Towards Human Pose Prediction using the Encoder-Decoder LSTM

Armin Saadat[1*] Nima Fathi[1*] Saeed Saadatnejad[2]

Sharif University of Technology (SUT), Tehran, Iran[1]
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland[2]

1. Introduction

Human pose prediction is defined as predicting the human keypoints locations for future frames, given the observed ones for past frames. It has numerous applications in various fields like autonomous driving. This task can be seen as a fine-grained task while human bounding box prediction deals with more coarse-grained information. The former has been investigated less and here, we modify one of the previously-used architectures of bounding box prediction to do a harder task of pose prediction in the SoMoF challenge. The results show the effectiveness of the proposed method in evaluation metrics.

2. Method

The proposed method is a sequence to sequence LSTM model based on [4], but it uses keypoints instead of bounding boxes and does not have an intention decoder. It takes as input the velocities and the positions of observed past joints and outputs the predicted velocities of the future joints, from which the future positions can be computed. As Figure 1 shows, the model encodes the position and the velocity of each person into a hidden layer which will be used as the initial state for the decoder. Using the encoded state, the decoder takes the velocity of the last observed frame as input and generates the predicted velocity for the first future frame which will be used as the input to the next LSTM cell. To train this model, the L1 loss between the predicted and ground-truth velocities is leveraged. Our method only uses keypoints, which makes it lightweight in comparison with methods using images.

3. Experiments

The evaluation metrics are VIM (Visibility-Ignored Metric) and VAM (Visibility-Aware Metric) introduced in [2]. The baselines are public benchmarks on SoMoF challenge: SC-MPF [1], TRiPOD [2] and Zero-Vel (zero velocity in prediction or keeping the last observation as prediction frames). The quantitative results of our experiments can be found in Table 1 for PoseTrack [3] and in Table 2 for 3DPW [5] datasets. The results show that our method outperforms the baselines in PoseTrack and gives good performance in 3DPW, especially in short-time horizon.

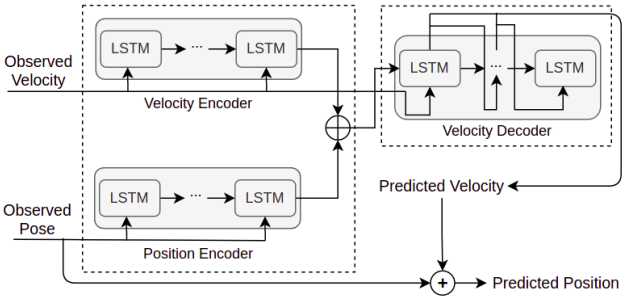

Figure 1: Network Architecture

Method	80 ms	160 ms	320 ms	400 ms	560 ms
SC-MPF	22.0/78.3	37.9/99.8	64.6/124.3	75.8/138.5	93.5/147.9
TRiPOD	15.2/30.0	26.7/49.6	48.1/80.3	58.6/93.3	71.1/110.3
Zero-Vel	13.1/26.5	24.0/45.1	43.3/72.9	52.1/83.8	65.6/97.3
Ours	**9.2/20.8**	**17.6/36.3**	**36.2/62.9**	**44.8/74.7**	**59.3/91.5**

Table 1: Results on PoseTrack (VIM/VAM)

Method	100 ms	240 ms	500 ms	640 ms	900 ms
SC-MPF	46.28	73.88	130.23	160.83	208.44
TRiPOD	30.26	51.84	**85.08**	**104.78**	146.33
Zero-Vel	29.35	53.56	94.52	112.68	**143.10**
Ours	**25.89**	**47.57**	86.39	106.65	148.28

Table 2: Results on 3DPW (VIM)

*Equal contribution
†https://github.com/Armin-Saadat/pose-prediction-autoencoder

4. Conclusion

We have presented an LSTM-based method for pose prediction for both 2D and 3D datasets. This method achieves better or comparative performance in evaluation metrics. We believe that for better predicting the future, interactions should be considered and we leave it for future studies.

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 1

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[3] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *CoRR*, abs/1710.10000, 2017. 1

[4] Smail Bouhsain, Saeed Saadatnejad, and Alexandre Alahi. Pedestrian intention prediction: A multi-task perspective. In *European Association for Research in Transportation (hEART)*, 2020. 1

[5] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 1