Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200726

3D Single Person Pose Estimation Method Based on Deep Learning

Xinrui Yuan[1], Hairong Wang, Jun Wang
North Minzu University, Yinchuan, China

Abstract. In view of the significant effects of deep learning in graphics and image processing, research on human pose estimation methods using deep learning has attracted much attention, and many method models have been produced one after another. On the basis of tracking and in-depth study of domestic and foreign research results, this paper concentrates on 3D single person pose estimation methods, contrasts and analyzes three methods of end-to-end, staged and hybrid network models, and summarizes the characteristics of the methods. For evaluating method performance, set up an experimental environment, and utilize the Human3.6M data set to test several mainstream methods. The test results indicate that the hybrid network model method has a better performance in the field of human pose estimation.

Keywords. Human pose estimation, deep learning, end-to-end, staged, hybrid network

1. Introduction

Human posture is an important feature of living beings and has been widely used in application scenarios such as behavior recognition [1], interactive analysis [2], video surveillance [3], entertainment [4], virtual reality [5], animation[6], etc. Through human pose estimation, it can assist in judging other people's behaviors, actions, and identification. In recent years, researchers have discovered that the human pose estimation method based on deep learning can effectively develop the accuracy of the estimation [7], resulting in a large number of research achievements. Chen Y et al. [8] gave a method for estimating human pose using structure-aware convolutional networks. The training of three sub-networks optimized the problems of joint occlusion and overlap. Chou CJ et al. [9] learn the structure and configuration of human body parts through confrontation training to achieve human pose estimation. Great achievements have been made in 2D single person pose estimation [10], but due to the single data set and the division of key points, the accuracy cannot be greatly improved. For the past few years, a breakthrough has been made in 3D single person pose estimation. Zhou X et al. [11] integrated the kinematics object model into deep learning and performed effective 3D single person pose estimation. Tekin B et al. [12] presented a novel deep learning regression architecture that combines an autoencoder with CNN to improve the prediction accuracy between various parts of the human body. These two methods belong to the end-to-end 3D single person pose estimation method [13].

[1] Corresponding Author, Hairong Wang, North Minzu University, Yinchuan, China; E-mail: bmdwhr@163.com.

Due to the constraints of the 3D human annotation data set, they can not be applied to more complex networks. In order to deal with complex and changeable scenes more accurately, researchers began to use the increasingly accurate 2D human pose estimation method as an aid to produce a staged 3D human pose estimation method. Zhou X et al. [14] combined a 2D part regressor based on deep learning and a sparse-driven 3D reconstruction method to design a 3D human pose estimation framework. Tome D et al. [15] presented a associated method, in order to estimate the human pose, combining the image appearance-based prediction offered by the 2D landmark detector with the geometric 3D skeletal information encoded in the new pre-trained model of the 3D human pose. Chen CH et al. [16] proposed a simple method for 3D human pose estimation by performing 2D pose estimation and then matching 3D examples. Moreno-Noguer F[17] formulated the 3D human pose estimation problem as a regression between the matrix encoding the 2D and 3D joint distances, and solved the problem of 3D human pose estimation from a single image. Martinez J et al. [18] established a system combining 2d joint positions to predict 3d positions, which greatly improves the previous results from 2D to 3D pose estimation. Pavlakos G et al. [19] used two separate networks to solve the problem of 3D human pose estimation from a single color image. With the development of the research process, the ambiguity problem in the process of 2D to 3D plane projection has appeared. Even if the deep learning network structure is optimized, it cannot be solved. Therefore, research scholars have begun to study the 3D single human pose estimation based on the hybrid network model. Zhou X et al. [20] extended the most advanced 2D pose estimation sub-network through the 3D deep regression sub-network. According to the correlation between the 2D pose and the depth estimation sub-task, the network was trained end-to-end, and 3D geometric constraints are introduced to predict the 3D human pose, which can better perform human pose estimation in the wild environment. Zhou X et al. [21] proposed a framework called MonoCap, which consists of a 2D part regression based on deep learning, a sparse-driven 3D reconstruction method and 3D temporal smoothness. Pavlakos G et al. [22] proposed a solution that can train end-to-end ConvNets for 3D human pose estimation in the absence of accurate 3D ground truth by using weaker supervision signals. Yang W et al. [23] came up with an adversarial learning framework to transform the 3D human pose structure learned from a fully noted data into a field image with only 2D pose annotations, and designed a original discriminator is used to enhance the generalization ability of of the 3D pose estimator. Pavllo D et al. [24] proved that a complete convolution model using time convolution of 2D keypoints can effectively estimate the 3D poses in the video. An ordinary and resultful semi-supervised training way is also introduced. It uses unlabeled video data to predict the 2D key points of the unlabeled video, then estimate the 3D pose, in the end input it into the 2D key points for back projection. Li C et al. [25] proposed a novel method with multiple feasible hypotheses to generate 3D poses from 2D joints. Experiments have shown that the 3D pose estimated by this method from 2D joint input is consistent in 2D reprojection.

In summary, although 2D single person pose estimation has achieved great accuracy, there are problems such as inaccurate recognition when the human pose is occluded or the data set is missing. 3D single person pose estimation uses 3D human skeletons to describe the human body pose. Through the improvement of the network model, the deep-level features can be effectively obtained from the image, which has gradually become the mainstream method. This article will review the 3D single person estimation methods using neural networks in the next section. In the third section, the

verification of the 3D single person estimation method based on neural network is carried out. A unified public data set is used to compare and analyze several mainstream methods, and some methods are evaluated by comparing experimental results. Finally, summarize the work done in this article and put forward prospects for the next research work.

2. 3D Single person pose estimation method

The current research on 3D single person pose estimation using deep learning neural networks can be classified through different training methods of 3D single person pose models, which are generally summarized in three types: end-to-end network model, staged network model and hybrid network model.

2.1. End-to-end 3D human pose estimation

An end-to-end network is a type of neural network. For a task, the input to the network is the original data, and the final output is what the task expects data. In the image classification task, the original picture is input of the neural network, and the output is the given image prediction category identification. As shown in figure 1, this method is an improved end-to-end CNN method. RGB images or videos are input to the neural network, and the output is the 3D coordinate position information of all human joint points corresponding to it. The end-to-end network actually uses a network structure to process the input content, there is no other data in the process of processing. Therefore, for this type of method, data preprocessing and network structure design are particularly important.

Tekin B et al. [12] generated a 3D pose from the input image through regression. By using the mapping obtained by the auto encoder and the output of the regular CNN, the dependencies can finally be encoded more effectively. In the paper, the regression architecture is used to analyze the structure, two high-dimensional embedded mappings are introduced, and the final pose estimation is completed after further adjustment of the network.

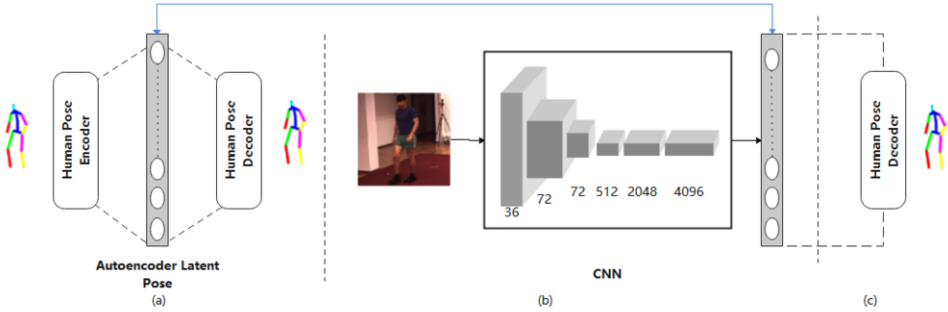

Figure 1. Structure diagram of the improved end-to-end CNN method.

Figure 1 shows a 3D human pose estimation architecture based on structured prediction. (a) It is a designed automatic encoder, in which the size of the hidden layer is larger than the size of its input and output layers. During the practice, adopt the CNN in (b) to map to the latent representation learned by the autoencoder in (a). (c) Use the decoder to map the latent representation back to the original pose space. This improved

end-to-end CNN structure combines traditional CNN networks with autoencoders for structured learning, which can resolve dependencies and improve the performance of pose estimation.

The data set applied to the human pose estimation by the end-to-end network is generally not constrained by the 3D human pose annotation data set, and therefore more complex networks cannot be realized. In order to perform more accurate and precise 3D single person pose estimation, the network structure must be able to cope with more complex and changeable scenes. 2D human pose estimation methods have become more and more accurate after recent years of development. Under such demand, researchers have begun to study the combined application of 2D human pose estimation methods and 3D human pose estimation methods.

2.2. Staged 3D human pose estimation

The staged 3D human pose estimation is a heterogeneous method proposed by researchers in order to solve the problem that the amount of labeled data encountered when using end-to-end network for 3D human pose estimation does not match the network scale. The staged method is similar to the end-to-end method in that 2D pose data is used. The difference is that the phased 3D human pose estimation method combines 2D data with 3D data. When acquiring data, a 2D human pose estimation method is needed for assistance, that is, a 2D human pose estimation model is used to obtain 2D human pose data from the original image data. Assuming that the generated 2D human pose is accurate, the 2D human pose is used as input, and the 3D human pose regression network is used to improve the dimensionality of the 2D data.

Martinez J et al. [18] designed a concise neural network as shown in Figure 2 below, which uses a linear layer to perform batch normalization in the network, and adds random dropout and ReLU activation functions twice. Input an array of 2D joint positions into system, and obtain a battery of 3D human joint positions through network operations.

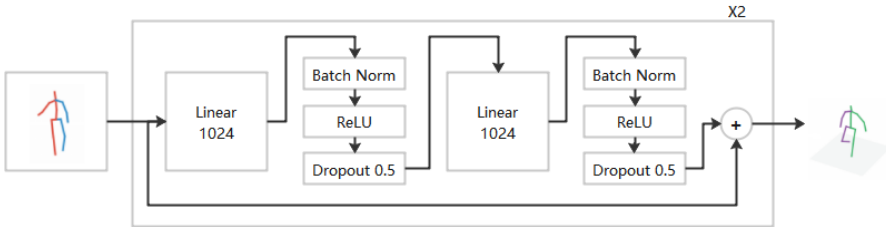

Figure 2. Improved staged neural network.

Based on the staged 3D human pose estimation, it is established on the 2D human pose estimation method. It uses 2D human pose as input to estimates the 3D human pose. Because 2D human pose can be considered as 3D human pose on a plane from a certain angle of view, there may be multiple different 3D human poses in the same projection. This is an ambiguity problem that cannot be solved by optimizing the deep learning network structure. At this time, more information needs to be introduced into the network to eliminate this ambiguity. Therefore some researchers began to research 3D human pose estimation of the hybrid network model.

2.3. 3D human pose estimation of the hybrid network model

Different from the end-to-end and staged 3D human pose estimation, the 3D human pose estimation of the hybrid network model adds additional image information and geometric constraints on the basis of pre-estimated 3D human pose. Then use these information to train a network model that promotes a 2D pose to 3D. The 3D human pose estimation of the hybrid network model uses accurate 2D human pose estimation methods while introducing more additional information including human joint points and motion characteristics. The network trained with this information can relieve the problem of 2D to a certain extent. The ambiguity problem of posture projection calculation of 3D human posture can also reduce the overfitting problem of end-to-end network.

At present, the more successful 3D human pose estimation method of hybrid network model is proposed by Pavllo D [24], which performs 3D human pose from video with convolution and semi-supervised training. A semi-supervised training means is introduced to boost the precision in the usability setting of annotated 3D ground truth pose data. The unlabeled video is combined with a ready-made 2D keypoint detector to develop the supervisory loss function with the back projection loss term. The designed encoder estimates the 3D pose through the 2D joint coordinates, and the decoder maps the 3D pose back to the 2D joint coordinates to solve the problem of automatic encoding of unlabeled data. As shown in Figure 3, the model takes a series of 2D poses that may be predicted as input, uses regression methods to obtain the 3D trajectory of the human pose, and adds soft constraints to make the average bone length of the unlabeled prediction and the average bone length of the labeled prediction match.

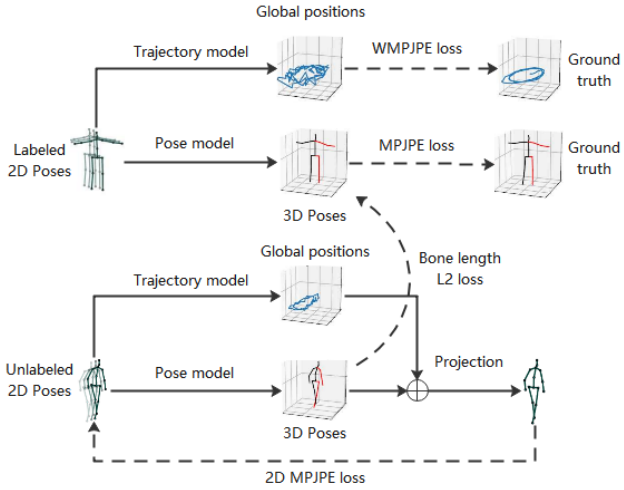

Figure 3. Semi-supervised training using 3D hybrid network model.

This chapter introduces the 3D single person pose estimation methods of three neural networks, gives the definition of different means, and illustrates merit and demerit of each estimation method through the method model in the paper. Among the three methods, using a hybrid network model to estimate the pose of a 3D single person is the best, and the second best is to use a staged method. With the development of technology, the end-to-end method will no longer be used alone. Hybrid network

models usually combine the end-to-end network with other methods to improve the accuracy of estimation. In the next chapter, we will verify the conclusions of this chapter through comparative experiments.

3. Method validation and comparative analysis

3.1. Dataset

The data set used in the human pose estimation method is usually divided into a 2D human pose estimation data set and a 3D human pose estimation data set.

Most of the data in the 2D human pose data set derived from image data online media, and these data are manually annotated. Since 3D human pose estimation is more complicated than 2D human pose estimation, the 3D human pose data set used in this article is also more complicated than the 2D human pose data set in the data acquisition and processing stage. Most of the data sets used in 2D human pose estimation are collected in natural environments, and the acquisition method of manual annotation is relatively simple. In contrast, the 3D human pose acquisition process requires a large number of cameras and sensors. Synchronous acquisition of multiple perspectives requires camera parameters. Therefore, many 3D human pose data sets are collected in laboratory environments. The final 3D human pose annotation data sets are far lower than the 2D human pose data sets in terms of diversity and quantity.

At present, common 3D human pose annotation data sets include HumanEva[26], Human3.6M[27], CMU Panoptic dataset[28], MPI-INF-3DHP[29], etc[30]. The following is a detailed introduction to these commonly used annotation data sets, and summarized in Table 1.

HumanEva data set[26]. The HumanEva data set [26] consists of two sub-data sets, HumanEva-I and HumanEva-II. The two data sets were captured using ViconPeak's commercial MoCap system and annotated with ground truth. The HumanEva-I data set has a total of 13.5GB, which contains 7 verified view video sequences synchronized with 3D human poses, including 4 grayscales and 3 colors. These 3D human posture marker data are collected through a motion capture system. In the 3m x 2m capture area, the bodies of four objects are tagged and six common movements such as walking, gesturing, throwing and catching, boxing and combination are performed. The HumanEva-II data set has a total of 4.54GB and contains only 2 objects. Its ground truth motion capture data is also obtained using ViconPeak's system.

Human3.6M data set[27]. The collection of Human3.6M data set [27] is carried out in the laboratory. The collection system is a precise MoCap system based on tags. There are 11 occupational actors wearing well-fitting clothes in the laboratory, including 5 women and 6 men. Collect 17 activities, including posing, discussing, smoking, taking pictures, making phone calls, etc. It has a total of approximately 127GB, containing 3.6 million 3D human poses, each of which corresponds to 4 images with diverse perspectives. The main capture equipment includes 4 digital cameras, 1 time-of-flight sensor, and 10 synchronized motion cameras. The capture area is approximately 4m x 3m. Figure 4 below is a schematic diagram of the motion capture system used by Human3.6M. In order to evaluate, it is divided into three protocols according to diverse training and test data: protocol 1, protocol 2, and protocol 3. There are two commonly used division criteria. The first is to use 1, 5, 6, 7, 8 as the training set, and 9, 11 as the test set. The second is to use 1, 5, 6, 7, 8, 9 as the

training set, and 11 as the test set. In this article, we mainly use the first protocol in Human3.6M to conduct comparative experiments on each model.

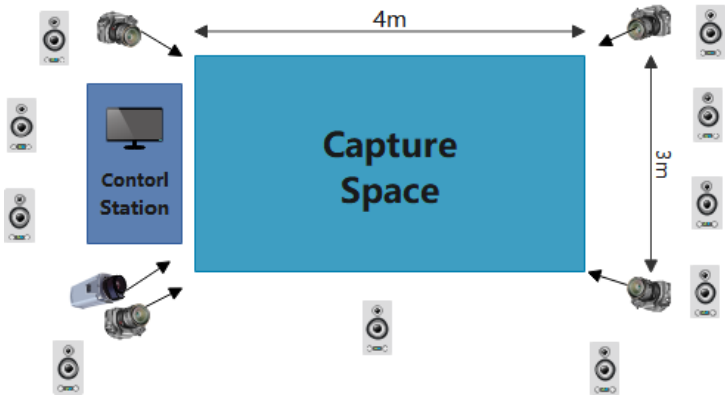

Figure 4. Human3.6M motion capture system schematic diagram.

CMU Panoptic dataset[28]. The CMU Panoptic dataset [28] is produced by CMU and is captured using a multi-view system with unmarked motion capture. The system has 480 VGA camera views, 31 HD views, 10 RGBD sensors and hardware synchronization system. The annotations include 3D key points, cloud points, optical flow, etc. The total size is about 41GB.

MPI-INF-3DHP data set[29]. The MPI-INF-3DHP dataset [29] was collected using an unmarked multi-camera MoCap system in indoor and outdoor sites. It contains 1.3 million frames from 14 different views. Eight subjects (4 women and 4 men) were noted to perform 8 events (such as walking, standing, exercising, sitting, squatting, stretching, exercising, etc.).

Table 1. Summary of commonly used data sets

Data Set Name	Scale	Collection Scene	Number Of People Collected	Number Of Samples	Capture Method
HumanEva[26]	18G	Indoor	4	≈8	Motion capture
Human3.6M[27]	127GB	Indoor	11	360	Motion capture
CMU Panoptic dataset[28]	41GB	Indoor	Multiplayer	unknown	Motion capture
MPI-INF-3DHP[29]	unknown	Indoor and Outdoor	8	＞130	Motion capture, Image synthesis

3.2. Evaluation index

Human3.6M dataset is currently the most widely used 3D human pose annotation dataset. There are three evaluation methods mentioned in the paper, namely MPJPE, MPJAE and MPJLE.

MPJPE is the mean per joint position error. It is currently the most widely used measure for performance evaluation of 3D attitude estimation. It calculates the Euclide distance (in millimeters) from the estimated 3D joint to the real condition of the ground, and then compares the results with the real value. The position of each joint point is represented in a 3D coordinate system. If it is used for 2D data, the unit used is pixel.

MPJAE is the mean per joint angle error. The function returns the joint angle instead of the joint position. It is also an MPJPE in nature, but some too small joint position errors are filtered and the errors are normalized.

MPJLE is the mean per joint localization error. MPJPE and MPJAE have two shortcomings. The first problem is that they do not have sufficient robust stability. Mispredicted joints will affect the error of the entire data set. The second disadvantage is that the final result may overemphasize errors that are hard for humans to detect. MPJLE is proposed to solve some of these problems.

The error evaluation method used in this article is the most commonly used MPJPE in 3D pose estimation. The calculation formula is as follows:

$$E_{MPJPE}(f,S)=\frac{1}{N_S}\sum_{i=1}^{N_S}\left\|m_{F,S}^{(f)}(i)-m_{gt,S}^{(f)}(i)\right\|_2 \tag{1}$$

Where f represents the frame, S represents the skeleton, N_S represents the quantity of nodes in the skeleton S, F represents the estimation result, gt represents the true value, and m represents a function, the input is the node number i, and the output is the 3D coordinate position of the node i, $\|*\|_2$ which represents 2 norm.

3.3. Comparative Test

Aiming at the three different human pose estimation methods listed in the previous chapter, this section uses the python programming language with the deep learning open source framework Tensorflow. Corresponding experimental environment was built on the Ubuntu system, and several models were verified and compared. The Human3.6M data set is used as the verification set for model verification. The verification set contains a total of 15 different human daily activity scenes of 11 different people. When dividing the training set and the verification set, consider the generalization ability of the human pose difference for the training model. It is not possible to put all 11 people in the training set, so division rules are formed for the training set and the verification set. This article uses protocol 1, which uses S1, S5, S6, S7, and S8 of the 17 joint bones as the training set and S9 and S11 as the test set. The evaluation method used in quantitative analysis is the average joint point position error MPJPE. After aligning the depth of the root joint (usually the pelvic joint), use the trained model to perform MPJPE calculation on the scene in the verification set, and then compare with the current method using the same data set, and finally summarize the experimental contrast results, as shown in the following Table 2 below.

Table 2. Estimated and verified results of each network model on the Human3.6M test set (MPJPE(mm))

Method	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Zhou X[11]	91.83	102.41	96.95	98.72	113.35	125.22	90.04	93.84
Tome D[15]	64.98	73.47	76.82	86.43	86.28	110.67	68.93	74.49
Chen C[16]	71.63	66.60	74.74	79.09	70.05	93.26	67.56	89.30
Moreno[17]	67.48	79.01	76.48	83.12	97.43	100.37	74.58	71.96
Martinez[18]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Zhou X[20]	54.82	60.70	58.22	71.41	62.03	65.53	53.83	55.58
Pavlakos[22]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9
Li C[25]	**43.8**	48.6	49.1	49.8	57.6	61.5	45.9	48.3
Pavllo[24]	45.2	**46.7**	**43.3**	**45.6**	**48.1**	**55.1**	**44.6**	**44.3**

Method	Sitting	Sitting Down	Smoking	Waiting	Walk Dog	Walking	Walk Together	Average
Zhou X[11]	132.16	158.97	106.97	94.41	126.04	79.02	98.96	107.26
Tome D[15]	110.19	173.91	84.95	85.78	86.26	71.36	73.14	88.39
Chen C[16]	90.74	195.62	83.46	71.15	85.86	55.74	62.51	82.72
Moreno[17]	102.40	116.68	87.70	94.57	82.72	75.21	74.92	85.64
Martinez[18]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhou X[20]	75.20	111.59	64.15	66.05	51.43	63.22	55.33	64.90
Pavlakos[22]	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Li C[25]	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
Pavllo[24]	**57.3**	**65.8**	**47.1**	**44.0**	**49.0**	**32.8**	**33.9**	**46.8**

According to Table 2, it is concluded that for the data obtained by experiments of the models in some papers in the Human3.6M data set, the phased 3D human pose estimation method is better than the end-to-end human pose estimation method. For the average MPJPE of all scenarios in the test set, the average MPJPE of Martinez et al. [18] is about 40mm lower than that of Zhou X et al. [11]. Moreover, the 3D human pose estimation method of the hybrid network model is also better than staged 3D human pose estimation method. In the chart, the average MPJPE of the method of Pavllo et al. [24] is 16.1 mm lower than that of the method of Martinez et al. [18].

In summary, in the 3D single person pose estimation method, the constantly evolving neural network model is combined with the data information of the 2D human pose joints and the geometric structural characteristics of the human pose structure itself to optimize the method performance of the model. This kind of mixed and matched network model greatly improves the accuracy of human pose estimation, and at the same time increases the possibility of application in more different complex scenes.

4. Summary

This article reviews the development of 3D single person pose estimation methods using neural networks, describes in detail the important method models, and compares the better theoretical models in previous studies, and gives the comparative experimental results. Make a summary of it.

The existing 3D single person pose estimation methods based on neural networks are relatively mature, but the accuracy of pose estimation in complex scenes needs to be improved. The current 3D single person pose estimation method based on neural networks cannot achieve satisfactory results in the case of multiple poses, which points out specific directions for future research work. In addition, a plenty of commonly used data sets are collected indoors, and there are too few types of human postures. Therefore, the way of generating data will be an immense challenge.

Acknowledgments

This work has been supported by the Natural science foundation of Ningxia province (No. 2020AAC03 218), and the North Minzu University key research project (No. 2019KJ26) and the National Innovation and Entrepreneurship Project under Grant (No. S2020-11407-010G).

References

[1] Wang K, Lin L, Jiang C, et al. 3D human pose machines with self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1069-1082.

[2] Véges M, Varga V, Lőrincz A. 3d human pose estimation with siamese equivariant embedding[J]. Neurocomputing, 2019, 339: 194-201.

[3] Cheng Y, Yang B, Wang B, et al. Occlusion-aware networks for 3d human pose estimation in video[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 723-732.

[4] Hwang D H, Kim S, Monet N, et al. Lightweight 3D Human Pose Estimation Network Training Using Teacher-Student Learning[C]//The IEEE Winter Conference on Applications of Computer Vision. 2020: 479-488.

[5] Huang F, Zeng A, Liu M, et al. DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image[C]//The IEEE Winter Conference on Applications of Computer Vision. 2020: 429-438.

[6] Mehta D, Sotnychenko O, Mueller F, et al. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera[J]. arXiv preprint arXiv:1907.00837, 2019.

[7] Soonchan Park,Sang-baek Lee,Jinah Park. Data augmentation method for improving the accuracy of human pose estimation with cropped images[J]. Pattern Recognition Letters,2020,136.

[8] Chen Y, Shen C, Wei X S, et al. Adversarial posenet: A structure-aware convolutional network for human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1212-1221.

[9] Chou C J, Chien J T, Chen H T. Self adversarial training for human pose estimation[C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018: 17-30.

[10] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 5693-5703.

[11] Zhou X, Sun X, Zhang W, et al. Deep kinematic pose regression[C]//European Conference on Computer Vision. Springer, Cham, 2016: 186-201.

[12] Tekin B, Katircioglu I, Salzmann M, et al. Structured prediction of 3d human pose with deep neural networks[J]. arXiv preprint arXiv:1605.05180, 2016.

[13] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[14] Zhou X, Zhu M, Leonardos S, et al. Sparseness meets deepness: 3d human pose estimation from monocular video[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4966-4975.

[15] Tome D, Russell C, Agapito L. Lifting from the deep: Convolutional 3d pose estimation from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2500-2509.

[16] Chen C H, Ramanan D. 3d human pose estimation= 2d pose estimation+ matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7035-7043.

[17] Moreno-Noguer F. 3d human pose estimation from a single image via distance matrix regression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2823-2832.

[18] Martinez J, Hossain R, Romero J, et al. A simple yet effective baseline for 3d human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2640-2649.

[19] Pavlakos G, Zhou X, Derpanis K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7025-7034.

[20] Zhou X, Huang Q, Sun X, et al. Towards 3d human pose estimation in the wild: a weakly-supervised approach[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 398-407.

[21] Zhou X, Zhu M, Pavlakos G, et al. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(4): 901-914.

[22] Pavlakos G, Zhou X, Daniilidis K. Ordinal depth supervision for 3d human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7307-7316.

[23] Yang W, Ouyang W, Wang X, et al. 3d human pose estimation in the wild by adversarial learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5255-5264.
[24] Pavllo D, Feichtenhofer C, Grangier D, et al. 3d human pose estimation in video with temporal convolutions and semi-supervised training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7753-7762.
[25] Li C, Lee G H. Generating multiple hypotheses for 3d human pose estimation with mixture density network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9887-9895.
[26] Sigal L, Balan A O, Black M J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International journal of computer vision, 2010, 87(1-2): 4.
[27] Ionescu C, Papava D, Olaru V, et al. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(7): 1325-1339.
[28] Joo H, Liu H, Tan L, et al. Panoptic studio: A massively multiview system for social motion capture[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3334-3342.
[29] Mehta D, Rhodin H, Casas D, et al. Monocular 3d human pose estimation in the wild using improved cnn supervision[C]//2017 international conference on 3D vision (3DV). IEEE, 2017: 506-516.
[30] Chen Y, Tian Y, He M. Monocular human pose estimation: A survey of deep learning-based methods[J]. Computer Vision and Image Understanding, 2020, 192: 102897.