3D hypothesis clustering for cross-view matching in multi-person motion capture

Miaopeng Li  
*State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China.*

Zimeng Zhou  
*State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China.*

Xinguo Liu  
*State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China.*

Follow this and additional works at: https://tsinghuauniversitypress.researchcommons.org/computational-visual-media

Part of the *Computational Engineering Commons, Computer-Aided Engineering and Design Commons, Graphics and Human Computer Interfaces Commons, and the Software Engineering Commons*

**Recommended Citation**  
Miaopeng Li, Zimeng Zhou, Xinguo Liu. 3D hypothesis clustering for cross-view matching in multi-person motion capture. *Computational Visual Media* 2020, 6(2): 147-156.
3D hypothesis clustering for cross-view matching in multi-person motion capture

Miaopeng Li¹, Zimeng Zhou¹, and Xinguo Liu¹

© The Author(s) 2020.

Abstract We present a multiview method for markerless motion capture of multiple people. The main challenge in this problem is to determine cross-view correspondences for the 2D joints in the presence of noise. We propose a 3D hypothesis clustering technique to solve this problem. The core idea is to transform joint matching in 2D space into a clustering problem in a 3D hypothesis space. In this way, evidence from photometric appearance, multiview geometry, and bone length can be integrated to solve the clustering problem efficiently and robustly. Each cluster encodes a set of matched 2D joints for the same person across different views, from which the 3D joints can be effectively inferred. We then assemble the inferred 3D joints to form full-body skeletons for all persons in a bottom-up way. Our experiments demonstrate the robustness of our approach even in challenging cases with heavy occlusion, closely interacting people, and few cameras. We have evaluated our method on many datasets, and our results show that it has significantly lower estimation errors than many state-of-the-art methods.

Keywords multi-person motion capture; cross-view matching; clustering; human pose estimation

1 Introduction

Multi-person motion capture estimates the articulated joint positions and/or angles for a group of people from video. It is an important yet challenging task with many applications, such as human–computer interaction, action recognition, emotion analysis, human performance analysis, and so on. The latest work shows that markerless motion capture is feasible for a single person in weakly controlled environments, but is very difficult for a group of people in uncontrolled environments, due to the increased complexity in occlusion, appearance, motion, shape, and scale.

In this paper, we focus on markerless motion capture for multiple people with a multiview setup. Past approaches typically solve this problem in two stages. The first stage detects 2D body keypoints or pose in each view for all persons, and the second stage matches them across views to reconstruct 3D poses. As deep-learning based keypoint and pose detection techniques have greatly advanced [1–6], the remaining challenge is to resolve the correspondences between detected keypoints or poses across different views and different persons. Most previous methods employ a 3D pictorial structure (3DPS) model to implicitly solve the correspondence problem by reasoning about all hypotheses in 3D that are geometrically compatible with the detected 2D information [7–11]. However, 3DPS-based approaches are computationally expensive due to the huge state space. In addition, they are not robust, especially when there are few cameras, as they link the 2D detected joints only based on multiview geometry, and appearance cues are ignored.

This paper presents a 3D hypothesis clustering technique to efficiently and robustly determine the cross-view correspondences between the detected joints. The proposed technique transforms the correspondence problem from 2D space to 3D, and solves it by a 3D hypothesis clustering algorithm incorporating appearance evidence, multiview geometry, and bone length information. Each resulting cluster is a set of 3D points, which
corresponds to a set of matched 2D detections of the same person in different views. The 3D joints can then be inferred from the matched 2D joints; full skeletons for all persons are formed by assembling the inferred 3D joints.

We have tested our algorithm on a number of benchmark datasets, including Panoptic [12], Shelf, and Campus [8]. The high accuracy of our method is shown by a quantitative comparison to many state-of-the-art methods. Some results of our method are shown in Fig. 1, while the Electronic Supplementary Material (ESM) provides more vivid results.

The main contributions of this paper are:

- A fully-automatic motion capture approach for multiple interacting people. No manual intervention or subject-specific priors are needed, enhancing flexibility. It determines the number of persons and estimates the bone length for each person automatically.
- A simple yet effective technique for cross-view 2D joint matching using 3D hypothesis clustering to match noisy 2D joint detections across multiple views. It is robust to cases with heavy occlusion, few cameras, and closely interacting people.
- Reliable multi-person pose reconstruction. Our method produces reliable 3D skeletons in a bottom–up way by combining multiview appearance evidence, geometric constraints, and bone length constraints.

2 Related work

There have been many approaches to 3D pose estimation for multiple people with a multiview setup. Most previous works are based on 3DPS models in which nodes represent 3D locations of body joints and edges encode pairwise relations between them [7–11]. The state space for each joint is often a 3D voxel grid obtained by discretizing a 3D space. The node-likelihood score of each voxel is computed by projecting the center of the voxel to all views and taking the average of the scores at the projected positions. Pairwise potentials between joints are given by skeletal constraints [8, 9] or body parts detected in 2D views [10, 11]. 3D poses of multiple people are jointly inferred by a posteriori estimation. In contrast to our approach, most of these approaches require the number of people to be known. Additionally, all these approaches consider all body joints for all people simultaneously, resulting in a huge state space and a large amount of computation needed for inferencing. Moreover, these approaches are usually sensitive to the number of cameras and noise in the 2D detections, as they implicitly match the 2D detections across views using only multiview geometry.

A recent approach first explicitly matches the detected 2D poses and then reconstructs 3D poses from the 2D poses belonging to each person [13]. It uses OpenPose of Cao et al. [4] as the single-view pose detector, and solves the cross-view 2D pose matching problem based on epipolar geometry. A limitation of this method is that simple epipolar geometry verification can be error-prone as the detected 2D poses are often incomplete and inaccurate due to occlusion and truncation. Furthermore, multi-person 2D pose parsing is performed in each view independently, which is time-consuming and ignores complementary information from other views. Also, as each possible pair of poses from two views are matched separately, inconsistent correspondences may result, i.e., two matched poses from two views may be associated with different people in another view. Such inconsistencies will lead to serious errors in 3D pose reconstruction. Dong et al. [14] address these challenges with a convex optimization based multi-way matching algorithm, which determines correspondences for all views at once. In their work, a human detector [15] and a state-of-the-art
single-person pose estimator [16] are applied to each view to estimate multi-person 2D poses. In the matching stage, a person re-identification network [17] is integrated to exploit the appearance information. This method achieves excellent performance, but is very time-consuming as it integrates so many network modules. Furthermore, pose-level 2D detection usually cannot provide reliable estimation in scenes with many closely interacting people [18], making it difficult to find correct correspondences between 2D poses.

In contrast, our approach solves the correspondence problem at joint level for all views at once, by combining multiview appearance cues, geometric constraints, and bone length constraints. The advantages are: firstly, only joint-level 2D detection is needed, which is robust to occlusion and truncation. Secondly, it performs pose parsing in 3D space only once rather than many times in 2D, which is more efficient and more reliable as complementary information from all views are exploited. Thirdly, it naturally avoids inconsistent correspondences as all views are considered simultaneously.

3 Approach

Our approach to multi-person 3D skeleton reconstruction is overviewed in Fig. 2. It takes as input a set of images from multiple calibrated and synchronized video cameras (Fig. 2(a)), and outputs 3D skeletons for all people in the scene (Fig. 2(h)). The skeleton for each person comprises 14 joints, and 13 limbs as shown in Fig. 3 (the dashed limb connections do not have constant bone length).

In the following sections, we assume a setup with $C$ calibrated cameras, providing input images $\{I_1, \ldots, I_C\}$ with corresponding projection matrices $\{P_1, \ldots, P_C\}$.

3.1 2D joint candidate detection

We detect 2D joint candidates in each image using the network proposed by Cao et al. [4]. It takes an RGB image as input, and outputs a set of confidence maps for each joint (Fig. 2(b)) and a set of part affinity fields (PAFs) for each limb (Fig. 2(c)).

Let $H_{i,j}$ denote the confidence map for joint $j$ in image $I_i$. Given $H_{i,j}$, the candidates of joint $j$ for all people can be obtained by performing non-maximum suppression (NMS), yielding a set of 2D points as shown in Fig. 2(d), denoted $\phi_{i,j} = \{x_{k,i,j}, k \in \{1, \ldots, n_{i,j}\}\}$. These detections can be noisy, due to occlusion, motion blur, ambiguity due to symmetry, distorted poses, etc. Extracting 3D joint positions directly from these detections is unreliable.

The PAFs are used to measure the confidence that two joint candidates form a limb of one person. Consider two 2D joint candidates in image $I_i$: $x_{i,a}$ and $x_{i,b}$, where joint indices $a$ and $b$ correspond to a limb in Fig. 3. The connectivity score that they come from the same person, $s(x_{i,a}, x_{i,b})$, is computed by integrating the dot product of the corresponding PAF and the unit vector from $x_{i,a}$ to $x_{i,b}$ over the line segment between them (see Eq. (10) in Ref. [4] for details).

The next two steps, 2D joint candidate matching across views (Section 3.2) and 3D joint candidate

![Fig. 2 Overview. The input is a set of synchronized images from multiple views (a); the final output comprises 3D poses in global world space (h). Each individual is followed over time marked with a fixed color as in (g) and (h).](image-url)
reconstruction (Section 3.3) are sequentially performed on the joints in the order shown on the skeleton model in Fig. 3. Note that the order used may affect the cross-view joint matching procedure. It is possible that optimizing the order might improve the result, but the fixed order used in our current implementation works well.

For joint \( j \), the reconstructed 3D candidates are denoted \( \Phi_j = \{X^k_j, k \in \{1, \ldots, N_j\}\} \).

### 3.2 2D joint candidate matching between views

2D joint candidate matching aims to find joint correspondences for each person across all views, which can be formulated as a multi-way matching problem. Classical approaches to this problem include spanning tree optimization [19], sampling inconsistent cycles [20], low-rank matrix recovery [21], etc. Most of them define feature similarity from 2D images and solve the problem with a complex optimization algorithm to guarantee cycle consistency. In our setup, all views are calibrated and synchronized, so it is possible and valuable to explore 3D cues for reliability and efficiency. Thus, we propose a 3D hypothesis clustering technique to match noisy 2D candidates across views, combining multiview geometric constraints, appearance cues, and bone length constraints. In this way, cycle consistency is naturally guaranteed.

#### 3.2.1 3D hypothesis space construction

Consider joint \( j \). Given the 2D candidates detected in each view, we triangulate each possible candidate pair from every pair of views, yielding a set of 3D points, forming a 3D hypothesis space. Let \( \Lambda_j \) be the hypothesis space for joint \( j \), consisting of a number of 3D points; see Fig. 4(a). Many wrong hypotheses exist stemming from the triangulation of 2D candidates which are wrongly detected or from different individuals, as their identities are not known. A basic way to detect outliers is to check if the reprojection error exceeds a fixed threshold \( \tau_1 \) (\( \tau_1 = 8 \) pixels in our experiments). However, some wrong hypotheses may pass this test provided that the two candidate locations satisfy the epipolar geometry constraint, as shown in Fig. 4(b). To detect such wrong hypotheses, we impose two further constraints on them.

The first constraint is PAF based. If there exists a limb: \( p \rightarrow j \) between joint \( j \) and its parent joint \( p \), then the 3D point \( \tilde{X}_j \in \Lambda_j \) must have a high connectivity score with one of the 3D parent candidates, denoted \( X_p, X_p \in \Phi_p \). Specifically, we define the connectivity score in 3D space by combining multiple views. Consider two 3D candidates: \( X_a \) and \( X_b \), where joint indices \( a \) and \( b \) correspond to a limb in Fig. 3. The connectivity score between \( X_a \) and \( X_b \), \( S(X_a, X_b) \), is computed by averaging the 2D connectivity score of the projected points in all views:

\[
S(X_a, X_b) = \frac{1}{C} \sum_{i=1}^{C} s(P_i X_a, P_i X_b) \quad (1)
\]

Then the constraint is defined as

\[
\exists X_p \in \Phi_p, \quad S(X_p, \tilde{X}_j) \geq \tau_2 \quad (2)
\]

where \( \tau_2 = 0.3 \) in our experiments.

The second constraint is bone length based. If \( p \rightarrow j \) is a length-constant limb, then the distance between \( X_p \) and \( \tilde{X}_j \) must approximately equal the corresponding bone length of one person. It is defined as

\[
\exists \ell_{p,j} \in L_{p,j}, \quad \|X_p - \tilde{X}_j\| - \ell_{p,j} \leq \tau_3 \ell_{p,j} \quad (3)
\]

where \( L_{p,j} \) is the length set of the limb \( p \rightarrow j \) from all persons, which are assumed to be known as priors for now; Section 3.6 will present an automatic method to quickly compute them. \( \tau_3 = 0.1 \) in our experiments.

Combining the two constraints along with the reprojection error constraint, the 3D hypothesis space after removing outliers is shown in Fig. 4(c).

#### 3.2.2 3D hypothesis space clustering

The hypothesis space for each joint is shared between all persons (Fig. 4(c)).
clustering aims to separate the hypothesis space into a component for each person. We adopt DBSCAN (density-based spatial clustering for applications with noise) [22]; other clustering algorithms which estimate the number of clusters could also be used. Figure 4(d) shows the clustered 3D hypothesis space. The 3D points in a given cluster correspond to a single person. The corresponding 2D joint candidates contributing the 3D points in a given cluster are naturally matched across views, yielding a single 3D joint candidate.

One challenge in cross-view matching is that the positions of some joints from different people can be very close (e.g., the right wrist when two people clap right hands with each other, as in the first example in Fig. 1). In such cases the clustering algorithm may be unable to distinguish them and so groups them into a single cluster, so one 3D candidate is missing. To solve this problem, if the center of a cluster has a plausible distance from more than one parent 3D candidate according to the corresponding bone length set, we split the cluster by re-grouping its 3D points according to their different optimum parent joints with reasonable bone length.

### 3.3 3D joint candidate reconstruction
Consider joint $j$ again. Once a set of 2D joint candidates $\{x_{i_1,j}, \ldots, x_{i_n,j}\}$ has been matched across views $\{i_1, \ldots, i_n\}$, we can determine a 3D point $X_j$ with a least-square optimization procedure, which minimizes the sum of reprojection error in the views.

It is known that least-squares are very sensitive to outliers, especially if there are very few input values, like in our case with few cameras. Thus, we use a weighted cost function based on the confidence in the 2D candidates:

$$X_j \leftarrow \arg \min_{i \in \{i_1, \ldots, i_n\}} \sum_{i \in \{i_1, \ldots, i_n\}} \left[ H_{i,j}(x_{i,j}) \| P_i X_j - x_{i,j} \| \right]$$

In this way, we give more importance to reprojection errors for good 2D candidates. This cost function is optimized using the Levenberg–Marquardt least-squares algorithm [23, 24]. The reconstructed 3D point $X_j$ is a candidate for the $j$th anatomical joint of one person, which we refer to as a 3D joint candidate for joint $j$. Figure 2(e) shows the reconstructed 3D joint candidates for all joints, for all persons.

### 3.4 3D pose parsing
Given the reconstructed 3D joint candidates (Fig. 2(e)), pose parsing aims to generate a full-body pose for each person by assembling the corresponding joint candidates. We extend the method proposed by Cao et al. [4] for 2D multi-person pose estimation: we perform pose parsing in 3D space instead of 2D. In turn, the connectivity score between possibly connected candidates is computed in 3D space as in Eq. (1). Moreover, each limb connection candidate must have the correct distance according to bone length. As more evidence is integrated from multiple views and bone lengths, our extended method is robust to occlusion with significantly overlapping people, etc. Figure 2(f) shows the parsed 3D poses for all persons.

### 3.5 Pose tracking
Since the above steps reconstruct multiple 3D poses for each time step independently, we use a pose tracking method to obtain the pose trajectories over time. In our setup, 3D pose for each person is reliably reconstructed, so is used for robust tracking.

Given the pose predictions in 3D space, we link them in time to obtain trajectories, which can be seen
as a data association problem over these predictions. We compute tracks by simplifying this problem to bipartite matching between each pair of adjacent frames (Figs. 2(f) and 2(g)). We initialize the tracks in the first frame and propagate labels forward using a greedy method, one frame at a time. The similarity weight for two poses in adjacent frames is measured by the average distance between torso joints (head, neck, left/right shoulder, left/right hip) in 3D space. Any pose that is not matched to an existing track instantiates a new track. The temporally consistent poses are shown in Figs. 2(g) and 2(h), where poses for the same person are marked in the same color.

3.6 Bone length estimation

To simplify system setup, we use an automatic technique to estimate bone lengths of all persons to be captured. As shown in Fig. 3, the skeleton has 11 length-constant bones. We estimate their lengths on-the-fly. First, we construct initial 3D poses for all persons in a similar way to that described above but without considering bone length information, and track them for a few seconds, producing hundreds of initial 3D poses for each subject. Then, for each subject, we compute bone lengths from each pose, yielding a number of length approximations for each bone. Though the distribution of the length approximations can be noisy, most approximations are near the ground truth. Then for each bone, we use the median length as the final estimate of length. This automatic method enables our motion capture system to conveniently adapt to different subjects without any subject-specific priors.

4 Experimental results

We have evaluated our proposed method on several benchmark datasets: Panoptic [12], Shelf, and Campus [8]. In the experiments, no ground truth data was used for training from the respective datasets. We rely solely on a generic 2D pose CNN (trained on the MPII dataset [25]) and knowledge of the geometry of the calibrated camera setup. Figure 5 shows some results for these datasets, demonstrating that our method can produce high-quality 3D skeletons in general scenarios with many interacting people (7 people in the 7th example), large variability in human physique (see the toddler in the 3rd example), severe occlusion (see the 4th example), and low-resolution images (360 × 288) using only 3 cameras (the 1st example).

4.1 Evaluation on benchmark datasets

Shelf and Campus datasets [8] contain 2–4 individuals captured by multiple cameras, indoors and outdoors. Our evaluation on the Shelf dataset
uses 300 frames, and on Campus, 220 frames, following previous works [9, 10, 14]. The evaluation metric is the percentage of correct parts (PCP) with \( \alpha = 0.5 \).

We compare our method to the state-of-the-art methods in Refs. [9, 10, 14]. Specifically, Belagiannis et al. [9] used a factor graph optimized with belief propagation to perform inference on 2D poses detected by DeepPose [1]. Ershadi-Nasab et al. [10] used Deepercut [3] as the 2D joint detector, defined a fully connected conditional random field, and then used loopy belief propagation for approximate inferring. Dong et al. [14] is a very recent method, which combines a human detector [15] and a single-person pose estimator [16] to obtain multi-person 2D poses in all views, and then uses a convex optimization based multi-way matching algorithm to match detected poses across views, from which the 3D pose of each person is inferred. Table 1 quantitatively reports the comparison results.

Both Refs. [9] and [10] are 3DPS model-based approaches. As Belagiannis et al. [9] did not use the most recent heatmap-based keypoint detectors, which perform better than DeepPose [1] (a direct coordinate regression-based detector). For a fair comparison, we also report the result in Ref. [14] which re-implements the approach of Belagiannis et al. [9] using the same state-of-the-art 2D keypoint detector [4] as our method, indicated as Belagiannis et al.*.

The results show that our method outperforms both 3DPS model-based methods [9, 10] by a large margin on the Campus dataset (3.61\%) and achieves a substantial improvement on the Shelf dataset (1.47\%). In particular, our approach significantly improves results for actor 3 in the Campus dataset, who suffers from heavy occlusion with only 3 cameras. The main reason is that the 3DPS model-based method implicitly matches 2D detections only using multiview geometry and ignores appearance cues. When the number of cameras is small or heavy occlusion exists in most views, multiview geometry consistency alone is sometimes insufficient to distinguish correct and incorrect correspondences, leading to incorrect 3D pose. Our approach explicitly leverages appearance cues (3D PAF score), multiview geometric constraints, and bone length constraints to find cross-view correspondences, yielding much more robust results. Moreover, our method avoids a complex inference procedure in a huge state space, which can be very time-consuming.

Our approach performs better or on par with the method of Ref. [14], in which, like in our method, 3D poses are reconstructed by direct triangulation. The advantages of our approach are as follows. Firstly, it is very compact and efficient as it avoids the many network modules in Ref. [14] (a human detector [15], a single-person pose detector [16], and a person re-identification CNN [17]): only a 2D pose CNN [4] is needed. Secondly, it is more robust to heavy occlusion and closely interacting people. In such cases, detection of 2D pose is usually unreliable, making it difficult to find correct correspondences between 2D poses. Our approach largely alleviates this problem as detection of 2D joints is usually feasible [18], and the fully skeletal poses are parsed in 3D space by integrating complementary information from all views. Thirdly, instead of requiring a complex convex optimization algorithm to guarantee cycle-consistency in cross-view matching, our approach uses a novel 3D hypothesis clustering algorithm, in which cycle-consistency is naturally guaranteed.

**Table 1** PCP score (in %) on the Shelf and Campus datasets [8]

| Shelf dataset | Actor1 | Actor2 | Actor3 | Average |
|---------------|--------|--------|--------|--------|
| Belagiannis et al. [9] | 75.26  | 69.68  | 87.59  | 77.51  |
| Belagiannis et al.* [9] | 98.1   | 91.1   | 92.8   | 94.0   |
| Ershadi et al. [10] | 93.29  | 75.85  | 94.83  | 87.99  |
| Dong et al. [14] | 97.9   | 89.5   | 97.8   | 95.1   |
| Ours          | 99.41  | 91.89  | 95.12  | 95.47  |

| Campus dataset | Actor1 | Actor2 | Actor3 | Average |
|----------------|--------|--------|--------|--------|
| Belagiannis et al. [9] | 93.45  | 75.65  | 84.37  | 84.49  |
| Belagiannis et al.* [9] | 84.8   | 89.0   | 71.5   | 81.8   |
| Ershadi et al. [10] | 94.18  | 92.89  | 84.62  | 90.56  |
| Dong et al. [14] | 90.6   | 89.2   | 97.7   | 92.5   |
| Ours          | 94.85  | 93.16  | 94.51  | 94.17  |

Panoptic is a large-scale indoor dataset [12], including 480 VGA video streams and 31 HD video streams of multiple people (more than 7) in social activities. We selected data from 5 activities (Toddler, Haggling, Band, Pizza, and Bang) for evaluation. For each activity, we selected 1 or 2 sub-sequences of 40 seconds (1200 frames), from HD cameras. To assess errors on this dataset, we used the mean per joint position error (MPJPE) as the evaluation metric.

Table 2 reports the error for each body joint as well...
as the average. Result show that our method has low errors (2.21 cm on average). As a comparison, the method in Ref. [12] produces average errors for the Haggling sequence of 4.88 cm, while our average is 2.10 cm, with fewer cameras used. The ESM provides further results for this dataset.

4.2 Ablation study

This section analyzes the impact of (i) the three constraints used in outlier detection (Section 3.2.1) and (ii) the number of views used for 3D pose reconstruction, using the Panoptic dataset.

4.2.1 Outlier detection constraints

We use three constraints to detect 3D hypothesis outliers, based on reprojection error (RE), PAF (PAF), and bone length (BL). We now test the impact on performance of these constraints, using the challenging sequence Pizza from the Panoptic dataset. It contains 6 people with frequent occlusion. The baseline is to detect outliers solely based on reprojection error.

To show how well bone length is preserved, we introduce a new error metric, mean bone length standard deviation (BoneStd). It measures stability of bone length; smaller is better. We compute BoneStd as follows: for each bone of each tracked subject, we compute the standard deviation of its lengths over all tested frames, and then average this over all bones and all subjects.

Table 3 shows the evaluation results. Compared with results only based on reprojection error, PAF-based and BL-based constraints each by itself improve estimation quality significantly. The best results are achieved when all three constraints are combined, which reduce MPJPE by 42.8% and BoneStd by 74.8%.

One can see that the BL constraint is more effective than the PAF constraint by both quality metrics, particularly BoneStd. The reason is that the PAF score may be unreliable when one part is invisible in most views.

4.2.2 Analysis on varying number of views

To evaluate the impact of the number of views, we performed our method using a varying number of views (6, 8, 12, 16) for several sub-sequences from the Panoptic dataset, with differing numbers of people (2, 3, 5, 6, 7). The views were sampled in an approximately uniform way.

As Table 4 shows, if a small number of views is used, the results have higher errors. If the number of views is fixed, the error depends on the complexity of the scene: the number of people involved, frequency of occlusion, etc. If the scene is simpler (e.g., with 2 or 3 people), a smaller number of views (e.g., 6 cameras) suffices. Overall, error cease to improve with around 12 views. As expected, more views benefit accurate motion capture, and help more if the scene contains more people.

4.3 Multi-person 3D pose dataset creation

We created a small-scale dataset with naturally interacting people performing daily activities both indoors and outdoors. It contains more than 20,000 frames captured from 8 camera views, covering daily activities and sports, such as walking, a discussion, a party, playing musical instruments, etc. Several sub-sequences are included in the ESM.

5 Conclusions

This paper presents a markerless motion capture method for multiple people, which fuses cues from 2D joint detection, multiview geometry, and auto-estimated bone length. Our key difference from
previous methods lies in a novel 3D hypothesis clustering technique to match 2D joint detections across views, which is robust to noise and limited cameras. Multi-person pose parsing is performed in 3D space instead of 2D, so that complementary information from multiple views can be exploited, improving both efficiency and reliability.

A major advantage of our method is that it is fully automatic and does not need subject-specific priors. The number of people in the scenes is automatically determined and may vary during capture. Our experiments demonstrate the robustness of our method to challenging issues such as significant occlusion, largely overlapping people, and closely interacting joints.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This work was partially supported by National Natural Science Foundation of China (No. 61872317) and FaceUnity Technology.

Electronic Supplementary Material Electronic supplementary material is available in the online version of this paper at https://doi.org/10.1007/s41095-020-0171-y.

References

[1] Toshev, A.; Szegedy, C. DeepPose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1653–1660, 2014.

[2] Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4724–4732, 2016.

[3] Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In: Computer Vision–ECCV 2016. Lecture Notes in Computer Science, Vol. 9910. Leibe B.; Matas J.; Sebe N.; Welling M. Eds. Springer Cham, 34–50, 2016.

[4] Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1302–1310, 2017.

[5] Li, M.; Zhou, Z.; Li, J.; Liu, X. Bottom-up pose estimation of multiple person with bounding box constraint. In: Proceedings of the 24th International Conference on Pattern Recognition, 115–120, 2018.

[6] Wang, H.; An, W. P.; Wang, X.; Fang, L.; Yuan, J. Magnify-net for multi-person 2D pose estimation. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 1–6, 2018.

[7] Amin, S.; Andriluka, M.; Rohrbach, M.; Schiele, B. Multi-view pictorial structures for 3D human pose estimation. In: Proceedings of the British Machine Vision Conference, 2013.

[8] Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; Ilie, S. 3D pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1669–1676, 2014.

[9] Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; Ilie, S. 3D pictorial structures revisited: Multiple human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 38, No. 10, 1929–1942, 2016.

[10] Ershadi-Nasab, S.; Noury, E.; Kasaei, S.; Sanaei, E. Multiple human 3D pose estimation from multiview images. Multimedia Tools and Applications Vol. 77, No. 12, 15573–15601, 2018.

[11] Joo, H.; Simon, T.; Li, X.; Liu, H.; Tan, L.; Gui, L.; Banerjee, S.; Godisart, T. S.; Nabbe, B.; Matthews, I. et al. Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 41, No. 1, 190–204, 2019.

[12] Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; Sheikh, Y. Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision, 3334–3342, 2015.
[13] Kadkhodamohammadi, A.; Padoy, N. A generalizable approach for multi-view 3D human pose regression. arXiv preprint arXiv:1804.10462, 2018.

[14] Dong, J.; Jiang, W.; Huang, Q.; Bao, H.; Zhou X. Fast and robust multi-person 3D pose estimation from multiple views. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7792–7801, 2019.

[15] Ren, S. Q.; He, K. M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 39, No. 6, 1137–1149, 2017.

[16] Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7103–7112, 2018.

[17] Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5157–5166, 2018.

[18] Li, M. P.; Zhou, Z. M.; Liu, X. G. Multi-person pose estimation using bounding box constraint and LSTM. IEEE Transactions on Multimedia Vol. 21, No. 10, 2653–2663, 2019.

[19] Huang, Q. X.; Zhang, G. X.; Gao, L.; Hu, S. M.; Butscher, A.; Guibas, L. An optimization approach for extracting and encoding consistent maps in a shape collection. ACM Transactions on Graphics Vol. 31, No. 6, Article No. 167, 2012.

[20] Zhou, T.; Jae Lee, Y.; Yu, S. X.; Efros, A. A. FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1191–1200, 2015.

[21] Zhou, X.; Zhu, M.; Danilidis, K. Multi-image matching via fast alternating minimization. In: Proceedings of the IEEE International Conference on Computer Vision, 4032–4040, 2015.

[22] Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the KDD-96, 1996.

[23] Levenberg, K. A method for the solution of certain nonlinear problems in least squares. Quarterly of Applied Mathematics Vol. 2, No. 2, 164–168, 1944.

[24] Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics Vol. 11, No. 2, 431–441, 1963.

[25] Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3686–3693, 2014.

Miaopeng Li is a Ph.D. student in the State Key Lab of CAD&CG, Zhejiang University, China. She received her bachelor degree from Northwestern Polytechnic University in 2016. Her research interests include markerless human motion capture, human pose estimation, and 3D reconstruction, and their applications.

Zimeng Zhou is a master student in the State Key Lab of CAD&CG, Zhejiang University. His research interests are computer vision and computer graphics, with a particular focus on human pose estimation.

Xinguo Liu received his bachelor and Ph.D. degrees in applied mathematics from Zhejiang University, in 1995 and 2001, respectively. He is a professor of computer science in the State Key Lab of CAD&CG, Zhejiang University. His research interests include geometry processing, realistic and image based rendering, deformable objects, and 3D reconstruction.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.