3D Human Pose Estimation Based on Multi-feature Extraction

Senlin Ge[1], Huan Yu[2], Yuanming Zhang[2], Huitao Shi[2], and Hao Gao[1](✉)

[1] College Automation, College Artificial Intelligence,
Nanjing University of Posts and Telecommunications, Nanjing, China
tsgaohao@gmail.com

[2] Shanghai Institute of Aerospace Electronic Technology, Shanghai, China

Abstract. As an important computer vision task, 3D human pose estimation has received widespread attention and many applications have been derived from it. Most previous methods address this task by using a 3D pictorial structure model which is inefficient due to the huge state space. We propose a novel approach to solve this problem. Our key idea is to learn confidence weights of each joint from the input image through a simple neutral network. We also extract the confidence matrix of heatmaps which reflects its feature quality in order to enhance the feature quality in occluded views. Our approach is end-to-end differentiable which can improve the efficiency and robustness. We evaluate the approach on two public datasets including Human3.6M and Occlusion-Person which achieves significant performance gains compare with the state-of-the-art.

Keywords: 3D human pose estimation · Multi-view · End-to-end

1 Introduction

Recovering 3D human pose and motion from multiple views has attracted a great deal of attention over the last decades in the field of computer vision, which has a variety of applications such as activity recognition, sports broadcasting [1] and retail analysis [2]. The ultimate goal is to estimate 3D locations of the body joints in a world coordinate system from multiple cameras. While remarkable advances have been made in reconstruction of a human body, there are few works that address a more challenging setting where the joints are occluded in some views.

The methodology for multi-view 3D pose estimation in many existing studies includes two steps. In the first step, it tries to estimate the 2D poses in each camera view independently, for example, by Convolutional Neural Networks (CNN) [3,4]. Then in the second step, it recovers 3D pose by aggregating the 2D poses from all of the views. One typical method is to use the 3D Pictorial Structures model (3DPS), which directly estimate the 3D pose by exploring an ample state space of all possible human keypoints in the 3D space [5,6]. However, this method has large quantization errors because of the huge state space needed to explore.

L. Fang et al. (Eds.): CICAI 2022, LNAI 13606, pp. 570–581, 2022.
https://doi.org/10.1007/978-3-031-20503-3_51

On the other hand, the mainstream two-stage approach does not make full use of the information in multiple camera views and the relationship between views. They simply use the traditional triangulation or 3DPSM to restore the 3D human skeleton [7] which makes the pose estimation performance is not well in some challenging scenarios.

In this paper, we propose to solve the problem in a different way by making the best of information which extracted from the multi-view images and intermediate features. The method is orthogonal to the previous efforts. The motivation behind our method is that a joint occluded in some views may be visible in other views. So it is generally helpful to evaluate a weight matrix for each joint of each view. To that end, we present a novel approach for the 3D human pose estimation. Figure 1 shows the pipeline.

We first obtain more accurate 2D poses by jointly estimating them from multiple views using a CNN based approach. At the same time, we calculate the confidence of each joint under each view through a simple neural network. If a joint is occluded in one view, its feature is also likely corrupted. In this case, we hope to give a small confidence to the joint so that the high-quality joint in the visible views is dominant which will improve the performance in the subsequent triangulation process.

Second, when we get the initial heatmap from the 2D pose estimation, using a learning weight network to extract the confidence matrix of heatmaps under each view. Then the heatmap from each view is weighted and fused by use of the confidence matrix to obtain the final heatmaps. We apply the SoftMax operator to get the 2D positions of keypoints. Finally, the 2D positions of keypoints with the joints' confidence calculated in the first step are passed to the triangulation module that outputs the 3D pose.

We evaluate our approach on two public datasets including Human3.6M [8] and Occlusion-Person Dataset [9]. It has achieved excellent results on the two datasets. Furthermore, we compare our method to a number of standard multiview 3D pose estimation methods to give more detailed insights. Our method is end-to-end differentiable which improves the efficiency and robustness on 3D pose estimation.

2 Related Work

In this section, we briefly review the related works that utilize the techniques of this paper.

2.1 Single-view 2D Human Pose Estimation

The goal of 2D human pose estimation is to localize human anatomical keypoints or parts in one RGB image. With the introduction of "DeepPose" by Toshev et al. [10], many existing deep learning-based methods have achieved amazing results [11–13].

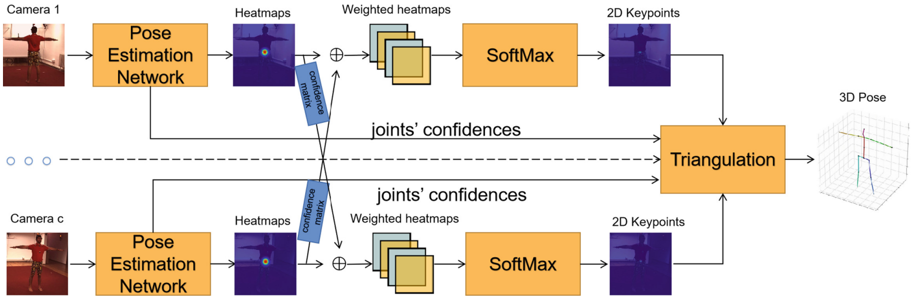

Fig. 1. The framework of our approach. It takes multi-view images as input and outputs the heatmaps and joints' confidences through the Pose Estimation Network. Then the heatmaps are weighted and fused by use of the confidence matrix to obtain the weighted heatmaps. The 2D positions of keypoints are inferred from the weighted heatmaps. Finally the 2D positions together with the joints' confidence are fed into the triangulation module to get the 3D pose position.

For 2D human pose estimation, current state-of-the-art methods can be typically categorized into two classes: top-down method and bottom-up method. In general, top-down methods [14–18] first detects people and then have estimated the pose of each person independently on each detected region. The bottom-up [19,20] methods jointly label part detection candidates and associated them to individual people by a matching algorithm. In our work, we choose the top-down method because of their higher accuracy. We adopt the SimpleBaseline [17] as the 2D human pose estimation backbone network.

2.2 Multi-view 3D Human Pose Estimation

Different from estimating from a single image, the goal of multi-view 3D human pose estimation is to get the ground-truth annotations for the monocular 3D human pose estimation. Most previous efforts can be divided into two categories. The first class is analytical methods [17,21–23] which explicitly model the relationship between a 2D and 3D pose according to the camera geometry. They first model human body by simple primitives and then optimize the model parameters through the use of multi-view images until the body model can be explained by the image features. The advantage of the analytical methods is that it can deal with the occlusion problem well because of the inherent structure prior embedded in human body model. However, due to the need to optimize all model parameters at the same time, the entire state space is huge, resulting in heavy computation in inference.

The second class is predictive methods which often flow a two-step framework by use of the powerful neural networks. They first detect the 2D human pose from all the camera views and then recover the 3D human pose by the use of triangulation or 3D Pictorial Structures model(3DPS). In [7], a recursive pictorial structure model was proposed to speed up the inference process. Recent work [25] has proposed to use 1 D convolution to jointly address the cross-view fusion and 3D pose reconstruction based on plane sweep stereo. [24] proposed a volumetric triangulation method to project the feature maps produced by the 2D human pose backbone into 3D volumes, which were then used to predict 3D poses. The shortcoming of this method is that not making full use of the information of images and feature maps which lead to the poor effect in the face of occlusion. On the contrary, our approach is efficient on multi-view 3D human pose estimation which benefits from the weight extraction network.

3 Method

In this section, we present our proposed approach for multi-view 3D human pose estimation. We assume that we have synchronized video streams from multiple cameras with known parameters which capture performance of a single person in the scene. The goal is to detect and predict human body poses in 3D given images captured from all views.

The overview of our approach is shown in Fig. 1. It first estimates 2D pose heatmaps and produces the joint confidence for each view. Then the heatmaps from all camera views are fused through the confidence matrix which extracted by a learning weight network. Finally, input the 2D positions of the keypoints and joints' confidence into the algebraic triangulation module to produce the 3D human pose.

3.1 2D Pose Detector and Joint Confidence

The 2D pose detector backbone h_p with learnable weight θ_p consists of a ResNet-152 network, followed by a series of transposed convolutions and a 1 * 1 kernel convolutional neutral network that outputs the heatmaps:

$$H_{c,j} = h_p(I_c;\theta_p), c = 1, 2, ..., C \tag{1}$$

where I_c denotes the image in the cth view, $H_{c,j}$ denotes the heatmap of the jth keypoint in the cth view.

In addition to output the heatmaps, we propose a simple network to extract the joint confidence for each view. The network structure as shown in Fig. 2. Starting from the images with known camera parameters, we apply some convolutional layers to extract features. Then the features are down-sampled by max pooling and feed into three fully connected layers that outputs the joint confidence:

$$\omega_{c,j} = h_\omega \left\{I_c;\theta_\omega\right\} \tag{2}$$

where h_ω denotes the joint confidence learnable module. θ_ω denotes the confidence learnable weight. $\omega_{c,j}$ denotes the confidence of the jth keypoint in the cth view.

It is worth noting that the joint confidence network is jointly trained end-to-end with the 2D pose detector backbone.

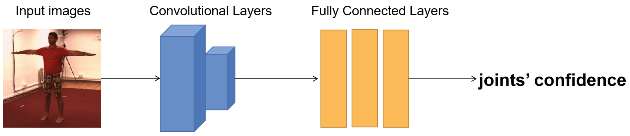

Fig. 2. The structure of the joints' confidence network.

3.2 Confidence Matrix for Heatmap Fusion

Our 2D pose detector takes multi-view images as input, generate initial pose heatmaps respectively for each, and then using a learning weight network to extract the confidence matrix of heatmaps which reflect the heatmap quality in each view. Finally the heatmaps are weighted and fused by use of the confidence matrix to obtain the final heatmaps. The core of this stage is to find the corresponding keypoints between all views.

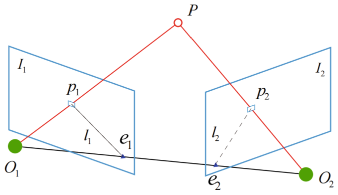

Fig. 3. Illustration of the epipolar geometry. For an image point p_1 in one view, we can find its corresponding point p_2 in another view and the corresponding point lies on the epipolar line.

We assume that there is a point P in 3D space as shown in Fig. 3. Note that we use homogeneous coordinate to represent a point. According to the equal up to a scale, the corresponding points p_1, p_2 in the two cameras are:

$$p_1 = KP \tag{3}$$

$$p_2 = K(RP + t) \tag{4}$$

where K is camera intrinsics, R and t are the rotation matrix and translation vector between two camera views.

From Fig. 3, we can find that the 3D point P, the image points p_1, p_2, and the camera centers O_1 and O_2 on the same plane which calls the epipolar plane. The l_1, l_2 which epipolar plane intersects with the two image planes calls epipolar line. In particular, we have

$$l_2 = Fp_1 \tag{5}$$

$$l_1 = F^T p_2 \tag{6}$$

where F is fundamental matrix which can be derived from K, R and t.

For each point p_1 in the first view, the epipolar geometry helps us to ensure the corresponding point p_2 in another view lie on the epipolar line l_2. However, since we do not know the depth of the point P, it is difficult to determine the exact location of point p_2. We decided to use the sparsity of heatmap to solve this problem. The heatmap has a small number of large responses near the joint location and a large number of zeros at other locations. So we select the largest response on the epipolar line as the corresponding point. Then we have:

$$\hat{H}_{c,j} = \lambda_c H_{c,j} + \sum_{u=1}^{N} \lambda_u max H_{u,j'} \tag{7}$$

where $\hat{H}$ denotes the weighted heatmap. λ is the weight of heatmap extracted from a learning weight network and N is the number of camera views.

To keep the gradients to flow back to heatmaps, we use soft-argmax to calculate the 2D positions of the joints:

$$x_{c,j} = e^{\hat{H}_{c,j}} / (\int_{q \in \Omega} e^{\hat{H}_{c,j}(q)}) \tag{8}$$

where $\hat{H}_{c,j}$ denotes the weighted heatmap of the jth keypoint in the cth view and Ω is the domain of the heatmap.

3.3 3D Pose Reconstruction

Given the estimated 2D poses and heatmaps from all views, we can reconstruct the 3D pose in several ways. The 3DPS model is one of them, but the large quantization errors result from exploring the huge state space may largely degrade the reconstruction. In order to fully integrate the information extracted from the multiview images and heatmaps, we make use of the point triangulation method with the joint confidence learned form neutral network for efficient inference.

The point triangulation is an efficient 3D pose estimation method with strong theoretical supports which reduces the finding of the 3D coordinates of a joint y_j to solving the overdetermined system of equations on homogeneous 3D coordinate vector of the joint $\widetilde{y}_j$:

$$A_j \widetilde{y}_j = 0 \tag{9}$$

where A_j is a matrix concatenating the homogeneous 3D vectors of all views for the jth keypoint.

However, traditional triangulation method can not solve the occlusion problem in views because it treats the joint in different views evenly without considering the joint may be occluded in some views, leading to unnecessary degradation of the final triangulation result. To deal with the problem, we add joint confidence which generated by a learnable module when triangulating and we have:

$$(\omega_j \circ A_j)\widetilde{y}_j = 0 \tag{10}$$

where ω_j is the joint confidence matrix which is in the same size of and $\circ$ denotes the Hadamard product.

We use Singular Value Decomposition of the matrix $E = U \sum V^T$ to solve the equation above. We set $\widetilde{y}$ as the last column of V. Then we can get the final 3D coordinates of a joint y_i by dividing the homogeneous 3D coordinate vector $\widetilde{y}$ by its fourth coordinate:

$$y = \frac{\widetilde{y}}{(\widetilde{y})_4} \tag{11}$$

3.4 Loss Function

For our method, the gradients pass from the output prediction to the input images which makes the method trainable end to end. The loss function for training the network contains two parts, the 3D mean square error (MSE) loss and the 2D joint smooth loss.

The 3D MSE loss between the estimated heatmaps and ground truth heatmaps is defined as:

$$L_{mse} = \sum_{c=1}^{C} \|H_k - H_{gt,k}\|_2 \tag{12}$$

The 2D joint smooth loss between the estimated 2D keypoints coordinates and the ground truth 2D keypoints coordinates:

$$L_{2d} = \sum_{c=1}^{C} \|y_c - y_{gt,c}\|_1 \tag{13}$$

The total loss of our method is defined as:

$$L = L_{mse} + L_{2d} \tag{14}$$

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on two standard datasets for multi-view 3D human pose estimation.

Human3.6M: Human3.6M is currently the largest multi-view 3D human pose estimation dataset. It provides synchronized images captured by four cameras which includes around 3.6 million images. Following the standard evaluation protocol used in the literature, subjects S1, S5, S6, S7, S8 are used for training and S9, S11 for testing [26–28]. To avoid over-fitting to the background, We also use the MPII dataset [29] to augment the training data.

Occlusion-Person: The Occlusion-Person dataset adopt UnrealCV to render multi-view images from 3D models. In particular, thirteen human models of different clothes are put into nine different scenes such as bedrooms, offices and living room. The scenes is captured by eight cameras. The dataset consists of thirty-six thousand frames and propose use objects such as sofas and desks to occlude some human body joints.

Metrics: The 2D pose estimation is measured by Percentage of Correct Keypoints (PCK) which measures the percentage of the estimated joints whose distance from the ground-truth joints is smaller than t times of the head length. Following the previous work, we set t to be 0.5 and head length to be 2.5% of the human bounding box width for all benchmarks.

The 3D pose estimation accuracy is measured by Mean Per Joint Position Error (MPJPE) between the estimated 3D pose and the ground-truth:

$$MPJPE = \frac{1}{M}\sum_{i=1}^{M}\left\|p_i^3 - \bar{p}_i^3\right\|_2 \tag{15}$$

where $y = [p_1^3, ..., p_M^3]$ denotes the ground-truth 3D pose, $\bar{y} = [\bar{p}_1^3, ..., \bar{p}_M^3]$ denotes the estimated 3D pose and M is the number of joints in a pose.

4.2 Results on Human3.6M

2D Pose Estimation Results. The 2D pose estimation results are shown in Table 1. We compare our approach to two baselines. The first is NoFuse which estimates 2D pose independently for each view without joints' confidence and weighted heatmaps. The second is HeuristicFuse which uses a fixed confidence for each heatmap according to Eq.(7). The patameter λ is set to be 0.5 by cross-validation. From the table, we can see that the performance of our approach is better than the two baselines. The average improvement is 10.6%. This demonstrates that our approach can effectively refine the 2D pose detection.

Table 1. The 2D pose estimation accuracy (PCK) of the baseline methods and our approach on the Human3.6M dataset

Methods	Root	Belly	Nose	Head	Hip	Knee	Wrist	Mean
NoFuse	95.8	77.1	86.4	86.2	79.3	81.5	70.1	82.3
HeuristicFuse	96.0	79.3	88.4	86.8	83.1	84.5	75.2	84.7
Ours	**96.5**	**94.9**	**96.3**	**96.4**	**96.0**	**92.5**	**85.9**	**94.1**

3D Pose Estimation Results. Table 2 shows the 3D pose estimation errors of the baselines and our approach. We also compare our approach with the RANSAC baseline which is the standard method for solving robust estimation problems. We can see from the table that our approach outperforms the other three baselines by 3.39, 0.89, 2.21mm respectively in average. Considering these baselines are already very strong, the improvement is significant. The results demonstrate that adding the confidence matrix of heatmaps and joints' confidence can significant improve the performance and model's robustness.

In addition, we also compare our approach with existing state-of-the-art methods for multi-view 3D pose estimation. The results are presented in Table 3.

From the table we can find that our approach surpasses the state-of-the-arts in average. The performance of our approach is 27.3mm with improvement of 3.6mm comparing with the second best method. The improvement is significant considering that the error of the state-of-the-art is already very small. We also show some 2D and 3D pose estimation results in Fig. 4.

Table 2. The 3D pose estimation error (mm) of the baseline methods and our approach on the Human3.6M dataset.

Methods	MPJPE,mm
NoFuse	30.7
HeuristicFuse	28.2
RANSAC	29.5
Ours	**27.3**

Table 3. The 3D pose estimation error (mm) of the state-of-the-arts and our approach on the Human3.6M dataset

Methods	MPJPE,mm
Pavlakos et al. [30]	56.9
Tome et al. [31]	52.8
Qiu et al. [7]	31.1
Gordon et al. [32]	30.9
Ours	**27.3**

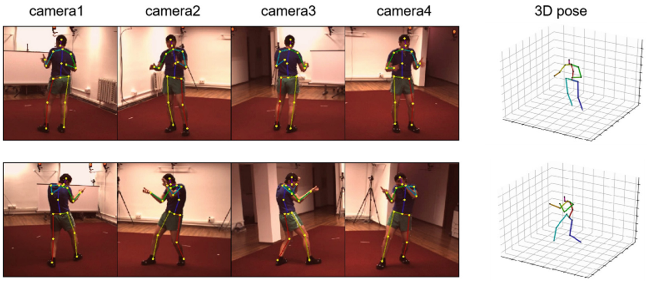

Fig. 4. Examples of 2D and 3D pose estimation on Human3.6M dataset.

4.3 Results on Occlusion-Person

2D Pose Estimation Results. Table 4 shows the results on the Occlusion-Person dataset. From the table we can find that our approach is significantly improved compared with NoFuse. This is reasonable because the features of the

Table 4. The 2D pose estimation accuracy (PCK) of the baseline methods and our approach on the Occlusion-Person dataset

Methods	Hip	Knee	Ankle	Shlder	Elbow	Wrist	Mean
NoFuse	63.4	21.5	17.0	29.5	14.6	12.4	30.9
HeuristicFuse	76.9	59.0	73.4	63.5	49.0	54.8	65.0
Ours	**97.7**	**94.4**	**91.0**	**97.9**	**91.0**	**93.1**	**94.2**

occluded joints are severely corrupted and the results demonstrate the advantage of our approach for dealing with occlusion.

3D Pose Estimation Results. The results of 3D pose estimation error (mm) are presented in Table 5. The result of NoFuse is 41.64mm which is a large error. The performance of our model is 9.67mm which means our approach can better handle occlusions than other baselines. Since there are very few works have report results on this new dataset, we only compare our approach to the three baselines. Figure 5 shows some examples of 2D and 3D pose estimation on Occlusion-Person dataset.

Table 5. The 3D pose estimation error (mm) of the baseline methods and our approach on the Occlusion-Person dataset.

Methods	MPJPE,mm
NoFuse	41.6
HeuristicFuse	13.4
RANSAC	12.4
Ours	**9.7**

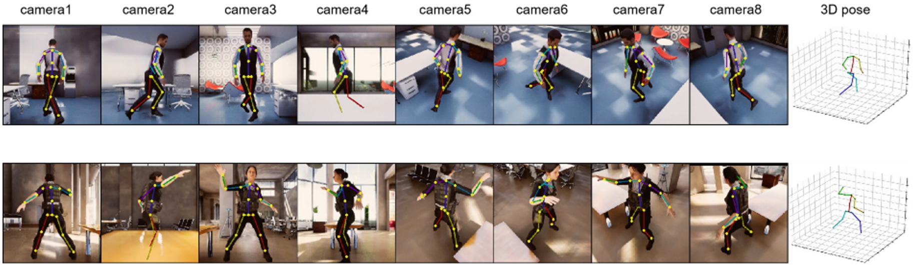

Fig. 5. Examples of 2D and 3D pose estimation on Occlusion-Person dataset.

5 Conclusion

In this paper, we present a novel approach for multi-view human pose estimation. Different from previous methods, we propose to extract weights of views and heatmaps to reflect their quality. The experimental results on the two datasets validate that the approach is efficient and robust to occlusion.

References

1. Bridgeman, L., Volino. M., Guillemaut, J.Y., et al.: Multi-person 3D pose estimation and tracking in sports. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE (2019)
2. Tu, H., Wang, C., Zeng, W.: VoxelPose: towards multi-camera 3D human pose estimation in wild environment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 197–212. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_12
3. Zhe, C., Simon, T., Wei, S.E., et al.: Realtime multi-person 2D pose estimation using part affinity fields. IEEE (2017)
4. He, K., Gkioxari, G., Dollár, P., et al.: Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. (2017)
5. Joo, H., Simon, T., Li, X., et al.: Panoptic studio: a massively multiview system for social interaction capture. IEEE Trans. Pattern Anal. Mach. Intell. 99 (2016)
6. Belagiannis, V., Sikandar, A., et al.: 3D pictorial structures revisited: multiple human pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. **38**, 1929–1942 (2016)
7. Qiu, H., Wang, C., Wang, J., et al.: Cross view fusion for 3D human pose estimation. University of Science and Technology of China; Microsoft Research Asia; TuSimple; Microsoft Research (2019)
8. Ionescu, C., Papava, D., Olaru, V., et al.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
9. Zhang, Z., Wang, C., Qiu, W., et al.: AdaFuse: adaptive multiview fusion for accurate human pose estimation in the wild. arXiv e-prints (2020)
10. Oberweger, M., Wohlhart, P., Lepetit, V.: DeepPose: human pose estimation via deep neural networks
11. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
12. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (2017)
13. Carreira, J., Agrawal, P., Fragkiadaki, K., et al.: Human pose estimation with iterative error feedback. IEEE (2015)
14. Ke, L., Chang, M.C., Qi, H., et al.: Multi-scale structure-aware network for human pose estimation (2018)
15. Fang, H.S., Xie, S., Tai, Y.W., et al.: RMPE: Regional Multi-person Pose Estimation. IEEE (2017)
16. Chen, Y., Wang, Z., Peng, Y., et al.: Cascaded pyramid network for multi-person pose estimation

17. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. arXiv e-prints (2018)
18. Li, J., Wang, C., Zhu, H., et al.: CrowdPose: efficient crowded scenes pose estimation and a new benchmark (2018)
19. Pishchulin, L., Insafutdinov, E., Tang, S., et al.: DeepCut: joint subset partition and labeling for multi person pose estimation. IEEE (2016)
20. Insafutdinov, E., Pishchulin, L., Andres, B., et al.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. arXiv e-prints (2016)
21. Amin S, Andriluka M, Rohrbach M, et al. Multi-view Pictorial Structures for 3D Human Pose Estimation[C]// British Machine Vision Conference 2013. 2013
22. Wang, C., Wang, Y., Lin, Z., et al.: Robust estimation of 3D human poses from a single image. arXiv e-prints (2014)
23. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D human pose from 2D image landmarks. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 573–586. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_41
24. Iskakov, K., Burkov, E., Lempitsky, V., et al.: Learnable triangulation of human pose. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (2020)
25. Lin, J., Lee, G.H.: Multi-view multi-person 3D pose estimation with plane sweep stereo (2021)
26. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. IEEE (2017)
27. Pavlakos, G., Zhou, X., Derpanis, K.G., et al.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: IEEE Conference on Computer Vision & Pattern Recognition. IEEE (2017)
28. Martinez, J., Hossain, R., Romero, J., et al.: A simple yet effective baseline for 3D human pose estimation. IEEE Computer Society (2017)
29. Andriluka, M., Pishchulin, L., Gehler, P., et al.: Human pose estimation: new benchmark and state of the art analysis. In: Computer Vision and Pattern Recognition (CVPR). IEEE (2014)
30. Pavlakos, G., Zhou, X., Derpanis, K.G., et al.: Harvesting multiple views for marker-less 3D human pose annotations. IEEE (2017)
31. Tome, D., Toso, M., Agapito, L., et al.: Rethinking pose in 3D: multi-stage refinement and recovery for markerless motion capture. IEEE (2018)
32. Gordon, B., Raab, S., Azov, G., et al.: FLEX: parameter-free multi-view 3D human motion reconstruction (2021)