Multi-view Pictorial Structures for 3D Human Pose Estimation

Sikandar Amin[1]
sikandar.amin@in.tum.de
Mykhaylo Andriluka[2]
andriluka@mpi-inf.mpg.de
Marcus Rohrbach[2]
rohrbach@mpi-inf.mpg.de
Bernt Schiele[2]
schiele@mpi-inf.mpg.de

[1] Intelligent Autonomous Systems
Technische Universität München
München, Germany

[2] Computer Vision and
Multimodal Computing
Max Planck Institute for Informatics
Saarbrücken, Germany

Abstract

Pictorial structure models are the de facto standard for 2D human pose estimation. Numerous refinements and improvements have been proposed such as discriminatively trained body part detectors, flexible body models, and local and global mixtures. While these techniques allow to achieve state-of-the-art performance for 2D pose estimation, they have not yet been extended to enable pose estimation in 3D. This paper thus proposes a multi-view pictorial structures model that builds on recent advances in 2D pose estimation and incorporates evidence across multiple viewpoints to allow for robust 3D pose estimation. We evaluate our multi-view pictorial structures approach on the HumanEva-I and MPII Cooking dataset. In comparison to related work for 3D pose estimation our approach achieves similar or better results while operating on single-frames only and not relying on activity specific motion models or tracking. Notably, our approach outperforms state-of-the-art for activities with more complex motions.

1 Introduction

In this paper we consider the task of articulated 3D human pose estimation from multiple calibrated cameras. Traditionally this task is addressed using 3D body models [4, 7, 15, 32] and involves complex inference in a high-dimensional space of 3D body configurations. Various mechanisms such as annealed particle filtering [7] or non-parametric belief propagation [32] have been proposed to address the search complexity. In this paper we argue that the search complexity can be reduced significantly by formulating the 3D inference problem as a joint inference over 2D projections of the pose in each of the camera views. To that end we build on the success of 2D pictorial structure models [21, 22, 36] that were shown to be effective for 2D human pose estimation. Reasoning purely in 2D allows to delay resolving 2D to 3D lifting ambiguities until the point when all image observations are taken into account. This is in contrast to approaches based on 3D body models that need to hypothesize 3D poses early in the inference process.

© 2013. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

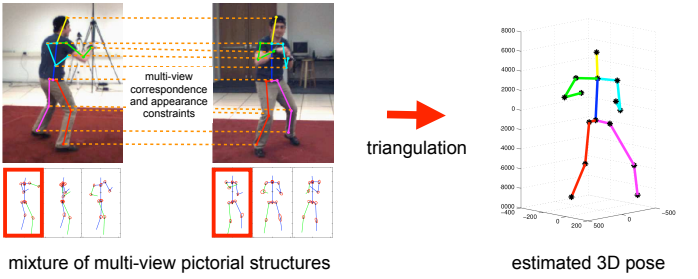


Figure 1: Our approach. (1) Projections of 3D pose in each view are jointly inferred using a mixture of multi-view pictorial structures models. The body layout priors of each mixture component are visualized below, activated components are highlighted in red. (2) 3D pose is recovered via triangulation.

Representing 3D pose as a collection of 2D projections allows to directly tap into recent literature on articulated 2D pose estimation. Following the state-of-the-art in this area we capture the appearance of 2D body projections using discriminatively-trained representations based on color and local gradient histograms that have been demonstrated to be robust against background clutter and variations in people appearance. This allows to apply our model to scenes with non-static backgrounds and estimate poses for a variety of people without adapting the model. Similarly to recent 2D pose estimation methods our approach is able to densely scan the space of all image locations and 2D body configurations in all views. This is in contrast to 3D models that typically rely on stochastic search and require initialization and temporal filtering to perform well [15, 34, 37].

As a first contribution of this paper, we propose a 2D pose estimation approach that extends our state-of-the-art 2D pictorial structures model [22] with color features and more effective spatial terms. Also, we generalize [22] to a mixture model and propose a novel approach for mixture component selection. The second and main contribution is to extend this 2D pose estimation model to a multi-view model that performs joint reasoning over people poses seen from multiple viewpoints. The output of this novel model is then used to recover 3D pose. We evaluate our approach on the HumanEva-I [30] dataset which is a standard benchmark for multi-view 3D pose estimation, and on our MPII Cooking [27] dataset which has an order of magnitude more activities than HumanEva-I and includes a significantly larger number of human subjects. On HumanEva-I our approach achieves accuracy on par or better than recent results from the literature [34, 37] that rely on activity-specific motion models and tracking, whereas our approach operates on single-frames only. On MPII Cooking our approach improves over our 2D approach [27] by a large margin demonstrating the advantages of jointly estimating pose across multiple views.

Related work. Articulated 3D human pose estimation has been considered in the literature in a variety of settings. Much recent work has focused on markerless motion capture and has been mainly applied to images acquired in controlled laboratory conditions [30]. Recent methods in this domain are typically relying on detailed body models [2, 15] and elaborate optimization strategies based on stochastic search [15, 32], local optimization [33], or a combination thereof [14]. Although impressive results have been obtained in this setting [16, 33, 35], the developed approaches appear to have difficulties to generalize beyond the motion capture domain. Furthermore, applications such as activity recognition, visual surveillance or mobile navigation typically rely on specialized pose estimation methods [17, 26, 27]. Various efforts have been made to bridge the gap between laboratory and real-world settings by relying on activity specific pose and motion priors [16], structure-

from-motion methods [23] or by combining visual observations with on-body inertial sensors [25]. These approaches build on techniques developed for controlled laboratory environments and extend them to more realistic settings. In this paper we pursue an alternative strategy. As a foundation for our approach we rely on our pictorial structures model [22] that has been shown to work for images of realistic complexity. We extend this model to incorporate recent improvements from the 2D pose estimation literature [9, 21, 36] and generalize it to multi-view to enable 3D pose estimation. Our approach is related to [1, 31] who also rely on pictorial structures to generate evidence for 3D pose estimation and to [5] who propose a pictorial structures model defined in 3D. Compared to [5] we rely on a more economic 2D representation, but require a loopy model to incorporate multi-view constraints. In contrast to our work, the approaches of [1, 31] have been formulated for the monocular setting and do not perform reasoning across viewpoints as we do here. We consider the more general task of 3D pose estimation of people across a variety of activities whereas our prior approach [1] is limited to pose estimation of walking people only and relies on activity specific motion priors and tracking. The approach of [31] builds on a simple observation model based on silhouette and color features and has only been evaluated in a laboratory setting. In contrast our approach builds on state-of-the-art appearance and spatial representations making it applicable to both, the laboratory setting and the MPII Cooking dataset that includes a variety of human subjects involved in a diverse set of activities, interacting with a large set of objects.

2 Single-view model

In the following we describe our approach to 2D pose estimation that relies on the pictorial structures model. We build on our formulation in [22] and extend it with several improvements motivated by related work to boost performance. Specifically, we introduce a *more flexible part configuration* and *multi-modal pairwise terms* [28, 36], *color features* [9], and *mixtures of pictorial structures* [20].

2.1 Pictorial structures model

The pictorial structures model, originally introduced in [11, 13], represents the human body 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 $l_i = (x_i, y_i, \theta_i)$, where (x_i, y_i) is the image position of the part, and θ_i is the absolute orientation. We formulate the model as a conditional random field, and assume that the probability of the part configuration L given the image evidence I factorizes into a product of unary and pairwise terms:

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

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 [6, 11, 22].

Flexible pictorial structures (FPS). We build on our publicly available implementation of pictorial structures [22], which consists of 10 parts that correspond to left/right body limbs, torso and head. We slightly change the part composition. 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 due to out-of-plane rotation [27, 28, 36]. Our new model has 14 parts that correspond to torso, head, as well as left and right wrist, elbow, shoulder, ankle, knee, and hip. For the MPII Cooking dataset we only use the 10 upper body parts as in [27].

2.2 Appearance representation

The part likelihood terms are represented with boosted part detectors that rely on the encoding of the image using a densely computed grid of shape context descriptors [3]. The feature vector is formed by concatenating the descriptors inside the part bounding box. Then the part likelihood term is given by $f(l_i;I) = \max\left(\sum_t \alpha_{i,t} h_{i,t}(e_i(l_i))/\sum_t \alpha_{i,t}, \varepsilon_0\right)$, where $e_i(l_i)$ is the feature vector corresponding to part i extracted at location l_i, $h_{i,t}$ are weak single-feature classifiers, and $\alpha_{i,t}$ are their corresponding weights learned with AdaBoost.

Color. We augment the shape context features 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 [8]. Additionally, certain colors are more likely to correspond to background than to one of the body parts [9]. For this we encode the color of the part bounding box using a multi-dimensional histogram of 10 bins for each of the dimensions of the RGB color space, which 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 Spatial model

The pairwise terms in Eq. 1 encode the spatial constraints between model parts and are modeled with a Gaussian distribution in the transformed space of the joint between two parts:

$$f_{ij}(l_i,l_j) = \mathcal{N}(T_{ji}(l_i) - T_{ij}(l_j)|\mu^{ij},\Sigma^{ij}), \tag{2}$$

where T_{ij} 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.

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 with term that maximizes over K modes and represent each mode with a Gaussian. The new multi-modal pairwise term is then given by:

$$f_{ij}(l_i,l_j) = \max_{k=1}^{K} \mathcal{N}(T^k_{ji}(l_i) - T^k_{ij}(l_j)|\mu^k_{ij},\Sigma^k_{ij}). \tag{3}$$

Note that this pairwise term is similar to the one used in [36], but has a somewhat different form as it incorporates both relative orientation and position of model parts whereas in [36] only the relative position is modeled and orientation is represented via an additional latent variable.

2.4 Mixtures of pictorial structures (Mixture PS)

Following [20, 21] we extend our approach to a mixture of pictorial structures models. We obtain the mixture components by 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. 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. 1 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 [24] who jointly solve the tasks of object detection and viewpoint classification. This approach is similar to DPM [10], but relies on a structured prediction formulation that encourages both correct localization and component detection. When applying [24] 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 [19] 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 $s(k,I) = \sum_{n=1}^{N} \|\mathrm{Cov}_n(k,I)\|_2$, where $\mathrm{Cov}_n(k,I)$ 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 $\hat{k} = \mathrm{argmin}_k s(k,I)$.

3 Multi-view model

In this section we describe our approach to 3D pose estimation, which consists of two steps. In the first step we jointly estimate the 2D projections of the 3D body joints in each view. As a basic tool for the representation and inference of the projected human pose we rely on the 2D model introduced in Sec. 2. In the second step, we use the estimated 2D projections and recover the 3D pose by triangulation [18].

For the sake of clarity we first present the multi-view model for the case of two views. We denote the 2D body configuration as L_m and the image evidence as I_m for view m. Similar to Eq. 1 the conditional posterior over body configurations in two views decomposes into a product of unary and pairwise terms. These define appearance and spatial constraints between parts independently for each view. In addition we introduce pairwise factors between every pair of corresponding parts in each view (see Fig. 1). The joint posterior over configurations in both views is given by

$$p(L_1,L_2|I_1,I_2) = \frac{1}{Z} f(L_1;I_1) f(L_2;I_2) \prod_n f_n^{app}(\mathbf{l}_n^1,\mathbf{l}_n^2;I_1,I_2) f_n^{cor}(\mathbf{l}_n^1,\mathbf{l}_n^2), \tag{4}$$

where $f(L_1;I_1)$ and $f(L_2;I_2)$ correspond to single-view factors and decompose into products of unary and pairwise terms according to Eq. 1. When more than two views are available we connect the corresponding 2D body parts in all pairs of views. The posterior in Eq. 4 then includes multi-view appearance and correspondence factors for each pair of connected parts in all views as well as within-view spatial and appearance factors. In the following we define the multi-view pairwise factors that encode appearance and correspondence constraints.

Multi-view appearance. The factor f_n^{app} 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 multiple views $e_n(\mathbf{l}_n^1,\mathbf{l}_n^2) = [e_n(\mathbf{l}_n^1), e_n(\mathbf{l}_n^2)]$ and train a boosted part detector using this representation. The appearance factor now depends on the locations of the part in each view. 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.

Multi-view correspondence. The factor f_n^{cor} encodes the constraint that part locations in each view should agree on the same 3D position. Given a pair of corresponding part locations $\mathbf{l}_n^1$ and $\mathbf{l}_n^2$ we first reconstruct the corresponding position of the part in 3D using linear triangulation [18]. The multi-view correspondence factor is then given by

$$f_n^{cor}(\mathbf{l}_n^1,\mathbf{l}_n^2) = \exp(-(\|\mathbf{l}_n^1 - \hat{\mathbf{l}}_n^1\|^2 + \|\mathbf{l}_n^2 - \hat{\mathbf{l}}_n^2\|^2)), \tag{5}$$

where $\hat{\mathbf{l}}_n^1$ and $\hat{\mathbf{l}}_n^2$ denote the projections of the reconstructed 3D point in each view.

3D mixture model. As in Sec. 2 our multi-view model employs mixtures of pictorial structures to represent 2D body configurations per view. However, in the multi-view case the mixture components correspond to groups of poses similar in 3D. In order to obtain such 3D mixture components we first cluster the 3D training poses with k-means. We then project the training data of each 3D cluster and learn 2D models from the projected data. We visualize the components learned on the boxing data in Fig. 1. Note that the resulting 2D mixture components are consistent across views by construction as they are learned from the projections of the same 3D poses. We exploit this fact by jointly selecting the best mixture component across all views and adapt the component selection procedure introduced in Sec. 2 accordingly. For the component detector we add the scores of the corresponding components across all views. For the uncertainty based criteria we add the uncertainty scores $s(k,I)$ defined in Section 2.3 for each of the corresponding components across all views.

Inference. The pictorial structures approach allows efficient and exact inference under the simplifying assumptions that the pairwise part dependencies have a tree structure and can be represented by Gaussian distributions. However, these assumptions limit the expressiveness of the model. For example the pairwise factors in Eq. 3 as well as the multi-view factors in Eq. 4 are not Gaussian and create loopy dependencies in the model structure. To perform inference with non-Gaussian factors and loopy model we rely on the approximate two-stage inference procedure introduced in [22]. In the first stage this procedure relies on the simplified tree-structured model with Gaussian pairwise factors (cf. Eq. 2) and simple shape and color appearance terms in order to generate proposals for body part locations. This stage can be seen as a search-space reduction step that is necessary in order to apply a more complex model [12, 29]. The inference in the first stage is performed with sum-product belief propagation which allows to compute marginal distribution for each body part $p(l_i|I)$. The inference is exact and efficient because the model is tree structured and messages can be computed with Gaussian convolutions. In the second stage we sample a sufficiently large set of locations from $p(l_i|I)$[1] and perform inference in the full model with all factors in the reduced state-space of the sampled part locations. Here we use the max-product belief propagation as it allows to obtain a consistent estimate for the whole body configuration. Finally, given the 2D projections estimated by the multi-view pictorial structures model we reconstruct the 3D pose using triangulation.

[1] In all experiments we sample 1,000 locations for each part and remove the duplicates. We found this number to be sufficient and increasing the number of sampled locations did not have significant influence on the final results.

Part likelihood terms	walking	box
FPS	114.3	77.4
+ mixtures PS	83.2	69.3
+ color	75.2	68.5
+ **multi-view** appearance	58.9	52.3
+ **multi-view** corresp.	55.5	49.1
+ multi-modal p.t.	**54.5**	**47.7**

(a) Part likelihood terms.

#	walking	box
1	75.2	68.5
2	68.3	64.7
3	**54.5**	**47.7**

(b) Number of camera views.

#	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**	-

(c) Mixture PS components: number, clustering, and selection strategy

Table 1: HumanEva-I. Test on S1, trained on S1,S2,S3, 3D error in mm.

4 Evaluation

We evaluate our approach on two datasets, HumanEva-I [30] which is a standard benchmark for 3D pose estimation in the laboratory setting, and on the more challenging MPII Cooking dataset [27] 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 [37], 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 [18] by using pose estimated from all 3 cameras. For the walking and box sequences we evaluate on the validation set, as [34, 37], and for *combo* sequence on the test set, as [34]. In order to compensate for the slight differences in positioning of joints in our model and in HumanEva 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 ground-truth.

MPII Cooking. We use the same training and test sets as in [27], evaluating 2D projections per camera and reporting percentage of correct parts (PCP). In contrast to [27] 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 5 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.

4.1 Evaluation of our approach

HumanEva-I. We start evaluating the effect of the various design choices on the overall performance of the model shown in Table 1. We begin by examining the different improvements for the part likelihood terms for the walking sequence (Table 1a, first column). Our model proposed in [22] with flexible parts (FPS) [27] and shape features has a 3D error of 114.3mm (first line, Table 1a). With our mixture of pictorial structures (with 16 components, 3D clustering, and min-var selection) the error significantly reduces to 83.2mm. 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. As an additional image feature we add color which further reduces error to 75.2mm.

So far we estimated pose individually per camera. We now examine the benefits of using our proposed multi-view model (see Sec. 3). Relating appearance across multiple views reduces error by 16.3mm to 58.9mm, adding location correspondence further reduces error to 55.5mm. Finally we add multi-modal pairwise terms for the parts which reaches the minimal error of 54.5mm. Table 1b compares the performance gain from a single camera

Part likelihood terms	cam 1	cam 2
FPS [27]	63.7	66.3
+ color	66.9	70.8
+ **multi-view** correspondence	72.6	73.8
+ **multi-view** appearance	74.3	75.4
+ mixtures PS	78.7	79.3
+ multi-modal pairwise terms	**80.0**	**80.5**

(a) Monocular and multi-view improvements

Model	Torso	Head	upper arm r	upper arm l	lower arm r	lower arm l	All
Cam-1							
FPS [27]	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 [27]	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**

(b) Results per part

Table 2: MPII Cooking, in percentage of correct parts (PCP).

over two to three cameras. The improvements show that our model can strongly exploit the appearance and spatial correspondences across views. Note that in our experiments the relative positions of the cameras remained the same for training and test runs. In the more general case of arbitrary positioned cameras these constraints are likely to be more difficult to incorporate, however, this remains for future work.

The second columns of Tables 1a and 1b show results for the boxing sequence. Compared to walking the error drops in all cases which can be explained by reduced variability in the viewpoint of the person. For this reason and only limited training data of 404 frames, we use only 8 instead of 16 components used for walking. Additionally we examine different options for the mixtures of pictorial structures in Table 1c. The first line gives the error without a mixture of a single component with 87.1/60.6mm for walking/box. Splitting the data into 8 components by clustering the data in camera 1 in 2D decreases the error to 59.2/52.1mm. A further decrease in error to 55.9/47.7mm can be achieved by clustering the data in 3D (line 4 in Table 1c). 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 59.4mm 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 54.5mm. This setup is used throughout all remaining experiments on HumanEva-I and also in Table 1a and 1b. In Fig. 2 we show qualitative results for diverse poses and activities.

MPII Cooking. Next we evaluate our design choices on the pose challenge from MPII Cooking as shown in Table 2a. Our FPS model [27] achieves 63.7 and 66.3 PCP (percentage of correct parts, higher is better) for camera 1 and camera 2, respectively. Adding color features improves performance to 66.9 and 70.8 PCP. If we add multi-view correspondence and appearance information between both cameras performance improves by a total of 7.4/4.6 to 74.3/75.4 PCP, which is consistent with the improvements on HumanEva-I. Additionally we add mixtures of PS with 5 components[2], 3D clustering and min-var component selection, gaining an additional 4.4 / 3.9 to 78.7 / 79.3 PCP. Finally, we include multi-modal pairwise part terms, achieving a PCP of 80.0 and 80.5 for camera 1 and camera 2, respectively.

4.2 Comparison to state-of-the-art

HumanEva-I. In Tables 3 we evaluate our model for the sequences and settings considered in related work and compare to state-of-the-art approaches of [34, 37]. We commence by describing the results for the walking sequence in Table 3a. 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

[2] 5 components was the maximum number which gave reasonable clusters with the limited training data.

Train	Test	CRBM [34]	imCRBM [34]	Yao [37]	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 [34]	Yao [37]	our
S1	75.4	74.1	**47.7**

(b) Box

Train	Test (S3)	CRBM [34]	imCRBM [34]	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: HumanEva-I. Comparison to state-of-the art, 3D error in mm.

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 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: 43.4-67.0, Yao: 41.6-64.0), indicating that our approach is less dependent on the respective setting.

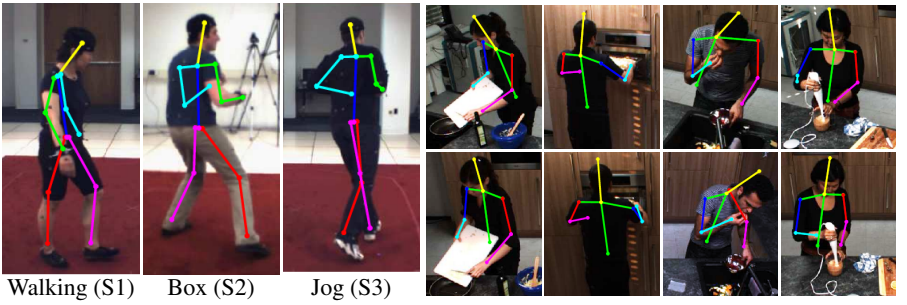
Next we examine the results for *box* in Table 3b. Here related work only reports results for subject S1: 75.4mm (CRBM) and 74.1mm (Yao *et al*.) error. Our model significantly reduces the error by 26.4mm to 47.7mm. In this case CRBM and Yao *et al*. cannot benefit from the strong motion prior used in the walking sequence as the box activity is less cyclic.

Finally we compare on the combo sequence in Table 3c. 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 [34] 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 3a), 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 the state-of-the-art approach is FPS [27]. In Table 2a we compare our approach in detail for all parts. We see that our approach improves for all parts and both cameras. Especially for the right and left lower arm we improve at least by 17.6 PCP (for right lower arm, cam-2). Overall we improve by 16.3 and 14.2 to an impressive 80.0 and 80.5 PCP for camera 1 and 2, respectively.

5 Conclusion

3D human pose estimation is traditionally addressed using 3D body models. In this work we follow an alternative avenue and rephrase the problem as inference over the set of 2D projections of the 3D pose in each camera view. This alternative formulation builds on the state-of-the-art pictorial structures model and allows to benefit from recent advances in 2D human pose estimation. By extending the original model we proposed in [22] with flexible parts, color features, multi-modal pairwise terms, and mixtures of pictorial structures, our 2D pose estimation approach significantly improves performance on both datasets used for evaluation. To exploit the multi-view information we augment the model with appearance and spatial correspondence constraints across views. Overall we achieve similar or better performance compared to state-of-the-art [34, 37] on HumanEva-I without using tracking or exploiting activity specific priors. Similar, on the pose challenge of MPII cooking, our



(a) HumanEva-I (b) MPII Cooking, top: cam 1, bottom: cam 2.

Figure 2: Example 3D pose estimation results from our approach (projected to 2D).

approach consistently improves for all parts over state-of-the-art.

Acknowledgements. This work was funded partially by the DFG project SCHI989/2-2.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR 2010*.

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

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

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

[5] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013.

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

[7] Jonathan Deutscher and Ian Reid. Articulated body motion capture by stochastic search. *IJCV*, 61:185–205, 2005.

[8] Marcin Eichner and Vittorio Ferrari. Better appearance models for pictorial structures. In *BMVC 2009*.

[9] Marcin Eichner and Vittorio Ferrari. Appearance sharing for collective human pose estimation. In *ACCV*, 2012.

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

[11] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[12] Vittorio Ferrari, Manuel Marin, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *CVPR 2008*.

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

[14] 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*.

[15] 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.

[16] Juergen Gall, Angela Yao, and Luc J. Van Gool. 2D action recognition serves 3D human pose estimation. In *ECCV*, 2010.

[17] Stephan Gammeter, Andreas Ess, Tobias Jaeggli, Konrad Schindler, Bastian Leibe, and L.J.V. Gool. Articulated multi-body tracking under egomotion. In *ECCV 2008*.

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

[19] 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*.

[20] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[21] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[22] Andriluka Mykhaylo, Roth Stefan, and Schiele Bernt. Discriminative appearance models for pictorial structures. *IJCV*, 2011.

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

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

[25] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV 2011*.

[26] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *CVPR 2013*.

[27] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[28] Ben Sapp, David Weiss, and Ben Taskar. Parsing human motion with stretchable models. In *CVPR 2011*.

[29] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[30] 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.

[31] Leonid Sigal and Michael J. Black. Predicting 3D people from 2D pictures. In *AMDO 2006*.

[32] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. Tracking loose-limbed people. In *CVPR 2004*.

[33] 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*.

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

[35] R. Urtasun and T. Darrell. Local probabilistic regression for activity-independent human pose inference. In *ICCV 2009*.

[36] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

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