Future Person Localization in First-Person Videos: Supplementary Material

Takuma Yagi
The University of Tokyo
Tokyo, Japan
tyagi@iis.u-tokyo.ac.jp

Karttikeya Mangalam
Indian Institute of Technology
Kanpur, India
mangalam@iitk.ac.in

Ryo Yonetani
The University of Tokyo
Tokyo, Japan
yonetani@iis.u-tokyo.ac.jp

Yoichi Sato
The University of Tokyo
Tokyo, Japan
ysato@iis.u-tokyo.ac.jp

1. Data Statistics

Figure 1 presents frequency distributions of lengths of the tracklets extracted from First-Person Locomotion Dataset and Social Interaction Dataset [2]. These statistics revealed that most people appeared only for a short time period. In our experiments, we tried to pick out tracklets which were 1) longer enough to learn meaningful temporal dynamics and 2) observed frequently in the datasets to stably learn our network. These requirements resulted in our 50,000 samples consisting of the tracklets longer than or equal to 20 frames (*i.e.*, 2 seconds at 10 fps) and our problem setting of 'predicting one-second futures from one-second histories'.

Details of sample division: We first calculated the mean of scale normalized lengths between the left hip and the right hip for the target person. If this mean is less than 0.25 we categorized the clip as **Across**. In the remaining clips, we labeled each frame of the clip as either **Toward** if X-coordinate of the left hip is larger than that of the right hip and **Away** otherwise. If the number of frames labeled **Toward** is more than 75% of the total number of frames in the clip, the clip is categorized as **Toward** and as **Away** if it is less than 25%.

T_{prev}	T_{future}	Walking direction			
		Toward	Away	Across	Average
6	10	111.39	78.54	98.41	79.77
10	10	109.03	75.56	93.10	77.26
6	6	53.12	46.49	52.75	46.16
10	6	52.69	46.10	53.15	45.92

Table 1. **Different Input/Output Lengths**. Final Displacement Error (FDE) for various combinations of input (T_{prev}) and output (T_{future}) lengths.

T_{prev}	Walking direction			
	Toward	Away	Across	Average
Social LSTM [1]	299.81	222.30	236.48	223.16
Ours	**184.62**	**125.41**	**169.01**	**124.85**

Table 2. **Predicting Two-Second Futures**. Final Displacement Error (FDE) where T_{future} was set to $T_{\mathrm{future}} = 20$.

Method	Walking direction			
	Toward	Away	Across	Average
X_{in}	136.43	**124.10**	117.56	127.40
$X_{in} + E_{in}$	136.52	124.22	115.00	127.28
$X_{in} + P_{in}$	133.10	124.57	114.80	125.78
Ours ($X_{\mathrm{in}} + E_{\mathrm{in}} + P_{\mathrm{in}}$)	**131.94**	125.48	**112.88**	**125.42**

Table 3. **Ablation Study on Social Interactions Dataset [2]**. Final displacement error (FDE) for various combination of input features. Notations were the same as those of Table 2.

2. Additional Results

2.1. Other Choices of Input/Output Lengths

In our experiments, we fixed the input and output lengths $T_{\mathrm{prev}}, T_{\mathrm{future}}$ to be $T_{\mathrm{prev}} = T_{\mathrm{future}} = 10$. Table 1 shows how performances changed for other choices of T_{prev} and T_{future}. Overall, longer input lengths led to better performance ($T_{\mathrm{prev}} = 6$ vs. 10). Also, predicting more distant futures becomes more difficult ($T_{\mathrm{future}} = 10$ vs. 6). To receive shorter inputs, we applied 1-padding to the first and second convolution layer in each stream.

We also compared our method against Social LSTM [1] on the task of predicting two-second futures (*i.e.*, $T_{\mathrm{future}} = 20$) in Table 2. We confirmed that our method still worked

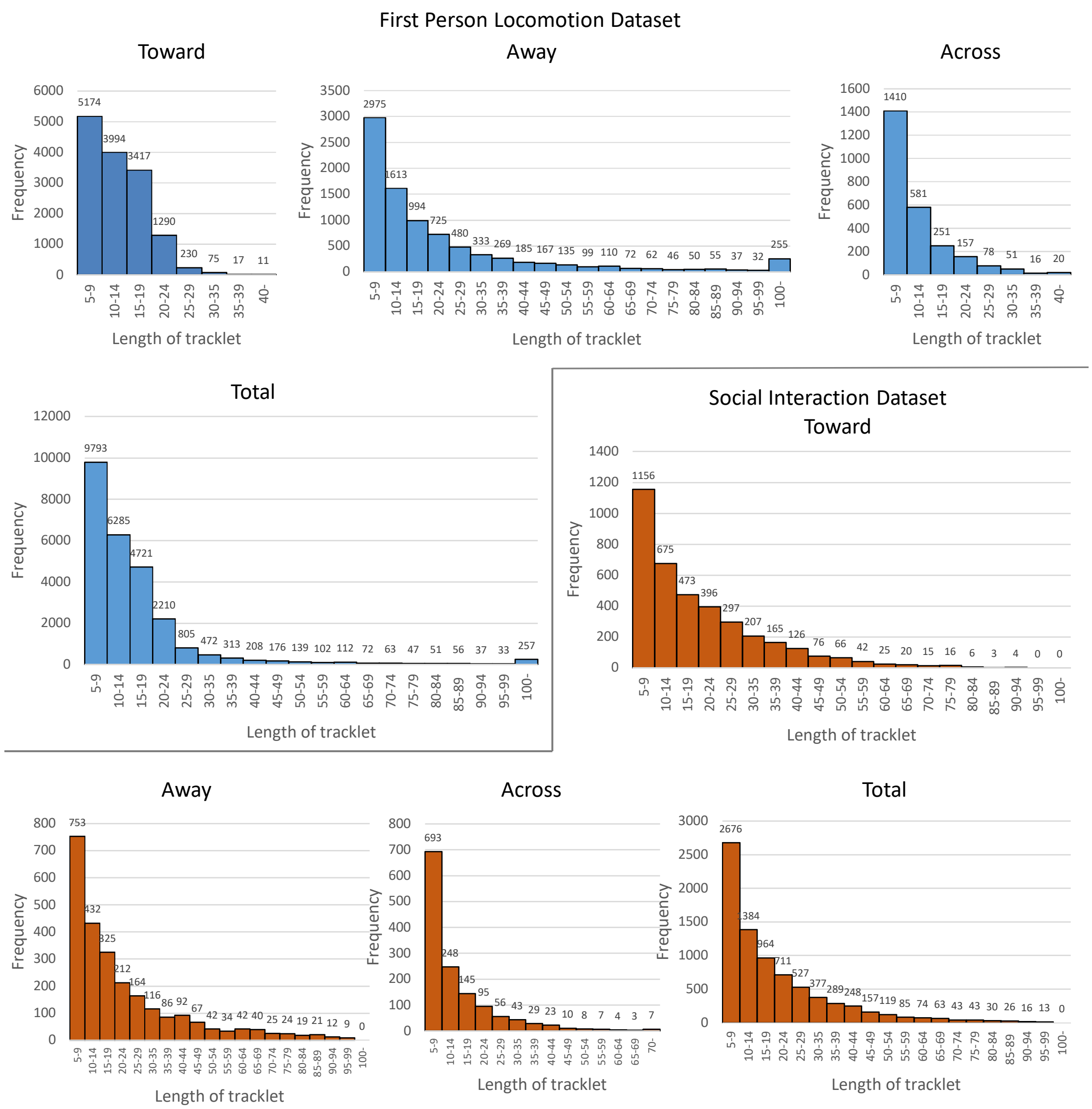

Figure 1. **Distributions of Tracklet Lengths**. Frequency distributions of various lengths of tracklets extracted from First-Person Locomotion Dataset and Social Interaction Dataset [2] for three walking directions and the entire database, respectively.

well on this challenging condition. To generate 20 frame prediction, we changed the kernel size of the deconvolution layers of 3, 3, 3, 3 to 3, 5, 7, 7.

2.2. Other Visual Examples

Figure 2 shows additional visual examples of how our method, as well as several baselines, predicted future locations of people.

2.3. Ablation Study on Social Interaction Dataset

We performed an ablation study on Social Interaction Dataset [2] in Table 3. While we computed ego-motion based on optical flows, the combination of ego-motion and pose cues contributed to performance improvements.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[2] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233, 2012.

Figure 2. **Qualitative Examples of Future Person Localization on First Person Locomotion Dataset**. (Row 1) Even though input sequence is almost static, our model is able to capture the left turn caused by the wearer's ego-motion. (Row 2, 3) In the input sequence, the target is changing the pose to move right. While compared model fails to predict because of being agnostic to the pose information, our model produces a better prediction. (Row 4) The behavior with respect to complicated ego-motion. In the input sequence, the wearer is turning left to avoid other pedestrians. However, in the future frames, the wearer moves to the opposite side to avoid contact with the target. In this case, our prediction is perturbed due to ego-motion and predicts worse than Social LSTM. (Row 5) Our model works well both in outdoor scenes as well as indoor scenes.