D pose reconstruction using multiple views and self-adaptation.

Like most other computer vision problems, human pose estimation and motion analysis as well started-off from the hypothesized model-driven approaches since the work of Hogg [58], O'Rourke and Badler [85], and Rohr [113]. However, such model-driven approaches were unable to cope with the real-world challenges such as variability in illumination and diversity in poses and appearance. To address these short-comings and with the significant improvements in both data representation and machine learning techniques, the data-driven approaches started to become popular and they dominate the field of human pose estimation to date.

2D Human Pose Estimation. Although, we have discussed the challenges in detail in section 1.1, but it is important to emphasize that the task of 2D human pose estimation from single viewpoint (monocular) specifically suffers from the 3D to 2D projection ambiguities, such as, foreshortening of body parts, and non-trivial foreground/background boundary recovery, etc. Moreover, large degrees of freedom, huge variance in appearance and the person viewpoint makes it even prohibitive. During the last 10-15 years, to address the scarcity of available data and since the introduction of efficient matching algorithms for pictorial structures [39] to images [36, 37], the part-based approaches [10, 12, 116] have nicely demonstrated their potential to perform well in a variety of settings. Literature on 2D human pose estimation is dominated by such part-based approaches. We will discuss these approaches in more detail in Chapter 2.

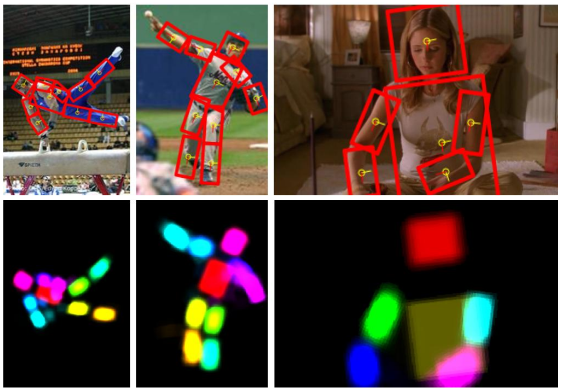


Figure 1.3: Andriluka *et al.* [10]

Earlier approaches tend to propose estimation of human poses based on simplistic graphical models *i.e.* star or tree, in addition to naive appearance representations [36, 37, 107, 110]. Such models are designed to illustrate the potential of computer vision to perform high level tasks, however are too simple to work in realistic images. To overcome the appearance related drawbacks, several researchers started to look into improving the quality of the local appearance representations of the individual parts [10, 106]. Notably, Andriluka *et al.* [10] proposed strong part detectors based on discriminatively trained feature representations. On the other hand, numerous approaches [120, 136, 141, 144, 153] have considered non-tree models to capture dependencies between non-adjacent human body parts, however this makes the overall model quite complex to solve.

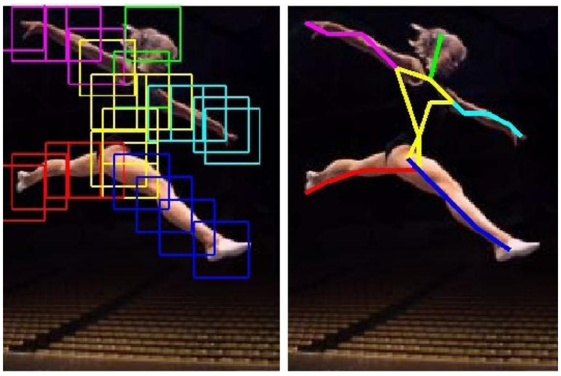


Figure 1.4: Yang and Ramanan [157]

In order to improve the model potential and to cope with the appearance and pose variation problems, the advanced approaches and the current state-of-the-art in human pose estimation in the limited data domain, utilize the concept of mixture components at different levels of their model. Some approaches propose mixture component at the level of body viewpoints (front, left, right, back, etc.) [65, 66], some treat the entire body as a mixture of templates [118], and other approaches even model individual body parts as a mixture of flexible shape templates [157], which provides combinatorial model richness at the local part level, however it does not address the problem at the level of full pose structure. Other approaches include evidence from mid-level representations in the estimation process [94]. Most of these approaches suffer from the rigidity of individual part representations. To address this short-coming, Zuffi *et al.* [164] proposed the idea of deformable structures where every part is represented using shape deformation parameters.

Due to the rise in the amount of labeled data and impressive computational power, very recently the deep learning based human pose estimation approaches have started to show remarkable results [61, 62, 86, 143]. Typically, most approaches rely on the specific type of deep learning architectures, called Convolutional Neural Networks (CNNs), which allows to learn a hierarchy of task specific feature maps. A few recent notable works in deep learning based 2D human pose estimation are [27, 59, 154]. Insafutdinov *et al.* [59] proposed image-dependent pairwise scores in addition to stronger part detectors based on ResNet deep learning architecture [56]. Wei *et al.* [154] proposed convolutional pose machines (CPMs) which employs a multi-stage model and iteratively incorporates global context to refine confidence maps for body parts. Very recently based on these CPMs, Cao *et al.* [27] proposed a part affinity fields (PAFs) representation which encodes unstructured inter-part pairwise relationships for variable number of people.

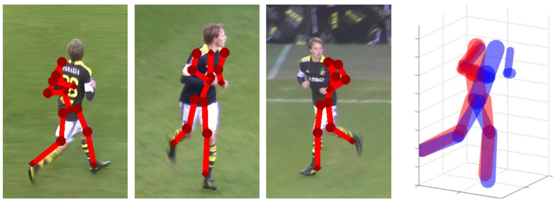


Figure 1.5: Burenius *et al.* [26]

3D Human Pose Estimation. Articulated 3D human pose estimation has been considered in the literature in a variety of settings. Most of the recent work has focused on motion capture scenarios in controlled laboratory conditions [127]. Recent methods in this domain typically rely on detailed body models [16, 42] and elaborate optimization strategies based on stochastic search [42, 128], local optimization [135], or a combination thereof [43]. Although impressive results have been obtained in this setting [44, 135, 146], the developed approaches appear to have difficulties to generalize beyond the motion capture domain. There are many different lines of work to recover the 3D human pose, *e.g.* monocular *vs* multi-view, single time-frame *vs* video sequence, etc.

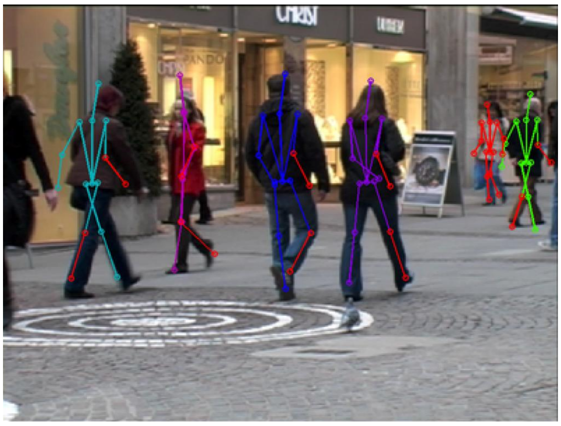


Figure 1.6: Andriluka *et al.* [11]

For the *monocular* approaches, the problem of estimating the 3D pose using images from single viewpoint is under-constrained, therefore the typical approaches in this domain rely on additional evidence to resolve the ambiguity during 2D to 3D lifting [11]. Some approaches consider deriving the 3D human pose by learning a mapping between poses in the 2D and 3D space [132]. Other approaches generate a set of 3D poses and prune them using the evidence from underlying 2D projections [102, 104].

Agarwal *et al.* [1, 2] recovered 3D pose by nonlinear regression using histogram-of-shape-contexts as image descriptors and explored several regression methods. Simmo-serra *et al.* [133] proposed to estimate the 3D human poses relying on pre-computed 2D part detections by propagating the uncertainity from 2D image observations to 3D space assuming reasonable accuracy of 2D parts detectors.

Several other approaches try to disambiguate the correct 3D pose from 2D detections in a single image [52, 102]. But most of these approaches mainly assume simple images with clean backgrounds in controlled laboratory conditions or synthetic images. However, some of the recent approaches tried to bridge the gap between laboratory and real-world settings, *e.g.* Andriluka *et al.* [11] proposed to estimate 3D human poses in outdoor scenes. To deal with the challenges of real-word settings they built upon their tracking-by-detection approach [9] imposing an activity specific temporal motion prior.

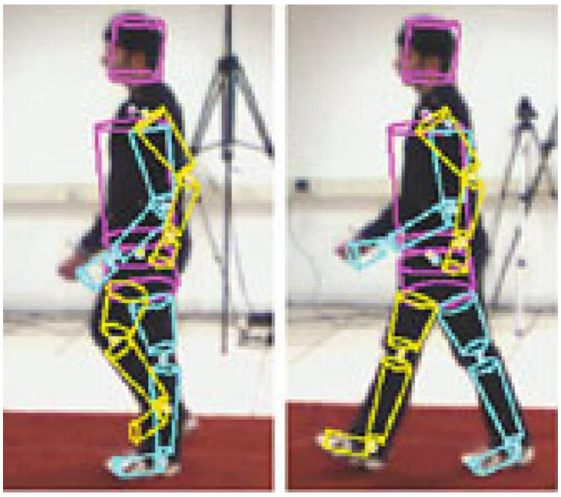


Figure 1.7: Taylor *et al.* [137]

The approaches which reason across *multiple views* naturally perform superior to the monocular approaches. These multi-view approaches often couple the pose estimation problem with tracking [22, 31, 42, 126, 137, 160] to achieve reasonable/desired accuracy. However, these approaches require reasonable initialization to work and cannot recover in case of tracking failures. Moreover, these approaches have been shown to work only in controlled laboratory settings. Various efforts have been made to adapt these methods and enable them for 3D pose estimation in more complex settings but ended up relying on other limiting factors. Such as activity specific pose and motion priors [44], structure-from-motion methods [55] and combining visual observations with on-body inertial sensors [99].

We also operate in multi-camera setup in realistic settings. But in contrast to the prior work, we consider the more general task of 3D human pose estimation for people across a variety of applications without relying on any activity specific motion priors, tracking and other sensors. Unlike previous approaches, in this work we choose an alternative strategy. As the foundation of our approach, we rely on the pictorial strutures model of Andriluka *et al.* [12]. We pursue this avenue due to the representational richness and simplicity of the pictorial structures approach. In prior work, the pictorial structures approach has been shown to work for the 2D human pose estimation case on images of realistic complexity. We extend the standard pictorial structures model to incorporate recent improvements from the 2D pose estimation literature [33, 66, 157]

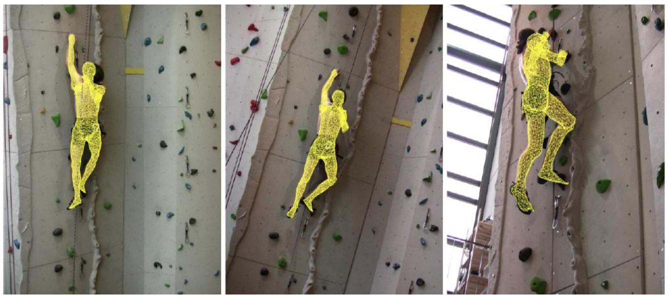


Figure 1.8: Hasler *et al.* [55]

and generalize it to multi-view to enable 3D pose estimation for single and multiple humans. More details on our multi-view pictorial structures approach for single and multiple humans is given in Chapter 3 and 5 respectively.

Model Adaptation. Specifically for human pose estimation the idea of model adaptation is rarely explored in the literature, likely because it is unclear how to robustly extract the person-specific appearance in the presence of noise in the pose estimation process. There are only few exceptions [33, 109, 124] that consider using the evidence from the test environment to improve the pose estimation performance. Ramanan *et al.* [109] detects the stylized poses and use them to track people in the rest of the image sequence. Eichner *et al.* [33] proposed to perform collective human pose estimation given multiple images by sharing color-based foreground/background appearance models. Perhaps the works of Shen *et al.* [124] and Jammalamadaka *et al.* [63] are the closest to what we aim to achieve in our approach for model adaptation. Jammalamadaka *et al.* [63] proposes various pose quality features and discriminatively train the outcome of the 2D pose estimation algorithms. Their model look to predict if the pose estimation algorithm has succeded or not. In our work, we follow the direction similar to [63] but also employ features based on the 3D pose reconstruction, which we find to be highly effective for filtering out incorrect pose estimates. On the other hand, Shen *et al.* [124] proposes an approach for model refinement in videos only. They propose to improve the performance of the pose estimation approach of Yang and Ramanan [157] by incorporating the assumed highly confident pose estimates from the test videos. More specifically, they simply re-train the part classifiers using those confident pose examples. Re-training the part classifiers with new examples seems like a very intuitive thing to do, but such an approach is doomed to degrade the recognition accuracy in the presence of false positives, since the accuracy of those confident examples is not guaranteed. Therefore, in addition to re-training the part classifiers in our model, we also propose another approach to directly use the evidence of the most confident and most similar examples in the pose estimation framework, hence avoiding the trouble of model re-training.

Apart from human pose estimation task, a large amount of methods have been studied for domain adaptation over the years. The main focus of such methods is to address the problem of dataset shift. Dataset shift is the ”gap” between the source and the target domains. The typical approach is the supervised domain adaptation [4, 151, 152, 156], which essentially requires a labeled data for the target domain as well, which they utilize for model refinement in different ways. On the other hand, a number of methods also consider unsupervised domain adaptation, where the training labels of the target domain are not given. Typically they aim to achieve this either by matching the feature distributions in the source and the target domains [48] or transforming the feature space to the target domain [15, 46, 49, 87].

1.3 Proposed Framework for 3D Human Pose Estimation

The overview of the framework of our approach for 3D human pose estimation and model adaptation is illustrated in Fig. 1.9. This framework offers a number of contributions for the task of human pose estimation in general, from single-view 2D pose estimation to multiple views 3D pose estimation. Later, we also propose a strategy to enable joint pose estimation of multiple humans in the scene. Although, it is not explicitly shown in the framework figure 1.9, but we discuss that in chapter 5 in detail.

In the following we discuss the different stages of our 2D and 3D pose estimation framework as shown in Fig. 1.9.

Stage I. First, we employ the state-of-the-art pictorial structures model of Andriluka *et al.* [10] with kinematic-tree prior and simple Gaussian pairwise terms to estimate the dense part marginals, in all views separately. It is important to note that, even though the inference is done separately for all views, the individual 2D pictorial structure models for all views come from our proposed 3D mixture models (section 3.3.3). Hence they implicitly carry some multi-view evidence even before applying any multi-view constraints. Thus the resulting part marginals are geometrically conistent in all views, unlike in the case of naive 2D pose estimations. We then extract sufficient samples from these part marginals to reduce the state-space size. This allows us to employ non-Gaussian constraints, and model complex-pairwise relationships, which is fundamental for our multi-view pictorial structures approach.

Stage II. In this second stage, building upon the success of the state-of-the-art pictorial structures approach [12] and inspired by the recent work in the human pose estimation literature, we propose several improvements to effectively capture the variance in the part appearance and inter-part dependencies for the monocular 2D pose estimation case. This includes, joint shape and color feature representation, flexible part configuration and multi-modal pairwise terms for more expressive representation of part-part relationships, and finally we introduce variance-based component selection criterion for the mixture of pictorial structures components [66].

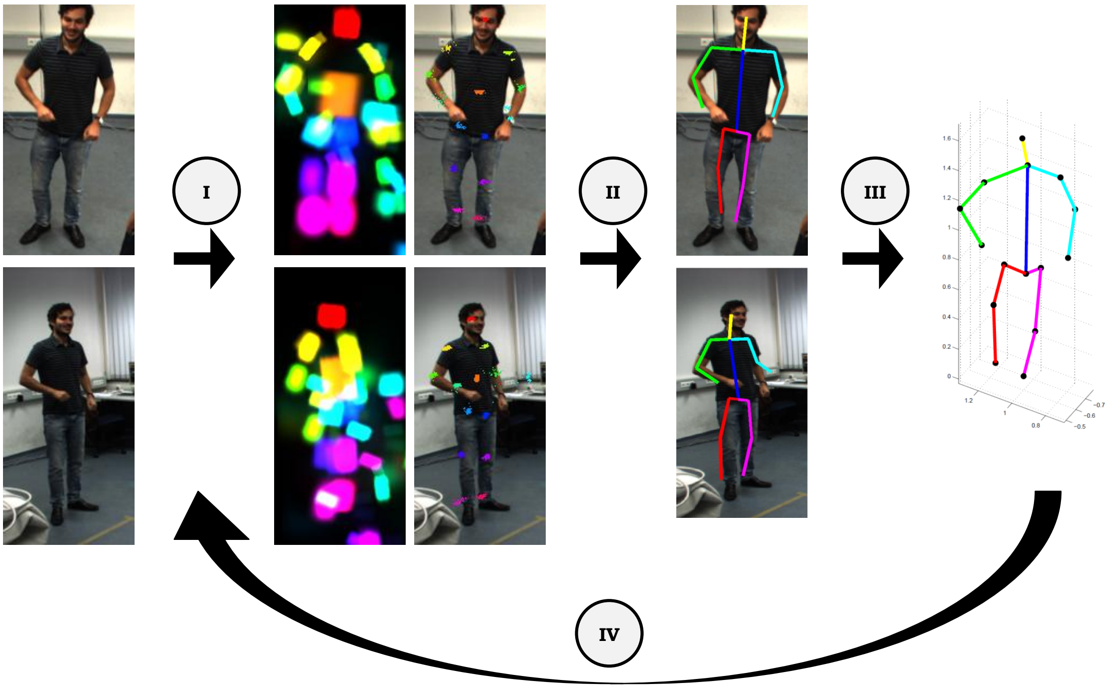


Figure 1.9: Overview of our approach for 3D Human Pose Estimation. (I) 2D pose estimation with simple kinematic-tree prior and sampling of part marginals. In multi-view settings, this kinematic-tree prior is consistent for all views due to our proposed 3D mixture model. (II) In this core step of our approach, we first employ several advancements for more expressive 2D pose estimation (see Chapter 2), and later introduce our multi-view pictorial structures approach to estimate the 2D poses jointly in all views (see Chapter 3). (III) Finally, the 3D pose is reconstructed using triangulation. (IV) In this step, we propose different strategies for improvement of 2D pose estimation model based on the quality of evidence provided by the output 3D pose.

Then as the main contribution in this step we propose a novel *multi-view pictorial structures* approach [6] to enable the human pose estimation in real-world 3D coordinates. We basically employ mutli-view constraints extending our effective monocular approach to estimate the 2D human poses jointly in all views. Our main motivation behind this approach is to avoid the complexity of inference in the 3D state-space.

Stage III. In the third stage, we simply reconstruct the final 3D pose using triangulation of 2D body parts obtained jointly for all views. One notices that this is the simplest stage in our framework, and we do not have to explicitly concern ourselves about the complexities of the inference in the 3D space. This is one of the key advantages of our approach.

Stage IV. Fourth, we build upon our multi-view pictorial structures approach and close the loop in our pose estimation pipeline [8]. To that end, we learn from the scene specfic information to adapt our model and incorporate the changes occured in the scene.

Moreover we also propose an efficient 3D pictorial structures approach in chapter 5 to perform the inference directly in the reduced 3D state-space. We show that this procedure allows us to estimate jointly the 3D poses of multiple human in the scene.

1.4 Proposed Datasets

During the course of this research work, we have introduced several datasets to address multiple vision related tasks including, 2D/3D human pose estimation, 2D/3D human pose tracking, activity detection and classification, and detection of human emotions based on their bodily expressions. Here is the list of datasets we proposed during the last few years,

- MPII Cooking
- MPII Emo
- Shelf

In the following we describe the details on recording of these datasets, and also discuss their exact purpose in our research.

1.4.1 MPII Cooking dataset

We propose the MPII Cooking dataset[1] in order to study the task of activity detection and classification. Although there does exist numerous other datasets with continuous image recordings (videos) [69, 70, 74, 82, 90, 123, 161], our focus is to address the problem of *fine-grained* activity detection. It is important to mention that activity detection is a much more complex task as compared to classification, since it requires identifying the activity in the video sequence and then classifying the same.

To collect our dataset containing fine grained activities, we video recorded 12 participants while cooking different dishes. Participants were asked to choose one to six of a total of 14 dishes. Our dishes include *omelet, fruit salad, salad, fried potatoes, sandwich, soup, pizza, hot drink, mashed potato, potato pancake, cake, snack plate, cold drink*, and *casserole*. We suggested the participants, kind of tools that could be useful for specific dishes, however using them or not was totally their prerogative. Also, we did not tell them the recipe or how to prepare a certain dish. This added more variability to the activities and made the overall setting more realistic. The resulting videos for same activities by different participants turned out to be quite dissimilar and

[1] https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-activities-dataset/

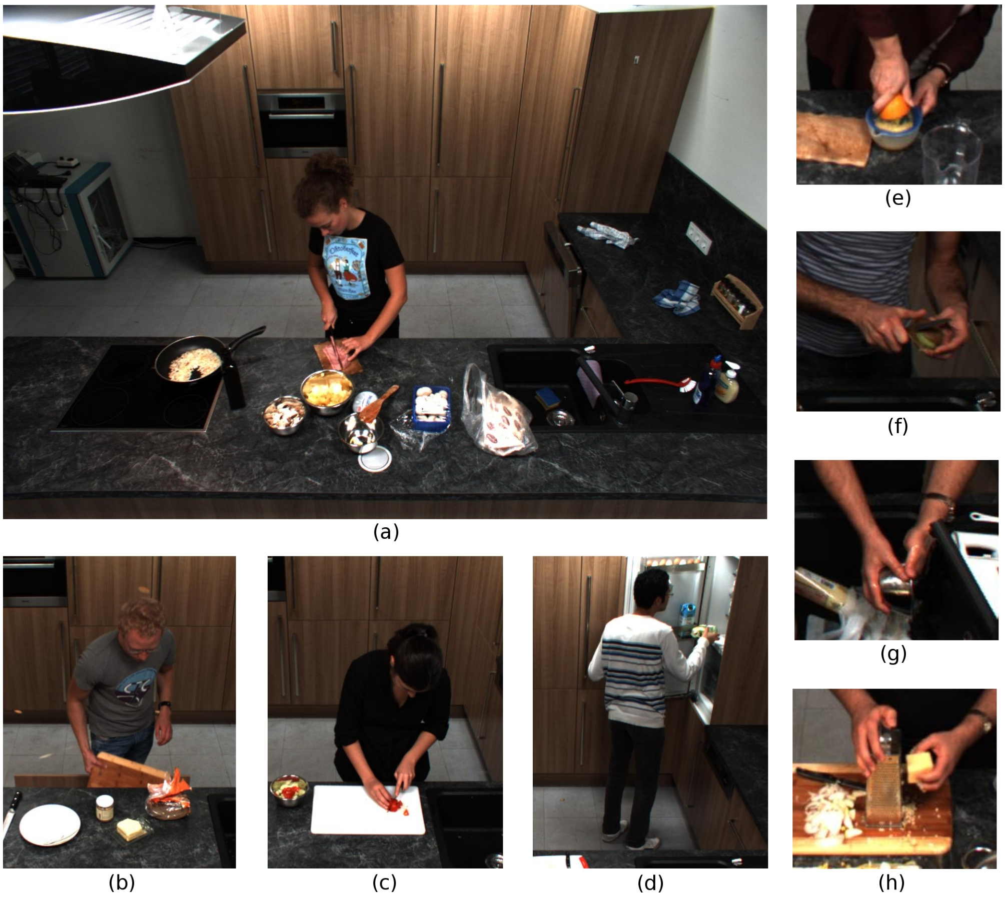


Figure 1.10: MPII-Cooking dataset: Variety of fine-grained cooking activities. (a) Full scene of *cut slices*, and crops of (b) *take out from drawer*, (c) *cut dice*, (d) *take out from fridge*, (e) *squeeze*, (f) *peel*, (g) *wash object*, (h) *grate*

challenging. Before the start of each recording, we let the participants familiarize themselves to our kitchen to feel at home. This way they got already used to the places for ingredients and the required tools. This made the overall recording appear fairly natural. The participants were paid for the recordings, and were typically university students with variable cooking experience (from novice to amateur chefs).

In total, we recorded 44 video sequences with overall length of 8 hours at a frame rate of 29.4fps. We used a camera setup from 4D View Solutions, with Point Grey Grashopper cameras with global shutter and an image resolution of 1624x1224 pixels. For most of the recordings we used multiple cameras, eight to be exact. However, for our activity detection task, we used images only from a single frontal camera.

All of the videos of cooking activities are annotated with fine-grained activity categories with corresponding time intervals. We employed a two stage review process for annotations in order to ensure the quality of activity annotations, since it can be a very subjective thing. For each video, we annotated the start and the end frame number for the activity, as well as its category.

We used Advene [14] annotation tool for this purpose. Since we did not control the activities performed by the participants, the recorded dataset is quite challenging in terms of large intra-class variability and small inter-class variability.

Multi-view Pose Challenge subset: We also provide a subset of our MPII Cooking dataset for the purpose of human pose estimation. As we explained earlier, this dataset was originally recorded for the task of fine grained activity recognition of cooking activities. Later in [6], we used it to benchmark the performance of our 3D pose estimation method. This evaluation dataset consists of 11 subjects with non-continuous images and two camera views. The training set includes 4 subjects and only 896 images and the test set includes 7 subjects and 1154 images. The sampling of frames was done uniformly for all activities, to have a variety of diverse body poses in our training and test sets. We annotate these selected frames with upper-body articulated human pose. In total we annotate 10 joints/parts of the person, namely shoulders, elbows, wrists, and hand joints as well as head and torso. This multi-view pose challenge subset of MPII Cooking dataset can be downloaded from the project website[2].

We refer the reader to our original paper [114] for more details on MPII Cooking dataset.

1.4.2 MPII-Emo dataset

This dataset corresponds to our work on finding emotion cues in bodily expressions and speech during dyadic interactions [79]. Our intuition stems from recent research in body language that suggests human emotions are embedded in their bodily expressions during regular interactions [67, 76, 98, 150]. For this task, we aimed at recording emotionally charged scenarios during natural interactions between couple of human subjects. Therefore, we opted for recording these dyadic interactions without any kind of on-body motion capture equipment.

In our MPIIEmo dataset[3], we propose 224 high quality (1624x1224 pixels @ 29.4 fps) video and audio recordings of couples of actors. Each couple performed in seven scenarios, with each scenario consisting of four subscenarios. This resulted in 224 video clips with a total length of 143 minutes or 252,457 frames. To the best of our knowledge, our proposed MPIIEmo dataset is the first of its kind with multi-camera videos of dyadic interactions of unaugmented people with full body visibility in realistic settings. Moreover the camera views are clock synchronous. In total we recruited 8 couples of actors in our recordings of emotionally charged interactions. The participating actors were university students with atleast one year of theatre training. Moreover, most of these actors were practicing improvisational theatre, which made them quite suitable

[2] http://www.datasets.d2.mpi-inf.mpg.de/amin13bmvc/mpii_cooking3d-1.0.zip

[3] https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpiiemo-dataset/

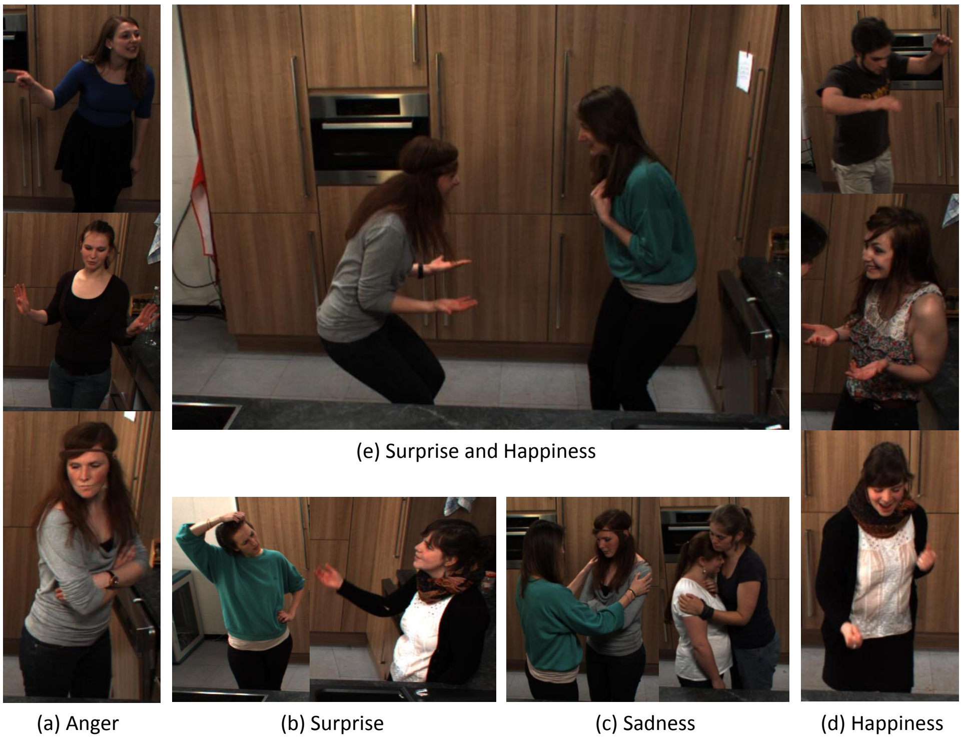


Figure 1.11: MPII-Emo dataset: Examples of bodily expressions associated with different types of emotion categories.

for our recording scenarios. We selected everday scenarios for the recordings to keep it simple and natural for the actors. We had two male only couples, three female only couples, and three mixed. This allowed us to record more variations in bodily expressions to demonstrate different levels of intensities for different emotion categories. Before the start of recording, we provided actors with a short descriptions of approximately 1-3 sentences about the scenarios. Furthermore, we also showed them the kitchen to make them feel comfortable, and get more natural expressions out of them. The actors always rehearsed and warmed up before the recordings. The provided dataset is completely annotated with two different models of emotion labels, *i.e*. categorical and continuous. We selected 4 categorical labels for our annotations *i.e*. anger, happiness, sadness, and surprise. On the other hand, the continuous labels include, valence, activation, power and anticipation.

We refer the reader to our original paper [79] for more details.

1.4.3 Shelf dataset

We introduced this dataset, mainly to study the task of multiple human 3D pose estimation in dynamic settings in our work [17]. The dataset depicts up to 4 humans interacting with each other while performing an assembly task. The Shelf dataset provides 668 and 367 annotated frames

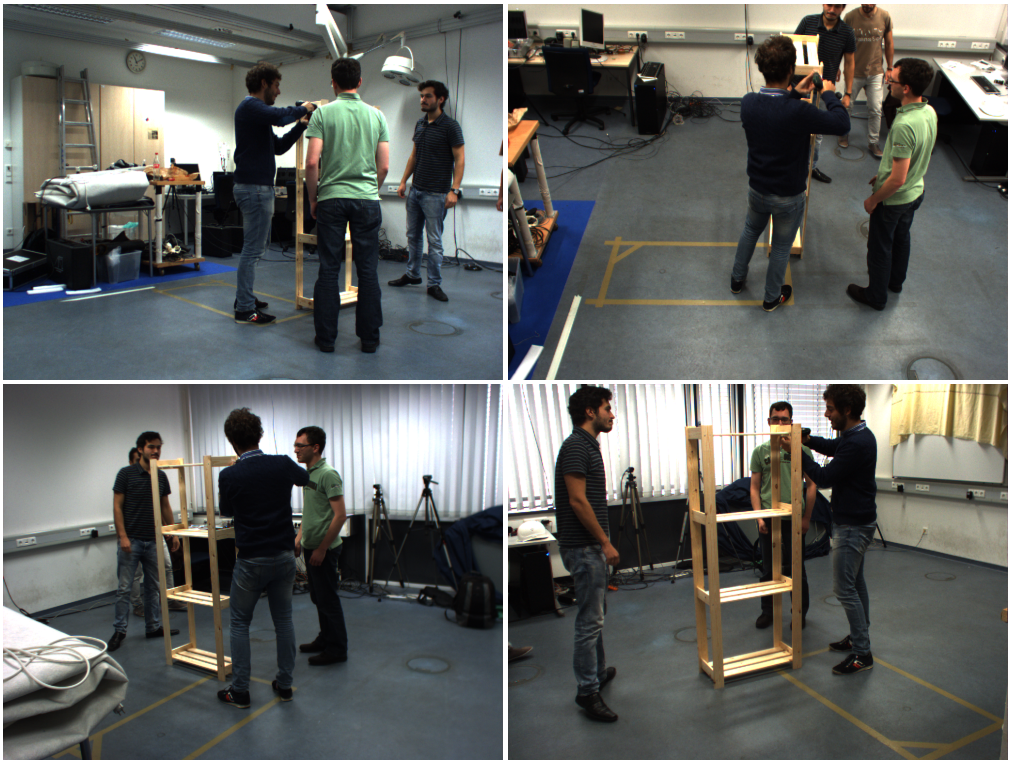


Figure 1.12: Shelf dataset: A benchmark for multiple human 3D pose estimation.

for training and testing respectively. For every frame each fully visible person is annotated in 3 camera views. For this Shelf dataset the training and test splits contain the same human subjects but in dynamic background (constantly changing background due to objects manipulation). We annotate the complete dataset with full-body articulated human pose. In total we annotate 14 joints/parts of the person, namely shoulders, elbows, wrists, hips, knees, and ankle joints as well as head and torso. The complete Shelf dataset can be downloaded from the project website[4].

As described earlier, all these datasets are publicly available. We hope these datasets will foster research for different tasks in computer vision.

1.5 Applications

There is a large variety of potential applications where visual analysis of humans can be utilized. Many real-world applications from surveillance and security to medical diagnosis require accurate prediction of the human body layout. Below we briefly discuss some of these prominent application domains.

[4] http://campar.cs.tum.edu/files/belagian/multihuman/Shelf.tar.bz2

1.5.1 Surveillance, Safety and Security

Surveillance of public places has become necessary for the law enforcement agencies to recognize and prevent criminal activities, monitor trespassing, and for post-crime investigations. Generally, this task is carried out either with the help of human observers or storing huge amount of video data for post event analysis. This procedure is obviously inefficient and sometimes become ineffective to prevent possible mishaps ahead of time.

Therefore during the last decade, to overcome the limitation of human observers to simultaneously see and analyze which is typically a multiple camera environment, the computer vision community has started to look into potential solutions for automatic video surveillance for abnormal event detection, and counting and tracking people in the scene. To that end, it is necessary to detect and recognize human activties automatically. In this work we also consider this application and argue that human body pose contain effective information for activity detection [114]. Other researchers have also found human body layout to be very useful for activity recognition task [64]. Biometrical information from the human pose and gait can also be used for the person identification [162] and re-identification. Moreover abnormal human gait w.r.t to the sence context also contains certain behavioral cues which might reveal the person's intentions.

Moreover, automated systems for human pose and gait estimation can also be deployed to analyze the behavior of elderly people for their *care and safety*. This could include recognizing abnormal postures and behaviours *e.g.* lying on the floor, or fall detection, etc.

Driver assistance systems: One of the critical real-world applicaiton of body pose estimation for human saftey is *Advanced Driver Assistance Systems* (ADAS). By observing the driver pose and the pedestrians on the street, the system can analyze the situation and reason about the possibility of the accidents. This technology enables to improve driver safety and driver experience. This is a very challenging task and requires high accuracy and efficiency.

1.5.2 Sports and Entertainment

Human motion capture has several applications in sports and entertainment industry.

In *sports*, human pose estimation or motion analysis is used for biomechanical study. Typically, performance enhancement in sports is achieved, either by correcting the player's technique for the execution of sporting skills or by discovering more effective techniques to perform those skills. Therefore, recovering the player's body pose in action is necessary for this task. For certain sports, bio-mechanical analysis is used to evaluate the legality of the player's technique, *e.g*. in Cricket, the bowlers are not allowed to straighten their arm more than 15° during their bowling action. This as well requires the qualitative and quantitative bio-mechanical analysis of the player during the action with high detail and accuracy. Generally motion capture for sport biomehanics is achieved using retro-reflective marker based approaches which are quite accurate and effective, but obviously can not be used for on-field sporting actions.

Video games industry is headed towards smart and natural user interfaces for advanced games. Vision based solutions to achieve this goal would certainly be inexpensive. Recent success of Microsoft's Kinect for Xbox, has boosted the interest in natural user interfaces for video games. Kinect relies on a depth sensor to recover the human pose [125], which is then used as a game controller.

In *film* industry, computer graphics is being extensively used for the last few decades. But only very recently with the advancements in MoCap technology, the computer graphics experts are able to generate real-life animations of virtual characters. To this end, first the MoCap technology is used to capture real human motion and then the virtual character is animated using this captured motion pattern. Marker-less motion capture technology would allow actors to perform in a more natural context without being distracted by special suits and/or markers. However, vision based 3D motion capture technology has not yet achieved the desired accuracy required for the realistic animation purposes [28]. However, with the success of deep learning techniques on various computer vision tasks, one shall witness purely vision-based seamless motion capture technology in the near future.

1.5.3 Robotics

One of the important applications for human pose estimation is in the domain of *Robotics*. For robots to effectively interact with humans it is necessary for them to recognize the human motions, understand their actions and predict what they are going to do in the immediate future. Especially, predicting the next future action is a critical characteristic for a robot to safely operate alongside humans. This ability can enable the robots to plan ahead of time and/or improvise according to the situation.

1.5.4 Medicine

Human pose estimation enables high level reasoning in the context of clinical analysis. Marker-less motion capture (MoCap) technology would be quite useful for non-invasive applications in medicine e.g., classification of gait pathologies. In some works researchers have tried to consider the task of gender classification given human motion patterns [53, 140]. Troje *et al*. [145] and

Johannes *et al.* [77] studied the associations between human gender, psychological condition and their gait patterns.

1.6 Outline of the thesis

Chapter 2: Expressive Pictorial Structures for 2D Human Pose Estimation.
In this chapter we propose several extensions to the original pictorial structures model of Andriluka *et al.* [12], making it more expressive for challenging settings. We evaluate different settings of our EPS model and its various components for the task of 2D pose estimation on two different benchmarks, MPII-Cooking pose challenge and ”Leads Sports People” (LSP) datasets. We find that with better appearance model (joint shape and color), the part detectors are able to disambiguate significantly between positive and negative part hypotheses. Also, more effective spatial terms and mixture of pictorial structures model allow us to represent larger range and several modes in the human pose space. As the final contribution in this chapter, we demonstrate the application of our 2D human pose estimations for the task of fine-grained activity detection and classification. To that end, we provide an implementation of a robust SIFT-based tracking algorithm for the tracking of human body-joints. A more detailed description of this activity recognition work can be found in our originial CVPR12 publication [114].

The contents of this chapter correspond to the **single-view** model of our BMVC 2013 publication *Multi-view Pictorial Structures for 3D Human Pose Estimation* [6]. Sikandar Amin is the lead author of this paper.

Chapter 3: Multi-view Pictorial Structures for 3D Human Pose Estimation.
In this chapter we introduce several extensions to our 2D pose estimation approach to enable human pose estimation in 3D, avoiding the complexity of inference in 3D state-space. To that end, we introduce 3D mixture model to encourage inference of consistent part marginals in all views. Then, we integrate two different multi-view constraints directly in the pictorial structures framework to estimate the 2D poses jointly in all views. This allows us to simply reconstruct the final 3D pose directly using these consistent 2D pose estimates. Our evaluations show that we achieve similar or better performance compared to state-of-the-art without using tracking or exploiting activity specific priors.

The contents of this chapter correspond to the **multi-view** model of our BMVC 2013 publication *Multi-view Pictorial Structures for 3D Human Pose Estimation* [6]. Sikandar Amin is the lead author of this paper.

Chapter 4: Test-time Adaptation for 3D Human Pose Estimation.
In this chapter we propose an approach to 3D pose estimation that adapts to the input data available at test time. We analyze two strategies for finding confident pose estimates: discriminative classification and ensemble agreement. We show that best results are achieved by combining both strategies. However, ensemble agreement is found to be the strongest cue to disambiguate between correct pose estimates and false positives. Also, our findings suggest that retraining the part detectors with confident examples is not always the best option. We show the significant improvements using our approach on two publicly available datasets.

The contents of this chapter correspond to our GCPR 2014 publication *Test-time Adaptation for 3D Human Pose Estimation* [8]. Sikandar Amin is the lead author of this paper.

Chapter 5: 3D Pictorial Structures for Multiple Human Pose Estimation.
In this chapter we present a 3D pictorial structures model for recovering 3D poses for multiple humans using the multi-view potential functions. We introduce a reduced state-space which allows fast inference. In our evaluation we show that our model is able to predict poses for single and multiple humans successfully without knowing the identity in advance. We do not require a background subtraction step and our approach relies on 2D body joint detections in each view, which can be noisy. In addition, we introduce two datasets for 3D body pose estimation of multiple humans.

The contents of this chapter corresponds to our CVPR 2014 publication *3D Pictorial Structures for Multiple Human Pose Estimation* [17]. Sikandar Amin is the second author of this paper, and contributed in generating initial set of part detections and collaborated to construct the reduced 3D state-space which is at the core of our 3D pictorial structures approach. Moreover, Sikandar Amin also contributed in the detection of multiple humans in the scene using multiple views.

Chapter 6: Conclusions and Future Work.
In this final chapter we summarize our contributions in this thesis. We discuss the advantages and disadvantages of our methods. Finally, we propose different future perspectives to extend or build upon our approach for human pose estimation. We also discuss other general future work.

Chapter 2

Expressive Pictorial Structures for 2D Human Pose Estimation

In this chapter, we consider the problem of 2D human pose estimation in challenging real world settings. We start with a brief introduction of state-of-the-art pictorial structures approach. Then we describe our improvements inspired by the recent advances in the field, specifically we employ *joint shape and color* features [33] for more discriminative appearance representation, *flexible part configuration* & *mutli-modal pairwise terms* [120, 157] to improve the expressiveness of the inter-part pairwise relationships. Finally, we also take advantage of *mixture* of pictorial structures [65] to better represent the multiple modes in the global human pose configurations, and propose a novel approach for mixture component selection.

Later we demonstrate the advantages of our proposed approach on two different 2D human pose estimation benchmarks. We show that our extended model outperforms the baseline and the state-of-the-art on the pose challenge of MPII-Cooking dataset. Moreover, we show that our approach outperforms most of the existing methods on LSP dataset, indicating the effectiveness of our approach in general and challenging real-world settings.

In Chapter 3, we build upon this proposed 2D pose estimation approach and discuss the extensions to incorporate multiple views for robust and efficient 3D human pose estimation.

2.1 Introduction and related work

In this chapter, we consider the challenging task of articulated 2D human pose estimation in monocular images. Part-based approaches are typically considered the state-of-the-art for this task [12, 95, 157]. The foundation of our work in this thesis as well correspond to this area of research. The need for part-based models such as pictorial structures arise due to the deficiency in amount of available data. Due to the large degree of freedom and deformable nature of human body, the human pose can take a huge range of configurations. Therefore collecting large amount of labeled data in this case, covering all possible poses and appearance, is impractical and leads to blind spots in the training data. The holistic approaches [60, 112, 146] are vulnerable due to the model rigidity and therefore perform poorly as a result of these training data blind spots. These holistic approaches are designed to estimate the positions of the body parts directly from the image evidence without relying on any intermediate part-based representation. These approaches have demonstrated reasonable results in controlled laboratory settings but have not yet shown any significant performance on the realistic images in natural settings with cluttered backgrounds. As well, such approaches require tremendous amount of labeled data covering a wide range of plausible human poses. To overcome this limitation the pictorial structrues, which is basically a piece-wise approach, is designed to capture the pose articulations which are not explicitly seen during training, assuming some local rigidity at the level of individual parts. The pictorial structures approach has proven most reliable in capturing the pose articulations and appearance variability in typical natural settings with limited data.

Problem of matching pictorial structures to images was first introduced by Fischler and Elschlager [39] more than 40 years ago. In 1973, they introduced the concept of representing an object as a configuration of rigid pictorial elements. They also considered inference tractability issues and discussed dynamic programing for efficient model inference. The idea was later rediscovered for the problem of human pose estimation in images by Felzenszwalb and Huttenlocher [36, 37]. They represented the model as a configuration of several body parts. Also they discussed the statistical understanding of the pictorial structures framework and framed 2D pose estimation as inference in a probabilistic model. This work of Felzenszwalb and Huttenlocher [37] proved to be the milestone in the research of 2D human pose estimation. Several researchers carried on with their basic pictorial structures model and proposed numerous advancements to overcome the challenges of generic real-world settings. Basically, there are two core components of pictorial structures model, *i.e. appearance representation* and *body prior model*. The typical approaches either propose stronger local part appearance or try to improve the body prior, making it more flexible and enabling it to successfully represent the general human poses. Few exceptions exist which consider both problems jointly [120, 157].

The first line of research aims at improving the local appearance model of the individual parts within the pictorial structures framework to better represent the most discriminative features. Earlier work [36, 37, 107, 110] severely suffered due to the use of very simple part detectors, based on the prior assumption about shape of the human body parts. Ren *et al.* [110] exploited the concept that the limbs of the human body can be characterized by the pair of parallel line segments. Ramanan and Forsyth [107, 108] proposed an approach for finding people by bottom up inference where the part appearance was based on convolution of edge templates. Naturally, the idea of hypothesizing a fixed part appearance makes these approaches sensitive to even slight variations. To overcome these drawbacks, researchers started to look into improving the quality of the appearance representation [10, 106]. In 2009, Andriluka *et al.* [10] proposed to integrate a powerful discriminative appearance model, based on boosted part detectors, directly in the pictorial structures framework for human pose estimation in generic 2D images. They represent the image evidence by a densely computed grid of *shape-context* descriptors [18, 78]. The other widespread feature representation to model the part appearance for pictorial structures has been *histogram of oriented gradients* (HOG) descriptor, introduced by Dalal and Triggs [30] for the task of human detection in images. Over the years, several approaches [65, 95, 106, 120, 157] have considered discriminatively trained appearance templates based on HOG feature descriptor. Also, several works [38, 95, 120] have argued that more specialzed models for body parts significantly improve the overall pose estimation performance.

The other important line of research focused on improving the prior model of the human pose. Over the years it has been proven that the performance of pictorial structures significantly depend upon whether the assumed prior model is flexible and informative enough to represent the huge variation in human poses or not. Typically, the body prior distribution is assumed to decompose as a tree structure to allow for efficient inference at test time. Yet, several approaches [120, 136, 141, 144, 153] have considered non-tree models to capture dependencies between non-adjacent human body parts, which makes the overall model quite complex to solve. Andriluka *et al.* [12] proposed a more generalized version of their model in order to cope with shortcomings of tree-based pictorial structures models *e.g.* overcounting of image evidence. Some methods [94] aimed to increase the flexibility of the tree-structured pictorial structure model relying on mid-level representations, such as *poselets* [21].

Recently, several approaches have been presented which focused at improving the overall model to improve the results. Yang and Ramanan [157] proposed an approach with mixture of discriminatively learned part templates, which provides combinatorial model richness at the local part level and the pairwise terms, however it does not address the problem at the level of full pose structure. Sapp *et al.* [120] proposed to use body joints instead of limbs in their pose model. This allowed them to naturally model the large variability in the limb length and implicitly model foreshortening. In our method, we adopt to similar flexible representation but we still search over the discretized set of orientations unlike [120] who assumed the joints are sort of rotation invariant.

Such standard single model pictorial structure approaches although efficient and robust for articulated pose estimation in images but do not perform well under extreme viewpoint variations. For example, a model trained for frontal poses can not perform for the non-frontal cases (side, back) due to the kinemtic constraints learned specific to the frontal poses. Also, the appearance of body part varies when a person is seen from a different viewpoint. Johnson and Everingham [65, 66] extended the pictorial structures approach and introduced mixture components to

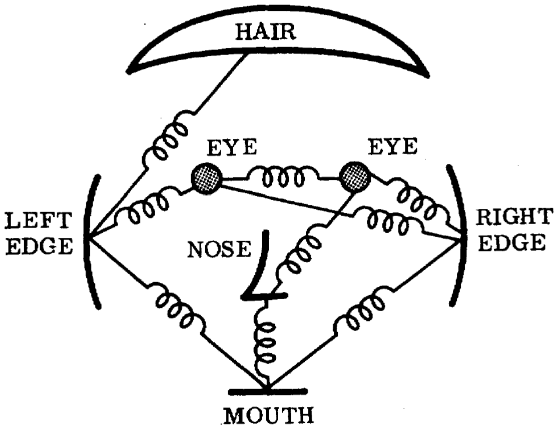


Figure 2.1: Pictorial structures representation of a Face, illustrating the different components of the model *i.e.* parts and springs (connections). This figure is reproduced from the original work of Fischler and Elschlager [39].

represent the multiple modes in the space of human poses in images. They clustered the pose annotations and learned a mixture of tree-models for each cluster.

Our contributions. In this work, we propose a 2D pose estimation approach that extends the state-of-the-art 2D pictorial structures model of [12]. As a first contribution, we propose a stronger appearance model based on joint shape and color feature representation. Secondly, motivated by the recent work we employ more effective spatial terms and we generalize [12] to a mixture model and propose a novel approach for mixture component selection. Consequently we obtain a more expressive pictorial structures model.

2.2 Pictorial structures model

The *Pictorial Structures* is a mathematical model to find objects in an image. It was first introduced by Fischler and Elschlager [39] in 1973. The fundamental concept behind this approach is to represent objects as a collection of parts arranged in a deformable configuration to efficiently capture variability in the shape/appearance of the object. The connection between two parts is typically characterized as a spring, to describe the deformation aspect. The example pictorial structure representation of a human face is shown in Figure 2.1 (reproduced from the original work of Fischler and Elschlager [39]). They formualted the matching of a pictorial structure to the given image as an energy minimization problem. Then, optimizing over a number of model parameters allows to find the single best match of pictorial struture model to the image evidence.

Later, Felzenszwalb and Huttenlocher [36, 37] described the statistical formulation for the pictorial structures framework and they have shown that the *maxmimum a posteriori* estimate of the object configuration given an observed image in this case, is equivalent to the energy minimization problem introduced by Fischler and Elschlager [39]. As the distribution over the object configuration is estimated in this case, therefore this allows to produce multiple potentially good object pose hypotheses.

Given an image evidence I, the objective is to infer the posterior probability $p(L|I)$ of the part configuration L. Estimating this posterior distribution directly is not obvious, therefore the standard way to approach such problems is using Bayes' theorem and work with the equivalent formulation. Applying Bayes' rule we can write the equivalent formulation of this posterior density as,

$$p(L|I) \propto p(L)\, p(I|L) \tag{2.1}$$

Here $p(I|L)$ represents the likelihood of the image evidence given a specific configuration L of the object. and $p(L)$ describes the prior distribution of the object part configurations. Note, the proportionality in Eq. 2.1 means that we do not consider $p(I)$ in the maximization of posterior $p(L|I)$, since the prior probability of the image evidence $p(I)$ is independent of L.

The choice of the prior model $p(L)$ is as critical as the representation of image likelihood $p(I|L)$. One of the simplest and straight-forward prior would be to consider all possible part configurations in the pose space occur with equal probability, a *uniform* prior. But such prior is not informative, therefore has no implication on the outcome. Mathematically, in this case maximizing the conditional probability in Eq. 2.1 would then lead to a *maximum likelihood* estimate instead of a *maximum a posteriori* estimate. We will discuss in Sec. 2.3.3 the type of body pose prior $p(L)$ we empoy in our model.

2.3 Expressive pictorial structures (EPS)

We aim to introduce more effective pictorial structures model that is not limited by the simple tree-structured kinematic pose prior. Therefore, as the foundation of our work, we rely on the implementation of the pictorial structures model by Andriluka *et al.* [12], which enables us to model non-Gaussian pairwise relationships and complex interactions between parts that result in a loopy graphical structure. In addition to typical kinematic pairwise constriants Andriluka *et al.* [12] also employ repulsive factors, typically between parts with similar appearance *e.g.* right and left lower legs. Such parts might otherwise end up at the same image location, in effect counting the same image region twice. This phenomenon is typically known as *double counting* problem. These repulsive constraints are shown in Fig. 2.2 with the dashed lines.

The human body is represented as a configuration $L = \{l_1, \ldots, l_N\}$ of N rigid parts and a set of pairwise part relationships E. The location of each part is given by $l_n = (x_n, y_n, \theta_n)$, where

(x_n, y_n) is the image position of the part, and θ_n is the absolute orientation. We formulate the model as a conditional random field (CRF), and assume that the probability of the part configuration L given the image evidence I factorizes into a product of unary and pairwise terms.

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(l_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(l_i, l_j). \tag{2.2}$$

where $f_n(l_n; I)$ represents the unary term for part n, and $f_{ij}(l_i, l_j)$ is the pairwise term between parts i and j. The assumption underlying this factorization is that the likelihood of the part configuration can be decomposed into the product of individual part likelihoods, making the inference tractable in practice [12, 29, 37]. Since the product of the factor scores in above Eq. 2.2 results in an un-normalized measure, therefore we write the partition function Z in the denominator to normalize it to the actual probability distribution.

In the following, we will discuss our improvements over the baseline model of [12].

2.3.1 Flexible pictorial structures (FPS)

Andriluka *et al.* [12] represents human pose as a configuration of parts corresponding to rigid body limbs (upper/lower arms and legs), torso and head. The learned appearance templates do not perform well due to foreshortening of body parts that occur as a result of out-of-plane rotations. Therefore, following the work of [120] we slightly change the part configuration. Instead of encoding body pose via a configuration of limbs we encode it via a configuration of body joints. The advantage of switching from limbs to joints is that the new model can better encode the foreshortening of body parts. Also this results in increased flexibility of representing the limbs. Depending on the task, the number of body parts may vary (see Fig. 2.3). In our flexible variant of PS model (FPS), for upper body pose estimation we use 10 parts corresponding to head, torso, as well as left and right shoulder, elbow, wrist and hand. In case of full body pose estimation we do not consider hand parts instead we use additional six lower body parts corresponding to (left/right hip joints, knees, ankles/feet), resulting in 14 parts in total.

In the following, we comprehensively discuss the components of the pictorial structures model from Equation 2.2 *i.e.* the part likelihood terms and the spatial constraints between pair of parts. Unlike Fischler and Elschlager [39], both of these terms are learned from the training data [10, 37]. Later we also discuss the mixture of pictorial structures and methods for the mixture component selection.

2.3.2 Joint appearance representation

The part likelihood term $f(l_n; I)$ are represented using boosted part detectors that rely on the encoding of the image using a densely computed grid of shape context descriptors [18]. The feature vector is formed by concatenating the descriptors inside the part bounding box. Then the part likelihood term is given by

$$f_n(l_n; I) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}(e_n(l_n))}{\sum_t \alpha_{n,t}}, \varepsilon_0\right) \tag{2.3}$$

where $e_n(l_n)$ is the feature vector corresponding to part i extracted at location l_n, $h_{n,t}$ are weak single-feature classifiers, and $\alpha_{n,t}$ are their corresponding weights learned with AdaBoost.

Color. We augment the shape context features $e_n(l_n)$ used in the boosted part detectors with color features. The intuition behind this is that certain body parts such as hands or the head frequently have a characteristic skin color [32]. Additionally, certain colors are more likely to correspond to background than to one of the body parts [33]. For this we encode the color of the part bounding box using a multi-dimensional histogram. To this end, we investigate two different color models which are quite popular in the literature, although there exist several other possibilities. The first one is the RGB (red, green, blue) color model. RGB is an additive color model which is typically used in most of the devices with color displays (TV, digital cameras, image scanners, LCD, etc.). The second one is the HSV (hue, saturation, value) color model which is the transformed and more intuitive representation of RGB from human perspective, since H component alone represents the color of the light. The V component of this HSV model represents the brightness of the pixel. In the literature [71], it has been found that giving relatively lesser importance to this brightness component (V) in the color representation (typically histogram) allows to better cope with the illumination variations in the scene.

We evaluate our model with different number of color histogram bins for each of the dimensions of the RGB/HSV color space in Sec. 2.4.1.2. For example, 10 bins for each dimension results in a $10^3 = 1,000$ dimensional feature vector. We concatenate the shape context with the color features and learn a boosted part detectors on top of this combined representation. Note that adding color information alongside the shape information allows us to automatically learn the relative importance of both features at the part detection stage.

2.3.3 Expressive spatial model

The other important aspect of the pictorial strutures approach is the spatial model which represents how the parts relate to each other in the probabilistic framework. As, discussed earlier, this component of the PS model also require considerable attention to recover the human poses. The

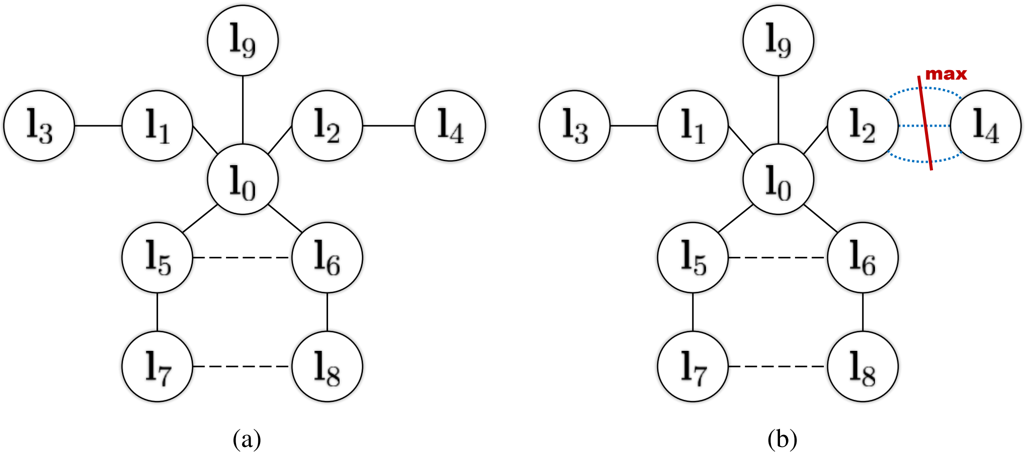


Figure 2.2: Comparison of the pictorial structures models: (a) Original model of Andriluka *et al*. [12] with kinematic factors (solid lines) and repulsive factors (dashed lines). (b) Our model with multi-modal pairwise terms ($\max$ over dotted lines).

importance of these spatial constraints intensify when part detectors are not strong enough which is typically the case in natural real-world settings, and enforcing a structure then emphasizes the plausible pose candidates.

The pairwise terms $f_{ij}(l_i, l_j)$ in Eq. 2.2 encode the spatial constraints between any pair of model parts (i, j) and are modeled with a Gaussian distribution in the transformed space of the joint between the two parts:

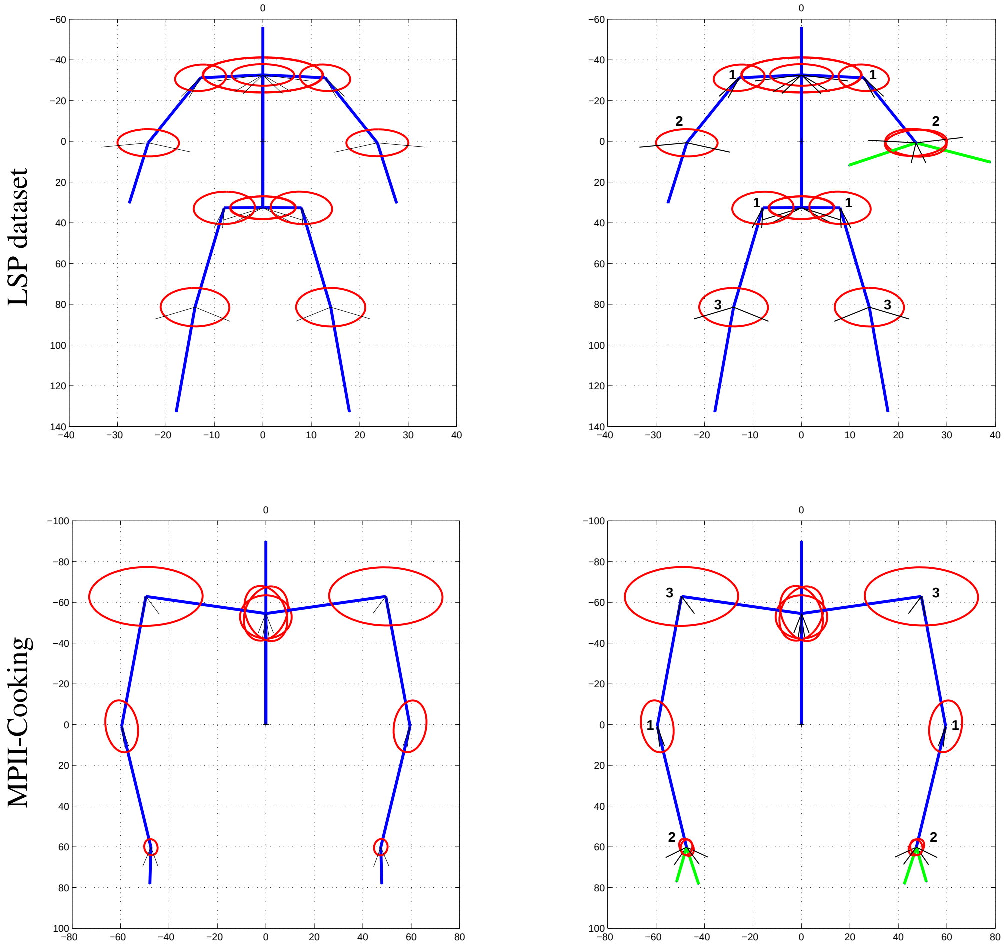
$$f_{ij}(l_i, l_j) = \mathcal{N}(T_{ij}(l_i) - T_{ji}(l_j)|\mu_{ij}, \Sigma_{ij}), \tag{2.4}$$

where T_{ij} (see Eq. 2.5) is the mapping between the location of part i and location of the joint between parts i and j, μ_{ij} represents the preferred relative orientation between parts in the transformed space and Σ_{ij} encodes the flexibility of the pairwise term. The parameters of the pairwise terms are learned in piecewise fashion using maximum likelihood estimation.

$$T_{ij}(l_i) = \begin{pmatrix} x_i + d_x^{ij}\cos\theta_i - d_y^{ij}\sin\theta_i \\ y_i + d_x^{ij}\sin\theta_i + d_y^{ij}\cos\theta_i \\ \theta_i \end{pmatrix} \tag{2.5}$$

Here, the $d^{ij} = (d_x^{ij}, d_y^{ij})^T$ is the position of the joint between parts i and j represented in the coordinate system of the part l_i. The value of d^{ij} is also learned from the training data during maximum likelihood estimation.

Representing the probability distribution of the pairwise part dependencies as a uni-modal gaussian distribution is naturally a very weak assumption. Since we use joint based body configuration which means the body limbs in our approach are represented using combination of two



(a)

(b)

Figure 2.3: Learned body prior models. Top row: Kinematic constraints learned on the LSP dataset for full-body pose estimation. Bottom row: Kinematic constraints learned on the MPII-Cooking dataset for upper-body pose estimation. (a) Prior on part configurations learned using the PS model of Andriluka *et al.* [12]. (b) In contrast our model learns potentially multiple modes for all pairwise terms. We show such multi-modal pairwise terms only for a few parts (in green color) for the sake of clarity. The numbers next to the joints in our model indicate the number of estimated modes K using Eq. 2.7.

neighboring joints in the graphical model. Therefore, the effect of foreshortening due to out-of-plane rotations is already reduced, as far as the part appearance is concerned. Still the huge variations and multi-modal nature of relative angle between parts lead to multiple modes in the projected part lengths $\|d^{ij}\|_2$, and influence the pairwise terms between parts.

Multi-modal pairwise terms. We extend our model by introducing mixture models at the level of these pairwise part dependencies. To that end we replace the unimodal Gaussian term in Eq. 2.4 with a term that maximizes over K modes and represent each mode with a Gaussian.

The new multi-modal pairwise term is then given by:

$$f_{ij}(l_i, l_j) = \max_{k=1}^{K} \mathcal{N}(T_{ij}^k(l_i) - T_{ji}^k(l_j)|\mu_{ij}^k, \Sigma_{ij}^k). \tag{2.6}$$

The number of modes K is computed using k-means. To that end, we first cluster the data w.r.t. the relative angle between the two parts with different number of clusters $\hat{K} = \{1, 2, \ldots, 5\}$[1]. Then we select the one which minimizes the sum of variances in all clusters for a given $\hat{K}$. This procedure is give by the following Eq. 2.7,

$$K = \underset{\hat{K}}{\operatorname{argmin}} \sum_{\hat{k}=1}^{\hat{K}} \operatorname{var}(\Phi_{ij}^{\hat{k}}) \tag{2.7}$$

Here, $\Phi_{ij}^{\hat{k}}$ represents the relative angle between parts i and j of the training poses corresponding to the cluster $\hat{k}$. Once the number of clusters K is selected then the parameters of the corresponding gaussian mixture model *i.e.* μ_{ij}^k and Σ_{ij}^k of Eq. 2.6 are learned directly using the *maximum likelihood estimation* from the data of the corresponding cluster k. The parameters of the gaussian mixture models can be equivalently estimated using expectation maximization. Note that this pairwise term is similar to the one used in [157], but has a somewhat different form as it incorporates both relative orientation and position of model parts whereas in [157] only the relative position is modeled and orientation is represented via an additional latent variable. Fig. 2.3 illustrates body priors in two different settings, full body 14 parts model for the LSP dataset and upper-body 10 parts model for the MPII-Cooking. In Fig. 2.3b we show the body prior model with our multi-modal pairwise terms in comparison to the one employed by Andriluka *et al*. [12] (cf. Fig. 2.3a). For the sake of clarity we only show the multiple modes for a few joints (in green), however the number of modes estimated (using Eq. 2.7) on the training set for all joints are given next to the joint locations in Fig. 2.3b.

2.3.4 2D Mixture PS

Previously we have discussed our proposals to allow more effective pairwise relationship by introducing the mixture model at the level of such pairwise part dependencies. We also illustrated that such multi-modal pairwise terms (see Eq. 2.6) are better representative of the training data in capturing multiple modes in the body joint articulations. Our evaluations (see Sec. 2.4) show that this approach is although quite effective but still fails to capture the multiple modes in the holistic pose space of human body. These multiple modes arise due to different viewpoints a person can take w.r.t. to the camera *e.g.* front, back, left, right, etc. To overcome this shortcoming in our model, we follow the work of Johnson and Everingham [65, 66] and extend our approach to a mixture of pictorial structures models. We obtain the mixture components by

[1] We empirically set the maximum number of clusters to 5.

clustering the training data with k-means and learning a separate model for each cluster. The components typically correspond to major modes in the data, such as various viewpoints of the person with respect to the camera. To cluster the data we first transform the poses of the training data in the coordinate framework of some pre-selected body joint (typically *neck-torso* joint). This transformation includes only translation and scale. We do not rotate the poses otherwise it will lead to single viewpoint cluster. We use the torso size to estimate the person scale during training.

The index of the component is treated as a latent variable that should be inferred at test time. We found that the value of the posterior in Eq. 2.2 is unreliable to predict the optimal mixture component, and propose two alternative strategies.

Component classifier. We train a holistic classifier that distinguishes the mixture component based on the contents of the person bounding box. For this we rely on the approach of [93] who jointly solve the tasks of object detection and viewpoint classification. This approach is similar to DPM [35], but relies on a structured prediction formulation that encourages both correct localization and component detection. When applying [93] to our setting we replace viewpoint classification with mixture component classification.

Minimum variance (min-var). We select the mixture component using criteria directly related to the quality of the pose estimation. Inspired by the recent work of [63] we select the best component with the minimal uncertainty in the marginal posterior distributions of the body parts. The criteria used to measure the uncertainty is given by

$$s(k, I) = \sum_{n=1}^{N} \|\text{Cov}_n(k, I)\|_2, \tag{2.8}$$

where $\text{Cov}_n(k, I)$ is a covariance matrix corresponding to the strongest mode of the marginal posterior distribution $p(\mathbf{l}_n|I)$ of the component with index k. The index $\hat{k}$ of the selected component is chosen as $\hat{k} = \text{argmin}_k s(k, I)$.

2.3.5 Inference

We can perform efficient and exact inference using pictorial structures model, as long as we limit our body model to a simplified tree-structure and represent the pairwise part dependencies by Gaussian distributions [10, 36, 37]. However, these simplifying assumptions limit the expressiveness of the model that we aim to improve in this work. Our proposed extensions to boost the flexibility and effectiveness of our model to capture even larger variations in the pose space create loopy dependencies in the structure of our model. Moreover the dependencies introduced by multi-modal pairwise terms in Eq. 2.6 are not Gaussian anymore. Therefore, to perform inference in such a loopy model with non-Gaussian pairwise factors we rely on the approximate two-stage inference procedure introduced in [12]. This procedure will also help us during inference in the multi-view case that we will discuss in the next chapter (Sec. 3.3.4). In the first stage, we employ a simplified tree-structured model with Gaussian pairwise factors (cf. Eq. 2.4) and simple shape and color appearance terms (see Sec. 2.3.2) to estimate the dense marginal distributions $p(l_n|I)$ for all body parts. The inference is performed with sum-product belief propagation algorithm. Andriluka *et al.* [10] showed that this inference is exact and efficient as the model is tree-structured and the belief propagation messages can be computed with simple Gaussian convolutions. Then, we generate the body part hypotheses by sampling the proposals from these densely estimated part marginals $p(l_n|I)$. In all experiments we sample the marginal distributions 1,000 times for each body part and remove the duplicates. We found this number to be sufficiently large in our experiments. In practice the number of duplicate part hypotheses retrieved during the sampling procedure is quite high. After removing duplicates typically we end up with around 500 hypotheses. This first stage of our adopted inference procedure, reduces the state space of the body part hypotheses significantly, and thus manageable for a more complex inference procedure [38, 119].

In the second stage we operate in the reduced state space of sampled part locations, hence it is feasible to employ more sophisticated factors in our model, including repulsive factors between symmetric parts in the same view, and non-Gaussian multi-modal pairwise terms. To perform inference in this case, we use max-product loopy belief propagation which allows to obtain a consistent pose estimate enforcing the structure of the typical human pose from the training set. The loopy belief propagation is not guaranteed to converge to the correct solution or indeed to converge at all [100]. However, in practice we achieve very reasonable results for many real world problems including human pose estimation as in our case.

2.4 Evaluation

In this section, we evaluate the performance of our *flexible* 2D pose estimation approach. First we examine the respective contribution of different components of our approach. After that, we compare the results of our approach to several state-of-the-art approaches on two different 2D human pose estimation benchmarks. We also discuss the advantages and disadvantages of our approach compared to the state-of-the-art with example illustrations.

Datasets. In order to validate our approach we use two different publicly available 2D pose estimation benchmarks. The training subsets of both datasets are completely disjoint from the

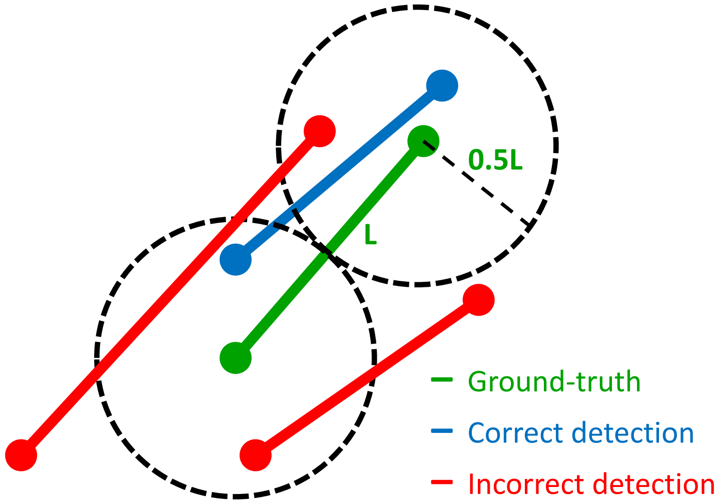


Figure 2.4: Example illustration of a correct part detection (blue) *vs* incorrect part detections (red) during PCP computation.

testing subsets. Such a setting quite challenging, since it requires to predict the poses for unseen human images, and therefore the approaches which lack generality, severely suffer in such settings.

First we evaluate our design choices and compare to the prior work on the pose challenge subset of our MPII-Cooking dataset [114]. The pose challenge subset of MPII-Cooking dataset includes 1071 images with 10 subjects for training (5 subjects are from separate recording) and 1277 images for testing with 7 different subjects. This dataset was originally recorded for the detection of fine-grained cooking activities in challenging kitchen settings. For MPII-Cooking dataset, we consider the task to estimate the upper body poses of the human subjects in the scene. We do not consider the detection of legs in this setting, since the legs are generally occluded due to the kitchen shelf. We follow [114] and use *person-centric* (PC) annotations for model evaluation, which means the right/left limbs in the image actually correspond to the anatomical right/left limbs of the human body.

Secondly, we also report the results on the ”Leads Sports Poses” (LSP) dataset, which exhibits strong variations in articulations and viewpoints. The images in this dataset are mainly collected from sports scenes. LSP dataset contains 1000 images for training and 1000 different images for testing the pose estimation model. Following Eichner and Ferrari [33], we use *observer-centric* (OC) annotations provided by the authors for evaluation. Observer-centric means the right/left of the person corresponds to the right/left position in the image regardless of actual anatomical right/left.

Performance Metric. To evaluate our approach in comparison to the state-of-the-art, we adopt the *percentage of correct parts* (PCP) measure proposed in [38]. PCP computes the percentage

Method	Torso	Head	upper arm r	 l	lower arm r	 l	All
Andriluka *et al.* [12]	80.1	80.0	67.8	69.6	48.9	49.6	66.0
+ *flex* (FPS [114])	78.5	79.4	61.9	64.1	62.4	61.0	67.9
+ *color*	80.6	84.7	64.6	66.2	**64.1**	62.5	70.4
+ *mixture PS*	**87.1**	**92.4**	62.7	67.3	61.4	62.9	72.3
+ *multi modal pairwise terms*	**87.1**	92.3	**68.4**	**70.1**	61.7	**63.2**	**73.8**

Table 2.1: Evaluation of model components. Improvements due to different components of our method over the baseline approach for 2D human pose estimation in percentage of correct parts (PCP) on **MPII-Cooking** dataset.

of correctly localized body parts by the pose estimation method. A body part is considered to be localized correctly if the predicted endpoints of the part are within half of the part length from their ground-truth positions. We also illustrate this in Fig. 2.4 for better understanding of the PCP measure. The PCP measure is pretty standard for evaluation of human pose estimation approaches. The main reason for its success and widespread adoption corresponds to the binary decision, if a part is correctly localized or not. Therefore, unlike other metrics (*e.g.* pixel error), the PCP measure is invaraint to the magnitude of error in the part localization. Hence every part contributes equally to the overall average PCP.

2.4.1 Evaluation of model components

In this section we evaluate our design choices on the pose challenge of the MPII-Cooking dataset presented in our earlier work on fine-grained activity detecion [114]. First we perform an in-depth analysis of our approach and report the improvements we get using different components of our approach. These results are demonstrated in Table 2.1. Our model consists of 10 upper body parts *i.e.* head, torso and 8 body joints (shoulders, elbows, wrists, and hands).

The improvements due to each of these model components are given in Table 2.1. We see our FPS model [114] with flexible part configuration achieves PCP of 67.9% as compared to 66.0% of the baseline model of Andriluka *et al.* [12]. We note that by modeling the local part appearance using joint shape and color features significantly improve the results for parts which are typically uncovered in natural settings *e.g.* head & hands, which in turn contributes to the overall pose estimation performance. For head it gives us improvement of 6% in PCP. Interestingly, color also helps to disambiguate the typical torso colors from the background leading to a gain of 3% in PCP. Overall, adding the color features in our model improves the performance to 70.4 PCP. Additionally we add mixture of PS models with 3 components and min-var component selection strategy, gaining an additional 1.9% to 72.3 PCP. Finally, our multi-modal pairwise terms, allow more flexibility to the pose estimation model making it more

Color descriptor	Torso	Head	upper arm r	 l	lower arm r	 l	All
FPS	78.5	79.4	61.9	64.1	62.4	61.0	67.9
+ RGB (5, 5, 5)	81.0	84.6	63.4	63.6	61.1	59.6	68.9
+ RGB (10, 10, 10)	80.9	83.7	64.0	64.8	63.4	61.1	69.6
+ RGB (20, 20, 20)	80.6	83.9	64.1	64.7	63.7	61.7	69.8
+ HSV (5, 5, 5)	81.2	84.0	63.3	63.7	62.1	60.3	69.1
+ HSV (10, 10, 10)	81.2	84.3	63.6	65.0	62.7	61.8	69.8
+ HSV (20, 20, 20)	**81.5**	83.9	63.8	65.5	62.7	62.3	70.0
+ HSV (25, 25, 25)	81.4	84.5	63.9	65.5	63.2	62.3	70.1
+ HSV (25, 25, 10)	80.6	**84.7**	**64.6**	**66.2**	**64.1**	**62.5**	**70.4**

Table 2.2: Evaluation of color descriptors. 2D pose estimation results (PCP) on the MPII-Cooking dataset for RGB and HSV color models with different bin sizes. The part detectors are always learned jointly with the shape-context descriptors as described in section 2.3.2. The baseline in this case is our flexible pictorial structures model (FPS).

expressive for challenging settings especially for the limbs. We get around 6/3% improvement for right/left upper arms respectively, achieving an overall PCP of 73.8%.

In the following, we one by one go through the different components of our model and investigate their contribution to the overall result.

2.4.1.1 Flexible pictorial structures (FPS)

This is the first and very basic component we employed to improve the flexibility of our model. Our model consists of 10 upper body parts *i.e.* head, torso and 8 body joints (shoulders, elbows, wrists, and hands) instead of the rigid limbs of original PS model [10]. We observe that having this *flexible* part configuration FPS (Sec. 2.3.1) already improves significantly for the lower arms, which are infact the most vulnerable parts of the human upper body. While overall the FPS model improves over baseline PS model [12] only by 1.9 PCP (66.0 PCP for PS models vs. 67.9 PCP for FPS), it improves the detection of lower arms by more than 11 PCP which are much more important for typical applications such as fined-grained activity recognition. This shows that using body joints as parts in the pictorial structures framework is able to cope better with the issue of foreshortening of body parts. Foreshortening has been pointed out as one of the main drawbacks in the approaches [10, 12, 96] that rely on the rigid limb templates.

2.4.1.2 Joint shape and color appearance

Here we would like to examine the performance of different settings for our color model. We aim to find out the reasonable bin sizes for different channels of RGB and HSV color spaces,

no. of clusters	selection	PCP
1	-	70.4
3	classifier	71.1
5	classifier	71.0
3	min-var	**72.3**
5	min-var	72.1

Table 2.3: Evaluation of 2D mixture PS with number of component clusters (1,3, and 5) and two different component selection strategies on **MPII-Cooking** dataset.

that would help in improving the pose estimation performance when combined together with the shape context features. In the following we discuss the promising results we obtain when we combine the color information with the shape-context feature descriptors and train the boosted part detectors using this joint representation. In Table 2.2, we list the pose estimation results (in PCP) using both RGB and HSV color histograms with different bin sizes. We choose our FPS [114] model as baseline for this evaluation.

We notice progressive improvement with increasing the number of bins for both RGB and HSV color models. However, we see that the improvement is more prominent in the HSV case. This result insists that the HSV color model is able to represent the image/scene better than the additive RGB model. Researchers have argued that since HSV and other similar color models uncouples the illumination or the brightness part of the color, therefore it allows to cope with the illumination variations better as compared to the additive color models such as RGB. Moreover, following the work of Lenz *et al*. [71], we reduce the number of bins for the V channel of HSV, in order to reduce the importance of illumination/brightness channel in the color model. Resultantly we see in Table 2.2 that our approach improves the PCP values considerably for almost all parts. For rest of the evaluations in this chapter we keep this setting of color model, *i.e*. HSV with bin sizes of (25,25,10) for its three channels. This leads to a 25x25x10 $= 6250$ dimensional color feature vector which we concatenate with the shape context feature descriptors to train the joint appearance model.

2.4.1.3 2D Mixture PS

In Sec. 2.4.1.3 we have discussed the contribution of the mixture of pictorial structures model for the 2D human pose estimation case. We have shown that a single pictorial structure model is a huge limitation in general settings with significant viewpoint variations and a mixture model definitely helps to overcome this shortcoming. We provide more insights about our model in this section. Using *mixture of pictorial structrue* components is an important milestone in the human pose estimation research. Our results reinforce the findings of [65]. We note that the mixture components significantly imrove the results for both torso and head and perform on-par for all other parts.

Following Johnson and Everingham [65], we also employ mixture of pictorial structures model to better represent different viewpoints of the human poses in the images. Johnson and Everingham [65] showed that they effectively captured different modes in the pose space.

We study two different strategies for the selection of the mixture component index in Sec. 2.3.4. We evaluate both strategies on MPII-Cooking dataset and compare their pose estimation results in Table 2.3. Interestingly, our approach which minimizes tha variance in the part marginals (min-var), performs considerably better that the component classifer. Since the min-var approach operates on the output of the PS mixture component, also the variance in the part marginals is strongly corelated with the pose estimation accuracy, as a result it predicts the index of the better performing mixture component. We believe with the increase in training data per component cluster, this gap in performance will reduce. On the other hand the advantage of the classifier approach is that we do not need to evaluate the model for each component instead the component is selected before the actual pose estimation. Then we predict the pose only using this selected selected mixture component.

In Table 2.3, we note that the classifier based approach only improve the overall PCP by 0.7% with 3 components and 0.6 with 5 components, on the MPII-Cooking pose challenge dataset. Although, the variance-based approach (min-var) works considerably better for 3/5 components. We get 1.9/1.7% improvement in overall PCP. But we observe that for both component selection strategies the performance degrades when we increase the number from 3 to 5 component clusters. This happens because Since the training and test sets for MPII-Cooking pose challenge contain pre-dominantly frontal poses. Increasing the number of clusters further results in some clusters with in sufficient training images.

Moreover, we show in Table 2.1, that the mixture of PS performs best for torso and the head parts. We believe this is because mixture of PS contribute in the recognition of correct viewpoints, hence the bigger and more stable parts perform significantly better. For more articulated parts the performance has also significantly improved.

2.4.1.4 Multi-modal pairwise terms

The *multi-modal pairwise terms* model the non-gaussian nature of the pairwise dependencies between parts. The structure The Gaussian assumption on the pairwise part dependencies limits the model to represent only certain kind poses in practice. As discussed in Sec. 2.3.3, we aim to drop this uni-modal Gaussian assumption and instead learn a mixture of Gaussian to model the pairwise relationships. Since we estimate the number of clusters or Gaussian modes K in our

Method	Torso	Head	upper arm r	 l	lower arm r	 l	All
Yang & Ramanan [157]	79.6	67.7	60.7	60.8	50.1	50.3	61.5
Andriluka *et al.* [10]	80.1	80.0	67.8	69.6	48.9	49.6	66.0
Rohrbach *et al.* [114]	78.5	79.4	61.9	64.1	**62.4**	61.0	67.9
Our (*Full Model*)	**87.1**	**92.3**	**68.4**	**70.1**	61.7	**63.2**	**73.8**

Table 2.4: Comparison of 2D upper body pose estimation methods on **MPII-Cooking** dataset in percentage of correct parts (PCP).

pairwise mixture model by minimizing the varaince in the data, therefore we do not consider the analysis of the number of clusters K on the performance. The estimated number of

Finally, our *multi-modal pairwise terms* allow more flexibility to the pose estimation model making it more expressive for challenging settings especially for the limbs. We get more than 6% improvement for both upper arms.

Moreover, we show in Table 2.1, we notice that although this mixture model at the pairwise level does not help in improving the recognition quality for the bigger parts like torso and head, but for the smaller and articulatable parts (upper and lower arms) it improves significantly. Finally we obtain a gain of 1.5%. For upper arms we get 5.7/2.8% for the upper arms (right/left) and for the lower arms it as well improves a bit. Overall, we achieve 73.8% PCP upper-body 2D pose estimation on MPII-Cooking dataset using our complete EPS model. In the next chapter we will discuss the use of this model as foundation for multi-view 3D human pose estimation.

2.4.2 Comparison to state-of-the-art

In this section we compare the results of our pose estimation method to the state-of-the-art for 2D human pose estimation.

MPII-Cooking dataset. First we compare the pose estimation results of our final model (EPS) to the state-of-the-art on MPII-Cooking dataset. Our model performs significantly better for all parts and overall we achieve improvement of around 8% PCP over the baseline approach of Andriluka *et al.* [12]. The flexible mixture of parts (FMP) by Yang and Ramanan [157] is quite efficient and powerful part detector and it has been found that its performance improves significantly with the increase in training data. In our evaluation, this FMP model fails to perform well since we operate in the limited data domain with 1071 training images with only 5 subjects and a completely disjoint test set with 1277 images with 7 different subjects. Overall FMP achieves 61.5 PCP which is considerably low compared to other methods. We attribute this lower performance of FMP model to the fact that it has significantly larger number of parameters to estimate during training and such a model is easier to overfit in the absence of sufficient amount of data.

Method	Torso	Head	upper leg	lower leg	upper arm	lower arm	All
Andriluka *et al.* [10]	80.9	74.9	67.1	60.7	46.5	26.4	55.7
Yang & Ramanan [157]	84.1	77.1	69.5	65.6	52.5	35.9	60.8
Eichner & Ferrari [33]	86.5	80.1	74.9	69.3	56.5	37.4	64.3
Pishchulin *et al.* [94]	87.5	78.1	75.7	75.7	54.2	33.9	62.9
Pishchulin *et al.* [95]	**88.7**	**85.6**	**78.8**	73.4	**61.5**	**44.9**	**69.2**
Our (*Full Model*)	85.4	76.7	78.6	**75.8**	50.4	37.5	64.7

Table 2.5: Comparison of 2D full body pose estimation methods using observer-centric (OC) annotations on **LSP** dataset in percentage of correct parts (PCP).

In Figure 2.6, we notice that our approach is much more flexible and expressive in representing the challenging articulations in the human poses compared to the baseline [12]. We also give the number of correctly estimated parts under the corresponding images, for the baseline approach [12] and our complete 2D model *i.e.* EPS.

Above we have shown and discussed the advantages of our method on challenging in-door settings for the task of upper-body pose estimation, but it is imperative to demonstrate the performance of our approach in more generic out-door scenes and compare to the recent state-of-the-art in such setting. This will help establish our argument about effectiveness and generality of our pose estimation approach.

LSP dataset. LSP dataset is basically one of the standard pose estimation benchmarks. Numerous researchers [10, 33, 65, 66, 94, 95, 157] have evaluated their model on this dataset and have achieved significant improvement over the years. The task is to estimate the full body pose given the images of the humans already localized in the images. Following the state-of-the-art, we evaluate using the observer-centric (OC) annotations. Our model consists of 14 parts *i.e.* head, torso and 12 body joints (wrists/hands, elbows, shoulders, hips, knees, and ankles/feet). The comparison of our model to the approaches from the recent literature is given in Table 2.5. We notice that the overall PCP value of our approach is significantly better than most of the current methods *i.e.* 64.7%. Our approach does not beat the state-of-the-art work from Pishchulin *et al.* [95] for the overall PCP of 69.2%. However, interestingly for the legs (upper and lower) our model significantly improves and performs on par with this best performing approach [95]. We achieve 78.6/75.8 PCP for upper/lower legs respectively. For all other parts, the model of Pishchulin *et al.* [95] works significantly better mainly due to their use of much stronger part detectors. To obtain these strong part detectors they train a separate DPM [35] for each body part with its own deformable latent parts structure, which makes the model extremely complex to train and evaluate at test time.

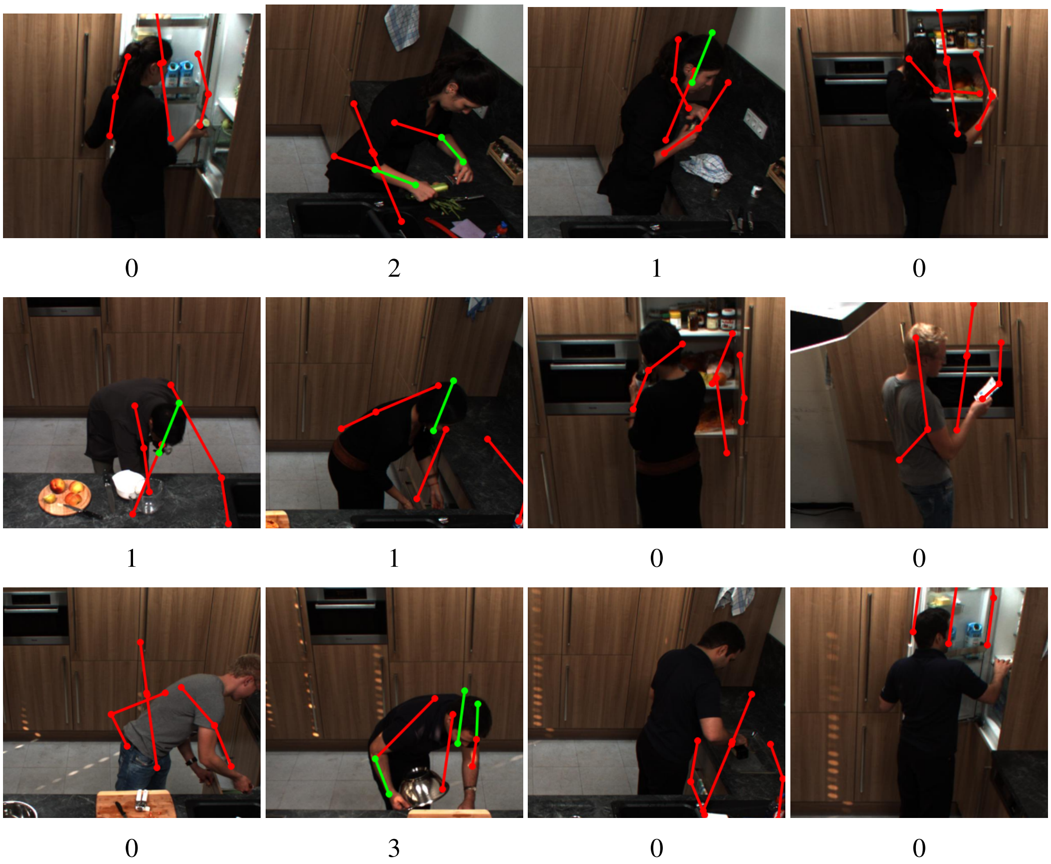


Figure 2.5: Failure Cases. Example images from **MPII-Cooking** dataset, for which our 2D pose estimation model (EPS) fails to predict the correct pose. The numbers under images show the number of correct parts.

2.4.3 Failure cases

MPII-Cooking dataset. We illustrate some of the failure cases on our MPII-Cooking dataset in Fig. 2.5. The major cases for which our pose estimation approach fails include, self-occlusion, non-typical poses for which there does not exist enough examples in the training set, low contrast due to challenging lighting conditions and the appearance of the cup-board, and severe illumination variations across different subjects.

LSP dataset. In Fig. 2.8, we demonstrate some of the images on which our 2D pose estimation approach completely fails for this LSP dataset. We observe that main reasons for failure are self-occlusions and multiple people in the images. The presence of multiple close by people makes it difficult for the pose estimation method to parse a pose containing parts from the same human in the images.

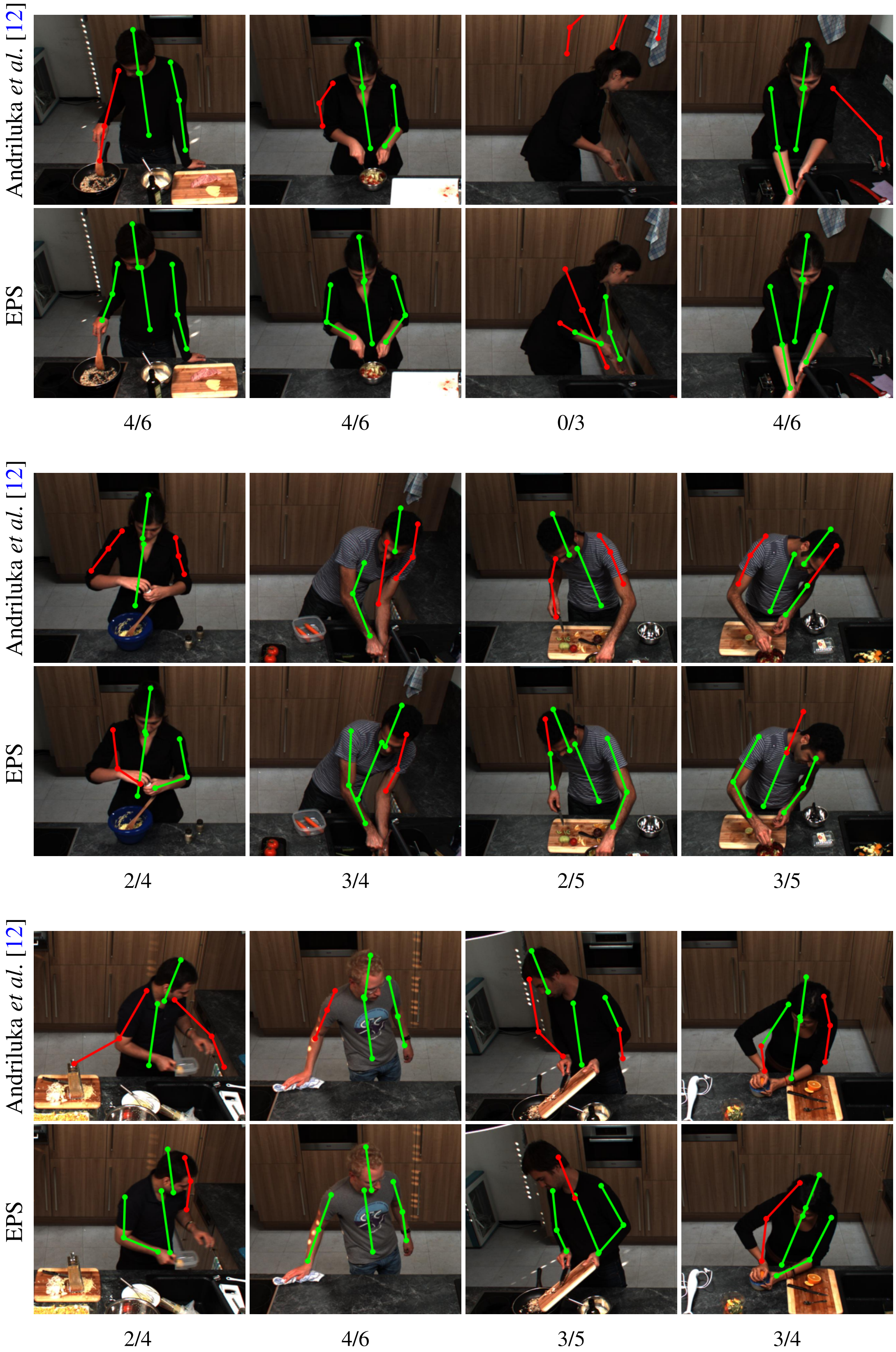


Figure 2.6: Pose estimation qualitative comparison. Example images from **MPII-Cooking** dataset, with poses estimated by Andriluka *et al*. [12] *versus* our complete 2D model (EPS). The numbers under images show the number of correct parts by [12]/EPS.

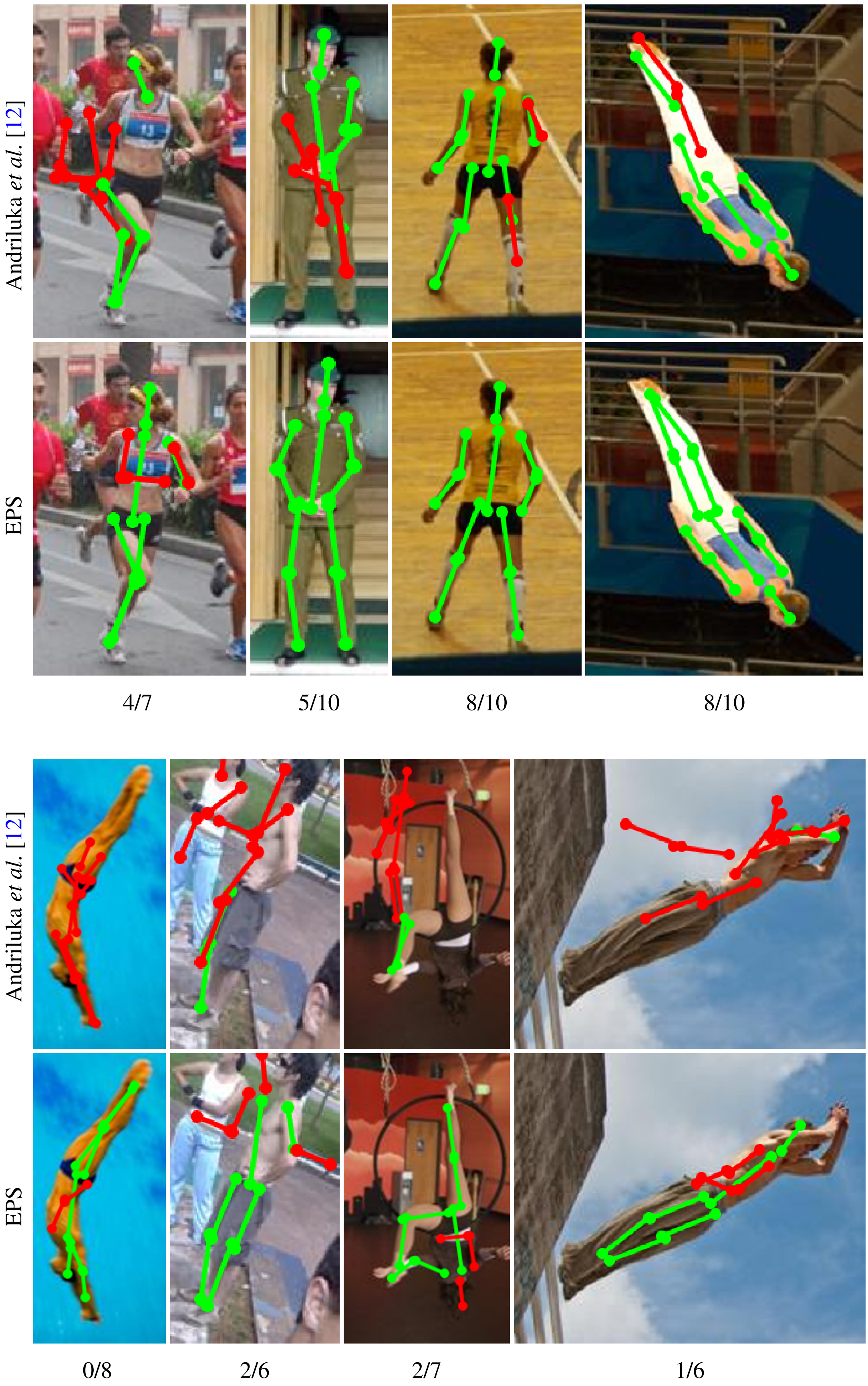


Figure 2.7: Pose estimation qualitative comparison. Example images from **LSP** dataset, with poses estimated by Andriluka *et al.* [12] *versus* our complete 2D model (EPS). The numbers under images show the number of correct parts by [12]/EPS.

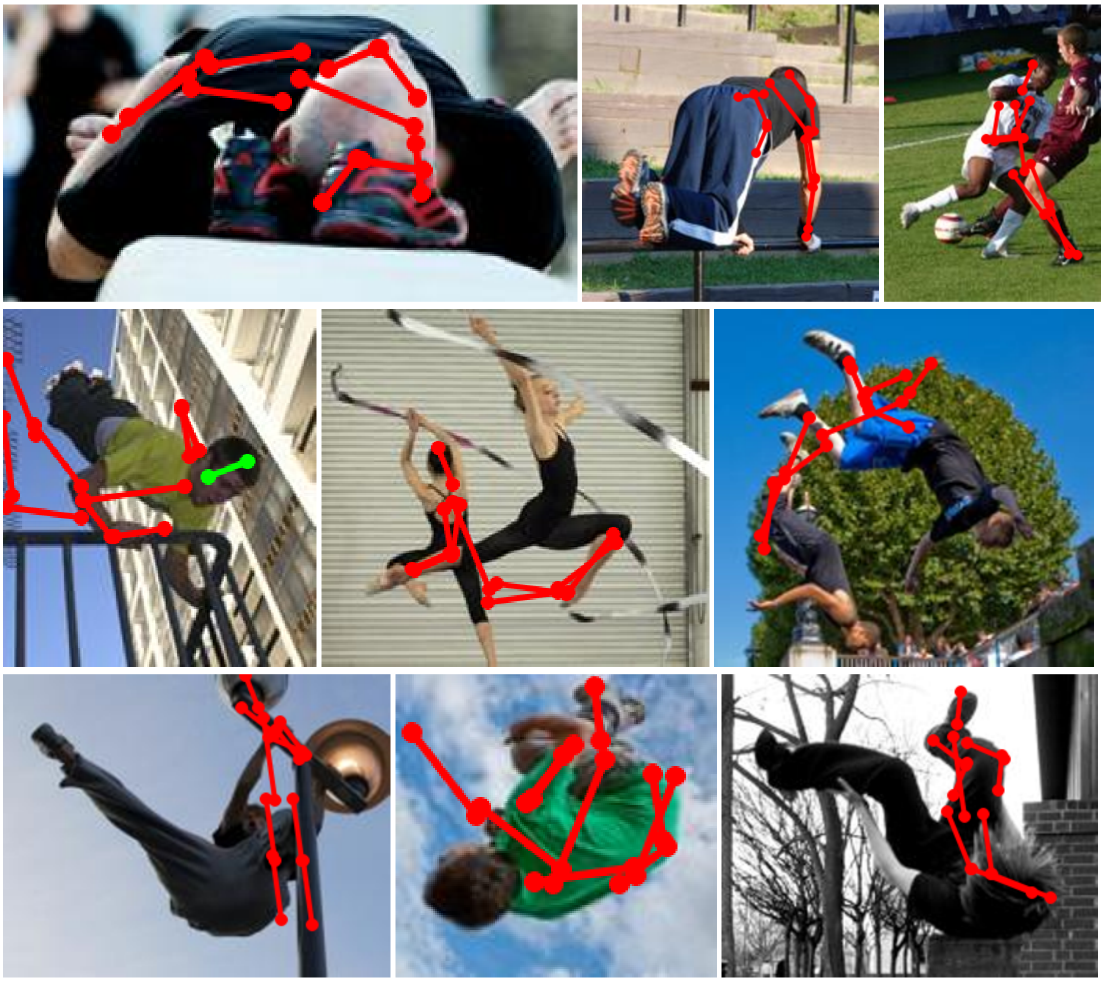


Figure 2.8: Failure Cases. Example images from **LSP** dataset, for which our 2D pose estimation model (EPS) fails to predict the correct pose.

2.5 Conclusion

Pictrorial structures approach have been shown to work well for the task of 2D human pose estimation in challenging real-word settings, and by design it performs significantly better than other approaches in limited data domain. Although the typical pictorial structure models significantly improve the pose estimation performance, they still fail on non-typical poses due to their limited expressive power. By extending the original model proposed in [12] with flexible parts, color features, multi-modal pairwise terms, and mixture of pictorial structures, our 2D pose estimation approach significantly improves over the baseline and most state-of-the-art methods on both MPII Cooking and LSP datasets. Both qualitative and quantitative evaluations corroborate the expressiveness of our model.

Chapter 3

Multi-view Pictorial Structures for 3D Human Pose Estimation

Pictorial structures have been used extensively for 2D human pose estimation in the literature. Over the years numerous improvements have been proposed to expand their range and enable them for a vast variety of realistic scenarios. The advancements range from richer appearance models, flexible body priors, and different level of mixture models to more accurate and efficient inference procedures. While these improvements does help pictorial structures based methods to achieve state-of-the-art results for the task of 2D pose estimation, interestingly they have not yet been extended to enable pose estimation in 3D.

Hence, in this work we propose a multi-view pictorial structures approach that incorporates appearance and geometry based muti-view evidence across viewpoints for an efficient and robust 3D human pose estimation. Moreover, our multi-view model also takes into account the recent advances in 2D pose estimation. We demonstrate that our proposed multi-view pictorial structures model achieves on par and/or significantly better results while We evaluate our multi-view pictorial structures approach on standard benchmark datasets, *i.e.* HumanEva-I, as well as novel datasets with realistic dynamic settings, *i.e.* MPII Cooking, and Shelf. Unlike state-of-the-art in 3D pose estimation our approach operates on single-frames only, and does not rely on activity specific motion models or tracking. However, our model still achieves similar or better results especially for the activity sequences with complex motions, while keeping the computational cost in check.

3.1 Introduction

In this chapter we address the task of articulated 3D human pose estimation using multiple cameras. The cameras are assumed to be calibrated w.r.t. a common global coordinate system.

In the literature, this problem is generally addressed in the 3D space employing complex 3D body models [23, 31, 42, 128] and sophisticated activity-specific motion models [137, 160]. In the end this leads to complicated inference procedures in the high-dimensional space of 3D pose configurations. In the literature one can find various efforts in order to reduce the search complexity of these methods, such as annealed particle filtering by Deutscher and Reid [31] or non-parametric belief propagation by Sigal *et al.* [128]. However, these approximation techniques degrade the pose estimation accuracy, therefore these modes have only been shown to work in very controlled laboratory settings.

Thus in this work we discuss that the inference complexity can be significantly reduced while achieving state-of-the-art results, by formulating the 3D pose estimation problem as a joint estimation of 2D poses in all camera views. This effectively reduces the search space, but requires additional modeling steps to represent the 2D projections and take foreshortening of limbs into account. Our proposed multi-view potential function ensure that the 2D poses are 3D consistent, hence can be directly triangulated to reconstruct the 3D pose. Instead of hypothesizing about 3D poses, relaxing the task to 2D pose estimation allows us to benefit from the recent advancements in articulated 2D pose estimation. In Chapter 2, we proposed several improvements and refinements for the 2D pictorial structures model. We named those advancements as Expressive Pictorial Structures (EPS) in section 2.3, to signify/suggest/indicate its expressive power in a range of realistic settings. Building on our success with EPS, we extend that model for joint inference of 2D pose projections in all views. The advantage of 2D inference is that our approach can densly search in all image locations for 2D pose configurations in all views while keeping the computation reasonably low.

Just like in the 2D case, we employ discriminatively trained representations for part appearances based on color and shape-context feature descriptors. We show that such representation is robust against backgound clutter and quite capable to deal with variance in appearance of human body parts, hence applicable to real world images. Thus our model is not person specific as various other approaches in the literature, and does not rely on the specific activity or the motion of the person. Moreover, our model The state-or-the-art in the literature such as [42, 137, 160], employ stochastic search and approximate inference procedures. Moreover, such methods depend on tracking and hence require initialization to preform. However, our approach is based on single-frames, and do not require any form of temporal filtering.

Our contributions. In this chapter, we present the extension of our 2D pose estimation model EPS (see Sec. 2.3) to multiple views. As a first contribution of this work, we propose a novel multi-view pictorial structures model which allows to integrate the multi-view constraints directly in the pictorial structures framework, hence performs joint reasoning over people poses

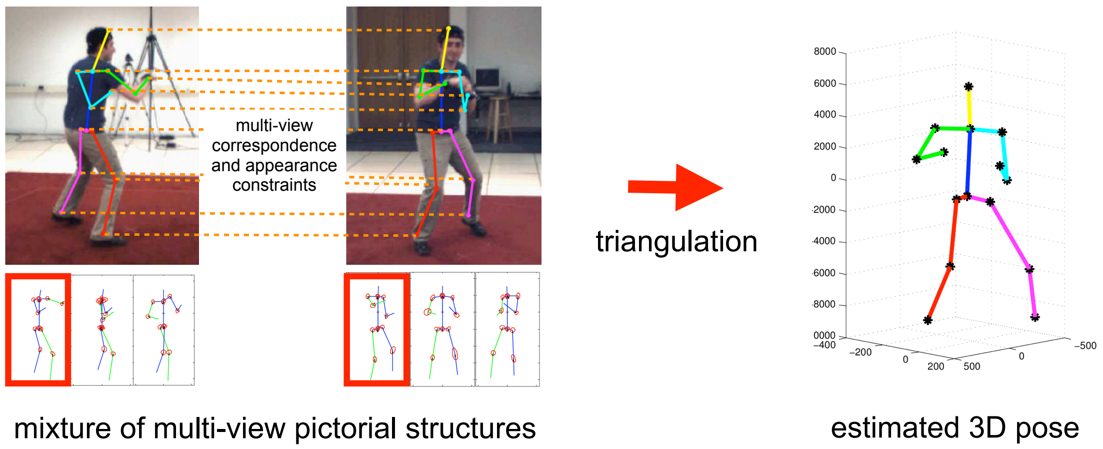


Figure 3.1: Our approach. (1) Projections of 3D pose in each view are jointly inferred using a mixture of multi-view pictorial structures models. The body layout priors of each mixture component are visualized below, activated components are highlighted in red. (2) 3D pose is recovered via triangulation.

seen from multiple viewpoints. The proposed multi-view constraints include multi-view appearance and geometric correspondence for all pairs of camera views. We study effectiveness of both constraints in detail. The second contribution is to propose a novel 3D mixture of pictorial structure models to enable inference of consistent part marginals in all views. This is the extension of 2D mixture model we discussed in chapter 2. As the final contribution, we benchmark our proposed approach on two datasets of quite different complexities. First we compare the 3D pose estimation performance of our model on the standard HumanEva-I [127] dataset. We show that our approach which does not depend on the activitiy-specific motion models and any kind of temporal filtering, achieves results on par or better than the state-of-the-art [137, 160]. Especially, our approach outperforms all approaches from the literature on the image sequences with multiple / complex motions. Then to demonstrate the ability of our proposed method to perform well in realistic scenarios, we also evalute on MPII Cooking [114] dataset which has an order of magnitude more human subjects, significantly larger number of activities, and dynamic backgrounds due to clutter in the scene. We show that by jointly estimating poses across multiple views, we can achieve significantly better results also in such challenging settings.

3.2 Related work

Articulated 3D human pose estimation is one of the most important challenges in computer vision. The search for state-of-the-art in 3D human pose estimation returns quite a diverse literature. However, we find that this task is mainly studied for markerless motion capture in controlled laboratory settings such as [127]. Modern approaches in this domain rely on complex body models in 3D [16, 42], and intricate inference and optimization strategies, such as stochastic search by [42, 128], local optimization by [135], or a combination of both by [43]. Although a number of methods [44, 135, 146] demonstrate quite impressive results in such settings, they seem to struggle to generalize beyond the motion capture domain.

In contrast to the laboratory setting, real-world images often feature dynamically changing backgrounds, multiple human subjects, occlusions of people by scene elements and interactions with scene objects. Recently, numerous techniques have been proposed to perform well in increasingly realistic real-world settings, such as methods based on structure-from-motion methods [55], by combining visual observations with on-body inertial sensors [99], or by utlizing activity specific pose and motion priors [44]. These approaches build on techniques developed for controlled laboratory environments and extend them to more realistic settings. In this work, in contrast to recent efforts of [44, 55, 99] who aim at adapting markerless motion capture methods to a more complex settings, we rely on our expressive pictorial structures model 2.3. In chapter 2, we demonstrated that such a model works quite well for images of realistic complexity. We extend this model to incorporate across-view constraints to enable articulated 3D human pose estimation in challenging scenarios. Our proposed approach is related to the works of [11, 129] due to the use of pictorial structures for 3D pose estimation, however these approaches do not account for the multi-view evidence and hence heavily rely on activity-specific motion models. To be more specific [11] is limited to the pose estimation of walking people only. And the approach of [129] employs only a simple observation model based on silhouette and color features, which do not perform well in dynamic scenes.

Arguably the most related approach to ours, is the 3D pictorial structures approach of Burenius *et al*. [26], who formulated pictorial structures for the 3D domain. Our approach offers several advantages over [26], including significantly more economic 2D representation due to joint 2D inference in all views, stronger part detectors, and effective body models inspired by the state-of-the-art in 2D pose estimation. These qualities make our approach applicable to both standard laboratory settings and more realistic settings such as MPII Cooking and Shelf datasets that include a variety of human subjects and dynamic scenes. In our evaluations we show that we achieve state-of-the-art results in all these challenging settings, that none of the other approaches has demonstrated.

In the last chapter (section 2.3), we discussed our appraoch to 2D pose estimation that relies on the pictorial structures formulation of Andriluka *et al*. [12]. We proposed several improvements motivated by the related work on pictorial structure models to deal with its different shortcomings and boost performance. Below we discuss the extension of our single-view model to incorporate evidence from multiple views for 3D pose estimation.

3.3 Multi-view pictorial structures model

Our objective is to estimate the 3D pose of the person given image observations obtained from multiple viewpoints. In order to estimate the 3D pose we proceed in two steps. In the first step we employ multi-view constraints and jointly estimate the 2D projections of the 3D body joints in each view. In the second step, we use the estimated 2D projections and recover the 3D pose by triangulation [54].

As a basic tool for the representation and inference of the 2D human pose in each view, we rely on the 2D model introduced in the last chapter (Sec. 2.3). We extend our single-view expressive pictorial structures formulation to incorporate evidence from multiple views. To this end, we employ different multi-view constraints including 3D mixture components which allows for inference of consistent posterior distributions in all views. These multi-view constraints introduce further cycles in our graphical model and increase the complexity of the model. However, to uphold our claim of reduction in model complexity in comparison to the prior work on 3D pose estimation, we restrict the multi-view constraints to pairwise dependencies across views. Also, our approach avoids the reasoning in the high-dimensional space of 3D poses as the inference is performed in 2D, jointly across views.

Let us represent the part configurations in all views jointly as a set $L = \{L_1, L_2, \dots, L_V\}$, where V is the total number of camera views. The individual part configuration L_v for any given view v is represented using the same formulation as described in the last chapter for the single view case *i.e.* $L_v = \{l_1^v, \dots, l_N^v\}$. The pose of each part n in view v is parametrized by $l_n^v = (x_n^v, y_n^v, \theta_n^v)$, where (x_n^v, y_n^v) is the image position of the part, and θ_n^v is the absolute orientation. We formulate our multi-view model as a conditional random field, and assume that the probability of the joint part configurations L given the image evidence of all views $I = \{I_1, I_2, \dots, I_V\}$, factorizes into a product of single-view and multi-view pairwise factors.

$$p(L_1, \dots, L_V | I_1, \dots, I_V) = \frac{1}{Z} \prod_v f(L_v; I_v) \cdot \prod_n \prod_{(a,b)} f_n^{cor}(l_n^a, l_n^b) f_n^{app}(l_n^a, l_n^b; I_a, I_b), \qquad (3.1)$$

where $\{(a, b)\}$ is the set of all view-pairs, $f(L_v; I_v)$ are the single-view factors for view v which decompose into products of unary and pairwise terms according to Eq. 2.2, and f_n^{cor} and f_n^{app} are multiview correspondence and appearance factors respectively for part n. These multi-view correspondence and appearance are complementary constraints, since the former requires camera geometry to work and the latter uses the image evidence instead. Both multi-view factors are given for all possible camera viewpairs. We do not consider muli-view constraints for triplets of camera views or other higher order terms as it substantially increases the complexity of our model for pose estimation. For example, if we have $V = 3$ camera views and $M = 1000$ part hypotheses per view, then we need to compute multi-view factor scores 1 Billion times (M^3) if we consider triplets of part hypotheses. Now consider pairwise terms only then this value comes down to only 3 Million ($3M^2$), since we get 3 viewpairs. An added advantage of modelling only the pairwise terms is that the uncertainity due to an occluded part in one view will not have significant effect during the inference as long as the part was captured by atleast two camera views.

The evidence of the same scene/object from multiple viewpoints gives a new perspective about that object, hence enables us to effectively discriminate the correct hypotheses from the incorrect ones. In the following we define the multi-view pairwise factors that encode correspondence and appearance constraints in our model.

3.3.1 Multi-view correspondence

The term multi-view correspondence refers to the constraint that the location of part hypothesis in one view is restricted by the location of the part hypothesis in another view. In our approach, we use the reprojection error to model this correspondence of part hypotheses for each pair of camera views.

Reprojection error measures the cost of estimating a real world 3D point from a correspondence of 2D image measurements. We use the originally estimated correspondence of part hypotheses $\mathbf{l}_n^a \leftrightarrow \mathbf{l}_n^b$ to reconstruct the position X_n of the part in 3D using linear triangulation [54]. Note that for the sake of clarity we use the term $\mathbf{l}_n^v = (x_n^v, y_n^v)$ to represent the position of a part hypothesis in view v instead of the term l_n^v which in addition to image position also includes the absolute orientation. This is because the reprojection error only requires the position of the part hypothesis in each view. The reprojection error for hypotheses of two views a and b is given by,

$$d_{rpe}(\mathbf{l}_n^a, \mathbf{l}_n^b) = ||\mathbf{l}_n^a - \hat{\mathbf{l}}_n^a||^2 + ||\mathbf{l}_n^b - \hat{\mathbf{l}}_n^b||^2 \quad s.t. \quad \hat{\mathbf{l}}_n^a = P_a X_n \;\&\; \hat{\mathbf{l}}_n^b = P_b X_n \tag{3.2}$$

Here, $\hat{\mathbf{l}}_n^a$ and $\hat{\mathbf{l}}_n^b$ are the reprojections of the 3D reconstructed point X_n, hence they represent the perfectly matched correspondences of part hypotheses in both views. P_a and P_b are the given projection matrices for both camera views.

The reprojection error, illustrated in Fig. 3.2, occurs because the original correspondence of part hypotheses $\mathbf{l}_n^a \leftrightarrow \mathbf{l}_n^b$ is often imprecise or wrong due to inaccuracies or failures in part detection process. Moreover, the angle between the backprojected rays of the correspondence $\mathbf{l}_n^a \leftrightarrow \mathbf{l}_n^b$ also determines the accuracy of the 3D reconstruction of the real world point X_n. Localization of 3D points becomes less precise as rays become more parallel [54]. This happens often for

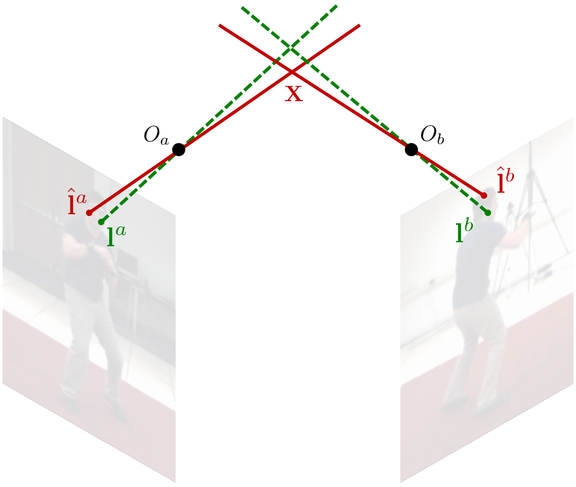


Figure 3.2: Illustration of the reprojection error. $\mathbf{l}^a \leftrightarrow \mathbf{l}^b$ shows the original noisy hypotheses correspondence across two views, and the dashed line indicates that the rays often do not intersect due to this noisy correspondence. X represents the actual 3D reconstruction as a result of the noisy correspondence and the reconstruction ambiguity. Finally, $\hat{\mathbf{l}}^a \leftrightarrow \hat{\mathbf{l}}^b$ is the exact reprojected hypotheses correspondence in two views.

too close cameras or opposite facing cameras in the scene. Therefore, using reprojection error to model our multi-view correspondence term allows us to penalize both, inaccurate point correspondences and non-optimal pair of camera views.

We assume the inaccuracies of the point correspondences in the images follow a Gaussian distribution, with zero mean and variance σ_c^2, around the ”true” points. This gives the probabilistic meaning to the reprojection error. Also, the variance σ_c^2 allows us to control the contrast between good and bad correspondences. The multi-view correspondence factor f_n^{cor} for any part hypothesis correspondence $\mathbf{l}_n^a \leftrightarrow \mathbf{l}_n^b$ is given by,

$$f_n^{cor}(l_n^a, l_n^b) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{d_{rpe}(\mathbf{l}_n^a, \mathbf{l}_n^b)}{2\sigma_c^2}\right), \tag{3.3}$$

If the relative translation and rotation of the two cameras in a viewpair is known, the corresponding epipolar geometry provides the constraints between the image points of both views. More specifically, a point in one view corresponds to the epipolar line in the other view. This means, although it is possible to constraint the location of corresponding points in two views using camera geometry, but this correspondence is not one-to-one. Therefore, we either require a third camera view for correction or some additional multi-view terms that could further constraint the location of the point correspondences in both views. Since we want to keep the correspondence constraints as pairwise terms for the practical/complexity reasons discussed earlier, therefore we instead introduce our multi-view appearance term which does not require any information about camera geometry, instead it exploits the multi-view image evidence.

3.3.2 Multi-view appearance

The projection of the objects on the 2D image plane results in the loss of 3D information and severely affects the recognition accuracy. Therefore, exploiting the multi-view image evidence we aim to improve the classification capability of our part detectors. To this end, we introduce the multi-view appearance term for all viewpairs. The image appearance of typical objects from different viewpoints is correlated, especially in the adjacent camera views. Hence, we look to exploit this appearance correspondence to improve our part detectors. Intuitively, this sounds like a very effective cue to discriminate the appearance of the desired objects from other objects and the background, however, it is not clear how this across view appearance dependency can be modeled. In our approach, we train this across view appearance dependency and it is represented by the factor f_n^{app} in our multi-view pictorial structures model (see Eq. 3.1) for all viewpairs. This pairwise factor encodes the color and shape of the body part seen from multiple viewpoints. We define the joint appearance feature vector by concatenating the features from both camera views in a viewpair $e_n(l_n^a, l_n^b) = [e_n(l_n^a), e_n(l_n^b)]$ and train a boosted part detector using this representation. Similar to the single-view case we rely on the Adaboost feature selection procedure of Viola and Jones [149] to choose the most discriminative subset of features. Note that in contrast to the single-view boosted part detectors the multi-view detector has access to features in all views during training and can exploit co-occurrence of features across views to learn a more discriminative detector. Following the representation of the single-view likelihood term (Eq. 2.3), we define our multi-view appearance factor score as,

$$f_n^{app}(l_n^a, l_n^b; I_a, I_b) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}([e_n(l_n^a), e_n(l_n^b)])}{\sum_t \alpha_{n,t}}, \varepsilon_0\right) \tag{3.4}$$

Here the presence of image evidence I_a and I_b shows that this term is image dependent.

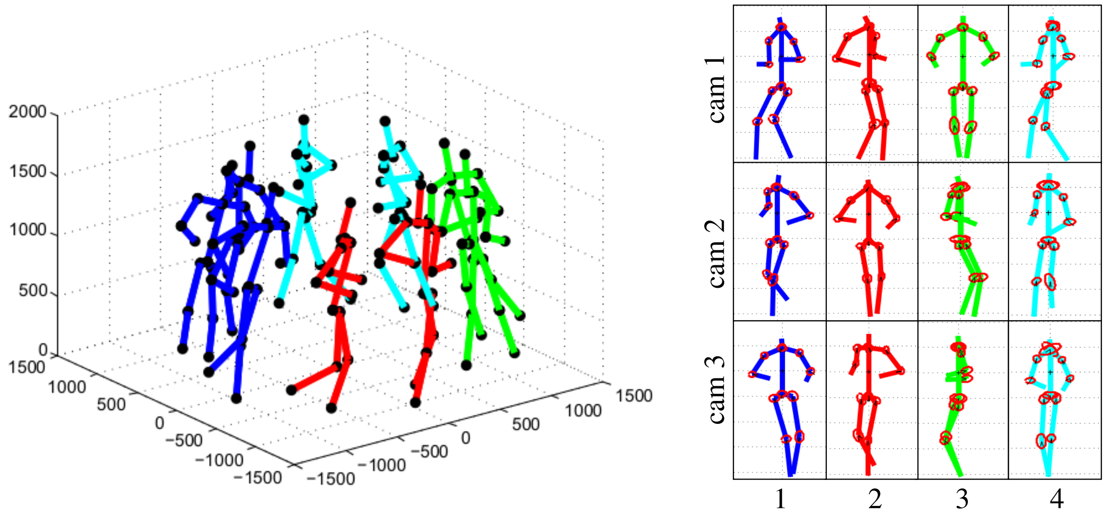
The positive examples to train the Adaboost classifier are obtained from the ground-truth annotations of the parts in corresponding views in a viewpair. On the other hand, we consider two types of negative examples for training the classifer. In the first type, we take the random examples in the negative images of both views in a viewpair. Note that the image region outside the person bounding box in a positive image is considered as the negative training image. The second type of negative training examples are the concatenations of positive part examples from the first view and negative example from the second view and vice versa. It is important to mention here that we include this second type of negative examples for training the classifier because the objective is to keep this multi-view appearance term to detect the appearance-based part correspondences across views. The single-view likelihood terms are supposed to take care of the cases when the part is occluded in one view and visible in the other.

Although, our multi-view appearance terms increase the discrimitative power of our part detectors but constraints the applicability to a static camera setup. If we intend to estimate the poses for relatively more dynamic settings compared to the setting at the training time, then we can simply omit this term from the proposed multi-view pictorial structures model. However, the multi-view correspondence term (Sec. 3.3.1) would still be valid and equally effective.

3.3.3 3D Mixture PS

Our multi-view model also employs mixtures of pictorial structures to represent 2D body configurations per view. In Chapter 2 we have discussed the advantage of employing mixture of pictorial structures for 2D human pose estimation. Keeping that in mind, here we aim to investigate the potential and applicability of mixture components in a multi-view setting. However in the multi-view setting, directly employing the 2D mixture model individually for each view is clearly a huge model limitation. Moreover, due to projective ambiguity often visibly different viewpoints of a person get clustered in the same group in the 2D case. This leads to training of sub-optimal mixture components for the 2D mixture PS (Sec.2.3.4). In contrast, in our 3D mixture PS, the mixture components correspond to groups of poses similar in 3D. Therefore, we notice better clustering in this case. In order to obtain such 3D mixture components we first reconstruct the 3D training poses from the 2D pose annotations in images and then cluster these 3D poses with k-means, illustrated in Fig. 3.3a. We then reproject the training data of each 3D cluster in all views and learn the 2D mixture models from the reprojected data. It is important to note that the part correspondences we get directly from 2D pose annotations are generally noisy. But the reprojection of reconstructed 3D poses leads to exact part correspondences in all views and therefore this also helps during training of the multi-view appearance terms, discussed in Sec. 3.3.2.

In Fig. 3.3, we visualize an example of our proposed 3D mixture PS learned on the Jogging sequence of HumanEva-I dataset. Fig. 3.3a illustrates the 3D pose clusters and Fig. 3.3b shows their corresponding 2D body configurations in all views. Note that the resulting 2D mixture components are also consistent across views by design, as they are learned from the projections of the same cluster of 3D poses. Therefore, at the test time this consistency of mixture components leads to consistent posterior distributions in all views. Moreover, we exploit this fact by jointly selecting the best mixture component across all views and adapt the component selection procedure introduced in Sec. 2.3.4 accordingly. For the classifier-based component selection strategy, we simply add the scores of the corresponding components across all views and select the one with the higher score. For the variance-based selection strategy, we add the uncertainty



(a) 3D clusters

(b) 2D body models

Figure 3.3: 3D Mixture PS. (a) 3D poses are reconstructed from 2D annotations and then clustered. (b) 3D poses from each cluster/component are then reprojected in all views, and we learn a separate PS model for each component. The numbers at the bottom indicate the component indices.

scores $s(k, I)$ defined in Eq. 2.8 for any component k, across all views V. Hence, $s(k, I)$ for our 3D mixture PS is given by,

$$s(k, I) = \sum_{v=1}^{V} \sum_{n=1}^{N} \|\text{Cov}_n(k, I)\|_2, \tag{3.5}$$

In our evaluations in Sec. 3.4.1.3, we show that the pose estimation performance using our proposed 3D mixture of pictorial structures is superior to the 2D mixture dicussed in the last chapter. Hence, it is one of the important contributions of our approach. However, similar to the multi-view appearance term the mixture of pictorial structure models is also applicable to the static camera setup, in general.

3.3.4 Inference

To perform inference in our multi-view case jointly for all camera views, we rely on the same procedure discussed in the last chapter (see Sec. 2.3.5). In addition to the repulsive factors, our multi-view factors (Eq. 3.1) introduce further loops in the graphical model, making the inference even more challenging. Therefore, in this case we again follow the approximate two-stage inference procedure introduced in [12]. In the first stage, we employ a simplified tree-structured model with Gaussian pairwise factors (cf. Eq. 2.4) to estimate the dense marginal distributions $p(l_n|I)$ for all body parts. Then, we generate the body part hypotheses by sampling

[1] the proposals from these densely estimated part marginals $p(l_n|I)$. Note that these simplified tree-structured models for each component in our multi-view case are consistent across views, since they are learned from the 2D projections of the same 3D pose clusters (see Fig. 3.3). Thus, although these models are employed independently per view, they generate geometrically consistent marginal distributions $p(l_n|I)$ each body part across in all views. As, explained in section 2.3.5, this first stage basically acts as a search-space reduction step. A reduced state-space allows us to apply models with more complex pairwise constraints between body parts.

In the second stage since we operate in the reduced state space of sampled part locations, it is feasible to employ more sophisticated factors in our model, including repulsive factors between symmetric parts in the same view, multi-modal pairwise terms, and multi-view appearance and correspondence factors between parts in all view-pairs. To perform inference in this case, we use max-product loopy belief propagation which allows to obtain a consistent estimate for the human pose configuration in all views. The loopy belief propagation is not guaranteed to converge to the correct solution or indeed to converge at all [100]. However, in practice we achieve very reasonable results for many real world problems including human pose estimation as in our case. Finally, given the 2D projections estimated by the multi-view pictorial structures model we reconstruct the 3D pose using triangulation.

3.4 Evaluation

We evaluate our approach on two datasets, HumanEva-I [127] which is a standard benchmark for 3D pose estimation in the laboratory settings, and on the more challenging MPII Cooking pose challenge dataset [114] that was recorded for the task of fine-grained activity recognition and features a larger number of subjects and interactions with objects.

HumanEva-I. Following [160], we use the three color cameras recorded in HumanEva-I. We use the provided evaluation scripts and report 3D error in millimetres. We compute the 3D poses with linear triangulation [54] by using 2D pose estimations from all 3 cameras. For the walking and box sequences we evaluate on the validation set, as [137, 160], and for *combo* sequence on the test set, as [137]. In order to compensate for the slight differences in positioning of joints in our model and in HumanEva-I we add a fixed offset to each joint. In order to estimate this offset we first manually fit our model to several training images and then compute the mean offset between our poses and the HumanEva-I ground-truth.

[1] In all experiments we sample the marginal distributions 1,000 times for each body part and remove the duplicates. We found this number to be sufficiently large in our experiments.

Setting	walking	box
FPS [114]	114.3	77.4
EPS (Sec. 2.3)	82.5	72.9
+ multi-view correspondence	73.9	62.8
+ multi-view appearance	75.7	66.2
+ 3D mixture PS	74.2	69.1
Our full model (Multi-view PS)	**54.5**	**47.7**

(a) Test on S1.

Setting	walking	box
FPS [114]	119.1	92.3
EPS (Sec. 2.3)	79.3	83.7
+ multi-view correspondence	70.8	65.8
+ multi-view appearance	74.9	72.1
+ 3D mixture PS	72.0	78.2
Our full model (Multi-view PS)	**50.2**	**57.3**

(b) Test on S2.

Table 3.1: Contribution of different design choices of our multi-view pictorial structures model to the overall pose estimation performance on **HumanEva-I** dataset. Trained on S1, S2, S3, 3D error in *mm*.

MPII Cooking. We use the same training and test sets as in [114], evaluating 2D projections per camera and reporting percentage of correct parts (PCP). In contrast to [114] where we evaluate only on a single camera, we restrict the training and test set to frames which are recorded by both, the first and second camera. For each view this results in 896 training images from 4 subjects and 1154 test images from 7 subjects (disjoint from training subjects). The two cameras are about 35° apart. Images and annotations are available on our website.

Performance Metric. For a fair comparison to the sate-of-the-art, we adopt the same metric as used in the prior work for the corresponding datasets. For HumanEva-I, we aim to minimize the mean 3D error in *mm* . To this end, we measure the average euclidean distance $d_{3d-error}$ in millimeters (*mm*) between the estimated 3D joint locations X_n^d and their corresponding 3D ground-truth locations X_n^g. This error is given below,

$$d_{3d-error} = \frac{1}{N} \sum_n \|X_n^d - X_n^g\|_2$$

On the other hand, for MPII-Cooking dataset we compute 2D and 3D *percentage of correct parts* (PCP). We have discussed the PCP measure in detail in Sec. 2.4, and this can also be visualized in Fig. 2.4. As PCP represents the percentage of body parts that have been detected

Setting	cam 1	cam 2	3D
FPS [114]	63.7	66.3	43.7
EPS (Sec. 2.3)	66.1	72.3	45.5
+ multi-view correspondence	75.1	75.3	64.1
+ multi-view appearance	75.5	76.1	64.3
+ multi-view appearance	67.3	71.6	53.7
+ 3D mixture PS	71.9	73.5	49.2
Our full model (Multi-view PS)	**80.0**	**80.5**	**68.2**

Table 3.2: Contribution of different design choices of our multi-view pictorial structures model to the overall pose estimation performance on **MPII-Cooking** pose challenge dataset. Results in 2D PCP for cam 1 & 2, and in 3D PCP for the 3D poses.

correctly, therefore unlike 3D error, the goal in this case is to maximize this percentage. Here it is important to mention that the state-of-the-art performance is not yet mature enough on MPII-Cooking dataset. Therefore, PCP metric seems more reasonable performance measure this setting, since an incorrect part does not affect the accuracy of all other parts using this metric.

3.4.1 Evaluation of model components

We start by evaluating the contribution of the various design choices on the overall performance of our model. Below we discuss one by one the performance on both HumanEva-I and MPII-Cooking pose challenge datasets. The impact of the various components of our model is shown in Table 3.1 for HumanEva-I and Table 3.2 for MPII-Cooking dataset. We consider our 2D model, expressive pictorial structures (EPS) as baseline for the evaluation of our design choices in multi-view pictorial structures model. We also compare the results to our flexible pictorial structures model (FPS) which is basically the model of Andriluka *et al.* [12] with flexible parts, presented in [114].

3.4.1.1 Multi-view correspondence

We first evaluate the effect of the variance parameter σ_c^2 of our multi-view correspondence term (Eq. 3.3) on the overall performance of our model. σ_c^2 models the allowed Gaussian noise in the part correspondences across views. Table 3.3 shows the comparison of our full model with all components using different values of σ_c^2, for both HumanEva-I and MPII-Cooking datasets. We observe that the varaince σ_c^2 around 1.0 seems to work reasonably well compared to other values. Too loose correspondence constraint (large σ_c^2) is equivalent to having no multi-view correspondence constraint at all. On the other hand, too tight correspondence constraint (small

Variance σ_c^2	HumanEva-I (*mm*) S1 - *walking*	S1 - *box*	MPII-Cooking (3D PCP) *pose-challenge*
5.0	58.9	52.3	59.0
2.0	56.0	52.1	66.3
1.0	**54.5**	**47.7**	**68.7**
0.5	55.3	48.2	**68.7**

Table 3.3: Evaluation of **multi-view correspondence** terms w.r.t. the variance σ_c^2 in Eq. 3.3.

σ_c^2) makes it difficult to find corresponding samples, since we are operating in reduced state-space for inference tractibility reasons as discussed in Sec. 3.3.4.

HumanEva-I. For HumanEva-I (Table 3.1), we see that adding this multi-view correspondence term to constraint the location of part hypotheses across views reduces the 3D error significantly. For walking/box sequences the mean 3D error comes down from 82.5/72.9*mm* to 73.9/62.8*mm* for S1 and 79.3/83.7*mm* to 70.8/65.8*mm* for S2. We observe that this multi-view correspondence term is the most effective component of our multi-view pictorial structures model. However, it is important to note that our baseline model (EPS) already employ 2D mixture of pictorial structures. We will discuss this in detail in the later part of this section (Sec. 3.3.3).

MPII-Cooking. The improvement we get on MPII-Cooking dataset due to multi-view correspondence constraints is consistent with the improvements on HumanEva-I. This is certainly the most important term in our multi-view approach on top of our 2D model (EPS). In Table 3.2, we note that by adding correspondence information between both cameras *i.e.* cam1/cam2, the 2D PCP improves by 9.0/3.0% over our baseline 2D model (EPS). Interestingly, the improvement is significantly more pronounced in 3D PCP, we get a gain of around 19%. This result demonstrates the effectiveness of having geometric correspondence constraints across views. Moreover, it tells us that 2D PCP with standard threshold (0.5) is fairly loose performance metric and therefore having accurate 2D pose estimates, by this definition, is not enough to reconstruct accurate 3D poses. We also need to have better correspondence in 2D locations of the part hypotheses. This is exactly why our multi-view correspondence constraint helps a lot as it looks for hypotheses with better across-view correspondence. Meaning that, even if the multi-view hypotheses are slightly off in 2D from ground-truth locations, they still have better across-view correspondences to obtain decent 3D reconstruction. Also from Table 3.3, we see that the performance on MPII-Cooking dataset with respect to the variance σ_c^2, is in sync with the results we obtained for HumanEva-I.

Setting	Boosting rounds 500	600	700
EPS (Sec. 2.3)	82.5	-	-
+ multi-view app.	75.7	75.6	75.7
Multi-view PS	54.5	54.5	54.4

(a) HumanEva-I. S1 - walking sequence. 3D error in *mm*.

Setting	Boosting rounds 500	600	700
EPS (Sec. 2.3)	45.5	-	-
+ multi-view app.	53.7	55.8	56.1
Multi-view PS	68.7	68.9	69.1

(b) MPII-Cooking. 3D *percentage of correct parts* (PCP).

Table 3.4: Evaluation of **multi-view appearance** terms w.r.t. the number of boosting rounds.

3.4.1.2 Multi-view appearance

In this part we consider the evaluation of our multi-view appearance constraint which encourages consistency in the appearance of our model parts across views.

HumanEva-I. Interestingly in Table 3.1, we see that for HumanEva-I this multi-view appearance term, which does not require camera geometry to work, is as well critically important. We improve by 6.8/6.7*mm* for S1 and 4.4/11.6*mm* for S2 on walking/box sequences. This reduction in 3D error is significant. We attribute this performance gain to the fact that for HumanEva-I the background is pretty static and the number of subjects is as well limited, therefore the learned part detectors are relatively quite powerful to discriminate between the correct and incorrect part hypotheses and establish consistency across views.

MPII-Cooking. On the other hand, in Table 3.2 we notice that for MPII-Cooking dataset adding this multi-view appearance term to our 2D model (EPS) does not result in any considerable improvement for 2D pose estimation. As a matter of fact, the performance is slightly degraded for cam 2 (from 72.3 to 71.6 PCP). One obvious difference in this case compared to HumanEva-I is that the setting is relatively complex, the scene is dynamic and the training and test sets are totally disjoint. Resultantly, the boosted part detectors are alone not strong enough to disambiguate between the parts and the background. This is true for both single-view and multi-view part likelihood terms. However, its encouraging to see that the performance in 3D has improved significantly, although not even close to the performance we obtain with multi-view correspondence term only. This shows that this across view appearance consistency clue is also effective in building correspodences across views. We also evaluate the improvement we get when we combine this appearance constraint on top of our geometric multi-view correspondence term. We note that doing this gives us slight but consistent improvement for both cameras and as well in 3D. This means, although the multi-view appearance term is not powerful enough alone, but having geometric correspondence in place it does help to distinguish between the smaller number of competing hypotheses.

Another difference compared to the HumanEva-I setting is that the images for MPII-Cooking dataset are of higher quality, hence the resulting feature vector size is relatively much larger. Hence, we believe that the default number of boosting rounds are not enough to learn reasonable multi-view appearance terms in this case. Therefore, in Table 3.4b we investiagte the effects of increasing the number of boosting rounds on the PCP performance of our model. We increase the number of boosting rounds just for training this multi-view appearance terms. The training of single-view appearance likelihood terms is untouched. Remember increasing number of boosting rounds mean selecting more features for the classification task. As expected, we obtain considerable improvement on MPII-Cooking dataset by adding more features during training of Adaboost classifiers for multi-view appearance terms. With 700 rounds we gain 2.4% in 3D PCP (from 53.7 to 56.1). However, there is literally no difference at all on HumanEva-I. Since the improvement is not significant for the full model on MPII-Cooking dataset (69.1 *vs* 68.7) and no improvement on HumanEva-I, therefore we persist with default number of boosting rounds *i.e*. 500 for all our experiments. It is also important to keep in mind that increasing number of rounds significantly increases the training and testing times.

Multi-view likelihood. In Fig. 3.4, we display the single-view likelihood maps for various body parts using our 2D model from Chapter 2. We notice that the single-view likelihood maps for torso and head are quite peaky since torso and head are relatively larger parts, hence easy to localize. Moreover, the image of a person's head contains quite distinct and stable landmarks or features that sets it apart from other body parts. On the other hand, the single-view likelihood maps for the smaller parts (left foot and right shoulder) look quite flat and contain several false modes. Infact, if we observe the likelihood images of cameras 2 and 3 closely in Fig. 3.4, we find out that the strongest peaks in the likelihood maps for the left foot part are actually at the right foot location. Therefore, it is clear that if we are to predict the part locations from this single-view likelihood alone, without employing the spatial body structure, we are doomed to fail.

Next we see in Fig. 3.5 that the likelihood maps for different body parts in cam 1 improve progressively as we add evidence from other camera views. We notice that with the introduction of appearance and geometric correspondence evidence from second view (cf. Fig. 3.5, 3rd column), we are already successful in eliminating most false modes from the likelihood maps, especially for smaller parts. Finally, as soon as we add the third camera view, all the parts get clearly localized (cf. Fig. 3.5, last column).

To compute the multi-view likelihood maps of Fig. 3.5, we employ both multi-view correspondence and appearance constraints. To that end, we only evaluate the multi-view appearance (see Eq. 3.4) for pixels with good geometric correspondence across views. Geometric correspondence is computed using Eq. 3.3. This also allows us to keep the computation feasible.

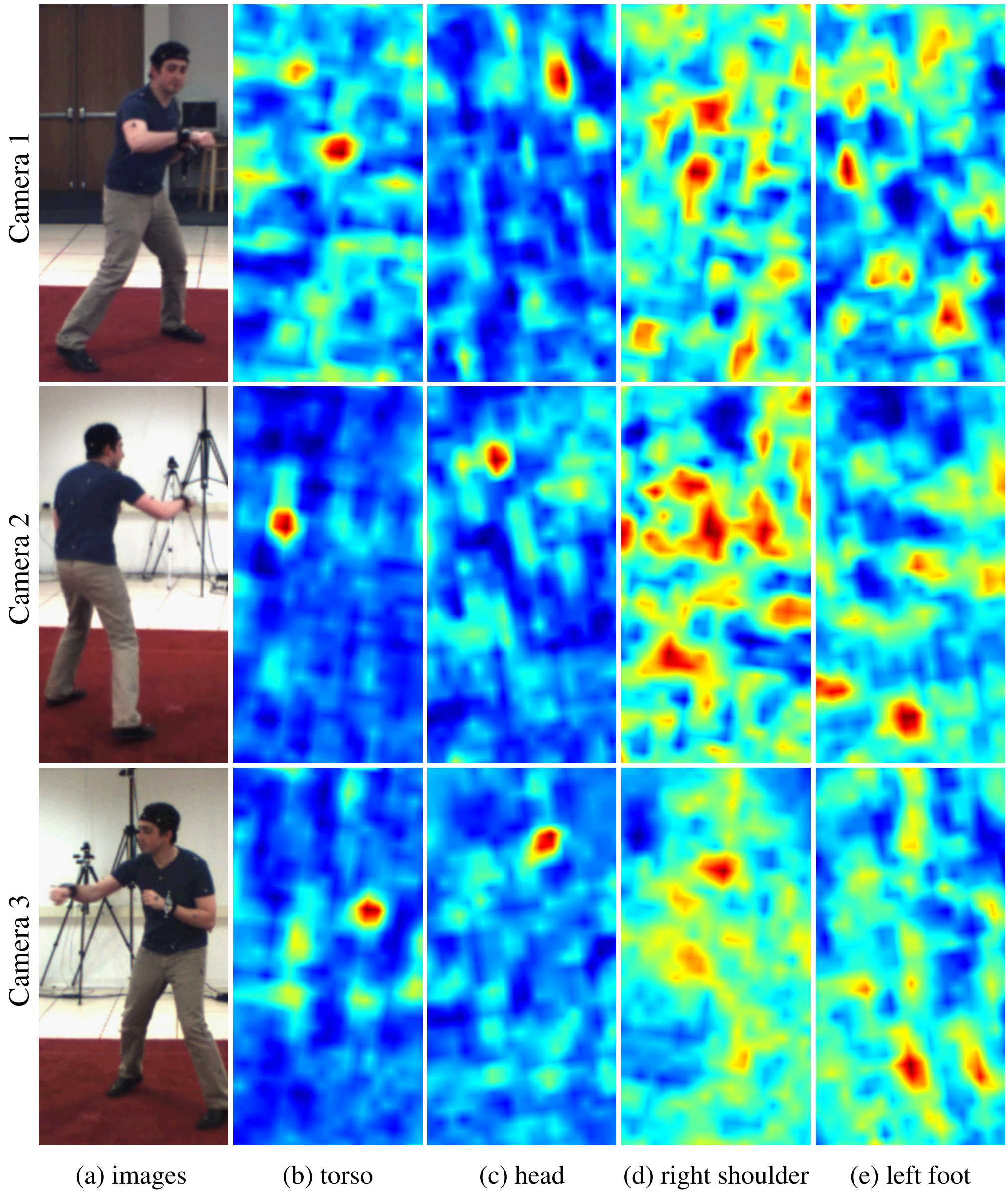
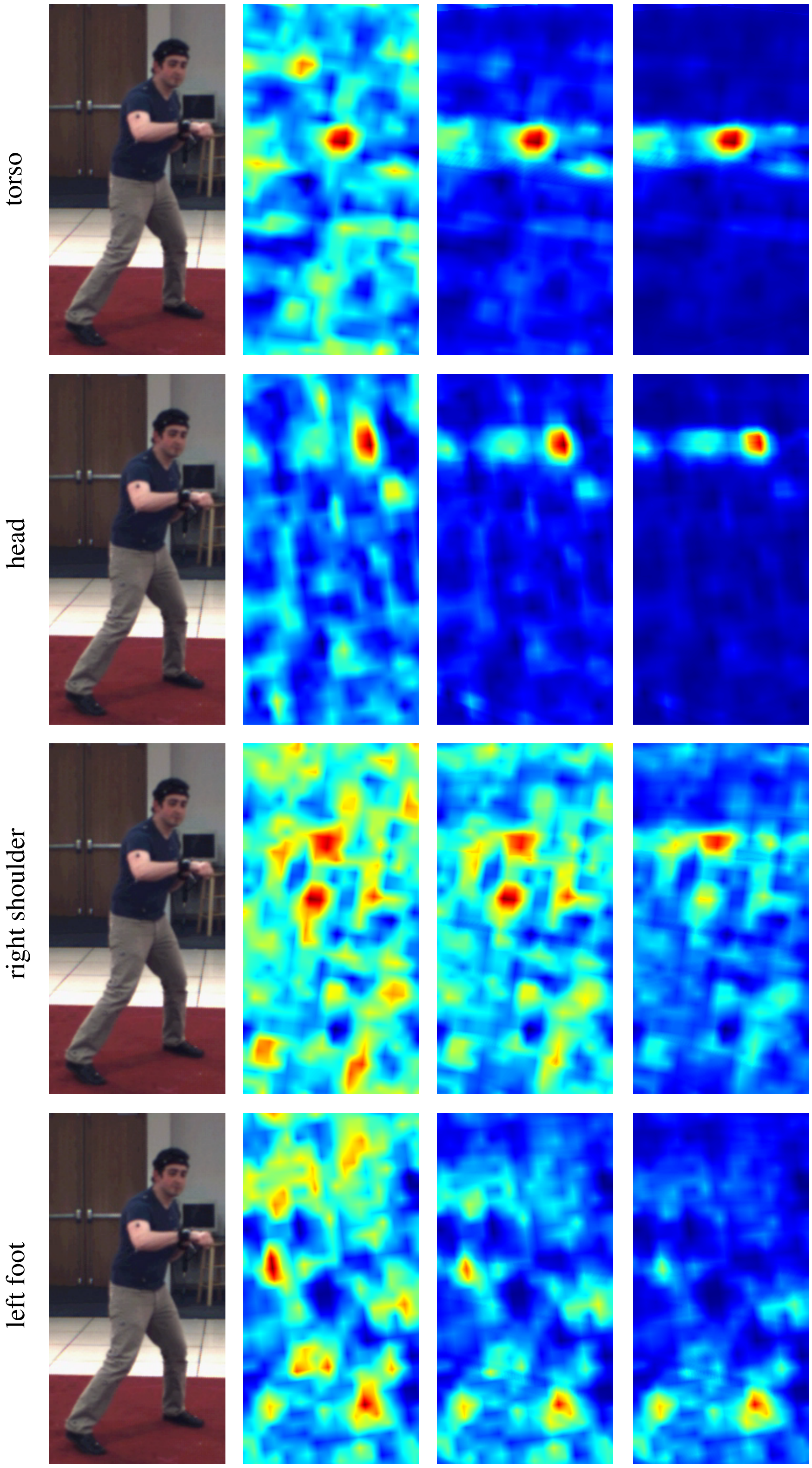


Figure 3.4: Single-view likelihood maps for different body parts in three camera views, according to Eq. 2.3



(a) images (b) single view likelihood (cam 1) (c) incl. 2^{nd} view evidence (d) incl. 2^{nd} & 3^{rd} views evidence

Figure 3.5: Multi-view evidence for more discriminative likelihood maps for images in (a). (b) single-view likelihood maps for first camera view. This corresponds to the first row in Fig. 3.4. (c) the confidence in the likelihood map increases around the correct part location when evidence from the second view is introduced. (d) using both second and third views the confidene further improves.

no. of clusters	clustering	selection	walking	box
1	-	-	87.1	60.6
8	2D	min-var	59.2	52.1
8	3D	classifier	59.4	-
8	3D	min-var	55.9	**47.7**
16	3D	min-var	**54.5**	-

Table 3.5: Evaluation of the **mixture of pictorial structures** models on **HumanEva-I** dataset. Settings: Number of clusters/components, clustering, and selection strategy. Test on S1, trained on S1, S2, S3, 3D error in *mm*.

3.4.1.3 3D Mixture PS

In Sec. 2.4.1.3, we discussed the contribution of the mixture of pictorial structures model for the 2D human pose estimation case. Although quite effective, it still suffered from pose clustering in 2D due to projective ambiguity. To overcome the shortcomings of such 2D mixture model and taking advantage of evidence from multiple views, we introduced a 3D mixture model of pictorial structures components in Sec. 3.3.3. Here we evaluate and discuss the advantages of switching from 2D mixture model to our 3D mixture model.

HumanEva-I. We note that switching from 2D to 3D mixture model (16 components, 3D clustering, and min-var selection) the error for walking/box sequences significantly reduces from 82.5/72.9*mm* to 74.2/69.1*mm* for S1 and from 79.3/83.7*mm* to 72.0/78.2*mm* for S2. We observe that the improvement is more pronounced for the walking sequence as compared to box. This strong improvement can be explained by the fact that the walking sequences shows subjects walking in a circle, *i.e.* they are seen from different view-points, thus a single component model cannot capture this variation well. On the other hand, for box sequence the subjects are mainly facing camera 1. Still our mixture model significantly helps for this sequence as well.

Additionally we examine different options for the mixtures of pictorial structures in Table 3.5. The first line shows the error for a single component *i.e.* 87.1/60.6*mm* for walking/box sequences. Splitting the data into 8 components by clustering the data separately in individual cameras in 2D decreases the error to 59.2/52.1*mm*. A further decrease in error to 55.9/47.7*mm* can be achieved by clustering the data in 3D (line 4 in Table 3.5). In both cases we use the minimum prediction variance to select the correct component. Using a classifier to select the right component (line 3) performs slightly worse with 59.4*mm* for walking, indicating that our min-var selection scheme is a reasonable choice. Finally, increasing the number of components to 16 for walking decreases the error slightly to 54.5*mm*. This setup is used throughout all remaining experiments on HumanEva-I and also in Tables 3.1 and 3.7. In Fig. 3.6, we show some example

no. of clusters	clustering	selection	cam 1	cam 2	3D
1	-	-	70.7	73.0	51.6
3	3D	min-var	**81.4**	**82.3**	**68.7**
5	3D	min-var	80.0	80.5	68.2

Table 3.6: Evaluation of the **mixture of pictorial structures** models on **MPII-Cooking** pose challenge dataset. Settings: Number of clusters/components, clustering, and selection strategy. Results in 2D PCP for cam 1 & 2, and in 3D PCP for the 3D poses. Last row demonstrates the settings in our paper Amin *et al.* [6].

no. of cameras	walking	box
1	75.2	68.5
2	68.3	64.7
3	**54.5**	**47.7**

Table 3.7: Evalution of **number of camera views** on **HumanEva-I** dataset. Test on S1, trained on S1, S2, S3. 3D error in *mm*.

qualitative results of our 2D and 3D pose estimates on sequences from different activities of HumanEva-I.

MPII-Cooking. Here we discuss our findings during evaluation of our 3D mixture PS on MPII-Cooking pose challenge benchmark. In Table 3.2 we notice, the improvement is more obvious in 2D PCP, since the marginally correct poses in 2D images generally do not have reasonable correspondence across views, hence the 3D reconstruction fails. Again reinforcing our argument that the 3D PCP is considerably more strict measure. For single component the pose estimation performance obviously suffers because single model is too restrictive to handle viewpoint variations. We find that 3 components is the maximum number which gives clusters of reasonable sizes with limited training data. This is as well clear from the results we obtain with 5 component clusters. The training data is limited (only 896 images, 2 camera views) and since the number images in the training set are not evenly balanced for different viewpoints of the person, the performance drops slightly for 5 components case as compared to the results we achieved using only 3 components. Overall with 3 components of the 3D mixture PS (clustering in 3D) and minimum variance (min-var) based component selection criterion, we achieve 2D PCP of 81.4/82.3% for cam 1/cam 2 and 68.7% in 3D PCP.

3.4.1.4 Number of camera views

Since HumanEva-I includes recordings from 3 cameras, therefore we evaluate the effect of number of cameras only on this dataset. Table 3.7 compares the performance gain from a single camera over two to three cameras. The improvements show that our model can strongly exploit

Train	Test	CRBM [137]	imCRBM [137]	Yao [160]	our
S1,2,3	S1	55.4	54.3	**44.0**	54.5
S1	S1	48.8	58.6	**41.6**	56.7
S1,2,3	S2	99.1	69.3	54.4	**50.2**
S2	S2	**47.4**	67.0	64.0	52.1
S1,2,3	S3	70.9	**43.4**	45.4	54.7
S3	S3	49.8	51.4	**46.5**	62.4

(a) Walking

Test	CRBM [137]	Yao [160]	our
S1	75.4	74.1	**47.7**

(b) Box

Train	Test (S3)	CRBM [137]	imCRBM [137]	our
S1,2,3	walking	61.84	80.72	**50.1**
	jogging	93.05	89.90	**54.0**
	combo	75.48	84.74	**51.8**
S3	walking	**48.12**	67.48	60.8
	jogging	75.67	86.44	**57.2**
	combo	60.17	75.77	**59.2**

(c) Combo

Table 3.8: Comparison to state-of-the art on **HumanEva-I** dataset. 3D error in *mm*.

the appearance and spatial correspondences across views. Using three cameras greatly improves the accuracy for both sequences walking and box. We know from epipolar geometry that a point in one view corresponds to an epipolar line in the other view. This ambiguity in correspondence can be typically resolved using the correspodence from a 3rd view. This is the reason, we observe significant improvement in Table 3.7, when we introduced the 3rd camera view.

3.4.2 Comparison to state-of-the-art

HumanEva-I. In Table 3.8, we evaluate our model for different sequences (walking, box, and combo) and settings considered in related work and compare to state-of-the-art approaches of [137, 160]. We commence by describing the results for the walking sequence in Table 3.8a. There are six different settings with respect to the training and test split of the data. Our approach performs best in one setting while Yao *et al*. perform best in three, CRBM and imCRBM perform best in one each. A closer inspection reveals that CRBM seems to overfit on the subject as it performs significantly better when the training subjects are limited to the test subjects (lines

Model	Torso	Head	upper arm r	 l	lower arm r	 l	All
Cam-1							
FPS [114]	88.4	84.7	42.9	61.6	46.4	58.3	63.7
our	**92.9**	**89.4**	**72.6**	**79.4**	**68.8**	**76.8**	**80.0**
Cam-2							
FPS [114]	84.5	90.9	56.1	56.3	55.3	54.6	66.3
our	**91.1**	**92.4**	**75.4**	**76.7**	**72.9**	**74.7**	**80.5**

Table 3.9: Detailed comparison to state-of-the art per body part in 2D PCP, on **MPII-Cooking** pose challenge dataset.

2,4,6). In contrast to it our model benefits from additional training data from other subjects, *i.e.* it seems to be able to generalize better. While our error ranges from 50.2 to 62.4, the other approaches vary stronger (CRBM: 48.8-99.1, imCRBM: 43.4-69.3, Yao: 41.6-64.0), indicating that our approach is less dependent on the respective setting.

Next we examine the results for *box* in Table 3.8b. Here related work only reports results for subject S1: 75.4mm (CRBM) and 74.1mm (Yao *et al.*) error. Our model significantly reduces the error by 26.4mm to 47.7mm. In this case CRBM and Yao *et al.* cannot benefit from the strong motion prior used in the walking sequence as the box activity is less cyclic.

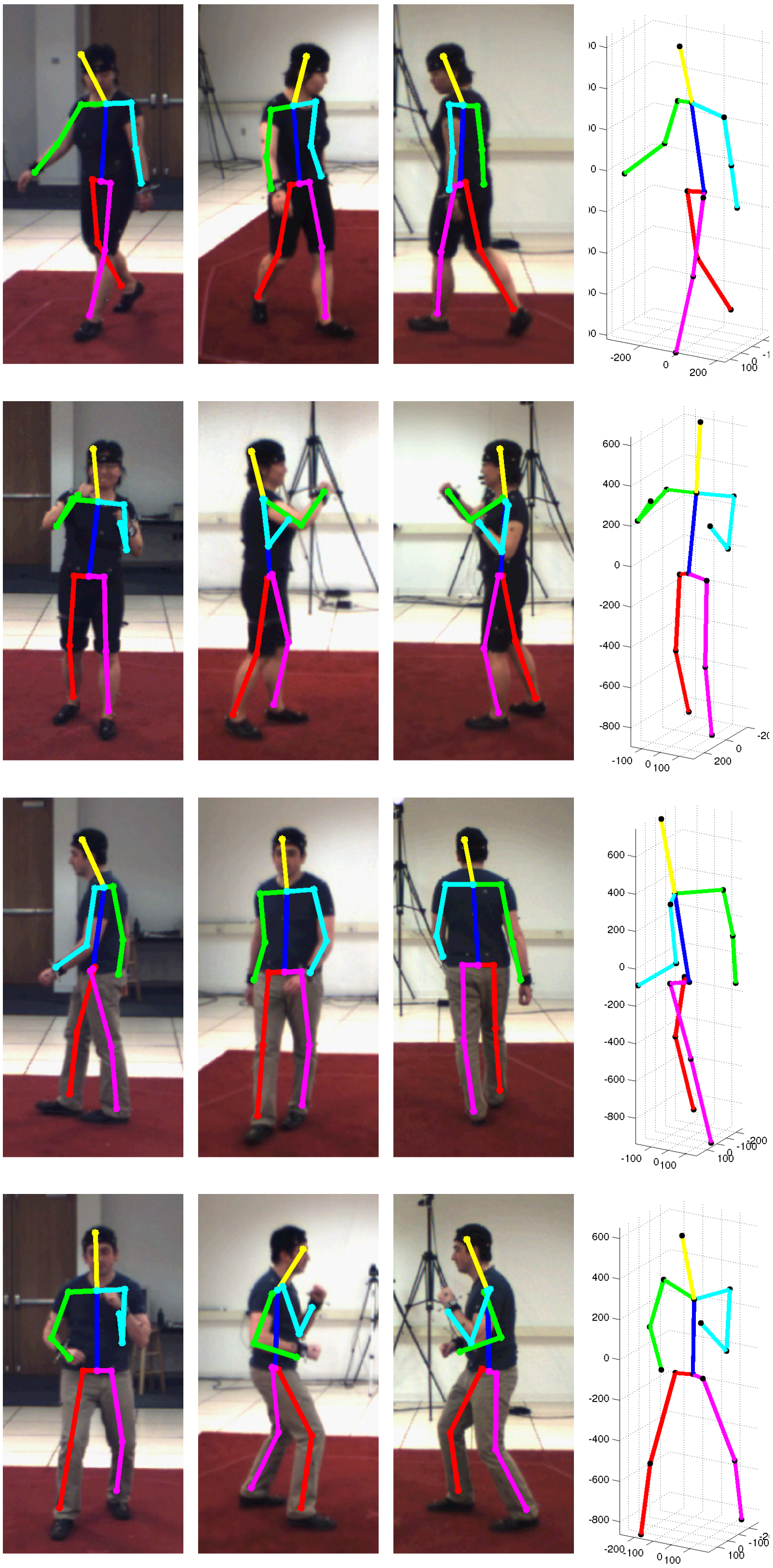
Finally we compare on the combo sequence in Table 3.8c. We first note that for all but one we improve over state-of-the-art. Most notably we achieve an error of only 51.8mm for *combo* (third line), while [137] report 75.48mm when training on all subjects. The challenge for this sequence is that the model is required to handle two different activities. While our model achieves a similar error compared to the walking sequence (see Table 3.8a), as it does not rely on a specific activity prior, especially imCRBM significantly drops in performance. Similar to walking, CRBM seems to overfit on the subject, showing much better results if the training is restricted to the test subject (upper versus lower part of the table).

MPII Cooking. For the MPII Cooking pose challenge dataset, the state-of-the-art approach is our FPS [114] model. In Table 3.9, we compare our approach in detail for all parts. We see that our multi-view approach improves for all parts and both cameras. Especially for the right and left lower arms we improve at least by 17.6 PCP (for right lower arm, cam-2). Overall we improve by 17.7 and 16.0 to an impressive 81.4 and 82.3 PCP for camera 1 and 2, respectively.

3.5 Conclusion

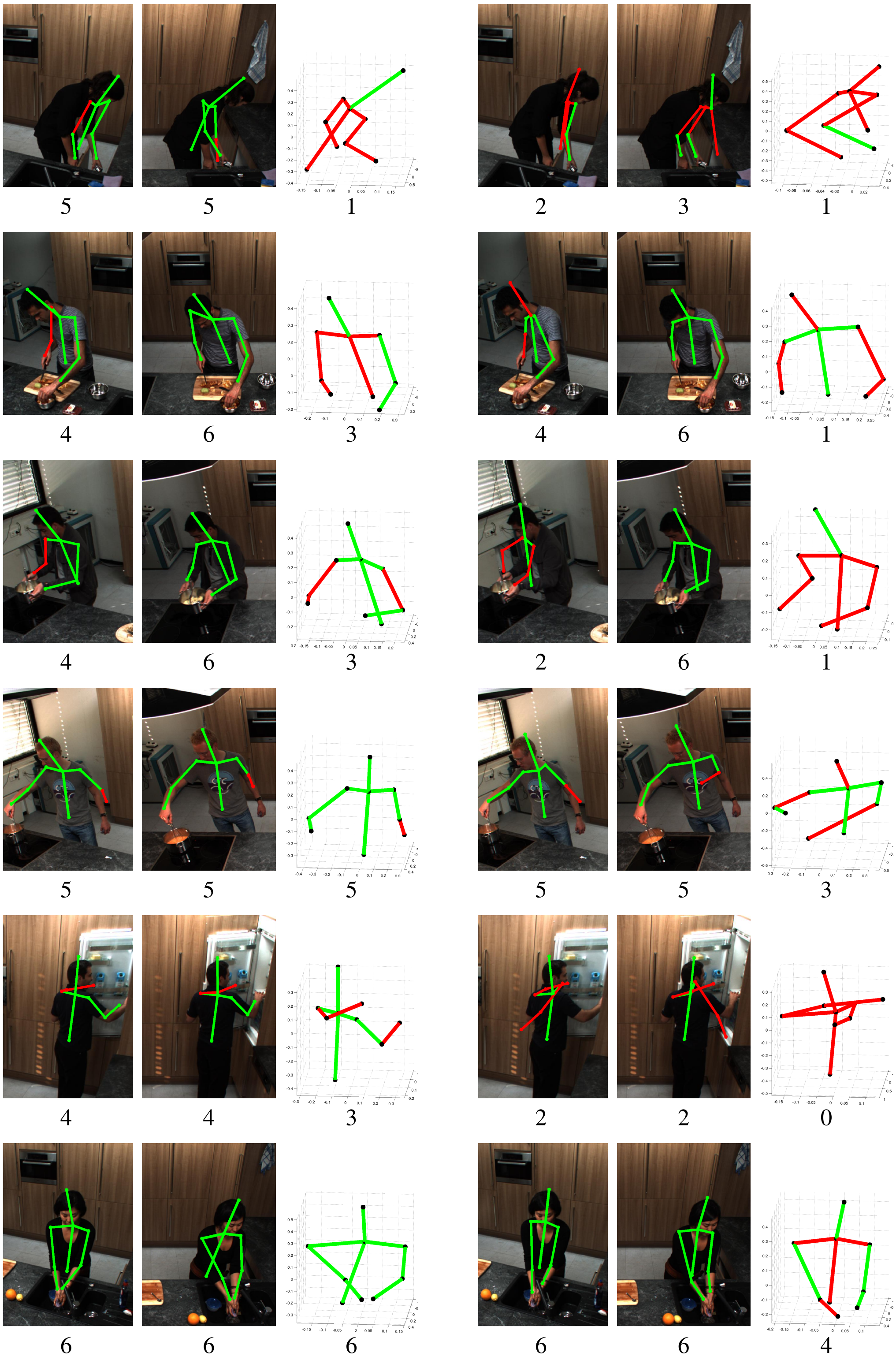
Traditional way of addressing the 3D human pose estimation problem relies on stochastic search in high dimensional space of 3D pose configurations, employing detailed motion priors. In contrast to the state-of-the-art, we have shown in this chapter, that similar or better 3D pose estimation performance can be achieved by formulating the problem as inference over the set of 2D projections of the 3D pose in all camera views. We show that such an approach is efficient and robust, since it builds upon the pictorial structures model which is considered as the de facto standard for 2D pose estimation in the ”wild”.

In this chapter, we extend our 2D pose estimation model *i.e.* EPS (see Sec. 2.3) with multi-view evidence. We augument our 2D model with appearance and geometric correspondence constraints across views. This significantly improves performance on both HumanEva-I and MPII-Cooking datasets with humans involved in various level of activities. Moreover, we also propose novel 3D mixture PS model to better represent the various viewpoints of the person, and to generate consistent part marginals in all views which is critically important to have reasonable part hypotheses during the state-space reduction step. Overall we achieve similar or better results compared to the state-of-the-art [137, 160] on HumanEva-I relying on single frames only, and without incorporating any activity specific priors. On MPII cooking multi-view pose challenge dataset we demonstrate consistent improvements for all parts over the baseline.



(a) cam 1 (b) cam 2 (c) cam 3 (d) 3D pose

Figure 3.6: Joint-2D and 3D pose estimation qualitative results on **HumanEva-I** dataset for different subjects and activity sequences.



(a) cam 1 (b) cam 2 (c) 3D (d) cam 1 (e) cam 2 (f) 3D

Multi-view pictorial structures Expressive pictorial structures (2D model)

Figure 3.7: 2D and 3D pose estimation qualitative comparison. Example images from **MPII-Cooking** pose challenge dataset, with poses estimated by our 2D model (EPS) *versus* our complete 3D model with mutli-view factors and 3D mixture components (MVPS). For 2D case (EPS) we do not employ any multi-view constraints but reconstruct the final 3D pose using triangulation of 2D estimates. The numbers under images show the number of correct parts.

Chapter 4

Model Adaptation for 3D Human Pose Estimation

We propose an automatic model adaptation approach for 3D human pose estimation, which utilizes test-scene specific human pose appearance and improves the model over time. The problem of predicting articulated 3D human poses is specially challenging in real-world scenes with dynamic backgrounds and multiple people. In this work we show that the typical approaches based on pre-trained models are at significant disadvantage in such settings. Although, recent approaches are beginning to show some reasonable results [17], we notice a huge gap in performance when we compare these results to what we achieve in the controlled settings [6].

To address the problem of 3D human pose estimation, we notice that a typical trend in current state-of-the-art is to build upon discriminatively trained part-based models and obtain a set of 2D body pose candidates in multiple views and then subsequently refine by reasoning in 3D [6, 17, 26]. Therefore, the performance of such methods is limited by the performance of underlying 2D pose estimation approaches. In our work in this chapter, we explore an avenue to accomodate the confidence of our 3D pose estimations to reinforce the 2D part detectors and adapt them to the scene at hand. The approach we propose is especially effective for the image sequences with re-occuring people, which is generally the case in videos. Our method tries to learn from the uniqueness of individuals' appearance to reinforce the part detectors, without relying on tracking or any other kind of temporal consistency.

We discuss two different modes of operation for model adaptation. In the *offline* mode, our method makes use of the appearance similarities from all over the sequence. However in the *online* mode, we show that the pose estimation performance on the new frames improve over time while conditioning only on the output of the previous frames. We propose a 2-stage approach. The former stage is tasked to identify the highly confident pose estimates, we call ”Key-Poses”. Whereas the latter stage relies on these supposedly accurate pose estimates to

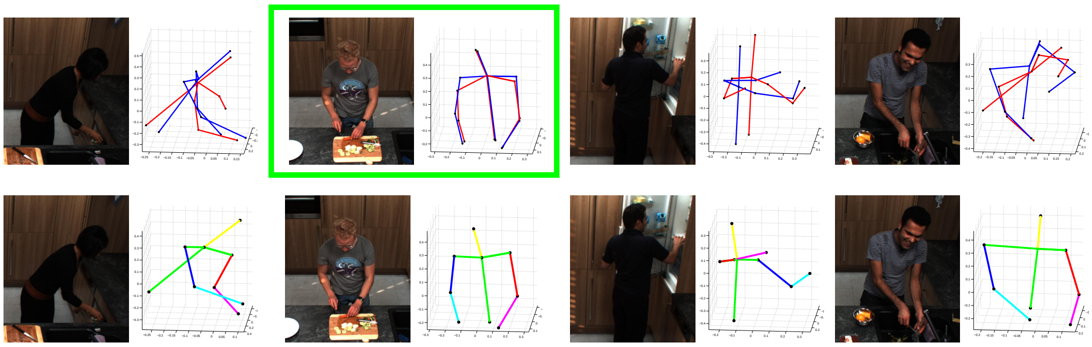


Figure 4.1: Overview of our approach. *Top row:* In the first stage we estimate 3D poses of people in each frame with an ensemble of multi-view pictorial structures models. Output of the models from the ensemble is shown in blue, red and black. We select highly confident key-frames either based on (1) agreement between models in the ensemble, or (2) using a classifier trained on features computed from these outputs. The green bounding box indicates a selected key-frame. *Bottom row*: Output of our final model that incorporates evidence from the keyframes.

enhance the model's ability to perform well. We propose multiple solutions to accomplish the tasks specified for both of these stages. For the identification of the true-positive key-pose hypotheses with high confidence, we dive into the concepts like ensemble agreement, and training a specialized classifier using a number of different features. We study the effect of all the considered features in detail. In our comparisons, we observe that the strongest feature in assessing the confidence of pose estimates comes from the ensemble agreement on 3D pose output. Later, to make use of the already "seen" evidence, we either retrain the 2D part detectors using the appearance templates provided by the key-pose hypotheses, or we use the similarity to these hypotheses as an additional cue in our conditional random field model to enhance the pose estimates. We show that the latter approach is useful in particular, since it allows for online model adaptation, unlike the retraining approach. We evaluate our approach on two publicly available benchmark datasets improving over the state-of-the-art in each case.

4.1 Introduction and related work

Articulated 3D human pose estimation from multiple views has been considered in the literature typically for the motion capture (MoCap) studio environments [34, 83, 127]. However in this work, we focus on the more general and uncontrolled settings with dynamic backgrounds and multiple people present in the scene. This task is certainly more complex and realistic compared to the MoCap settings. One of the key challenges in this setting is that the appearance of the people and background is more diverse and dynamic unlike in the MoCap environment where the people wear clothes which are easily distinguishable from the static background. Moreover, the MoCap studios are well lit which doesn't allow the people appearance to vary significantly. This controlled nature of studios restricts the illumination based natural appearance variations to occur. Therefore, in such settings it is quite common to rely on representing observations based on simple image processing techniques like background subtraction. However, our quest to be able to predict human poses in increasingly realistic environments doesn't allow us to rely on such simplistic image processing means as the main component of pose estimation procedure.

Recent results in 2D pose estimation on natural images [12, 157] have inspired researchers to explore part-based models also for human 3D pose estimation in real-world settings [6, 17, 26, 68]. Some approaches [26, 68] use the 2D part detectors to model the likelihood of the 3D parts, while others [6, 17] in addition also use the 2D part detectors to obtain an initial set of part hyptheses which are subsequently refined by reasoning in 3D. Such approaches have accomplished reasonable results on state-of-the-art pose estimation benchmarks, but essentially rely on effectiveness of the 2D appearance models. For this reason, it is vital to improve the performance of the 2D pose estimation component of these approaches.

Hence in this work, building upon our previous work on multi-view pictorial structures (Chapter 3), we propose an approach to tune the 2D pose estimation model in a way to incorporate varitions of the dynamic scene at the test time. In a typical scenario, we observe same human subjects and the background for several frames. Yet, the appearance of humans and the scene background vary continuously due to person pose variations (depending on the type of the activity), object manipulation, and ofcourse due to subtle changes as a result of constantly changing ilumination conditions. Intuitively, any approach should give a superior performance if it could tune the model continuously using the evidence from the constantly changing scene instead of relying just on a pre-trained model. However, with a few exceptions [33, 109], this intuition is rarely explored in the literature, possibly because it requires accurately and robustly mining the person-specific appearance evidence in the presence of noise using the current state-of-the-art pose estimation methods, which struggle in the dynamically changing settings. There are some exceptions that include works from Ramanan *et al.* [109] to detect stylized poses in video sequence, learn the appearance model, and track them in subsequent frames. Eichner and Ferrari [33] proposed to share appearance for collective human pose estimation.

As described above, the foremost issue is to estimate the confidence of the pose predictions (reliably identify the correct pose predictions) in order to gather some person-specific evidence. Different approaches have considered a similar problem, of robustly detecting the person pose in the image. For instance, Yang and Ramanan [157] proposed to train dicriminative models jointly for person detection and pose estimation. On the other hand, Jammalamadaka *et al.* [63] proposed a more specialized method that predicts the confidence of the pose estimation model based on the combination of various features as a post-processing step. In this paper we follow the direction similar to [63] but also employ features based on the 3D pose reconstruction, which we find to be highly effective for filtering out the incorrect pose estimates.

Our contributions. Our main contribution in this chapter is a novel, effective, however simple framework which can be easily employed for model adaptation in both offline and online modes of operation. To that end, we extend our previous work *i.e.* multi-view pictorial structures model for articulated 3D human pose estimation [6] to enable it to handle the observations available at test time. Our approach has an interesting property that it is (weakly-) conditioned on all of the available images. This means our approach is able to utilize the potential evidence from every image it sees (or has access to). In the *offline* mode this means the entire test set, however in the *online* mode this means all of the previous frames in the sequence. This is in contrast to prior works [6, 17, 26, 68] that typically operate on single-frames only or are limited to temporal smoothness constraints which are effective only in a small temporal neighborhood of each frame [11, 120].

As a second important contribution, we introduce two approaches to assess the credibility of the 3D pose estimates. The first approach is inspired by the work of Jammalamadaka *et al.* [63], in which we train a discriminative model based on various pose quality features. In the second approach we consider the agreement of an ensemble of several independently trained models on a single 3D pose output. Such an approach is naturally motivated by the general acceptance of potential of boosting methods. An interesting finding of our evaluation is that pose agreement alone performs on-par or better than the discriminatively trained confidence predictor. However, a combination of both approaches typically improves the results.

4.2 Overview of our approach

As described earlier, this work on model adaptation is based on our multi-view pictorial structures approach for articulated 3D human pose estimation [6], which is also described in detail in Chapter 3. That approach first jointly estimates the body part hypotheses in each view, and then recovers the final 3D pose by triangulation. This current work for model adaptation operates in either one of two different modes, *i.e. offline adaptation* and *online adaptation*.

Offline adaptation. The offline mode works in multiple stages. First we detect poses using the pretrained model [6] on the complete test-set, then we identify key-poses (section 4.4.1), after that we refine the pose estimation model using either one of the two refinement approaches discussed in section 4.4.2, and finally perform the pose estimation again on the complete test-set.

Online adaptation. On the other hand the online mode only predicts the poses once on the complete test-set. For the first frame in the test-set it uses the pre-trained model [6]. For the later frames it uses the evidence from the pose estimates of all the previous frames. So essentially online mode works in the same way as the offline mode, except that the evidence used for refinement only comes from the previous frames in the sequence. Hence, saving the time and effort of performing pose estimations twice for every image.

Before we move on to describe our proposed extensions to [6], in the following section we briefly recap the Multi-view Pictorial Structures model. However, we refer the reader to Chapter 3 of this thesis to get a more detailed understanding of the model. Then we move on to describing our methods for the mining of confident examples (=key-poses), and model refinement. Later we discuss the results of our approach and improvements we achieve on two publicly available datasets.

4.3 Summary of multi-view pictorial structures model

The pictorial structures model is a part-based model which represents the configuration of an objects's shape as a collection of several rigid parts and a set of pairwise part relationships [37, 39].

The pictorial structures model is explained in detail in the previous chapters, 2 and 3. Following the same notation we used in the earlier chapters, we denote the part configuration as $L = \{l_i | i = 1, \dots, N\}$, where $l_i = (x_i, y_i, \theta_i)$, where $l_i = (x_i, y_i, \theta_i)$ describes the image position and absolute orientation of each part. It is assumed that the pairwise part relationships have a tree structure and the conditional probability of the part configuration L given the image evidence I factorizes into a product of unary and pairwise terms as follows:

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(l_n; I) \cdot \prod_{(i,j) \in E} f_{ij}(l_i, l_j). \tag{4.1}$$

here $f_n(l_n; I)$ represents the likelihood term for part n, $f_{ij}(l_i, l_j)$ is the pairwise term for parts i and j and Z is a partition function.

Multi-view model: As described in [6] and Chapter 3, we extend this model to the case of multiple views in order to reconstruct the human poses in 3D. The multi-view pictorial structures approach that we introduced in Chapter 3 essentially generalizes the single-view model by jointly reasoning about the images of human body parts in each view. Let L_v denote the 2D body configuration and I_v the image observations in view v. Multiview constraints are modeled as additional pairwise factors in the pictorial structures framework that relate locations of the same joint in each view. The resulting multiview pictorial structures model corresponds to the

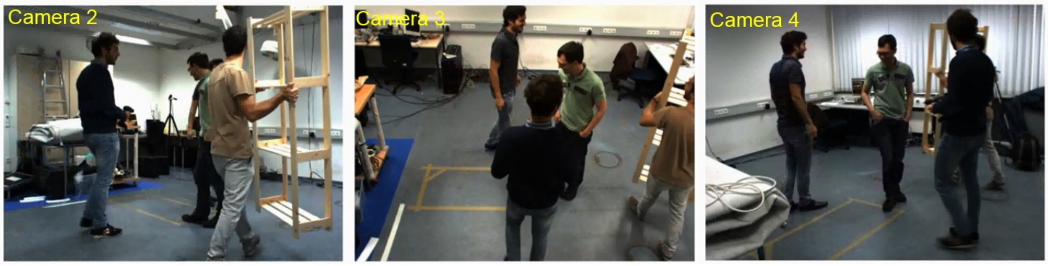


Figure 4.2: Dynamically changing settings.

following decomposition of the posterior distribution:

$$p(L_1, ..., L_V | I_1, ..., I_V) = \frac{1}{Z} \prod_v f(L_v; I_v) \prod_{(a,b)} \prod_n f_n^{app}(l_n^a, l_n^b; I_a, I_b) f_n^{cor}(\mathbf{l}_n^a, \mathbf{l}_n^b), \tag{4.2}$$

where $\{(a, b)\}$ is the set of all view-pairs, $\mathbf{l}_n^v$ represents the image position of part n in view v, in contrast l_n^v in addition to image position also includes the absolute orientation, $f(L_v; I_v)$ are the single-view factors for view v which decompose into products of unary and pairwise terms according to Eq. 4.1, and f_n^{app} and f_n^{cor} are multiview appearance and correspondence factors for part n. The inference is done jointly across all views. The 2D pose estimation resutls are then triangulated to reconstruct the final 3D pose.

4.4 Model adaptation

We notice the monocular and multi-view appearance terms in our multi-view pictorial structures model 4.2 rely on the pre-trained part-detectors which are trained only for a very limited number of ground truth pose examples, hence not quite suitable for dynamically varying settings such as in Fig. 4.2. In such dynamic settings, the distribution of the input image keeps drifting, therefore the pre-trained models find it difficult to provide reasonable estimates in this case. In order to equip our model to support robust pose estimation in those dynamic settings, we propose a model adaptation approach that operates in two stages. In the first stage we aim to automatically extract evidence of the scene specific pose appearance from the test-images (images for which we do not provide any ground truth poses). The scene specific information can only be obtained if we are able to correctly analyse the scene at hand and parse the important information out of the provided images. In our case, this means to correctly identify the true-positive pose estimates from the already processed images. Then in the second stage we use this novel evidence to refine our multi-view pose estimation model, in turn improving the over all pose estimation performance. In the following, we describe these two stages in detail.

4.4.1 Confident examples mining using ensemble methods

Predicting the quality of the reconstructed human poses, or the quality of any image recognition result for that matter, is a non-trivial task and an open research problem in computer vision. As discussed earlier, our objective in the first stage is to correctly identify the examples from the test set for which the initial (pre-trained) model succeeded in estimating the human body pose. This requires a mechanism to assess the quality of the pose estimates. To that end, we consider two methods which we discuss in more detail below.

Before we move on to describing the details of the methods, it is important to understand that quality assessment of pose estimates is an open problem as discussed in detail in the related work section 4.1. We found that the key to finding the correct body pose estimates for both of our proposed approaches discussed in sections 4.4.1.1 and 4.4.1.2, is training an ensemble of several pose estimation models. Several studies [84, 97, 115] have discussed the usefulness of ensemble methods to boost the predictive performance of the model compared to any of the constituent method alone. Importantly, the success of Random Forests [41] and Boosting algorithms for object detection in images has added to the faith in the concept of ensemble models and bagging. Motivated by the effectiveness of ensemble prediction, in our work we proceed by training an ensemble of multi-view pictorial structure models. To train a single model of the ensemble set, we sample a subset of images from the training set. We repeat this sampling multiple times and in total train an ensemble of M multi-view pictorial structure models. It is important to mention here that we randomly sample the images from the training set, and that means the image subsets for different models in the ensemble may not be disjoint. We also experiment by just partitioning our training data in M disjoint subsets, however doing so results in a slightly lower performance. We provide more details on the effects of data partitioning on the pose estimation performance in the evaluation section 4.5.2.

4.4.1.1 3D pose agreement

Our first method evaluates the prediction accuracy by comparing the 3D pose hypotheses estimated by the M ensemble models to the single 3D pose hypothesis estimated by the reference model. The rationale behind keeping the reference model in addition to the ensemble is that the reference model will typically provide a stronger hypothesis about the human pose as it is trained using all available training examples. However, the ensemble models will in turn provide weaker but independent hypotheses about the same 3D pose, as a result increasing the odds of finding the right 3D pose.

To measure the prediction accuracy in practice, we compute the euclidean distance based agreement between 3D pose hypotheses estimated by different models. In Eq. 4.3, we define the quality score based on pose agreement *i.e.* s_{pa}.

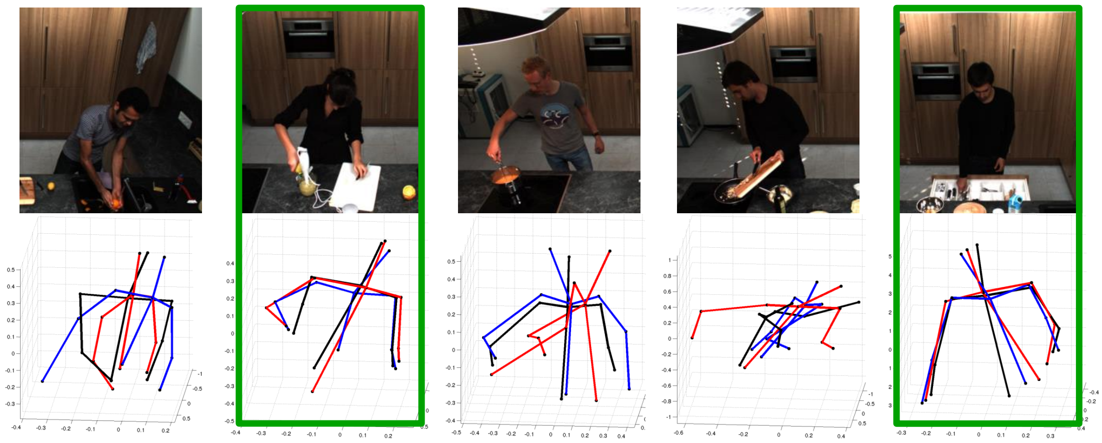


Figure 4.3: Ensemble Agreement

$$s_{pa} = \exp\left(-\frac{\sum_n \sum_m \|\mathbf{x}_n^m - \hat{\mathbf{x}}_n\|_2^2}{N}\right), \tag{4.3}$$

where $\mathbf{x}_n^m$ represents the 3D position of part n estimated with the ensemble model $m \in \{1, ..., M\}$, and $\hat{\mathbf{x}}_n$ is the location of part n estimated with the reference model. We find this approach (comparing output of ensemble models to the reference model) more practical as compare to comparing all pairs of pose hypotheses with one another, since it requires less number of comparisons and it involves comparing multiple weak pose estimates to a strong one. We believe, agreement between two different weak pose estimates is ineffective and unnecessary.

4.4.1.2 Pose classification

In this alterantive approach, we opt to train a discriminative classifier to identify the true positive pose estimates. This requires coming up with a reasonable feature representation, however it is not obvious what kind of features could be useful for that purpose. Typically, researchers rely on the confidence score of the output, however generally it turns out that such a score is not enough to tell whether the algorithm succeeded or failed. The main reason behind this phenomenon is the fact that we rely on approximate inference algorithms (loopy belief propagation) and do not explicitly estimate the true probability distributions. Our inference procedure is explained in detail in Section 3.3.4. To perform exact inference one would need to assume a relatively simple acyclic pose prior (e.g. a tree), which cannot model the complex inter-part dependencies and therefore fails to predict the right pose altogether.

To address this issue of suitable representation, we however study and apply several features that are explicitly related to the quality of pose estimates, we call them ”pose quality features”.

Pose Quality Features

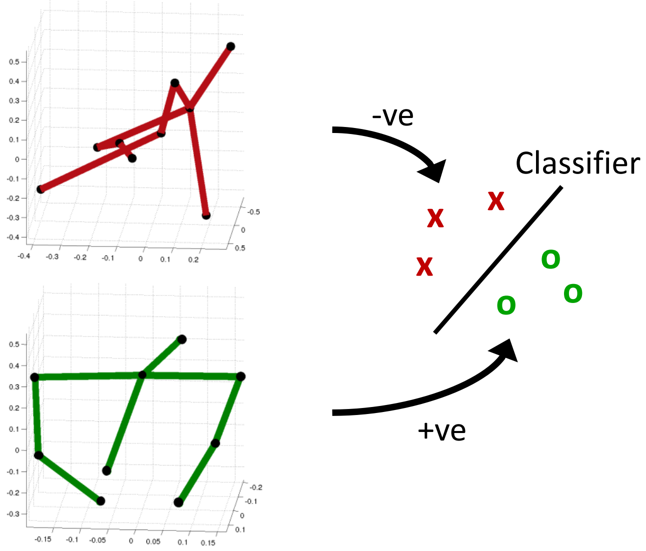


Figure 4.4: Pose Classification

1. *3D pose features.* The 3D pose feature help us discriminate between the plausible and implausible body layouts in 3D *i.e.* these features encode if an estimated 3D human pose is valid or not. The features we use to encode the plausibility of the 3D pose correspond to, torso, head and limb lengths, distance between shoulders, angles between upper and lower body limbs, and angles between head and shoulder parts. We extract these features from the correctly detected 3D poses and all of the groundtruth annotations from the validation sets during the training phase. At test-time the feature vector naturally comes from the detected pose. We found such features are particularly useful for the 3D case as compared to the 2D, where forshortening of body-parts due to image projections render body pose features ineffective.

2. *Prediction uncertainty features.* Intuition suggests that for a confident and successful prediction, the posterior probability density has to be peaky, *i.e.* should have a peak at the location of the searched object (body part in our case). This also implies that a flat posterior distribution results due to prediciton failures, or in difficult cases (clutter, occlusions, etc.). In other words, we can say that the pose estimation method is not confident enough about its prediction. This concept was as well used in [63] to introduce variance based features. To construct features from the marginal posterior distributions of body parts, we first compute the covariance matrix corresponding to the strongest mode in the distribution of each part in all views. Then to encode the uncertainty given by covariance matrix as a feature dimension, we take the L2 norm of this matrix. This is the same criteria we used for component selection in [6], and is similar to the features used in [63]. Note that for symmetrical parts (shoulders, arms, hands, and legs), we generally observe two strong

peaks in the marginal distributions of these parts. Therefore in such cases, we always take the minimum covariance matrix norm of the two strongest peaks.

3. *Posterior.* Although, we mentioned above that the approximated posterior probability itself fails to determine the correctness of the prediction. We still include it as a separate feature and let the classifier decide the importance of its value as a discrimiative feature. The value of posterior score corresponding to the final estimated 2D poses in all cameras is given by Eq. 4.2.

We concatenate these three types of features to produce a combined feature vector of the size $20 + 2VN$ for the upper-body and $26 + 2VN$ for the full-body case, where V is the number of views and N is the number of parts in the pictorial structures model.

Training. We train the pose classifiers in a cross-validation fashion. For example, for training a classifier for m^{th} model of the ensemble, we use the model m to predict the 3D poses for images used to train all other models in the ensemble $\hat{m} \neq m$. Hence for the training of a single pose classifier, data for one model in the ensembe acts as the training set and data for all other models acts as the validation set. The pose quality features extracted from the pose predictions for all those images in the validation set are used to train the m^{th} pose classifier. We repeat this procedure for each model in the ensemble, and end up with M different pose classifiers. We consider detection 3D poses with all body parts estimated correctly to be positive examples, and all others as negative examples. We use the PCP metric with threshold 0.5 to evaluate the correctness of the 3D pose as explained in section 2.4.

For classification, we studied the performance of Adaboost, Linear SVM, and SVM with RBF kernel. RBF kernel SVM tends to overfit to the training data, however linear SVM and Adaboost classifiers both give reasonable and equivalent performance. We compare the performance of different classifier choices for the task of key-pose identification in terms of average precision values. Motivated by slightly better results, we adopt the Adaboost classifier for our final approach for determining the correctness of the pose estimates.

In this case, the classifier score corresponding to the m^{th} pictorial structure model $s_{abc,m}$ is given by the weighted sum of the weak single-feature classifiers $h_{m,t}$ with weights $\alpha_{m,t}$ learned using AdaBoost:

$$s_{pc,m} = \frac{\sum_t \alpha_{m,t} h_{m,t}(x_m)}{\sum_t \alpha_{m,t}} \tag{4.4}$$

$$s_{pc} = \sum_m w_m s_{pc,m}, \tag{4.5}$$

where the weights of the classifier scores are given by $w_m = \frac{\sum_{\hat{m}\neq m}\phi_{\hat{m}}}{(M-1)\sum_{\hat{m}}\phi_{\hat{m}}}$, and $\phi_{\hat{m}}$ are given by the training time cross-validation error. s_{pc} is the final score for this approach which tells the quality of pose estimates.

4.4.1.3 Combined approach

We show in our experiments in section 4.5 that both of the approaches discussed above are quite useful to order the estimated poses w.r.t the correctness. Now, to understand if these approaches are as well complementary would be of immense interest. Therefore, we find it reasonable to consider a weighted combination of pose agreement and classification scores, as another potentially viable option to identify the true positive pose estimates with high confidence.

$$s_{comb} = w_{pa}s_{pa} + w_{pc}s_{pc}, \tag{4.6}$$

Here s_{comb} is basically just the average of both scores, since w_{pa} and w_{pc} are both set to 0.5. In the interest of reducing complexity we did not learn these weights in the current method, and leave it as a potential future study. We demonstrate in section 4.5 that such a naive combination already improves results over using each approach individually.

4.4.2 2D model refinement

In the previous section, we described our method to identify the high scoring pose estimates from the test-set. We choose the pose estimates with the quality score, s_{comb}, above empirically selected threshold of 0.5. We discuss the quality of the pose estimates w.r.t the quality score in the evaluation section. We find that, visibly the pose estimates above this threshold look quite reasonable. We refer to these selected poses as ”key-poses”. Next step in our pipeline is to study the different ways to use the evidence provided by these key-poses to improve our pose estimation model. To that end, we investigate two different avenues to exploit the potential evidence from the selected key-poses for 2D model refinement.

4.4.2.1 Retraining part detectors (Baseline)

In the first approach we retrain the discriminative part classifiers by augmenting the training data with new examples given by the body parts of the key-poses. This approach is an obvious choice to improve the model. Intuitively, this should increase the performance since we not just have more data but also the data is more relevant to the test domain. In practice, we augment the positive part examples using the best part hypotheses coming from the key-poses along with

$b = 5$ of their nearest neighbors. We believe the effect of adding nearest neighbors is similar to the typical data augumentation techniques like image jittering. However choosing nearest neighbors this allows us to not care about manually jittering the part examples as typically done in the state-of-the-art, but rely on natural image variations. We study the effect of b during our evaluations (see Table 4.1). We mine the nearest neighbors from the entire test set. We compute the nearest neighbors in the feature space. We encode each part hypothesis using the shape-context and color features, just as described in section 2.3.2. The shape-context descriptor is computed on a regular grid within the part bounding box and concatenated with color histogram. The nearest neighbors are then found using euclidean distance in this feature space. We also select 10% the part hypotheses with the worst quality score (s_{comb}) to augument the set of our negative part examples for classifier retraining. These examples provide information about scene background, and can be seen as the hard negatives, since they were considered as the best estimates for the pose in the test image during our initial pose estimation step. However, they could not pass our confidence test based on quality score. We consider this retraining approach as the baseline in our quest for finding a better strategy to exploit the evidence from the selected key-poses. For the rest of the paper, we refer to this retraining approach as **RT**.

Although this baseline approach is intuitive and straight forward, but it involves computational overhead due to retraining of all part classifiers with larger number of training examples. This retraining approach is only feasible for our first mode of operation (offline adaptation). In order to reduce the computational overhead and enable online model adaptation, we propose our second approach which is also quite intuitive and as well simple to implement.

4.4.2.2 Appearance similarity unaries

In this second approach we introduce additional unary term for each part in the pictorial structures model that encourages similarity to the key-poses.

The novel term works similar to the per part unary terms of Equation 4.1. Hence, the resulting pictorial structures model per view becomes,

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(l_n; I) f_{SIM}(l_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(l_i, l_j). \tag{4.7}$$

Here $f_{SIM}(l_n; I)$ refers to the factor score for a part hypothesis l_n.

We compute the factor score corresponding to this novel unary term for all the part hypotheses that are sampled for the second stage of the inference procedure (see section 2.3.5 for details). To compute the score for a given part hypothesis in practice, we compare it to the corresponding part of all the identified key-poses and select the one with the most similar appearance, *i.e.* minimum

Dataset	$b = 0$	$b = 5$	$b = 10$	$b = 20$
MPII-Cooking	69.3	69.6	69.6	**69.7**
Shelf	76.5	77.3	**77.5**	77.0

Table 4.1: Effect of the number of nearest neighbours parameter b on the overall performance. b is the parameter used in our retraining (RT) approach.

euclidean distance in the feature space, in order to compute the factor score. The factor score for similarity term (f_{SIM}) for a hypothesis l_n of a given part n is:

$$f_{SIM}(l_n; I) = \exp\left(-\frac{\min_j \|e(l_n) - e(\mathrm{a}_{nj})\|_2^2}{2 * \sigma_n^2}\right), \tag{4.8}$$

where part hypothesis l_n refers to its image position and the absolute orientation, a_{nj} is a hypothesis for part n from the j-th key-pose, and $e(.)$ computes the shape-context and color features at the given location. σ_n^2 is estimated based on variance of all the appearance feature vectors corresponding to part n of the training set. We refer to this approach as **SIM** later in text. Here it is also important to mention that this approach by design helps us to avoid learning based on some wrong evidence due to the potential false positives in our key-poses. We describe the advantage of this property in the evaluation section.

The *online* mode operation only considers the key-poses generated from the pose estimates of the previous frames in the test-set sequence, however the *offline* approach takes into account all of the key-poses for computing the appearance similarity factor score. Note that for our *online* mode of operation, its only feasible to use the appearance similarity approach for model refinement, and it works quite efficiently. Its an overkill to retrain after pose estimation in every image of the test sequence unless one considers online training, which we suggest as a future work to our approach. On the other hand, feasibility is typically not an issue for our *offline* mode of operation. However, we explain later in the experiments section that re-training is not the best option in any case.

4.5 Experiments

4.5.1 Datasets

In Chapter 3, we discussed our basic multiview pictorial structures model and its variant 3D pictorial structures. We showed that our approach achieved state-of-the-art 3D human pose estimation performance on a standard benchmark dataset, *i.e.* HumanEva, without any temporal reasoning like the predecessors. However we concluded that the HumanEva presents a relatively

	Number of Ensemble Models				
Dataset	$E = 2$	$E = 3$	$E = 4$	$E = 5$	$E = 6$
MPII Cooking	**69.7**	69.5	69.6	69.6	69.1
Shelf	75.4	**77.7**	77.3	76.0	74.9

Table 4.2: Number of ensemble models (or ensemble size) E *vs* 3D PCP score. Using the RT+SIM approach for model refinement.

easier scenario that does not demonstrate the challenges of real-world settings. Therefore we introduced two novel multiview datasets for 2D and 3D human pose estimation, *i.e.* MPII Cooking and Shelf, that exhibit settings that one would encounter in the uncontrolled real-world scenarios. However, our evaluations revealed that the excellent pose estimation results of HumanEva dataset could not be translated to these novel datasets. Therefore, in this chapter we focus on 3D human pose estimation in such realistic and dynamically changing settings, importantly with limited amount of available training data.

Multi-view MPII Cooking dataset: We discussed this dataset in detail in Chapter 1. This dataset was originally recorded for the task of fine grained activity recognition of cooking activities. Later in [6] we used it to benchmark the performance of our 3D pose estimation. This evaluation dataset consists of 11 subjects with non-continuous images and two camera views. The training set includes 4 subjects and only 896 images and the test set includes 7 subjects and 1154 images. This multi-view pose challenge subset of MPII Cooking dataset can be downloaded from the project website[1].

Shelf dataset: We introduced this dataset, mainly to study the task of multiple human 3D pose estimation in dynamic settings in our work [17]. The dataset depicts up to 4 humans interacting with each other while performing an assembly task. The Shelf dataset provides 668 and 367 annotated frames for training and testing respectively. For every frame each fully visible person is annotated in 3 camera views. For the Shelf dataset the training and test splits contain the same human subjects but in dynamic background (constantly changing background due to objects manipulation). In our evaluation we use the standard dataset splits for training and testing our model adaptation approach. Moreover, we also employ the same evaluation protocols as used in our original publications [6, 8, 17]. The complete Shelf dataset can be downloaded from the project website[2].

4.5.2 Significance of Ensemble Size

We explained in the section 4.4.1 that ensemble prediction is the fundamental concept behind our proposed key-pose mining strategies. In this section, we analyze the effects of ensemble size. Since, we only have a fixed number of training images, therefore increasing the number of models in the ensemble decreases the number of training examples per ensemble model, and consequently individual models become less representative and weak. However, larger number of models in an ensemble means larger number of independent predictors. In Table 4.2, we show the significance of number of models in the ensemble for the final pose estimation results in terms of 3D PCP on both MPII Cooking and Shelf datasets.

In the results, we notice that less number of models in the ensemble seem to perform better, with the exception of $E = 2$ on Shelf dataset. The reason for this performance drop is the fixed and limited amount of training data. Therefore, by increasing the ensemble size E, we end up significantly reducing the number of training examples per model. Resultantly, the individual models are not representative enough of the general human pose appearance and articulations. Thus, models become less effective to generate reasonable predictions.

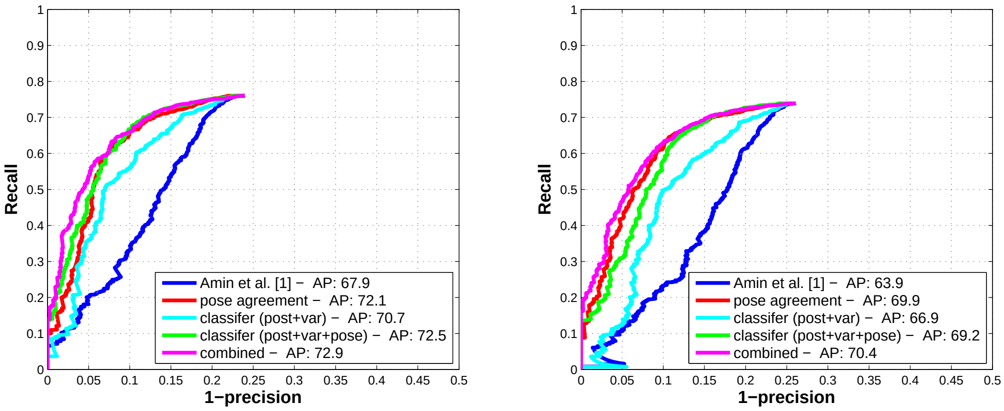
We performed these experiments for analysing the significance of ensemble size and number of images used to train a single model, for the overall pose estimation performance. However, we did not actually use this outcome to select the number of ensemble models to employ. For all our experiments we empirically choose 4 models as the ensemble size, based on the criterion that number of images per model should not go below 150. Note that in addition to the ensemble models, we also train a single ”reference model”. In contrast to the ensemble models, the reference model is trained using the complete training set, and this is the model used for final human pose estimation. This is explained in detail in section 4.4.1.

4.5.3 Analysis of Key-poses

Here it is important to mention that we are only analyzing the individual parts of the key-pose, and and not the key-pose as a whole. This makes sense since we intend to use single parts of the key-poses and perform part-level model refinement.

To analyze the performance of our approaches for confident-examples mining (*i.e.* key-pose detection), we compute recall-precision curves and average precision values. The results are shown in Fig. 4.5 and 4.6. In this analysis we focus on the more challenging parts that are generally difficult to detect due to their size, occlusions, and the amount of imaging variations

[1] http://www.datasets.d2.mpi-inf.mpg.de/amin13bmvc/mpii_cooking3d-1.0.zip
[2] http://campar.cs.tum.edu/files/belagian/multihuman/Shelf.tar.bz2



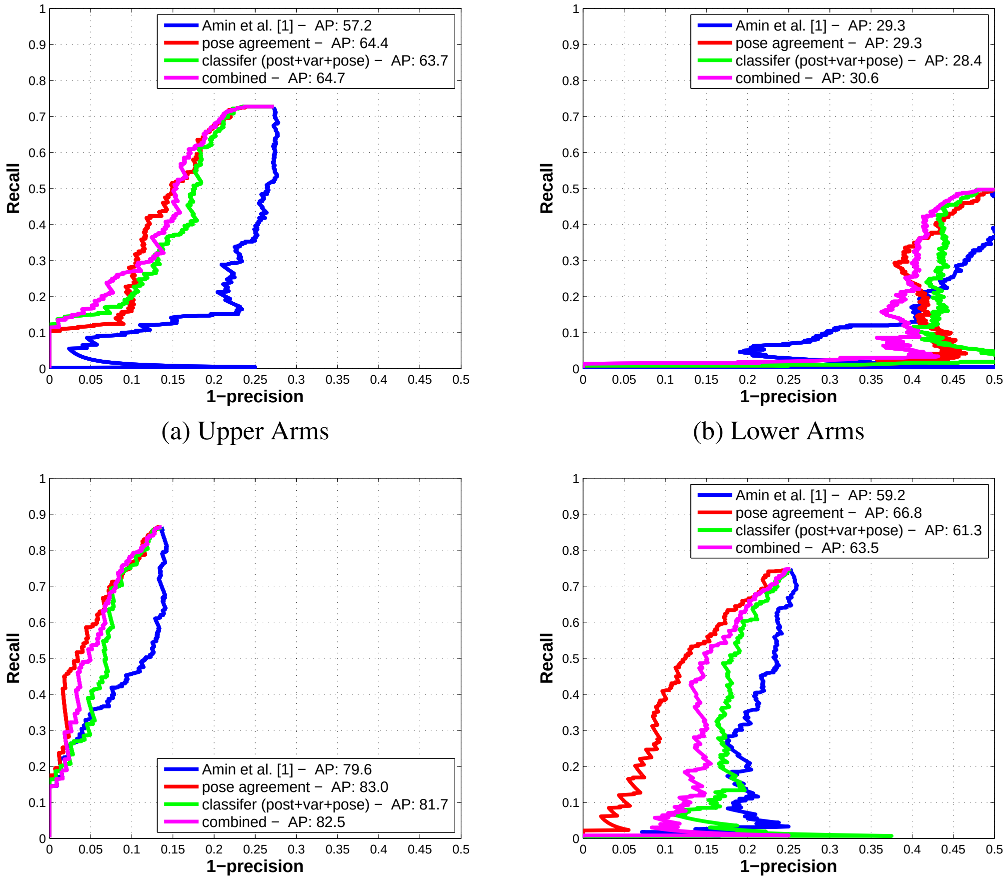
(a) Upper Arms

(b) Lower Arms

Figure 4.5: MPIICooking dataset: Key-frame selection using score based on the posterior marginal of the model from [6] (blue) compared to variants of our approach. Best result corresponds to combination of ensemble agreement and pose classification scores given by Eq. 4.6 (magenta).

they go through. We omit the bigger and more static parts *i.e.* head and torso from the evaluation since they are almost perfectly localized for pose estimates in all images, and therefore show very slight performance differences in terms of average precision results. For all other parts, which are smaller in size hence more susceptible to noise, we observe that directly using the marginal posterior of the PS model as pose confidence leads to poor results (blue curve in Fig. 4.5 and 4.6). This poor performance makes them interesting candidates for the evaluation of our key-pose identification approaches.

Multi-view MPII Cooking dataset: For the multi-view pose challenge subset of MPII Cooking dataset, using only the marginal posterior as the confidence measure, we get AP of 67.9/63.9 for upper/lower arms respectively. Next, instead of using just marginal posterior directly, we train the pose classifier of section 4.4.1.2 using this marginal posterior score along with the variance of the marginal posterior density as features. Doing this, we improve the AP to 70.7/66.9 respectively. The performance of the pose classifier approach further increases to 72.5/69.2 when we extend the feature vector with 3D pose cues. This result underlines the importance of 3D pose features for classification. Interestingly the simpler approach for finding key-frames, the 3D pose agreement from section 4.4.1.1, alone gives us better results *i.e.* 72.1/69.9, as compared to training the pose classifiers. Naively combining scores of both approaches (classifier & pose agreement) as given by Eq. 4.6, we slightly step up the average precision of the detected key-poses to 72.9/70.4. This result shows the importance of different levels of complementary features in the process of extracting key-poses with high confidence. Moreover, this also shows that there is a definite potential for further improvement with a more wise combination strategy. For example training the weights of the combination in Eq. 4.6 could result in a more effective way to incorporate the complementary advantages of the two approaches.



(a) Upper Arms

(b) Lower Arms

(c) Upper Legs

(d) Lower Legs

Figure 4.6: Shelf dataset: Key-frame selection using score based on the posterior marginal of the model from [6] (blue) compared to variants of our approach.

As discussed earlier, it makes sense to analyze the smaller and more difficult-to-detect body parts. As an example in Fig. 4.7, we illustrate the typical detections of the right-hand part of our identified key-poses w.r.t the key-pose scores. This helps us visually understand the hard cases. We notice that using the baseline approach (marginal posterior score) to rate the part estimates, leads to many false positives in the high scoring region. Moreover,l several correct part estimates end up in the low scoring region. On the other hand, using our key-pose score from Eq. 4.6 the order of part estimates is much more reasonable in the scoring range between [0,1] *i.e.* correct ones are generally found in the high scoring half and the incorrect ones in the low scoring half. The score $\hat{s}_{comb}$ in Fig. 4.7 is the normalized form of the score in Eq. 4.6.

Shelf dataset: The recall-precision curves for the Shelf dataset are shown in Fig. 4.6. Although, here it appears that the detection of lower arms is significantly more challenging, but still the order of AP curves are generally in line with the ones from MPII Cooking dataset. One exception in this case, the key-pose results for legs using the pose agreement approach alone outperforms the combined score of Eq. 4.6 i.e., 83.0/66.8 as compared to 82.5/63.5 for upper/lower

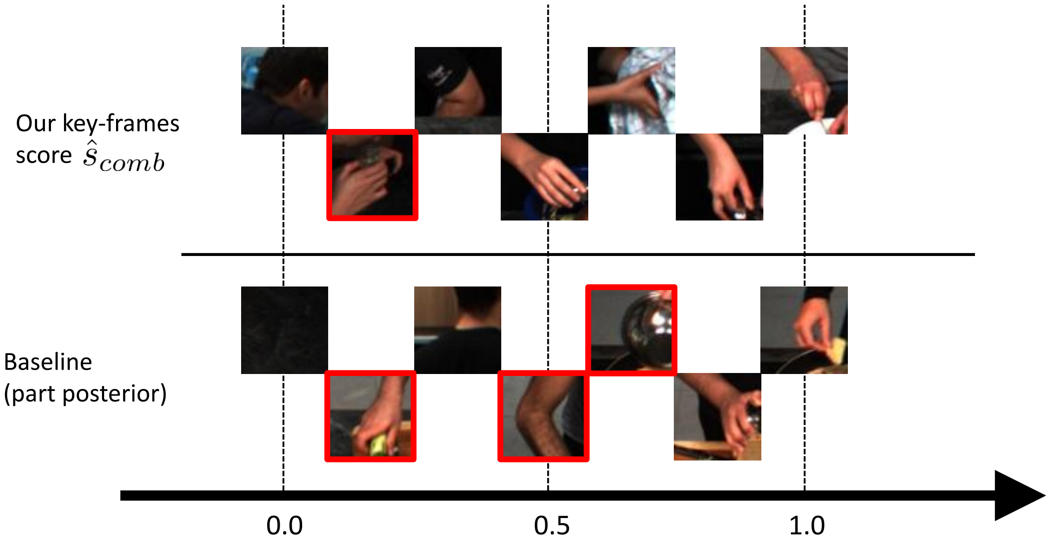

Our key-frames score $\hat{s}_{comb}$
Baseline (part posterior)
0.0
0.5
1.0

Figure 4.7: Example of typical key-frames for the right hand part. The red bounding boxes indicate the wrong classifications. *Top row:* Our key-frames mining score $\hat{s}_{comb}$ from Eq. 4.6 normalized between [0,1]. *Bottom row:* Mining key-frames only using the part marginals.

legs. The reason for this behavior is the significantly worse performance of the pose classifier of Eq. 4.4. This result suggests the need to learn the weights w_{pa} and w_{pc}, when combining both scores in Eq. 4.6. We already mentioned to investigate this in our future work.

4.5.4 Pose estimation results

In this section we discuss the improvements in the final pose estimation performance conditioned on the identified key-poses. To identify the key-poses we used the combined score s_{comb}, since we have observed the best results using that score in the previous section. We study both approaches RT and SIM to incorporate the appearance evidence from the underlying images. We use body-joints as parts in our pictorial structures model, instead of actual limbs. Such representation has been adopted in multiple papers [6, 114, 120]. This representation is commonly referred to as flexible pictorial structures model (FPS), since it makes the model more flexible as it allows to counter the effects of limb foreshortening.

Multi-view MPII Cooking dataset: Since MPII Cooking dataset basically focuses on upper body of the human subjects, we use the standard pose configuration, i.e., 10 upperbody parts. We introduced this configuration in [114].

We report the results in percentatge of correct parts for the 2D projections of the poses per camera (i.e., 2D PCP) and also the 3D PCP for the final pose based on multiple cameras in

Table 4.3. In Chapter 3, we discussed the results based on our original multi-view pictorial structures model 3.1. We consider those results as the baseline for our evaluations in this section.

First we look at the proposed retraining approach (RT), which includes the part appearances coming from the key-poses as additional positive training examples. Our RT approach achieves 83.6/84.2 PCP for cameras 1 and 2 respectively, and 68.9 in 3D PCP. RT shows improvement also for all individual parts. This improvement shows the classifier is able to learn the person or scene specific information provided mined examples using key-poses. Quite interestingly, our second approach (SIM), which involves introducing a novel unary term to our pictorial structures model based on appearance similarity to the key-poses, achieves similar or better PCP results compared to RT. For 2D case we get 83.4/84.4 overall PCP and 69.3 in 3D PCP. This indicates that though it might be difficult for a classifier to learn some generic featuers in such difficult cases, a simple feature matching approach could still help. Observing closely, we see that the improvement in this case is more pronounced on the lower arms compared to the retraining approach. Moreover, it is important to note that this second approach SIM does not require retraining of any model parameters. As discussed before this allows the approach to be practical enough to be used for the online model adaptation case.

Additionally, for the completion of results, we also evaluate the combination of the two approaches i.e. RT+SIM. For this approach, we not only retrain the part classifiers using appearance features from the key-poses, but also introduce the novel appearance similarity based unary term f_{SIM} to the multiview pictorial structures framework. As expected we achieve the best results using this combination approach since it combines the benefits of both approaches and results in stronger unaries for the part hypotheses, however the gain we achieve is not as significant, since the raw source of evidence is the same for both RT and SIM, and we are theoretically not introducing any independent evidence for the second approach in the combination.

In Figure 4.10, we illustrate some example improvements of our current approach, over the original multi-view model of [6] on MPII Cooking dataset. Furthermore, in Fig. 4.11, we demonstrate some typical failure cases of our current approach.

Shelf dataset: Here in this case, we use a full body model with 14 body parts following our work in [17] for multiple human 3D pose estimation. Since our approach from [17] is a cascaded approach, and is based on the 2D part detectors of [6], therefore the accuracy of that approach of [17] is bounded by the performance of the employed 2D part detectors, which are pre-trained as usual. Such an approach is unable to recover once the 2D part detectors fail to fire in the first stage of the pipeline. In comparison our current approach, which is able to utilize the test scene specific information for model adaptation, achieves far better results in terms of percentage of correct parts (PCP) in 3D. In Table 4.4 we list the 3D PCP results for all parts and compare to the results we get with our baseline approaches of [6] and [17]. Our retraining approach (RT)

Model	Torso	Head	upper arm r	upper arm l	lower arm r	lower arm l	All
Cam-1							
Amin et al. [6]	92.9	89.4	72.6	79.4	68.8	76.8	80.0
RT (offline)	95.2	93.8	**75.3**	**83.4**	73.6	80.9	83.6
SIM (offline)	**95.8**	92.5	74.0	82.6	73.5	82.2	83.4
RT+SIM (offline)	**95.8**	**94.0**	74.8	83.3	**74.2**	**82.3**	**84.0**
SIM (online)	93.1	91.6	73.7	81.3	71.1	78.9	81.6
Cam-2							
Amin et al. [6]	91.1	92.4	75.4	76.7	72.9	74.7	80.5
RT (offline)	92.1	95.5	79.2	82.8	76.9	78.8	84.2
SIM (offline)	92.4	**96.2**	**79.5**	81.6	**77.1**	79.6	84.4
RT+SIM (offline)	**92.6**	**96.2**	78.9	**83.3**	**77.1**	**79.7**	**84.7**
SIM (online)	91.3	94.2	78.3	79.9	76.8	79.1	83.3
3D							
Amin et al. [6]	75.6	82.0	**59.4**	67.5	59.6	64.1	68.0
RT+SIM (offline)	**76.9**	**83.1**	58.8	**71.2**	**60.2**	**67.6**	**69.6**
SIM (online)	76.1	82.6	59.1	69.7	59.5	65.3	68.7

Table 4.3: MPII Cooking: accuracy measured using percentage of correct parts (PCP) score [38]. We compare the model of Amin et al. [6] with variants of our approach. *RT* stands for model retraining, *SIM* stands for a model augmented with similarity factors.

	MVPS [6] (Baseline)	RT (offline)	SIM (offline)	RT+SIM (offline)	SIM (online)
Actor1	66	68.5	**72.1**	72.0	68.8
Actor2	65	67.2	69.4	**71.3**	66.7
Actor3	83	83.9	84.9	**85.7**	83.1
Average	71.3	74.4	77.0	**77.3**	73.1

Table 4.4: Shelf dataset: accuracy measured using 3D PCP score. We compare the model of Belagiannis et al. [17] with variants of our approach. *RT* stands for model retraining, *SIM* stands for a model augmented with similarity factors.

outperforms the baseline of [17] by around 3% in PCP. Quite interestingly, we achieve 77.0 PCP (2.6% more than the RT approach) with our second approach SIM which includes just an extra similarity term for model inference only at test-time. SIM approach even outperforms the RT approach by further 2.6% in PCP. This result is quite significant and hints at the potential of simpler means to incorporate scene specific evidence. Finally, just like in the case of evaluations on MPII Cooking, we also study the combined approach i.e. RT+SIM, and obtain an additional gain of 0.3 PCP. Some examples of the qualitative results for our model adaptation approach on Shelf dataset are depicted in Fig. 4.9.

To summarize our evaluations on the two datasets, we observe that the simpler approach SIM achieves similar or better results compared to the RT approach which is computationally expensive, time consuming, and especially even impractical for *online* model adaptation. The superior

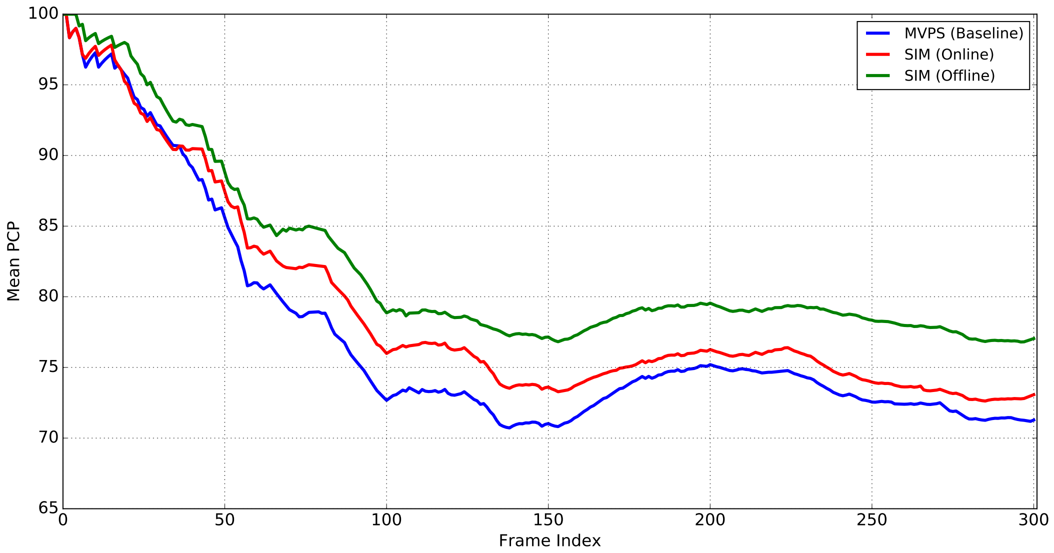


Figure 4.8: Average PCP results on Shelf dataset. The x-axis gives the frame index until which the average was computed.

performance of the SIM approach can be attributed to the fact that it is based on the euclidean distance based similarity between the part hypotheses and the example parts of the identified key-poses. Although euclidean distance is a cummulative measure but it considers all dimensions of the appearance feature vectors when looking for the similarities. Furthermore, it also shows that there does exist some features which do not perform well during the feature selection process of AdaBoost when learnt using part examples of the key-poses together with all the examples of the training set. However these under-performing features are potentially quite useful when comparing only individually the test hypotheses and part examples of the mined key-poses. This phenomenon is quite understandable, since neighboring images or atleast appearance of individual parts could turn out to be quite similar. As part of the future work, we intend to study the impact of other distances like L1-norm, for the appearance similarity. The worse performance of our retraining approach can also be explained by the fact that the presence of false positives in the key-frames results in noisy positive examples for the training of boosted part classifiers. Note that our appearance similarity approach SIM is typically not affected by the false positives in the key-frames, as we only look for the similarity to the closest key-poses in terms of part appearance.

4.5.5 Online model adaptation

In this section we discuss the results of our *online* mode of operation which allows us to adapt the model on the fly. In Tables 4.3 and 4.4, we compare the results of this online mode which employs only the appearance similarity (SIM) for boosting the scene specific appearance representation of the model. Overall we achieve an average 3D PCP of 68.7 on MPII Cooking dataset and 73.1 on Shelf dataset. We notice that these results are worse as compared to all of the *offline* approaches, *i.e*. RT, SIM, and RT+SIM. Besides, the overall gain of this online adaptation approach with respect to the MVPS baseline [6] is not as significant as in the offline case. The obvious reason is that the pose estimation in this case is only conditioned on the past data, *i.e*. the key-poses only come from the previous frames in an image sequence or a video, whereas a better match for the faultering pose prediction could be in some future frame. On the other hand, for the offline mode the pose estimation in each frame is conditioned on the pose estimation results of the complete test-set. Currently, we keep all the key-frames from the history, since the datasets are not so large.

We dig deeper to analyze the results in detail in comparison with the MVPS (baseline) and the SIM (offline) approaches. We illustrate this comparison in Fig. 4.8 in terms of average PCP results on the Shelf dataset. We observe that the two curves for MVPS (baseline) and SIM (online) are quite similar in the beginning (*i.e*. until around frame index 40), however the curve for SIM (online) improves over time. This is understandable since there is not enough key-pose evidence available in the beginning, besides the mined evidence in the start is most likely not diverse enough to solve the difficult cases. Moreover, the significantly better results of SIM (offline) approach, which are depicted as well by the superior average 3D PCP curve in Fig. 4.8, also suggest that for a system to meaningfully adapt to the deployment scenario it needs to see a variety of evidence through mined key-poses. Since we only have a limited number of frames (*i.e*. 301) in the Shelf dataset, the online approach does not achieve the results as good as in the offline case in the end. However one could expect the online approach would reach the accuracy of the offline approach in the long run. We would like to investigate this intuition as a future work.

4.6 Conclusion

In this chapter we propose an approach for 3D human pose estimation that collects the scene specific evidence from the input test images and adapts the model accordingly, for improved predictions. We described two different modes of operation, *offline* and *online*, for model adaptation. We show that the model adaption comes in handy in the real-world settings, where the image distribution continuously varies over time. Our approach operates by selecting poses which are predicted with high confidence, and then uses the evidence from the underlying image to perform pose estimation in the other image frames with the improved model. In order to identify true-positive pose estimates with high confidence (key-poses), we analyzed two strategies: (i) discriminative classification; and (ii) ensemble agreement on the output 3D poses. In our evaluations, we observe that a naive combination of both strategies already achieves the best

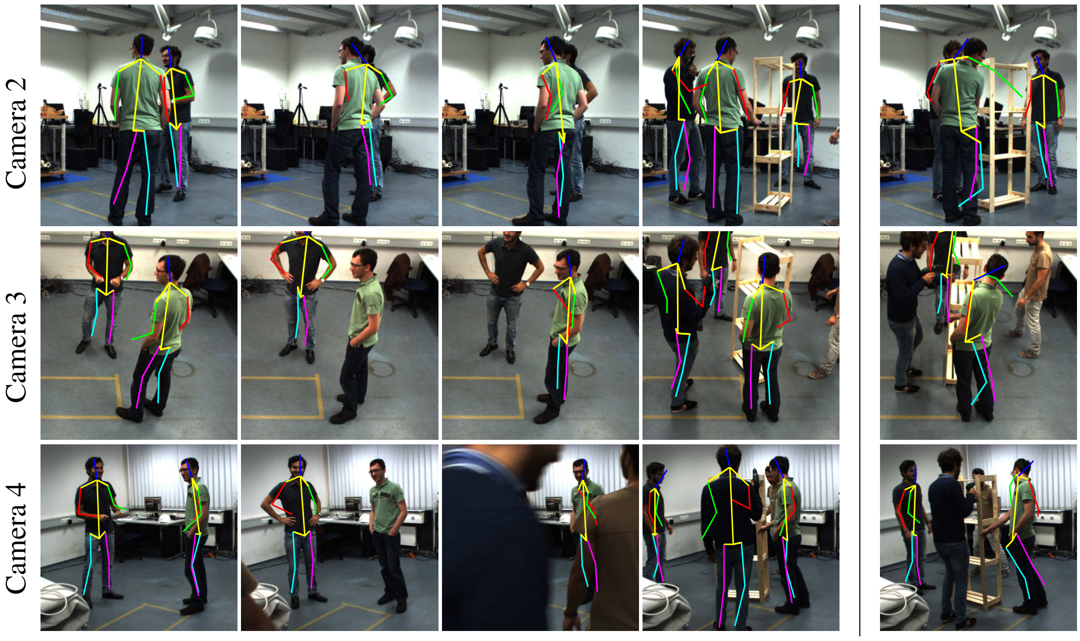


Figure 4.9: Examples of pose estimation results of our approach on the Shelf dataset. Last column shows an example of the failure case.

results. Hence, there is a definite potential for further improvement. Having said that, the ensemble agreement alone improves considerably over the baseline confidence measure which is based on the pictorial structures output.

Moreover, we also propose two alternative approaches for taking into account the potential evidence offered by the identified key-poses. We show that our second approach which offers an alternative way to achieve model adaptation without having to retrain any model parameters, performs on par or better than the retraining approach. We discussed that how this second approach allowed us to completely avoid the overhead of retraining, while introducing only relatively insignificant amount of extra computations at test time. As an additional benefit, it makes *online* model adaptation feasible. It is important to note that, *online* model adaptation is not viable with the baseline model refinement approach, i.e. RT, due to excessive overhead of retraining the part detectors every time a new keypose is selected.

We have shown the effectiveness of our approach on two publicly available datasets. As a future work, we are looking into generalizing our approach to multiple rounds of confident examples mining, and we will be exploring other approaches for automatic acquisition of training examples from unlabeled images.

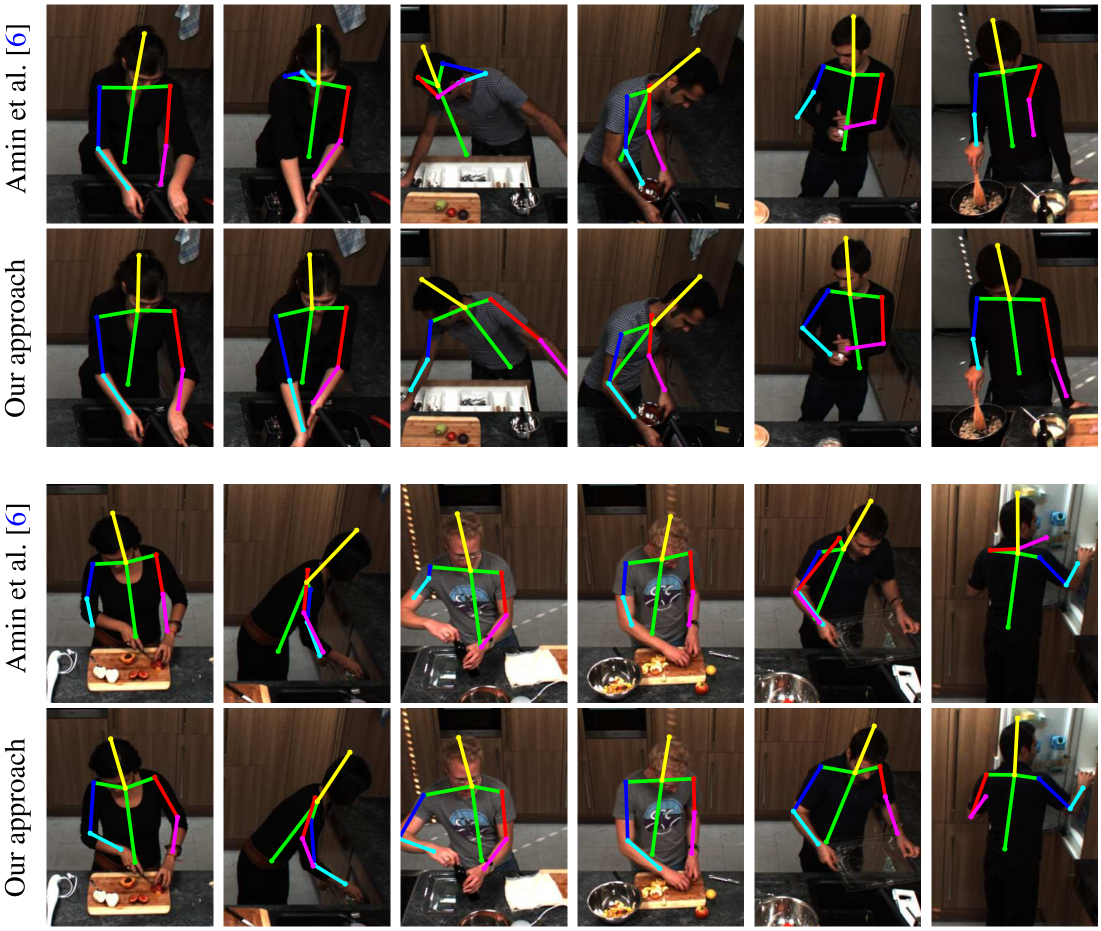


Figure 4.10: Examples of pose estimation results obtained with our approach and comparison to the state-of-the-art approach of Amin et al. [6] on the MPII Cooking dataset.

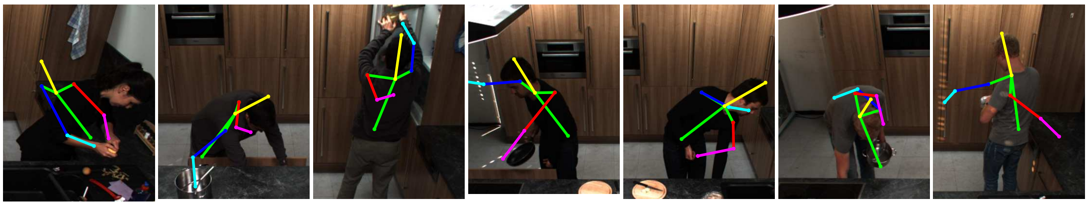


Figure 4.11: Examples of pose estimation failures on the MPIICooking dataset.

Chapter 5

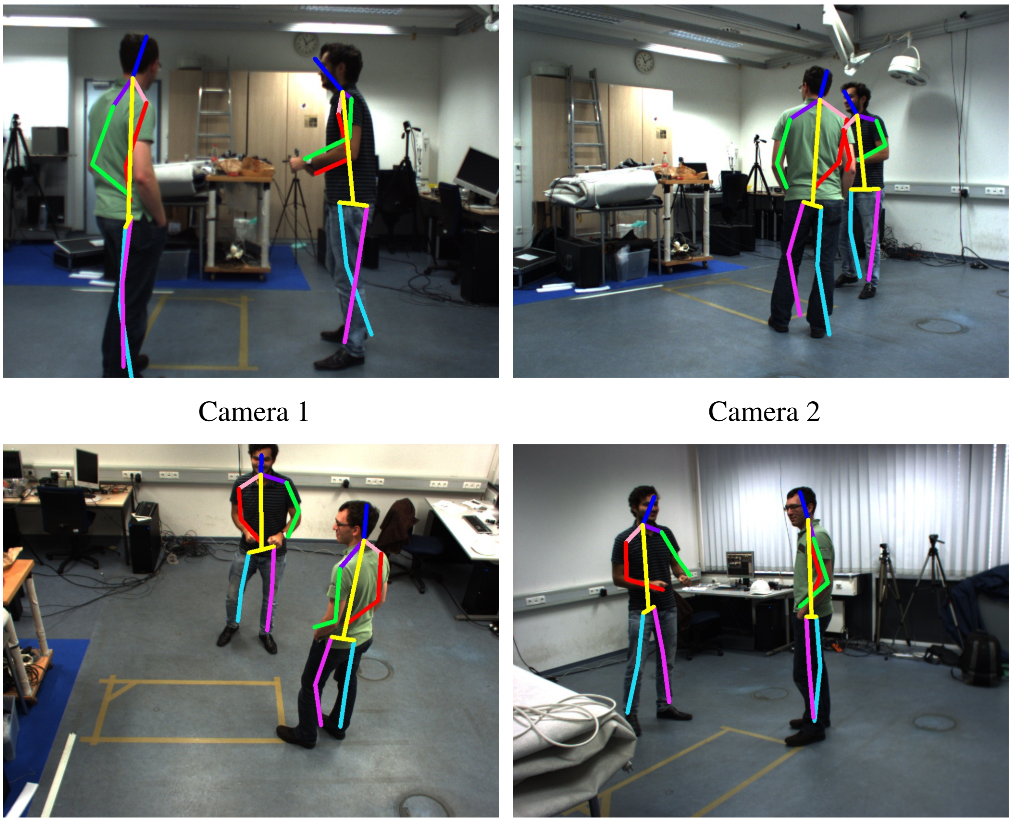
3D Pictorial Structures for Multiple Human Pose Estimation

In this chapter, we address the problem of 3D pose estimation of multiple humans from multiple views. Multiple human pose estimation is a more challenging problem due to a larger state space, person-person occlusions, and without having the prior knowledge about the across view human identities for data association. To address these challenges, we begin with generating a reduced state-space in 3D for the human body joints. To that end, we use a pre-trained 2D part detector and triangulate the corresponding body joints in pairs of camera views to obtain body joint hypotheses in 3D. Finally, to resolve the across view ambiguities and parse human poses in the reduced 3D state space of body joints, we introduce a novel 3D pictorial structures (3DPS) model. Our proposed model is efficient, generic, and applicable to both single and multiple human 3D pose estimation, which is important to be able to compare to the state-of-the-art.

We first evaluate our proposed method on single human 3D pose estimation benchmarks, such as HumanEva-I [127] and KTH Multiview Football Dataset II [26] datasets. Finally, we introduce and evaluate our approach on two datasets for multiple human 3D pose estimation. We provide the multi-human body-joint annotations in multiple views for both of these datasets, and benchmark our methods to inspire the research on multiple human 3D pose estimation.

5.1 Introduction

3D pose estimation of articulated objects especially humans is one of the most active area of research in computer vision. Human pose estimation Estimating the human body pose, in particular in 3D, opens up possibilities/opportunities for many vision-based applications such as human motion capture and analysis, and activity detection, which facilitates human-computer



Camera 1

Camera 2

Camera 3

Camera 4

Figure 5.1: **Shelf dataset**: Projections of our estimated 3D body poses across 4 out of 5 camera views, on our proposed multi-view multiple human dataset.

interaction. Several approaches have been proposed in the past to address the task of human pose estimation in 3D, based on different input modalities [6, 11, 26, 125, 131]. However, these approaches do not explicitly address the task of jointly estimating multiple human 3D poses from multiple views.

One of the major challenges of 3D world is the extent of its state-space. Therefore, most approaches begin with targeting the complexity of 3D space to make the 3D pose estimation feasible. A number of approaches such as [131] rely on background subtraction for that purpose. On the other hand, [26] assume fixed limb lengths and uniformly distributed rotations of body parts. Burenius et al. [26] also investigated the potential of using a discretized volume of 3D state-space for estimating the 3D body pose in order to avoid the continuous nature of 3D world. However in their analysis this still turned out to be an expensive task due to the six degrees of freedom (6 DoF) of each body part, especially for the level of discretization required to achieve reasonable results.

We however introduce a more efficient approach, which relies on pre-trained 2D part detectors of [6] to create a set of 2D part hypotheses in each view. We then triangulate the corresponding parts in pairs of views to generate a 3D state-space of body joints. This clearly makes our task much more simple. Since, we then only require to parse the human 3D poses in this reduced state space, instead of exploring a large continuous or discretized space of all possible translations and rotations.

Human pose estimation approaches such as [26, 131] suffer from another common problem of confusing left-right and front-back of the human body anatomy depending on camera pose. This problem excerbates in the presence of multiple humans in the scene due to the similar body parts of different humans. Such challenges of across view associations and mixing of multiple human body parts results in further ambiguities in the 3D pose estimation process. For example, a right leg of a person in one view could have several corresponding candidates in the other views coming from right and left leg of all the humans present in the scene and not only of the same human. Moreover there generally exist a large number of false positive part detections that make this task further challenging. Naively parsing a human pose in this case would lead to fake and implausible 3D skeletons in the space. The limitation of our approach is that the body parts need to be detected atleast in two views to be abe to triangulate the given 2D hypotheses and generate the 3D one.

In order to address this problem of ambiguities, we propose a novel 3D pictorial structures (3DPS) model that jointly infers poses of multiple humans from our reduced 3D state space of body joint hypotheses. The model our proposed 3DPS as a conditional random field (CRF) with multi-view and inter-part potential functions. The unary potentials in the CRF are computed based on the confidence of the per-view 2D part detectors and the reprojection error as introduced in our previous work [6]. Moreover, we also include unary potentials based on the length of the parts for resolving geometric ambiguities, and the visibility of the parts for modelling occlusions. The inter-part relations are modelled as pairwise potential functions. Furthermore, we also introduce inter-part pairwise collision potential to prevent part collisions in the 3D space. In the end, we perform the inference on our graphical model using loopy belief propagation. Finally, by sampling the marginal distributions we are able to infer the 3D human poses. Our model is generic and naturally also applicable to the single human 3D pose estimation, however we assume the count of humans to parse in the scene. For single human 3D pose estimation, our evaluations show that we obtain results on par with the state-of-the-art approaches [6, 26].

Our contributions: As the first and main contribution, we introduce a novel 3D pictorial structures model (3DPS) that enables joint 3D pose estimation of multiple humans in the scene. It is important to note that we do not assume the identity of humans across views rather this is determined as part of the inference procedure. Secondly, we propose an efficient approach to reduce the complex 3D state space and do not rely on the granularity of 3D state space discretization. To that end, we only assume the detections of 2D body joints are given for triangulation.

As the third and final contribution, we propose a new dataset, *i.e*. Shelf, with ground-truth annotations for the purpose of evaluating our proposed multiple human 3D pose estimation approach. In addition we also provide ground-truth annotations for multiple human poses, and evaluate our method on a standard Campus [19] dataset, which was originally introduced for the purpose of multi-view person detection and tracking using multiple cameras.

5.2 Related work

We refer the reader to the related work sections of chapters 1, 2, and 3 of this thesis for a detailed review of literature on 2D and 3D human pose estimation. Moreover, one can also have have a look at the review papers *e.g*. [81, 130]. In this section however, we would like to briefly point out a few most relevant works on 3D human pose estimation.

The literature can be mainly categorized in discriminative, generative, and pictorial structures based approaches. The discriminative approaches use image evidence (e.g. silhouettes, edges, or depth) to learn a mapping which classifies the between good or bad human poses [3, 50, 57, 125, 134, 137]. Since these methods are purely based on data (image evidence), hence require large number of training images or otherwise would not be able to model data variations (unknown poses) and noise. Therefore such methods are quite unreliable due to classification failures in the realistic settings. However, training with depth images stands out as an exception and typically generalizes for unknown poses [125]. The shortcoming of such depth-based methods is the sensor itself, since it is quite limited in range and does not provide reliable depth information outdoors.

On the other hand, the generative approaches (often called top-down approaches) rely on learning the distribution of the pose space. In these methods the task of pose estimation is typically coupled with tracking in order to get some reasonable 3D human poses estimates [22, 31, 42, 126, 137, 160]. Due to the dependency on tracking, these methods require pose initialization and continuous image sequences for pose prediction, and cannot recover once tracking fails. Moreover, most of such methods rely on motion models and are able to predict poses only for a learned type of motion.

Finally, there is another family of pose estimation approaches, so called pictorial structures or bottom-up approaches [11, 131]. Pictorial structures was originally introduced as a generic framework for object detection [37, 39]. Such approaches are a mixture of both generative and discriminative approaches For these approaches the human body is represented as a configuration of rigid parts. The representation of parts are learned discriminatively based on underlying image evidence (feature descriptors *e.g*. shape context, HOG, etc.). On the other hand, the parts are connected with one another in a graph using flexible pair-wise conenctions. Andriluka *et* *al*. [11] proposed a pictorial structures based approach by learning a mapping between 2D and 3D poses, but the approach is limited to specific and simple types of motions.

In the last few years, a number of approaches have been proposed that extend the 2D pictorial structure approach to 3D multi-view case. Approaches like [6, 26] are most relevant to our work in this chapter. Burenius *et al*. [26] showed that the pictorial structures approach can be used in the 3D space by discretizing the 3D state space. However, their analysis showed that the level of granularity, for the discretization of 3D space, required for a reasonable performance is still computationally quite expensive due to large degrees of freedom (6 DoFs) in translation and rotation of the body parts. Later building upon the same approach and employing better part detectors Kazemi et al. [68] achieved slightly better results with still the similar computational cost [26]. Extending such an approach for joint pose estimation of multiple humans becomes infeasible. Both of these papers evaluate only on a very limited football dataset with cropped player images and simple backgrounds, that they introduced themselves.

Finally, in our previous work on multi-view pictorial structures [6], we introduced several extensions to the pictorial structures framework such as, multi-view potentials, and 3D mixture models for an improved 2D pose estimation jointly in all camera views. The 3D pose is obtained by triangulation of 2D poses. We demonstrate impressive results on standard single human dataset,*i.e*. HumanEva-I [127]. However, the main drawback of that approach was the dependency of our novel multi-view potentials on the static camera setup. In contrast to this prior work, in this chapter we propose 3D pictorial structures (3DPS) model that operates in the 3D state space and hence independent of the camera setup.

5.3 Method

We model our 3D pictorial structures (3DPS) model as a conditional random field (CRF). First we describe the 3DPS model and the unary and pairwise potentials involved. After that we discuss our reduced state space for efficient model inference, and parsing of multiple humans. It is important to note that our proposed model is able to parse body poses of multiple humans whose body parts lie in a common 3D space.

5.3.1 3D pictorial structures model

Our 3D pictorial structure (3DPS) model represents the human body pose as a CRF of n random variables $Y_i \in \mathbf{Y}$ in an undirected graphical model (Fig. 5.2). The nodes of the graph or the random variables of the CRF correspond to a single body part. Two body parts are connected by an edge in the graph which describes the conditional dependence of these parts. The entire human

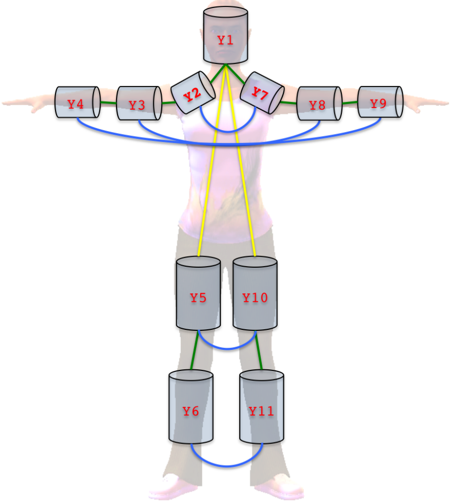


Figure 5.2: **Graphical model of the human body**: We use 11 nodes in our graphical model to represent the human body pose. We show the kinematic and collision constrains as pairwise edges in the graph. Here green for rotation, yellow for translation, and the blue for the collision constrains between symmetric parts only.

body pose can then be written as a configuration of parts in the 3D space $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. The individual variables Y_i describe the state of the body parts *i.e.* $Y_i = [\chi_i^{pr}, \chi_i^{di}]^T \in \mathbb{R}^6$. Here $\chi_i^{pr} \in \mathbb{R}^3$ represents the 3D position of the proximal joint, and $\chi_i^{di} \in \mathbb{R}^3$ represents the 3D position of the distal joint in the global coordinate system (Fig. 5.4). The posterior of the 3D pose output $\mathbf{y} \in \mathbf{Y}$ given the image observation $\mathbf{x} \in \mathbf{X}$ is given by,

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_i^n \phi_i^{conf}(y_i, \mathbf{x}) \cdot \prod_i^n \phi_i^{repr}(y_i, \mathbf{x}) \cdot \\ \prod_i^n \phi_i^{vis}(y_i, \mathbf{x}) \cdot \prod_i^n \phi_i^{len}(y_i, \mathbf{x}) \cdot \prod_{(i,j) \in E_{kin}} \psi_{i,j}^{tran}(y_i, y_j) \cdot \\ \prod_{(i,j) \in E_{kin}} \psi_{i,j}^{rot}(y_i, y_j) \cdot \prod_{(i,j) \in E_{col}} \psi_{i,j}^{col}(y_i, y_j) \end{aligned} \quad (5.1)$$

where $Z(\mathbf{x})$ is the partition function. $\phi_i^{conf}(y_i, \mathbf{x})$ is the unary potential given by the confidence of the 2D part detectors. $\phi_i^{repr}(y_i, \mathbf{x})$ is the unary potential based on the reprojection error, $\phi_i^{vis}(y_i, \mathbf{x})$ is another unary potential which tells the visibility of the body part in multiple views, and finally the unary potential $\phi_i^{len}(y_i, \mathbf{x})$ represents the geometric constraint on the part lengths. The pairwise potential functions encode the pairwise kinematic constraints between different parts. For symmetric body parts we have an extra pairwise constraint to avoid their collision in the 3D space.given by the potential function $\psi_{i,j}^{col}(y_i, y_j)$. E_{kin} represents the set of inter-part edges in the graph that model the kinematic constraints between the body parts, and E_{col} is the set of inter-part edges that model collision between the symmetric parts *i.e.* left and right limbs.

Next, we describe these unary and pairwise potential functions.

Unary potentials: In the following we briefly describe the unary potential functions used in our 3DPS model 5.1.

Detection confidence: Since a single part hypothesis in 3D space is infact a triangulation of 2D part detections in two views, therefore we take the mean confidence of the part detector in the two corresponding views to represent the detection confidence function $\phi_i^{conf}(y_i, \mathbf{x})$. This detection confidence is the most important potential function.

Reprojection error: The reprojection error tells the across-view geometric consistency of the 2D part hypotheses used for triangulation. Given accurate poses of cameras, a higher reprojection error indicates these 2D hypotheses do not correspond to the same point in 3D, therefore should be penalized. In previous work [6], we showed the effectiveness of the reprojection error based multi-view potential function in our CRF as a pairwise potential between parts of multiple views. In light of that experience, we keep this reprojection error based potential in the model, however as a unary term, since we are operating directly in a single 3D state space. Given 2D positions of the body joint hypotheses in two views *i.e.* $\mathbf{p}$ and $\mathbf{p}'$, the reprojection error as defined by [54] is,

$$C(x_i) = \|\mathbf{p}, \hat{\mathbf{p}}\|^2 + \|\mathbf{p}', \hat{\mathbf{p}}'\|^2 \tag{5.2}$$

where $\|.\|$ represents the euclidean distance, and $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}'$ are the projections of the hypotheses of the body joints in the two views. Finally, in order to express the reprojection error $C(x_i)$ as the score of a single 3D body joint hypothesis, a sigmoid function is employed.

$$\phi_i^{repr}(y_i, \mathbf{x}) = \frac{1}{1 + \exp(C(x_i))}. \tag{5.3}$$

Multi-view visibility potential: To take further advantage of the multi-view information, we introduce a visibility potential $\phi_i^{vis}(y_i, \mathbf{x})$ which describes the importance of a 3D hypothesis based on the number of observed views. To that end, we project the 3D hypothesis in all views and search if a corresponding 2D hypothesis is available in a neighborhood of 5 pixels radius. Finally, we normalize the number of observed views with respect to the total number of cameras. We observe that this potential is quite useful in order to disregard false positive hypotheses, which usually obtain a smaller visibility weight. It is important to note that this visibility term is complementary to the reprojection error term, since it contains information from all views.

Body part length: We model the typical length of the body parts as a separate unary potential function $\phi_i^{len}(y_i, \mathbf{x})$. We use the ground-truth data to compute the mean and standard deviation of the part lengths for all body parts. This potential function helps to penalize parts of unusual lengths. In a multiple human setting, these unusual parts appear often due to across-human ambiguities.

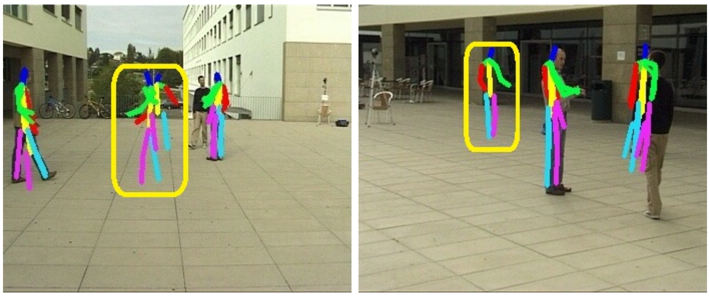


Figure 5.3: **Body parts state space**: Projections of our estimated 3D poses across views. One can also observe the presence of fake pose hypotheses (in yellow bounding boxes). Such fake poses may appear in rare cases, since we do not assume the identity of humans across views when generating our 3D state space.

Pairwise potentials Following, the state-of-the-art in 2D human pose estimation [11], we model the rotation and translation between different body parts in our graphical model using Gaussian distributions. Although limited due a single mode, gaussian distribution has been quite successful in modelling the inter-part pairwise relationships. Moreover following prior works, in our model the translation and rotation of a body part are given in the local coordinate system of the parent part.

For simplicity we assume separate priors for both part rotation and translation. Hence, we define the pairwise potentials separately for both transformations.

Rotation potential: We model the rotation only along one axis, therefore we can model it by a uni-variate Gaussian distribution as,

$$\psi_{i,j}^{rot}(y_i, y_j) = \mathcal{N}(y_{ij}^R \mid \mu_{ij}^R, \sigma_{ij}^R) \tag{5.4}$$

where μ_{ij}^R is the mean and σ_{ij}^R is the variance of the gaussian distribution.

Translation potential: On the other hand, the pairwise potential for translation needs to modelled as a multivariate Gaussian distribution.

$$\psi_{i,j}^{tran}(y_i, y_j) = \mathcal{N}(y_{ij}^T \mid \mu_{ij}^T, \Sigma_{ij}^T) \tag{5.5}$$

where μ_{ij}^T is the mean, and Σ_{ij}^T is the covariance learned using the ground-truth translation vectors.

To learn the distribution parameters for both rotation and translation potential functions, we use the ground-truth 3D pose data. Since we operate in local coordinate system of parts, we are not dependent on the camera setup. This allows us to learn our prior model on one camera setup and apply on another.

Collision potential: Furthermore, we also employ pairwise collision potentials which prevent the symmetric parts from colliding in the 3D space. Due to the presence of noise in the 2D part detections and especially the false positives of symmetric parts, we end up generating 3D part hypotheses of different body parts which overlap in the 3D space. Such part hypothesis are obviously not compatible with each other for the final pose estimate, because two body parts cannot take the same space in 3D. To that end, we define a body part as a pair of spheres, where the spheres are centered at the location of two joints of the part. We detect collision by computing sphere-to-sphere intersection [72]. In case the collision is detected the two part hypotheses are penalized with a constant δ.

$$\psi_{i,j}^{col}(y_i, y_j) = \delta \cdot inter(y_i, y_j) \tag{5.6}$$

where $inter(y_i, y_j) \in \{0, 1\}$ is the sphere-to-sphere intersection function, and 1 indicates collision detection.

Discrete state space One of the main advantage of our method is that we operate in a reduced state space of 3D body joints. This allows us not to worry about the level of granularity for discretization.In order to generate the reduced state space we rely on pre-trained 2D part detectors of [6], which gives the 2D part hypotheses in each view. We assume a calibrated camera system with c camera views. We then generate a 3D state space by triangulating the 2D part hypotheses for all view pairs. We know that the minimum number of views required to triangulate a 3D point is two. The intuition behind considering only pairs of camera views to generate the set of 3D hypotheses, is to be able to get a 3D part hypothesis even if the body part is occluded in some of the views. This also means our approach relies on this assumption that the body parts should be estimated correctly in atleast a pair of camera views by the 2D part detectors. Otherwise, we will not be able to estimate 3D pose correctly in case of a missing 3D part hypothesis. Moreover, using view-pairs for 3D state space generation, scales the state space by a factor of $\binom{c}{2}$, where c represents the number of cameras. But in our evaluations, we observe that the state space remains small enough for fast inference.

It is important to note that since we do not assume the identity of humans across views, we end up generating noisy 3D hypotheses due to trangulation of 2D part hypotheses of different humans. In some cases, these wrong 3D hypotheses could result in fake human skeletons in the 3D space where there is infact no human present. Such a case is demonstrated in Fig. 5.3.

5.3.2 Inference of multiple humans

Once we have computed all the potential functions for the hypotheses in the 3D state space, next step is to perform inference to estimate the posterior disctribution of Eq. 5.1. Typically,

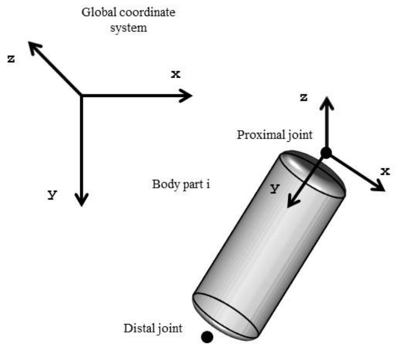


Figure 5.4: **Body part structure**: The body parts in our model are composed two joint positions, *i.e*. proximal and distal. The origin of the local coordinate system is at its proximal joint.

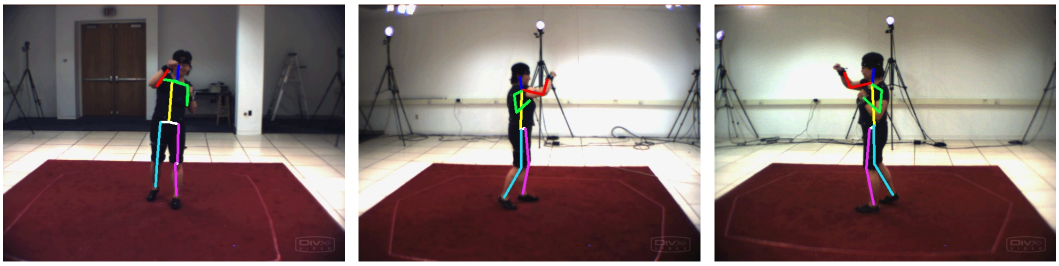
the inference becomes intractable for a graphical model like ours, which is a non-tree model (*i.e*. with loops). However, since we operate in a discrete state space, approximate inference could be employed. We use the loopy belief propagation algorithm of [20] for estimating the marginal distributions of the body parts. We assume that the number of humans in the scene is known. We then, sample the human body poses one by one from the estimated posterior distribution. Since our approach rely on the 2D part detectors, it is possible to miss some of the parts of the body pose in case we do not have any plausible 2D hypothesis for that part in atleast a pair of views. Therefore, in order to parse reasonable poses we allow our human 3D pose to lack body parts, in case some of the parts could not be obtained. Our proposed framework is also applicable for single human 3D pose estimation.

5.4 Experiments

We evaluate our approach on four different datasets. One of them, *i.e*. HumanEva-I [127], is the standard for single human 3D pose estimation. However, the dataset demonstrate controlled laboratory settings. We also evaluated our model on the KTH Multiview Football II [26] datasets which also contain only single well localized human in the images. We compare our results with two relevant multi-view approaches [6, 26]. Benchmarking our model on these single human datasets is important to compare to the state-of-the-art, and moreover to demonstrate the applicability of our method for single human pose estimation.

Since we want to perform 3D pose estimation of multiple humans ”in the wild”, and there does not exist a standard dataset with realistic settings, we ourselves introduce two novel datasets for this purpose, *i.e*. Shelf (Fig. 5.1) and Campus [19] (Fig. 5.7). Campus dataset was originally introduced in [19] for the task of detection and tracking of multiple humans from multiple calibrated camera views. We however annotated this dataset with human bosy pose in all views to be able to evaluate 3D pose estimation.

Our body model for the experiments consists of 11 body parts (Fig. 5.2). For each dataset, we use its training sequences to train our model's appearance terms, however we learn the body prior only once and use for all datasets. The body prior for the pairwise potentials are learned using the training set of the Campus dataset [19], and employed for evaluations on all datasets. For 2D part detection we rely on the detectors of our previous approach [6]. We learn the body prior for the pairwise potentials from a training subset of the Campus dataset [19] and use it during all the evaluations.

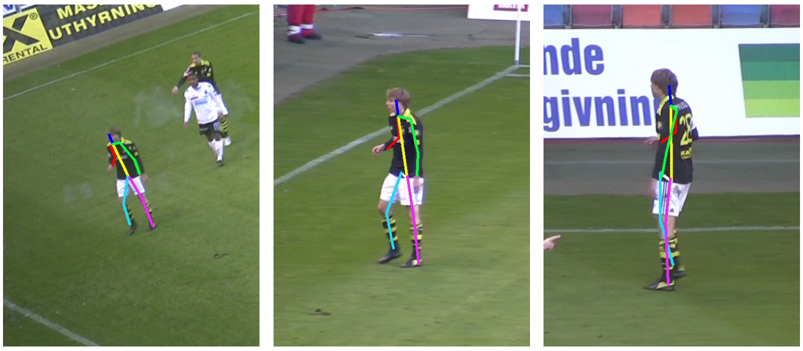


Camera 1 Camera 2 Camera 3

Figure 5.5: **HumanEva-I**: 2D projections of our estimated 3D pose across views for the Box sequence.

5.4.1 Single human pose estimation

We first demonstrate that our method performs on par to the state-of-the-art approaches [6, 26] for multi-view single human 3D pose estimation. Here we want to highlight that our approach is able to compete with other methods without the need to learn a dataset specific body prior.



Camera 1 Camera 2 Camera 3

Figure 5.6: **KTH Multiview Football II**: 2D projections of our estimated 3D pose across views for the player 2 sequence.

HumanEva-I: We compare our results to the approaches [6, 131] on the Box and Walking image sequences of HumanEva-I [127] dataset. We rely on the 2D appearance terms of [6], however employ our unique body model which is learned only once. In Table 5.1, we summarize the pose estimation results in terms of average 3D joint error in *mm*. We notice that, our original multi-view pictorial stuctures approach of [6] achieves very low 3D error, however with this current approach 3DPS we are able to achieve on par results without employing dataset specific body prior.

Sequence	Walking	Box
Amin et al. [6]	54.5	47.7
Sigal et al. [131]	89.7	-
Our method	68.3	62.7

Table 5.1: **Human-Eva I**: Average 3D joint error in millimetres (mm).

KTH Multiview Football II: We follow the same evaluation procedure as in the original work [26]. In this case, we evaluate in terms of percentage of correct parts (PCP), explained in detail in section 2.4. Unlike 3D error of HumanEva, a higher PCP value is better. In Table 5.2, we summarize the pose estimation results in PCP scores. Note that we outperform the approach of [26] for upto two cameras, however our method performs slightly worse with 3 cameras. On analysis we observe that failure to detect the part hypotheses in atleast two views is the main cause of failure in those cases especially for the legs, which are sometimes not quite clearly visible in the images due to motion blur. However, we still perform better than [26] for the detection of upper and lower arms. Overall for the full body pose, we obtain equivalent results with significantly less computations due to our reduced state space.

	Bur. [26]	Our	Bur. [26]	Our
Body Parts	C2	C2	C3	C3
Upper Arms	53	64	60	68
Lower Arms	28	50	35	56
Upper Legs	88	75	100	78
Lower Legs	82	66	90	70
All Parts (average)	62.7	63.8	71.2	68.0

Table 5.2: **KTH Multiview Football II**: The 3D PCP (percentage of correct parts) values of our method *versus* [26].

5.4.2 Multiple human datasets and pose estimation

As we discussed earlier, the task of articualted 3D pose estimation for multiple humans from multiple views has not been studied extensively in the literature. Moreover, unlike single human pose estimation there does not exist a standard benchmark dataset for multiple human pose estimation. To this end, in this work we introduce our own *Shelf* dataset. This dataset demonstrate a cooperative scenario for disassembling a shelf by two to four individuals at a time (Fig. 5.1).

The annotate the joints of all the humans in the scene for all camera views. Furthermore, we have also annotated the Campus dataset [19] with full body human poses for three humans in all views.

We also compare to our previous work Amin *et al*. [6] which employs single human pose estimation since it requires the humans to be localized in multiple views. This way our current method 3DPS is certainly at disadvantage, because for 3DPS we do not assume the across view identity of humans and perform inference in joint space where 3D part hypotheses of all humans lie.

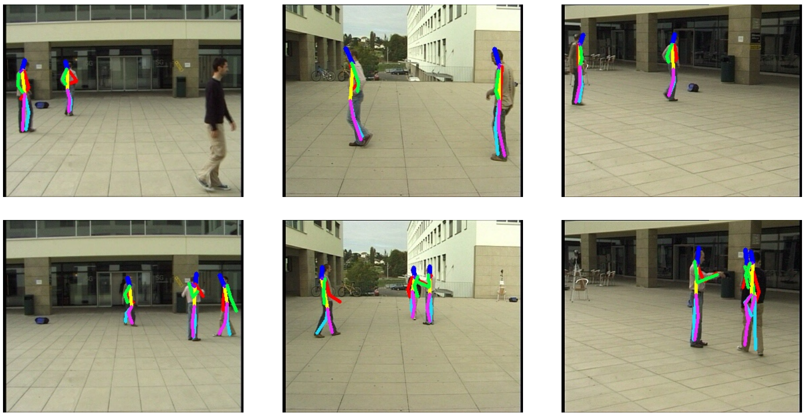


Figure 5.7: **Campus**: 2D projections of our estimated multiple human 3D poses across views.

Campus: In Table 5.3, we summarize the overall PCP results on Campus dataset. We see that our method which performs joint inference of multiple humans perform similar or better than our single human approach of [6]. First we assume that the identity of each human is known across views, and evaluate both our previous [6] and current (3DPS) methods. In Table 5.3, we show that we achieve similar results (*single human* column). Interestingly, we notice that the pose estimation performance stays similar even when we do not assume any across view human identity and perform infer multiple human poses from a joint 3D state space. This shows the ability of our model to handle multiple humans jointly.

Inference	Single Human		Multiple Human
	Amin et al. [6]	Our	Our
Actor 1	81	82	82
Actor 2	74	73	72
Actor 3	71	73	73
Average	75.3	76	75.6

Table 5.3: **Campus**: The 3D PCP (percentage of correct parts) values of our 3DPS method *versus* our MVPS method (Amin *et al*. [6])

Shelf: For our proposed Shelf[1] dataset, we also follow the same evaluation protocol as for Campus dataset. In Table 5.4, we observe that our pose estimation results on Shelf dataset are well in line with the results on Campus dataset, *i.e.* we perform similar or better for single human case, and do not lose any accuracy (in terms of PCP) for a more challenging task of multiple human pose estimation in a joint state space, where all humans lie together.

Inference	Single Human		Multiple Human
	Amin et al. [6]	Our	Our
Actor 1	65	66	66
Actor 2	62	65	65
Actor 3	81	83	83
Average	69.3	71.3	71.3

Table 5.4: **Shelf**: The 3D PCP (percentage of correct parts) values of our 3DPS method *versus* our MVPS method (Amin *et al.* [6])

5.5 Conclusion

In this chapter, we proposed a 3D pictorial structures (3DPS) model for the inference of multiple humans 3D pose estimation jointly, using multiple camera views. We proposed several multi-view potential functions in our model for the posterior distribution. We show that our proposed potential functions are able to handle self and person-person occlusions, and does not get confused by the mixing of body parts from multiple humans. We have shown that our model is generic, and therefore equally applicable for both single human and multiple human 3D pose estimation tasks. We also introduced an approach to perform efficient inference in the 3D space of body part hypotheses. As a future work, we want to reduce reliance of our 3DPS approach on single view 2D part detectors, since 3DPS cannot recover a 3D part hypothesis if 2D part detectors fail to fire in the first place. Finally, we introduced two new multiple human pose estimation datasets with full body pose annotations of humans in multiple views to stimulate active research on this challenging task.

[1] http://campar.cs.tum.edu/files/belagian/multihuman/Shelf.tar.bz2

Chapter 6

Conclusion and Future work

Computer vision as a field has grown significantly in the past 10-15 years. The state-of-the-art accuracy on several computer vision tasks has hit the acceptable levels, and industry has started to acknowledge its importance in the shaping of our near future. We are witnessing computer vision algorithms increasingly in our daily-use products. Although there is still much work to be done before this artificial vision could take over the human vision, but the idea of it is by no means an inconceivable task anymore. Keeping the recent progress in mind it is not at all ambitious to anticipate machines challenging humans on higher-level vision tasks, within just 10 years in the future. We discussed in the introduction of this thesis, that as an important milestone to realize such a future, the machines would need to reliably estimate and understand the human pose in real-time.

6.1 Summary and Discussion

Human pose estimation is necessary for several applications discussed in section 1.5. Our goal in this work has been to improve the human pose estimation in general settings and avoiding the complexity of inference in 3D state-space. During the course of this research work, we studied multiple applications to our 2D and 3D pose estimation approaches. For example, we proposed fine grained activity detection in [114], and study of emotions based on bodily expressions in [79].

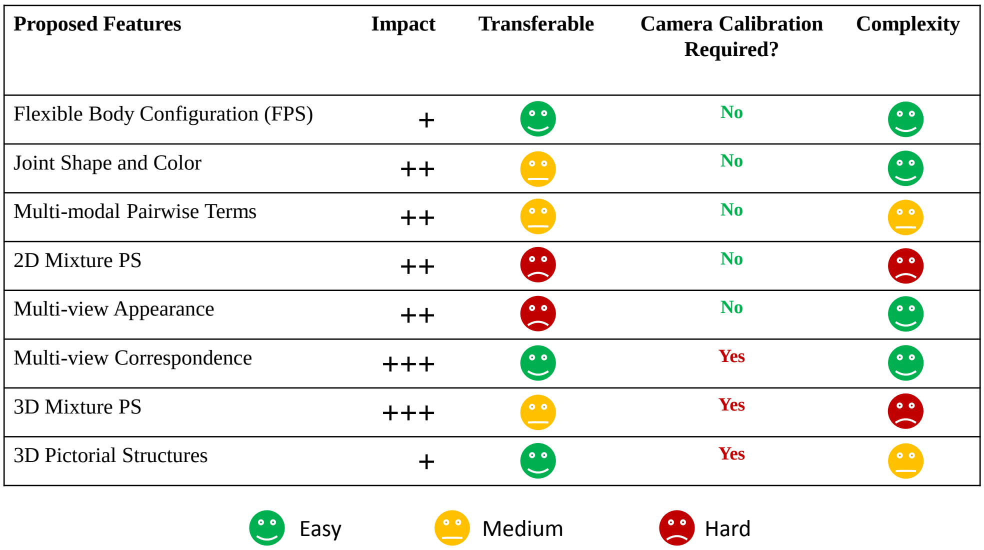
In section 6.2, we disuss some of the potential extentions and future perspectives of our research work on human pose estimation. Before we come to that, in the following we summarize our contributions in this work with a discussion about their benefits and limitations.

6.1.1 2D/3D Human pose estimation

In this thesis we proposed effective 2D and 3D human pose estimation methods for challenging real world settings with only limited amount of available labelled data for training of the models. We build upon the state-of-the-art pictorial structures framework and demonstrate significant progress on multiple pose estimation benchmarks with single or multiple humans. First in Chapter 2, inspired by the state-of-the-art, we proposed several advancements for the orignial pictorial structures approach of Andriluka *et al.* [12] and achieved a more expressive pictorial structures model which allowed us to improve the 2D pose estimation performance significantly for more complex human body layouts. We suggest the reader to look at the Fig. 2.7 for a summary of challenges that our 2D model overcame for the task of 2D pose estimation as compared to the baseline. The proposed extensions include, flexible part configuration, joint shape and color feature representation, multi-modal pairwise terms to model inter-part dependencies more effectively, and 2D mixture of pictorial structures to better represent the different viewpoints of the human body. In our evaluation on two different pose estimation benchmarks, we show that the proposed model extensions are complementary, in a sense that each of them solve a particular problem common with simple pictorial structures approaches.

Next, in Chapter 3 we introduced a novel multi-view pictorial structures approach that makes use of the evidence across views, to enable joint human pose estimation in several camera views and ultimately outputs the 3D human pose. To this end, we employed multi-view appearance and geometric correspondence conrtaints in the pictorial structures framework. We find in our evaluations that multi-view correspondence term which tries to find the geomterically consistent part hypotheses in all views, is the strongest component of our model. This result was as well expected since this requires the knowledge of camera calibration parameters. Interestingly, we observed that the multi-view appearance term, which looks for the part hypotheses that are consistent in appearance across views also performs well but in the simplistic settings *i.e.* HumanEva-I, mainly due to relatively cleaner backgound. Finally, we also show that in the multi-view settings we can employ 3D mixture of pictorial structures which improves considerably compared to simple 2D mixture models, due to their ability to impose geometrically consistent body priors even when performing inference separately in each view. This is true since for a 3D mixture model framework, for each view we employ the 2D body priors which are learned from the 2D projections of same set of ground-truth 3D poses.

We notice that one of the important features of our approach, *i.e.* multi-view appearance, is only applicable to the static camera setup. It is not straight forward to use this multi-view constraint in a setting different from the training set. This is certainly the most obvious disadvantage of our approach. The 3D mixture PS, which is another important feature in our framework for pose estimation, though is not directly transferable to another camera setup but can be used after re-learning of the model parameters. We base our argument on the fact that the learning



Proposed Features	**Impact**	**Transferable**	**Camera Calibration Required?**	**Complexity**
Flexible Body Configuration (FPS)	+	🙂	No	🙂
Joint Shape and Color	++	😐	No	🙂
Multi-modal Pairwise Terms	++	😐	No	😐
2D Mixture PS	++	☹	No	☹
Multi-view Appearance	++	☹	No	🙂
Multi-view Correspondence	+++	🙂	Yes	🙂
3D Mixture PS	+++	😐	Yes	☹
3D Pictorial Structures	+	🙂	Yes	😐

🙂 Easy 😐 Medium ☹ Hard

Figure 6.1: Summary of the proposed features with performance ratings in different aspects.

of generative body priors is quite fast, and in addition we only require to project 3D training poses in the different views of the new camera setup, before learning the parameters. On the other hand, multi-view correspondence feature is directly employable to any multi-view camera setup, since it does not require any parameters learning during model training, and only looks for the geometric consistency of the pose hypotheses at test-time.

In Fig. 6.1, we summarize the relative impact of several proposed features on the overall 2D and 3D pose estimation performance of our approach. We also list the dependence of the features on the camera calibration setup, ability to transfer to camera setups other than training setup, and complexity of the proposed feature in terms of implementation, and computation. We notice that the two most important features are multi-view correspondence and 3D mixture PS, both somehow require camera geomtery to work. Next, we also see that 3D pictorial structures is quite useful to parse multiple human poses jointly from a combined state-space of body parts for all humans, however it does not achieve a significant gain in terms of pose estimation performance. We notice that the multi-view correspondence constraints based on geometry are easily transferable and is relatively easier to implement, however require the knowledge of camera pose. Finally, Joint shape and color features learning is almost as good as employing the 2D mixture of pictorial structures model [65, 66], which is relatively difficult to realize.

Finally, we also proposed a 3D pictorial structures approach to infer the human poses directly in the 3D space. In Chapter 5, we discussed this approach in detail. We have shown that as the main advantage of such an approach, we are able to parse poses of multiple humans from a joint state space of 3D body parts of multiple humans. In addition to our formulation of 3D pictorial structures probabilitic framework, we demonstrated that the complexities of inference in the 3D space can be significantly avoided by employing inference in a reduced 3D state space

Dataset	HumanEva-I	MPII Cooking	Shelf
3D Error (*mm*)	55	178	155

Table 6.1: Difficulty comparison of different datasets in terms of mean 3D error in *mm*.

of multiple human body parts. We also showed that such a reduced state space can be efficiently generated using 2D part detectors and triangulation of hypotheses in pairs of camera-views.

6.1.2 Model adaptation

In Table 3.8, we show that our 3D pose estimation approach works great on HumanEva-I [127], and also achieve state-of-the-art results on more challenging datasets like MPII-Cooking (see Chapter 3), and Shelf (see Chapter 5). However, we observe a significant gap in the results of two regimes, *i.e*. HumanEva-I *versus* MPII Cooking and Shelf, in terms of 3D error in *mm*. In Table 6.1, we compare the 3D error (in mm), on all three datasets. We see that on HumanEva-I the mean 3D error is atleast three times lower as compared to the mean 3D error on other datasets. We observe that the dynamic nature of the MPII Cooking and Shelf datasets sets them apart; and a pre-trained model in such settings often misfires and thus fails to achieve results comparable to what we get in controlled settings like HumanEva-I.

Since we are operating in limited data domain, learning a generic model that captures all possible variations associated with a dynamically changing scene, is not yet possible. Therefore, in Chapter 4 we follow an alternate avenue and try to build upon the intuition that if we observe the same scene repeatedly, we should be able to learn something more specific about that scene and the persons in it.

To this end, we proposed two different approaches to utilize the scene specific evidence. In our evaluations we found that retraining part detectors for model adatation is not the best way to approach this problem. Infact, we showed that a simpler approach could lead to equivalent or better results than computationally expensive retraining approach. In this other approach we proposed to utilize the appearance similarity evidence that can be incorporated directly in our pictorial structures model as an additional unary term. This way we enable *online* model adaptation with respect to the already seen evidence. This evidence naturally comes from the selected pose estimates based on higher confidence score, we call *key-poses*, just as in the case of retraining approach. In our evaluations we have shown that the online model adaptation approach is though not as good as the offline one, but still improves consistently over time as it is expected to.

We also evaluated different approaches to mine these key-poses from the test-set (evaluation set without pose labels). We found that ensemble agreement of several independent models is the best cue to identify such key-poses. Besides, we concluded that 3D pose reconstruction based features perform generally well compared to the 2D features used in the literature.

However, there are several possible limitations of our approach. First of all, our proposed framework for selection of confident examples (key-poses), is currently limited to characterizing the single best pose estimate in every image, as a key-pose or not. This limits the achievable recall of our key-pose mining approach to the number of only correctly estimated poses on the complete dataset. We demonstrate in Fig. 4.5 and 4.6 that our approach is quite successful at scoring the correct pose estimates higher than the incorrect ones. Therefore, we could potentially extend this method to include multiple best pose estimates in the selection procedure, in turn improving the final pose estimation performance directly even before model refinement step. Second, in Eq. 4.6 the simple addition of scores coming from two separate approaches is certainly not ideal. One should as well consider the weights *i.e*. the effectiveness of each approach when combining different scores. We have also pointed out this issue in the dicussion during key-pose analysis (see Sec. 4.5.3). We have seen the performance in some cases degraded as a result of this combination. Moreover in our model adaptation framework, training ensemble of pictorial structures models, and learning discriminative pose quality parameters in a cross-validation fashion is a bit arduous.

6.2 Future perspectives

Tracking (ongoing work).
The biggest challenge in human motion analysis for a video sequence is to detect the person and estimate its initial body pose. We exactly addressed this challenge in this thesis, and proposed efficient and robust framework for human pose detection. The next important step in human motion analysis is to track the detected pose to obtain smooth pose trajectories, and build a complete track of the person motion. This task is typically addressed using temporal constraints in consecutive time frames. Human pose tracking has been addressed extensively in the literature [11, 34, 43, 45, 105, 107, 108, 109, 128, 135, 137, 139, 147, 160]. Besides, we have also proposed a robust SIFT-based tracking algorithm in our publication [114] to obtain part trajectories over time to study the task of pose-based activity recognition. We demonstrated that our part-tracking approach is quite robust, and is able to generate smooth trajectories over a neighborhood of 50 frames forward and backward. However, the proposed tracking approach works as the post-processing step for our human pose estimation algorithm.

Nevertheless, we are currently looking into ways to integrate our SIFT-based tracking approach directly in the pictorial structures framework (in the direction similar to [120]), also making use of the evidence from multiple views. We believe that incorporating temporal constraints directly in our multi-view pictorial structures framework would result in superior single-shot human motion capture over a fixed temporal neighborhood.

Body-language understanding.

Human body langauge is a rich source of information for emotion-related cues in typical social interactions. In past few years, activity recognition using human pose features in videos has emerged as an interesting computer vision challenge. However, characterization of bodily expressions to extract emotional information remains largely unexplored. Currently we are investigating the effectiveness of our 3D pose-based features, first to detect the distinct bodily expressions, and then to recognize the corresponding human emotions. We have published our initial results in [79].

3D pose estimation with uncalibrated camera setup.

In Chapter 3 we discussed our multi-view pictorial structures approach that requires camera parameters to enforce geometric correspondence constraints. In our evaluations we show that this geometric correspondence is the most important term in our model. Moreover, in the final step, to reconstruct the 3D pose from the 2D pose estimations we again require these camera parameters. While we achieve state-of-the-art results, but this dependency on camera geometry information limits the applicability of our 3D pose estimation approach to only a static camera setup or in case of camera motion we require continuous calibration information for each frame.

A more general solution would be to estimate the 3D human pose and the camera parameters jointly for the scene. This requires establishing strong across-view correspondences and minimizing some well-defined error *e.g.* reprojection error. Given the accurate enough correspondences, optimization for the camera parameters is well studied in the literature. SIFT matching is the de-facto standard for finding a large number of correspondences in multiple views for a stereo setup or in setup with continuous camera motion. Therefore, SIFT matching has been successfully employed in applications like simultaneous localization and mapping (SLAM) and structure from motion (SfM), etc. However, it has been also proven that the SIFT matching is not applicable for views with larger baselines [163], as typically the case in our multi-view setup. For rigid objects with well-defined 3D shapes (available 3D models, strong shape priors, etc.), quite reasonable results have been achieved in recent vision research [92, 93, 163]. However, it is certainly quite difficult for deformable or articualted objects like humans to obtain sufficiently many and accurate correspondences in muliple wide baseline views. We have done some initial experiments in establishing accurate correspondences across views using our multi-view appearance term. Although, we notice reasonable correspondences in simplistic settings like HumanEva-I but this approach does not perform well in relatively more dynamic settings *e.g.* MPII-Cooking.

Automatic pose annotations.

In Chapter 4 we proposed an approach that adapts the pose estimation model to the scene at hand. To this end, first we identify the confident examples from the test scene and then propose different strategies to use the potential evidence from such mined examples to improve the model over time. In the future we would like to generalize such an approach to multiple rounds of mining confident examples, and subsequent model refinement. The idea is that by repeating this process one can automatically obtain annotated training data from unlabeled images. However, we plan to explore and incorporate other approaches as well to allow for this automatic pose annotation in a more efficient and effective way.

Unsupervised human pose estimation.

Another possible research direction would be, learning to parse human poses in the absence of training pose annotations. Such an unsupervised technique, although quite ambitious, would certainly have a huge impact on general vision research. One could benefit from a very related unsupervised approach proposed by Brox and Malik [24] to segment objects in motion. Following this work, we could build and analyze the 2D/3D point trajectories for videos in short or long term using optical flow, in order to group the pixels belonging to individual rigid body parts.

Additional feature modalities.

In Chapter 2 we have seen significant improvement in estimating human poses using color as an additional feature in our model. Therefore, potentially in the future work we would also like to explore other effective cues to help boost the pose estimation performance. In the past few years, several works have been proposed which exploit temporal consistency and use optical flow very effectively as an additional cue for parsing human poses in video sequences [25, 40, 120]. Moreover, in past few years several methods have been proposed which use depth information to estimate 3D human poses, motivated by the work of [125].

Multi-view appearance using deep learning.

Recently deep learning framework has gained much popularity in computer vision due to its ability to learn optimal features specific to the task at hand. In addition to the state-of-the-art performance for generic object detection, deep learning has also started to show remarkable results for single and multiple humans 2D pose estimation in realistic settings [27, 61, 62, 86, 88, 103, 143]. Other approaches have also shown significant results for task of 3D human pose estimation [89, 138, 142]. The sudden rise of deep learning framework can be attributed to the significant growth in the amount of labeled data and impressive computational power available today in the form of high performance GPUs. In the future work, it would be interesting for us to investigate the potential of deep learning to learn multi-view appearance representations 3.3.2. Note that the un-calibrated multi-view camera setups could benefit given reasonable such representations.

Dense multi-view pictorial structures.

In our work we only considered multi-view pictorial structures for 3D human pose estimation in the reduced state-space to avoid computational complexity. We have shown that this approach works fine in the simplistic settings with cleaner backgrounds and limited number of subjects like HumanEva. But for more dynamic settings with cluttered backgrounds this certainly limits the potential of multiple views, since we ignore the across-view constraints in the initial inference step (see Sec. 3.3.4). We believe this idea should be explored in the future work if we want to take greater advantage of multi-view settings.

Real-time performance.

Andriluka *et al*. [12] pointed out in their work that the Gaussian convolutions during the initial inference procedure are the most time consuming steps. Since, these convolutions can be performed in parallel, therefore this step can be optimized with a GPU implementation. However, in the second step (see Sec. 3.3.4) we deal with a more complex graphical model that requires loopy belief propagation for inference. The computational efficiency in this case depends on the number of sampled part hypotheses. Therefore, the need is to explore faster ways to filter out incorrect part hypotheses before performing the inference step. This is something we plan to look into in the near future.

List of Figures

List of Tables

Publications

[1] Multi-view Pictorial Structures for 3D Human Pose Estimation.
Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, Bernt Schiele. 24th British Machine Vision Conference (BMVC), Bristol, UK, September 2013.

[2] Test-time Adaptation for 3D Human Pose Estimation.
Sikandar Amin, Philipp Müller, Andreas Bulling, Mykhaylo Andriluka. 36th German Conference on Pattern Recognition (GCPR), Münster, Germany, September 2014.

[3] 3D Pictorial Structures for Multiple Human Pose Estimation.
Vasileios Belagianis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, Slobodan Ilic. IEEE Conference on Computer Vision & Pattern Recognition (CVPR), Columbus, USA, June 2014.

[4] A Database for Fine Grained Activity Detection of Cooking Activities.
Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka and Bernt Schiele. IEEE Conference on Computer Vision & Pattern Recognition (CVPR), Providence, USA, June 2012.

[5] Emotion recognition from embedded bodily expressions and speech during dyadic interactions. Philipp Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. In Proc. of the 6th International Conference on Affective Computing and Intelligent Interaction (ACII), 2015.

[6] Geometric Proposals for Faster R-CNN [7].
Sikandar Amin, Fabio Galasso. IEEE Conference on Advanced Video and Signal based Surveillance (AVSS) 2017.

[7] UA-DETRAC 2017: Report of AVSS2017 & IWT4S Challenge on Advanced Traffic Monitoring [73]. IEEE Conference on Advanced Video and Signal based Surveillance (AVSS) 2017.

[8] Script Data for Attribute-based Recognition of Composite Activities.
Marcus Rohrbach, Michaela Regneri, Micha Andriluka, Sikandar Amin, Manfred Pinkal, Bernt Schiele. 12th European Conference on Computer Vision (ECCV), Springer, Oct. 2012.

[9] Coherent Multi-Sentence Video Description with Variable Level of Detail.
A. Senina, M. Rohrbach, W. Qiu, A. Friedrich, Sikandar Amin, M. Andriluka, M. Pinkal, B. Schiele. 36th German Conference on Pattern Recognition (GCPR), Münster, Germany, September 2014.

[10] A Distributed Many-Camera System for Multi-person Tracking, A modular, scalable and distributed Architecture for People Tracking in a Human Robot Collaboration Scenario.

C. Lenz, T. Röder, M. Eggers, Sikandar Amin, T. Kisler, B. Radig, G. Panin, A. Knoll. Joint Conference for Ambient Intelligence (AmI), November 2010.

[11] Robust People Detection and Tracking across Multiple Ceiling-Mounted Cameras.
Sikandar Amin. Institute for Media Technology, Technische Universität München.

Bibliography

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[2] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):44–58, Jan 2006.

[3] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *TPAMI*, 2006.

[4] L. T. Alessandro Bergamo. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Neural Information Processing Systems (NIPS)*, Dec 2010.

[5] R. D. Alexander Kasper, Zhixing Xue. The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. In *IJRR*, 2012.

[6] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *British Machine Vision Conference (BMVC)*, 2013.

[7] S. Amin and F. Galasso. Geometric proposals for faster r-cnn. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, August 2017.

[8] S. Amin, P. Müller, A. Bulling, and M. Andriluka. Test-time adaptation for 3D human pose estimation. In *German Conference on Pattern Recognition (GCPR)*, Münster, Germany, September 2014.

[9] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[10] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[11] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[12] M. Andriluka, S. Roth, and B. Schiele. Discriminative appearance models for pictorial structures. *IJCV*, 2011.

[13] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005.

[14] O. Aubert and Y. Prié. Advene: an open-source framework for integrating and visualising audiovisual metadata. In *ACM Multimedia*, 2007.

[15] M. Baktashmotlagh, M. Harandi, B. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 769–776, Dec 2013.

[16] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007.

[17] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, 2014.

[18] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, 2000.

[19] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *TPAMI*, 2011.

[20] C. M. Bishop et al. *Pattern recognition and machine learning*. springer New York, 2006.

[21] L. D. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV'09*, pages 1365–1372, 2009.

[22] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, 1998.

[23] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *IJCV*, 56(3):179–194, 2004.

[24] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *Computer Vision – ECCV 2010*, volume 6315 of *Lecture Notes in Computer Science*, pages 282–295. Springer Berlin Heidelberg, 2010.

[25] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-d pose tracking: Exploiting contour and flow based constraints. In *Computer Vision – ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 98–111. Springer Berlin Heidelberg, 2006.

[26] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013.

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] T. Captury. Markerless motion capture technology. http://www.thecaptury.com/.

[29] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, 2005.

[30] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[31] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61:185–205, 2005.

[32] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.

[33] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV*, 2012.

[34] A. El Hayek, C. Stoll, N. Hasler, K.-i. Kim, H.-P. Seidel, and C. Theobalt. Spatio-temporal motion tracking with unsynchronized cameras. In *CVPR*, 2012.

[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32, 2010.

[36] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Computer Vision and Pattern Recognition (CVPR)*, 2000.

[37] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[38] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[39] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[40] K. Fragkiadaki, H. Hu, and J. Shi. Pose from flow and flow from pose. In *CVPR*, 2013.

[41] J. Gall, N. Razavi, and L. Van Gool. An introduction to random forests for multi-class object detection. In *Outdoor and large-scale real-world scene analysis*, pages 243–263. Springer, 2012.

[42] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture: A multi-layer framework. *IJCV*, 87(1–2), 2010.

[43] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009.

[44] J. Gall, A. Yao, and L. J. V. Gool. 2D action recognition serves 3D human pose estimation. In *ECCV*, 2010.

[45] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. Gool. Articulated multibody tracking under egomotion. In *ECCV*, 2008.

[46] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation, September 2014.

[47] J. S. Goddard. Pose and motion estimation from vision using dual quaternion-based extended kalman filtering, 1997.

[48] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 222–230. JMLR Workshop and Conference Proceedings, 2013.

[49] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006, Nov 2011.

[50] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, 2003.

[51] D. Grest, T. Petersen, and V. Krüger. A comparison of iterative 2d-3d pose estimation methods for real-time applications. In *Image Analysis*, volume 5575 of *Lecture Notes in Computer Science*, pages 706–715. Springer Berlin Heidelberg, 2009.

[52] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image, 2009.

[53] A. Halevina and N. F. Troje. Sex classification of point-light walkers: Viewpoint, structure, kinematics. *Journal of Vision*, 7(9):483, 2007.

[54] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[55] N. Hasler, B. Rosenhahn, T. Thormaehlen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, 2009.

[56] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[57] M. Hofmann and D. Gavrila. Multi-view 3d human pose estimation in complex environment. *IJCV*, 2012.

[58] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983.

[59] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*, 2016.

[60] C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In *Computer Vision, 2011. ICCV 2011. IEEE 11th International Conference on*, 2011.

[61] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. In *International Conference on Learning Representations (ICLR)*, April 2014.

[62] A. Jain, J. Tompson, Y. LeCun, and C. Bregler. Modeep: A deep learning framework using motion features for human pose estimation, September 2014.

[63] N. Jammalamadaka, A. Zisserman, M. Eichner, V. Ferrari, and C. V. Jawahar. Has my algorithm succeeded? an evaluator for human pose estimators. In *ECCV*, 2012.

[64] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proc. ICCV*, pages 3192–3199, 2013.

[65] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[66] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[67] A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen. Gesture-based affective computing on motion capture data. In *Proc. of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2005.

[68] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *BMVC*, 2013.

[69] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[70] I. Laptev and P. Pérez. Retrieving actions in movies. In *ICCV*, 2007.

[71] C. Lenz, T. Röder, M. Eggers, S. Amin, T. Kisler, B. Radig, G. Panin, and A. Knoll. A distributed many-camera system for multi-person tracking. In *Ambient Intelligence*, volume 6439 of *Lecture Notes in Computer Science*, pages 217–226. Springer Berlin Heidelberg, 2010.

[72] M. Lin and S. Gottschalk. Collision detection between geometric models: A survey. In *Proc. of IMA Conference on Mathematics of Surfaces*, 1998.

[73] S. Lyu, M. C. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. D. Coco, P. Carcagnì, D. Anisimov, E. Bochinski, F. Galasso, F. Bunyak, G. Han, H. Ye, H. Wang, K. Palaniappan, K. Ozcan, L. Wang, L. Wang, M. Lauer, N. Watcharapinchai, N. Song, N. M. Al-Shakarji, S. Wang, S. Amin, S. Rujikietgumjorn, T. Khanova, T. Sikora, T. Kutschbach, V. Eiselein, W. Tian, X. Xue, X. Yu, Y. Lu, Y. Zheng, Y. Huang, and Y. Zhang. Ua-detrac 2017: Report of avss2017 iwt4s challenge on advanced traffic monitoring. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7, August 2017.

[74] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[75] K. Matzen and N. Snavely. Nyc3dcars: A dataset of 3d vehicles in geographic context. In *Proc. Int. Conf. on Computer Vision*, 2013.

[76] M. D. Meijer. The contribution of general features of body movement to the attribution of emotions. In *Journal of Nonverbal behavior*, 1989.

[77] J. Michalak, N. F. Troje, J. Fischer, P. Vollmar, T. Heidenreich, and D. Schulte. Embodiment of sadness and depression—gait patterns associated with dysphoric mood, June 2009.

[78] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, Oct 2005.

[79] P. Müller, S. Amin, P. Verma, M. Andriluka, and A. Bulling. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *Proc. of the 6th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.

[80] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

[81] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 2006.

[82] P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *CVPR*, 2008.

[83] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *WACV*, 2013.

[84] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[85] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-2(6):522–536, Nov 1980.

[86] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[87] S. J. Pan, I. Tsang, J. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2):199–210, Feb 2011.

[88] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy. Towards accurate multi-person pose estimation in the wild. *CoRR*, 2017.

[89] S. Park, J. Hwang, and N. Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *European Conference on Computer Vision Workshop (ECCV)*, 2016.

[90] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. In *BMVC*, 2010.

[91] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:677–695, 1997.

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

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

[94] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 588–595, 2013.

[95] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[96] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[97] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[98] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford. Perceiving affect from arm movement. In *Cognition*, 2001.

[99] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Mueller, H.-P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV*, 2011.

[100] S. Prince. *Computer Vision: Models, Learning and Inference*. Cambridge University Press, 2012.

[101] Qualisys. Motion capture system. http://www.qualisys.com/.

[102] I. Radwan, A. Dhall, and R. Goecke. Monocular image 3d human pose estimation under self-occlusion. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1888–1895, Dec 2013.

[103] U. Rafi, I.Kostrikov, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2016.

[104] V. Ramakrishna, T. Kanade, and Y. A. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision, 2012*, October 2012.

[105] D. Ramanan. Learning probabilistic non-linear latent variable models for tracking complex activities. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[106] D. Ramanan. Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[107] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Computer Vision and Pattern Recognition (CVPR)*, 2003.

[108] D. Ramanan and D. A. Forsyth. Tracking people by learning their appearance. *PAMI*, 2007.

[109] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, 2005.

[110] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 824–831 Vol. 1, Oct 2005.

[111] D. Roetenberg, H. Luinge, and P. Slycke. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors, 2013.

[112] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr. Randomized trees for human pose detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[113] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94 – 115, 1994.

[114] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[115] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–39, 2010.

[116] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV*, 2002.

[117] B. Rosenhahn. Pose estimation revisited, 2003.

[118] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimations. In *CVPR*, 2013.

[119] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[120] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[121] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.

[122] S. Savarese and L. Fei-Fei. Multi-view object categorization and pose estimation. In *Computer Vision*, volume 285, pages 205–231. Springer Berlin Heidelberg, 2010.

[123] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[124] H. Shen, S.-I. Yu, Y. Yang, D. Meng, and A. Hauptmann. Unsupervised video adaptation for parsing human motion. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 347–360. Springer International Publishing, 2014.

[125] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[126] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. *ECCV*, 2000.

[127] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2), 2010.

[128] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, 2004.

[129] L. Sigal and M. J. Black. Predicting 3D people from 2D pictures. In *AMDO*, 2006.

[130] L. Sigal and M. J. Black. Guest editorial: state of the art in image-and video-based human pose and motion estimation. *IJCV*, 2010.

[131] L. Sigal, M. Isard, H. Haussecker, and M. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *IJCV*, 2011.

[132] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. *CVPR*, 2013.

[133] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[134] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *CVPR*, 2005.

[135] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *ICCV*, 2011.

[136] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *International Conference on Computer Vision (ICCV)*, ICCV '11, pages 723–730, Washington, DC, USA, 2011. IEEE Computer Society.

[137] G. Taylor, L. Sigal, D. Fleet, and G. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010.

[138] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016.

[139] M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE Workshop on THEMIS*, 2009.

[140] I. M. Thornton, D. W. Cunningham, N. F. Troje, and H. H. Bülthoff. "you can tell by the way i use my walk…": New studies of gender and gait. *Journal of Vision*, 1(3):354, 2001.

[141] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, 2010.

[142] D. Tomè, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CoRR*, 2017.

[143] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[144] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, pages 227–240. Springer Berlin Heidelberg, 2010.

[145] N. F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5), 2002.

[146] R. Urtasun and T. Darrell. Local probabilistic regression for activity-independent human pose inference. In *Computer Vision, 2009. ICCV 2009. IEEE 11th International Conference on*, 2009.

[147] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006.

[148] Vicon. Tracker; oxford metrics, oxford, uk. http://www.vicon.com/.

[149] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[150] H. G. Wallbott. Bodily expression of emotion. In *European journal of social psychology*, 1998.

[151] M. Wang, W. Li, and X. Wang. Transferring a generic pedestrian detector towards specific scenes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3274–3281, June 2012.

[152] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3401–3408, June 2011.

[153] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '11, pages 1705–1712, Washington, DC, USA, 2011. IEEE Computer Society.

[154] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[155] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[156] J. Xu, S. Ramos, D. Vazquez, and A. Lopez. Domain adaptation of deformable part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(12):2367–2380, Dec 2014.

[157] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[158] A. Yao, J. Gall, G. Fanelli, and L. V. Gool. Does human action recognition benefit from pose estimation? In *BMVC*, 2011.

[159] A. Yao, J. Gall, and L. V. Gool. Coupled action recognition and pose estimation from multiple views. *IJCV*, 2012.

[160] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *NIPS*, 2011.

[161] J. S. Yuan, Z. C. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, 2009.

[162] Z. Zhang and N. F. Troje. View-independent person identification from human gait. *Neurocomputing*, 69(1–3):250 – 256, 2005. Neural Networks in Signal Processing 2003 {IEEE} International Workshop on Neural Networks for Signal Processing.

[163] Z. Zia, M. Stark, K. Schindler, and B. Schiele. Revisiting 3d geometric models for accurate object shape and pose. In *3rd International IEEE Workshop on 3D Representation and Recognition (3dRR, ICCV Workshop)*, November 2011.

[164] S. Zuffi, O. Freifeld, and M. J. Black. From pictorial structures to deformable structures. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3546–3553. IEEE, June 2012.