Electronic Research Archive

ERA, 31(3): 1485–1497.
DOI: 10.3934/era.2023075
Received: 01 November 2022
Revised: 14 December 2022
Accepted: 27 December 2022
Published: 16 January 2023

http://www.aimspress.com/journal/ERA

Research article

Regress 3D human pose from 2D skeleton with kinematics knowledge

Longkui Jiang[1], Yuru Wang[2,*] and Weijia Li[2]

[1] Technology School, Jilin Business and Technology College, Changchun, China
[2] School of Information Science and Technology, Northeast Normal University, Changchun, China

*** Correspondence: Email:** wangyr915@nenu.edu.cn; Tel: 08613604407105.

Abstract: 3D human pose estimation is a hot topic in the field of computer vision. It provides data support for tasks such as pose recognition, human tracking and action recognition. Therefore, it is widely applied in the fields of advanced human-computer interaction, intelligent monitoring and so on. Estimating 3D human pose from a single 2D image is an ill-posed problem and is likely to cause low prediction accuracy, due to the problems of self-occlusion and depth ambiguity. This paper developed two types of human kinematics to improve the estimation accuracy. First, taking the 2D human body skeleton sequence obtained by the 2D human body pose detector as input, a temporal convolutional network is proposed to develop the movement periodicity in temporal domain. Second, geometrical prior knowledge is introduced into the model to constrain the estimated pose to fit the general kinematics knowledge. The experiments are tested on Human3.6M and MPII (Max Planck Institut Informatik) Human Pose (MPI-INF-3DHP) datasets, and the proposed model shows better generalization ability compared with the baseline and the state-of-the-art models.

Keywords: 3D human pose estimation; temporal convolution; human kinematics; knowledge model

1. Introduction

Human pose estimation is a key technology in the field of computer vision. Its output is the basis of down-stream tasks such as action recognition, visual tracking and action analysis. The early work in human pose estimation was mainly limited to a 2D plane, and the goal is to get the body joints' 2D coordinates from 2D images or videos. In recent years, 3D body pose estimation has become popular because it provides more accurate data with depth information. 3D pose estimation can be categorized into three types, according to the input: from a monocular image [1–4], from multi-

camera images [5–6] and from a depth image [7–9]. Monocular 3D human pose estimation is the most popular, and it is widely used in applications such as virtual reality, intelligent video analysis and human-computer interaction.

At present, there are two main branches for monocular 3D human pose estimation. One is the so-called two-stage method, which first estimates the 2D human pose and then lifts it to a 3D human pose. An example is the weakly supervised model [10]. The other is the end-to-end 3D human pose estimation method, which predicts the 3D human pose directly from images or videos. Examples are the adversarial learning method [11], the self-supervised approach [12] and the famous Transformer [13]. Because the human pose shows spatial correlation, some work tried to extract skeleton features in the spatial domain. Liu et al. [14] employed graph networks with weight sharing to do 3D pose estimation. The stacked graph hourglass model [15] tried to capture multi-scale spatial correlation. With the goal of capturing both the spatial and temporal correlation of a human pose, Zhang et al. [16] proposed a spatial-temporal encoder to learn spatial-temporal correlations. In comparison with the two-stage model, the end-to-end model regresses the 3D pose directly from the 2D image, which provides the model with rich information. However, it usually requires the support of large-scale human pose datasets. The 2D pose datasets includes Leeds Sports Pose Dataset (LSP) [17], Frames Labeled In Cinema (FLIC) [18], Max Planck Institut Informatik (MPII) [19] and Microsoft COCO: Common Objects in Context (MSCOCO) [20]. The 3D pose datasets include HumanEva [21], MPI-INF-3DHP and human 3.6M [22]. For 3D human poses, it is a very challenging task to obtain large-scale labels. Therefore, most of the existing data are collected in the laboratory using motion capture systems (such as human 3.6M), and the backgrounds are relatively simple and very limited in number. Due to these limitations, the end-to-end methods usually perform better in some specific scenarios but cannot generalize well to applications in natural scenes. In order to improve the model's generalizability, Gholami et. al. [23] proposed adapting the training data to the test dataset, such as camera viewpoint, position, human actions and body size.

The two-stage model estimates the 3D pose by two steps. It first gets a 2D pose by a 2D detector, and then regresses the 3D pose from the 2D. Obviously, the two-stage model will heavily rely on the 2D detector; however, the 2D pose datasets are more sufficient than 3D and contain much in-the-wild data. Thus, the 2D detector will be trained by more diverse data, and the two-stage model can be expected to show better generalizability. In addition, the two-stage model also has an advantage of low complexity. Therefore, in this paper, we adopt it to predict the 3D human pose with the 2D human skeleton as input.

The main contributions of this work include the following:

1) We design a multi-stage supervision temporal convolution network to capture human dynamics by temporal continuity constraints. In addition, the network is trained in a multi-stage supervision manner to improve the model.

2) We impose that the model is to be consistent with general human pose dynamic knowledge and introduce human body pose geometry to the network training step, so as to improve the model generality.

2. Related works

The task of 3D pose estimation is more challenging than 2D pose estimation, because it needs to regress relative depth between body joints, which suffers from severe ambiguity.

For the two-stage 3D pose estimation models, Martinez et. al. [24] designed a simple multi-layer network and regressed the 3D pose from the 2D pose skeleton. At first, this work analyzed the reasons for low accuracy in 3D pose estimation, and determined two aspects: the low accuracy of the 2D estimation, and the mapping from 2D to 3D. In fact, the 2D detector has achieved very high performance, so this paper focused on estimating the 3D pose from the 2D. Therefore, at second, they designed a very simple and lightweight network, and achieved good performance. Their work demonstrated the effectiveness of the two-stage model.

Based on Martinez's work, Fang et al. [25] extended it by a pose semantic network to code joints' dependency and correlations. Because recurrent neural network is good at learning temporal correlations, it was also introduced in Martinez's model. Hossian and Little [26] used Long short-term memory (LSTM) to capture the temporal continuity and achieved accuracy improvement. However, this model cannot deal with long-term sequential data, because it will lead to the gradient vanishing and gradient explosion problems.

Temporal convolution networks provide a new way to capture temporal continuity, so they are also used in the field of pose estimation. WaveNet [27] proved the convolution model's advantages in capturing temporal information. WaveNet is constructed by 1D convolution, and it can prevent the problems of gradient vanishing and gradient explosion. In addition, it is of high efficiency, because it can process temporal data in parallel. Based on this model, Pavllo et. al. [28] designed a temporal convolution model to estimate 3D body pose. This model generates the 2D pose sequence first by the 2D detector and then estimates the 3D pose. In comparison with WaveNet, the temporal convolution model is advantaged in learning the implicit kinematics knowledge. Instead of estimating 3D human pose from monocular images, videos can provide temporal information to improve accuracy and robustness. Several works [29–31] utilized spatial-temporal relationships and constraints such as bone-length and left-right symmetry to improve performance. In this paper, we employed the temporal convolution model and improved it in two ways: First, we design a multi-stage supervision model to further explore the periodic motion pattern; second, we introduce the prior geometry knowledge to generalize the model.

The performance of data-driven model is limited by the dataset, so prior knowledge is imposed to the deep models in many computer vision fields [32–34]. With the goal of decreasing depth ambiguity, some work tried to introduce human geometric knowledge into the 3D pose estimation model. Belagiannis et. al. [35] imposed kinematic constraints on the translation and rotation between body parts in the 3D pictorial model, and the symmetric body parts are constrained not to collide with each other. Ronchi et. al. [36] imposed limb length loss and measured the difference in length between the predicted limb and the predefined reference length. In fact, we can develop many kinematic constraints, such as limb lengths, limb length proportions, joint angles, occlusion constraints, appearance constraints and temporal smoothness constraints. Kinematics is crucial prior knowledge for the deep models, and it constrains the model predictions to be reasonable when measured by body geometry.

3. Periodical temporal convolution network with multi-stage supervision

In the temporal domain, human movement always shows continuity according to human kinematics. If we can capture the temporal continuity, it will provide important information for the 3D pose estimation model. In this paper, we employed the temporal convolution network to model the periodical human kinematics.

The model structure is shown in Figure 1. The network takes a 2D human pose sequence of size 243×34 (17 joints' 2D pose coordinates, and 243 is the sequence length) as input. The input sequence is passed through four same consecutive modules, which are composed of a 1D convolution with 3 convolution kernels and 1024 output channels, batch normalization layer, Rectified Linear Unit (ReLu) and dropout. Each module is added by a residual connection, as shown in the upper part of Figure 1. The dimension of the input data is directly reduced through a specific slice function (Slice), and the results are added to the output data of the module. The channel number is 1024 in the module. Each module contains two different convolution layers. The first convolution layer applies extended convolution, which is mainly used to extract data features. The kernel size is 3, and the expansion rate is 3 (3d3 in the figure). With the increase of modules, the kernel size is fixed, but the expansion rate increases exponentially. As shown in this figure, the expansion rate of the second module is 9, while it is 27 for the third and 81 for the last. Different from the first convolution layer, the kernel size for the second one is 1, which is used to increase the depth of the network and improve the nonlinearity. At the end of the network, a convolution layer is applied to output 3D human posture.

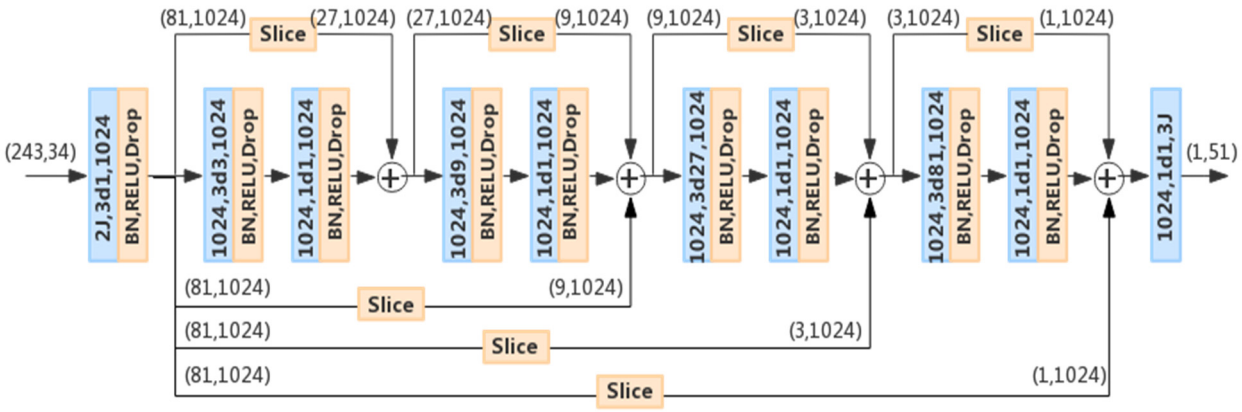


Figure 1. Temporal convolutional network structure based on multi-stage supervision.

In order to avoid the gradient vanishing problem, we add residual connection to enhance the gradient propagation. As shown in Figure 1, we add multi-level residual connection to supervise the network in multiple stages. In the process of forward propagation, multi-level residual connection enables shallow features to be directly propagated to the upper layer. The features of the shallow layer are combined with high-level features as input to the next layer. Combining features at different levels helps to reduce network degradation and improve network generalization performance. In the process of back propagation, the gradient can be transmitted to the lower layer faster without too much intermediate weight matrix transformation, so it can effectively alleviate gradient vanishing. The time-series convolution network based on multi-stage supervision makes the feature information more

smoothly spread in the forward and backward directions, so the network has better optimization performance, which will further improve the model accuracy.

The input is the joints' 2D coordinates of the consecutive frames $x_{2d}(x \in R^{2n})$, and the model will output the joints' 3D coordinate estimations $(y_{3d}, y \in R^{3n})$. The loss function is defined as the Euclidean distance between the estimated 3D pose and the ground-truth $(y_t, y \in R^{3n})$:

$$L_{mpjpe} = \frac{1}{N}\sum_N \|y_{3d} - y_t\| \tag{1}$$

in which N is the batch size and is set to be 1024 in our case. We use gradient descent to optimize the model and exponential decay to update the learning rate. The learning rate is set to be 0.001, and the decay rate is 0.95.

4. Geometry constraints

Relying on joints' coordinates only tends to cause ambiguity when restoring 3D coordinates from 2D. Therefore, the geometric prior knowledge of human kinematics is introduced in this part. We employed the distance $L_{p-mpjpe}$ between the ground-truth coordinates and the estimated joints' coordinates after translation, rotation and scale transformation.

$$L_{p-mpjpe} = \frac{1}{N}\sum_N \|T(y_{3d}) - y_t\| \tag{2}$$

where T is the transformation operation. In addition, we also introduced the geometry consistency as constraints, including bone length symmetry and proportions. The skeleton of a normal human body is symmetrical: for example, the bone length of the left shoulder is the same as that of the right shoulder.

The bone length symmetry constraint:

$$L_{sym} = \sum_i \frac{1}{|S_i|}\sum_{e\in S_i}(l_e - l_{sym(e)})^2 \tag{3}$$

The bones connecting human joints are divided into four groups $S = \{S_{arm}, S_{leg}, S_{shoulder}, S_{hip}\}$. S_{arm} includes left and right upper arms and left and right lower arms. S_{leg}includes the left and right thighs and the left and right calves. $S_{shoulder}$ contains the left shoulder bone and the right shoulder bone, and S_{hip}contains the left hip bone and the right hip bone. l_e is the length of the selected bone. The length of the bone at the symmetrical position is obtained through the sym function. For example, if l_e is the right thigh, $l_{sym(e)}$ will be the left thigh. l_{sym} controls the bone length by calculating the error of each group of bones and constrains the bone length at symmetrical positions.

The bone length proportions constraint:

$$L_{rat} = \sum_i \frac{1}{|R_i|}\sum_{e\in R_i}\left(\frac{l_e}{\overline{l_e}} - \overline{r_i}\right)^2 \tag{4}$$

$$\overline{r_i} = \frac{1}{|R_i|}\sum_{e\in R_i}\frac{l_e}{\overline{l_e}} \tag{5}$$

The bones are also grouped into four groups $R = \{R_{arm}, R_{leg}, R_{shoulder}, R_{hip}\}$. l_e is the bone length, $\overline{l_e}$ is its corresponding mean length in its datasets. The proportion $l_e/\overline{l_e}$ should be consistent in the same group. $\overline{r_i}$ is the mean proportion in group i.

The constraints and the accuracy loss are integrated as the final loss:

$$L = \lambda_1 L_{mpjpe} + \lambda_2 L_{p-mpjpe} + \lambda_3 L_{sym} + \lambda_4 L_{rat} \quad (6)$$

in which $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the weights of each constraint.

5. Experimental results and analysis

5.1. Dataset and metrics

We tested our method on the Human3.6M and MPI-INF-3DHP datasets. The Human3.6M dataset contains 15 actions of 11 testers, with a total of 3.6 million video frames. For the task of 3D human pose estimation, there are mainly three standard evaluation protocols based on this dataset: Protocol 1 (MPJPE) is the average joint position error in millimeters, which is the Euclidean distance between the predicted joint position and the real position. Protocol 2 (P-MPJPE) is the error after the predicted joint position is aligned with the real position after translation, rotation and retraction. Protocol 3 (N-MPJPE) is the error after aligning the predicted joint position with the real position only after scaling. Among the three protocols, Protocol 1 (MPJPE) is the most widely used. However, for the method of predicting 3D human pose based on sequence, absolute position error cannot measure the smoothness of prediction over time. In order to evaluate this, Pavllo et al. [28] measured the joint velocity error (MPJVE), which is a time-based velocity motion measurement and the first-order derivative of MPJPE's 3D pose error. For the Human 3.6M dataset, we employed the 17 joint skeletons, used 5 testers (S1, S5, S6, S7, S8) for training and 2 testers (S9, S11) for testing and trained a general model for 15 actions. The MPI-INF-3DHP test set [37] provides images in three different scenarios: studio with a green screen (GS), studio without green screen (noGS) and outdoor scene (Out- door). We use this dataset to test the generalization ability of our model and use 3D-PCK and AUC as evaluation metrics.

5.2. 3D pose estimation results based on ground-truth 2D data

In this section, we employ the 2D ground-truth data (2D skeletons of 243 frames) as the input to the second stage. The channel number is set to be 1024. As shown in Tables 1–3, the proposed model with multi-stage intermediate supervision achieved lower error when evaluated by all three protocols.

Table 1. Protocol 1: reconstruction error (MPJPE).

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
Pavllo et al. [28]	26.0	29.8	24.6	27.0	25.8	29.4	29.2	26.7
Ours	25.7	29.2	25.1	27.0	25.8	30.4	28.7	25.9
	Sit.	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg
Pavllo et al. [28]	31.7	34.6	27.4	27.3	27.9	21.5	21.8	27.4
Ours	30.7	35.0	27.2	26.8	27.7	21.4	22.6	27.3

Table 2. Protocol 2: reconstruction error (P-MPJPE).

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
Pavllo et al. [28]	26.0	29.8	24.6	27.0	25.8	29.4	29.2	26.7
Ours	25.7	29.2	25.1	27.0	25.8	30.4	28.7	25.9
	Sit.	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg
Pavllo et al. [28]	31.7	34.6	27.4	27.3	27.9	21.5	21.8	27.4
Ours	30.7	35.0	27.2	26.8	27.7	21.4	22.6	27.3

Table 1 shows the evaluation by Protocol 1 (MPJPE). The proposed method shows performance improvement on most actions, even for some relatively difficult actions, such as "sitting", "sittingD", and "Discussion". Averagely, the error is reduced by about 1 mm, and the prediction accuracy is increased by 2.7%. Table 2 shows the evaluation by Protocol 2 (P-MPJPE). The proposed model achieved 0.1 mm error reduction and about 1 mm for the actions of "Sitting". Table 3 is based on Protocol 3 (N-MPJPE), where the proposed model achieved about 0.7 mm error reduction and about 1.7 mm for the actions of "Sitting" and "Discussion". The proposed model also achieved lower error in joint velocity, as shown in Table 4, which means better temporal smoothness.

Table 3. Protocol 3: reconstruction error (N-MPJPE).

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
Pavllo et al. [28]	36.0	39.1	31.4	35.6	33.5	38.0	40.5	34.7
Ours	34.5	37.4	31.7	34.8	33.5	39.1	39.2	33.2
	Sit.	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg
Pavllo et al. [28]	41.8	41.4	34.9	36.8	34.5	26.4	27.1	35.5
Ours	40.1	40.9	33.9	35.4	34.3	26.6	27.7	34.8

Table 4. Velocity error over the generated 3D poses on Human 3.6M.

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
Pavllo et al. [28]	1.92	1.97	1.48	2.27	1.42	1.79	1.85	2.16
Ours	1.92	1.94	1.45	2.24	1.39	1.75	1.80	2.11
	Sit.	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg
Pavllo et al. [28]	1.11	1.53	1.40	1.59	2.68	2.29	1.91	1.83
Ours	1.07	1.49	1.37	1.56	2.64	2.27	1.90	1.79

5.3. Results based estimated 2D Pose

In order to test performance of regressing the 3D pose directly from the 2D image, we employed a Cascade Pyramid Network (CPN) [38] as the 2D detector in the two-stage model, and the predicted 2D skeletons' sequence is input to the 3D estimator. Table 5 compares the proposed model with the state-of-the-art models, where "U" is the model with multi-stage supervision, and "U+L" is the model with multi-stage supervision and geometry constraints. Our method achieved the best result on almost all the actions. For the actions "Phone" and "Photo", our model performs worse than the baseline model. For these two actions, the kinematics feature is not as obvious as other actions in both spatial and temporal domain, especially for the

action "Phone," so our model did not show advantages. For the action "Photo," the "U+L" model performs better than "U" model, which means the geometry constraints are effective for this action.

With predicted 2D pose as input, the model shows less accuracy than the model with ground-truth 2D pose as input. As shown in Table 5, the average prediction error is reduced by about 2.3%. In comparison, the "U+L" model shows better performance than the "U" model, and averagely achieved 0.8% improvement. The model with "U+L" achieved the lowest error on almost all the actions. We can see that the "U+L" model has the same number of parameters as the "U" model but better generalization ability.

We applied the model trained on Human 3.6m to the MPI-INF-3DHP dataset, to test the model's generalizability. Table 6 shows the results and comparison with the state-of-the-art models. Trained only on the Human 3.6M dataset, our model shows good generalizability, due to the general knowledge being data independent.

Table 5. Comparison of experimental results under protocol 1 (MPJPE, bold: best, underline: second best) on Human3.6M.

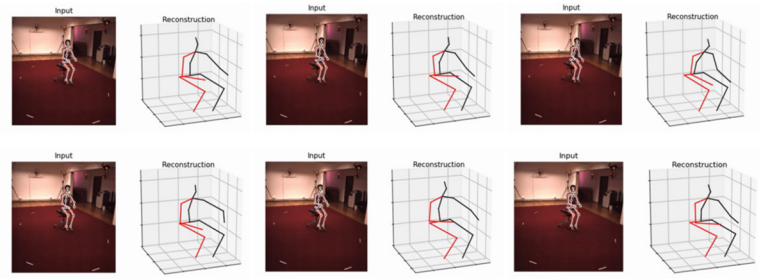
	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
Martinez et al. [24]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Sun et al. [39]	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6
Fang et al. [25]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7
Pavlakos et al. [40]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9
Yang et al. [41]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7
Luvizon et al. [42]	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7
Hossain and Little [26]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7
Lee et al. [43]	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	43.0
Pavllo et al. [28]	45.2	46.7	43.3	45.6	**48.1**	**55.1**	44.6	44.3
Liu et al. [14]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0
Xu and Takano [15]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3
Ours (U)	44.6	<u>46.5</u>	<u>43.0</u>	<u>45.4</u>	<u>48.4</u>	57.3	**43.9**	<u>43.7</u>
Ours (U+L)	<u>44.3</u>	**46.1**	**42.5**	**45.2**	<u>48.4</u>	<u>56.0</u>	**43.9**	**43.5**
	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT	Avg.
Martinez et al. [24]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Sun et al. [39]	71.7	86.7	61.5	53.4	61.6	47.1	53.4	59.1
Fang et al. [25]	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Pavlakos et al. [40]	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Yang et al. [41]	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Luvizon et al. [42]	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
Hossain and Little [26]	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee et al. [43]	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Pavllo et al. [28]	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Liu et al. [14]	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.7
Xu and Takano [15]	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Ours (U)	**56.6**	**64.3**	<u>47.0</u>	<u>43.9</u>	<u>49.2</u>	<u>32.7</u>	**33.7**	<u>46.7</u>
Ours (U+L)	<u>56.7</u>	<u>64.6</u>	**45.6**	**43.6**	**48.9**	**32.6**	**33.7**	**46.4**

Table. 6 Results on the MPI-INF-3DHP test set.

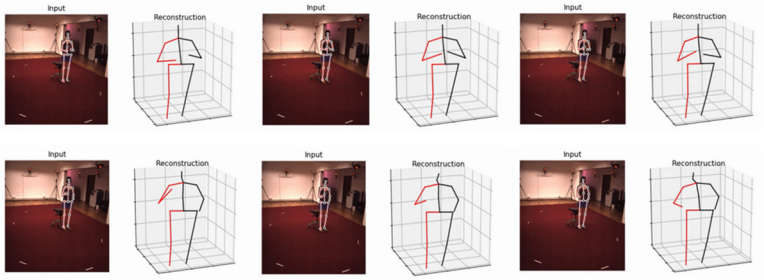
	Training Data	GS	noGS	Outdoor	All (PCK)	All (AUC)
Martinez et al. [24]	Human3.6M	49.8	42.5	31.2	42.5	17.0
Zhou et al. [44]	Human3.6M+MPII	75.6	71.3	80.3	75.3	38.0
Xu and Takano [15]	Human3.6M	81.5	81.7	75.2	80.1	45.8
Ours (U)	Human 3.6M	81.7	81.6	75.4	80.2	46.0
Ours (U+L)	Human3.6M	81.6	81.7	75.0	80.1	45.8

5.4. Results visualization

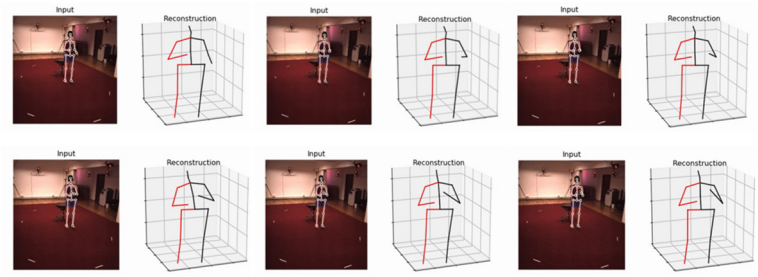
In addition to the above quantitative experimental results, we also visualize the 3D pose results for the Human3.6M dataset. Figure 2 shows the prediction effects of some actions, including eating, talking on the phone and smoking. It can be seen from the figure that the proposed model effectively restores the human pose in 3D space with a high prediction accuracy.



(a) Eating



(b) Phoning



(c) Smoking

Figure 2. The visualized prediction results of some typical action in the Human3.6M database.

6. Conclusions

For 3D pose estimation, this paper proposed developing human kinematics in two ways. First, we employed the temporal convolution network to extract the temporal continuity and supervised by constructing multi-stage intermediate connections to alleviate gradient vanishing. Second, we introduced geometry constraints to improve the model generalizability. When tested on two public datasets, the proposed model showed comparable performance with the state-of-the-art models. Developing human kinematics is important information for a data-driven model. This paper presents a preliminary study, and developing more kinematics will provide the data-driven model with more effective prior knowledge, which is also our future work.

Acknowledgments

We would like to thank for the support by the Science and Technology Development Program of Jilin Province (20220101102JC) and Jilin Province Professional Degree Postgraduate Teaching Case Construction Project.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, *IEEE Trans. Pattern Anal. Mach. Intell.*, **28** (2006), 44–58. https://doi.org/10.1109/tpami.2006.21
2. J. Cho, M. Lee, S. Oh, Single image 3D human pose estimation using a procrustean normal distribution mixture model and model transformation, *Comput. Vis. Image Und*., **155** (2017), 150–161. https://doi.org/10.1016/j.cviu.2016.11.002
3. T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, M. Magnor, Optical flow-based 3D human motion estimation from monocular video, in *German Conference on Pattern Recognition*, **10496** (2017), 347–360. https://doi.org/10.1007/978-3-319-66709-6_28
4. X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, K. Daniilidis, Sparseness meets deepness: 3D human pose estimation from monocular video. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 4966–4975. https://doi.org/10.1109/CVPR.2016.537
5. A. Shafaei, J. J. Little, Real-time human motion capture with multiple depth cameras, in *2016 13th Conference on Computer and Robot Vision (CRV)*, (2016), 24–31. https://doi.org/10.1109/CRV.2016.25
6. D. Michel, C. Panagiotakis, A. A. Argyros, Tracking the articulated motion of the human body with two RGBD cameras, *Mach. Vision Appl.*, **26** (2015), 41–54. https://doi.org/10.1007/s00138-014-0651-0
7. Y. Zhu, K. Fujimura, Bayesian 3D human body pose tracking from depth image sequences, in *Asian Conference on Computer Vision*, **5995** (2009), 267–278. https://doi.org/10.1007/978-3-642-12304-7_26

8. X. Zheng, M. Fu, Y. Yang, N. Lv, 3D Human poses recognition using Kinect, in *2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics*, (2012), 344–347. https://doi.org/10.1109/IHMSC.2012.92
9. Y. Guo, Z. Li, Z. Li, X. Du, S. Quan, Yi Xu, PoP-Net: Pose over parts network for multi-person 3D pose estimation from a depth image, in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2022), 3917–3926. https://doi.org/10.1109/WACV51458.2022.00397
10. X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3D human pose estimation in the wild: A weakly-supervised approach, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 398–407. https://doi.org/10.1109/ICCV.2017.51
11. W. Yang, W. Ouyang, Xi. Wang, J. Ren, H. Li, X. Wang, 3D human pose estimation in the wild by adversarial learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 5255–5264. https://doi.org/10.1109/CVPR.2018.00551
12. J. N. Kundu, S. Seth, P. Ym, V. Jampani, A. Chakraborty, R. V. Babu, Uncertainty-aware adaptation for self-supervised 3D human pose estimation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 20416–20427. https://doi.org/10.1109/CVPR52688.2022.01980
13. C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, 3D human pose estimation with spatial and temporal transformers. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 11636–11645. https://doi.org/10.1109/ICCV48922.2021.01145
14. K. Liu, R. Ding, Z. Zou, L. Wang, W. Tang, A comprehensive study of weight sharing in graph networks for 3D human pose estimation, in *European Conference on Computer Vision*, (2020), 318–334. https://doi.org/10.1007/978-3-030-58607-2_19
15. T. Xu, W. Takano, Graph stacked hourglass networks for 3D human pose estimation, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 16100–16109. https://doi.org/10.1109/CVPR46437.2021.01584
16. J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 13222–13232. https://doi.org/10.1109/CVPR52688.2022.01288
17. L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Strong appearance and expressive spatial models for human pose estimation, in *2013 IEEE International Conference on Computer Vision*, (2013), 3487–349. https://doi.org/10.1109/ICCV.2013.433
18. B. Sapp, B. Taskar, Modec: Multimodal decomposable models for human pose estimation, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 3674–3681. https://doi.org/10.1109/CVPR.2013.471
19. M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: new benchmark and state of the art analysis, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 3686–3693. https://doi.org/10.1109/CVPR.2014.471
20. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: Common objects in context, in *European Conference on Computer Vision*, (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

21. L. Sigal, A. O. Balan, M. J. Black, HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, *Int. J. Comput. Vision*, **87** (2010). https://doi.org/10.1007/s11263-009-0273-6
22. Ionescu C., D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2014), 1325–1339. https://doi.org/10.1109/TPAMI.2013.248
23. M. Gholami, B. Wandt, H. Rhodin, R. Ward, Z. J. Wang, AdaptPose: Cross-dataset adaptation for 3D human pose estimation by learnable motion generation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 13065–13075. https://doi.org/10.1109/CVPR52688.2022.01273
24. J. Martinez, R. Hossain, J. Romero, J. J. Little, A simple yet effective baseline for 3D human pose estimation, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2659–2668. https://doi.org/10.1109/ICCV.2017.288
25. H. Fang, Y. Xu, W. Wang, X. Liu, S. Zhu, Learning pose grammar to encode human body configuration for 3D human pose estimation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **32** (2018), 6821–6828. https://doi.org/10.1609/aaai.v32i1.12270
26. M. R. I. Hossain, J. J. Little, Exploiting temporal information for 3D pose estimation, in *European Conference on Computer Vision*, **11214** (2018), 69–86. https://doi.org/10.1007/978-3-030-01249-6_5
27. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., WaveNet: A generative model for raw audio, preprint, arXiv:1609.03499.
28. D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7745–7754. https://doi.org/10.1109/CVPR.2019.00794
29. R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, A. Jain, Learning 3D human pose from structure and motion, in *European Conference on Computer Vision*, (2018), 679–696. https://doi.org/10.1007/978-3-030-01240-3_41
30. Y. Cai, L. Ge, J. Liu, J. Cai, T. Cham, J. Yuan, et al., Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 2272–2281. https://doi.org/10.1109/ICCV.2019.00236
31. Z. Li, X. Wang, F. Wang, P. Jiang, On boosting single-frame 3D human pose estimation via monocular videos, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 2192–2201. https://doi.org/10.1109/ICCV.2019.00228
32. Z. Cui, T. Song, Y. Wang, Q. Ji, Knowledge augmented deep neural networks for joint facial expression and action unit recognition, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, (2020), 14338–14349.
33. Q. Chen, B. Zhong, Q. Liang, Q. Deng, X. Li, Teacher-student knowledge distillation for real-time correlation tracking, *Neurocomputing*, **500** (2022), 537–546. https://doi.org/10.1016/j.neucom.2022.05.064
34. X. Sun, X. Zhang, L. Cao, Y. Wu, F. Huang, R. Ji, Exploring language prior for mode-sensitive visual attention modeling, in *Proceedings of the 28th ACM International Conference on Multimedia*, (2020), 4199–4207, https://doi.org/10.1145/3394171.3414008

35. V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic, 3D pictorial structures revisited: Multiple human pose estimation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2016). https://doi.org/10.1109/TPAMI.2015.2509986
36. M. R. Ronchi, O. M. Aodha, R. Eng, P. Perona, It's all relative: Monocular 3D human pose estimation from weakly supervised data, preprint, arXiv:1805.06880.
37. D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, et al., Monocular 3D human pose estimation in the wild using improved CNN supervision, in *2017 International Conference on 3D Vision (3DV)*, (2017), 506–516. https://doi.org/10.1109/3DV.2017.00064
38. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 7103–7112. https://doi.org/10.1109/CVPR.2018.00742
39. X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2621–2630. https://doi.org/10.1109/ICCV.2017.284
40. G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3D human pose estimation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 7307–7316. https://doi.org/10.1109/CVPR.2018.00763
41. W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, X. Wang, 3D human pose estimation in the wild by adversarial learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 5255–5264. https://doi.org/10.1109/CVPR.2018.00551
42. D. C. Luvizon, D. Picard, H. Tabia, 2D/3D pose estimation and action recognition using multitask deep learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 5137–5146. https://doi.org/10.1109/CVPR.2018.00539
43. K. Lee, I. Lee, S. Lee, Propagating lstm: 3D pose estimation based on joint interdependency, in *European Conference on Computer Vision*, **11211** (2018), 119–135. https://doi.org/10.1007/978-3-030-01234-2_8
44. K. Zhou, X. Han, N. Jiang, K. Jia, J. Lu, Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, 2344–2353.

AIMS Press