International Journal of Information Technology and Applications
Volume 1, Issue 3, September 2024 ISSN 3007-7397 © P-Learn Press

A Multiview Video Dataset for Action and Pose Analysis

Haoyu Li, Longfei Zhang*, Zehong Chen, Yufeng Wu, Gangyi Ding

Key Laboratory of Digital Performance and Simulation Technology, Beijing Institute of Technology, Beijing, China.
longfeizhang@bit.edu.cn

Abstract. The purpose of this study is to construct a dataset of human pose. In computer vision (CV), human pose estimation (HPE) is an important direction. After getting pose information, we can better analyze human motion and action. High-quality datasets are necessary during the work to train the HPE models. Nowadays, many kinds of datasets have been developed to meet these needs, and based on them, many methods have achieved excellent performance. However, pose estimation is still tricky for some unique and complex actions or motions, such as dance and sports actions. Aiming to solve this problem, we propose constructing a multiview-based pose dataset to analyze human action. In this work, we use filtering as an optimization approach to improve the accuracy of results. By filtering data, those wrongly recognized vital points can be corrected to some extent, and the missing key points or poses can be estimated based on the continuity of motion. The filter mothed can also improve the smoothness of the curve, so after filtering, the pose sequences can be more fluent and then conform to the kinetics. Sometimes, some errors in the 2D pose may result in images from some views caused by occlusions, environmental changes, and other reasons. In this situation, choosing 2D pose results from appropriate views can further improve the accuracy of the final results. Besides, for the environment where there are several persons, we add the work of matching humans across different views, so we finish estimating poses of multi-person. From the experiments, the feasibility and reusability of the proposed method are demonstrated. With this method, the accuracy of results can be maintained while the work is simplified, which can help improve the efficiency of related research.

Keywords: Pose Estimation, Action Analysis.

1. Introduction

Human pose data is essential to understanding motion or action information. In the CV field, 3D pose estimation is becoming an important technology. Based on the results, many works of other applications, e.g., 3D human modeling, human-computer interaction, behavior recognition, and motion learning, can be better finished. Skeleton data is a widely used human model to present pose information. After extracting a 3D human skeleton, with information on each human key point, the primary temporal and spatial motion information can be learned well. Getting accurate 3D HPE results is an essential step to finishing the work. Usually, the related data is obtained by complex motion capture devices, which are challenging to deploy and have very high costs. In this paper, we find a more straightforward method. Combining deep learning, CV technologies, and geometry knowledge, we can get high-quality 3D human spatial information with RGB images.

The primary process of 3D pose estimation involves inputting a single RGB image or video (monocular or muti-view) and constructing 3D skeletons based on them.

3D HPE based on monocular images or videos has been an emerging direction in the CV field. This kind of approach mainly includes the operations of predicting the depths of critical points from the 2D pose results [1, 2] or regressing the 3D pose with the adjacent frames in the video (or image sequences) [3, 4]. However, the results obtained by this method are not always accurate enough because of the missing spatial information. Another 3D HPE method based on multiview images or videos is used to solve the problem. With the parameters of each view, we can get extra spatial information to get more accurate 3D results.

Many pose estimation datasets [9, 10, 11, 12, 13, 14, 15, 16] are used for training, testing, and evaluating the HPE baseline, and many works based on them receive SOTA results. 2D pose datasets such as MSCOCO [14] and MPII [16] have been widely used. However, compared to the dataset of 2D pose, there are still some shortcuts to 3D pose datasets: (1) The number is limited, and getting enough data to train and test new baseline is still hard for researchers; (2) The lack of complexity and diversity. Most available datasets mainly contain simple daily behaviors but few professional motions. (3) Getting 3D pose data requires complex devices and strict environments, which restricts the operation of constructing datasets.

This paper introduces a method of constructing a 3D pose dataset in a multiview environment with pose estimation and reconstruction approaches. The main steps are: (1) Calibrating cameras and getting their intrinsic and extrinsic parameters; (2) Getting the 3D pose ground truth of the human skeleton from a motion capture system and obtaining images of each camera at the same time; (3) Finishing the work of 2D pose estimation of each view; (4) Constructing 3D human pose with parameters of each camera.

With the results of the 3D pose, we use a filtering operation to optimize keypoint data and compare them with ground truth. Moreover, for the dataset that includes more than one person, we finish the work of matching persons in different views and estimating the poses of multi-persons.

The main contributions of this paper can be summarized as follows:

We propose a baseline for constructing a 3D pose estimation dataset based on multiview images. Compared to most existing approaches to dataset construction, the method simplifies the steps while maintaining the accuracy of 3D pose results.

The filtering operation smooths the results, reducing the jitters and making the curves of the critical points more fluent.

By predicting the depths of root keypoints, we finish the work of matching the persons in different views. This method makes full use of image information, including 2D pose data, human detection results, and camera parameters. Also, it provides a new strategy for estimating multi-person poses.

This paper is organized as follows: Section 2 reviews the related works. Section 3 details the methods. Section 4 presents the experiment's results and analysis. Section 5 summarizes the work and looks forward to future work.

2. Related Works

In this chapter, we briefly summarize the classic work on pose estimation and related techniques. We combine these methods to build a new baseline to finish the dataset construction work.

2.1.Pose Estimation

Many 2D pose estimation methods [17, 18, 19, 20] have been used to locate the 2D coordinates of human critical points in images and videos. Based on the results of the 2D pose, the 3D pose is further reconstructed. This paper uses the top-down approach AlphaPose [18] to get the 2D pose. YoloV4 [21] provides the human detection results, and each person's pose is further estimated.

As mentioned above, many 3D pose estimation methods based on monocular images or videos are now available [1, 2, 3, 4]. However, these methods are limited for some complicated scenes: (1) The occlusions of different sizes will cause the error in 3D pose results; (2) Under some views, different 3D poses may have the same 2D pose projection results. Therefore, the uniqueness of the results of the 3D pose cannot be ensured. 3D pose estimation methods based on multiview images can avoid these problems. After getting the 2D poses from each view, we can finish the 3D human skeleton reconstruction by combining the 2D key points with the parameters of each camera.

2.2. Datasets

There have been lots of datasets used for pose estimation or motion recognition. For example, Human 3.6M [9] covers several everyday scenarios from 11 professional actors. LSP (Leeds Sports Pose) [11] dataset contains 2000 images of sportspersons to estimate sports poses. In [10], based on videos, the FineGym dataset is built to understand and analyze gymnastic actions. Moreover, there are also some datasets for pose estimation of multi-persons, 2D [14, 15] and 3D [12, 13].

2.3. Object Tracking

Object tracking is an essential topic in the CV field; in videos or image sequences, the trajectory and position of objects are recognized. This paper mainly analyzes the interesting motion data, so we must focus on the scene's unique human. We combined the object tracking module with pose estimation during the experiment to locate the motion data.

2.4. Motion Capture

In order to evaluate the method's accuracy, we need to compare the 3D pose result with ground truth. This paper uses 24 motion capture cameras to get the human surface and skeleton data (see Figure 1). Information about the human surface is constructed from the mark points on the human, and the skeleton data is calculated.

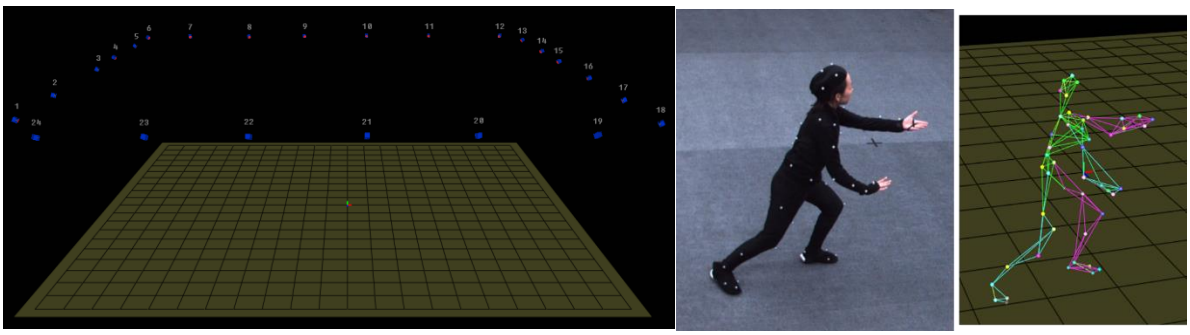

Figure 1. The layout of motion capture system (left) and the construction of human surface (right).

2.5. Data optimization

In the experiment, there will be some errors in the camera calibration and pose estimation results, which will cause noise and jitter and influence the accuracy of the final result of 3D Pose estimation. Therefore, we choose the filter method to process the pose data to make it smoother and conform to the kinematics. In addition, the critical points detected wrongly can be corrected in this way. Moreover, the missed pose data can also be filled for the frames where the pose is not estimated, or the humans are not detected.

We combine the pose estimation approaches in different dimensions based on the existing methods to construct a 3D pose dataset. Moreover, by adding object tracking and filtering operations, we build connections within the pose sequences, improving accuracy and motion fluency.

3. Materials and Methods

Figure 2 represents the framework of this paper. It includes traditional 2D and 3D pose estimation methods, as well as data optimization.

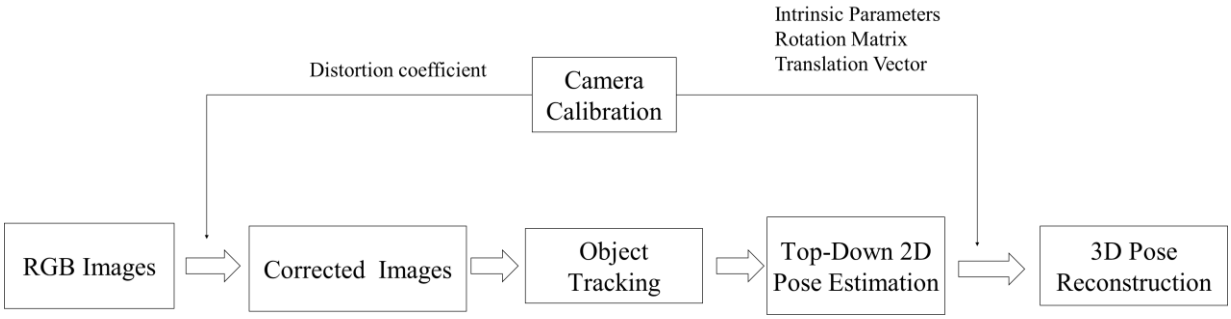

Figure 2. The framework of pose data collection.

3.1. Camera Calibration

After getting the result of the 2D pose, it's necessary to exploit the parameters of each RGB camera when reconstructing 3D key points, so first, we need to calibrate the cameras in the scene. In the experiment, the size of space is 12m×12m×3m, and the layout of the devices is shown in Figure 3. We utilize the method in [23] to finish calibration. We collect images containing chessboard and calculate each camera's parameters. The resolution of the images is 1936×1216.

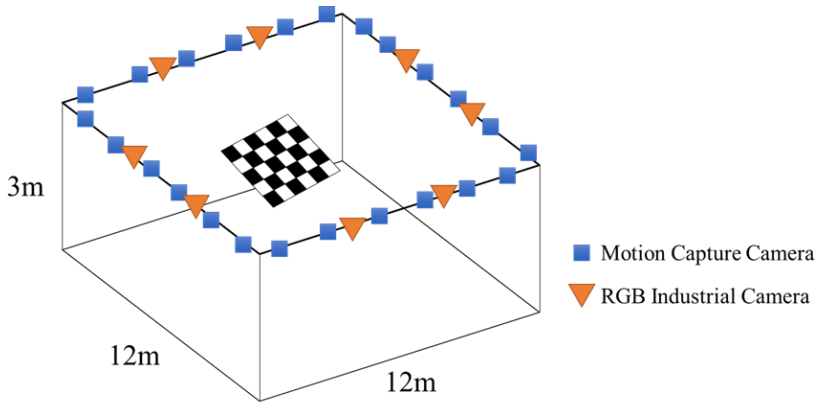

Figure 3. The layout of the system which contains 24 motion capture cameras to get the ground truth of pose data, and 8 RGB industrial cameras. The chessboard is used for RGB cameras calibration.

For each camera, we firstly get its intrinsic matrix

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

and the distortion coefficient $D = [k_1, k_2, k_3, p_1, p_2]$. Then we jointly calibrate all RGB cameras to get their rotation matrix $R_{3\times3}$, as well as the translation vector $T = [t_x, t_y, t_z]$. After getting these data, we use them to finish image correction and reconstruction of 3D key points with 2D pose from each view.

3.2. Getting the Results of 2D Pose Estimation

The method is based on 2D pose estimation, so we concentrate on the performance of 2D pose estimation approach. Meanwhile, we focus on the motion from the special person, which we need to locate. Therefore, the top-down method is helpful to achieve this target. In this paper, we choose AlphaPose [18], which exploits the Regional Multi-Person Pose Estimation (RMPE). This method consists of Symmetric Spatial Transformer Network (SSTN), Parametric Pose Non-Maximum-Suppression (NMS) and Pose-Guided Proposals Generator (PGPG). With these three components, we can get more accurate results of object detection and 2D estimation.

Moreover, human motion data has the attribute of continuity. In the sequences, the poses in each frame can be related to the ones in adjacent frames. During the process, we pay attention to the position changes of the whole human and key points. Therefore, we utilize an object tracking module to improve the accuracy of human detection and pose estimation and the performance of motion recognition.

3.3. 3D Pose Reconstruction

After obtaining 2D poses sequences from each view, we choose the human of interest based on the tracking result. With each camera's parameters, we finish 3D reconstruction as shown in Figure 4.

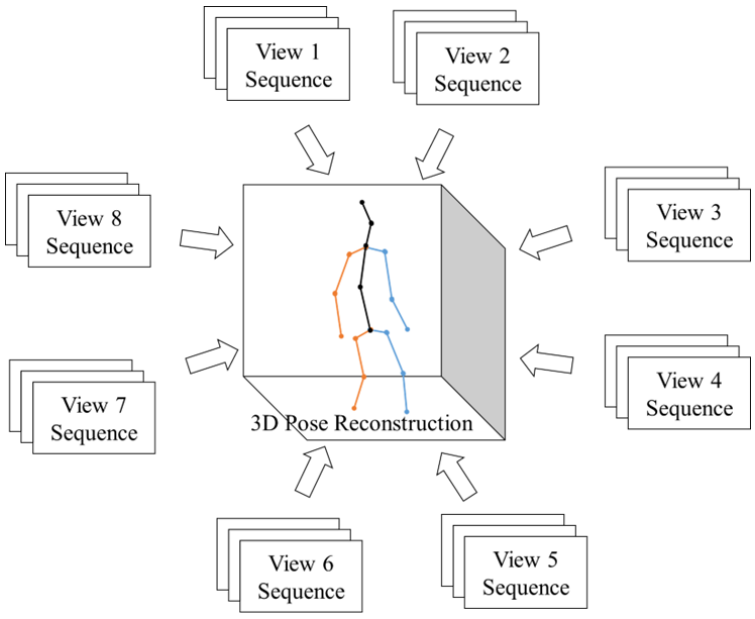

Figure 4. 3D reconstruction base on muti-view

The 2D coordinates of 3D point Q in each view are (x_1, y_1), (x_2, y_2), … (x_i, y_i), … (x_n, y_n), and the parameters of n cameras are K_1, K_2, … K_n, R_1, R_2, … R_n, T_1, T_2, … T_n, respectively. In theory, the 3D coordinate of point Q satisfy the equation $AX = 0$, where:

$$A = \begin{bmatrix} x_1 P_{1(3)} - P_{1(1)} \\ y_1 P_{1(3)} - P_{1(2)} \\ \vdots \\ x_i P_{i(3)} - P_{i(1)} \\ y_i P_{i(3)} - P_{i(1)} \\ \vdots \\ x_n P_{n(3)} - P_{n(1)} \\ y_n P_{n(3)} - P_{n(2)} \end{bmatrix}, n \geq 2 \tag{2}$$

And Pi is the projection matrix of the i-th camera, where:

$$P_i = K_i[R_i \quad T_i] = [P_{i(1)} \quad P_{i(2)} \quad P_{i(3)}]^T \tag{3}$$

In practice, because of errors, we need to use least square method to find the optimal solution of the equation, that is, calculate the eigenvector of A^TA. After getting the result vector, we normalize it by $X' = \tilde{X}/\tilde{X}[4]$, and see $(X'[1], X'[2], X'[3])$ as the coordinate of point Q.

3.4. Smooth Filtering

Due to false detection or missing detection in the pose sequence, the 2D key points in the time domain will appear jitter and deviation. In order to make the 3D pose more accurate and respect the kinematic constraint, we optimize the 2D and 3D data by smooth filtering and compare the results. Some simple actions have regular changes in practice, while most professional motions are nonlinear aperiodic. According to these attributes, we choose the Savitzky-Golay filter [24]. This flexible and efficient method fits successive sub-sets of adjacent data points with a low-degree polynomial by linear least squares. This filter can choose the window length and polynomial degree to meet different smoothness needs while keeping the data tendency.

For a set of data $\{x[i] | i=-m, \ldots, -1, 0, 1, \ldots, m\}$, an nth order polynomial is constructed to fit it:

$$f(i) = \sum_{k=0}^{n} b_{n0} + b_{n1} i + b_{n2} i^2 + \cdots + b_{nn} i^n \tag{4}$$

the result of residual sum of squares is:

$$E = \sum_{i=-m}^{m} (f(i) - x[i])^2 = \sum_{i=-m}^{m} \left(\sum_{k=0}^{n} b_{nk} i^k - x[i] \right)^2 \tag{5}$$

to get the smallest result, the partial derivatives of E to each b_{nr} $(r=0, 1, \ldots, n)$ are all 0:

$$\frac{\partial E}{\partial b_{nr}} = 2 \sum_{i=-m}^{m} \left(\sum_{k=0}^{n} b_{nk} i^k - x[i] \right) i^r = 0 \tag{6}$$

$$\sum_{k=0}^{n} h_{nk} \sum_{i=-m}^{m} i^{k+r} = \sum_{r=-m}^{m} x[i] i^{r} \tag{7}$$

when given the data set x, the window length $2m+1$ and the polynomial order n, we can calculate each coefficient b_{nr} and then get the polynomial function $f(i)$.

3.5. Human Matching

In [1], a method of multi-person pose estimation in monocular 3D pose is proposed. Base on the result of human detection, the scale of image and the focal length of camera, the depth of each person's root key point is predicted by RootNet, and then the 3D pose results of all persons are estimated (see Figure 5). The results are still accurate enough to express the 3D poses, but we can use the depth prediction results to finish the work of matching human cross views.

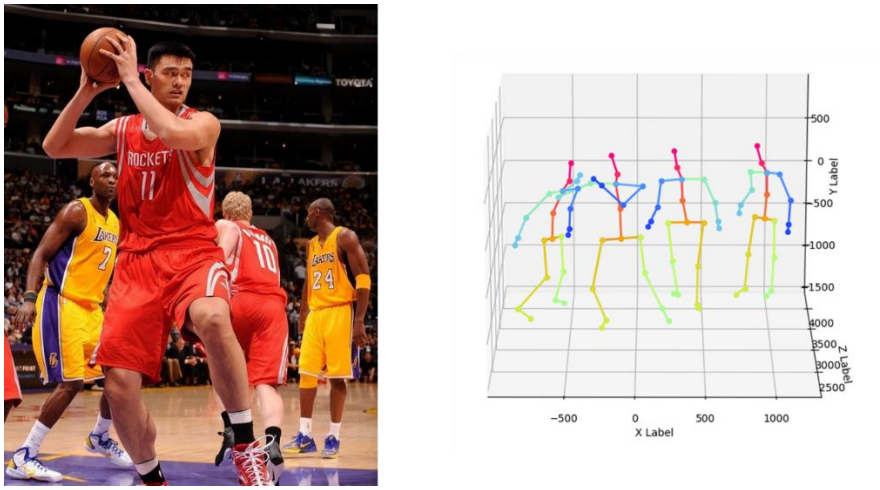

Figure 5. Using depth prediction to estimate 3D poses of multi persons in monocular image.

The thinking of matching human is shown in Figure 6. And the specific algorithm is as follows:

Algorithm 1. Matching persons cross different views.

Input:
Images from different views $img_0, img_1, \dots img_n$;
The 3D positions of root keypoint set rootKps={}
Process:
for $i = 0,1, \dots, n$ *do*
  Get the human detection results of img_i $boxes = \{box_0, box_1, \dots, box_m\}$
  for $j = 0,1, \dots, m$ *do*
    Predict the depth of root keypoint rkp_{ij} *in* box_j
    Calculate the 3D position of rkp_{ij} *in camera coordinate system* $rkp_{ij(cam)}$
    Calculate the 3D position of rkp_{ij} *in world coordinate system* $rkp_{ij(world)}$
    $rootKps.add(rkp_{ij(world)})$
  end for
end for
Match the human according to the distances between all the points in set rootKps
Output: The matching results

With the human detection results from Darknet and the intrinsic parameters of the cameras to which the

images correspond, we can get the depths of root key points. For a root keypoint rkp, with its 2D coordinate (x, y) and depth Z, we can calculate its position in the camera coordinate system (xcam, ycam, zcam) based on the intrinsic parameter.

$$\begin{bmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_{world} \\ Y_{world} \\ Z_{world} \\ 1 \end{bmatrix} \tag{8}$$

We can further calculate its position in world coordinate system (x_{world}, y_{world}, z_{world}) based on the rotation matrix R and translation vector T. According to this information, we can finally match the human in different views.

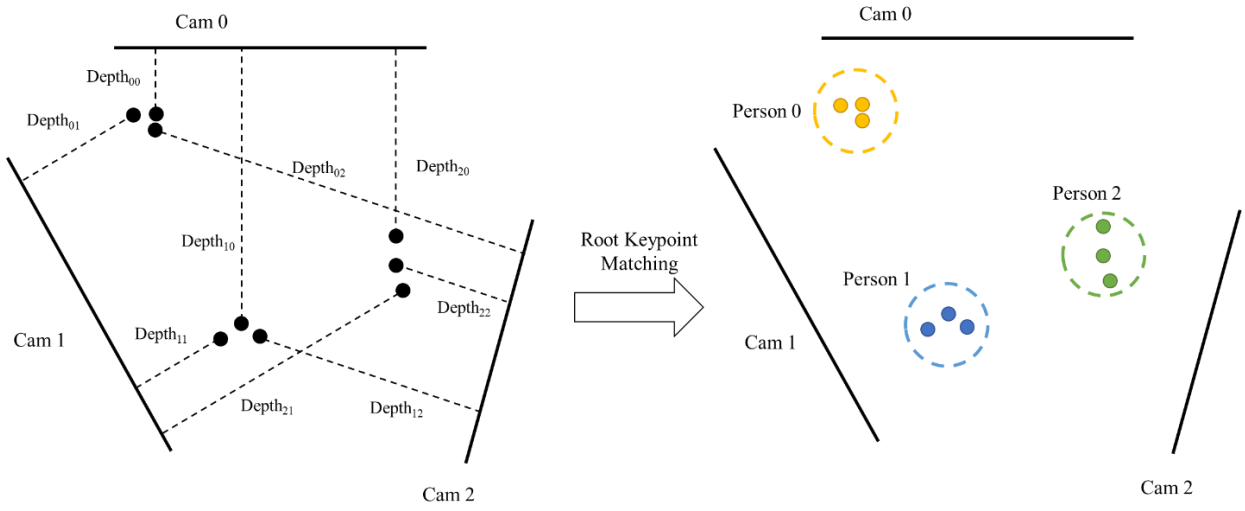

Figure 6. Matching persons in different views from the depth of root key points. $Depth_{ij}$ means depth of the ith person's root key point under camera j.

4. Results and Discussion

4.1. Experiment Data Description

In this paper, we first exploit our method on the Human 3.6M dataset (See Figure 7). Moreover, we also construct a sports dataset to research professional poses and motions (See Figure 8). The dataset includes the motions in the sports items Baduanjin and Tai Chi, and the reasons we select them are: (1) They are professional and complex enough to demonstrate the performance of our method; (2) These two items are comprehensive for they both consist of several sub-motions which can involve different parts of human to ensure the coverage of the whole body. After calibrating the RGB cameras and getting their parameters, we simultaneously collect the surface key points, the calculated skeleton data of humans, and the RGB pictures from each view. Then, we finish the work of 2D pose estimation and 3D reconstruction of the skeleton. We compare the original results with those filtered to determine the effect of the filtering operation on improving the data. In addition, we also compared the 3D results under different numbers of views; that is, we selected 2D poses from some views to calculate the 3D skeleton and see the influence of view number on results.

Furthermore, we also get the 3D pose of the Campus dataset under different approaches to see the results in a multi-person environment (see Figure 9). Based on tracking, we get the multi-person 2D poses in different

views and match them to reconstruct 3D poses. Moreover, with VoxelPose [6], we also get the multi-dimensional poses in the environment.

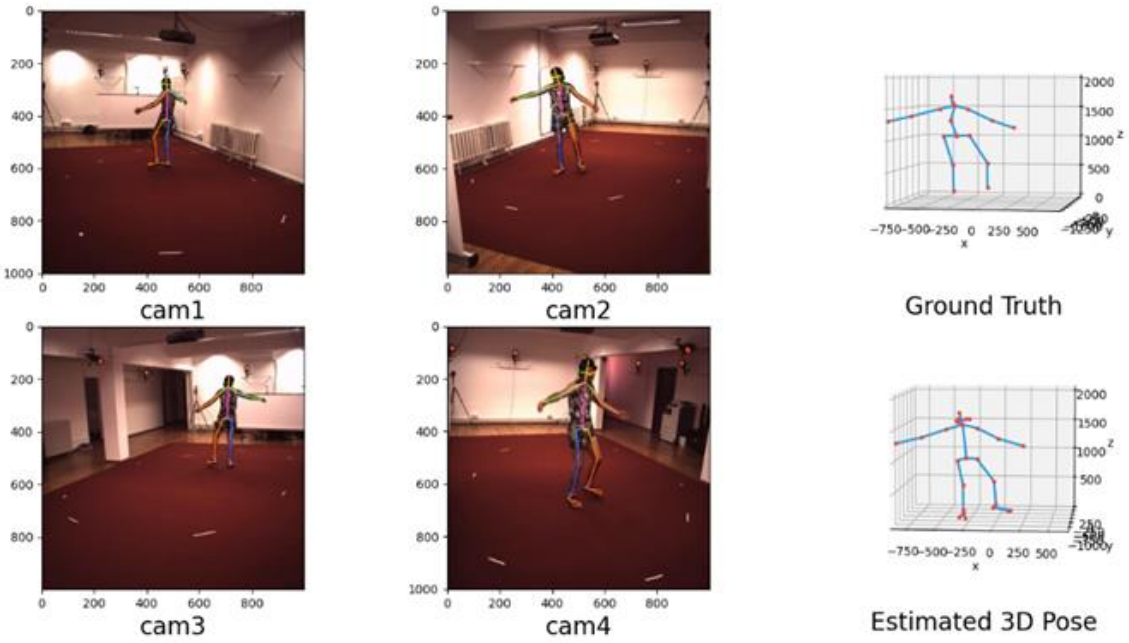

Figure 7. The results 3D pose estimation on Human 3.6M.

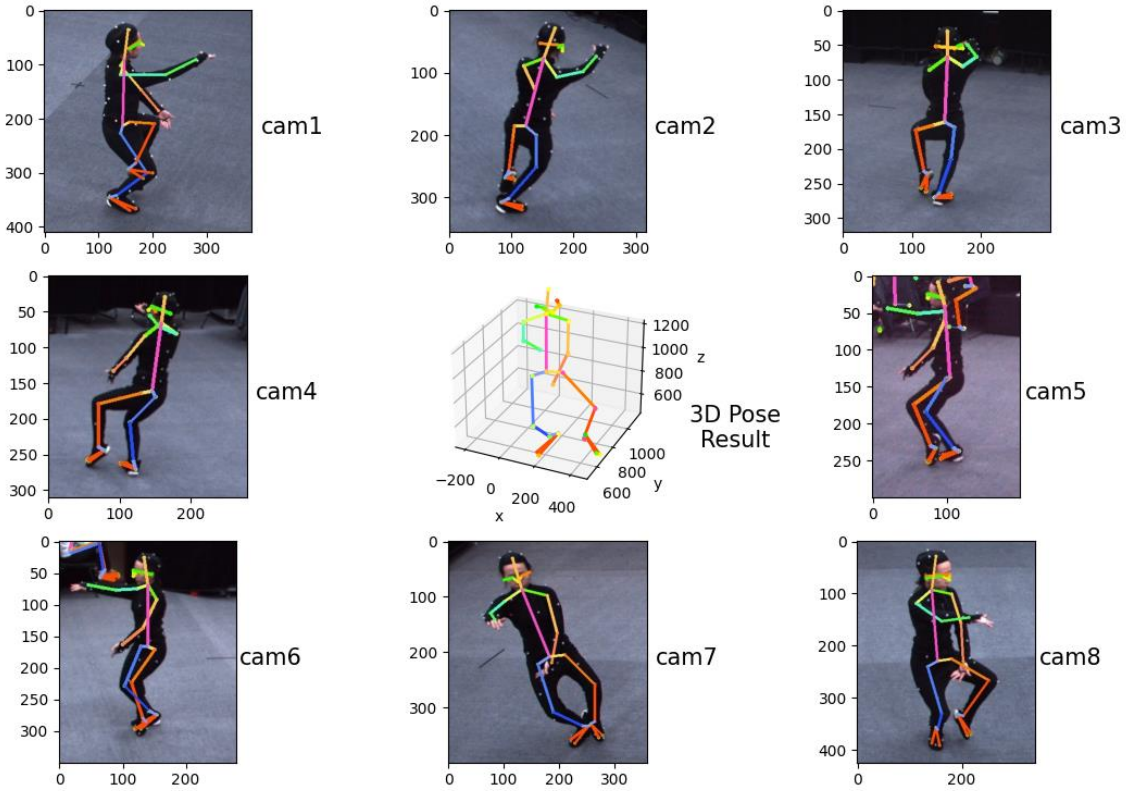

Figure 8. The 3D pose results of self-construct dataset.

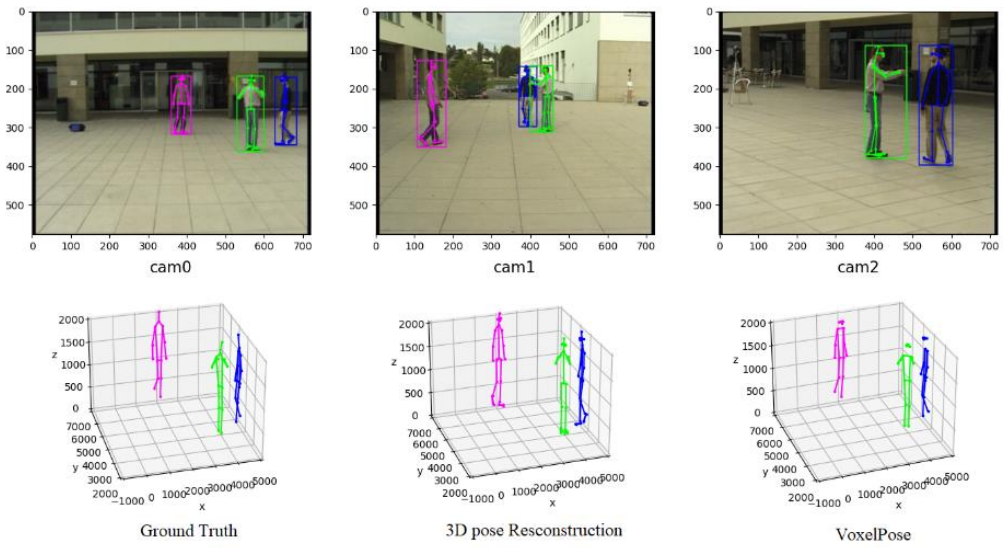

Figure 9. Multi-person 3D pose results on Campus.

Finally, for the results of all the methods, we filter the keypoints to optimize data and see the accuracy of every human respectively. The performance of filter can be seen in Figure 10 (Human 3.6M, the human skeleton contains 17 keypoints), which shows the errors of each keypoint.

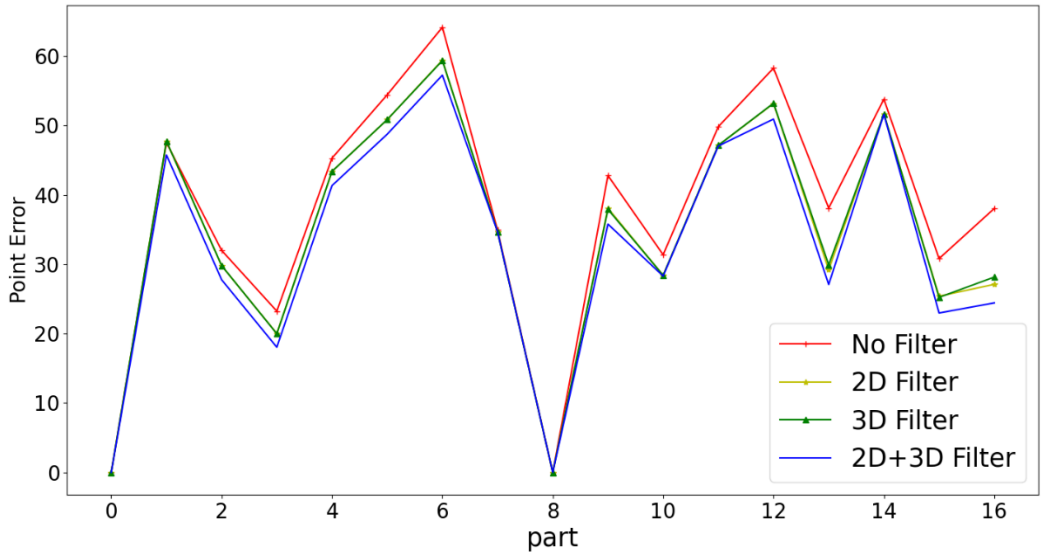

Figure 10. The performance of filtering operation in different dimensions.

4.2. Experiment Results Analysis

4.2.1. Ablation Study

For the purpose of demonstrating the effect of filtering, we process the poses data in 2D and 3D space, and compare the accuracy of filtered and original results. In order to see the relationship between the number of views and results, we select 3D reconstruction result from different numbers of views respectively and compare their difference from ground truth.

4.2.2. Evaluation

The results of Human 3.6M are shown in Table 1. When calibrate the RGB cameras, we move the chessboard so the RGB cameras and motion capture cameras are related to different coordinate system, so for the results of self-build dataset, we need to align their data before comparing them. In our experiment, we normalize the root keypoint of skeleton (hip joint) and human orientation. The results are shown in Table 2.

Table 1. The results on Human 3.6M under different operations.

methods	AP_{50}	AP_{150}
2views	89.5	93.4
2D filter-2views	91.1	94.7
3D filter-2views	91.0	94.5
2D + 3D filter-2views	92.1	95.6
4views	89.7	94.6
2D filter-4views	92.8	95.7
3D filter-4views	93.1	95.7
2D + 3D filter-4views	94.7	96.0

Table 2. The results of Baduanjin and Taichi under different operations.

methods	Baduanjin		Taichi	
	AP_{50}	AP_{150}	AP_{50}	AP_{150}
2 views	76.8	86.3	71.1	83.4
2D filter-2 views	78.2	86.3	72.5	85.0
3D filter-2 views	78.4	86.5	73.6	86.1
2D + 3D filter-2 views	78.4	86.8	73.5	86.4
4 views	83.7	87.9	78.3	88.7
2D filter-4 views	83.6	88.3	78.1	89.6
3D filter-4 views	84.5	88.2	82.1	89.9
2D + 3D filter-4 views	84.5	88.7	83.8	89.8
8 views	86.4	91.3	86.7	90.1
2D filter-8 views	86.1	92.5	87.6	91.7
3D filter-8 views	86.5	93.1	87.8	91.7
2D + 3D filter-8 views	87.2	94.0	88.4	91.9

The Campus dataset contains more than person. In this dataset, the pose data of three persons so we evaluate the results of each person (as shown in Table 3).

Table 3. The results of 3D pose under different methods on Campus.

(a) The results under tracking- and matching-based methods

methods	Actor1		Actor2		Actor3		Average	
	AP_{50}	AP_{150}	AP_{50}	AP_{150}	AP_{50}	AP_{150}	AP_{50}	AP_{150}
No filter	54.9	86.9	65.4	96.1	56.8	93.4	58.8	91.8
2D filter	55.8	87.2	66.5	96.6	57.2	93.9	59.6	92.2
3D filter	55.8	87.3	66.9	96.6	56.7	94.1	59.6	92.4
2D+3D filter	55.8	87.4	66.9	96.7	56.8	94.2	59.7	92.6

(b) The results based on VoxelPose

methods	Actor1		Actor2		Actor3		Average	
	AP_{50}	AP_{150}	AP_{50}	AP_{150}	AP_{50}	AP_{150}	AP_{50}	AP_{150}
No filter	55.6	89.3	56.4	94.0	59.0	95.0	56.8	94.1
3D filter	55.8	91.0	57.2	93.9	59.2	96.2	57.2	94.8

From the Tables, we can see that the filter method can improve the accuracy of 3D pose estimation, especially in 3D space. Moreover, the accuracy of the survey is increasing with the number of views. In theory, 2D poses in 2 views are enough to reconstruct a 3D pose; however, due to missing false key points, a 3D skeleton based on only 2 views always contains incorrect vital points. Therefore, when constructing a 3D pose dataset in a multiview environment, we need to determine the number of views according to the complexity of motion and select 2D poses under appropriate views to improve the quality of the ultimate dataset. In a multi-person environment, the filtering operation can also help get more accurate results, whether for humans or all persons.

5. Conclusion

In this paper, aiming to address the problems of pose estimation and motion evaluation, we propose constructing a 3D pose dataset based on a multiview environment. The method combines the 2D and 3D pose estimation in a multiview environment. With top-down 2D pose estimation and tracking, we can locate the human of interest in single-person scenes and finish human matching in multi-person scenes. By experimenting with the methods on different datasets, we can see that filtering operations in different dimensions can effectively improve the accuracy of 3D pose estimation results. Compared with the traditional methods, ours is simple and flexible. After completing the previous works of deploying devices and calibration, we can quickly finish capturing images and get the precious 3D position and pose results of humans or other interested objects just by computing techniques. In the future, we will research how to increase the speed of operation and improve realtime performance.

References

[1] Moon G, Chang J Y, Lee K M. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 10133-10142.

[2] Weinzaepfel P, Brégier R, Combaluzier H, et al. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild[C]//European Conference on Computer Vision. Springer, Cham, 2020: 380-397.

[3] Kocabas M, Athanasiou N, Black M J. Vibe: Video inference for human body pose and shape estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5253-5263.

[4] 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/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7753-7762.

[5] Dong J, Jiang W, Huang Q, et al. Fast and robust multi-person 3d pose estimation from multiple views[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7792-7801.

[6] Tu H, Wang C, Zeng W. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment[C]//European Conference on Computer Vision. Springer, Cham, 2020: 197-212.

[7] Qiu H, Wang C, Wang J, et al. Cross view fusion for 3d human pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4342-4351.

[8] Chen L, Ai H, Chen R, et al. Cross-view tracking for multi-human 3d pose estimation at over 100 fps[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3279-3288.

[9] 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.

[10] Shao D, Zhao Y, Dai B, et al. Finegym: A hierarchical video dataset for fine-grained action understanding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2616-2625.

[11] Johnson S, Everingham M. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation[C]//bmvc. 2010, 2(4): 5.

[12] Belagiannis V, Amin S, Andriluka M, et al. 3D pictorial structures for multiple human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1669-1676.

[13] 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.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[15] Lin W, Liu H, Liu S, et al. Human in events: A large-scale benchmark for human-centric video analysis in complex events[J]. arXiv preprint arXiv:2005.04490, 2020.

[16] Andriluka M, Pishchulin L, Gehler P, et al. 2d human pose estimation: New benchmark and state of the art analysis[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014: 3686-3693.

[17] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[18] Fang H S, Xie S, Tai Y W, et al. Rmpe: Regional multi-person pose estimation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2334-2343.

[19] Cai Y, Wang Z, Luo Z, et al. Learning delicate local representations for multi-person pose estimation[C]//European

Conference on Computer Vision. Springer, Cham, 2020: 455-472.

[20] Zhang F, Zhu X, Dai H, et al. Distribution-aware coordinate representation for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7093-7102.

[21] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[22] Zhou K, Yang Y, Cavallaro A, et al. Omni-scale feature learning for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3702-3712.

[23] Zhang Z. A flexible new technique for camera calibration[J]. IEEE Transactions on pattern analysis and machine intelligence, 2000, 22(11): 1330-1334.

[24] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures[J]. Analytical chemistry, 1964, 36(8): 1627-1639.

[25] Liu W, Mei T. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective[J]. ACM Computing Surveys (CSUR), 2022.

[26] Cheng B, Xiao B, Wang J, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5386-5395.

[27] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[28] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[29] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[30] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[31] Zhang Y, Wang C, Wang X, et al. Fairmot: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.

[32] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[C]//European Conference on Computer Vision. Springer, Cham, 2020: 107-122.