# Multi-person 3D Pose Estimation from Monocular Image Sequences

Ran Li, Nayun Xu, Xutong Lu, Yucheng Xing, Haohua Zhao, Li Niu, and Liqing Zhang(✉)

Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China
{liran920526,xunayun,luxutong,haoh.zhao,ustcnewly}@sjtu.edu.cn, ericxing0430@gmail.com, zhang-lq@cs.sjtu.edu.cn

**Abstract.** This article tackles the problem of multi-person 3D human pose estimation based on monocular image sequence in a three-step framework: (1) we detect 2D human skeletons in each frame across the image sequence; (2) we track each person through the image sequence and identify the sequence of 2D skeletons for each person; (3) we reconstruct the 3D human skeleton for each person from the detected 2D human joints, by using prelearned base poses and considering the temporal smoothness. We evaluate our framework on the Human3.6M dataset and the multi-person image sequence captured by ourselves. The quantitative results on the Human3.6M dataset and the qualitative results on our constructed test data demonstrate the effectiveness of our proposed method.

**Keywords:** 3D human pose estimation · 2D human pose estimation · Human tracking

## 1 Introduction

With the rapid development of visual action recognition, 3D human skeleton reconstruction in a single image and image sequences has attracted plenty of attention in recent years. Compared with 2D human skeleton, 3D human skeleton generally leads to better performance of action recognition, due to the rotation invariance of 3D human skeleton. Several works have been done for action recognition based on 3D human skeleton. For example, the work in [17] explored Lie group theory to represent dynamics of the 3D human skeletons. Following [17], Lie group theory is combined with a deep network architecture to learn more representative features in [6]. The work [8] and [9] proposed to use CNN and LSTM to extract the spatio-temporal feature from 3D human skeleton.

Although action recognition based on 3D human skeleton has achieved great success, collecting 3D human skeletons with wearable devices is very expensive and sometimes not accurate. An alternative way is to reconstruct 3D human skeleton from 2D human skeleton because 2D human skeletons are more accessible. However, the reconstruction of 3D human skeleton from 2D human skeleton


T. Gedeon et al. (Eds.): ICONIP 2019, LNCS 11954, pp. 15–24, 2019.
https://doi.org/10.1007/978-3-030-36711-4_2

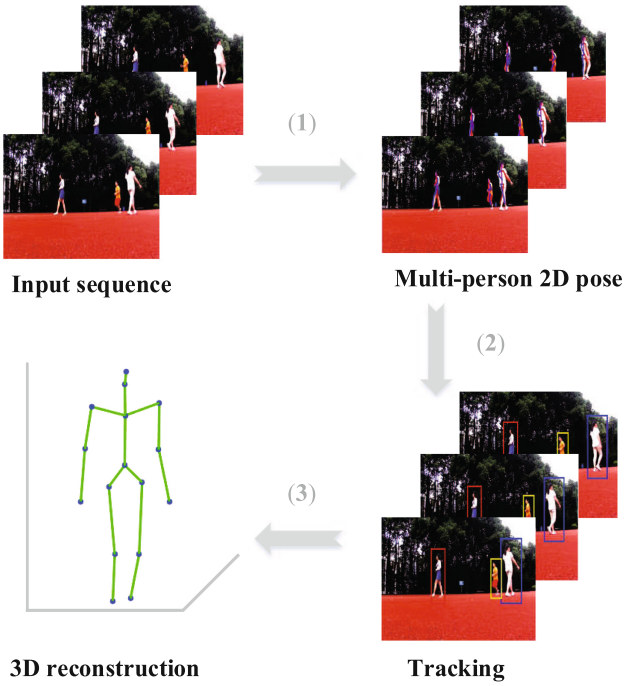



**Fig. 1.** Framework overview: (1) We first estimate the 2D pose in each frame using Regional Multi-person Pose Estimation (RMPE). (2) Then, we track each person using Discriminative Scale Space Tracker (DSST). (3) Finally, we reconstruct 3D human pose from estimated 2D human pose for each person in each frame.

is a very difficult task, because one 2D pose may correspond to multiple 3D poses with different camera parameters. To address this problem, some recent works [18,21] alternately update the 3D pose and camera parameters, yielding the estimated camera parameters and the corresponding 3D pose.

Nevertheless, 2D human skeletons could also be unavailable in the real world. In this case, our goal is to reconstruct 3D human skeleton from monocular image sequence, in which we need to initially estimate the 2D human skeletons based on the image sequence. Moreover, multiple persons may appear simultaneously in one image sequence, so we need to separate the sequence of 2D human skeletons belonging to different persons. Therefore, the design of an integrated framework remains an open and challenging problem.

In this paper, we propose a multi-person 3D pose estimation framework based on monocular image sequences, which integrates a 2D human pose estimator, a human tracker, and a 3D reconstruction method based on the estimated 2D human pose. Specifically, we first detect 2D human skeletons by locating 2D human joints in each frame across the image sequence using Regional Multi-person Pose Estimation (RMPE) [5]. Then, we track each person through the image sequence and identify the sequence of 2D skeletons for each person by

using Discriminative Scale Space Tracker (DSST) [4]. Finally, we reconstruct 3D human skeletons from detected 2D skeletons by using prelearned base poses and considering the temporal smoothness. Our framework is illustrated in Fig. 1.

Our experiments are conducted on the single-person Human3.6M dataset and multi-person image sequence collected by ourselves. The results on the Human3.6M dataset demonstrate that the proposed method outperforms the current state-of-art methods. Besides, the estimated 3D human pose in multi-person image sequence shows the advantage of our proposed framework in a qualitative fashion.

## 2 Related Work

In this section, we will discuss the related works on 3D human pose estimation based on single image or image sequence.

### 2.1 3D Human Pose Estimation Based on Single Image

Most papers on 3D pose estimation assume that 3D poses can be represented by a linear combination of a set of base poses, and learn the combination coefficients of bases poses. Some works [13,15] require manually labeled 2D joint locations as input while other works [14,19,20] only require a single image as input. For the works only requiring a single image, they either jointly estimate the 2D pose and the 3D pose [14], or initially estimate the 2D pose followed by 3D pose reconstruction [20]. However, all the above works focus on a single image while our framework focuses on image sequence.

### 2.2 3D Human Pose Estimation Based on Image Sequence

There also exist some works [18,21] using monocular image sequence as input. Compared with those works based on a single image, they exploit the relation between neighboring frames in the image sequence. For instance, Wandt *et al.* [18] make a strong periodic assumption on 3D human poses for periodic motion (*e.g.*, walking and running), and use the variances of all bone lengths as a regularizer for non-periodic motion. Zhou *et al.* [21] applied the discrete temporal derivative operator to make the 3D human poses across the sequences smoother. However, all these approaches only deal with single-person image sequence while our framework can handle the multi-person image sequence.

## 3 Our Framework

Our framework consists of three steps. In the first step, we adopt the Regional Multi-person Pose Estimation (RMPE) [5] to obtain the 2D joint locations in each frame across the image sequence, given the fact that RMPE has shown excellent performance for multi-person 2D pose estimation. In the second step, with

the detected 2D human skeletons in each frame, we use the Discriminative Scale Space Tracker (DSST) [4] to identify the sequence of bounding boxes belonging to each person, leading to a sequence of 2D human skeletons belonging to each person. The reason of choosing DSST tracker can be explained as follows. In the real world, people may move towards different directions, closer to or farther from the camera, resulting in different scales of person in different frames. Thus, we adopt the DSST, which is robust with various scales, to track each person. In the third step, for 3D human pose estimation, we use a 3D human skeleton reconstruction algorithm, which takes a sequence of 2D poses of each person as input and output a sequence of 3D poses for each person.

### 3.1 Representation of 3D Poses Using Base Poses

We use $\mathbf{Y}_t \in \mathbb{R}^{3\times a}$ to denote the 3D pose in $t$-th frame with $\mathbf{a}$ being the number of joints. Following [2,13], to regulate reconstructed 3D pose, we assume that 3D pose can be represented as a linear combination of the $\mathbf{K}$ base poses $\mathbf{Q}_k \in \mathbb{R}^{3\times a}, k = 1, 2...K$:

$$\mathbf{Y}_t = \sum_{k=1}^{K} (\mathbf{w}_{kt} \times \mathbf{Q}_k), \tag{1}$$

where $\mathbf{w}_{kt}$ is the coefficient corresponding to the $k$-th base pose for the $t$-th frame. We learn $K$ ($K = 64$ in our experiments) base poses from the motion capture dataset, following the method used in [2,13]. The learned base poses form an overcomplete dictionary, which means that the number of the base poses is large and the combination coefficients of a given 3D pose are sparse.

### 3.2 3D Pose Reconstruction Based on a Sequence of 2D Poses

Given a 3D pose of the $t$-th frame $\mathbf{Y}_t$ and camera parameters, we can obtain the corresponding 2D pose $\mathbf{X}_t \in \mathbb{R}^{2\times a}$. In particular, with the camera parameters including the projection matrix $\mathbf{M}_t \in \mathbb{R}^{2\times 3}$ and translation vector $\mathbf{T}_t \in \mathbb{R}^2$, the 3D pose $\mathbf{Y}_t$ can be projected to the 2D pose $\mathbf{X}_t$ as follows,

$$\mathbf{X}_t = \mathbf{M}_t \times \mathbf{Y}_t + \mathbf{T}_t \mathbf{1}^T, \tag{2}$$

where $\mathbf{1}$ is an $a$-dim column vector. However, the equation in (2) only holds in ideal cases. Considering the representation error of 3D poses $\mathbf{Y}_t$ based on the linear combination of based poses, we tolerate the projection error to a certain degree and aim to minimize the following projection error:

$$P(\mathbf{W}, \mathbf{M}, \mathbf{T}) = \frac{1}{2} \|\mathbf{X}_t - \mathbf{M}_t \times \mathbf{Y}_t - \mathbf{T}_t \mathbf{1}^T\|_F^2, \tag{3}$$

where Frobenius norm is used to calculate the projection error, $\mathbf{X}$ (*resp.*, $\mathbf{W}$, $\mathbf{M}$, and $\mathbf{T}$) is the collection of $\mathbf{X}_t$ (*resp.*, $\mathbf{W}_t$, $\mathbf{M}_t$, and $\mathbf{T}_t$).

Considering that for the neighboring frames in the sequence, their camera parameter $\mathbf{M}$ and representation coefficient of base poses $\mathbf{W}$ should vary smoothly.

**Table 1.** Quantitative comparison with state-of-the-art results on Human3.6M dataset. We report the mean per joint errors (mm) of the test subjects S9 and S11.

| | 3D (mm) |
|---|---|
| LinKDE [7] | 162.14 |
| Tekin et al. [16] | 125.28 |
| Ours | **119.76** |

We impose first-order temporal smoothness regularizer $\|\nabla_t \mathbf{W}\|_F^2$ (*resp.*, $\mathbf{M}\|_F^2$) on $\mathbf{W}$ (*resp.*, $\mathbf{M}$), in which $\nabla_t$ stands for the derivative on temporal factor. Moreover, the representation coefficient of 3D poses $\mathbf{W}$ should be sparse according to the analysis in Sect. 3.1, so we add a L1 norm $\|\mathbf{W}\|_1$ to ensure the sparsity of $\mathbf{W}$. To this end, we collect the penalty terms as follows,

$$R(\mathbf{W}, \mathbf{M}) = \alpha\|\mathbf{W}\|_1 + \beta\|\nabla_t \mathbf{W}\|_F^2 + \gamma\|\nabla_t \mathbf{M}\|_F^2, \tag{4}$$

in which $\alpha$, $\beta$, and $\gamma$ are trade-off parameters and empirically fixed as 0.1, 5, and 0.5 respectively in our experiments.

By combining (3) and (4), we reach our final formulation:

$$\min_{\mathbf{W},\mathbf{M},\mathbf{T}} P(\mathbf{W}, \mathbf{M}, \mathbf{T}) + R(\mathbf{W}, \mathbf{M}). \tag{5}$$

By solving (5), we can obtain $\mathbf{W}_t$ and recover $\mathbf{Y}_t$ based on (2).

### 3.3 Optimization

The problem in (5) can be solved using block coordinate descent, which means alternatively updating one variable while fixing other variables.

**Update the Representation Coefficients W:** The subproblem *w.r.t.* $\mathbf{W}$ can be written as

$$\min_{\mathbf{W}} P(\mathbf{W}, \mathbf{M}, \mathbf{T}) + \alpha\|\mathbf{W}\|_1 + \beta\|\nabla_t \mathbf{W}\|_F^2, \tag{6}$$

which can be solved via accelerated proximal gradient (APG) [11]. Since this problem is convex, the global minimum can be guaranteed.

**Update the Rotation Matrix M:** The subproblem *w.r.t.* $\mathbf{M}$ can be written as

$$\min_{\mathbf{M}} P(\mathbf{W}, \mathbf{M}, \mathbf{T}) + \gamma\|\nabla_t \mathbf{M}\|_F^2, \tag{7}$$

which is a manifold optimization problem and can be solved via the matlab toolbox Manopt following [3].

**Update the Translation Matrix T:** The subproblem *w.r.t.* $\mathbf{T}$ can be written as

$$\mathbf{T}_t = \mathbf{1}^T \left( \mathbf{X}_t - \mathbf{M}_t \sum_{k=1}^{K} (\mathbf{w}_{kt} \times \mathbf{Q}_k) \right), \tag{8}$$

which computes the average of rows in $\mathbf{X}_t - \mathbf{M}_t \sum_{k=1}^{K} (\mathbf{w}_{kt} \times \mathbf{Q}_k)$.

We alternatingly update three variables until the objective function in (5) converges. In each step, the objective function is non-increasing, so the convergence of this algorithm is ensured. Besides, considering the impact of initialization on our solution, we opt for the initialization method proposed in [1], which proves to be very effective.

## 4 Experimental Results

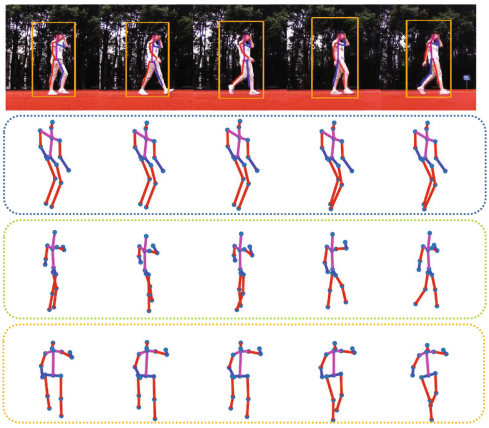
In this section, we first demonstrate the superiority of our framework on a recently published single-person dataset Human3.6M, because there is no available multi-person 3D pose estimation dataset based on image sequence as far as we are concerned. Then, we collect multi-person image sequences and compare our framework with the state-of-the-art methods in a qualitative fashion by showing the reconstructed 3D skeletons of two persons, which again verifies the effectiveness of our framework.

### 4.1 Implementation Details

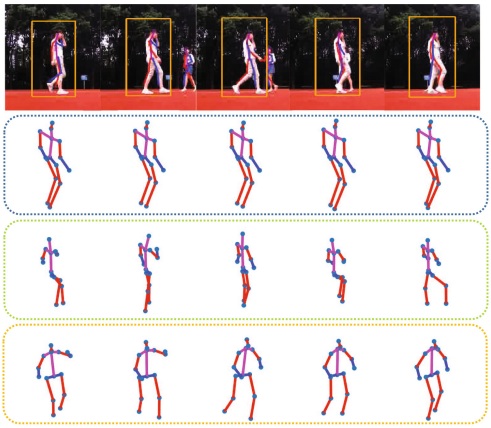
In our framework, for 2D human pose estimation, we use Regional Multi-person Pose Estimation (RMPE) [5] as the 2D human pose estimator to locate 16 2D human joints for each person, in which the stacked hourglass network structure [12] is adopted. For human tracking, we use Discriminative Scale Space Tracker (DSST) [4] with two correlation filers composed of one 1-dim scale filer and one 2-dim translation filter, in which Fast Fourier Transform (FFT) is applied to significantly improve the tracking efficiency. To initiate the tracking process, we employ VGG-based Single Shot Detector (SSD) [10] to detect the persons in the first frame of each image sequence, and then increase the detected human proposals by 30% both at length and width to ensure the intactness of each detected person. For 3D human skeleton reconstruction from a sequence of 2D human skeletons, we impose the periodic assumption [18] to ensure the smooth transition between neighboring frames across the sequence.

### 4.2 Evaluation on the Human3.6M Dataset

The Human3.6M dataset contains 11 subjects performing 17 actions, such as walking, smoking, and eating. All the videos are captured by a MoCap system from 4 different viewpoints in a controlled environment, in which each video has only one person. Accurate 2D human skeletons are also provided for all videos. The frame rate is downsampled from 50 fps to 10 fps. We use two subjects (S9, S11) to evaluate different methods and other subjects to learn $K$ base poses following previous work.



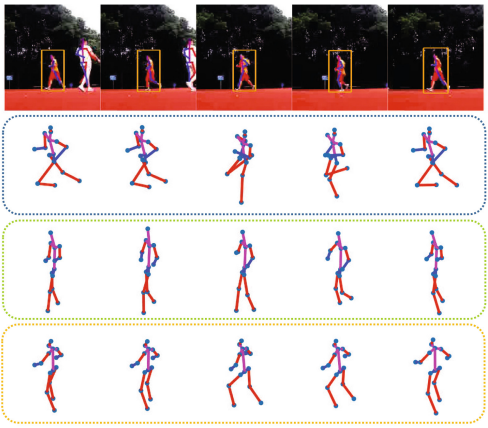
(a)The reconstruction result of walking person for the first five frames.



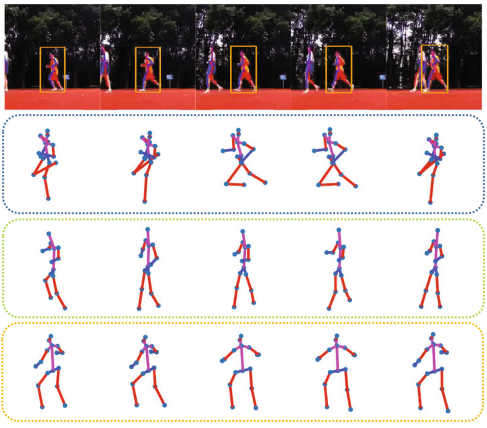
(b)The reconstruction result of walking person for the subsequent five frames.

**Fig. 2.** The comparison results of one running person within 10 frames.

In this experiment, we compare our framework with two state-of-the-art baselines on the evaluation set S9 and S11 of all the aspects. The first baseline [7] is based on the single frame regression. It is provided with the Hman3.6M dataset. The second baseline [16] explores motion information from consecutive frames in short image sequences. The results are summarized in Table 1, which demonstrates that our framework outperforms other methods on the large-scale Human3.6M dataset, which shows the effectiveness of integrating 2D human pose estimator and sequential 3D human skeleton reconstruction into our framework.



(a)The reconstruction result of running person for the first five frames.



(b)The reconstruction result of running person for the subsequent five frames.

**Fig. 3.** We select 10 frames of two persons in our captured multi-person image sequence to show the comparative results. In both Figs. 3 and 2, the first row is the original image sequences with 2D human pose estimator and human tracking. The second row is the reconstruction result of RMPE+DSST+[18]. The third row is the reconstruction result of RMPE+DSST+[13]. The fourth row is the reconstruction result of our proposed framework.

### 4.3 Evaluation on Multi-person Image Sequence

To the best of our knowledge, there is no available multi-person 3D pose estimation dataset based on image sequence. Hence, we sample 10 frames for each person from one captured video clip with two persons and 91 frames and evaluate our framework on this multi-person image sequence. We compare our framework with two baseline methods [13,18], in which the former method reconstructs the

3D human skeleton from monocular image sequences with the periodic assumption and the latter method estimates the 3D human poses from single images. By combining these two methods with our used 2D human pose estimator (*i.e.*, RMPE) and tracker (*i.e.*, DSST), we can also obtain the multi-person 3D human skeletons from the image sequence. The selected images with tracking and 2D pose detector and the comparative results are showed in Figs. 3 and 2.

The comparative results of one walking person are illustrated in Fig. 3. In the second row, we can observe that the results obtained by [18] fails to reconstruct the 3D skeletons corresponding to hand raising during walking, and the reconstructed skeletons rarely move. In the third row, we can see that the results obtained by [13] contain some strange poses like the 6th and the 9th frame. Our reconstruction results are shown in the last row, in which the 3D skeletons corresponding to hand raising bear a strong resemblance to the original image sequences and the transition between neighboring frames is smoother compared with the other two methods.

The comparative results of one running person are illustrated in Fig. 2. In the second row, we can observe that the results obtained by [18] always use the wrong legs and the motion is a little dramatic. In the third row, it can be seen that the results obtained by [13] contain sharp changes and strange poses. For example, the angle of the left knee is unrealistic in the 9th frame on the 3rd row. In contrast, as shown in the last row, our framework can achieve both smooth and realistic 3D human skeletons.

## 5 Conclusion

In this paper, we have proposed a multi-person 3D pose estimation framework based on monocular image sequences. Our proposed framework has integrated a 2D human pose estimator, a human tracker, and 3D human skeleton reconstruction in a coherent and effective manner. According to the quantitative comparison on the recently published large-scale dataset Human3.6M and the qualitative analyses on our captured multi-person image sequence, our framework achieve better results than state-of-the-art baseline methods, which clearly demonstrate the validness of our proposed framework.

**Acknowledgements.** The work was supported by the National Basic Research Program of China under Grant (No. 2015CB856004) and the Key Basic Research Program of Shanghai Science and Technology Commission, China under Grant (Nos. 15JC1400103, 16JC1402800).

## References

1. 3D shape estimation from 2D landmarks: a convex relaxation approach (2015)
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR, pp. 1446–1455 (2015)
3. Boumal, N.: Manopt, a matlab toolbox for optimization on manifolds. JMLR **15**(1), 1455–1459 (2014)

4. Danelljan, M., Häger, G.: Accurate scale estimation for robust visual tracking. In: BMVC. BMVA Press (2014)
5. Fang, H.: RMPE: regional multi-person pose estimation. CoRR abs/1612.00137, 4321–4330 (2016)
6. Huang, Z., Wan, C., Probst, T., Van Gool, L.: Deep learning on lie groups for skeleton-based action recognition. In: CVPR, pp. 1243–1252 (2017)
7. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. T-PAMI **36**(7), 1325–1339 (2014)
8. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: CVPR, pp. 4570–4579 (2017)
9. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention LSTM networks for 3D action recognition. In: CVPR, vol. 7, p. 43 (2017)
10. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
11. Nesterov, Y., et al.: Gradient methods for minimizing composite objective function (2007)
12. 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
13. 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
14. Simo-Serra, E., Quattoni, A.: A joint model for 2D and 3D pose estimation from a single image. In: CVPR, pp. 3634–3641 (2013)
15. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In: CVPR, pp. 1677–1684 (2000)
16. Tekin, B., Sun, X., Wang, X., Lepetit, V., Fua, P.: Predicting people's 3D poses from short sequences. CoRR abs/1504.08200 (2015)
17. Vemulapalli, R., Arrate, F.: Human action recognition by representing 3D skeletons as points in a lie group. In: CVPR, pp. 588–595 (2014)
18. Wandt, B.: 3D reconstruction of human motion from monocular image sequences. T-PAMI **38**, 1505–1516 (2016)
19. Wang, C., Wang, Y.: Robust estimation of 3D human poses from a single image. In: CVPR, pp. 2369–2376 (2014)
20. Yasin, H., Iqbal, U.: A dual-source approach for 3D pose estimation from a single image. In: CVPR, pp. 4948–4956 (2016)
21. Zhou, X., Zhu, M.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: CVPR, pp. 4966–4975 (2016)