3D Pictorial Structures Revisited: Multiple Human Pose Estimation

Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka,
Bernt Schiele, Nassir Navab, and Slobodan Ilic

Abstract—We address the problem of 3D pose estimation of multiple humans from multiple views. The transition from single to multiple human pose estimation and from the 2D to 3D space is challenging due to a much larger state space, occlusions and 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 pairs of body parts obtained by part detectors for each camera view. In order to resolve ambiguities of wrong and mixed parts of multiple humans after triangulation and also those coming from false positive detections, we introduce a 3D pictorial structures (3DPS) model. Our model builds on multi-view unary potentials, while a prior model is integrated into pairwise and ternary potential functions. To balance the potentials' influence, the model parameters are learnt using a Structured SVM (SSVM). The model is generic and applicable to both single and multiple human pose estimation.
To evaluate our model on single and multiple human pose estimation, we rely on four different datasets. We first analyse the contribution of the potentials and then compare our results with related work where we demonstrate superior performance.

Index Terms—Human pose estimation, 3D pictorial structures, part-based models.

1 Introduction

The problem of human pose estimation has drawn attention to the computer vision community for many years. Determining the 3D human body pose has been of particular interest, because it facilitates many applications such as human tracking, 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 [1], [2], [3], [4], [5], [6]. Nevertheless, estimating jointly the 3D pose of multiple humans from multi-views, remains an open problem (Figure 1).

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. [3]. In order to reduce the complexity of the 3D space, many approaches rely on background subtraction [6] or assume a simplified human model with fixed limb lengths and uniformly distributed rotations of body parts [3]. 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 in all pairs of camera views. In this way, our task becomes simpler and requires inferring a correct human body pose from a set of 3D body part hypotheses without exploring all possible rotations and translations of the body parts.

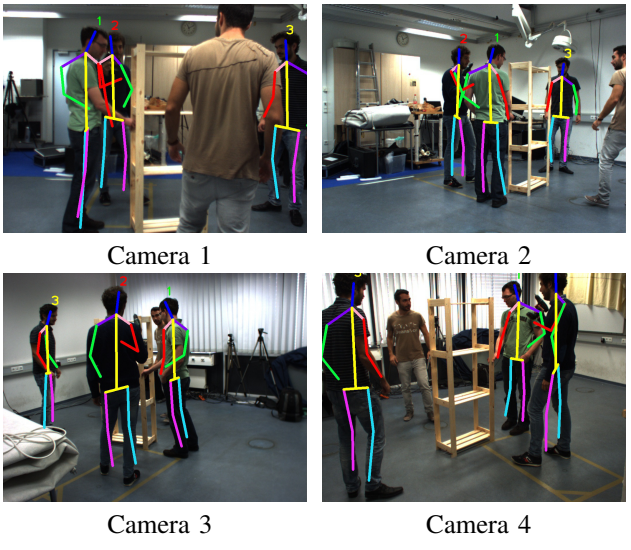


Fig. 1: **Shelf dataset**: Our results on 3D pose estimation of multiple individuals projected in 4 out of 5 views of the Shelf dataset [7].

- *V. Belagiannis and N. Navab are with Computer Aided Medical Procedures, Technische Universität München, Germany. E-mail: see http://campar.in.tum.de*
- *S. Amin is with Intelligent Autonomous Systems, Technische Universität München, Germany. E-mail: see http://ias.cs.tum.edu*
- *M. Andriluka is with Stanford University, USA. E-mail: see http://www.d2.mpi-inf.mpg.de*
- *B. Schiele is with Max Planck Institute for Informatics, Saarbrücken, Germany. E-mail: see http://www.d2.mpi-inf.mpg.de*
- *S. Ilic is with Siemens AG, Munich, Germany. E-mail: see http://campar.in.tum.de*

Another common problem, which has been particularly addressed in single human approaches (i.e. [1], [3]), is the separation between left-right and front-back of the body anatomy because of the different camera views. This problem becomes more complicated in multiple human 3D pose estimation, since we assume that the identity of each individual is not given in advance. Thus, an association between the individuals across all views is required to avoid mixing the 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 incorrect body part hypotheses that can lead to fake body poses in the 3D space.

Moreover, multiple human pose estimation has common challenges such as occlusion between individuals and different type of motion (e.g. running or walking), which should be described from a single model. When considering also dynamic environments, where background subtraction is not feasible, makes the problem more complicated in comparison to human pose estimation in a studio setup [4], [6].

In order to address these challenges, we introduce a 3D pictorial structures (3DPS) model that infers body poses of multiple humans from a reduced state space of body part 3D hypotheses. The 3DPS model is based on a conditional random field (CRF) in which a random variable corresponds to a body part. A body part is defined from three parameters that stand for its 3D position. The unary potential functions are computed from the confidence of the 2D part-based detectors and reprojection error of the corresponding body parts. We propose additionally the visibility unary potential for modelling occlusions and resolving geometrical ambiguities. Furthermore, we introduce the temporal consistence unary term for constraining the 3D body part hypotheses with respect to the inferred body poses. The pairwise and ternary potential functions integrate a human body prior in which the relation between the body parts is modelled. We constrain the symmetric body parts to forbid collisions in 3D space by introducing an extra pairwise collision potential. Since we employ multiple potential functions, the necessity to weight them correctly arises. For that reason, we rely on a Structured SVM (SSVM) [8] to learn the parameters of the model. Finally, the inference on our graphical model is performed using belief propagation. We parse the 3D pose of each individual by first localizing it and then sampling from the marginal distributions. Our only assumption is to have correctly detected every body part from at least two views in order to recover the part during triangulation. We build our model on our earlier work on 3D pictorial structures [7], but with a different body part parametrisation. Instead of defining the body part in terms of 3D position and orientation as in [7], we keep only the position parameters and implicitly encode the orientation in the factor graph. The 3DPS 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 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. Second, we learn the parameters of our model with a Structured SVM (SSVM) formulation. Third, we introduce a discrete state space for fast inference using 2D part detectors, instead of exploring a finely discretized 3D space. Very importantly, we do not assume that we have information about the number and identity of the humans in advance. During the evaluation, we perform an extensive analysis of all terms of the 3DPS model for highlighting their contribution. Experimental results on HumanEva-I [9], KTH Multiview Football II [3], Campus [10] and Shelf [7] datasets demonstrate state-of-the-art results in comparison to related work, for single and multiple 3D human pose estimation.

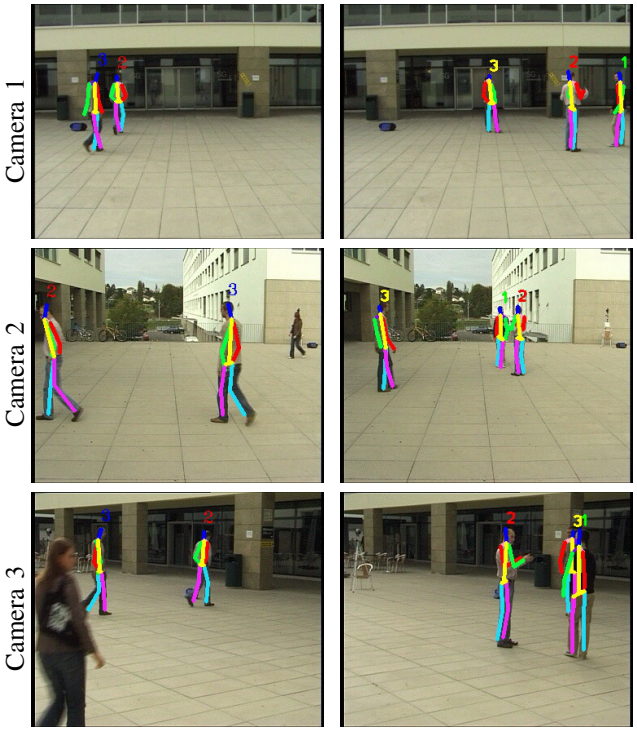


Fig. 2: **Campus dataset**: Our results on 3D pose estimation projected in all views for the Campus dataset [10]. On the result of Camera 3 on the right column, the projected poses of Actor 1 and 3 overlap in the image plane.

2 Related Work

There is plethora of literature on human pose estimation [11], [12]. Due to the relevance to our work, we focus on single and multiple human 3D pose estimation work.

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 or depth features and 3D human body poses is learnt [13], [14], [15], [5], [16], [17], [18]. 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 [19], [5]. However, current depth sensors are not useful for providing reliable depth information outdoors, where single and multiple cameras are still widely accessible. More recently, the resurrection of the neural networks, within the context of deep learning, has been proven to be the most prominent discriminative method for 2D human pose estimation [20], [21], [22]. However, deep learning has been only recently introduced to the 3D human pose estimation using a single camera [23].

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. In such approaches, the human skeleton is represented either in a high-dimensional state space [24], [25], [26], [4], [27], [28], [17], embedded in low dimensional manifolds bound to the learnt types of motion [18] or building hierarchically the body skeleton [29], [30]. 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 [31], [32], [2], [6]. 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 that has been extensively explored for 2D human body pose estimation [33], [2], [34], [35], [36], [37]. Deriving the 3D human pose from a single-view is possible by learning a mapping between poses in the 2D and 3D space [16] or lifting 2D poses [2], but this is not generic enough and is restricted to particular types of motion. Based on a multi-view setup, several recent approaches have been introduced that extend the pictorial structures to 3D human body pose estimation. The main challenge in extending pictorial structures to the 3D space is the large state space that has to be explored. Burenius et al. [3] have 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 computation of the pairwise potentials. Consequently, the method is bound to single human pose estimation and an extension to multiple humans is not obvious. The follow-up work of Kazemi et al. [38] has introduced better 2D part detectors based on learning with randomized forest classifiers, but still relied on the same optimization as in [3]. In both works, the optimization is performed several times due to the inability 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 works [3], [38] have evaluated on a football dataset that they have introduced and it includes cropped players with simple background. We have evaluated our approach on this dataset as well. Another approach for inferring the 3D human body pose of a single person from multiple views has been proposed by Amin et al. [1]. 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 [9], the main drawback of the method is the dependency on the camera setup in order to learn pairwise appearance terms. Moreover, the inference is performed in the 2D space for each view separately. In contract, we propose a 3D model in which the inference is performed directly in the 3D space.

Finally, similar to our model, the loose-limbed model of Sigal et al. [6] 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 [39] 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 human pose estimation. We resolve ambiguities imposed by multiple human body parts. Additionally, we operate on a reduced state space that make our method fast.

3 METHOD

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

In our earlier work [7], we have introduced a 3DPS model for estimating the human body pose inspired by the original work on pictorial structures to define the body part as a limb and model its position and orientation. In our revisited 3DPS model, we reduce the parameterization of the body part to include only the 3D position. A body part can be interpreted as a physical body joint, other than the head body part. To model the relation between body limbs in terms of translation and rotation, we define accordingly factors of pairs or triplets of parts in our factor graph (Figure 3). Consequently, the orientation is implicitly encoded in the factor graph. This human body parameterization facilitates the inference task, and besides,

it has demonstrated state-of-the-art results in 2D human pose estimation [37]. Finally, similar to other pictorial structures methods [33], [2], [3], [34], [1], [6], we have equally weighted the potential functions of the model in our earlier work on 3D pictorial structures [7]. In this work, we learn the parameters of our model using a Structured SVM (SSVM) solver [8] in order to balance the influence of the potential functions.

3.1 3D pictorial structures model

The 3D pictorial structure (3DPS) model represents the human body as an undirected graphical model. In particular, we model the human body as a CRF of n random variables in which each variable Y_i corresponds to a body part. An edge between two variables denotes conditional dependence of the body parts and can be described as a physical constraint. For instance, the lower limb of the arm is physically constrained to the upper one. To model the body constraints, a set of potential functions has been defined based on pairs or triplets of body parts (Figure 3). The body pose in 3D space is given by the configuration $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, where the state of each variable $Y_i \in \Lambda_i$ represents the 3D position of the body part and is taken from the discrete state space $\Lambda_i \subset \mathbb{R}^3$. The state space Λ_i is constructed based on 2D body part detection across multiple views.

Considering an instance of the observation $\mathbf{x} \in \mathbf{X}$ that corresponds to multiple-image part detections' evidence, a parameter vector $\mathbf{w} \in \mathbb{R}^D$, a set of reference poses $\mathbf{p}$ and a body configuration $\mathbf{y} \in \mathbf{Y}$, we define the posterior as:

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}, \mathbf{p}) = \frac{1}{Z(\mathbf{x}, \mathbf{w}, \mathbf{p})} \prod_{i}^{n} (\phi_i^{conf}(y_i, \mathbf{x}) \cdot \phi_i^{repr}(y_i, \mathbf{x}) \cdot \phi_i^{vis}(y_i, \mathbf{x}) \cdot \phi_i^{temp}(y_i, p_i))^{w_i} \prod_{(i,j) \in E_{tran}} \psi_{i,j}^{tran}(y_i, y_j)^{w_{ij}} \prod_{(i,j,k) \in E_{rot}} \psi_{i,j,k}^{rot}(y_i, y_j, y_k)^{w_{ijk}} \prod_{(i,j) \in E_{col}} \psi_{i,j}^{col}(y_i, y_j)^{w_{ij}} \tag{1}$$

where $Z(\mathbf{x})$ is the partition function, E_{tran} and E_{rot} are the graph edges that model the kinematic constraints between the body parts and E_{col} are the edges that model the collision constraints between symmetric parts. The reference body poses $\mathbf{p}$ correspond to inferred poses from previous time steps.

The unary potentials are composed of the detection confidence $\phi_i^{conf}(y_i, \mathbf{x})$, reprojection error $\phi_i^{repr}(y_i, \mathbf{x})$, multi-view part visibility $\phi_i^{vis}(y_i, \mathbf{x})$ and the temporal consistence potential functions $\phi_i^{temp}(y_i, p_i)$. The pairwise and ternary potential functions encode the body prior model by imposing kinematic constraints on the translation $\psi_{i,j}^{tran}(y_i, y_j)$ and rotation $\psi_{i,j,k}^{rot}(y_i, y_j, y_k)$ between the body parts. Symmetric body parts are constrained not to collide with each other by the collision potential function $\psi_{i,j}^{col}(y_i, y_j)$. The parameters w_i, w_{ij} and w_{ijk} of the model correspond to weights for the unary, pairwise and ternary potential functions. In total, our model has 14 unary, 19 pairwise and 6 ternary potential fucntions ($D = 39$).

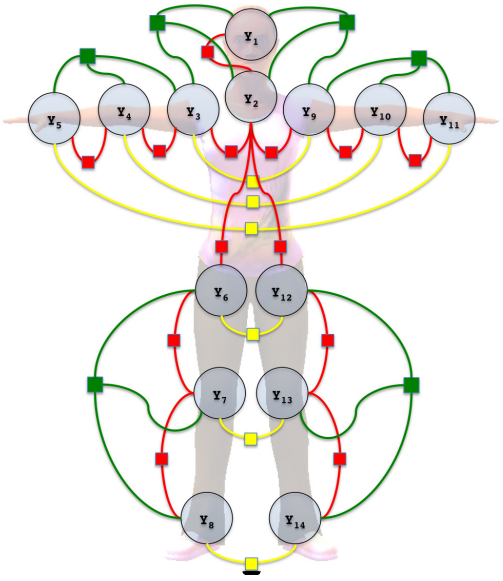


Fig. 3: **Factor graph for the human body**: We use 14 variables in our graphical model to represent the body parts. A body part in our model corresponds to a physical body joint, other than the head part. The factors denote different types of constraints and are illustrated with different colours. The kinematic constrains are presented with red (translation) and green (rotation) edges (factors). The collision constraints are represented with yellow edges. The unary factors have not been drawn for simplicity reasons.

Next, we first define the state space, unary, pairwise and ternary potential functions and secondly describe how we learn the parameters of our model. Finally, we conclude with the inference of single or multiple individuals.

State space The state space Λ_i of a variable Y_i comprises the h hypotheses that each variable can take. A hypothesis corresponds to a 3D body part position in the global coordinate system. In order to be computationally efficient, we discretise the 3D space using body part detectors in each view separately. The 2D part detectors produce a posterior probability distribution of body parts in the 2D space. By sampling a number of samples from this distribution, we create 2D body part hypotheses in every view.

Assuming a calibrated system of c cameras, the 3D discrete state space is formed by triangulation of corresponding 2D body parts detected in multi-views. The triangulation step is performed for all combinations of view pairs. For each body part state space Λ_i, there is a number of hypotheses 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 individuals. Note that such wrong hypotheses can look correct in the 3D space and even create a completely fake body skeleton when different people are in a similar pose, as shown in Figure 4. Furthermore, the 2D part detectors produce many false positive detections which result in the creation of wrong hypotheses. The total number of 3D hypotheses is given by:

$$h = n * {n_{samples}}^2 * c * (c-1)/2 \tag{2}$$

where $n_{samples}$ are the number of samples of the part detector. Generally, the number of hypotheses scales with the number of camera views c, and the number of sampled 2D body parts $n_{samples}$, but in general remains small enough for fast inference (Figure 5).

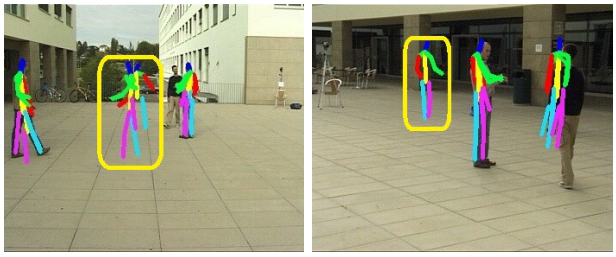


Fig. 4: **State space**: The body part hypotheses are projected in two views. Fake hypotheses which form reasonable human bodies are observed in the middle of the scene (yellow bounding box). These are created by intersecting the body parts of different humans with similar poses because the identity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

Unary potentials In our approach, the unary potential functions are designed to score in a mult-view setup with multiple humans. Each 3D body part hypothesis contributes to the estimation of the unary functions. Firstly, every hypothesis has a confidence which is defined as the average of the pairs of the triangulated 2D part detectors' confidence. The average confidence forms the detection confidence function $\phi_i^{conf}(y_i, \mathbf{x})$. Secondly, given the triangulated hypothesis $y_i \in \mathbb{R}^3$ of the body part i detected in two camera views A and B at the locations $\mathbf{p_A} \in \mathbb{R}^2$ and $\mathbf{p_B} \in \mathbb{R}^2$, the reprojection error [40] is measured from the following geometric error cost function:

$$C(y_i; \mathbf{x}) = d(\mathbf{p_A}, \pi(y_i, \mathbf{x}, A))^2 + d(\mathbf{p_B}, \pi(y_i, \mathbf{x}, B))^2 \tag{3}$$

where d corresponds to the Euclidean distance, and $\pi(y_i, \mathbf{x}, A)$ and $\pi(y_i, \mathbf{x}, B)$ are the projections of the part y_i in view A and B. In order to express the reprojection error as a score, a sigmoid function is employed and integrated into the reprojection error potential function $\phi_i^{repr}(y_i, \mathbf{x})$. The final potential function becomes:

$$\phi_i^{repr}(y_i, \mathbf{x}) = \frac{1}{1 + \exp(C(y_i; \mathbf{x}))}. \tag{4}$$

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 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.

The above unary potential functions are computed based on the observation from the current time step. To impose temporal consistency with previous inferred body poses $\mathbf{p}$ and the current 3D hypotheses, we introduce the temporal consistence function $\phi_i^{temp}(y_i, p_i)$, which acts as a regulariser between the inferred and candidate 3D part hypotheses. However, wrongly inferred body parts can occur as well. To account for both situations, we propose to consider the geometric distance between the 3D hypothesis of the part i and the inferred part p_i, if it is below a threshold c, which we have set experimentally to 10 cm. The role of the threshold c is to define a perimeter in which correct hypotheses can lie and thus it not a hard constraint. Finally, the geometric distance is reformulated as a score using a sigmoid function:

$$\phi_i^{temp}(y_i, p_i) = \begin{cases} \frac{1}{1+\exp(d(y_i, p_i))} \text{ if } d(y_i, p_i) < c \\ \epsilon \text{ otherwise} \end{cases} \tag{5}$$

where $d(y_i, p_i)$ is the Euclidean distance between the 3D part hypothesis and previously inferred parts and ϵ a small constant for numerical stability during the inference.

The main benefit of the unary potential functions' formulation is to make use of the multi-view information. 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 part, with a high detection confidence, should normally have a high reprojection error. Hence, the score of the reprojection potential of a false positive is low. Furthermore, 3D 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 part detections of different individuals create wrong body part hypotheses with high detection confidence, but they are penalized by the body prior potential functions.

Pairwise and ternary potentials The paradigm of pictorial structures in the 2D space has successfully modelled the relations between body parts in terms of location and orientation [2], [35], [36]. Recently, the body parts have been defined only using the location parameters and the orientation has been encoded in the body prior [1] or in a mixture of parts [37]. In the revisited 3DPS model, we follow the same idea of body part parametrisation and model the constraints between physical body limbs in the factor graph (Figure 3). In particular, we define two type of constraints: kinematic and collision. The kinematic constraints are modelled using translation and rotation transformations, while the collision constraints prevent the symmetric body parts from colliding with each other due to false positive detections.

Starting with the kinematic constraints, the translation potential models the translation of the part i to the local coordinate system of the part j. A multivariate Gaussian is

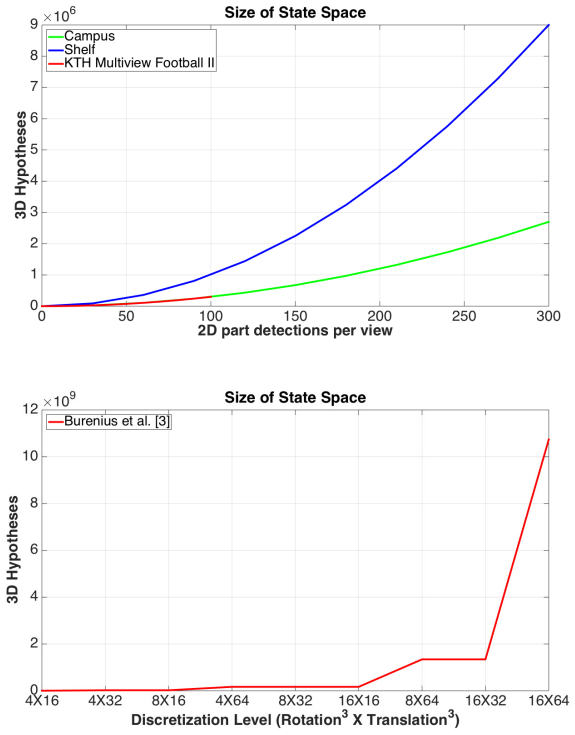


Fig. 5: **Size of the state space**: On the top graph, the size of the state space for three different datasets is presented, based on the number of sampled 2D part detections per view and individual. On the bottom graph, the size of the state space is presented according to the 3D space discretisation of [3]. It is clear that using a part detector as input results in magnitudes smaller state space in comparison to 3D space discretisation in terms of rotation and translation. In both cases, 10 body parts have been considered for the computation of the final number of 3D hypotheses. In [3], a discretisation of $8^3 X 32^3$ ($Rotation^3 \times Translation^3$) has been chosen as a compromise between performance and speed. In our case, we have sampled 40 2D parts for all the experiments.

used to capture this transformation and it is given by:

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

where $y_{ij}^T = y_i - y_j$, μ_{ij}^T is the mean and Σ_{ij}^T is the covariance. The main diagonal of the covariance is only used for relaxing the computations during the inference. Assuming that each body part belongs to a body limb, the translation brings one limb to the local coordinate system of the other. For the rotation, we consider the case of a hinge joint (i.e. 1DoF) between two body limbs. To that end, a triad of body parts is used to form two body limbs with a shared joint. The rotation across one axis can be easily computed from the dot product of the two body part vectors. The rotation potential function is modelled using a one dimensional Gaussian distribution:

$$\psi_{i,j,k}^{rot}(y_i, y_j, y_k) = \mathcal{N}(y_{ijk}^R \mid \mu_{ijk}^R, \sigma_{ijk}^R), \tag{7}$$

where $y_{i,j,k}^R = arccos(dot(y_i - y_j, y_k - y_j))$, μ_{ijk}^R is the mean and σ_{ijk}^R the variance. Moreover, we consider positive angles for the computation of the potential function. In order to model the whole rotational space, a von Mises distribution would be required. In our experiments, we have seen that an approximation with a Gaussian is sufficient.

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 more often occur for the symmetric body parts. We model the relation of the symmetric body parts by learning their Euclidean distance using another one dimensional Gaussian distribution which expressed from the collision potential function:

$$\psi_{i,j}^{col}(y_i, y_j) = \mathcal{N}(d(y_{ij}^{col}) \mid \mu_{ij}, \sigma_{ij}^{col}), \tag{8}$$

where $d(y_{ij}^{col})$ corresponds to the Euclidean distance between the part i and j, μ_{ij}^{col} is the mean and σ_{ij}^{col} the variance.

We use ground-truth data to learn the pairwise and ternary potential functions as well as the number of the previous time steps for the temporal consistence potential function. Since the world coordinate system is cancelled by modelling the relation of the body parts in terms of local coordinate systems, we are less dependent on the camera setup, in contrast to [1]. Moreover, our prior model is stronger than a binary voting for a body part configuration [3] and less computational expensive than [6]. During inference of multiple humans, our prior model constrains the body parts of each individual to stay connected.

3.2 Margin-based parameters learning

The 3DPS model has several potential functions of different nature and magnitude. Moreover, some potential functions are more error prone than others. Consequently, arises the necessity to balance the influence of the potentials within the final model. The parameters $\mathbf{w}$ of the the model weight accordingly the unary, pairwise and ternary potential functions. To learn the parameters, we pose our problem as regularised risk minimisation and use a Structured SVM (SSVM) solver.

Our goal is to learn a weight for each potential function, given a set of training samples S with labels $y^s \in \{-1, 1\}$. For each training sample, a feature vector $\mathbf{\Phi}(\boldsymbol{\phi}^s, \boldsymbol{\psi}^s)$ with the concatenation of all potential functions is formed. Finally, the $0-1$ loss function has been chosen and integrated into the energy function, which is given by:

$$\min_{\mathbf{w}} \frac{1}{2} \parallel \mathbf{w} \parallel^2 + \frac{C}{S} \sum_{s=1}^{S} \xi^s$$
$$s.t. \quad \max(0, 1 - y^s \langle \mathbf{w}, \mathbf{\Phi}(\boldsymbol{\phi}^s, \boldsymbol{\psi}^s) \rangle) \leqslant \xi^s. \tag{9}$$

where ξ^s are the slack variables and C a constant. Finally, the optimisation is done with the cutting plane algorithm [41]. Margin-based parameter estimation has been successfully applied in segmentation [42], [43] and more recently in 2D human pose estimation [20] with different type of loss functions. During the experiments, we have observed that the $0-1$ loss has fitted well to our problem. Moreover, during the evaluation we demonstrate that weighting the potential function has an important impact on the final result.

3.3 Inference of multiple humans

The final step for obtaining the 3D pose of multiple individuals 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 potential functions count for these problems and penalize each hypothesis accordingly, allowing us to parse each human correctly.

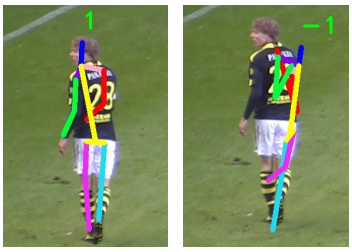
Here, we seek to estimate the posterior probability (1) using belief propagation. Since our graphical model does not have a tree structure, we employ the loopy belief propagation algorithm [44] to estimate the marginal distributions of the body part hypotheses. Sampling a solution directly from the marginals is not possible, since the hypotheses of different individuals lie together in the same state space. For that reason, we introduce a human localisation prior $\mathbf{h}$ for sampling from the marginals. We first localise each individual in each view and associate the localized bounding boxes across different views in order to recover the identity of each individual. To that end, we could use a human detector similar to [7], but we employ a multi-view human tracker [10] for more accurate bounding-box localisation and individual across-view association, as in [45]. Given the localization bounding boxes $\mathbf{h}$ of all individuals across all views, we look for the samples of each individual with the highest probability from the marginals:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}, \mathbf{p}, \mathbf{h}) \tag{10}$$

where $\hat{\mathbf{y}}$ is a subset of hypotheses that corresponds to the body parts of each individual. For each individual and for each body part, we look for the sample with the highest probability which at the same time is projected inside the localization bounding boxes across all or most views. Since the number of hypotheses in the state space is limited and the marginals are sorted, the above computation is inexpensive. Gradually, the 3D body poses of all individuals are parsed.

Despite the fact that the tracker provides each individual's identity, we do not make use of it for forming a local state space for each individual, in the beginning. Instead, we prefer to build a global state space by triangulating all the possible combinations of body part detections between view pairs. The reason is that the localisation can result in bounding boxes $\mathbf{h}$ with body parts of different individuals in case of occlusion. In that case, the inference from a local state space would result on inferred poses with body parts from different individuals, which would still look realistic. However, this problem does not occur with the global state space, where our model resolves the ambiguities between different individuals with the geometric potential functions.

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



Positive Negative

Fig. 6: **Training sample**: On the left column a positive training sample is presented, while on the right column a negative one. We choose negative samples which form reasonable human poses, instead of randomly sampling from the image space.

4 Results

In this section, we evaluate our approach on single and multiple human pose estimation on four datasets. Our model is composed of several potential functions which contribute differently to the final result. At first, we perform an evaluation of the potential functions and afterwards compare our method to related approaches. For single human pose estimation, we use the HumanEva-I [9] and KTH Multiview Football II [3] datasets, while we evaluate on the Campus [10] and Shelf [7] datasets for multiple human pose estimation.

The model that we employ for the experiments is composed of 14 body parts (Figure 3). For each evaluation dataset, we use the training data to learn our model's unary, pairwise and ternary terms as well as the model parameters. For learning the parameters of the model, we generate positive and negative examples according to the ground-truth of each dataset. On one hand, we consider as positive, samples that are very close to the ground-truth body pose in each view. On the other hand, the negative samples still form human body poses, but they do not correspond to the correct one, as it is depicted by Figure 6. Our part detector is based on the 2D part detector of [1] for all datasets other than KTH Multiview Football II [3]. In the KTH Multiview Football II dataset, there is a big amount of 2D training data which allows us to train a deep part detector similar to [22], [23]. The human localisation is done with the tracker of [10]. For all the experiments, we have set the number of loaded samples ($n_{samples}$) of the 2D part detector to 40, since we have experimentally observed that it is a good compromise between performance and speed. Finally, we employ the the PCP (percentage of correctly estimated parts) performance measure for evaluating our results. During the evaluation, the PCP score of a limb, defined by a pair of joints, is considered correct if the distance between the two predicted joint locations and true limb joint locations is at most 50% of the length of the ground-truth joints, as in [46].

4.1 Potential functions contribution

The 3DPS models is composed of unary, pairwise and ternary potential functions. We perform an in-depth evaluation of the potential functions and analyse their contribution

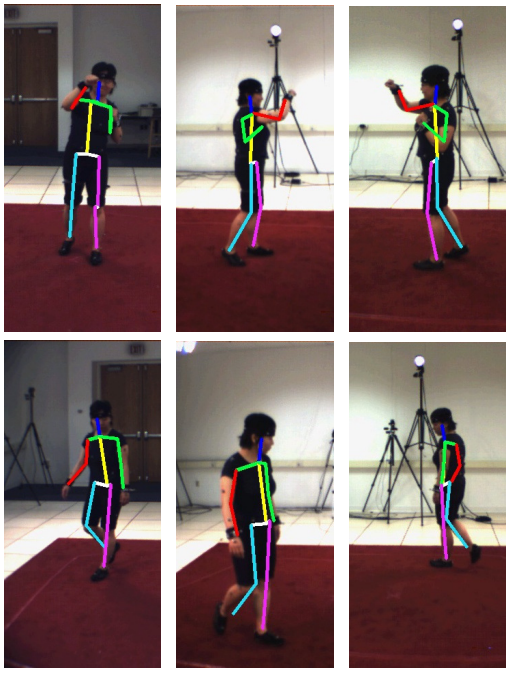


Fig. 7: **HumanEva-I dataset**: The 3D estimated body pose is projected across each view for the Box and Walking sequences.

to the model. On single human pose estimation, we use the KTH Multiview Football II [3] dataset for analysing the behaviour of the potential functions. On multiple human pose estimation, we choose the Campus [10] and Shelf [7] datasets. Since, we observe different human motion across the individuals in both datasets, the evaluation is done for each individual separately.

Our model has in total seven potential functions: the confidence, reprojection, visibility, temporal consistence, translation, collision and the rotation. We start with a basic model that is composed of a single potential function. Gradually, we aggregate in the model all the potentials and report the performance of different body parts at each step. Finally, the full model is composed of all unary, pairwise and ternary terms. The results are summarised in Tables 1 and 2. Below, it follows an analysis for each term separately. In addition, the results for each dataset are summarised in Figure 8.

Confidence: This is the most important potential function of the 3DPS model. The part-detectors' confidence contributes to the state space generation and confidence potential as well. In general, the performance of the part-detectors' is reflected in this potential function. As a result, a weak part detector would have a big effect on the whole performance. However, this is not the case in our model. The two/multi-view potentials and strong body prior efficiently penalises 3D hypotheses, which occurred from false positive part detections. For instance, the weak performance of part detectors in the Campus dataset for Actor 2 and 3 (Table 1c and 1d) is surpassed with the use of the other potentials. On the other hand, the already good performance of the part detectors in the Shelf dataset (Table 2a, 2b and 2c) or KTH Multiview Football II (Table 1a) has the most dominant influence to the final result, which does not improve significant by adding the other potential functions.

Reprojection & visibility: The reprojection error potential function makes use of two-views for estimating a score, while the visibility makes use of all views. Consequently, these terms are affected by the accuracy of the camera calibration. Moreover, geometric ambiguities due the camera poses (e.g. opposite cameras) can negatively affect the behaviour of these potentials (Figure 11). This is particularly the case for the reprojection potential in the Shelf dataset where the lower arms of Actor 1 (Table 2a) or the upper arms of Actor 2 (Table 2b) loose some performance due to geometric ambiguities. This type of ambiguities occur less often by better camera positioning (e.g. Campus dataset) or employing more views. Finally, the visibility potential, which relies on all views, always improves the final result.

Temporal consistency: Sustaining the consistency within the inferred body poses has a positive impact to the performance of our model, for most of the cases. In particular, the temporal consistence term has bigger contribution to the model in cases of uniform motion. For instance, the performance is improved more in walk gait (Table 1c and 1d) than playing football (Table 1a). However, the performance of our model on some body parts (e.g. Actor 1 in Campus - Table 1b) decreases by adding the temporal consistence term. The reason for this result is the threshold c that we have set during training. This threshold defines if an inferred part will be considered as correct or not. In order to keep the number of the model parameters low, we have set a single c for all body parts and all evaluation datasets as well. The motion of the body parts is nevertheless different. Furthermore, the motion between different individuals varies as well. Therefore, setting a single threshold for all body parts is not optimal, but it guarantees a more generic model.

Translation & collision: The body prior is divided into two pairwise and a ternary potentials. The most influential prior term is the translation potential function. It improves the performance in all datasets. The second pairwise term, the collision potential, helps to identify 3D hypotheses which came out from the triangulation of false positive detections of symmetric body parts. In some cases, the collision potential has small contribution or cuts down the performance of the upper legs (i.e. Actor 3 of Shelf dataset - Table 2c) because of false positive detections, which still fit well to the human body model.

Rotation: This ternary term requires triplets of body parts in order to be computed. Thus, it is the most expensive potential, in terms of computations. However, the contribution of this potential to the final result is not proportional to its cost. For example, the performance is improved around 1% in the Shelf dataset (Table 2a, 2b and 2c), while the improvement is much less in Campus (Table 1b, 1c and 1d), where the pose variation is confined to walking. On the other hand, it appears to be more valuable in the case

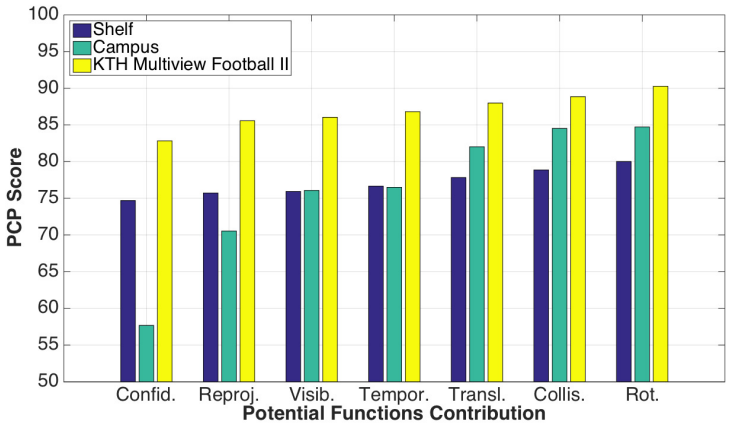


Fig. 8: **Potentials' contribution**: The contribution of each potential function is presented for the KTH Multiview Football II [3], Campus [10], Shelf [7] datasets. The performance measurement is the PCP score. The horizontal axis corresponds to the aggregation of the potential functions (confidence, reprojection, visibility, trermporal consistency, translation, collision, rotation). For the Campus and Shelf datasets, the average PCP score of all individuals is presented. Adding more potential functions to the base model (only confidence) gives considerable improvement to the KTH Multiview Football II and Campus datasets, while the improvement is smaller in the Shelf dataset.

of large body pose variation such as in the KTH Multiview Football II dataset. However, there are some cases in which the rotation potential reduces the performance of some body parts due to incorrect hypotheses which fit well to the rotation prior model.

Overall performance: Through the above analysis of the potential functions, fruitful conclusions are drawn. In general, the confidence of the part detectors is very crucial to the final result, but a weak part detector can be significantly refined using our 3DPS model. Moreover, the reprojection error potential is more affected from the camera pose, while the visibility term compensates in cases of geometric ambiguity. Finally, the body prior mainly benefits from the translation and collision potentials, while the rotation potential contributes more in case of large body pose variation (Figure 8). In order to modulate less the model, we have used the same prior model, in terms of the translation, rotation or collision, for all body parts. However, the results highlight that some body parts do not benefit in all cases from this assumption. For example, using a single Gaussian distribution as rotation potential has reduced the head performance for half of the examined cases. Thus, a combination of individual prior models for different body parts could result in better performance, but create a less generic model.

Discussion: One fundamental assumption of the 3DPS model is a calibrated multi-view setup. As it is reflected from the results, calibration errors or geometric ambiguities influence the model performance. Thus, a further step would be to assume an uncalibrated setup, where the goal would be to infer both the 3D body pose and camera pose at the same time [47]. Furthermore, more robust part detectors would have a big impact on the model performance. At this stage, deep learning methods [20], [48] could contribute to more robust part detections for a single-view or combined views.

Fig. 9: **KTH Multiview Football II dataset**: The 3D estimated body pose is projected across each view. The results comes from the inference with all cameras.

4.2 Single human pose estimation

We evaluate our method on single human 3D pose estimation for demonstrating that it performs as well as start-of-the-art multi-view approaches [1], [3]. The purpose of this experiment is to highlight that we can achieve similarly good or even better results than other methods with an enriched, in terms of potentials, human model.

HumanEva-I: We evaluate on Box and Walking sequences (Figure 7) of the HumanEva-I [9] dataset and compare with [1] and [6]. We share similar appearance term only for the 2D single view part detection with [1] and employ different body models. Table 3 summarizes the results of the average 3D joint error in millimetres. In this dataset, we have used the aforementioned evaluation measure in order to keep up with the related work. Notably, Amin et al. [1] report very low average error, and we also

TABLE 1: **Potentials's aggregation**: The aggregated PCP (percentage of correctly estimated parts) scores are presented for the potential functions. Each column corresponds to an additional potential function.

(a) **KTH Multiview Football II**

	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Upper Arms	89.49	92.06	92.29	93.69	96.26	96.83	97.96
Lower Arms	56.78	63.55	64.02	65.89	66.82	68.93	71.86
Upper Legs	96.50	97.43	97.66	97.90	98.83	99.30	99.40
Lower Legs	88.55	89.25	90.19	89.72	90.02	90.32	91.80
Average	82.83	85.57	86.04	86.80	87.98	88.85	90.26

(b) **Campus - Actor 1**

	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Head	68.97	96.55	96.55	100.00	100.00	100.00	96.55
Torso	75.86	82.76	89.66	93.10	89.66	89.66	93.10
Upper Arms	63.79	86.21	93.10	75.86	96.55	98.28	96.55
Lower Arms	53.45	72.41	75.86	65.52	86.21	91.38	86.21
Upper Legs	79.31	87.93	96.55	89.66	96.55	93.10	93.10
Lower Legs	82.76	91.38	89.66	87.93	89.66	89.66	96.55
Average	70.69	86.21	90.23	85.35	93.11	93.68	93.68
All parts	70.35	85.52	89.66	83.10	92.76	93.45	93.45

(c) **Campus - Actor 2**

	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Head	74.12	87.65	94.71	91.18	100.00	100.00	98.24
Torso	41.18	46.47	47.06	45.88	46.47	47.65	48.82
Upper Arms	66.76	76.47	81.47	89.41	89.12	93.82	97.35
Lower Arms	13.24	17.94	20.88	33.82	32.35	41.47	42.94
Upper Legs	60.00	65.29	69.41	68.53	75.59	75.59	75.00
Lower Legs	86.76	87.65	87.65	90.00	90.29	90.59	89.41
Average	57.01	63.58	66.86	69.80	72.30	74.85	75.29
All parts	56.88	62.88	66.06	70.06	72.12	75.06	75.65

(d) **Campus - Actor 3**

	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Head	30.69	67.65	81.55	80.58	89.32	94.17	93.20
Torso	30.69	56.86	69.90	73.79	79.61	83.50	85.44
Upper Arms	51.49	66.67	75.73	79.13	79.61	88.83	89.81
Lower Arms	41.09	50.49	59.71	66.99	68.45	77.18	74.76
Upper Legs	60.40	66.67	75.24	80.10	85.92	84.95	91.75
Lower Legs	57.43	62.75	64.56	65.53	81.07	81.55	76.21
Average	45.30	61.85	71.12	74.35	80.66	85.03	85.20
All parts	48.22	61.77	70.19	73.79	79.90	84.27	84.37

TABLE 3: **Human-Eva I**: The average 3D joint error in millimetres (mm) is presented.

Sequence	Walking	Box
Amin et al. [1]	54.5	47.7
Sigal et al. [6]	89.7	-
Proposed	68.3	62.7

TABLE 4: **KTH Multiview Football II**: The PCP (percentage of correctly estimated parts) scores using 2 and 3 cameras are presented. One can observe that we have mainly better results for the upper limbs. In addition, learning the parameters of the CRF helps to improve the final result in comparison to [7].

	2 Cameras			3 Cameras			
	Bur. [3]	Bel. [7]	Proposed	Bur. [3]	Bel. [7]	Kaz. [38]	Proposed
Upper Arms	53	64	96	60	68	89	98
Lower Arms	28	50	68	35	56	68	72
Upper Legs	88	75	98	100	78	100	99
Lower Legs	82	66	88	90	70	99	92
Average	62.7	63.8	87.5	71.2	68.0	89.0	90.3

achieve similar results. Cases in which we have observed failures are related to lack of correct detected parts from at least two cameras.

KTH Multiview Football II: In this dataset, we evaluate on Player 2 as in the original work of Burenius et al. [3] and the follow up work of [38]. We follow the same evaluation protocol and estimate the PCP (percentage of correctly estimated parts) scores for each set of cameras (Figure 9). The results are summarized in Table 4. We outperform the method of [3] on both cameras setups, using a richer body model and a radically smaller state space (Figure 5). In [3], the 3D space is discretised in terms of rotation and translation at different discretisation levels. However, a fine discretisation is required for accurate results. On the other hand, our discrete state space is significantly smaller without the cost of a reduced performance. Finally, the more accurate part detectors of [38] improve the results, but we still obtain superior performance.

In addition, learning the parameters of the model brings

TABLE 2: **Potentials's aggregation**: The aggregated PCP (percentage of correctly estimated parts) scores are presented for the potential functions. Each column corresponds to an additional potential function.

(a) **Shelf - Actor 1**

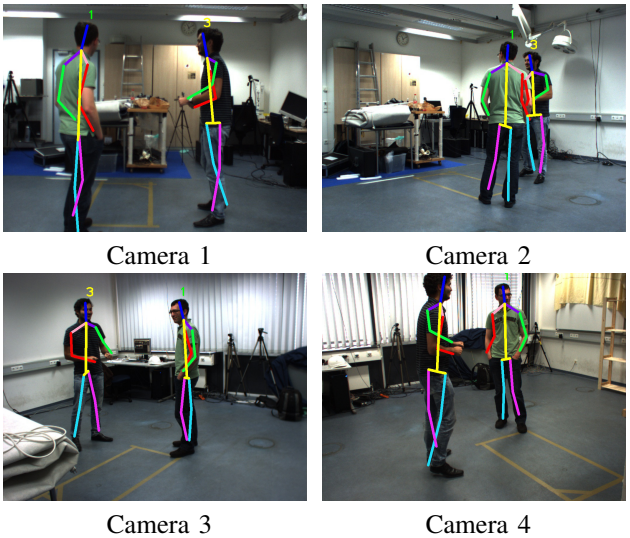
	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Head	92.39	94,02	94.02	94.57	95.65	96.17	96.29
Torso	100.00	100,00	100.00	100.00	100.00	100.00	100.00
Upper Arms	79.08	80,71	80.43	80.98	80.16	82.24	82.24
Lower Arms	57.34	56,79	58.97	62.23	62.50	65.30	66.67
Upper Legs	40.49	40,76	40.49	41.58	44.29	42.90	43.17
Lower Legs	80.43	81,25	81.25	82.34	85.60	85.79	86.07
Average	74.96	75,59	75.86	76.95	78.03	78.73	79.07
All parts	70.71	71,30	71.63	72.88	74.08	74.86	75.26

(b) **Shelf - Actor 2**

	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Head	68.42	57.89	57.89	57.89	68.42	68.42	78.95
Torso	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Upper Arms	84.21	81.58	81.58	84.21	84.21	84.21	82.58
Lower Arms	31.58	36.84	36.84	36.84	34.21	42.11	47.37
Upper Legs	47.37	47.37	47.37	47.37	47.37	50.00	50.00
Lower Legs	73.68	76.32	76.32	76.32	78.95	78.95	78.95
Average	67.54	66.67	66.67	67.11	68.86	70.62	72.98
All parts	64.21	64.21	64.21	64.74	65.79	67.90	69.68

(c) **Shelf - Actor 3**

	Unary				Pairwise		Ternary
	Confidence	Reprojection	Visibility	Temporal	Translation	Collision	Rotation (Full model)
Head	74.00	92.00	94.00	94.00	95.00	96.00	98.00
Torso	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Upper Arms	90.00	90.00	90.00	91.00	91.50	92.00	93.15
Lower Arms	85.00	86.50	86.00	89.50	88.00	90.00	92.30
Upper Legs	49.50	50.00	50.00	50.00	50.00	52.30	56.50
Lower Legs	91.00	91.00	91.50	91.00	95.00	96.00	97.00
Average	81.58	84.92	85.25	85.92	86.58	87.72	89.49
All parts	80.50	82.70	82.90	83.70	84.40	85.66	87.59



Camera 1 Camera 2

Camera 3 Camera 4

Fig. 10: **Shelf dataset**: Our results projected in 4 out of 5 views of the Shelf dataset [7].

a considerable improvement in comparison to our previous work [7]. Our approach runs on around 1 fps for single human 3D pose estimation, given the 2D detections. All the experiments are carried out on a standard Intel i7 2.40 GHz laptop machine and our method is implemented in C++ with loop parallelizations.

4.3 Multiple human pose estimation

The problem of multiple human 3D pose estimation has not been extensively addressed yet. Most of the related work has focused on single human 3D pose estimation [1], [3], [6]. Moreover, the number of available datasets in the literature for multiple human pose estimation from multiple views is very limited. Recently, we have proposed two multiple human datasets [7]: the Shelf and Campus. We evaluate our method on these datasets using the PCP (percentage of correctly estimated parts) and compare with related work as well. Since we are not aware of another method which performs multiple human 3D pose estimation, we choose single human approaches [1], [49] 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.

Campus: In this dataset, three different individuals (Figure 2) share a common state space. In [7], we have demonstrated that putting the 3D hypotheses of all individuals to the same state space does not result in a reduced performance using the 3DPS model. In addition, in this work we show that learning the parameters of the model improves further the final result (Table 5a). The performance of the 3DPS model on Actor 1 distinguishes itself from the

other two for the accurate body pose estimation in most of the evaluated frames. Actor 2 follows with similar results, while Actor 3 looses some performance due to the weak localisation of the torso and lower arms. The reason for the reduced performance can be found in the analysis of the potential functions in Table 1c. It is observed that the part detectors for the torso and lower arms are weak for the Actor 2. In comparison to [1] where the inference is done separately for each Actor, we have in general better limb localisation and we perform similar for the rest of the body parts.

Shelf[1]**:** Similarly to the Campus dataset, three individuals compose the Shelf dataset (Figure 1 and 10). The head and torso are localised correctly for all individuals for most of the time, while the lower arms and upper legs are the most difficult parts to localise. Going back to the analysis of the potential function in the Tables 2a,2b and 2c, ones observes weak behaviour of the part detectors for these body parts. This is a common fact for all three Actors. Comparing to [49], we perform mainly better on the arms and lower legs. Furthermore, the inference is done separately for each individual in [49], while we assume a common state space for all individuals. Finally, we demonstrate in this dataset as well that parameter learning using a Structured SVM (SSVM) considerably improves the final result in comparison for our earlier work on 3DPS [7].

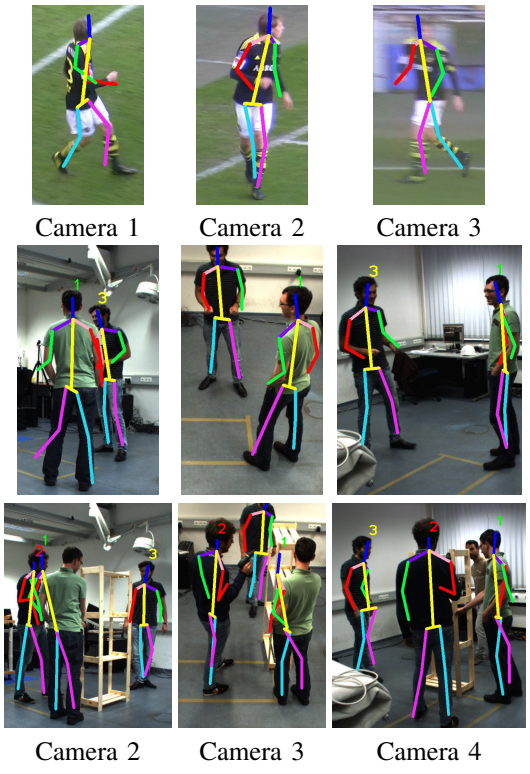


Fig. 11: **Failure cases**: On the top row, it is presented a wrongly inferred body pose due to geometric ambiguities and false part localisation (KTH Multiview Football II dataset). On the middle row, the lower limb of Actor 1 looks correct from Camera 3 and 4, but it is actually wrongly localised again due to geometric ambiguity (Shelf dataset). On the bottom, the body pose of Actor 1 is wrong due to 3D hypotheses which occurred from false positive part detections.

5 Conclusion

We have presented a 3D pictorial structures (3DPS) model for recovering the 3D human body pose of multiple individuals from multiple camera views. We have introduced a discrete state space which allows fast inference. Our model is composed of a set of potential functions, which make use of two- and multi-view observations. To weight correctly, the influence of the potential functions, we have used a Structured SVM (SSVM) to learn the parameters of our model. Our model has been applied to multiple humans without knowing the identity in advance. Self and natural occlusions can be handled by our algorithm, while we rely only on noisy part detectors for each view. The model is also applicable to single human pose estimation where we have demonstrated state-of-the-art results.

Acknowledgments

This work was funded by the DFG Project "Advanced Learning for Tracking and Detection in Medical Workflow Analysis". The authors would like to thank Xinchao Wang for providing the tracking information, as well as, Maximilian Baust and Peter Gehler for their helpful discussions.

References

[1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view pictorial structures for 3d human pose estimation," in *British Machine Vision Conference*, vol. 2, 2013.

1. The dataset and additional material is available at: http://campar.in.tum.de/Chair/MultiHumanPose.

[2] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 623–630.

[3] M. Burenius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3618–3625.

[4] J. R. Mitchelson and A. Hilton, "Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling." in *BMVC*, 2003, pp. 1–10.

[5] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[6] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International journal of computer vision*, vol. 98, no. 1, pp. 15–48, 2012.

[7] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D pictorial structures for multiple human pose estimation," in *Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on*. IEEE, 2014.

[8] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 104.

[9] 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," *International journal of computer vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[10] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *Pattern Analysis and*

TABLE 5: **State-of-the-art comparison**: The PCP (percentage of correctly estimated parts) scores are presented for different related work and the proposed method. The global score of all individuals takes additionally into consideration the number of occurrence for each individual.

(a) **Campus dataset**

	Amin et al. [1]			Belagiannis et al. [7]			Proposed		
	Actor 1	Actor 2	Actor 3	Actor 1	Actor 2	Actor 3	Actor 1	Actor 2	Actor 3
Head	64.58	78.84	38.52	93.62	97.40	81.26	96.55	98.24	93.20
Torso	100.00	100.00	100.00	49.94	41.13	69.67	93.10	48.82	85.44
Upper Arms	94.80	84.66	83.71	82.85	90.36	77.58	96.55	97.35	89.81
Lower Arms	66.67	27.25	55.19	77.80	39.65	61.84	86.21	42.94	74.76
Upper Legs	100.00	98.15	90.00	86.23	73.87	83.44	93.10	75.00	91.75
Lower Legs	81.25	83.33	70.37	91.39	89.02	70.27	96.55	89.41	76.21
All body parts	85.00	76.56	73.70	82.01	72.43	73.72	93.45	75.65	84.37
Average (Actors)		78.42			75.79			84.49	
All individuals (global PCP)		76.61			73.82			81.08	

(b) **Shelf dataset**

	Amin et al. [49]			Belagiannis et al. [7]			Proposed		
	Actor 1	Actor 2	Actor 3	Actor 1	Actor 2	Actor 3	Actor 1	Actor 2	Actor 3
Head	93.75	100.00	85.23	89.30	72.10	94.66	96.29	78.95	98.00
Torso	100.00	100.00	100.00	90.20	92.80	96.35	100.00	100.00	100.00
Upper Arms	73.08	73.53	86.62	72.16	80.11	91.00	82.24	82.58	93.15
Lower Arms	32.99	2.94	60.31	60.59	44.20	89.00	66.67	47.37	92.30
Upper Legs	85.58	97.06	97.89	37.12	46.30	45.80	43.17	50.00	56.50
Lower Legs	73.56	73.53	88.73	70.61	71.80	94.50	86.07	78.95	97.00
All body parts	72.42	69.41	85.23	66.05	64.97	83.16	75.26	69.68	87.59
Average (Actors)		75.69			71.39			77.51	
All individuals (global PCP)		77.3			71.75			79.00	

Machine Intelligence, IEEE Transactions on, vol. 33, no. 9, pp. 1806–1819, 2011.

[11] 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*, vol. 104, no. 2, pp. 90–126, 2006.

[12] L. Sigal and M. J. Black, "Guest editorial: state of the art in image- and video-based human pose and motion estimation," *International Journal of Computer Vision*, vol. 87, no. 1, pp. 1–3, 2010.

[13] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 1, pp. 44–58, 2006.

[14] K. Grauman, G. Shakhnarovich, and T. Darrell, "Inferring 3d structure with a statistical image-based shape model," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 641–647.

[15] M. Hofmann and D. M. Gavrila, "Multi-view 3d human pose estimation in complex environment," *International journal of computer vision*, vol. 96, no. 1, pp. 103–124, 2012.

[16] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3d human motion estimation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 390–397.

[17] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3d human pose tracking," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 631–638.

[18] A. Yao, J. Gall, L. V. Gool, and R. Urtasun, "Learning probabilistic non-linear latent variable models for tracking complex activities," in *Advances in Neural Information Processing Systems*, 2011, pp. 1359–1367.

[19] J. Lallemand, O. Pauly, L. Schwarz, D. Tan, and S. Ilic, "Multi-task forest for human pose estimation in depth images," in *3DTV-Conference, 2013 International Conference on*. IEEE, 2013, pp. 271–278.

[20] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in Neural Information Processing Systems*, 2014.

[21] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in Neural Information Processing Systems*, 2014.

[22] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on*. IEEE, 2014.

[23] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision–ACCV 2014*, 2014.

[24] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *CVPR, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, 1998, pp. 8–15.

[25] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, 2005.

[26] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel, "Optimization and filtering for human motion capture," *International journal of computer vision*, vol. 87, no. 1-2, pp. 75–92, 2010.

[27] R. Plankers and P. Fua, "Articulated soft objects for multi-view shape and motion capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, 2003.

[28] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3d human figures using 2d image motion," in *Computer VisionâĂŤECCV 2000*. Springer, 2000, pp. 702–718.

[29] M. W. Lee and R. Nevatia, "Human pose tracking using multi-level structured models," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 368–381.

[30] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1208–1221, 2004.

[31] K. Alahari, G. Seguin, J. Sivic, and I. Laptev, "Pose estimation and segmentation of people in 3d movies," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2112–2119.

[32] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[33] ——, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1014–1021.

[34] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 228–242.

[35] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[36] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, 1973.

[37] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1385–1392.
[38] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multiview body part recognition with random forests," in *British Machine Vision Conference*, 2013.
[39] E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," *Communications of the ACM*, vol. 53, no. 10, pp. 95–103, 2010.
[40] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
[41] T. Finley and T. Joachims, "Training structural svms when exact inference is intractable," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 304–311.
[42] A. Lucchi, Y. Li, K. Smith, and P. Fua, "Structured image segmentation using kernelized features," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 400–413.
[43] S. Nowozin, P. V. Gehler, and C. H. Lampert, "On parameter learning in crf-based approaches to object class image segmentation," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 98–111.
[44] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 1.
[45] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, and N. Navab, "Multiple human pose estimation with temporally consistent 3D pictorial structures," in *Computer Vision–ECCV 2014, ChaLearn Looking at People Workshop*. Springer, 2014.
[46] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.
[47] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt, "Outdoor human motion capture by simultaneous optimization of pose and camera parameters," in *Computer Graphics Forum*. Wiley Online Library, 2014.
[48] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision–ECCV 2014*. Springer International Publishing, 2014, pp. 297–312.
[49] S. Amin, P. Müller, A. Bulling, and M. Andriluka, "Test-time adaptation for 3d human pose estimation," in *German Conference on Pattern Recognition (GCPR/DAGM)*, Münster, Germany, September 2014.

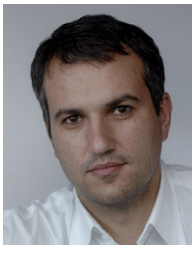
Vasileios Belagiannis is PhD student at Technische Universität München (TUM) and part of the computer vision group of the Computer Aided Medical Procedures (CAMP) institute. He received his masters in Computational Science and Engineering from Technische Universität München, Germany in 2011 and his Diploma in Production Engineering from Democritus University of Thrace, Greece in 2009. His main research interests are computer vision, machine learning and data analysis. In particularly, his current work is focused on multiple human pose estimation in real-world environments, activity recognition and monocular object tracking and detection.

Sikandar Amin has received his Master's degree in Communications Engineering from Technische Universität München, Germany in 2009. Currently, he is a PhD candidate in Intelligent Autonomous Systems group in Technische Universität München, Germany. Since 2013, he is working as a visiting researcher with the Max Planck Institute for Informatics in Saarbrücken, Germany. His research interests include computer vision and machine learning. Specifically, he is working on 2D and 3D human pose estimation in complex scenes for higher level tasks including activity recognition and studying human emotions during dyadic interactions in challenging real-world settings.

Mykhaylo Andriluka has received his Bachelor in Applied Mathematics from Odessa I.I.Mechnikov National University, Ukraine in 2001. He has also received his Diplom in Mathematics from Darmstadt University of Technology in 2006. In 2010, he obtained his PhD from Darmstadt University of Technology in computer vision. Since 2011, he is research scientist in Max Planck Institute for Informatics in Saarbrücken, Germany. Currently, he is visiting assistant professor in Stanford University, USA. His research interest include human pose estimation, people tracking-by-detection and machine learning for computer vision.

Bernt Schiele received his masters in computer science from Univ. of Karlsruhe and INP Grenoble in 1994. In 1997 he obtained his PhD from INP Grenoble in computer vision. He was a postdoctoral associate and Visiting Assistant Professor at MIT between 1997 and 2000. From 1999 until 2004 he was an Assistant Professor at ETH Zurich and from 2004 to 2010 he was a full professor of computer science at TU Darmstadt. In 2010, he was appointed scientific member of the Max Planck Society and a director at the Max Planck Institute for Informatics. Since 2010 he has also been a Professor at Saarland University. His main interests are computer vision, perceptual computing, statistical learning methods, wearable computers, and integration of multi-modal sensor data. He is particularly interested in developing methods which work under real-world conditions.

Nassir Navab is a professor of computer science and founding director of Computer Aided Medical Procedures and Augmented Reality (CAMP) laboratories at Technische Universität München (TUM) and Johns Hopkins University (JHU). He also has faculty appointments at TUM and JHU medical schools. He received the Ph.D. degree from INRIA and University of Paris XI, Paris, France, and enjoyed two years of postdoctoral fellowship at MIT Media Laboratory before joining Siemens Corporate Research (SCR) in 1994. At SCR, he was a distinguished member and received the Siemens Inventor of the Year Award in 2001. In 2012, he was elected as a fellow member of MICCAI society. He has served on the program committee of over 80 international conferences. His current research interests include robotic imaging, computer aided surgery, computer vision and augmented reality.



Slobodan Ilic is currently senior key expert research scientist at Siemens Corporate Technology in Munich, Perlach at the Sensor Technology Department. He is also a visiting researcher and lecturer at Computer Science Faculty of TUM and closely works with the Vision Group at the CAMP Chair, where he supervises a number of PhD students. From 2009 until end of 2013 he was senior researcher and leader of the Computer Vision Group of CAMP at TUM, and before that he was a senior researcher at Deutsche Telekom Laboratories in Berlin. In 2005 he obtained his PhD at EPFL in Switzerland under supervision of Pascal Fua. His research interests include: 3D reconstruction, deformable surface modelling and tracking, real-time object detection and tracking, human pose estimation and semantic segmentation.