Multiple Human Pose Estimation with Temporally Consistent 3D Pictorial Structures

Vasileios Belagiannis[1], Xinchao Wang[2], Bernt Schiele[3],

Pascal Fua[2], Slobodan Ilic[1,4], Nassir Navab[1,5]

[1]Computer Aided Medical Procedures, Technische Universität München, Germany
{belagian, slobodan.ilic, navab}@in.tum.de
[2]Computer Vision Laboratory, EPFL, Lausanne, Switzerland
{xinchao.wang, pascal.fua}@epfl.ch
[3]Max Planck Institute for Informatics, Saarbrücken, Germany
{schiele}@mpi-inf.mpg.de
[4]Siemens AG, Munich, Germany
[5]Computer Aided Medical Procedures, Johns Hopkins University, USA

Abstract. Multiple human 3D pose estimation from multiple camera views is a challenging task in unconstrained environments. Each individual has to be matched across each view and then the body pose has to be estimated. Additionally, the body pose of every individual changes in a consistent manner over time. To address these challenges, we propose a temporally consistent 3D Pictorial Structures model (3DPS) for multiple human pose estimation from multiple camera views. Our model builds on the 3D Pictorial Structures to introduce the notion of temporal consistency between the inferred body poses. We derive this property by relying on multi-view human tracking. Identifying each individual before inference significantly reduces the size of the state space and positively influences the performance as well. To evaluate our method, we use two challenging multiple human datasets in unconstrained environments. We compare our method with the state-of-the-art approaches and achieve better results.

Keywords: human pose estimation, 3D pictorial structures, part-based pose estimation.

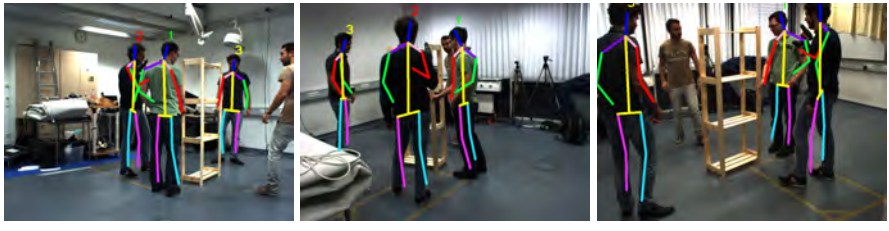
1 Introduction

The problem of human pose estimation has drawn the attention of computer vision researchers for many years. Determining the body pose of multiple human has a wide range of potential applications such as motion capture, activity recognition and human interaction. In every application, the human motion remains consistent over time. Different approaches have been proposed for multiple human pose estimation on 2D [1, 4, 10, 24] or 3D space [5, 6, 14, 20, 21]. Nevertheless, the utility of temporal consistency for pose estimation has not been sufficiently addressed yet.

Defining the body pose as a constellation of parts has become the standard model in human pose estimation using image data [4, 24, 25]. Pictorial structures is the most common part-based model for estimating the 2D body pose of single human [3, 11, 13]. The model has been extended to the 3D space, in order to cope with mutli-view camera setups as well [2, 9, 16]. Recently, pictorial structures have been successfully modelled for multiple human 3D pose estimation [6]. However, the temporal consistency, between the estimated body poses of different individuals among subsequent frames, has never been addressed within the framework of pictorial structures. This property is of high importance in multiple human pose estimation, where the trajectory of each individual is directly connected to the body pose.

In this work, we propose a temporally consistent 3D Pictorial Structures model (3DPS) for multiple human pose estimation from multiple camera views. We build our model on a Conditional Random Field (CRF), which is composed of unary, pairwise and ternary potential functions. The unary potentials incorporate the observation to our model, while the pairwise and ternary model the human body as a prior. In addition, we introduce the temporal function based on an additional potential that ensures temporal consistency between the human poses over time. In order to propagate the inferred poses with the correct identity, we rely on a multi-view state-of-the-art human tracker [7], which has been proven to work reliably and recently evolved to track interacting objects [28].

Our contributions are summarised as follows: We propose a temporally consistent 3DPS model which is applied to multiple human pose estimation. Our model takes input from a multi-object tracker, but it is not subjected to a particular one. Given the track of each individual, we signicantly reduce state space of each joint, which leads to much faster inference. On two challenging datasets for multiple human pose estimating from multiple views, we obtain the best results in comparison to previous work [6].



Camera 1 Camera 3 Camera 5

Fig. 1. Shelf dataset result: The 3D estimated body poses are projected across the camera views. The identity of each individual is derived by the tracker.

2 Related Work

The problem of multiple human pose estimation has been studied from a 2D and 3D perspective. Moreover, it has often been coupled with human tracking. We review the most related work and focus on multiple human pose estimation from multiple views. We refer the reader to [22, 26] for a more in-depth analysis of human pose estimation.

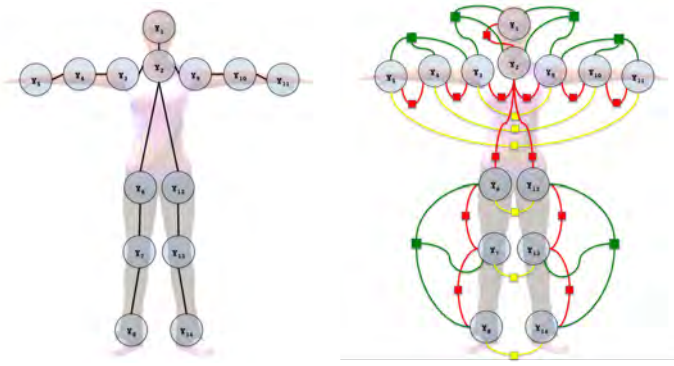
Part-based models have proved to be a powerful solution for 2D pose estimation [1, 4, 10, 24]. Among many of the proposed models, pictorial structures is the most widely used one [3, 11, 13]. In [24], the main focus has been articulated human tracking with manual initialization, based on the pictorial structures. Similarly in [4], pictorial structures have been integrated into a tracking-by-detection framework. More oriented towards pose estimation, the human interaction has been modelled using the pictorial structures framework in [10]. However, each frame has been treated independently without considering the temporal component. Recently, a segmentation approach has been combined with a part-based model for estimating the human pose from stereo data [1]. Despite the available 3D information, the final pose is in the 2D space.

In the 3D space, the problem of multiple human pose estimation has been addressed using monocular [5, 18, 29], stereo [14, 23] or multi-camera input information [21, 20]. In [29] and [18], the monocular 3D pose estimation has been combined with tracking. Both approaches rely on a blob detector, which can be unreliable for individuals with similar appearance. Richer appearance models have been introduced in [5] by building on the pictorial structures. Since the approach is monocular, the final 3D pose is inferred by 2D pose lifting. In [14], a two-stage algorithm is applied on stereo data for detecting human and recovering their pose. Similar to our framework, a multi-view system has been employed in [20, 21]. In [21], the proposed method can estimate the pose up to two people in a studio environment. Our model does not have such limitations and it is mainly applied to unconstrained environments. Finally, a model fitting approach has been proposed in [20]. The learned body model is fitted into a voxel representation.

The 3DPS model for multiple human pose estimation [6] is similar to our model. In this approach, a model for inferring multiple human body poses without knowing their identity has been proposed. However, keeping all the individuals in a common state space results in additional computations. Furthermore, the model does not consider the temporal consistency of the inferred poses and consequently cannot identify the individuals. In our method, we first recover the identity of each individual using tracking and afterwards infer the pose. Moreover, we introduce a temporal term for regularising our solution. This term keeps the inferred poses consistent over time. In the experiments we directly compare to [6] and observe that these differences lead not only to a significantly reduced state space and thus faster inference, but also to significantly improved performance. Finally, our model is advisable for multiple human pose estimation, where there are different individuals.

3 Method

In this section, we introduce the temporally consistent 3D pictorial structures (3DPS) model. We build our model on a Conditional Random Filed (CRF) that is composed of unary, pairwise and ternary potential functions. The unary potentials include geometric and appearance features, as well as temporal features that encode the identity and pose of the individuals. The pairwise and ternary potentials impose physical constrains and act as a body prior. In the following subsections, we present our model, the state space, the potential function (Subsection 3.1) and conclude with the inference of multiple human 3D poses (Subsection 3.2).



Body model Graph factors

Fig. 2. Graphical model of the human body: We used 14 variables in our graph to represent the body joints. On the left, the undirected graphical model that corresponds to the human body is presented. On the right, the graph factors are illustrated with different colours. The kinematic constrains are presented with red (translation) and green (rotation) edges (factors). The collision constrains are represented with yellow edges.

3.1 Temporally Consistent 3DPS Model

In the temporally consistent 3D pictorial structure (3DPS) model, the body is represented as an undirected graphical model, wherein each node corresponds to a body joint (Figure 2). The edges of the graph denote relation between the body joints. In comparison to the original pictorial structure model, we have interchanged the notion of the body parts with the joints to avoid the foreshortening effect [2]. Let the random variable Y_i denote a joint location in the global 3D coordinate system, i.e., $Y_i = [\delta_i, \pi_i, \vartheta_i]^T \in \mathbb{R}^3$. The body pose is then defined by the configuration $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, where n is the number of joints. Each random variable Y_i takes its values from a state space $\Lambda_i \in \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is

the global state space. We model the observations and interaction between the variables using a factor graph [17]. The factors are divided into unary, pairwise and ternary ones (Figure 2). The unary factors describe the relation between the state space and the random variables. The pairwise and ternary factors are categorized into kinematic and collision ones. The kinematic factors model the body prior, while the collision factors model the relation between symmetric body joints.

The posterior probability of the body configuration $\mathbf{y} \in \boldsymbol{\Lambda}$ given an observation instance $\mathbf{x} \in \mathbf{X}$ and the temporal body pose $\mathbf{p} \in \mathbf{P}$ of an individual is defined as:

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

where $Z(\mathbf{x})$ is the partition function, E_{kin} and E_{col} are the edges corresponding to the kinematic factors and the collision factors, respectively. The temporal body pose $\mathbf{P}$ corresponds to inferred poses from a previous time step. A set of joints forms the temporal pose $\mathbf{P}$ of an individual. The unary terms include the detection confidence $\phi_i^{conf}(y_i, \mathbf{x})$, reprojection error $\phi_i^{repr}(y_i, \mathbf{x})$, joint multi-view visibility $\phi_i^{vis}(y_i, \mathbf{x})$ and the temporal consistence potential functions $\phi_i^{temp}(y_i, p_i)$. The pairwise potential functions are divided into the translation $\psi_{i,j}^{tran}(y_i, y_j)$ and collision $\psi_{i,j}^{col}(y_i, y_j)$, while the rotation $\psi_{i,j,k}^{rot}(y_i, y_j, y_k)$ is modelled as a ternary potential function. Next, we first define the state space $\boldsymbol{\Lambda}$, then explain the potential functions and temporal consistency that we obtain by using tracking.

State space generation The state space Λ_i of a variable Y_i comprises the values that correspond to a candidate joint location in a discretised 3D space. In order to be computationally efficient, we discretise the 3D space using 2D body joint detectors. We first sample 2D joint locations for each view using a body joint detector [2]. Next, we create the 3D locations by triangulation of the corresponding 2D body joints detected in all combinations of view pairs. Our only assumptions are that the camera system is calibrated and there is at least one true positive detection of each joint from a view pair. Finally, the global state space $\boldsymbol{\Lambda} = \{\Lambda_1, \Lambda_2, \ldots \Lambda_n\}$ includes locations which correspond either to true or false positive body joints. The false positive candidate locations usually occur by triangulating one or two false positive joint detections. Knowing the identity of each individual significantly reduces the cardinality of the state space in comparison to [6].

Unary potentials Given the state space, we extract a set of features for each candidate joint 3D location. These features form the observation $\mathbf{X}$, which is provided to the unary potential functions. The first feature is the detection confidence, which is used by $\phi_i^{conf}(y_i, \mathbf{x})$. It corresponds to the mean confidence of the joint detector in the two views that have been used during triangulation. Note here, we consider all combinations of two views. To account for the errors of triangulation between two views [15], we introduce the reprojection error function $\phi_i^{repr}(y_i, \mathbf{x})$, which is defined as follows:

$$\phi_i^{repr}(y_i, \mathbf{x}) = \frac{1}{1 + \exp(C(y_i))} \quad (2)$$

where $C(y_i)$ is the reprojection error in two views, which is computed using the Euclidean distance similar to [6]. The first two potential functions reason about combinations of two views. To also profit from the multi-view setup, the multi-view visibility function $\phi_i^{vis}(y_i, \mathbf{x})$ computes the number of views in which a candidate has been correctly detected. To that end, we project the candidate joint 3D location to each view and search for a detection instance in a small radius (5 pixels) around the projected location. We then accumulate the number of visible views and normalize the result with respect to the total number of cameras. Thus we obtain a score between 0 and 1, which we take as $\phi_i^{vis}(y_i, \mathbf{x})$. Note that, candidates that resulted from ambiguous views or false positive detections are implicitly penalised by obtaining a low visibility score.

Temporal consistency The features of the observation are extracted from a single frame and thus do not account for temporal consistency. To assure consistency between the inferred joints, we introduce the temporal consistence function $\phi_i^{temp}(y_i, p_i)$. This is a potential that relies on a temporal body joint location p_i and acts as a regulariser. The aim of this function is to impose temporal consistency between the inferred poses. For this reason, it penalises candidates that geometrically differ significantly from the temporal joint. As temporal pose, we choose the inferred pose from the previous frame and assume small changes between two subsequent frames. However, the temporal poses can include wrongly inferred joints or identity. To address these problems, the geometric distance between the candidate and the temporal joint is not considered if it exceeds a threshold. The distance is expressed as a score using a sigmoid function and is given as follows:

$$\phi_i^{temp}(y_i, p_i) = \begin{cases} \frac{1}{1+\exp(d(y_i,p_i))} \text{ if } d(y_i, p_i) < c \\ \epsilon \text{ otherwise} \end{cases} \tag{3}$$

where ϵ is a small constant for numerical stability and $d(y_i, p_i)$ is the Euclidean distance between the candidate 3D joint location and the temporal joint and c the distance threshold, which we set to $10cm$. The temporal prior is another observation in our posterior (1). We choose this computationally inexpensive formulation instead of formalising the problem using a dynamic CRF [27].

Pairwise and ternary potentials The pairwise and ternary potential functions model the interaction between the random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ of our model. The interaction is interpreted as a body prior with kinematic and collision constraints. In the kinematic constraints, we impose physical constraints between the body joints. For that purpose, we define two types of transformation: translation and rotation. The collision constraints are introduced to mainly handle false positive detections. This phenomenon is usually observed in the symmetric joints where the classifier is occasionally triggered for the wrong joint.

The translation potential models a joint i in the local coordinate system of a joint j, which is defined using a multivariate Gaussian distribution:

$$\psi_{i,j}^{tran}(y_i, y_j) = \mathcal{N}(y_{ij}^T \mid \mu_{ij}^T, \Sigma_{ij}^T), \tag{4}$$

where μ_{ij}^T is the mean and Σ_{ij}^T is the covariance. We keep the covariance as a diagonal matrix for relaxing the computations.

The rotation transformation ternary potential denotes the rotation between two body parts. To model the transformation, a triad of joints is initially chosen. Then two

vectors with the same origin are created. The origin corresponds to a hinge joint (1 DoF) and the vectors to body parts. Since a hinge joint allows the rotation only along one axis, we model the rotation potential using a univariate Gaussian distribution:

$$\psi^{rot}_{i,j,k}(y_i, y_j, y_k) = \mathcal{N}(y^R_{ijk} \mid \mu^R_{ijk}, \sigma^R_{ijk}), \tag{5}$$

where μ^R_{ijk} is the mean and σ^R_{ijk} the variance. The von Mises distribution is the actual distribution for the rotation but using a Gaussian has been proven to generalise well in [6]. For that reason, we follow the same formulation.

The collision constraints are introduced to ensure the spatial exclusion between joints. To force body joints not to collide and respect a minimum distance, we introduce the collision function $\psi^{col}_{i,j}(y_i, y_j)$. The joints are modelled as spheres and the collision function estimates if there is intersection between the spheres:

$$\psi^{col}_{i,j}(y_i, y_j) = \begin{cases} 1 \text{ if } inter(y_i, y_j) = 0 \\ \epsilon \text{ otherwise} \end{cases} \tag{6}$$

where ϵ is again a small constant for numerical stability and $inter(y_i, y_j) \in \{0, 1\}$ is the sphere-sphere intersection function [19]. This is a hard constraint, but in our experiments it functioned well.

In our model, we have cancelled the notion of the global coordinate system and express every variable relation in a local coordinate system. This is the principle of the original work on the pictorial structures [11, 13]. For this reason, our prior model is applicable to any multi-view setup independently of the system calibration. To learn the prior model, we have used ground-truth data from a single training dataset. Then in the experimental phase, we apply this prior model on different setups.

3.2 3D Pose Inference and Tracking

The last step for obtaining the pose of different individuals is the inference. We seek to maximize the posterior probability in (1) for each individual h, given as follows:

$$\mathbf{y_h}^* = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x_h}, \mathbf{p_h}) \tag{7}$$

where $\mathbf{y_h}^*$ are the configurations that maximize the posterior and h is the identity of each individual. To recover the identity, we introduce a human tracker in our framework. We rely on the work of [7] to derive the tracks of each individual. As as result, each individual has its own observation $\mathbf{x_h}$ and lies in its own state space using the tracking information. In addition, the identity of the inferred body poses $\mathbf{p_h}$ is propagated over time using the tracking output. Knowing the identity of each individual reduces the computations in the temporal consistence function. Theoretically, relying exclusively on the tracking information can result in drifts. Nevertheless, in practice the tracker performed very reliably in our experiments.

Our graphical model includes loops and thus the inference is performed with loopy belief propagation using the sum-product algorithm [8]. Finally, the MAP estimate for each individual gives the body pose. In addition, we profit from tracking and obtain the trajectory of each individual.

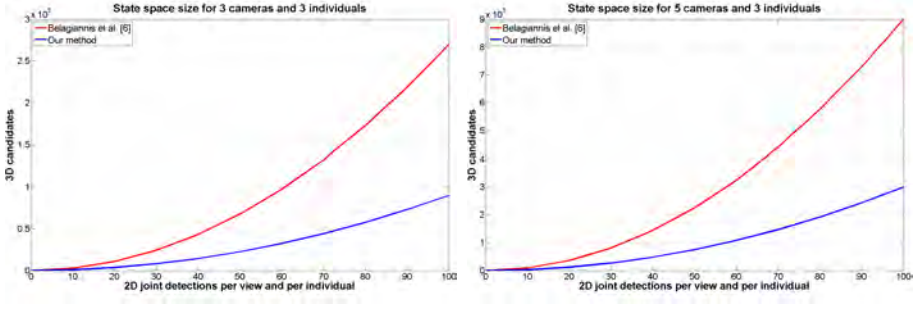
4 Experiments

Multiple human 3D pose estimation from multiple camera views in unconstrained environments is an active research topic, which is still growing. For that reason, the number of evaluation datasets is limited. To evaluate our method, we have used two challenging datasets [6], Campus and Shelf, which are the only available datasets for this problem to the best of our knowledge. The Campus dataset has 3 cameras and involves 3 individuals, while Shelf has 5 cameras and includes 4 individuals. Since one individual is mostly occluded and invisible from most of the views, we evaluate on the other three, as in [6].

Our model is composed of 14 body joints (Figure 2). We use the joint detector from [2] and learn the body prior using the ground-truth data from the KTH Multiview Football II dataset [9]. The evaluation is divided into three tasks: analysis of the state space, evaluation on the human detection and the body pose estimation. Since the method of [6] is closely related to our work, we compare our results with it for each experiment.

4.1 State Space Analysis

One of our main contributions is to impose the temporal consistence to the model. This is achieved by taking people tracking results as input. Given the identity of each individual, we can significantly reduce the number of candidates for each joint state space. This reduction results in faster inference. Our algorithms runs at 1 fps, given the tracks, for inferring up to 3 individuals in comparison to [6], which runs at 1 fps for single human pose estimation, given the detections. In Figure 3, we show the number of 3D candidates of a body joint versus the number of 2D joint detection samples.



Campus dataset Shelf dataset

Fig. 3. State space size: We show the number of recovered 3D candidates versus the number of 2D detection samples on both datasets. The number of 3D candidates is computed by summing up the triangulation instances of all combinations of all view pairs. The number of candidates is significantly lower than [6], because of the unknown identity of each individual.

The number of recovered 3D candidates is the aggregation of the triangulation instances of all combinations of view pairs, given different number of 2D body joint

detection. In the case of [6], the triangulation is performed between all the individuals due to the unknown identity. This results in a much larger state space where inference is computationally more expensive.

4.2 People Detection

In our framework, we consider the task of people detection separately from pose estimation. Therefore, our model is dependent on the detection result. By taking the advantage of tracking, we not only improve the people detection results, but also obtain the identity of each individual in comparison to frame-by-frame detection [6]. To evaluate the people detection results, we employ the PCP evaluation score [12]. For each individual, we define a line in the centre of the ground-truth cube which is perpendicular to the ground. It corresponds to the height of each individual in the 3D space. In the evaluation stage, we derive the same line from the inferred pose by fitting a cube and estimate the PCP score. From our experiments, we have found that the results of this mean of evaluation is equivalent to cube intersection but the computation of the PCP score is way faster. The results are summarised in Table 1.

	Campus		**Shelf**	
	Belagiannis et al. [6]	Our method	Belagiannis et al. [6]	Our method
Recall	98.05	99.30	90.50	97.82

Table 1. Detection results:The detection recall is estimated using the PCP score. The threshold α is set to 0.5 for both methods.

In both datasets, the recall is quite high for both methods. However, the temporal consistence of our method gives an improvement of around 10% in the Shelf dataset over the frame-by-frame detection of [6]. The only failures that we have observed were small drifts over time. The precision of our method is totally corrected using the tracker information and thus we do not include it in the comparison.

Camera 1 Camera 2 Camera 3

Fig. 4. Campus dataset result: The 3D estimated body poses are projected across the camera views. The identity of each individual is derived by the tracker.

4.3 Body Pose Estimation

In our last task, we evaluate the final accuracy of pose estimation on the two datasets, both of which include several individuals and have been captured in unconstrained environments. We compare our results with the ones of a related work [6], which is a state-of-the-art approach. Unlike our model, the human model in [6] does not consider the temporal consistence over time. Furthermore, the unknown identity of each individual results in more false positive candidates. These differences between the two methods are reflected in their performance, presented in Tables 2. Below, we discuss the results for each dataset separately.

	Campus		**Shelf**	
	Belagiannis et al. [6]	Our method	Belagiannis et al. [6]	Our method
Actor 1	82	83	66	75
Actor 2	72	73	65	67
Actor 3	73	78	83	86
Average	75.6	78	71.3	76

Table 2. Pose estimation results: The PCP score are presented for both datasets. The threshold α is set to 0.5 for both methods.

Campus dataset: In this dataset, we achieve substantially better results for the Actor 3 and slightly better for the other two individuals. The results are demonstrated in Figure 4. The support of the tracker facilitates the pose recovery of the Actor 3 who often undergoes occlusions. For the same reason, there is a small predominance for the other two individuals. The reduced state space and the temporal consistence of our model improve the precision of the inferred poses, as it is depicted in Figure 5.

Shelf dataset: In this dataset, our method again achieves better results on all individuals. In particular, there is a big difference between our method and [6] in the result of Actor 1, thanks to the temporal consistence. False positive candidates are penalised by the temporal potential function, and therefore the performance is improved. Qualitative results are presented in Figure 1 and Figure 5.

5 Conclusion

We have presented a temporally consistent 3D pictorial structures model. Our model applies to unconstrained environments for multiple human pose estimation from multiple views. We have introduced tracking and temporal consistency to our formulation for recovering the 3D human pose. Knowing the identity of each individual results in a small state space which allows efficient inference. Moreover, the temporal consistency helps to penalise false positive candidates of the state space. To demonstrate the advances of our model, we have evaluated on two challenging datasets and achieved state-of-the-art results.

Fig. 5. Qualitative comparison: We show the results of our method on the top row, and those from [6] on the bottom row. We have chosen different frames and camera views from the Campus and Shelf datasets to illustrate the advances of our model. In all cases, the inferred poses of our model are more precise due to the regularisation of the temporal potential function and the reduced state space.

6 Acknowledgements

This work was supported in part by the Swiss National Science Foundation and by DFG - Deutsche Forschungsgemeinschaft under the project "Advanced Learning for Tracking and Detection in Medical Workflow Analysis". The authors would like to thank Iro Laina for helping with the data preparation.

References

1. Alahari, K., Seguin, G., Sivic, J., Laptev, I.: Pose estimation and segmentation of people in 3d movies. In: ICCV (2013)
2. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3d human pose estimation. In: BMVC (2013)
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR (2009)
4. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR. pp. 1–8. IEEE (2008)
5. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR (2010)
6. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3D pictorial structures for multiple human pose estimation. In: CVPR. IEEE (2014)
7. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. TPAMI (2011)
8. Bishop, C.M., et al.: Pattern recognition and machine learning. springer New York (2006)
9. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: CVPR (2013)
10. Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: ECCV. Springer (2010)

11. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV (2005)
12. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)
13. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. Computers, IEEE Transactions on (1973)
14. Gammeter, S., Ess, A., Jäggli, T., Schindler, K., Leibe, B., Van Gool, L.: Articulated multi-body tracking under egomotion. In: ECCV. Springer (2008)
15. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision, vol. 2. Cambridge Univ Press (2000)
16. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: BMVC (2013)
17. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. Information Theory, IEEE Transactions on 47(2), 498–519 (2001)
18. Lee, M.W., Nevatia, R.: Human pose tracking using multi-level structured models. In: ECCV. Springer (2006)
19. Lin, M., Gottschalk, S.: Collision detection between geometric models: A survey. In: Proc. of IMA Conference on Mathematics of Surfaces (1998)
20. Luo, X., Berendsen, B., Tan, R.T., Veltkamp, R.C.: Human pose estimation for multiple persons based on volume reconstruction. In: ICPR. pp. 3591–3594. IEEE (2010)
21. Mitchelson, J.R., Hilton, A.: Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling. In: BMVC. pp. 1–10 (2003)
22. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding (2006)
23. Plankers, R., Fua, P.: Articulated soft objects for multi-view shape and motion capture. IEEE PAMI 25(10) (2003)
24. Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: CVPR. IEEE (2003)
25. Sigal, L., Isard, M., Haussecker, H., Black, M.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. IJCV (2011)
26. Sigal, L., Black, M.J.: Guest editorial: state of the art in image-and video-based human pose and motion estimation. IJCV (2010)
27. Sutton, C., McCallum, A., Rohanimanesh, K.: Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. The Journal of Machine Learning Research 8, 693–723 (2007)
28. Wang, X., Türetken, E., Fleuret, F., Fua, P.: Tracking interacting objects optimally using integer programming. ECCV (2014)
29. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. TPAMI (2004)