社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

位置と姿勢の分離に基づく一人称視点映像中の人物運動予測

呉　東昊†　八木　拓真†　松井　勇佑†　佐藤　洋一†

† 東京大学　〒153–8505 東京都目黒区駒場 4–6–1
E-mail: †{wu-dh,tyagi,ymatsui,ysato}@iis.u-tokyo.ac.jp

Egocentric Pedestrian Motion Forecasting for Separately Modelling Pose and Location

Donghao WU†, Takuma YAGI†, Yusuke MATSUI†, and Yoichi SATO†

† The University of Tokyo　4–6–1 KOMABA MEGURO-KU, Tokyo, 153–8505 Japan
E-mail: †{wu-dh,tyagi,ymatsui,ysato}@iis.u-tokyo.ac.jp

Abstract　We study the problem of forecasting human's motion captured from egocentric videos. We propose a novel learning approach by separately modeling human pose and its corresponding scale and position with two deep learning modules, whose outputs are later combined to make the final prediction. Our proposed method successfully forecasts the position and body pose of the target person with an ideal scale, relieving from the mean convergence problem. The experiment is evaluated based on First-Person Locomotion (FPL) dataset. The predictions show the separate modeling approach has plausible-looking visualization results upon egocentric settings, outperforming the state-of-the-art methods which only consider modeling single pose granularity of human motion that suffers from the mean convergence results.

Key words　egocentric vision, human dynamics, motion forecasting, neural network, deep learning

1. Introduction

Human motion prediction has become an increasingly important topic in the computer vision, which is widely applied to human-computer interaction [6]. To enable machines to perceive and have interactions with moving people, knowledge of how the targets move is requested. Most of the previous work conducted research based on fully-supervised skeletons, such as Motion-Capture data [3] and studied based on images taken from fixed cameras [4], which makes their proposed pose forecasting methods hardly to be practically applied in a wild natural environment.

To further extend human motion forecasting to daily scenarios, *e.g.*, dynamic observations on the human motion in a crowded commercial district, in this work, we carried on our research with the wearable monocular RGB camera – viewing people in an egocentric vision to simulate the situations where we observe people in our daily-life. Different from previous data settings, we observed a more complicated dynamics dependencies inherent in human body. For the motion sequence observed in egocentric videos, not only the human pose is varying temporally but it is also accompanied by the changes of scale and location of the target person. For example, the moving target's pose varies while its corresponding size turns larger when the target is moving toward to the observer. (see Figure 1).

Since we have empirically observed that current human motion modeling approaches have difficulties obtaining good prediction performance in egocentric settings (causing severe mean convergence prediction results), we propose a separate modeling approach to tackle the problem. Specifically, we separately learn the pose and its corresponding scale and location with two forecasting modules based on different neural networks. For the pose forecasting module, we train the network modeling based on standardized input pose sequences (with normalised scale and centralized position) aiming to only learn kinetic dependencies of the human pose; for the scale-location forecasting module, we extend the convolution-

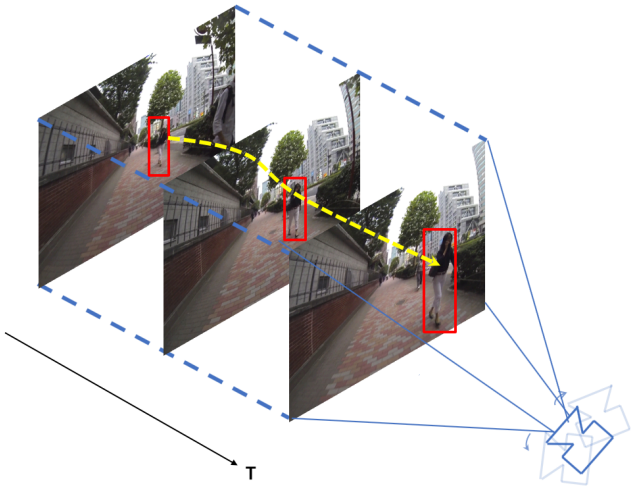

Figure 1 **Observing Human Motion from Egocentric Videos**. The target's pose varies temporally, accompanied by the changes of corresponding scale and location of the body viewed from a moving camera.

deconvolution network proposed in [5] to forecast fine-grained pose sequence while capturing the target' s scale and location granularities to learn visual spacial dependencies of the human motion. We adopt pose displacement cue that contains the temporal variance of body joints, describing how pose changes positions and scales in 2D space in the observed time sequence. With the help of this input cue, we achieved obtaining future pose sequence with expected scale and location comparing with the ground truth. After a combination process with outputs of the two modules, we could generate a complete future pose sequence. Both quantitative and qualitative results demonstrate the effectiveness of our proposed approach.

2. Related Work

Articulated motion forecasting A number of recent works predicted motion focusing on fully supervised skeleton data with joint positions and joint angles. These works proposed structure including Encoder-Recurrent-Decoder (ERD) [7], Structural RNN [10], Residual RNN [9] and convolutional layers with sequence-to-sequence architecture [11] to capture structural-temporal dependencies of body joints. Other methods involves model with adversarial geometry-aware based on geodesic loss [17], joint rotations parameterisation with unit-length quaternions in [15], etc.. Since the motion data utilized in these works have invariant scale and position, we refer to their network architectures to only model pose granularity of the human motion viewed from egocentric videos.

Modeling motion captured from the RGB camera The human motion data extracted by pose estimation algorithm through images was usually used in this case, where it was inconvenient to collect highly supervised human body data. A typical work by Chao et al. [14] , proposed a PFNet as the first study on forecasting human dynamics from single RGB images. Jacob et al. [1] combined the strength of VAE to incorporate pose velocity and using conditional GAN [13] to generate future poses in pixel level. Wu et al. [18] adjusted the residual neural network to evaluate human pose from RGB frames and conducted pose forecasting for Virtual Reality use with a stack-LSTM forecasting structure . However, these works achieved pose forecasting mainly based on fixed visions without taking moving observation into account. Our settings involve apparent observer motion, which is different from these studies.

3. Proposed Method

We frame the task of human motion forecasting in egocentric videos as a sequence-to-sequence problem. The goal of our work is to predict future pose sequence of target person based on the observed pose sequences in the temporal domain. Suppose $\boldsymbol{p}_t$ denotes the human pose at time t, which contains N two-dimensional joints in the image space, *i.e.*, $\boldsymbol{p}_t = ((u_t^1, v_t^1), (u_t^2, v_t^2), ..., (u^N, v^N))^\top \in \mathbb{R}_+^{2N}$. We aim to forecast the target person' s future poses $\boldsymbol{q}_t \in \mathbb{R}_+^{2N}$ in the subsequent T_f time steps with the observed poses $\boldsymbol{p}_t \in \mathbb{R}_+^{2N}$ in the past T_p time steps as input. This results in the input sequence and the output sequence as bellow:

$$\boldsymbol{P}_{in} = \left(\boldsymbol{p}_{t-T_p+1} \ \cdots \ \boldsymbol{p}_t\right) \in \mathbb{R}_+^{2N \times T_p} \tag{1}$$

$$\boldsymbol{P}_{out} = \left(\boldsymbol{q}_{t+1} \ \cdots \ \boldsymbol{q}_{t+T_f}\right) \in \mathbb{R}_+^{2N \times T_f} \tag{2}$$

Our approach is illustrated in Figure 2. We start with pose estimation on series of video frames to extract body joints. Then we separately modeling pose and scale-location granularities of human motion with two modules that are forecasted using encoder-decoder based architecture. These predictions finally are combined to construct the future human motion.

3.1 Modeling structural-temporal dependencies of body joints

The above part of Figure 2 shows our pose forecasting pipeline. Since human motion data is composed of high dimensional vectors with complicated structural dependencies inherent in body joints, it is crucial for the network to have capability of learning the local movements of the body limbs. In our work, we both explored RNN-based and CNN-based sate-of-the-art networks (RRNN [9] and CS2S [11]) with sampling based loss calculation to predict future pose sequence

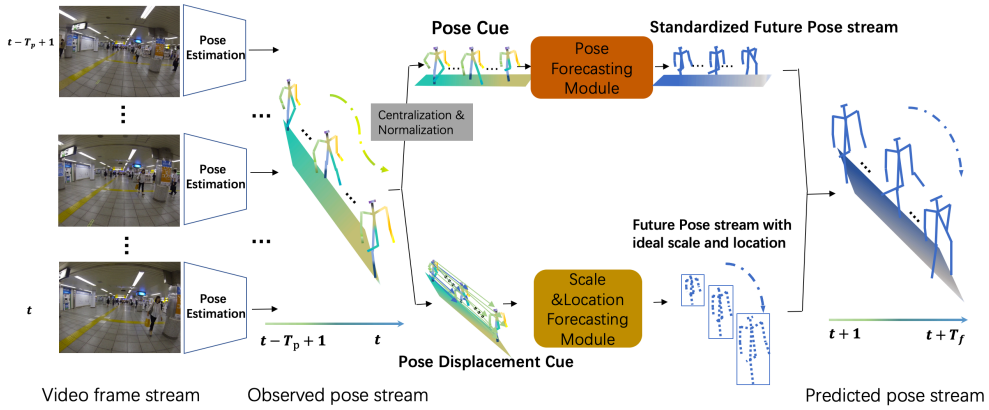

Figure 2 **Proposed Separate Modelling Approach**.
For forecasting future pose stream in egocentric videos. The model mainly consists of pose forecasting module and the scale-location forecasting module to respectively accept a pose stream and a pose-displacement stream for modelling motion dynamics and the variance of pose' s scale and location. With the combination of the outputs from the two modules, we generate the final predicted pose stream.

in the pose forecasting module. Figure 3 illustrates the architecture of these two networks. In the original formulation, the human motion sequence in egocentric settings is quite complex, consisting of various scales and positions information. To better modeling structural-temporal dependencies of body joints, we eliminate the complex scale and position granularities of human motion with a standardization process. By centralizing the human motion to the fixed position and normalizing the pose with the same scale, we convert the original motion sequence $\boldsymbol{p}_t$ to $\overline{\boldsymbol{p}}_t$ as the input to the pose forecasting network. The $\overline{\boldsymbol{p}}_t$ is derived from $\boldsymbol{p}_t$ (shown in (3)):

$$\overline{\boldsymbol{p}}_i = \frac{\boldsymbol{p}_i - \boldsymbol{l}_i^{stack}}{\boldsymbol{s}_i^{stack}} \in \mathbb{R}^{2N} \tag{3}$$

where $\boldsymbol{l}_i^{stack}$ is the stack of $l_i = (u_i^{neck}, v_i^{neck}) \in \mathbb{R}^2$ (the position of the pose at time step i, defined by the coordinates of the neck joint of the body) with the same dimensional of pose sequence for computation convenience, and $\boldsymbol{s}_i^{stack}$ is the stack of $s_i \in \mathbb{R}^2$ which represents the scale of the pose and is defined by the $L2$ distance between the neck joint and the waist joint.

Loss Metric The input sequence in the pose forecasting module is trained with the $L2$ loss between the predicted pose and the standardised ground truth pose. Thus the output pose sequence is also scale and position invariant.

$$\sum_{i=t+1}^{t+t_f} \left(\overline{\boldsymbol{q}}_{\boldsymbol{i}} - \overline{\boldsymbol{q}}_{\boldsymbol{GT_i}}\right) \tag{4}$$

where $\overline{\boldsymbol{q}}_{\boldsymbol{GT_i}}$ represents the standardized ground truth poses in the future time sequence.

3.2 Modeling temporal dependencies of body scales and positions

The below part of Figure 2 shows our scale-location forecasting pipeline. For predicting ideal scale and location of future pose sequences, we refer to the Conv-Deconv network proposed in [5], who firstly proposed a method on learning complicated dynamics of human trajectories in egocentric videos. Unlike the original multi-stream structure , in our work, we only use single-stream input which accepts the following pose-displacement cue.

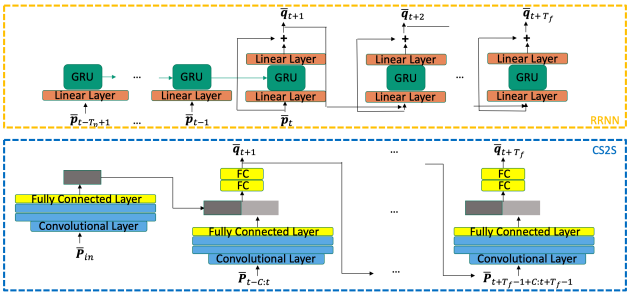

Figure 3 **Architecture of our pose forecasting networks**. The above is RNN-based network; the below is CNN-based network.

Pose displacement cue We define a novel cue called pose displacement cue as input in the scale-location forecasting module. The definition of pose-displacement stream shows as below:

$$\boldsymbol{P}_{D_{in}} = \left(\boldsymbol{p}_{t-T_p+2} - \boldsymbol{p}_{t-T_p+1} \cdots \boldsymbol{p}_t - \boldsymbol{p}_{t-T_p+1}\right) \in \mathbb{R}_+^{2N \times T_p - 1} \tag{5}$$

As shown in (5), the pose-displacement stream represents the displacements between the pose joints at the initial time step and the body joints at the subsequent time steps. It contains the temporal variance of pose with joints, describing how pose changes positions and scales in 2D space. We let scale-location forecasting module to accept pose-displacement stream (with normalization to have zero mean and unit variance) to generate future scale and location as pose displacement prediction.

Loss Metric We calculate $L2$ loss between the predicted

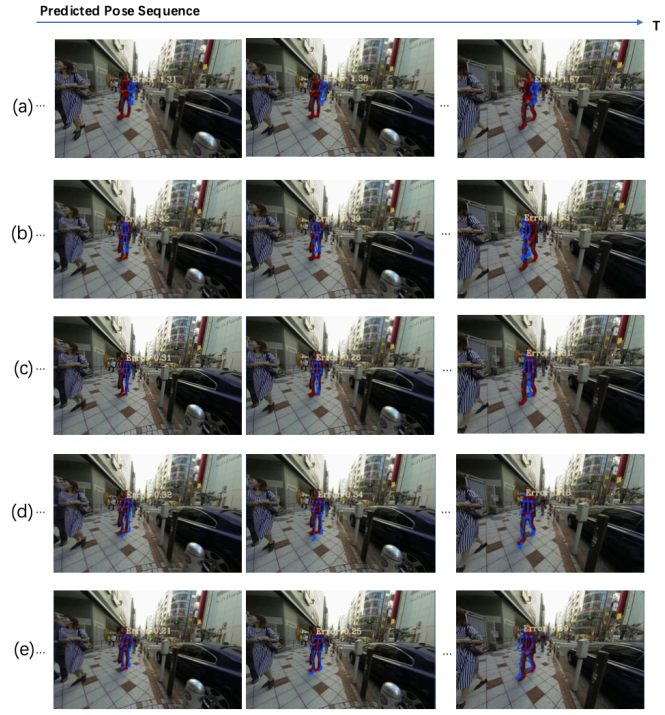

Figure 4 **Visualization results of predicted poses**. (a) – (e) respectively represents the visualization results of **RRNN**, **CS2S**, **Conv-Deconv**, **Ours 1**, and **Ours 2**. Red poses represent future ground truth pose sequence and blue poses are the predicted results.

pose and the ground truth pose with scale normalization. Details of using scale normalization are shown in Section 4.1

$$\sum_{i=t+1}^{t+t_f} \left(\widetilde{\boldsymbol{q}}_i / s^{\boldsymbol{stack}}_{\boldsymbol{q}_{GT_i}} - \boldsymbol{q}_{GT_i} / s^{\boldsymbol{stack}}_{\boldsymbol{q}_{GT_i}} \right) \tag{6}$$

where variables with GT represents features of ground truth poses in the future time sequence.

3.3 Pose recombination

During the pose recombination, we endow the standardized pose sequence output from pose forecasting module with the predicted scale and location. Since the pose displacement cue helps the model to capture the most active movements of body joints, which are contained in the direct output (a pose displacement vector) generated by scale-location forecasting module, we add this information with adjusting its scale and location to the standardized poses to strengthen the flexibility of body limbs.

4. Experiments

In this section, we adopt our separate training approach on a egocentric dataset to learn pedestrian motion dynamics and compare the results with prior works.

First-Person Locomotion Dataset [5] This dataset contains various locomotion of people captured from a wearable camera. It consists of 4.5 hours egocentric video frames with more than 5,000 observed people. In our experiment, we followed the data preprocessing procedure of work [5], including video undistortion, median filtering and pose imputation to relieve the erroneous or missing detection due to occlusion and motion blur. The final video frame after preprocessing is 25 fps. We conduct the prediction with the input and output length from 400 milliseconds to one second.

Millionseconds	400ms	600ms	800ms	1000ms
RRNN [9]	1.068	1.191	1.223	1.376
CS2S [11]	0.664	0.767	0.837	0.928
Conv-Deconv [5]	0.404	0.628	0.678	0.707
Ours 1	0.357	0.453	0.569	**0.628**
Ours 2	**0.343**	**0.450**	**0.567**	**0.628**

Table 1 **Comparisons to baseline methods.** Each score describes the scale-normalized $L2$ distance between bodyjoints of predicted pose and the ground truth. The error is represented in percentage with respect to the frame size of 960×1280 pixels based on the unit length of pose scale. For our proposed methods, **Ours 1** represents the approach with RRNN working in the pose forecasting module and conv-deconv working in the scale-location forecasting module; **Ours 2** represents CS2S working in the pose forecasting module.

4.1 Implementation

We train the pose forecasting network with the SGD optimizer with a learning rate 0.01. The scale-location forecasting network with ADAM optimiser with a learning rate 0.001. We use scale-normalised average displacement of body joints $\sum_{i=t+1}^{t+t_f} \left(\boldsymbol{q}_i / s^{\boldsymbol{stack}}_{\boldsymbol{q}_{GT_i}} - \boldsymbol{q}_{GT_i} / s^{\boldsymbol{stack}}_{\boldsymbol{q}_{GT_i}} \right)$ to evaluate the prediction performance. According to our empirical findings, the distance between the target person and the camera concerns with the scale of the motion that we observe from the video. For the pose closed to the camera, even though the output and the ground truth fits well, it would make the value of the distance between the body joints very large because the large target shown in the frame takes up more pixel units. We need to eliminate this influence brought by adding scale normalization for evaluating the prediction performance.

4.2 Performance Evaluation

Following previous convention of human motion forecasting works, we consider various prediction horizons. Table 1 compares prediction errors with single-module approaches as baseline methods. Since we tried both RNN and CNN based network working in the pose forecasting module, we show two results in the table. We can tell from the error score that our two models basically have same-level modeling performance and achieve much better prediction results comparing to the baselines whatever in the short-term or long-term prediction.

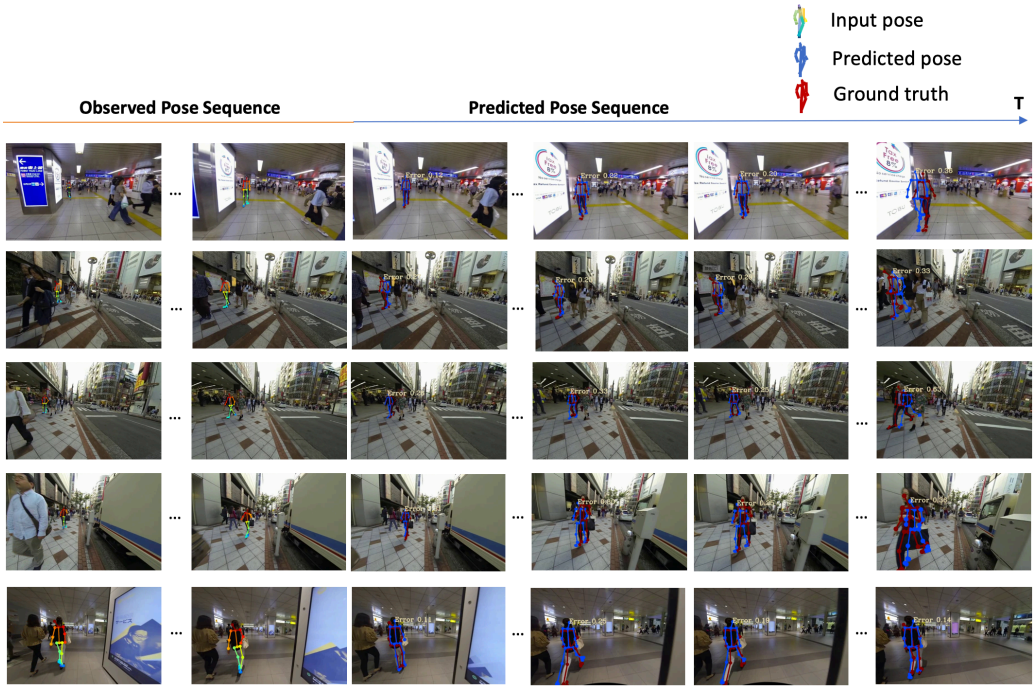

Figure 5 **Prediction results with Ours 2.**
The figure shows the result with 15 frames input and 15 frames output. Predicted pose sequence supports the effectiveness of our separate modeling approach that achieves learning both pose dynamics as well as scale and location dependencies temporally without mean convergence problem

	With Pose-displacement	With scale-location
Pose	**0.1787**	0.1841
Scale	**6.3407**	7.1237
Location	27.4471	**26.6623**

Table 2 **Separate evaluation on the pose, the scale and the location granularity.**
The pose error is based on $L2$ distance, calculated from standardized predicted poses and standardized ground truth poses, where we fix the location of the pose stream and normalize pose with its own scale in each frame. The scale error and location error are also measured by $L2$ distance with respect to the frame length and width.

We also provide qualitative results by visualising predicted pose sequence of the test data observed in egocentric visions, which are shown in Fig 4 and Fig 5.

4.3 Explorations on scale-location modeling

Since there are few works focusing on human scale-location forecasting in egocentric videos, we extensively test several approaches to ensure the effectiveness of our method with the use of pose displacement cue.

Approach 1 Input pose cue instead of pose-displacement cue in the scale-location forecasting module.
Approach 2 Following [5], we input the scale-location cue to obtain future scales and locations of human motion in the scale-location forecasting module.

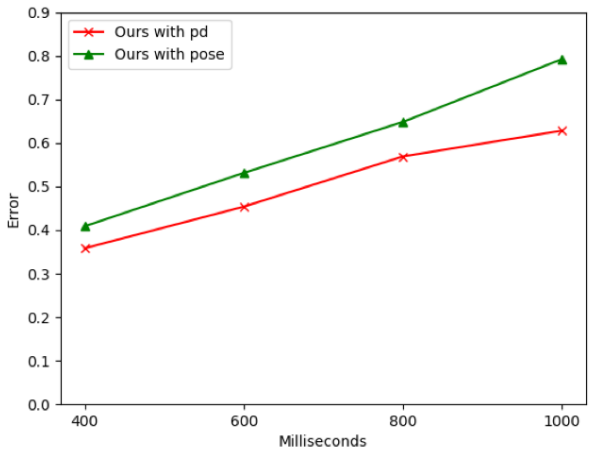

Figure 6 **Error comparisons between using pose displacement cue or pose cue.**
We observe obvious improvement of using pose displacement cue as input sequence in scale-location forecasting module.

Fig 6 shows the error curves of using pose sequence and pose displacement sequence as input stream in our scale location forecasting module with the prediction length from 400ms to 1s. The result reports that pose displacement sequence behaves more effectively than pose sequence on capture temporal dependencies of scales and locations in egocentric visions.

Figure 7 **Visualization of predicted pose sequence (with standardization).**
(a) represents the ground truth future pose sequence. (b) and (c) represents the results with the use of pose displacement cue or scale-location cue. The figure shows the result with the input 15 frames (red-black skeletons) and output 14 frames (green-blue skeletons).

Table 2 shows that our approach and Approach 2 basically have same-level quantitative performance on scale-location forecasting, however, with the use of pose displacement cue, we can obtain better pose prediction results. To further understand the difference, Fig 7 shows qualitative results of predicted pose sequence with body joints. Notice the walking pace of the poses, we can find out that the prediction with our approach is more fit to the ground truth. But for Approach 2, the walking pace in the output sequence is not that obvious. This is because the combination process of Approach 2 only endows the outputs from pose forecasting module with future scales and locations, the active movements of body joints are failed to be incorporated without the use of pose displacement.

5. Conclusions

To deal with the prediction problem of the egocentric human motion with complex pose, scale and location granularities, we propose a novel motion forecasting pipeline, modeling human pose and scale-position features separately with two forecasting modules, in which the final predictions are obtained by recombing the two outputs from the networks. We design a standardization process to remain the scale and position of the input pose sequence invariant for pose forecasting. And we present a pose displacement cue which includes the information about the variance of scale and the position of the observed motion sequence to learn their temporal dependencies. Considering the pose displacement cue also helps to capture the active movements of body joints, we propose to incorporate the generated results with the use of pose displacement to the output sequence from pose forecasting module during the pose recombination, to strengthen the flexibility of body limbs . The experiment shows the effectiveness of our approach and achieve a plausible-looking visualization results in egocentric view, much relieving from the mean convergence problem.

Reference

[1] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision* (*ICCV*), 2017.

[2] A.Tözeren. Human body dynamics: classical mechanics and human movement. In *Springer Science Business Media*, 1999

[3] C.Ionescu, D.Papava, V.Olaru, and C.Sminchisescu. Human 3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325-1339, jul 2014.

[4] S.L.Pintea, J. Gemert, and A.W.Smeulders. Motion prediction in static images. In *Proceedings of the European Conference on Computer Vision* (*ECCV*), 2014.

[5] T. Yagi, M. Karttikeya, Y. Ryo, and S. Yoichi. Future person localization infirst-person videos. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2018.

[6] H. Chiu, E. Adeli, B. Wang, and J.C. Niebles. Action-agnostic human pose forecasting. In *Winter Conference on Applications of Computer Vision* (*WACV*), 2019

[7] K.Fragkiadaki, S.Levine, P.Felsen, and J.Malik. Recurrent network models for human dynamics. In *IEEE International Conference on Computer Vision* (*ICCV*), 2015.

[8] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv*:1812.08008, 2018.

[9] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2017.

[10] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2016.

[11] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2018.

[12] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating the future by watching unlabeled video. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2016.

[13] M. Mirza and S. Osindero. Conditional generative adversarial nets. In *CoRR*, volume abs/1411.1784, 2014.

[14] Y.W.Chao, B. J.Yang, S.Cohen, and J.Deng. Forecasting human dynamics from static images. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2017.

[15] Wang, Hongsong Feng, Jiashi. VRED: A Position-Velocity Recurrent Encoder-Decoder for Human Motion Prediction. In *arXiv preprint arXiv*:1906.06514, 2019

[16] S. Ma, L. Sigal, and S. Sclaroff. Learning activity progression in lstms foractivity detection and early detection. In *IEEE Computer Vision and Pattern Recognition* (*CVPR*), 2016.

[17] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and Jose MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision* (*ECCV*), 2018.

[18] E. Wu and H. Koike. Future pose – mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera. In *Winter Conference on Applications of Computer Vision* (*WACV*), 2019.