# 3D Pictorial Structures for Multiple Human Pose Estimation

Vasileios Belagiannis[1], Sikandar Amin[2,3], Mykhaylo Andriluka[3],
Bernt Schiele[3], Nassir Navab[1], and Slobodan Ilic[1]

[1]Computer Aided Medical Procedures, Technische Universität München, Germany
[2]Intelligent Autonomous Systems, Technische Universität München, Germany
[3]Max Planck Institute for Informatics, Saarbrücken, Germany
*{belagian, sikandar.amin, navab, slobodan.ilic}@in.tum.de, {andriluka, schiele}@mpi-inf.mpg.de*

## Abstract

*In this work, we address the problem of 3D pose estimation of multiple humans from multiple views. This is a more challenging problem than single human 3D pose estimation due to the much larger state space, partial occlusions as well as across view ambiguities when not knowing the identity of the humans in advance. To address these problems, we first create a reduced state space by triangulation of corresponding body joints obtained from part detectors in pairs of camera views. In order to resolve the ambiguities of wrong and mixed body parts of multiple humans after triangulation and also those coming from false positive body part detections, we introduce a novel 3D pictorial structures (3DPS) model. Our model infers 3D human body configurations from our reduced state space. The 3DPS model is generic and applicable to both single and multiple human pose estimation.*

*In order to compare to the state-of-the art, we first evaluate our method on single human 3D pose estimation on HumanEva-I [22] and KTH Multiview Football Dataset II [8] datasets. Then, we introduce and evaluate our method on two datasets for multiple human 3D pose estimation.*

## 1. Introduction

Articulated objects and especially humans are an active area in computer vision research for many years. Determining the 3D human body pose has been of particular interest, because it facilitates many applications such as tracking, human motion capture and analysis, activity recognition and human-computer interaction. Depending on the input modalities and number of employed sensors different methods have been proposed for single human 3D pose estimation [2, 4, 8, 20, 24]. Nevertheless, estimating jointly the 3D pose of multiple humans from multi-views, has not been fully addressed yet (Figure 1).

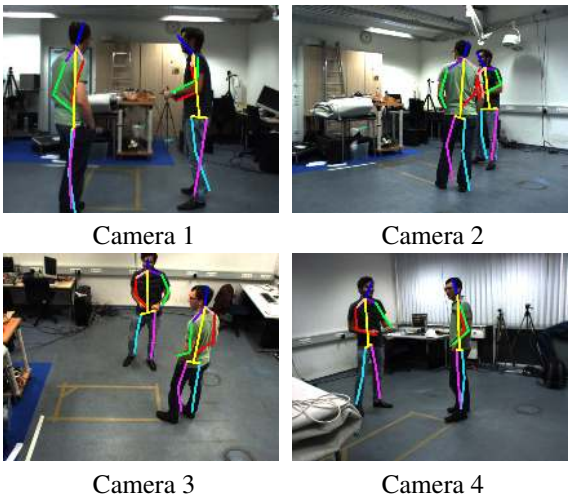



Figure 1: **Shelf dataset**: Our results projected in 4 out of 5 views from our proposed multi-view dataset.

In a multi-view setup, the 3D space can be discretized into a volume in which the human body is defined as a meaningful configuration of parts. Estimating the 3D body pose can be an expensive task due to the six degrees of freedom (6 DoF) of each body part and the level of discretization, as it has been analyzed by Burenius et al. [8]. In order to reduce the complexity of the 3D space, many approaches rely on background subtraction [24] or assume fixed limb lengths and uniformly distributed rotations of body parts [8]. Instead of exploring a large state space of all possible translations and rotations of the human body parts in 3D space, we propose a more efficient approach. We create a set of 3D body part hypotheses by triangulation of corresponding body joints sampled from the posteriors of

2D body part detectors [2] in all pairs of camera views. In this way, our task becomes simpler and requires inferring a correct human skeleton from a set of 3D body part hypotheses without exploring all possible rotations and translations of body parts.

Another common problem in single human approaches [2, 8] is the separation between left-right and front-back of the body anatomy because of the different camera positions. This problem becomes more complicated in multiple human 3D pose estimation, given similar body parts of different humans in each view. In this way, not knowing in advance the identity of the humans and consequently their body parts in each view results in more ambiguities because of the mixing of body parts of different individuals. For example, a left hand of one person in one view will have multiple left hand candidates in other camera views coming not only from the same person, but also from other individuals and potential false positive detections. In practice, this will create fake body parts and can lead to fake skeletons in 3D space.

In order to resolve these ambiguities, we introduce a novel 3D pictorial structures (3DPS) model that infers skeletons of multiple humans from our reduced state space of 3D body part hypotheses. The 3DPS model is based on a conditional random field (CRF) with multi-view potential functions. The unary potentials are computed from the confidence of the 2D part-based detectors and reprojection error of the joint pairs of the corresponding body parts. We propose additionally the part length and visibility unary potentials for modelling occlusions and resolving geometrical ambiguities. The pairwise potential functions integrate a human body prior in which the relation between the body parts is modelled. Our body prior is learned from one camera setup but it works with any other setup. We constrain the symmetric body parts to forbid collisions in 3D space by introducing an extra pairwise collision potential. Finally, the inference on our graphical model is performed using belief propagation. We parse each human by sampling from the marginal distributions. Our only assumption is to have correctly detected every body part joint from at least two views in order to recover the part during inference. Our model is generic and applicable to both single and multiple human pose estimation. Moreover, inference of multiple human skeletons does not deteriorate despite the ambiguities which are introduced during the creation of the multi-human state space.

This work has the following contributions: First, we propose the 3D pictorial structures (3DPS) model that can handle multiple humans using multi-view potential functions. Very importantly, we do not assume that we have information about the identity of the humans in each view other than 2D body part detections. Experimental results on HumanEva-I [22] and KTH Multiview Football II [8] datasets demonstrate that our model is on par with state-of-the-art methods [2, 8] for single human 3D pose estimation. Secondly, we introduce a discrete state space for fast inference, instead of exploring a finely discretized 3D space. Finally, we propose two new datasets (Campus [5] and Shelf) with ground-truth annotations and evaluate our multiple human pose estimation method.

### 1.1. Related work

Reviewing the entire literature on human pose estimation is beyond the scope of this paper [19, 23]. Due to the relevance to our work, we focus on literature for 3D human body pose estimation.

The categorization in discriminative and generative approaches is common for both 2D and 3D human body pose estimation. In the discriminative category, a mapping between image (e.g. silhouettes, edges) or depth observations and 3D human body poses is learned [1, 14, 16, 20, 26, 28, 30]. These types of methods are unstable to corrupted data because of classification failures. They also only generalize up to the level in which unknown poses start to appear. Nonetheless, training with depth data has been proven to generalise well to unknown poses [20]. However, current depth sensors, such as Kinect, are not useful for providing reliable depth information outdoors, where single and multiple cameras are still widely accessible.

Most of the generative approaches rely on a kinematic chain where the parts of the object are rigidly connected. The problem is often coupled with tracking [7, 9, 13, 21, 28, 30]. In such approaches, which are also called top-down methods, the human skeleton is represented either in a high-dimensional state space or embedded in low dimensional manifolds bound to the learned types of motion. Since these methods rely on tracking, they require initialisation and cannot recover in case of tracking failures.

There is another family of generative approaches, also called bottom-up, in which the human body is assembled from parts [4, 24]. These methods are referred to as pictorial structures and they do not imply rigid connections between the parts. Pictorial structures is a generic framework for object detection which has been extensively explored for 2D human body pose estimation [3, 4, 10, 12, 29]. Deriving the 3D human pose is possible by learning a mapping between poses in the 2D and 3D space [25] or lifting 2D poses [4], but this is not generic enough and is restricted to particular types of motion. Recently, several approaches have been introduced that extend pictorial structure models to 3D human body pose estimation. The main challenge in extending pictorial structures to 3D space is the large state space that has to be explored. Burenius et al. [8] have recently introduced an extension of pictorial structures to the 3D space and analysed the feasibility of exploring such a huge state space of possible body part translations and rotations. In order to

make the problem computationally tractable, they impose a simple body prior that limits the limb length and assumes a uniform rotation. Adding a richer body model would make the inference much more costly due to the computations of the pairwise potentials. Consequently, the method is bound to single human pose estimation and the extension to multiple humans is not obvious. The follow-up work of Kazemi et al. [17] introduces better 2D part detectors based on learning with randomized forest classifiers, but still relies on the optimization proposed in 3D pictorial structures work [8]. In both works, the optimization is performed several times due to the ambiguity of the detector to distinguish left from right and front from back. As a result, the inference should be performed multiple times while changing identities between all the combinations of the symmetric parts. In case of multiple humans, either having separate state spaces for each person or exploring one common state-space, the ambiguity of mixing symmetric body parts among multiple humans becomes intractable. Both papers evaluate on a football dataset that they have introduced and it includes cropped players with simple background. We have evaluated our approach on this dataset. Another approach for inferring the 3D human body pose of a single person is proposed by Amin et al. [2]. Their main contribution lies in the introduction of pairwise correspondence and appearance terms defined between pairs of images. This leads to improved 2D human body pose estimation and the 3D pose is obtained by triangulation. Though this method obtained impressive results on HumanEva-I [22], the main drawback of the method is the dependency on the camera setup in order to learn pairwise appearance terms. In contrast, our body prior is learned once from one camera setup and is applicable to any other camera setup.

Finally, similar to our 3DPS model, the loose-limbed model of Sigal et al. [24] represents the human as a probabilistic graphical model of body parts. The likelihood term of the model relies on silhouettes (i.e. background subtraction) and applies only to single human pose estimation. This model is tailored to work with the Particle Message Passing method [27] in a continuous state space that makes it specific and computationally expensive. In contrast, we propose a 3DPS model which is generic and works well both on single and multiple humans. We resolve ambiguities imposed by multiple human body parts. Additionally, we operate on a reduced state space that make our method fast.

## 2. Method

In this section, we first introduce the 3D pictorial structures (3DPS) model as a conditional random field (CRF). One important feature of the model is that it can handle multiple humans whose body parts lie in a common 3D space. First, we present how we reduce the 3D space to a smaller discrete state space. Next, we describe the potential functions of the 3DPS model, emphasizing on how this model addresses challenges of multiple human 3D pose estimation in multi-views. Finally, we discuss the inference method that we employ to extract 3D human body skeletons.

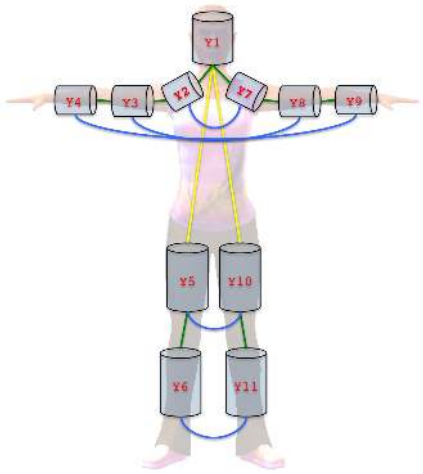



Figure 2: **Graphical model of the human body**: We use 11 variables in our graph to represent the body parts. The kinematic constrains are expressed in green (rotation) and yellow (translation) edges, while the collision constrains are drawn with blue edges.

### 2.1. 3D pictorial structures model

The 3D pictorial structure (3DPS) model represents the human body as an undirected graphical model (Figure 2). In particular, we model the human body as a CRF of $n$ random variables $Y_i \in \mathbf{Y}$ in which each variable corresponds to a body part. An edge between two variables denotes conditional dependence of the body parts and can be interpreted as a physical constraint. For instance, the lower limb of the arm is physically constrained to the upper one. The body pose in 3D space is defined by the body configuration $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Each variable $Y_i$ defines a body part state vector $Y_i = [\chi_i^{pr}, \chi_i^{di}]^T \in \mathbb{R}^6$ as the 3D position of the proximal $\chi_i^{pr} \in \mathbb{R}^3$ and distal $\chi_i^{di} \in \mathbb{R}^3$ joint in the global coordinate system (Figure 3) and takes its values from the discrete state space $\Lambda_i$.

Considering now an instance of the observation $\mathbf{x} \in \mathbf{X}$ (i.e. body part hypotheses) and a body configuration $\mathbf{y} \in \mathbf{Y}$, the posterior becomes:

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

where $Z(\mathbf{x})$ is the partition function, $E_{kin}$ are the graph edges that model the kinematic constraints between the body parts and $E_{col}$ are the edges that model the collision between symmetric parts. The unary potentials are composed of the detection confidence $\phi_i^{conf}(y_i, \mathbf{x})$, reprojection error $\phi_i^{repr}(y_i, \mathbf{x})$, body part multi-view visibility $\phi_i^{vis}(y_i, \mathbf{x})$ and the body part length $\phi_i^{len}(y_i, \mathbf{x})$ potential functions. The pairwise potential functions encode the body prior model by imposing kinematic constraints on the translation $\psi_{i,j}^{tran}(y_i, y_j)$ and rotation $\psi_{i,j}^{rot}(y_i, y_j)$ between the body parts. Symmetric body parts are constrained not to collide with each other by the collision potential function $\psi_{i,j}^{col}(y_i, y_j)$.

Next, we first define the discrete state space, unary and pairwise potential functions and secondly conclude with the inference and parsing of multiple humans.

**Discrete state space** The state space $\Lambda_i$ of a body part variable $Y_i$ comprises the $h$ hypotheses that the variable can take. A hypothesis corresponds to a 3D body part's position and orientation. In order to create our global state space of multiple human body parts $\mathbf{\Lambda} = \{\Lambda_1, \Lambda_2, \dots \Lambda_n\}$, we employ 2D part detectors in each view separately. We rely on the approach of [2], which produces a posterior probability distribution of the body part position and orientation in the 2D space. By sampling a number of samples from this distribution, we create 2D body part hypotheses in every image. In practice, the detected body parts of [2] correspond to human body joints.

Assuming a calibrated system of $c$ cameras, the 3D discrete state space is formed by triangulation of corresponding 2D body joints detected in multi-views. The triangulation step is performed for all combinations of view pairs. To create the actual global state space $\mathbf{\Lambda}$, which is composed of body parts and not only joints, we create a 3D body part from a pair of 3D joints. One 3D joint corresponds to the proximal and the other to the distal joint of the body part, as depicted in Figure 3. The proximal joint defines the position of the 3D body part, while its orientation is derived using the distal joint. For each body part state space $\Lambda_i$, there is a number of hypotheses $\Lambda_i = \{\lambda_i^1, \lambda_i^2, \dots, \lambda_i^h\}$ that can be associated to it. Not knowing the identity of humans creates wrong hypotheses stemming from the triangulation of the corresponding body parts of different people. Note that such wrong body part hypotheses can look correct in the 3D space and can even create a completely fake skeleton when different people are in a similar pose, as shown in Figure 4. Finally, the number of hypotheses of the state space scales with the number of views, and with a number of input 2D body joints sampled from the posteriors of the 2D part detector, but in general remains small enough for fast inference.

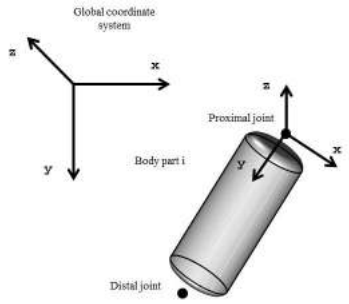



Figure 3: **Body part structure**: Each body part is composed of the proximal and distal joint position. A local coordinate system is attached to its proximal joint.

**Unary potentials** In our approach, the unary potential functions are designed to score in a mult-view setup with multiple humans. Every body part hypothesis is defined by the 3D position of its joints and part orientation. In addition, it includes the detection confidence and reprojection error of the joints from which it occurred. We propose to use these measurements to the estimation of the unary potential functions.

At first, the detection confidence function $\phi_i^{conf}(y_i, \mathbf{x})$ is the mean confidence of the part detector in two views. Secondly, given two joint positions $\mathbf{p}$ and $\mathbf{p}'$, either proximal or distal, of the body part $i$ observed from two views and the triangulated point $\chi_i \in \mathbb{R}^3$, the reprojection error [15] is measured from the following geometric error cost function:

$$C(\chi_i) = d(\mathbf{p}, \hat{\mathbf{p}})^2 + d(\mathbf{p}', \hat{\mathbf{p}}')^2 \tag{2}$$

where $d$ corresponds to the euclidean distance, and $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}'$ are the projections of the joint $\chi_i$ in the two views. In order to express the reprojection error as the score of a hypothesis, a sigmoid function is employed. Since the error is always positive, the function is reformulated and integrated into the reprojection error potential function $\phi_i^{repr}(y_i, \mathbf{x})$. The final potential function becomes:

$$\phi_i^{repr}(y_i, \mathbf{x}) = \frac{1}{1 + \exp(\bar{C}(\chi_i))}. \tag{3}$$

To take advantage of the multi-view information, we introduce the body part multi-view visibility potential $\phi_i^{vis}(y_i, \mathbf{x})$ which weights a hypothesis based on the number of views in which it has been observed. To compute the number of views, we project the hypothesis to each view and search in a small radius (5 pixels) for an instance of the part detector. Then, we normalize the estimated number of visible views with respect to the total number of cameras. Consequently, hypotheses that occur from ambiguous views (e.g. opposite cameras) or false positive hypotheses (Figure 4) are implicitly penalized by obtaining a smaller visibility weight. Thus, the visibility term is complementary to the

reprojection error. Finally, we model the length of a body part with the length potential function $\phi_i^{len}(y_i, \mathbf{x})$. We use a one dimensional Gaussian distribution and ground-truth data to learn the mean and standard deviation of the length of each body part. This potential function mainly penalizes body parts that occur from joints of different individuals.

In the formulation of the posterior (1), we consider the dependence between unary potential functions. The confidence of the part detector, which also contributes to the creation of the 3D hypotheses, is the most important potential function. However, false positive detections or triangulations with geometric ambiguity should be penalized. This is achieved by the reprojection and multi-view visibility potential functions. For instance, a wrongly detected 2D joint, with a high detection confidence, should normally have a high reprojection error. Hence, the score of the reprojection potential of a false positive part is low. Furthermore, part hypotheses that have been created from different individuals with similar poses can have small reprojection error but they are penalized from the multi-view visibility potential. Finally, true positive joint detections of different individuals create wrong body part hypotheses with high detection confidence but they are penalized by the part length potential function.

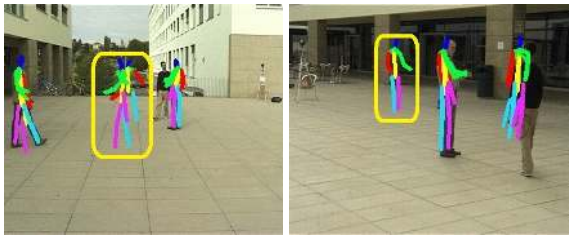


Figure 4: **Body parts state space**: The body part hypotheses are projected in two views. Fake hypotheses which form reasonable human bodies are observed in the middle of the scene (yellow bounding box). These are created by intersecting the joints of different humans with similar poses because the identity of each person is not available.

**Pairwise potentials** The paradigm of pictorial structures in the 2D space has successfully modelled the relations between body parts [4, 10, 12]. We follow the same idea and express a body part in the local coordinate system of a neighbouring part (Figure 2). We model the rotation or translation between the body parts using Gaussian distributions. Furthermore, the symmetric parts are forced not to collide for recovering from false positive detections.

Initially, the state vector $Y_i$ of the part $i$ is expressed in a local coordinate system. To define the local coordinate system, we build on the geometric vectors, which are defined from the proximal and distal joints of the part $i$ and its neighbour $j$. Then, the matrix transformation $H_i(Y_i) \in \mathbb{R}^{4\times4}$ includes the rotation and translation of the part $i$ from its local to the global coordinate system. The inverse transformation $H_i^{-1}(Y_i)$ maps the part $i$ back to the local coordinate system. We denote as $Y_{ij} \in \mathbb{R}^{4\times4}$ the transformation for expressing the part $i$ to the local coordinate system of the part $j$ and it is given from:

$$Y_{ij} = H_j^{-1}(Y_i) \cdot H_i(Y_i). \tag{4}$$

We assume independence between the rotation $Y_{ij}^R$ and the translation $Y_{ij}^T$ of the $Y_{ij} = [Y_{ij}^R, Y_{ij}^T]$ transformation and learn two different priors, based on the type of the constraint (Figure 2). For the rotation $y_{ij}^R$, we consider only the case of hinge joints for imposing fewer constraints to our prior model. Thus, we fix the two axes of rotation and learn a prior for the third one. Since the prior captures the rotation only along one axis, it is modelled by a Gaussian distribution:

$$\psi_{i,j}^{rot}(y_i, y_j) = \mathcal{N}(y_{ij}^R \mid \mu_{ij}^R, \sigma_{ij}^R) \tag{5}$$

where $\mu_{ij}^R$ is the mean and $\sigma_{ij}^R$ the variance. In order to model the whole rotational space, a von Mises distribution would be required. But in our experiments, we have seen that an approximation with a Gaussian is sufficient. The translation $y_{ij}^T$ is modelled using a multivariate Gaussian distribution:

$$\psi_{i,j}^{tran}(y_i, y_j) = \mathcal{N}(y_{ij}^T \mid \mu_{ij}^T, \Sigma_{ij}^T) \tag{6}$$

with mean $\mu_{ij}^T$ and covariance $\Sigma_{ij}^T$. For relaxing the computations, the diagonal of the covariance is only estimated.

In addition, we model the relation between the symmetric body parts to avoid collisions between them. This problem occurs because of false positive (FP) detections that can occur. To that end, a body part is defined as a pair of spheres where each sphere is centred on the part's joints. Then, the collisions of symmetric parts are identified by estimating the sphere-sphere intersection [18]. We model this relation by penalizing the collided part hypotheses with a constant $\delta$:

$$\psi_{i,j}^{col}(y_i, y_j) = \delta \cdot inter(y_i, y_j) \tag{7}$$

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

We use ground-truth data to learn the pairwise potential functions. Since the world coordinate system is cancelled by modelling the relation of the body parts in terms of local coordinate systems, we are not dependent on the camera setup, in contrast to [2]. Thus, we can learn the prior model from one dataset and use it during inference to any other dataset. Moreover, our prior model is stronger than a binary voting for a body part configuration [8] and less computational expensive than [24]. During inference of multiple humans, our prior model constrains the body parts of each individual to stay connected.

### 2.2. Inference of multiple humans

The final step for obtaining the 3D pose of multiple humans is the inference. The body part hypotheses of all humans share the same state space. In addition, the state space includes completely wrong hypotheses due to the unknown identity of the individuals and false positive detections as well. However, our body prior and the scores of the unary potentials allow us to parse each person correctly.

Here, we seek to estimate the posterior probability of equation (1). Since our graphical model does not have a tree structure, we employ the loopy belief propagation algorithm [6] for estimating the marginal distributions of the body parts. Estimating the number of humans jointly in all views using a detector [11], we know how many skeletons we have to build. The body parts of each individual are sampled from the marginal distributions and projected to all views. We choose views with small overlap ($< 30\%$) between the detection boding boxes for avoiding mixing up the body parts of different individuals. Gradually, all the 3D poses are parsed based on the detection input. Body parts that have not been detected from the part detectors from one or any view, are not parsed. As a result, we allow a 3D human pose to lack body parts.

Our framework for multiple human 3D pose estimation applies exactly the same on single humans. In the next section, we demonstrate it by evaluation our model both on single and multiple human 3D pose estimation.

## 3. Experiments

In this section, we evaluate our approach on single and multiple human pose estimation on four datasets. At first, we use the HumanEva-I [22] and KTH Multiview Football II [8] datasets to demonstrate that our model is directly applicable to single human 3D pose estimation. We compare our results with two relevant multi-view approaches [2, 8]. Since we are not aware of a multiple human dataset, we have annotated the Campus dataset [5] (Figure 7) and introduce our own Shelf dataset for multiple human evaluation (Figure 1).

The model that we employ for the experiments is composed of 11 body parts (Figure 2). For each evaluation dataset, we use the training sequences to learn our model's appearance term but the body prior is learned only once. Our part detector is based on the 2D part detector of [2] and the human detector of [11]. Since our body prior is not dependent on the camera setup and consequently on the evaluation dataset, we learn the body prior for the pairwise potentials from a training subset of the Campus dataset [5] and use it during all the evaluations.

Camera 1 Camera 2 Camera 3

Figure 5: **HumanEva-I**: The 3D estimated body pose is projected across each view for the Box sequence.

### 3.1. Single human evaluation

We first evaluate our method on single human 3D pose estimation for demonstrating that it performs as well as start-of-the-art multi-view approaches [2, 8]. The purpose of this experiment is to highlight that we can achieve similarly good or even better results than other methods without the need to learn a calibration-dependent body prior [2] or a weak prior [8] for relaxing the computations.

Camera 1 Camera 2 Camera 3

Figure 6: **KTH Multiview Football II**: The 3D estimated body pose is projected across each view for the player 2 sequence.

**HumanEva-I:** We evaluate on Box and Walking sequences of the HumanEva-I [22] dataset and compare with [2, 24]. We share similar appearance term only for the 2D single view part detection with [2] and employ different body models. Table 1 summarizes the results of the average 3D joint error. Notably, Amin et al. [2] report very low average error but we also achieve similar results. Cases in which we have observed failures are related to lack of correct detected joints from at least two cameras.

| Sequence | Walking | Box |
|---|---|---|
| Amin et al. [2] | 54.5 | 47.7 |
| Sigal et al. [24] | 89.7 | - |
| Our method | 68.3 | 62.7 |

Table 1: **Human-Eva I**: The results present the average 3D joint error in millimetres (mm).

**KTH Multiview Football II:** In this sequence, we evaluate on Player 2 as in the original work [8]. We follow the same evaluation process as in [8] and estimate the PCP (per-

centage of correctly estimated parts) scores for each set of cameras. The results are summarized in Table 2. We outperform the method of [8] on two cameras and lose some performance for the legs using three cameras due to detection failures. Note that overall we obtain similar results with significant fewer computations due to our discrete state space. Our approach runs on around 1 fps for single human 3D pose estimation, given the 2D detections. The experiments are carried out on a standard Intel i5 2.40 GHz laptop machine and our method is implemented in C++ with loop parallelizations.

| | Bur. [8] | Our | Bur. [8] | 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 |

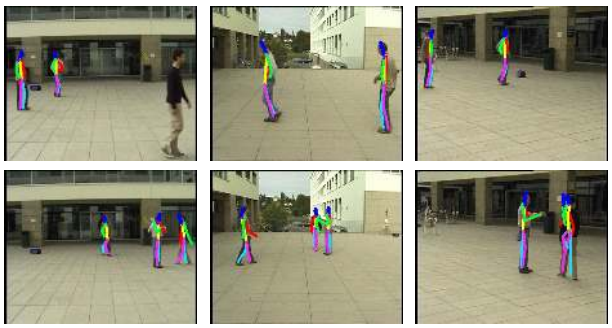
Table 2: **KTH Multiview Football II**: The PCP (percentage of correctly estimated parts) scores, for each camera, are presented for our method and [8]. One can observe that we have mainly better results for the upper limbs.

### 3.2. Multiple human datasets and evaluation

Multiple human 3D pose estimation is a problem which has not yet been extensively addressed. One can observe that from the available literature and evaluation datasets. While for single humans there are standard evaluation datasets such as HumanEva [22], there is no standard benchmark on multiple human 3D pose estimation. In this work, we propose our own Shelf dataset which consists of disassembling a shelf (Figure 1). The Shelf dataset includes up to four humans interacting with each other. We have produced manual joint annotation in order to evaluate our method. Furthermore, we have annotated the Campus dataset [5] which is composed of three humans performing different actions. We evaluate our method on both datasets.

Since we are not aware of another method which performs multiple human 3D pose estimation, we chose a single human approach [2] to compare to and perform 3D pose estimation for each human separately. Of course, this way of evaluation is not to our favour because evaluating on each human separately, knowing their identity, excludes body part hypotheses that belong to other humans and simplifies the inference. In our method, the body parts of all humans lie in the same state space. We evaluate our method for multiple humans simultaneously and for each one separately.

**Campus:** Assuming first that the identity of each human is known, we have evaluated our method and the one from [2] to each human separately and achieve similar results. This is the single human inference (Table 3). More interesting are the results when we apply our framework by considering all the humans together and with unknown identities. This is the multiple human inference (Table 3). We have achieved the same good results. This proves that our model is robust to including the body parts of all humans, without knowing their identity, in the same state space.



Camera 1 Camera 2 Camera 3

Figure 7: **Campus**: The 3D estimated body pose is projected across each view.

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

Table 3: **Campus**: The 3D PCP (percentage of correctly estimated parts) scores are presented. On single human inference, the identity of each actor is known. On the multiple human inference, the body parts of all actors lie in the same state and the identity of each actor is unknown.

**Shelf**[1]**:** On the proposed dataset, we follow the same evaluation protocol of single and multiple human inference. First, we detect humans in all views and then extract their body parts. Next, we run our method and finally evaluate on the detections. We obtain better results than [2] for single and multiple human inference (Table 4). In cases of occlusion, our model better recovers 3D human poses compared to [2], because of the multi-view potential terms. In the multiple human inference, we have achieved similar results as in the single human inference. This proves that including the body parts of different individuals in a common state space did not result in reduced performance. The actors are correctly inferred under self-occlusion or under occlusion by other objects.

[1] http://campar.in.tum.de/Chair/MultiHumanPose

| Inference | Single Human | | Multiple Human |
|---|---|---|---|
| | Amin et al. [2] | 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 4: **Shelf**: The 3D PCP (percentage of correctly estimated parts) scores are presented. On single human inference, the identity of each actor is known. On the multiple human inference, the body parts of all actors lie in the same state and the identity of each actor is unknown.

## 4. Conclusion

We have presented the 3D pictorial structures (3DPS) model for recovering 3D human body poses using the multi-view potential functions. We have introduced a discrete state space which allows fast inference. Our model has successfully been applied to multiple humans without knowing the identity in advance. The model is also applicable to single humans where we achieved very good results during evaluation. Self and natural occlusions can be handled by our algorithm. 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 have introduced two datasets for 3D body pose estimation of multiple humans.

## References


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

[2] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3d human pose estimation. In *BMVC*, 2013.

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

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

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

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

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

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

[9] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 2005.

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

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

[12] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 1973.

[13] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *IJCV*, 2010.

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

[15] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2000.

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

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

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

[19] 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.

[20] 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.

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

[22] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010.

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

[24] 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.

[25] 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.

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

[27] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *CVPR*, 2003.

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

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

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