Towards Generalization of 3D Human Pose Estimation in the Wild

Renato Baptista(✉), Alexandre Saint(✉), Kassem Al Ismaeil(✉), and Djamila Aouada(✉)

Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg, Luxembourg
{renato.baptista,alexandre.saint,kassem.alismaeil,djamila.aouada}@uni.lu

Abstract. In this paper, we propose 3DBodyTex.Pose, a dataset that addresses the task of 3D human pose estimation in-the-wild. Generalization to in-the-wild images remains limited due to the lack of adequate datasets. Existent ones are usually collected in indoor controlled environments where motion capture systems are used to obtain the 3D ground-truth annotations of humans. 3DBodyTex.Pose offers high quality and rich data containing 405 different real subjects in various clothing and poses, and 81k image samples with ground-truth 2D and 3D pose annotations. These images are generated from 200 viewpoints among which 70 challenging extreme viewpoints. This data was created starting from high resolution textured 3D body scans and by incorporating various realistic backgrounds. Retraining a state-of-the-art 3D pose estimation approach using data augmented with 3DBodyTex.Pose showed promising improvement in the overall performance, and a sensible decrease in the per joint position error when testing on challenging viewpoints. The 3DBodyTex.Pose is expected to offer the research community with new possibilities for generalizing 3D pose estimation from monocular in-the-wild images.

Keywords: 3D human pose estimation · 3DBodyTex.Pose · Synthetic data · In-the-wild

1 Introduction

In the past couple of years, human pose estimation has received a lot of attention from the computer vision community. The goal is to estimate the 2D or 3D position of the human body joints given an image containing a human subject. This has a significant number of applications such as sports, healthcare solutions [3], action recognition [3,6], and animations.

Due to the recent advances in Deep Neural Networks (DNN), the task of 2D human pose estimation a great improvement in results [4,5,14]. This has been mostly achieved thanks to the availability of large-scale datasets containing 2D annotations of humans in many different conditions, e.g., in the wild [2]. In contrast, advances in the task of human pose estimation in 3D remains limited. The

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12661, pp. 72–81, 2021.
https://doi.org/10.1007/978-3-030-68763-2_6

Fig. 1. Examples of the 3D body scans used to generate in-the-wild images with 2D and 3D annotations of humans.

main reasons are the ambiguity of recovering the 3D information from a single image, in addition to the lack of large-scale datasets with 3D annotations of humans, specifically considering in-the-wild conditions. Existent datasets with 3D annotations are usually collected in a controlled environment using Motion Capture (MoCap) systems [9], or with depth maps [7,17]. Consequently, the variations in background and camera viewpoints remain limited. In addition, DNNs [28] trained on such datasets have difficulties generalizing well to environments where a lot of variation is present, e.g., scenarios in the wild.

Recently, many works focused on the challenging problem of 3D human pose estimation in the wild [16,18,26,27]. These works differ significantly from each other but share an important aspect. They are usually evaluated on the same dataset that has been used for training. Thus, it is possible that these approaches have been over-optimized for specific datasets, leading to a lack of generalization. It becomes difficult to judge on the generalization, and more precisely for in-the-wild scenarios where variations coming from the background and camera viewpoints are always present.

In order to address the aforementioned challenge, this paper presents a new dataset referred to as *3DBodyTex.Pose*. It is an original dataset generated from high-resolution textured 3D body scans, similar in quality to the ones contained in the 3DBodyTex dataset introduced in [19] and later on presented in the *SHARP2020* challenge [20,21]. 3DBodyTex.Pose is dedicated to the task of human pose estimation. Synthetic scenes are generated with ground-truth information from real 3D body scans, with a large variation in subjects, clothing, and poses (see Fig. 1). Realistic background is incorporated to the 3D environment. Finally, 2D images are generated from different camera viewpoints, including challenging ones, by virtually changing the camera location and orientation. We distinguish extreme viewpoints as the cases where the camera is, for example, placed on top of the subject. With the information contained in 3DBodyTex.Pose, it becomes possible to better generalize the problem of the 3D human pose estimation to in-the-wild images independently of the camera viewpoint as shown experimentally on a state-of-the-art 3D pose estimation approach [27]. In summary, the contributions of this work are:

(1) 3DBodyTex.Pose, a synthetic dataset with 2D and 3D annotations of human poses considering in-the-wild images, with realistic background and standard to extreme camera viewpoints. This dataset will be publicly available for the research community.
(2) Increasing the robustness of 3D human pose estimation algorithms, specifically [27], with respect to challenging camera viewpoints thanks to data augmentation with 3DBodyTex.Pose.

The rest of the paper is organized as follows: Sect. 2 describes the related datasets for the 3D human pose estimation task. Section 3 provides details about the proposed 3DBodyTex.Pose dataset and how it addresses the challenges of in-the-wild images and extreme camera viewpoints. Then, Sect. 4 shows the conducted experiments, and finally Sect. 5 concludes this work.

2 Related Datasets

Monocular 3D human pose estimation aims to estimate the 3D joint locations from the human present in the image independently of the environment of the scene. However, usually not all camera viewpoints are taken into consideration. Consequently, the 3D human body joints are not well estimated for the cases where the person is not fully visible or self-occluded. In order to use such images for training, labels for the position of the 2D human joints are needed as ground-truth information [2,10]. Labeling such images from extreme camera viewpoints is an expensive and difficult task as it often requires manual annotation. To overcome this issue, MoCap systems can be used for precisely labeling the data. However, they are used in a controlled environment such as indoor scenarios. The Human3.6M dataset [9] is widely used for the task of 3D human pose estimation and it falls under this scenario. It contains 3.6M frames with 2D and 3D annotations of humans from 4 different camera viewpoints. The HumanEva-I [23] and TotalCapture [24] datasets are also captured in indoor environments. HumanEva-I contains 40k frames with 2D and 3D annotations from 7 different camera viewpoints. TotalCapture contains approximately 1.9M frames considering 8 camera locations where the 3D annotations of humans were obtained by fusing the MoCap with inertial measurement units. Also captured within a controlled environment, the authors of [13] proposed the MPII-INF-3DHP dataset for 3D human pose estimation which was recorded in a studio using a green screen background to allow automatic segmentation and augmentation. Consequently, the authors augment the data in terms of foreground and background, where the clothing color is changed on a pixel basis, and for the background, images sampled from the internet are used. Recently, von Marcard *et al.* [11] proposed a dataset with 3D pose in outdoor scenarios recorded with a moving camera. It contains more than 51k frames and 7 actors with a limited number of clothing style.

An alternative proposed with SURREAL [25] and exploited in [15], is to generate realistic ground-truth data synthetically. SURREAL places a parametric body model with varied pose and shape over a background image of a scene to

Table 1. Comparison of datasets for the task of 3D human pose estimation. (⋆) indicates that clothing was synthetically added to the dataset.

	3DBodyTex. Pose (Ours)	HumanEva-I	Human3.6M	MPII-INF–3DHP	TotalCapture	3DPW	SURREAL
# of subjects	405	4	11	8	5	7	n/a
# of samples	81k	40k	∼3.6M	>1.3M	∼1.9M	>51k	∼6.5M
Ground-truth pose	2D+3D	2D+3D	2D+3D	3D	3D	3D	2D+3D
Real people	Yes	Yes	Yes	Yes	Yes	Yes	No
Background	Indoor & outdoor	Indoor	Indoor	Green screen	Indoor	Outdoor	Indoor
Clothing	Realistic	Realistic	Realistic	Realistic$^{(\star)}$	No	Limited	No
# of total camera viewpoints	200	7	4	14	8	n/a	n/a
# of challenging viewpoints	70	0	0	3	0	n/a	n/a

simulate a monocular acquisition. Ground-truth 2D and 3D poses are known from the body model. To add realism, the body model is mapped with clothing texture. A drawback of this approach is that the body shape lacks details.

The 3DBodyTex dataset [19] contains static 3D body scans from people in close-fitting clothing, in varied poses and with ground-truth 3D pose. This dataset is not meant for the task of 3D human pose estimation. However, it is appealing for its realism: detailed shape and high-resolution texture information. It has been exploited for 3D human body fitting [22] and it could also be used to synthesize realistic monocular images from arbitrary viewpoints with ground-truth 2D and 3D poses. The main drawback of this dataset is the fact that it contains the same tight clothing with no variations.

3 Proposed 3DBodyTex.Pose Dataset

In contrast with 3DBodyTex, the new 3DBodyTex.Pose dataset contains 3D body scans that are captured from 405 subjects in their own regular clothes. From these 405 subjects, 204 are females and 201 are males. Having different clothing style from different people adds more variation to the dataset when considering in-the-wild scenarios. Figure 1 shows a couple of examples of 3D body scans with different clothing. In this work, the goal is to use the 3D body scans to synthesize realistic monocular images from arbitrary camera viewpoints with its corresponding 2D and 3D ground-truth information for the task of 3D human pose estimation. The principal characteristics of 3DBodyTex.Pose are compared to state-of-the-art datasets in Table 1.

The 3DBodyTex.Pose dataset aims to address the challenges of in-the-wild images and the extreme camera viewpoints. Given that the only input is the set of 3D scans, we need to estimate the ground-truth 3D skeletons, to synthesize the monocular images from challenging viewpoints and to simulate an in-the-wild environment. These three stages are detailed below.

Ground-Truth 3D Joints. To estimate the ground-truth 3D skeleton, we follow the automatic approach of 3DBodyTex [19] where body landmarks are first detected in 2D views before being robustly aggregated into 3D positions. Hence, for every 3D scan we have the corresponding 3D positions of the human body joints that is henceforth used as the ground-truth 3D skeleton.

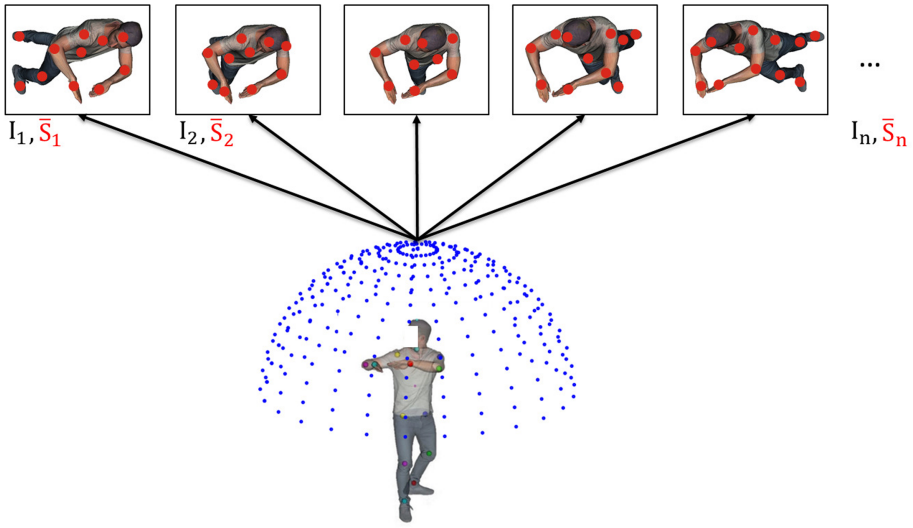

Fig. 2. Extreme camera viewpoints images (top row) from a single 3D body scan. The blue dots represent the camera locations for each camera viewpoint. Better visualized in color. (Color figure online)

Challenging Viewpoints. We propose to change the location and orientation of the camera in order to create monocular images where also extreme viewpoints are considered, see Fig. 2. Considering a 3D body scan $\mathsf{P} \in \mathbb{R}^{3\times K}$, where K is the number of vertices of the mesh, the 3D skeleton with J joints $\mathsf{S} \in \mathbb{R}^{3\times J}$, and also the homogeneous projection matrix $\mathsf{M_n}$ for the camera position n, we can back-project the 3D skeleton into the image $\mathsf{I_n}$ by

$$\bar{\mathsf{S}}_n = \mathsf{M}_n \cdot \mathsf{S}, \tag{1}$$

where $\bar{\mathsf{S}}_n \in \mathbb{R}^{3\times J}$ represents the homogeneous coordinates of the projected 3D skeleton into the image plane I_n, corresponding to a 2D skeleton. In this way, we are able to generate all possible camera viewpoints around the subject and easily obtain the corresponding 2D skeleton. In summary, each element of the 3DBodyTex.Pose is composed of image I_n, 2D skeleton $\bar{\mathsf{S}}_n$, and 3D skeleton S in the camera coordinate system.

In-the-Wild Environment. In order to address the challenge of the in-the-wild images with ground-truth information for the task of 3D human pose estimation, we further propose to embed the 3D scan in an environment with cube mapping [8] which in turns adds a realistic background variation to the dataset. An example texture cube is shown in Fig. 3(a). The six faces are mapped to a cube surrounding the scene with the 3D body scan at the center, see Fig. 3(b). Realistic textures cubes are obtained from [1].

(a) (b)

Fig. 3. (a) Example of an unfolded cube projection of a 3D environment (extracted from [1]). (b) Example of a 3D body scan added to the 3D environment of a realistic scene.

To have variation in the data, for each image, we randomly draw a texture cube, a camera viewpoint and a 3D scan. The proposed 3DBodyTex.Pose dataset provides reliable ground-truth 2D and 3D annotations with realistic and varied in-the-wild images while considering arbitrary camera viewpoints. Moreover, it offers a relatively high number of subjects in comparison with state-of-the-art 3D pose datasets, refer to Table 1. It also offers richer body details in terms of clothing, shape, and the realistic texture. Figure 4 shows the data generation overview.

4 Experimental Evaluation

In what follows, we use the approach proposed by Zhou *et al.* [27] to showcase the impact of the 3DBodyTex.Pose dataset in improving the performance of 3D pose estimation in the wild. We note that, in a similar fashion, 3DBodyTex.Pose can be used to enhance any other existent approach. Our goal is to share this new dataset with the research community and encourage (re-)evaluating and (re-)training existent and new 3D pose estimation approaches especially considering in-the-wild scenarios with a special focus on extreme viewpoints.

4.1 Baseline 3D Pose Estimation Approach

The work in [27] aims to estimate 3D human poses in the wild. For that, the authors proposed to couple together in-the-wild images with 2D annotations with indoor images with 3D annotations in an end-to-end framework. The authors provided the code for both training and testing the network.

The network proposed in [27] consists of two different modules: (1) 2D pose estimation module; and (2) depth regression module. In the first module, the

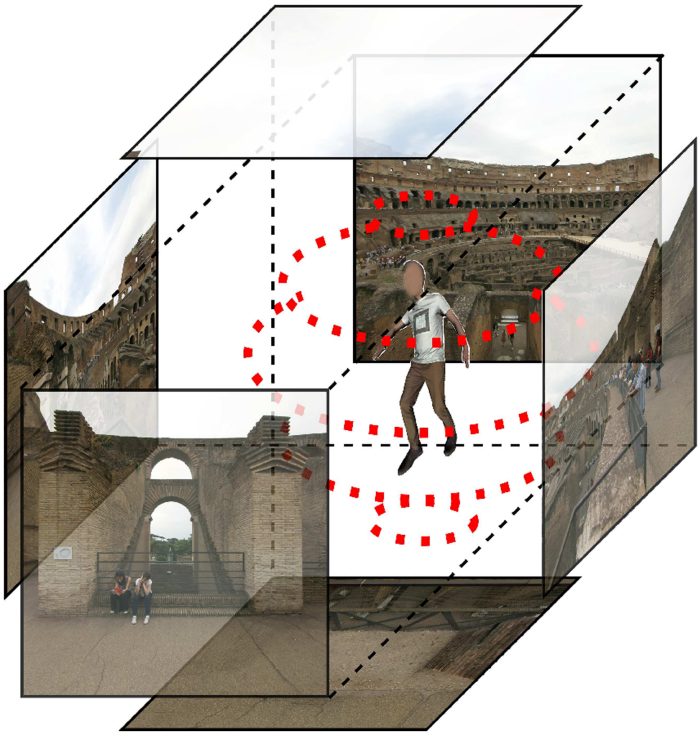

Fig. 4. Data generation overview. The 3D body scan is placed in the center of the cube mapping environment. Different camera viewpoints (in red) are considered in order to capture the scene from multiple angles. Better visualized in color. (Color figure online)

goal is to predict a set of J heat maps by minimizing the L^2 distance between the predicted and the ground-truth heat maps where only images with 2D annotations were used (MPII dataset [2]). Secondly, the depth regression module learns to predict the depth between the camera and the image plane by using the images where 3D annotations are provided (Human3.6M dataset [9]). Also within the second module, the authors proposed a geometric constraint which serves as a regularization for depth prediction when the 3D annotations are not available. At the end, the network is built in a way that both modules are trained together.

4.2 Data Augmentation with 3DBodyTex.Pose

We propose to retrain the network presented in [27] by adding the 3DBodyTex.Pose data to the training set originally used in [27]. Specifically, 60k additional RGB images from 3DBodyTex.Pose and their corresponding 2D skeletons were used to increase the variation coming from realistic background and camera viewpoints.

Table 2. Quantitative results of the MPJPE in millimeters on the Human3.6M dataset following the same protocol as in [27]. The average column represents the average error value of all actions in the validation set.

Methods	Average (mm)
Zhou *et al.* [27]	64.9
Zhou *et al.* [27] ++ (Ours)	**61.3**
Martinez *et al.* [12]	62.9
Rogez *et al.* [18]	61.2
Yang *et al.* [26]	<u>58.6</u>

Table 3. Results of the MPJPE while testing on challenging camera viewpoints only.

Methods	Average (mm)
Zhou *et al.* [27]	292
Zhou *et al.* [27] ++ (Ours)	**267**

We first follow the same evaluation protocol as in [27] by testing on the Human3.6M dataset [9], and using the Mean Per Joint Position Error (MPJPE) in millimeters (mm) as an evaluation metric between 3D skeletons. Table 2 shows the results of retraining [27] by augmenting with 3DBodyTex.Pose (**Zhou *et al.*** [27] ++) along with other reported state-of-the-art results as a reference. Without using 3DBodyTex.Pose, the average error between the estimated 3D skeleton and the ground-truth annotation is 64.9 mm, and when retrained with the addition of our proposed dataset, the error decreases to 61.3 mm. This result is a very promising step towards the generalization of 3D human pose estimation for in-the-wild images. Despite the fact that testing in Table 2 is on Human3.6M (indoor scenes only), retraining with 3DBodyTex.Pose helps bring the performance of [27] closer to the top performing approaches [18,26] and even beating others, i.e., [12].

As one of the aims of this paper is to mitigate the effect of challenging camera viewpoints, we tested the performance of [27] on a new testing set containing challenging viewpoints only. These were selected from the 3DBodyTex.Pose dataset and reserved for testing only[1]. Table 3 shows that adding the 3DBodyTex.Pose to the training set in the network of [27] performs better when testing with challenging viewpoints only. Note that the relative high values of the errors, as compared to Table 2, are due to the fact that the depth regression module is learned with the 3D ground-truth poses of the Human3.6M dataset only.

5 Conclusion

This paper introduced the 3DBodyTex.Pose dataset as a new original dataset to support the research community in designing robust approaches for 3D human

[1] Never seen during training.

pose estimation in the wild, independently of the camera viewpoint. It contains synthetic but realistic monocular images with 2D and 3D human pose annotations, generated from diverse and high-quality textured 3D body scans. The potential of this dataset was demonstrated by retraining a state-of-the-art 3D human pose estimation approach. There is a significant improvement in performance when augmented with 3DBodyTex.Pose. This opens the door to the generalization of 3D human pose estimation to in-the-wild images. As future work, we intend to increase the size of the dataset covering more camera viewpoints and realistic backgrounds, and by adding different scaling factors with respect to the camera location in order to increase the generalization over the depth variation.

Acknowledgements. This work was funded by the National Research Fund (FNR), Luxembourg, under the projects C15/IS/10415355/3DACT/Bjorn Ottersten and AFRPPP/11806282. The authors are grateful to Artec3D, the volunteers, and to all present and former members of the CVI^2 group at SnT for participating in the data collection. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg.

References

1. Humus Cubemap. http://www.humus.name/index.php?page=Textures. Accessed 29 Jan 2020
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: CVPR (2014)
3. Baptista, R., et al.: Home self-training: visual feedback for assisting physical activity for stroke survivors. CMPB **176**, 111–120 (2019)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. arXiv:1812.08008 (2018)
5. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR (2017)
6. Demisse, G.G., Papadopoulos, K., Aouada, D., Ottersten, B.: Pose encoding for robust skeleton-based action recognition. In: CVPRW (2018)
7. D'Eusanio, A., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R.: Manual annotations on depth maps for human pose estimation. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) ICIAP 2019, Part I. LNCS, vol. 11751, pp. 233–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30642-7_21
8. Greene, N.: Environment mapping and other applications of world projections. IEEE CG&A **6**(11), 21–29 (1986)
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE TPAMI **36**(7), 1325–1339 (2014)
10. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC, vol. 2, p. 5. Citeseer (2010)
11. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part X. LNCS, vol. 11214, pp. 614–631. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_37

12. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017)
13. Mehta, D., et al.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 3DV (2017)
14. 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, Part VIII. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
15. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: CVPR (2018)
16. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR (2019)
17. Pini, S., D'Eusanio, A., Borghi, G., Vezzani, R., Cucchiara, R.: Baracca: a multimodal dataset for anthropometric measurements in automotive. In: International Joint Conference on Biometrics (IJCB) (2020)
18. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net++: multi-person 2D and 3D pose detection in natural images. IEEE TPAMI **42**(5), 1146–1161 (2019)
19. Saint, A., et al.: 3DBodyTex: textured 3D body dataset. In: 3DV (2018)
20. Saint, A., Kacem, A., Cherenkova, K., Aouada, D.: 3DBooSTeR: 3D body shape and texture recovery. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12536, pp. 726–740. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66096-3_49
21. Saint, A., et al.: SHARP 2020: the 1st shape recovery from partial textured 3D scans challenge results. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12536, pp. 741–755. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66096-3_50
22. Saint, A., Shabayek, A.E.R., Cherenkova, K., Gusev, G., Aouada, D., Ottersten, B.: Bodyfitr: robust automatic 3D human body fitting. In: ICIP (2019)
23. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV **87**(1–2), 4 (2010)
24. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3D human pose estimation fusing video and inertial sensors. In: BMVC, vol. 2, p. 3 (2017)
25. Varol, G., et al.: Learning from synthetic humans. In: CVPR (2017)
26. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3D human pose estimation in the wild by adversarial learning. In: CVPR (2018)
27. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)
28. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., Jégou, H. (eds.) ECCV 2016, Part III. LNCS, vol. 9915, pp. 186–201. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_17