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%.

2. Additional Results

2.1. Other Choices of Input/Output Lengths

In our experiments, we fixed the input and output lengths $T_{prev}, T_{future}$ to be $T_{prev} = T_{future} = 10$. Table 1 shows how performances changed for other choices of $T_{prev}$ and $T_{future}$. Overall, longer input lengths led to better performance ($T_{prev} = 6$ vs. 10). Also, predicting more distant futures becomes more difficult ($T_{future} = 10$ 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., $T_{future} = 20$) in Table 2. We confirmed that our method still worked.
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


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.