UDC 004.89

IMPROVED 3D HUMAN POSE ESTIMATION FROM VIDEO BASED ON MIXSTE MODEL

Tongrui LI, Sergey ABLAMEYKO
(Belarusian State University, Minsk)

Abstract. *3D human pose estimation is an important branch in the field of computer vision. Due to depth blur and self-occlusion, the accuracy of existing 3D human pose estimation methods is low. In order to improve this problem, we propose an improved human pose estimation model based on the Transformer model suitable for processing human skeleton data. We add a pre-training stage to perform the task of recovering 3D skeletons from noisy 2D observations. It helps the model to understand and learn the underlying three-dimensional structure of the human body. By conducting experiments on the public 3D pose estimation data set Human3. 6M and comparing it with currently popular 3D pose estimation methods, it is verified that the above algorithm has a high accuracy.*

Keywords: *3D HPE, transformer, Pretraining, MixSTE, time series.*

Introduction. Existing 3D pose estimation methods can be roughly divided into two categories. One is to directly return the three-dimensional coordinates of human body joint points from two-dimensional images [1, 2]. The other is to use a two-dimensional human body pose estimation algorithm to estimate the two-dimensional human body joint points, and then restore the two-dimensional pose to a three-dimensional pose through a neural network [3, 4]. Since estimating the human body's three-dimensional posture directly through images has a high degree of nonlinearity and a large output space, in the task of human body three-dimensional posture estimation, the model based on two-dimensional posture regression to the three-dimensional posture is more applicable. However, recovering three-dimensional poses from two-dimensional poses is an ill-posed problem. One two-dimensional skeleton may correspond to multiple valid three-dimensional skeletons, and the accuracy of existing algorithms needs to be improved.

3D human pose estimation recovers lost depth information from 2D visual observations. Inspired by this, we design a preprocessing link using large-scale 3D motion capture data. We first obtain some 2D skeleton data by randomly projecting the 3D motion capture data, and then randomly mask out some key points and add noise to the 2D skeleton data to simulate occlusion, detection failure and error phenomena in real data. Zhang et al. [5] proposed the MixSTE model, which takes into account the spatial and temporal correlations of joints during motion. The spatial Transformer and temporal Transformer modules are stacked end to end into multiple cycles and then returned to obtain the three-dimensional position of the keypoints. We apply the preprocessing module to this model to build a new model, improving prediction accuracy of 3D human pose estimation.

Methodology. The main principle of pretraining as shown in Figure 1, we utilize a motion encoder to learn human motion representation by recovering 3D human motion from corrupted 2D skeleton sequences. Our network takes a concatenated 2D coordinates $C_{N,T} \in \mathbb{R}^{N \times T \times 2}$ with N joints and T frames as input, where the channel size of the input is 2. Firstly, we project the input keypoints sequence $C_{N,T}$ to high-dimensional feature $P_{N,T} \in \mathbb{R}^{N \times T \times d}$ with feature dimension d for each joint representation. Then we utilize the position embedding matrix for retaining the position information of the spatial and temporal domains. The proposed MixSTE takes the $P_{N,T}$ as input and aims to alternately learn the spatial correlation and separate temporal motion. Finally, we use a regression head to concatenate the outputs $X \in \mathbb{R}^{N \times T \times d}$ of encoder, and take the dimension d to 3 to get the 3D human pose sequence $\mathrm{Out} \in \mathbb{R}^{N \times T \times 3}$.

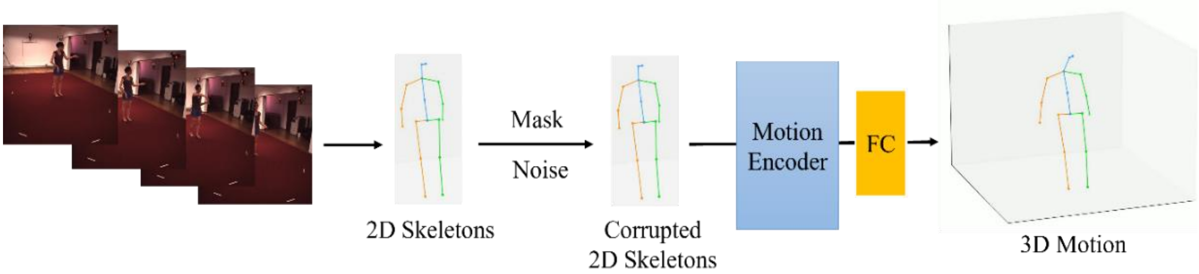

Figure 1. – Pretrain Framework

We utilize the MixSTE to model spatial dependency and temporal motion for a given 2D input keypoint sequence, respectively. MixSTE consists of a Spatial Transformer Block (STB) and a Temporal Transformer Block (TTB). Here, the STB computes the self-attention between joints and aims to learn the body joint relations of each frame, while the TTB computes the self-attention between frames and focuses on learning the global temporal correlation of each joint.

To inject efficient motion trajectories into the learned representations, we consider the temporal correspondence of each joint in order to model the correlation on the same joint over the dynamic sequence. We will separate different joints in the time dimension so that the trajectory of each joint is an individual token $p \in \mathbb{R}^{1 \times T \times d}$, and model different joints of the body in parallel to better represent the temporal correlation. The joint separation operation is as follows:

$$X_l^t = Concat\left(\mathcal{F}\left(p_{i,1}, p_{i,2}, p_{i,1}, \ldots, p_{i,T}\right)\right), \quad i \in N, \tag{1}$$

where $p_{i,j} \in P_{N,T}$, denotes the *i* joint in the *j* frame, $\mathcal{F}$ represents the temporal encoder function, and the output of the *l* TTB encoder is $X_l \in \mathbb{R}^{N \times T \times d}$.

We use the spatial transformer block (STB) to learn spatial correlations among joints in each frame. Given 2D keypoints with *N* joints, we consider each joint as a token in spatial attention. Firstly, we take 2D keypoints as input and project each keypoint to a high-dimensional feature with the linear embedding layer. The feature is referred to as a spatial token in STB. We then embed the spatial position information with a positional matrix $E_{s-pos} \in \mathbb{R}^{N \times d}$. After that, spatial tokens $P_i \in \mathbb{R}^{N \times d}$ of the *i* frame is fed into spatial self-attention mechanism of STB to model dependencies across all joints and output the high-dimensional tokens $X_l^s \in \mathbb{R}^{N \times T \times d}$ in *l* STB.

Experimental results. We evaluate our model on a commonly used 3D HPE datasets, Human3.6M. It is the most widely used indoor dataset for 3D single person HPE. There are 11 professional actors performing 17 actions such as sitting, walking, and talking on the phone. We adopt the experiment setting: all 15 actions are used for training and testing, the model is trained on five sections (S1, S5, S6, S7, S8) and tested on two subjects (S9 and S11). The two commonly used evaluation metrics (MPJPE and P-MPJPE) are involved in this dataset.

Our model is implemented with PyTorch toolkit and runs on a server with NVIDIA RTX 4090 GPUs. In the experiments, two kinds of input 2D pose sequences are utilized the 2D pose estimated by CPN [6]. For model training, we set each mini-batch as 128 sequences. The network parameters are optimized for 20 epochs by Adam optimizer with basic learning rate of 0.001 and decayed by 0.96 after each epoch. We consider the repeat time L of modules, the hidden embedding channel C, and the number of head H in attention block as free parameters that we tailor to the scale of network.

As shown in Table, our model outperforms the vast majority of previous methods, including other Transformer-based spatiotemporal modeling designs, especially when the model is measured against MPJPE. It shows the effectiveness of the proposed new transformer model in learning 3D geometric structures and temporal dynamics. Figure 2 shows that our method achieves lower errors through the proposed pre-training stage and figure 3 shows results of 3D human pose estimation from video.

Table. – Protocol 1: reconstruction error (MPJPE). Protocol 2: reconstruction error after rigid alignment with the ground truth (P-MPJPE), where available. “*” denotes the pre-training module proposed in

Protocol #1	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
MHFormer	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3
STCFormer-L	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2
P-STMO	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9
Ours	37.6	40.2	39.1	33.9	42.4	51.2	37.2	35.9
Ours*	37.1	40.2	39.2	33.7	40.7	48.4	37.7	35.7
Protocol #1	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
MHFormer	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
STCFormer-L	52.5	56.8	41.8	38.4	40.2	26.2	27.7	40.5
P-STMO	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
Ours	51.2	56.1	41.9	38.4	37.3	26.9	27.1	39.7
Ours*	51.1	57.5	41.3	37.6	36.9	25.7	26.0	39.2
Protocol #2	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.
MHFormer	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2
STCFormer-L	29.3	33.0	30.7	30.6	32.7	38.2	29.7	28.8
P-STMO	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5
Ours	31.1	33.4	33.4	28.8	35.4	39.9	30.2	30.9
Ours*	31.3	33.4	33.4	28.8	33.7	38.3	30.7	31.3
Protocol #2	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
MHFormer	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
STCFormer-L	42.2	45.0	33.3	29.4	31.5	20.9	22.3	31.8
P-STMO	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
Ours	42.9	50.1	36.3	30.9	31.5	22.8	23.4	33.4
Ours*	43.1	50.1	35.6	30.5	31.7	21.9	22.7	33.1

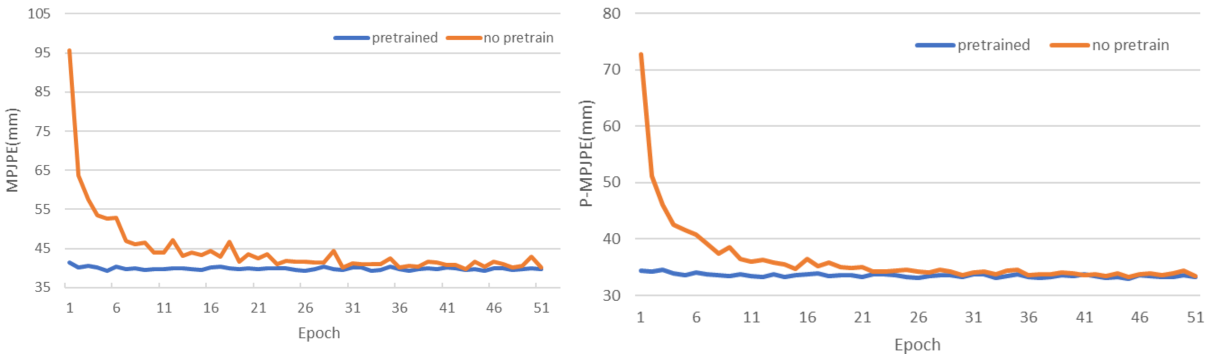

Figure 2. – Learning curves

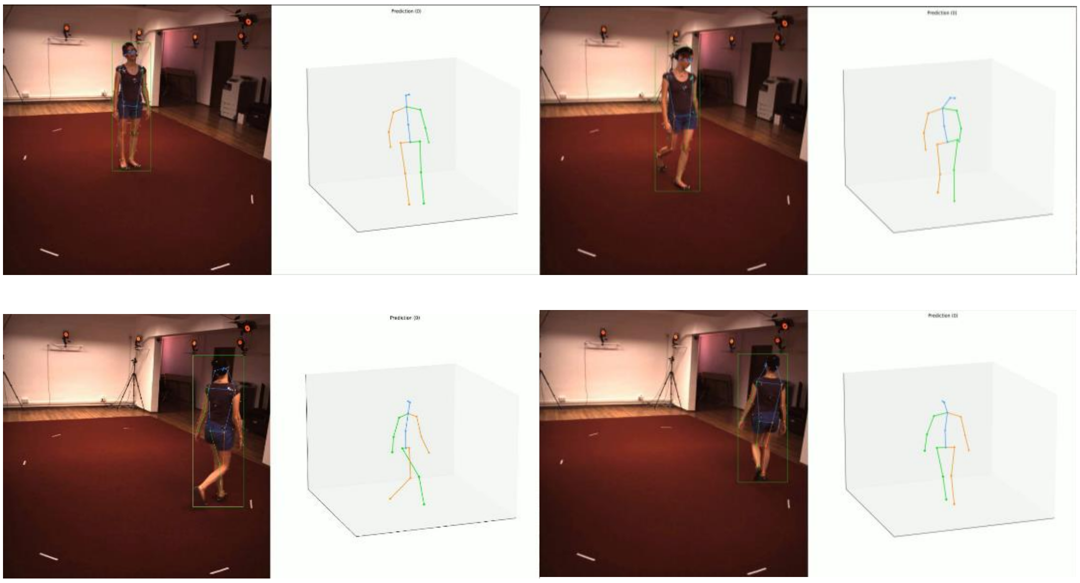

Figure 3. – Results of 3D human pose estimation

Conclusion. We proposed an improved human pose estimation model based on the Transformer model suitable for processing human skeleton data. We added a pre-training stage to perform the task of recovering 3D skeletons from noisy 2D observations. It helps the model to understand and learn the underlying three-dimensional structure of the human body. By conducting experiments on the public 3D pose estimation data set Human3. 6M and comparing it with currently popular 3D pose estimation methods, it is verified that the above algorithm has a high accuracy.

REFERENCES

1. Coarse-to-fine volumetric prediction for single-image 3D human pose. IEEE Conference on Computer Vision and Pattern Recognition., New York. 2017 / PAVLAKOS G [et al.]. – IEEE Press, 2017. – 1263–1272p.
2. 3D human pose estimation from monocular images with deep convolutional neural network. Computer Vision – ACCV. 2014 / LI S J [et al.]. – Cham: Springer International Publishing, 2015. – 332–347 p.
3. VIBE: video inference for human body pose and shape estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition., New York. 2020 / KOCABAS M [et al.]. IEEE Press, 2020.– 5252–5262p.
4. 3D human pose estimation in video with temporal convolutions and semi-supervised training. IEEE/CVF Conference on Computer Vision and Pattern Recognition., New York. 2019 / PAVLLO D [et al.]. – IEEE Press, 2019.– 7745–7754 p.
5. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video / Zhang, J. [et al.]. – arXiv e-prints, 2022. – arXiv.2203.00859.
6. Cascaded Pyramid Network for Multi-person Pose Estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition., Salt Lake City, UT, USA, 2018 / Y. Chen [et al.]. – 7103–7112p.