Master Thesis

Future Person Localization in First-Person Videos
(一人称視点映像における人物位置予測)

Takuma Yagi

The University of Tokyo
Department of Information and Communication Engineering
Graduate School of Information Science and Technology

Advisor: Prof. Yoichi Sato

Abstract

Assistive technologies are attracting increasing attention as a promising application of first-person vision—which aims to extract visual knowledge through wearable cameras. Since first-person vision techniques can capture the world much like our eyes, it could be used as a system that perceives the world around the wearer and assists them to decide on what to do next. In this thesis, we focus on assisting a user to navigate in crowded spaces. Notably, we study a new problem of future person localization task—to predict the future position of a pedestrian appearing in first-person videos.

We made two main observations: (1) the wearer's ego-motion is observed in the form of global motion of the first-person video (2) The pose of a person indicates how that the person is moving and will be located in the future. We propose a prediction framework of a multi-stream convolutional neural network which takes pose and ego-motion information as inputs. By capturing the interaction between the wearer's ego-motion and the pose of the target person, our proposed method enables future person localization in the scenario of both the wearer and the target person are walking.

To validate the effectiveness of our method, we constructed a new dataset of first-person videos called First-Person Locomotion (FPL) dataset which captured various walking scenes in crowded places. Experimental results showed that proposed pose feature and ego-motion feature contributes to prediction performance in a complementary manner. We confirmed that our proposed method predicts one-second future more accurate than the previous method designed for fixed-view videos on our new dataset as well as on public social interaction dataset.

Keywords: first-person vision, future person localization, convolutional neural network, assistive vision

Contents

List of Figures

List of Tables

1. Introduction

1.1. Overview

Assistive technologies are attracting increasing attention as a promising application of *first-person vision*—computer vision using wearable cameras such as Google Glass and GoPro HERO. Much like how we use our eyes, first-person vision techniques can act as an artificial visual system that perceives the world around camera wearers and assist them to decide on what to do next. Recent work has focused on a variety of assistive technologies such as blind navigation [53, 89], object echo-location [87], and personalized object recognition [40].

In this work, we are particularly interested in helping a user to navigate in crowded places with many people present in the user's vicinity. Consider a first-person video stream that a user records with a wearable camera. By observing people in certain frames and predicting how they move subsequently, we would be able to guide the user to avoid collisions. As a first step to realizing such safe navigation technologies in a crowded place, this work proposes a new task that predicts locations of people in future frames, *i.e.*, *future person localization*, in first-person videos as illustrated in Figure 1.1[1].

To enable future person localization, this work makes two key observations. First, ego-motion of a camera wearer is observed in the form of global motion in first-person videos. This ego-motion should be incorporated into the prediction framework as it greatly affects future locations of people. For example, if a camera wearer is moving forward, apparent vertical locations of people in the first-person video will be moving down accordingly. Moreover, if the camera wearer is walking towards other people, they would change walking direction slightly to avoid a collision. This type of interacting behaviors would also affect the future locations of people.

Another key observation is that the pose of a person indicates how that person is moving and will be located in the near future. First-person videos can be used effectively to get access to such pose information as they often capture people up-close.

Based on these key observations, we propose a method to predict the future locations of a person seen in a first-person video based on poses, scales, and locations of the person in the present and past video frames and ego-motion of the video (also refer to Figure 1.1). Specifically, we develop a deep neural network that learns the history of the above cues in several previous frames and predicts locations of the target person in the subsequent

[1]Parts of faces in the paper were blurred for preserving privacy.

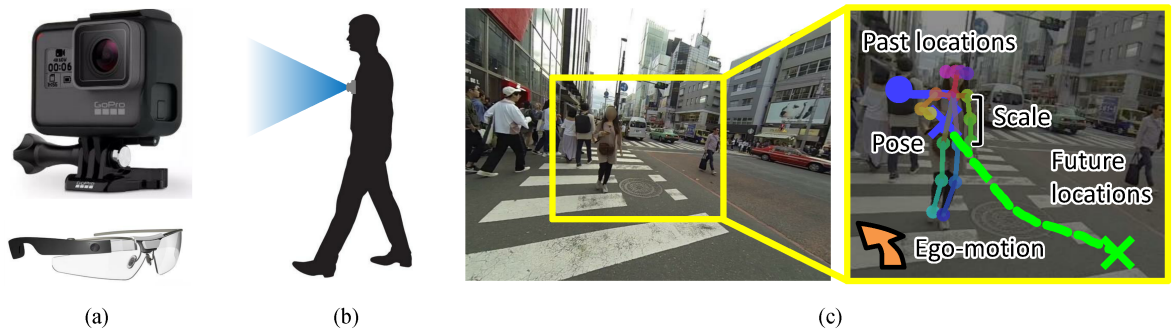

Figure 1.1.: Future person localization. (a) An example of a wearable camera (upper: GoPro HERO, lower: Google glass.) (b) Camera configuration used in this work. A wearable calera is mounted on the wearer's chest. (c) Given a first-person video of a certain target person, our network predicts where the target person will be located in the future frames based on the poses and scales of the person as well as the ego-motions of the camera wearer.

future frames. A convolution-deconvolution architecture is introduced to encode and decode temporal evolution in these histories.

To validate our approach, we develop a new dataset of first-person videos called First-Person Locomotion (FPL) Dataset. The FPL Dataset contains about 5,000 people seen at diverse places. We demonstrate that our method successfully predicts future locations of people in first-person videos where state-of-the-art methods for human trajectory prediction using a static camera such as [1] fail. We also confirmed a promising performance of our method on a public first-person video dataset [20].

1.2. Contributions

The main contributions of this thesis are summarized as follows:

- We introduce a new task of future person localization in first-person videos. To the best of our knowledge, this is the first work to predict the future position of a person from a single first-person video. By capturing the interaction between the wearer's ego-motion and the pose of the target person, our proposed method enables future person localization in the scenario that both the wearer and the target person are walking.
- We present a multi-stream convolutional neural network which takes ego-motions of the video, poses, scales, and locations of the person appearing in the present and past video frames. The network predicts the future locations where the person would be seen from the wearer in subsequent frames.
- We collect a new First-Person Locomotion (FPL) dataset which contains diverse

walking behavior in crowded places. We automatically generate the trajectories using the latest human pose estimation method and person tracking.

1.3. Thesis Outline

In Chapter 2, we review the recent works on first-person vision and trajectory prediction. In Chapter 3, we introduce our proposed method. We first formulate the problem of future person localization and then explain the intuition and implementation of three salient cues (location-scale, pose, and ego-motion cues) used as inputs to the network. In Chapter 4, we introduce the newly collected First-Person Locomotion (FPL) dataset and explain the sample generation procedure. In Chapter 5, we evaluate the effectiveness of our method on the FPL dataset and the social interaction dataset. Chapter 6 summarizes this thesis with possible future directions. In Appendix, we provide additional qualitative results and the dataset statistics as supplementary information.

2. Related Work

In this chapter, we review the important works on first-person vision and trajectory prediction.

2.1. First-Person Vision

First-person vision covers computer vision techniques using wearable cameras such as Google Glass and GoPro HERO [41]. In contrast to traditional surveillance setting with fixed, third-person cameras, first-person vision offers additional information about the camera wearer—their action and interaction with other people and objects. A typical problem setting involving first-person vision is to recognize activities of camera wearers. Recently, some work has focused on activity recognition [21, 55, 58, 73], activity forecasting [19, 24, 69, 77], person identification [34], gaze anticipation [101] and grasp recognition [7, 14, 54, 82].

This research can be regarded as a category of *second-person vision*—predicting the activity of individuals appearing in first-person videos. The configuration of first-person videos enables efficient recognition of facial expression, head/body movements, and other non-verbal behavior compared to third-person videos. Although few works focused on this topic, detection of social interactions [20, 70, 26, 4, 3], action recognition [81, 80, 100], joint attention detection [45, 37] and facial attribute learning [95] are studied.

2.2. Trajectory Prediction

We are interested in predicting the short-term behavior of a pedestrian. In this section, we review the previous work modeling pedestrian dynamics mainly on short-term interaction such as collision avoidance.

Modeling Pedestrian Dynamics

The task of predicting future locations of people itself has been studied actively in crowd analysis, computer vision and robotics [86, 65, 22, 91]. Early attempts date back to 1970s [32, 31] analyzing the holistic dynamics of walking people in large crowds.

The seminal work of social force model [30] assumes a walking person as a particle-like instance which would subject to attraction and repulsion forces called 'social forces.' Four types of hand-crafted force functions are introduced, and its effectiveness is verified by simulation [93], abnormal crowd behavior detection [61] and tracking [56]. Their approach can predict future trajectory typically assuming a constant velocity Kalman filter model. However, the model has no learnable hyperparameters and is only able to forecast short-term collision avoidance with careful initialization. Regarding the advance in optimization techniques and computational power, the idea of learning pedestrian dynamics from real-world data become popular.

Several works focus on learning social factors of pedestrian dynamics from real-world video data, using an energy-based model. They mainly focus on reducing the search space of data association in multi-target tracking [71, 84, 72]. Scovanner and Tappen [84] introduce an energy-based model which can learn the parameters of the model from real-world pedestrian movement. Pellegrini *et al.* [71] propose *Linear Trajectory Avoidance* which predicts the short-term future taking future destination and collision avoidance into account.

Alahi *et al.* [2] propose Social Affinity Maps which models the motion affinities of neighbouring pedestrians with Origin-Destination prior. However, its representation was rather hand-crafted and not fully data-driven. Recent methods have tried to model social interaction directly from trajectories, using Deep Neural Networks [99, 1, 50].

Capturing grouping behavior among a small group of people is one of the important topics in pedestrian behavior modeling. Choi *et al.* [17] present a new spatio-temporal local descriptor represented as a histogram of surrounding people and their poses, which captures the social relationship between people. Ge *et al.* [27] present an agglomerative clustering method to detect small groups from pedestrian trajectories. Yamaguchi *et al.* [97] propose a learnable energy-based model which considers potential destination and group behavior in addition to social force terms. Leal *et al.* [49] formulate multiple object tracking as a minimum-cost network flow problem and propose using grouping behavior determined by inter-person distance.

Beyond modeling trajectories, leveraging the information of environmental cues such as roads and curbs, and its interaction between pedestrians is also important. Kitani *et al.* [47] model the preference of walkable area in a surveillance setting. Ballan *et al.* [6] adopt a knowledge transfer scheme to transfer the interaction between the target person and their surrounding environment to a novel scene never observed before. Huang *et al.* [36] introduce a deep learning-based model of inferring the reward map of the environment.

Another promising direction is to consider the additional information extracted from the pedestrians themselves. We humans do not only interact using the spatial relationship but also use non-verbal signal extracted from themselves. For example, we can easily infer the agility of a person from their appearance—whether they are old or young, male or female, for example. Head pose and gaze also provide strong cue where the person would move next. Huang *et al.* [36] train a convolutional neural network to predict

the target's orientation to capture its temporal contexts. Ma *et al.* [59] leverage the visual appearance of the pedestrian themselves such as age and gender. Su *et al.* [85] show that predicting future gaze direction and joint attention significantly improves future localization performance. Hasan *et al.* [29] show that multi-task formulation of trajectory prediction and future head pose estimation improves forecasting performance, supporting the effectiveness of using pose information.

As a related task, a scene-specific situation of road crossing has been studied in intelligent vehicle domain. Bonnin *et al.* [13] leverage contextual information such as zebra crossing. Keller *et al.* [43] compare various models of path prediction in a road crossing scenario. Kooij *et al.* [48] model pedestrian situational awareness, situation criticality and spatial layout of the environment into a single Dynamic Bayesian Network (DBN).

Although beyond our focus, it is valuable to mention modeling the long-term goal of a pedestrian. In surveillance camera setting, Xie *et al.* [96] assume that pedestrians will be attracted to one's potential destination and propose a model of predicting walkable area and the expected destination in a relatively large environment (scores of meters square). Karasev *et al.* [42] model long-term motion policy of a pedestrian using a Markov decision process framework.

Prediction Framework

Trajectory prediction methods can also be categorized by the prediction framework they use. Antonini *et al.* [5] modeled pedestrian behavior as Discrete Choice models. Early methods assumed trajectory problem as an extension of linear dynamical systems such as Kalman Filter [56], Extended Kalman Filter [71, 83] and Dynamic Bayesian Network [48]. Gaussian Process-based formulation are also proposed [94, 88, 18, 91, 92, 43]. Some methods adopted global optimization approach which optimizes the parameters of the model using the entire training data [97, 71, 78]. Given both locations of start and destination, work based on inverse reinforcement learning (IRL) [67, 104] can forecast in-between paths [47, 62, 42, 51, 59]. Although IRL-based approach can naturally model human behavior as a Markov decision process model, it typically cannot handle high-dimensional features while there exist the notable exception of the use of deep Q-learning [12].

More recently, deep learning based approach, directly solving the trajectory prediction problem as a supervised learning formulation, has appeared. Most methods use either recurrent neural networks (RNNs) [1, 50, 11, 29] or convolutional neural networks (CNNs) [36, 99]. Rehder *et al.* [75] adopted a planning-based approach which predicts the expected destination and its path at once. The model was represented by a single network combining recurrent mixture density networks and fully-convolutional networks.

Alahi *et al.* [1] proposed a Long Short-Term Memory (LSTM) based model which learns the interaction between multiple walking people. In contrast to the previous approaches manually defining pedestrian interaction using known heuristics [30], they proposed an end-to-end model which learns human-human interaction in a fully data-driven manner.

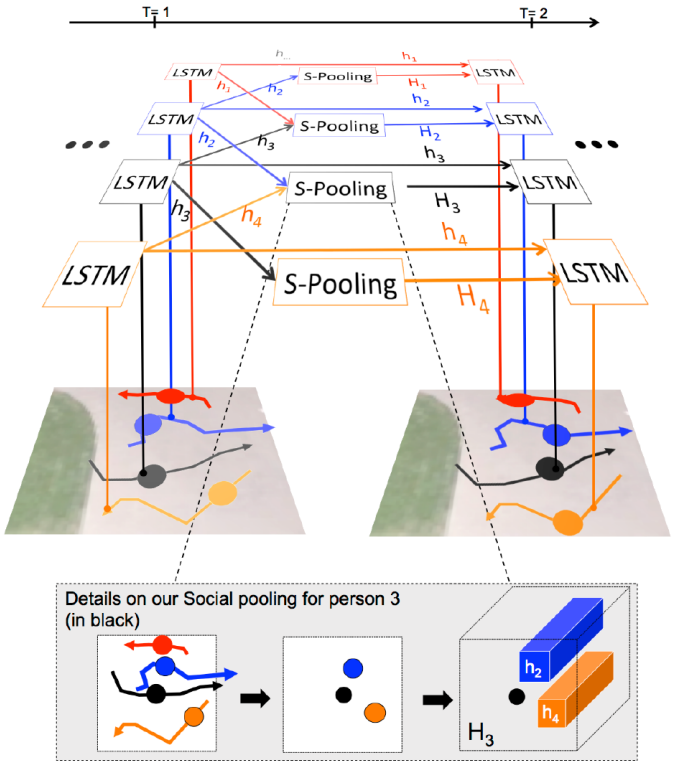

Figure 2.1.: Overview of the Social LSTM model. Each trajectory has a separate LSTM. When calculating the hidden state of the next timestep, the social pooling layer squashes the hidden states of the surrounding people into a fixed size vector (The image is taken from [1].)

The authors proposed a new "Social" pooling layer to share neighboring people's hidden state inside the LSTM. The social pooling layer squash the hidden states of the neighboring people into a fixed-size vector (Figure 2.1). The resulting model can learn the interaction between pedestrians with minimal heuristics.

These methods are, however, not designed to deal with first-person videos where significant ego-motion affects the future location of a certain person. Also, while the fixed camera setting assumed in these methods can suffer from oblique views and limited image resolutions, the egocentric setting provides strong appearance cues of people. Our method focuses on one-by-one interaction between the camera wearer and their surrounding pedestrian from the first-person view, utilizing ego-motion, scale and pose information which is unique in egocentric setting. Such cues are previously not considered because of the lack of person resolution taken from third-person cameras and less reliability of human pose estimation techniques.

2.3. Datasets in Trajectory Prediction

Dataset	Year	#People	#Scenes	Duration	View	Target
ETH [52]	'07	750	2	5 min	Top-down	Pedestrian
UCY [71]	'09	786	3	25 min	Top-down	Pedestrian
Edinburgh [60]	'09	95998	1	120 days	Top-down	Pedestrian
Town Centre [9]	'11	230	1	N/A	Bird's eye	Pedestrian
VIRAT [68]	'11	4021	11	8.5 hrs	Bird's eye	Pedestrian
Social Interaction [20]	'12	N/A	N/A	N/A	First-person	Pedestrian
Central Station [98]	'15	12600	1	1 hr	Bird's eye	Pedestrian
Stanford Drone [78]	'16	11216	8	8.5 hrs	Top-down	Pedestrian, car, bus, biker, skater
EgoMotion [69]	'16	N/A	26	9.1 hrs	First-person	Pedestrian
Basketball [85]	'17	N/A	N/A	10.5 hrs	First-person	Basketball player
Continuous Activity [77]	'17	N/A	17	N/A	First-person	Pedestrian
Ours	'18	5164	N/A	4.5 hrs	First-person	Pedestrian

Table 2.1.: A comparison of the datasets used in trajectory prediction.

Responding to the increasing demand for data-driven approaches, many datasets for trajectory prediction had been proposed. A summary is given in Table 2.1. Heavily used datasets such as ETH [52] and UCY [71] was recorded in a typical surveillance setting from a fixed camera. EgoMotion dataset [69] and First-Person Continuous Activity dataset [77] are designed for predicting the future action of the wearer itself, and they do not contain other pedestrians within the videos. Basketball dataset [85] collects various movement of basketball players from the first-person view, although, its dynamics are different from natural walking scenarios. Social interactions dataset [20] is the only dataset which collects diverse walking people from head-mounted cameras, and we use this dataset for additional analysis. However, their work's objective is to capture social interaction thus not all scenes contain walking behavior.

Due to the lack of walking behavior captured in first-person videos, we collect a new dataset which contains diverse walking behavior, recorded by a chest-mounted camera. This dataset contains more than 5,000 people mainly walking in crowded places and includes various types of interactions between the camera wearer. We explain the details of this dataset in Chapter 4.

2.4. Trajectory Prediction in First-Person Vision

Few works focus on the problem of predicting the camera wearer's future movement from first-person videos. Park *et al.* [69] first consider this task using an RGB-D first-person observation. Thanks to the depth information, they propose an EgoRetinal map which encodes the information of surrounding obstacles. An EgoRetinal map has a ground-plane like coordinate naturally capturing the perspective (3D) effect of the scene and 2D visual appearance. However, this representation heavily relies on geometric cues and requires computationally expensive stereo setting to predict the trajectory. Bokhari *et al.* [12] focus on the problem of forecasting longer time horizons (*e.g.,* several minutes)

using deep-Q learning, achieving path prediction. Rhinehart *et al.* [77] propose an inverse reinforcement learning based method which predicts the future location where the camera wearer would go next in an indoor activity scenario. Bertasius *et al.* [10] present a model which generates 3D location and head motion trajectory taking a single RGB first-person image as an input. They assume a one-on-one basketball game scenario and show that verifying whether the pre-defined goal is accomplished by the generated trajectories or not is important to generate plausible trajectories of goal-oriented behavior. Concurrent with our work, Bhattacharyya *et al.* [11] study the problem of pedestrian trajectory prediction from vehicle onboard cameras, different setting but partially share its nature. They proposed a two-stream architecture which predicts future bounding box estimate and vehicle ego-motion (speed and steering angle) at once.

To the best of our knowledge, this work is the first to address the task of predicting future locations of people in first-person videos. Our task is different from *egocentric future localization* [69] that predicts where the camera wearers will be located in future frames. One notable exception is the recent work by Su *et al.* [85]. They present a model to predict the future location and gaze of basketball players from multiple first-person videos. Their key idea is to reconstruct the entire 3D scene using a 3D reconstruction technique. They adopted a group trajectory selection scheme which considers social cue such as gaze and proximity. Although they proposed a method to predict future behaviors of basketball players in first-person videos, their method requires *multiple* first-person videos to be recorded collectively and synchronously to reconstruct accurate 3D configurations of camera wearers. This requirement of multiple cameras is in contrast to our work (*i.e.*, using a single camera) and does not fit for assistive scenarios where no one but the user on assistance is expected to wear a camera. To achieve this, we formulate the prediction problem within the view of the wearer—in image coordinates.

3. Proposed Method

In this chapter, our proposed method of future person localization using salient first-person cues is introduced.

3.1. Problem Formulation

We first formulate the problem of predicting future locations of people in first-person videos. Consider a certain *target* person seen in a current frame of a first-person video recorded on the street. Our goal is to predict where the target person will be seen in subsequent frames of the video based on the observation up to the current frame. Formally, let $\boldsymbol{l}_t \in \mathbb{R}^2_+$ be the 2D location of the person in the frame t. As illustrated in Figure 3.1, we aim to predict the person's relative locations in the subsequent T_{future} frames from the current one at t_0 (red frames in the figure), based on observations in the previous T_{prev} frames (blue ones). This could be written as below:

$$\boldsymbol{L}_{\text{out}} = \left(\begin{array}{cccc} \boldsymbol{l}_{t_0+1} - \boldsymbol{l}_{t_0} & \boldsymbol{l}_{t_0+2} - \boldsymbol{l}_{t_0} & \dots & \boldsymbol{l}_{t_0+T_{\text{future}}} - \boldsymbol{l}_{t_0} \end{array} \right) \in \mathbb{R}^{2\times T_{\text{future}}}.$$

The key technical interest here is what kind of observations can be used as a salient cue to better predict $\boldsymbol{L}_{\text{out}}$. Based on the discussions we made in Chapter 1 (also refer to Figure 3.1), we focus on c-1) locations and c-2) scales of target people, d) ego-motion of the camera wearer, and e) poses of target people as the cues to approach the problem. In order to predict future locations from those cues, we develop a deep neural network that utilizes a multi-stream convolution-deconvolution architecture shown in Figure 3.2. Input streams take the form of fully-convolutional networks with 1-D convolution filters to learn sequences of the cues shown above. Given a concatenation of features provided from all input streams, the output stream deconvolutes it to generate $\boldsymbol{L}_{\text{out}}$. The overall network can be trained end-to-end via back-propagation. In the following sections, we describe how each cue is extracted and how they would be combined to improve prediction performance. Concrete implementation details and training strategies are discussed in Section 5.1.

3.2. Location-Scale Cue

The most straightforward cue to predict future locations of people $\boldsymbol{L}_{\text{out}}$ is their previous locations up to the current frame t_0. For example, if a target person is walking in a

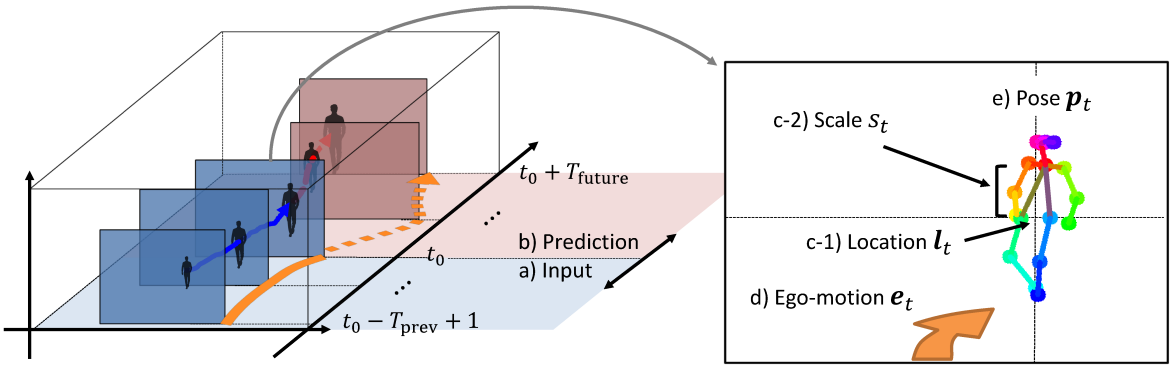

Figure 3.1.: Problem setting. Given a) T_{prev}-frames observations as input, we b) predict future locations of a target person in the subsequent T_{future} frames. Our approach makes use of c-1) locations and c-2) scales of target persons, d) ego-motion of camera wearers and e) poses of the target persons as a salient cue for the prediction.

certain direction at a constant speed, our best guess based on only previous locations would be to expect them to keep going in that direction in subsequent future frames too. However, visual distances in first-person videos can correspond to different physical distances depending on where people are observed in the frame.

In order to take into account this perspective effect, we propose to learn both locations and scales of target people jointly. Given a simple assumption that heights of people do not differ too much, scales of observed people can make a rough estimate of how large movements they made in the actual physical world. Formally, let $\boldsymbol{L}_{\text{in}}$ be a history of previous target locations shown as below:

$$\boldsymbol{L}_{\text{in}} = \begin{pmatrix} \boldsymbol{l}_{t_0-T_{\text{prev}}+1} & \dots & \boldsymbol{l}_{t_0} \end{pmatrix} \in \mathbb{R}^{2\times T_{\text{prev}}}.$$

Then, we extend each location $\boldsymbol{l}_t \in \mathbb{R}^2_+$ of a target person by adding the scale information of that person $s_t \in \mathbb{R}_+$, *i.e.*, $\boldsymbol{x}_t = (\boldsymbol{l}_t^\top, s_t)^\top$. Then, the *'location-scale'* input stream in Figure 3.2 learns time evolution in $\boldsymbol{X}_{\text{in}}$ and the output stream generates $\boldsymbol{X}_{\text{out}}$ as shown in below:

$$\boldsymbol{X}_{\text{in}} = \begin{pmatrix} \boldsymbol{x}_{t_0-T_{\text{prev}}+1} & \dots & \boldsymbol{x}_{t_0} \end{pmatrix} \in \mathbb{R}^{3\times T_{\text{prev}}},$$

$$\boldsymbol{X}_{\text{out}} = \begin{pmatrix} \boldsymbol{x}_{t_0+1} - \boldsymbol{x}_{t_0} & \dots & \boldsymbol{x}_{t_0+T_{\text{future}}} - \boldsymbol{x}_{t_0} \end{pmatrix} \in \mathbb{R}^{3\times T_{\text{future}}}.$$

3.3. Pose Cue

One of the notable advantages of using first-person videos is the ability to observe people up-close. This makes it easier to capture what poses they take (*e.g.*, which directions

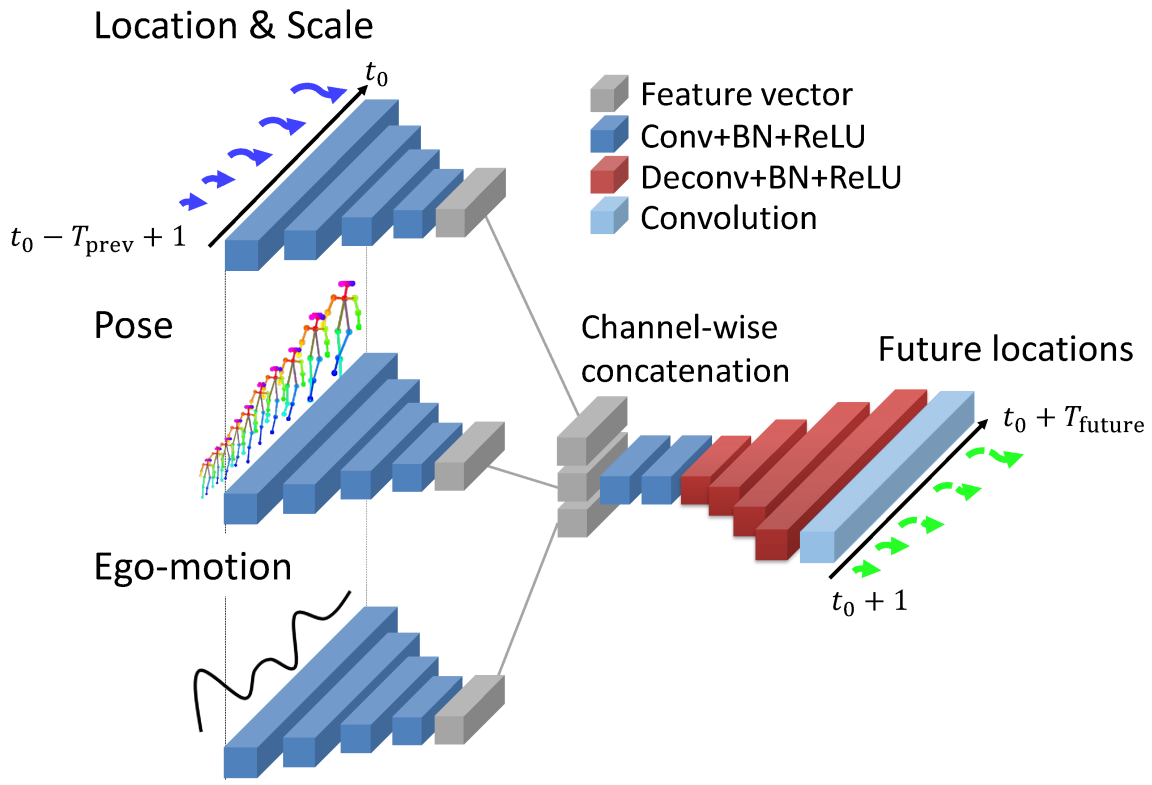

Figure 3.2.: Proposed network architecture. Blue and red blocks correspond to convolution and deconvolution layers followed by batch normalization and rectified linear unit (ReLU), respectively. The light blue block corresponds to a separate convolution layer. Gray blocks describe intermediate deep features.

they orient), which could act as another strong indicator of the direction they are going to walk along.

The *'pose'* stream in Figure 3.2 is aimed at encoding such pose information of target people. More specifically, we track temporal changes of several body parts of target people including eyes, shoulders, and hips as a feature of target poses. This results in an input sequence $\boldsymbol{P}_{\text{in}}$ as shown in below:

$$\boldsymbol{P}_{\text{in}} = \left(\begin{array}{ccc} \boldsymbol{p}_{t_0 - T_{\text{prev}} + 1} & \cdots & \boldsymbol{p}_{t_0} \end{array} \right) \in \mathbb{R}^{2V \times T_{\text{prev}}},$$

where $\boldsymbol{p} \in \mathbb{R}^{2V}$ is a $2V$-dimensional vector stacking locations of V body parts.

While the scale of the raw pose sequence differs by where the person appears, we only want to capture the difference among their poses—not their scale. Therefore $\boldsymbol{p}_t$was normalized by subtracting the absolute location $\boldsymbol{l}_t$ and divided by the scale s_t.

3.4. Ego-Motion Cue

While $\boldsymbol{X}_{\text{in}}$ explicitly describes how a target person is likely to move over time, the direct prediction of $\boldsymbol{X}_{\text{out}}$ from $\boldsymbol{X}_{\text{in}}$ is still challenging due to significant ego-motion present in first-person videos. More specifically, the coordinate system to describe each point $\boldsymbol{l}_t$ changes dynamically as the camera wearer moves. This makes $\boldsymbol{X}_{\text{in}}$ and $\boldsymbol{X}_{\text{out}}$ quite diverse depending on both walking trajectories of the target person and ego-motion of the camera wearer.

Moreover, ego-motion of camera wearers could affect how the target people move as a result of interactive dynamics among people. For instance, consider a case where a target person is walking towards the camera wearer. When the target person and the camera wearer notice that they are going to collide soon, they will explicitly or implicitly condition themselves to change their walking speed and direction to avoid the potential collision. Although some recent work has tried to incorporate such interactive behaviors into human trajectory prediction [1, 50, 59, 78], their approaches need all interacting people to be observed in a static camera view and cannot be applied directly to our case.

In order to improve future localization performance for first-person videos, we propose to learn how the camera wearer has been moving, *i.e.*, the ego-motion cue. Specifically, we first estimate the rotation and translation between successive frames. Rotation is described by a rotation matrix $\boldsymbol{R}_t \in \mathbb{R}^{3\times3}$ and translation is described by a 3D vector $\boldsymbol{v}_t \in \mathbb{R}^3$ (*i.e.*, x-, y-, z-axes), both from frame $t-1$ to frame t in the camera coordinate system at frame $t-1$. These vectors represent the local movement between the successive frames, however, does not capture the global movement along multiple frames. Therefore, for each frame t within the input interval $[t_0 - T_{\text{prev}} + 1, t_0]$, we accumulate those vectors to describe time-varying ego-motion patterns in the camera coordinate system at frame $t_0 - T_{\text{prev}}$:

$$\boldsymbol{R}'_t = \begin{cases} \boldsymbol{R}_t & (t = t_0 - T_{\text{prev}} + 1) \\ \boldsymbol{R}_{t-1}\boldsymbol{R}'_t & (t > t_0 - T_{\text{prev}} + 1) \end{cases}$$

$$\boldsymbol{v}'_t = \begin{cases} \boldsymbol{v}_t & (t = t_0 - T_{\text{prev}} + 1) \\ \boldsymbol{R}'^{-1}_t \boldsymbol{v}_t + \boldsymbol{v}'_{t-1} & (t > t_0 - T_{\text{prev}} + 1) \end{cases}$$

We form the feature vector for each frame by concatenating the rotation vector $\boldsymbol{r}'_t$ (*i.e.*, yaw, roll, pitch) converted from $\boldsymbol{R}'_t$ and $\boldsymbol{v}'_t$, resulting in a 6-dimensional vector $\boldsymbol{e}_t$. Finally, we stack them to form an input sequence $\boldsymbol{E}_{\text{in}}$ for the *'ego-motion'* stream shown in Figure 3.2:

$$\boldsymbol{e}_t = ((\boldsymbol{r}'_t)^\top, (\boldsymbol{v}'_t)^\top)^\top \in \mathbb{R}^6,$$

$$\boldsymbol{E}_{\text{in}} = \begin{pmatrix} \boldsymbol{e}_{t_0 - T_{\text{prev}}+1} & \dots & \boldsymbol{e}_{t_0} \end{pmatrix} \in \mathbb{R}^{6\times T_{\text{prev}}}.$$

3.5. Multi-Stream Convolution-Deconvolution Architecture

Now we are ready to combine the above cues into a single framework. When modeling time-series data, *recurrent* models are usually used. However, in this work, we chose to rely on *structured prediction* approach which predicts the entire sequence at once, using a feed-forward network. Specifically, we adopt the use of one-dimensional convolutional neural networks (1D-CNN) which learns a *temporal* filter. The reasons to take 1D-CNN architecture are two-fold: (1) it avoids the problem of error accumulation (2) it can capture multi time-scale temporal patterns.

The dynamics of the trajectory observed in first-person videos are more complicated than that of third-person videos since the camera is moving. Our videos taken from a chest-mounted camera contain oscillation patterns caused by locomotion and the corresponding trajectories also contains such shaky motions. In such cases, recurrent models can suffer from error accumulation effect [23]. Since we observed the same effect reported in [23], we decided to predict the entire future sequence directly.

Meanwhile, the justification of the use of 1D-CNN stems from the hierarchical structure of locomotion. Typical locomotion consists of rapid periodical motion while generating slow changes in global moving direction. When there are no obstacles ahead, the movement is smooth. However, once an event occurs, they might suddenly change their direction suddenly. To capture such hierarchical motion dynamics, a single time-scale recurrent connection from time t to $t+1$ is insufficient. One possible solution is to utilize multi-scale RNNs (*e.g.*, [16]). However, they only consider capturing two time-scales and partially resolve this issue. To overcome this issue without losing flexibility and efficiency, we propose to learn the short-term temporal patterns as a set of temporal convolution layers. Given feature sequence of length T_{prev}, we iteratively apply temporal convolution without padding. The length of the intermediate feature become shorter as we apply convolution. Consequently, weights in each layer can capture different time-scale. This also applies to the sequence generation phase, and it could be similarly represented as a set of deconvolution layers.

Another design choice left is how to fuse the information of different modalities. In this work, we separately train CNNs for each modality and then concatenate the features along channels. The following 1×1 convolution layers after concatenation learn the relationship between different modalities. We confirm that this simple strategy improves future localization performance.

4. First-Person Locomotion Dataset

To the best of our knowledge, most of the first-person video datasets comprise scenes where only a limited number of people are observed, *e.g.*, CMU Social Interaction Dataset [70], JPL Interaction Dataset [81], HUJI EgoSeg Dataset [74]. In this work, we introduce a new dataset which we call *First-Person Locomotion (FPL) Dataset*. The FPL Dataset consists of about 4.5 hours of first-person videos recorded by people wearing a chest-mounted camera and walking around in diverse environments. Some example frames are shown in Figure 4.1. The number of observed people is more than 5,000 in total. In this chapter, we explain the details about the dataset construction.

4.1. Data Collection

Camera configuration A single participant collected all the videos in this dataset. The participant wore a GoPro HERO3 camera on his chest (Figure 4.2). At the beginning of each session, the participant was asked to adjust the elevation angle of the camera that the horizon appears around the middle of the frame when he stands still. We recorded all the video data at 960p/100 fps. We set the frame rate higher so that higher shutter speed will be selected.

Recording We recorded all the videos at Tokyo, Japan. We selected the shooting location which contains enough number of pedestrians and has enough density so that natural interaction between pedestrians would occur—down-towns, stations, shopping street for example (see Figure 4.1 for visual example).

4.2. Data Preprocessing

In this section, we describe the details of the data preprocessing procedure as shown in Figure 4.3.

Video preprocessing Videos have been downsampled to 20 fps. We undistorted the frames using pre-calculated distortion coefficients and camera parameters using a chessboard pattern [102].

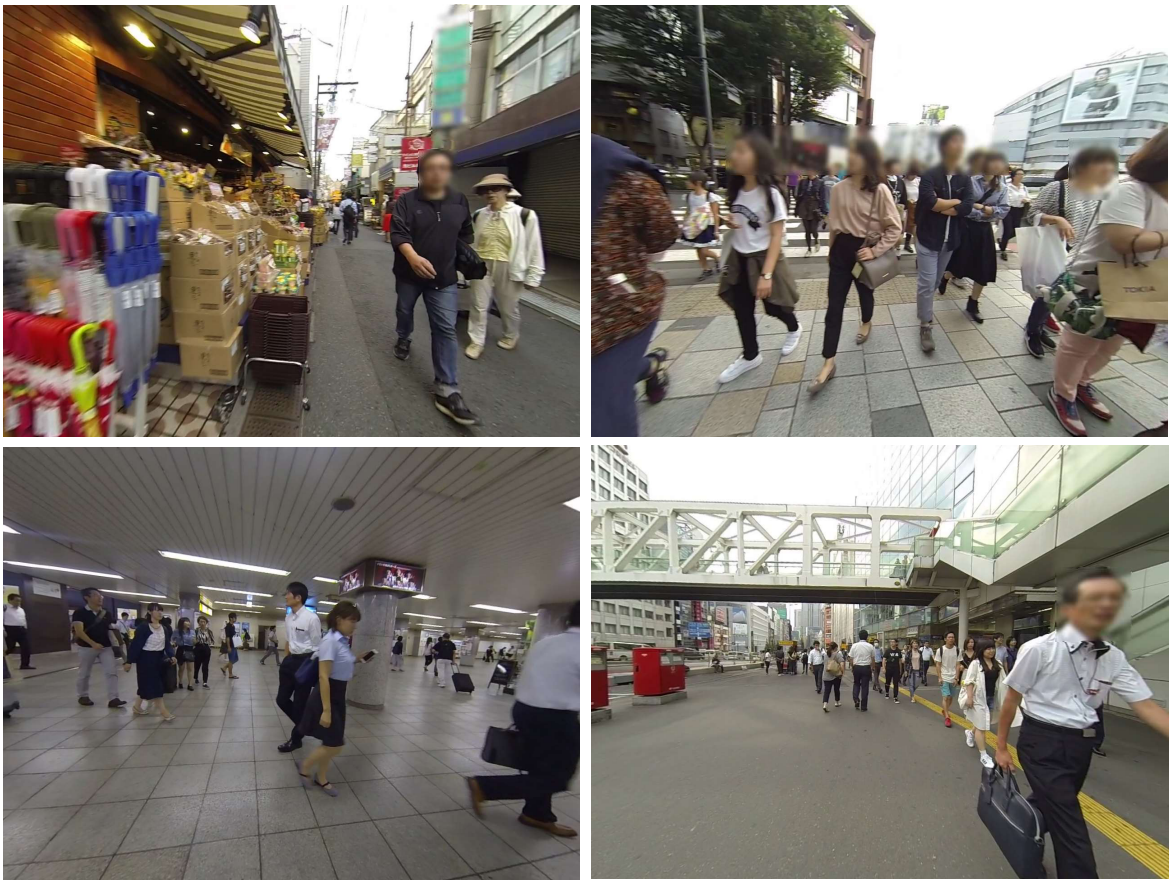

Figure 4.1.: First-Person Locomotion Dataset recorded by wearable chest-mounted cameras under diverse environments, which comprises more than 5,000 people in total.

Feature extraction For each frame, we detected people with OpenPose [15]. The detector extracts 18 2D keypoints per person. For frame alignment, we calculated a homography matrix between frames similar to [55]. First, a frame is split into 4×4 grids, and then feature points are detected using ORB detector [79] for each grid. By running an ORB detector separately, feature points will be uniformly distributed along the frame. If there are not enough matches between adjacent frames, then additional feature points are detected using a SURF detector [8]. Given matches, homography matrix between adjacent frames is calculated.

Tracklet generation We tracked the upper body of detected people (namely represented as a bounding box covering the upper body) over time using the kernelized correlation filter [33] after two consecutive frames aligned with homography. We terminated the tracking if subsequent detection results were not found within a certain pre-defined spatiotemporal range. To reduce the number of false positive association, we abort tracking when there are more than one close candidates. As a result of this

Figure 4.2.: Camera configuration: the camera is mounted on the center of the wearer's chest.

tracking, we obtained many short tracklets[1]. These tracklets were then merged to generate longer ones with the conditions 1) if the detected person at the tail of one tracklet is visually similar to that at the head of the other tracklet and 2) if these tracklets were also spatiotemporally close enough. A cosine distance of deep features extracted by Faster R-CNN [76] was used to measure visual similarity.

Error removal Due to the greedy merging strategy, some tracklets have duplicate detections among frames. Therefore, for each tracklet, we select the first bounding box as "reference detection" and compare the visual similarity of the deep features for every duplicate detections. Bounding boxes with a low score are removed here. Also, inappropriate detections (*e.g.,* bounding boxes suddenly jumping to impossible location) are removed by a pre-defined threshold. Finally, for each tracklet, we calculate the visual similarity of bounding box in the beginning, the middle and the end of it and reject the tracklet when a significantly low visual similarity between them appeared.

Data cleansing Obtained tracklets still have several problems which affect future localization performance. First, the raw pose sequence contains erroneous or missing

[1]Out of 830,000 human poses detected first, approximately 200,000 (24.1%) poses were successfully associated to form the valid samples.

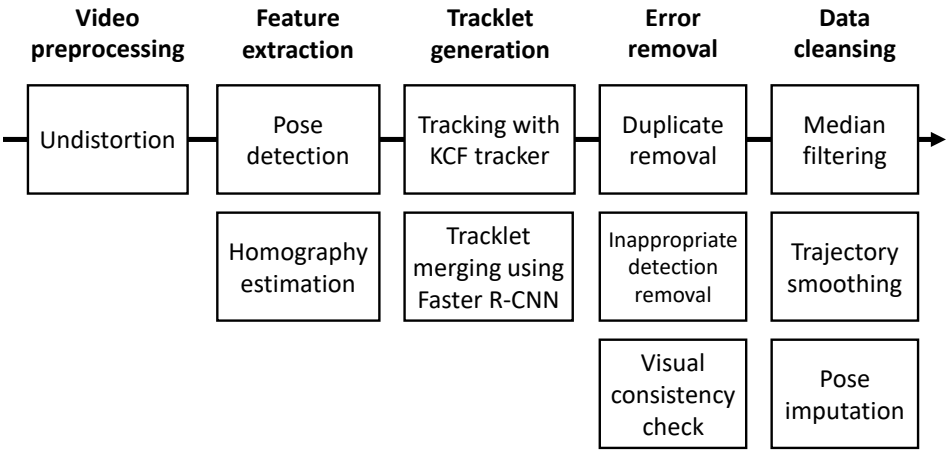

Figure 4.3.: Data preprocessing procedure.

detection because of occlusion and motion blur. Second, oscillation exists because of the wearer's locomotion. Thus, (1) median filtering, (2) pose imputation and (3) trajectory smoothing are applied. (1) 20 fps tracklets are downsampled into 10 fps here. For every frame, median filtering is applied for every keypoint. This suppresses sudden erroneous detections. (2) We applied nearest neighbor imputations to impute missing keypoint values. (3) Finally, we smoothed the y coordinate of the trajectory with a window size of five. This reduces the effect of oscillation induced by the wearer's locomotion.

Sample generation Finally, we generated fixed-length samples from the tracklets. For each tracklet, we extracted locations $\boldsymbol{l}_t$, scales s_t, poses $\boldsymbol{p}_t$, and ego-motion $\boldsymbol{e}_t$ as follows. First, we extracted 18 body parts using OpenPose [15]. $\boldsymbol{l}_t$ was then defined by the middle of two hips. Also, s_t was given by the distance between the location of the neck and $\boldsymbol{l}_t$. Furthermore, we obtained $\boldsymbol{p}_t$ as a 36-dimensional feature (*i.e.*, $V = 18$), which was normalized by subtracting $\boldsymbol{l}_t$ and divided by s_t. $\boldsymbol{e}_t$ was estimated by the unsupervised ego-motion estimator [103]. Finally, we applied sliding window to generate multiple fixed length (*i.e.*, 2 seconds) samples. As a result of this procedure, we obtained approximately 50,000 samples in total. Detailed analysis of the dataset is described in the Appendix. We note that all the features could be extracted from RGB frames observed from a single wearable camera.

Choice of ego-motion feature In this work, we used the latest ego-motion estimator [103], which predicts 6-DoF transformation matrices directly from input frames. We used their work because it only requires short snippet (5 frames) to estimate local ego-motion and works robustly regardless of the quality of the images. In our preliminary experiment, we used monocular ORB-SLAM [63] to the entire sequence to extract the ego-motion. However, the model suffered from severe scale drift when there are not enough feature points up-close (*e.g.*, waiting in front of the large intersection.) Furthermore, ORB-SLAM frequently failed to track feature points when motion blur appeared

due to rapid camera motion. On the other hand, [103] can robustly estimate the rough ego-motion and we found it enough to use it to extract input features.

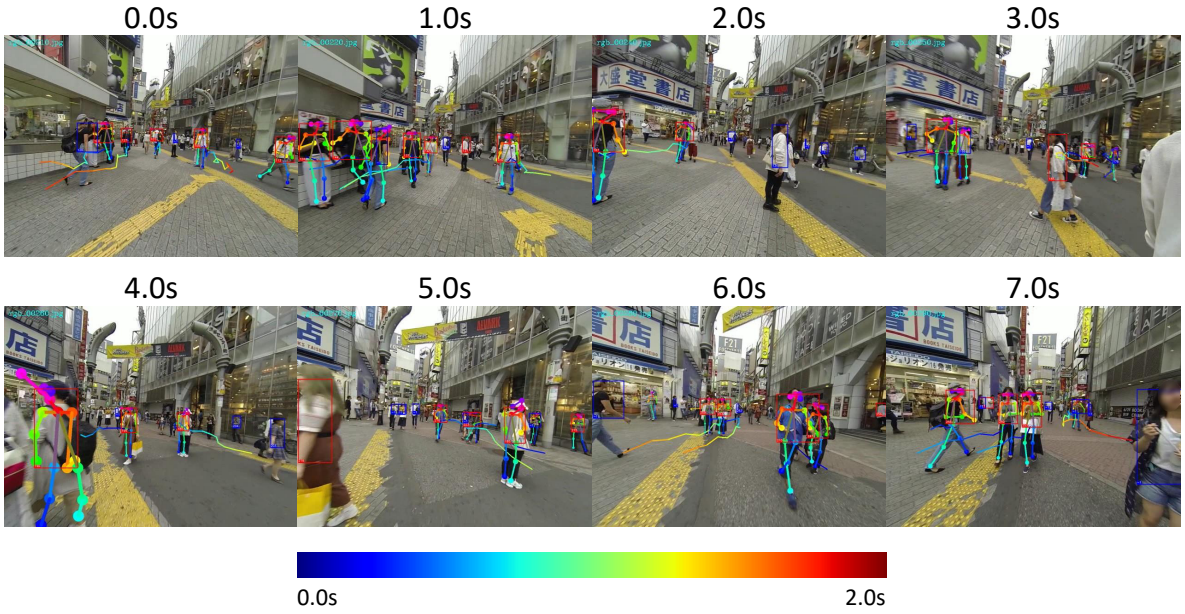

Figure 4.4.: A representative example of an extracted tracklet. red and blue bounding boxes denote detected person. Their detected pose is overlayed on the frame. Tracklet with red bounding box represents tracklet tracked two seconds or more. Tracklet with blue bounding box represents tracklet tracked less than two seconds. Trajectories are shown as a line with colormap and their color correspond to the timing when the person arrives at the specified location.

We show an example of the generated tracklets in Figure 4.4. We can find that our tracklet generation procedure can robustly associate detections even in cluttered scenes with occluded or missing detections while avoiding false positive associations. Thanks to the tracklet merging scheme, our procedure can even handle short-term occlusions (*e.g.,* the third person from the left in 0.0s: confirm that he is correctly tracked at 1.0s although a woman passed by in front of him). However, we did not deal with long-term occlusions since such occlusion does not occur much and we are mainly interested in predicting the future location of a person in front of the wearer. One possible extension is to introduce a global optimization approach [35].

5. Experiments

In this chapter, we report the experimental result and show the effectiveness of our approach.

5.1. Implementation

5.1.1. Formulation

Training and testing samples are given in the form of a tuple $(\boldsymbol{X}_{\text{in}}, \boldsymbol{E}_{\text{in}}, \boldsymbol{P}_{\text{in}}, \boldsymbol{X}_{\text{out}})$, where $\boldsymbol{X}_{\textbf{in}}$ is location-scale, $\boldsymbol{E}_{\text{in}}$ is camera ego-motion, $\boldsymbol{P}_{\text{in}}$ is pose, and $\boldsymbol{X}_{\text{out}}$ is relative future location-scale with respect to $\boldsymbol{x}_{t_0}$. $\boldsymbol{X}_{\text{in}}, \boldsymbol{E}_{\text{in}}, \boldsymbol{P}_{\text{in}}$ are available both in training and testing times and defined in interval $[t_0 - T_{\text{prev}} + 1, t_0]$. On the other hand, $\boldsymbol{X}_{\text{out}}$ serves as ground-truth defined in $[t_0 + 1, \ldots, t_0 + T_{\text{future}}]$, which we can access only during the training time. In this experiment, if not specified, we set $T_{\text{prev}} = T_{\text{future}} = 10$ at 10 fps, *i.e.*, a time window of one second for both observation and prediction.

5.1.2. Architecture choice

The full specification of the proposed network architecture is shown in Table 5.1. Each input stream feeds $D \times 10$-dimensional inputs (where D changes depending on which cues we focus on) to four cascading 1D temporal convolution layers of different numbers of channels, each of which is followed by batch normalization (BN) [39] and rectified linear unit (ReLU) activation [64]. Then, 128×2-dimensional features from the input streams are concatenated and fed to the output stream consisting of two 1D convolution layers with BN and ReLU, four cascading 1D deconvolution layers also with BN and ReLU, and one another 1D convolution layer with linear activation.

5.1.3. Optimization

To train the network, we first normalized $\boldsymbol{X}_{\text{in}}$ and $\boldsymbol{X}_{\text{out}}$ to have zero mean and unit variance. We also adopted a data augmentation by randomly flipping samples horizontally. The loss functions to predict $\boldsymbol{X}_{\text{out}}$ was defined by the mean squared error (MSE). We optimized the network via Adam [46] for 17,000 iterations with mini-batches of 64 samples, where a learning rate was initially set to 0.001 and halved at 5,000, 10,000, 15,000 iterations. All implementations were done with Chainer [90].

Layer type	Channel	Kernel size	Output size
Input streams (Location-scale, pose, and ego-motion)			
Input	-	-	$D \times 10$
1D-Conv+BN+ReLU	32	3	32×8
1D-Conv+BN+ReLU	64	3	64×6
1D-Conv+BN+ReLU	128	3	128×4
1D-Conv+BN+ReLU	128	3	128×2
Output stream			
Concat	-	-	384×2
1D-Conv+BN+ReLU	256	1	256×2
1D-Conv+BN+ReLU	256	1	256×2
1D-Deconv+BN+ReLU	256	3	256×4
1D-Deconv+BN+ReLU	128	3	128×6
1D-Deconv+BN+ReLU	64	3	64×8
1D-Deconv+BN+ReLU	32	3	32×10
1D-Conv+Linear	3	1	3×10

Table 5.1.: Our network architecture where BN: batch normalization [39] and ReLU: rectifier linear unit [64]. The network consists of three input streams and one output stream, where inputs have different dimensions D depending on the streams: $D = 3$ for the location-scale stream, $D = 6$ for the ego-motion stream, and $D = 36$ for the pose stream.

5.2. Evaluation Protocols

5.2.1. Data splits

We adopted five-fold cross-validation by randomly splitting samples into five subsets. We ensured that samples in training and testing subsets were drawn from different videos. Training each split required about 1.5 hours on a single NVIDIA TITAN X. Also when evaluating methods with testing subsets, we further divided samples into three conditions based on how people walked (*i.e.*, walking directions): target people walked a) **Toward**, b) **Away** from, or c) **Across** the view of a camera.

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%.

5.2.2. Evaluation metric

Although our network predicts both locations and scales of people in the future frames, we measured its performance based on how accurate the predicted locations were. Similar to [1], we employed the final displacement error (FDE) as our evaluation metric. Specifically, FDE was defined by the L2 distance between predicted final locations $\boldsymbol{l}_{t_0+T_{\text{future}}}$ and the corresponding ground-truth locations.

5.2.3. Baseline methods

Since there were no prior methods that aimed to predict future person locations in first-person videos, we have implemented the following baselines.

- **Constant**: We use location at the t_0-th frame as the prediction.
- **ConstVel**: Inspired by the baseline used in [69], this method assumes that target people moved straight at a constant speed. Specifically, we computed the average speed and direction from $\boldsymbol{X}_{\text{in}}$ to predict where the target would be located at the $t_0 + T_{\text{future}}$-th frame.
- **NNeighbor**: We selected k-nearest neighbor input sequences in terms of the L2 distance on the sequences of locations $\boldsymbol{L}_{\text{in}}$ and derived the average of k-corresponding locations at frame $t_0 + T_{\text{future}}$. In our experiments, we set $k = 16$ as it performed well.
- **Social LSTM** [1]: We also evaluated Social LSTM, one of the state-of-the-art approaches on human trajectory prediction, with several minor modifications to better work on first-person videos. Specifically, we added the scale information to inputs and outputs. The estimation of Gaussian distributions was replaced by direct prediction of $\boldsymbol{X}_{\text{out}}$ as it often failed on the FPL Dataset. The neighborhood size N_o used in the paper was set to $N_o = 256$.

5.3. Results

5.3.1. Quantitative evaluation

Table 5.2 reports FDE scores on our FPL Dataset. Overall, all methods were able to predict future locations of people with the FDE less than about 15% of the frame width (approximately 19° in horizontal angle). We confirmed that our method (**Ours**) has significantly outperformed the other baselines. Since walking speeds and directions of people were quite diverse and changing dynamically over time, naive baselines like **ConstVel** and **NNeighbor** did not perform well. Moreover, we found that **Social LSTM** [1] performed poorly. Without explicitly taking into account how significant ego-motion affects people locations in frames, temporal models like LSTM would not be

Method	Walking direction			
	Toward	Away	Across	Average
Constant	16.51	7.52	11.59	8.54
ConstVel	13.98	7.70	9.50	8.37
NNeighbor	12.95	7.02	9.67	7.69
Social LSTM[1]	13.65	8.66	11.11	9.25
Ours	**8.06**	**5.76**	**7.03**	**6.04**

Table 5.2.: Comparisons to baseline methods. Each score describes the final displacement error (FDE) in percentage with respect to the frame width of 1280 pixels.

able to learn meaningful temporal dynamics, ultimately rendering their predictions quite unstable. Note that without our modification shown in Section 5.2, the performance of vanilla Social LSTM was further degraded (*i.e.*, 11.9 FDE on average). Comparing results among walking directions, **Toward** was typically more challenging than other conditions. This is because when target people walked toward the view of a camera, they would appear in the lower part of frames, making variability of future locations much higher than other walking directions.

5.3.2. Error analysis

We investigated the distribution of the errors. With our method, 73% samples received error less than 100 pixels (10° in horizontal angle). There were only 1.4% samples suffered from significant error larger than 300 pixels (30° in horizontal angle). Additionally, we calculated the errors normalized by each sample's scale. By assuming that the length between the center hip and the neck of a person to be 60 cm, the average error obtained by our method approximately corresponded to 60 cm in the physical world.

5.3.3. Qualitative evaluation

Figure 5.1 presents several visual examples of how each method worked. Examples (a), (b), and (c) are results drawn respectively from **Toward**, **Across**, and **Away** subsets. Especially, significant ego-motion of the camera wearer to turn right was observed in Example (b), making predictions of baseline methods completely failure. Another case where ego-motion played an important role was when target people did not move, such as the person standing still in Example (d). Example (e) involves not only significant ego-motion but also changes in walking direction of the target. Our method successfully performed in this case as it could capture postural changes of target persons for prediction.

Method	Walking direction			
	Toward	Away	Across	Average
$\boldsymbol{L}_{\text{in}}$	11.44	6.32	8.25	6.89
$\boldsymbol{X}_{\text{in}}$	9.26	6.02	7.77	6.40
$\boldsymbol{X}_{\text{in}} + \boldsymbol{E}_{\text{in}}$	8.80	5.80	7.35	6.15
$\boldsymbol{X}_{\text{in}} + \boldsymbol{P}_{\text{in}}$	8.38	6.00	7.61	6.29
Ours ($\boldsymbol{X}_{\text{in}} + \boldsymbol{E}_{\text{in}} + \boldsymbol{P}_{\text{in}}$)	**8.06**	**5.76**	**7.03**	**6.04**

Table 5.3.: Ablation study. $\boldsymbol{L}_{\text{in}}$: locations, $\boldsymbol{X}_{\text{in}}$ location-scales, $\boldsymbol{E}_{\text{in}}$: ego-motion, and $\boldsymbol{P}_{\text{in}}$: poses. Each score describes the final displacement error (FDE) in percentage with respect to the frame width of 1280 pixels.

5.3.4. Ablation study

We made an ablation study to see how each of scales, ego-motion, and poses contributed overall prediction performances. Specifically, we started from the only location information $\boldsymbol{L}_{\text{in}}$, then added scale information to use $\boldsymbol{X}_{\text{in}}$. For these two conditions, we learned a single-stream convolution-deconvolution architecture. Then, we evaluated the combination of $\boldsymbol{X}_{\text{in}} + \boldsymbol{E}_{\text{in}}$ (locations, scales, and ego-motion) and that of $\boldsymbol{X}_{\text{in}} + \boldsymbol{P}_{\text{in}}$ (locations, scales, and poses) by learning two-stream convolution-deconvolution architectures. Results are shown in Table 5.3. We confirmed that all of the cues helped individually to improve prediction performances. Especially, significant performance gains were observed on the **Toward** subset from $\boldsymbol{L}_{\text{in}}$ to $\boldsymbol{X}_{\text{in}}$, *i.e.*, by introducing scale information, and from $\boldsymbol{X}_{\text{in}}$ to $\boldsymbol{X}_{\text{in}} + \boldsymbol{P}_{\text{in}}$, *i.e.*, by further combining pose information. By adding ego-motion information $\boldsymbol{E}_{\text{in}}$, a modest improvement was observed on the **Away** subset and achieved the best performance by our full model.

5.3.5. Effect of prediction length

We conducted further analysis of the effect of prediction length. Results are shown in Table 5.4 and Figure 5.2. From Table 5.4, we found that the prediction error linearly increases with the prediction length in all categories. From Figure 5.4, we found that also Social LSTM follows the linear increase trend, but their increment is lower in our proposed method. These results indicate that the proposed method consistently performs better than the baseline regardless of the prediction length.

5.3.6. Failure cases

Figure 5.3 shows several typical failure cases. On both examples, our method and other baselines did not perform accurately as camera wearers made sudden unexpected ego-motion. In the case of the first row, the model predicted that the target persons walk left

Walking direction	Prediction length (second)				
	0.2	0.4	0.6	0.8	1.0
Toward	1.46	2.72	4.00	5.58	8.06
Away	1.41	2.50	3.52	4.61	5.76
Across	1.53	2.82	4.01	5.28	7.03
Average	1.42	2.53	3.59	4.73	6.04

Table 5.4.: Effect of prediction length. Each score describes the final displacement error (FDE) in percentage with respect to the frame width of 1280 pixels. In all conditions, the prediction error linearly increases with the prediction length.

(the wearer goes right towards the target) while the target person walks right in reality. From the observation, it is likely to predict so since the current location of the person is off to the left. However, there is an obstacle (signboard) on the right, and the wearer cannot change the direction to the right. To overcome this failure, scene information should be incorporated while we do not consider in the proposed method. In the case of the second row, the model predicted that the target person would go to the left from the view of the wearer regarding the wearer's direction change to the right, but it does not in reality. We think that this is because the current model models the mixture of the wearer's movement and the target person's movement, by using the image coordinates seen from the wearer's view. This assumption can omit the procedure of mapping one's position into world coordinate but makes capturing the independent motion difficult.

5.4. Evaluation on Social Interaction Dataset

Finally, we evaluate how our approach works on First-Person Social Interaction Dataset [20]. This dataset consists of several first-person videos taken in an amusement park and involves a variety of social moments like communicating with friends, interacting with a clerk, and waiting in line, standing for a more general and challenging dataset. In our experiment, we manually extracted a subset of videos where camera wearers kept walking while sometimes interacting with others. From this subset, we collected approximately 10,000 samples in total. Similar to the previous experiment, we adopted five-fold cross-validation to evaluate how our method and other baselines performed.

5.4.1. Training setup

In this dataset, camera wearers frequently turned their head to pay their attention to various different locations. This made ego-motion estimator [103] completely inaccurate as it was originally trained to estimate ego-motion of vehicle-mounted cameras, where

Method	Walking direction			
	Toward	Away	Across	Average
Ours with grid flow	8.32	5.96	7.50	6.25
Ours full model	**8.06**	**5.76**	**7.03**	**6.04**

Table 5.5.: Flow-based ego-motion feature. Each score describes the final displacement error (FDE) in percentage with respect to the frame width of 1280 pixels.

such frequent turning was hardly observed in their training datasets. To cope with this, instead of the velocity and rotation used in Section 3.4, we made use of optical flows to describe ego-motion cues. More specifically, we computed dense optical flows using [38] and divided them into 4×3 grids. We then computed average flows per grid and concatenated them to obtain 24-dimensional vector for describing ego-motion per frame. For the training, we first pre-trained our network on FPL Dataset with the same training strategies shown in Section 5.1 but with the above flow-based ego-motion feature. Table 5.5 show the performance using the flow-based feature. Our network with flow-based features resulted in 6.25% FDE on FPL dataset, *i.e.*, 0.21% performance drop from the full model. One possible reason for the better performance using ego-motion features based on [103] is that they can capture yaw rotations (*i.e.*, turning left and right) of camera wearers more accurately. We then fine-tuned this trained network on the Social Interaction Dataset for 200 iterations using Adam with a learning rate of 0.002.

5.4.2. Results

FDE scores are shown in Table 5.6. Similar to the previous experiment, we divided testing datasets into three subsets, Toward, Away, and Across, based on walking directions of target people. Although performances of all methods were rather limited compared to the previous results in Table 5.2, we still confirmed that our method was able to outperform other baseline methods including Social LSTM [1]. Some visual examples are also shown in Figure 5.4.

Since this dataset is captured from head-mounted cameras, head motions are far more complicated than our FPL dataset, making a prediction in image coordinates difficult. However, we observed that our method still have the ability to capture the interaction between the wearer and the target person as shown in Figure 5.4. In the second example, while the trajectory itself is moving to the right, the proposed model correctly captured the direction change.

Method	Walking direction			
	Toward	Away	Across	Average
Constant	12.89	10.47	10.60	11.65
ConstVel	13.57	13.81	10.42	13.34
NNeighbor	13.06	12.44	11.63	12.65
Social LSTM [1]	14.08	13.79	13.26	13.90
Ours	**9.29**	**8.96**	**8.53**	**9.10**

Table 5.6.: Results on social interactions dataset [20]. Each score describes the final displacement error (FDE) in in percentage with respect to the frame width of 1280 pixels.

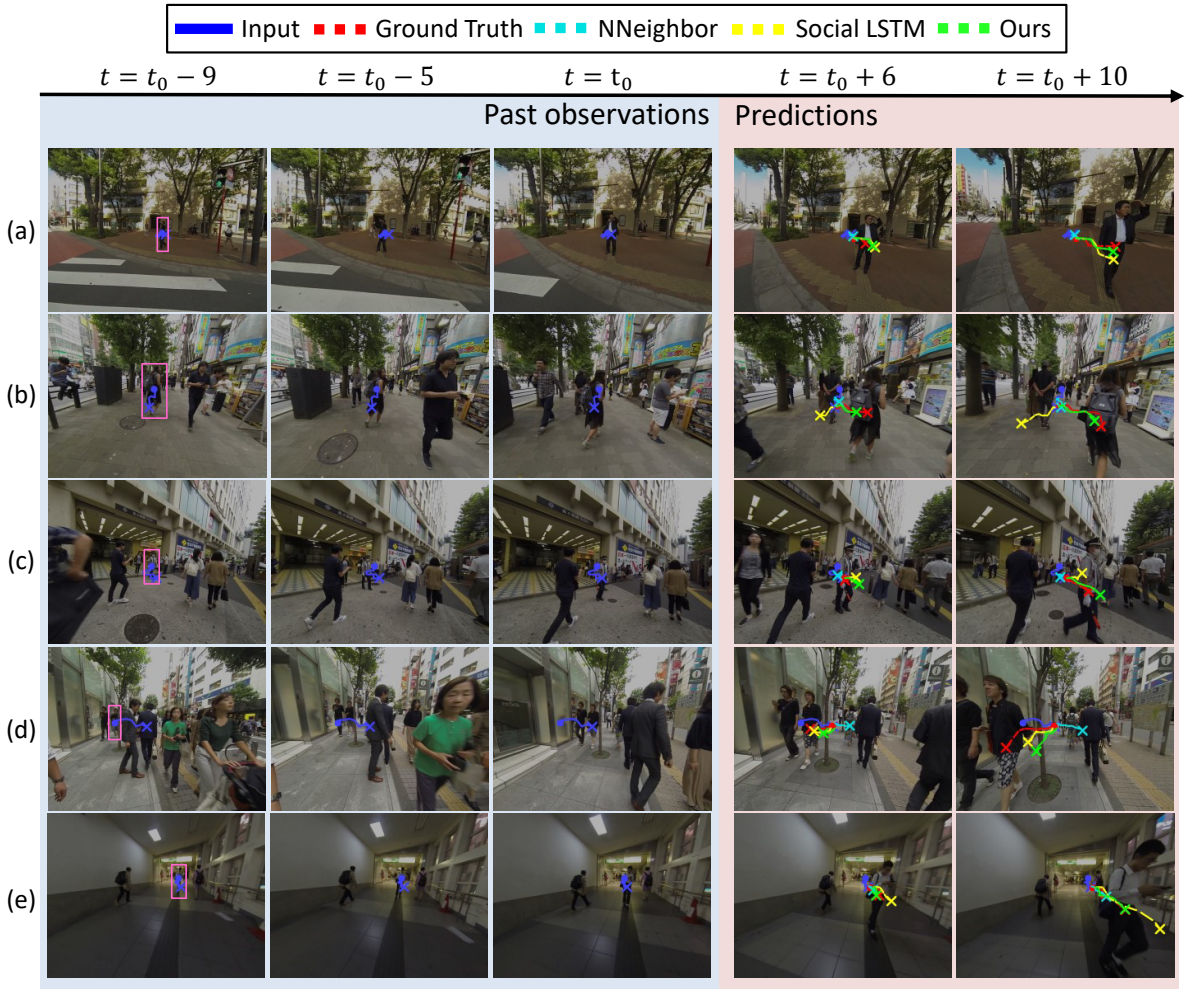

Figure 5.1.: Visual prediction examples. Using locations (shown with solid blue lines), scales and poses of target people (highlighted in pink, left column) as well as ego-motion of camera wearers in the past observations highlighted in blue, we predict locations of that target (the ground-truth shown with red crosses with dotted red lines) in the future frames highlighted in red. We compared several methods: **Ours** (green), **NNeighbor** (cyan), and **Social LSTM** [1] (yellow).

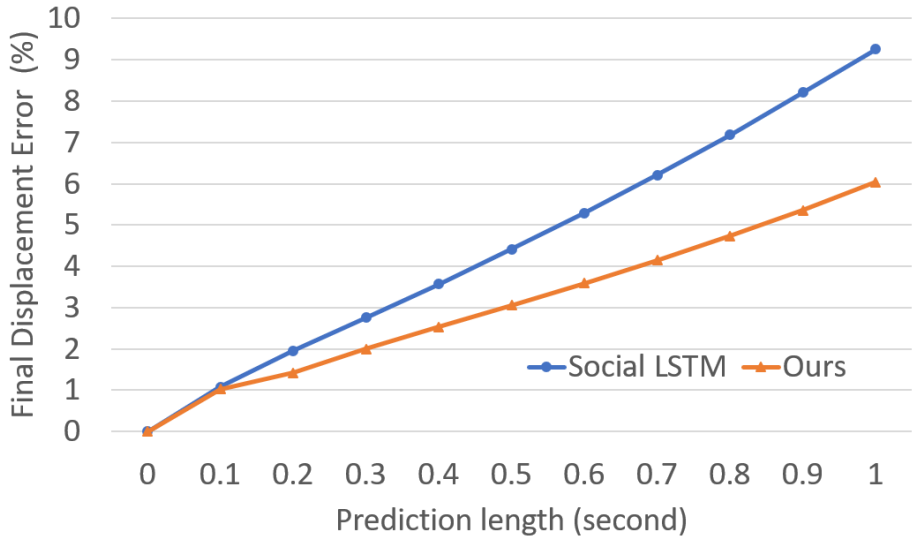

Figure 5.2.: Effect of prediction length. Quantitative comparison of the final displacement error (FDE) in percentage with respect to the frame width of 1280 pixels. Our proposed method's amount of error increment is lower than the Social LSTM baseline.

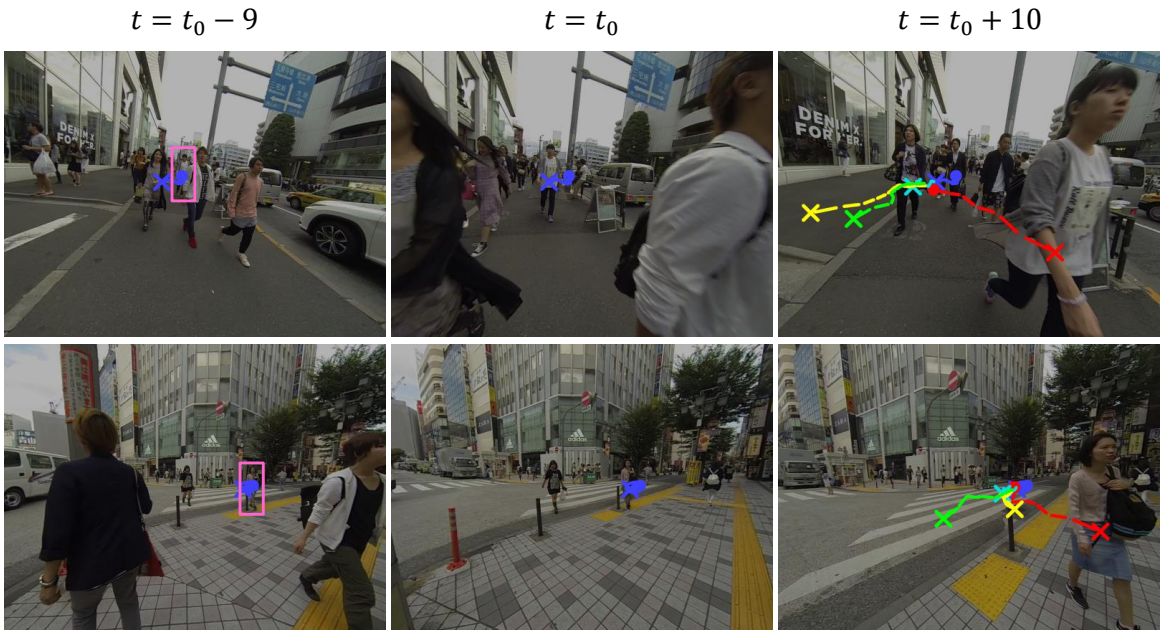

Figure 5.3.: Failure cases. Given previous locations (blue) of of target people (pink bounding boxes), predictions by our method (green) and Social LSTM [1] (yellow) both deviated from ground-truth future locations (red).

Figure 5.4.: Visual prediction examples on Social Interaction Dataset [20]. previous locations (blue lines) of target people (pink bounding boxes); predictions by our method (green lines); and ground-truth future locations (red lines).

6. Conclusion

6.1. Summary

In this thesis, we have presented a new task of future person localization in first-person videos. This work first tackles the problem of predicting the future position of a person observed in first-person videos, providing a margin for preparation before the actual event happens. Unlike previous approaches which only considered the positional information, our work considered additional visual cues obtained from first-person videos, namely the pose of target people and ego-motion of camera wearers. The pose information works as a strong prior of where the person is likely to move next. The ego-motion information explicitly takes the interaction between the wearer and the target person. Experimental results have revealed that ego-motion of camera wearers, as well as scales and poses of target people, were all necessary ingredients to accurately predict where target people would appear in future frames. All the features are extracted from RGB frames captured from a single wearable camera.

6.2. Limitation and Future Work

We discuss the limitations found in our method and show possible future directions for this work.

Cross-subject evaluation

In this study, we only collected walking scenes from a single wearer. However, different wearers might exhibit different interaction between the surrounding people due to the difference in their physical and psychological characteristics. Different camera configuration also affects performance. It is worth evaluating the cross-subject performance—how performance changes when videos from a new wearer who did not appear in training data are given.

Inference in 3D coordinates

In this work, we assumed that only a single RGB camera is available. Thus, we constructed a prediction framework taking 2D image coordinates as inputs and outputs,

ignoring the depth information—how far the people appear. The scale information partially compensates the lost perspective effect, works as an approximation of how far the people appear. However, the approximation error causes a drop in future localization performance. Pose and ego-motion representation also ignore or only approximate the depth information, which leads to suboptimal performance.

To overcome this issue, we can consider using richer sensors so that we can measure precise depth information [69, 57]. Park *et al.* [69] utilized a stereo camera and constructed an EgoRetinal map based on the depth information calculated from stereo images. Luo *et al.* [57] used the 3D point cloud data captured from a laser scanner. By using more accurate depth information, we can map the 3D position of people into a global world coordinate by running a Simultaneous Localization and Mapping (SLAM) algorithm and run inference on 3D space. One interesting direction is to simultaneously solve global mapping and motion forecasting, where preliminary work is done in an indoor static environment [77].

Separation of wearer motion and target Motion

In this work, we have modeled future person localization on the space of apparent motion observed from the wearer's view. This motion results from the interplay between the wearer's motion and the target person's motion. This assumption makes the formulation easier because it comes down to a problem of predicting a single time-series data. However, in reality, it is rather natural to assume that the wearer and the target person has their own motion model. For example, when the distance between the wearer and the target person are far away (*e.g.*, 5 meters), one's motion might not affect another person's motion anymore—their motions are independent. In such case, we might want to model the motion of the wearer and the target person separately rather than their mixture. Also, apparent motion prediction only forecasts the relative position between the wearer and the target person, making it impossible to leverage the information of the static obstacles. In other words, the current model can predict where the person would appear looking from the wearer's view but cannot predict where he or she will moves.

Therefore, formulating the problem as a set of two motion forecasting procedures—motion forecasting of the wearer and the target person, respectively—would be one promising extension. Concretely, we can think of the problem as separately predicting the wearer's motion and the target person's on a shared space (*e.g.,* top-down world coordinate, the EgoRetinal map [69]). Separate modeling makes it easier to incorporate contextual information such as scene information since we now predict the future absolute position of the wearer and the target person separately. This problem can be solved as an instance of multi-task learning problem, simultaneously learning wearer motion and target motion.

Use of scene context

As we showed in the failure example in Figure 5.3, scene context plays an important role also in the first-person setting. The model should be able to avoid obstacles or unwalkable area by considering the scene information. Effectively incorporating both geometric cues [69] and semantic cues [47] into a prediction framework of human-human interaction is still an open question.

Forecasting under uncertainty

This work adopted a simple deterministic prediction scheme since we aimed to evaluate the amount of prediction error. However, in practical applications, we often would like to know the *confidence* or *uncertainty* of the prediction. When two pedestrians walk while facing each other, they are equally likely to avoid either left or right, and its probability distribution will be uniform distribution if we can demonstrate the same situation enough number of times. Such an uncertain condition frequently occurs in the trajectory prediction problem, and the model should be able to forecast multi-modal trajectory distributions. Previous works adopted using a generative model which can sample new trajectory conditioned by a random variable [50, 28]. Although this approach can determine the dispersion of predicted samples by generating many samples, it is computationally inefficient and cannot measure the *uncertainty* of the prediction in a principled way.

Although there are several approaches to measure such uncertainty*, Bayesian deep learning* approaches [66] command considerable attention in recent days [25, 44]. This Bayesian modeling approach categorizes uncertainty into two categories—*aleatoric* uncertainty and *epistemic* uncertainty—and provides principled guide for understanding uncertainty appearing in the real world. It is known that both types of uncertainty can be captured by a single Bayesian neural network (BNN) [44].

Their model provided a natural formulation of capturing both aleatoric uncertainty and epistemic uncertainty from a small number of sampled outputs, however, only can predict single-modal prediction using Gaussian distribution. In the case of trajectory prediction, this assumption is inapplicable since multiple likely trajectories will appear. Bhattacharyya *et al.* [11] adopted BNN for predicting pedestrian movement. However, their model cannot capture switching actions, whether a pedestrian should cross the road or not, for example. One interesting future direction is to develop a method which can predict a multi-modal distribution with a minimum number of samples.

Online personalization

This work assumes *assistive vision* as one of the potential applications. Assistive vision using wearable cameras can assume that the wearer continuously wears the device for longer horizons. In such case, we can receive benefit from personalizing the model for a

specific wearer—adapting the model to give better performance for the user—resulting in improved performance for the user. Such personalization model is not sufficiently studied in the field of computer vision, and future person localization problem is a suitable task of deploying it. We can adaptively update the model during the use of the device and can reduce the regret of the model.

Acknowledgments

まず指導教官である佐藤洋一教授に多大なる感謝を申し上げます．佐藤先生にはテーマ選定，実験遂行から論文執筆まで隅々にわたる指導をいただきました．佐藤先生が示した高い理想と指導があって初めて，自分が取り組んだ研究を国際会議の場に押し上げることができました．私が進む方向に悩んだ時は，常に私を適切な方向に導いてくださいました．佐藤先生の研究に対する真摯な姿勢には大いに影響を受けました．

直接指導をいただいた米谷先生には研究上の一般的な指導に加え，実験上の細かい点の決定にあたって数えきれないほどのアドバイスをいただきました．米谷先生の具体的かつ的確なアドバイスは実験や執筆を進めるうえで不可欠のものでした．この場を借りて改めて感謝申し上げます．

また，佐藤研究室で共に過ごしたスタッフ及び学生の皆様に感謝申し上げます．樋口先生には飲み会・研究室旅行などの研究外でのイベントを中心に，私たちが研究室生活を送るうえでのご支援を多数いただきました．松井先生には本稿およびスライド発表のご指導をいただきました．共に研究に取り組んだ Karttikeya と Chinmoy からは，英語でコミュニケーションを取りながら研究することの楽しさ・難しさを学びました．同期の加藤君，松下君，土田君，Li 君には快適な研究室生活を提供してもらいました．ありがとうございました．

最後に，常に私を見守り，支えてくれた両親に感謝して，本稿を閉じたいと思います．

八木 拓真

2019/01/31

A. Data Statistics

Figure A.1 and A.2 presents frequency distributions of lengths of the tracklets extracted from First-Person Locomotion Dataset and Social Interaction Dataset [20]. 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) long enough to learn meaningful temporal dynamics and 2) frequently observed 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.'

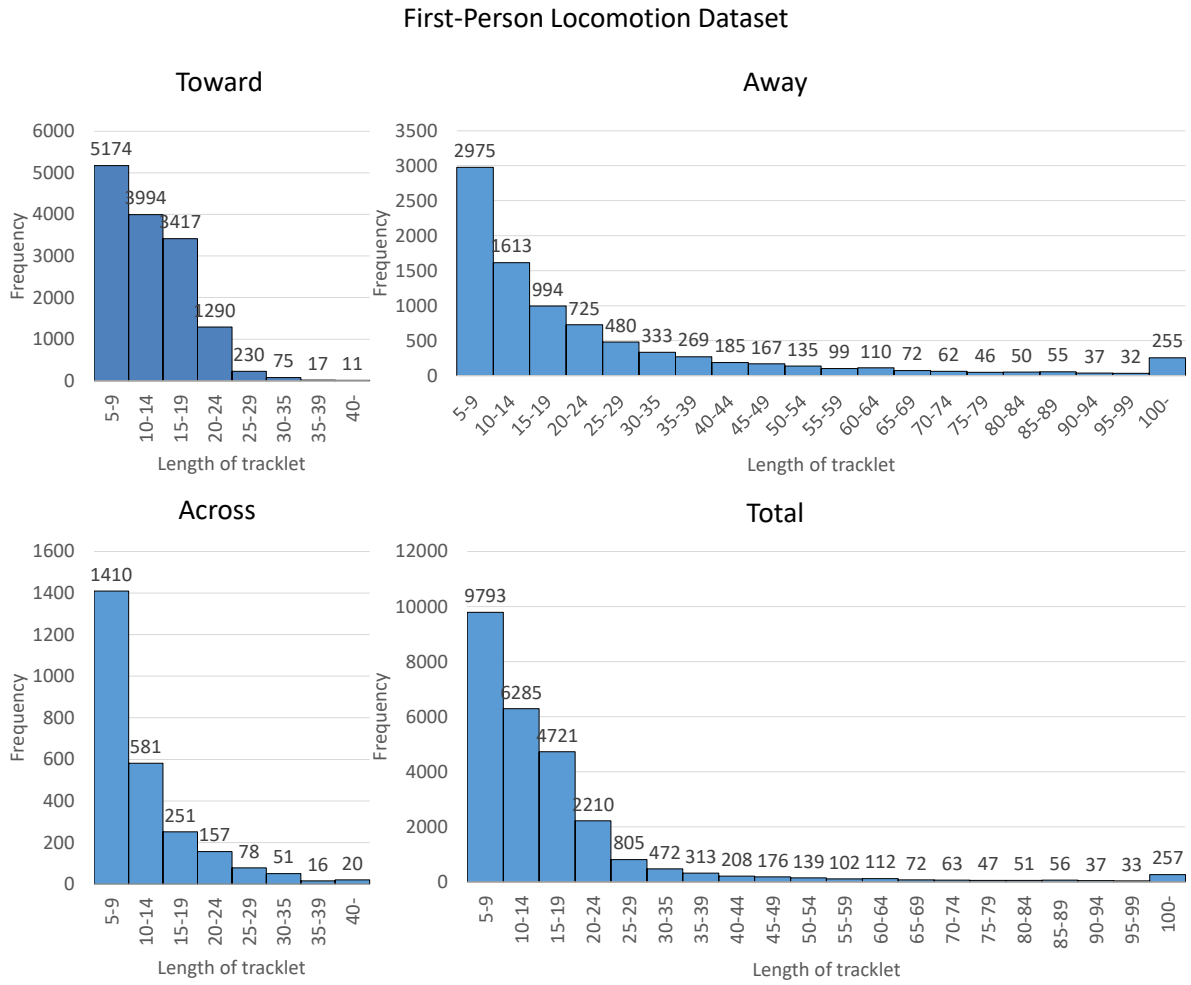

Figure A.1.: Distributions of tracklet lengths (FPL). Frequency distributions of various lengths of tracklets extracted from First-Person Locomotion Dataset for three walking directions and the entire database, respectively.

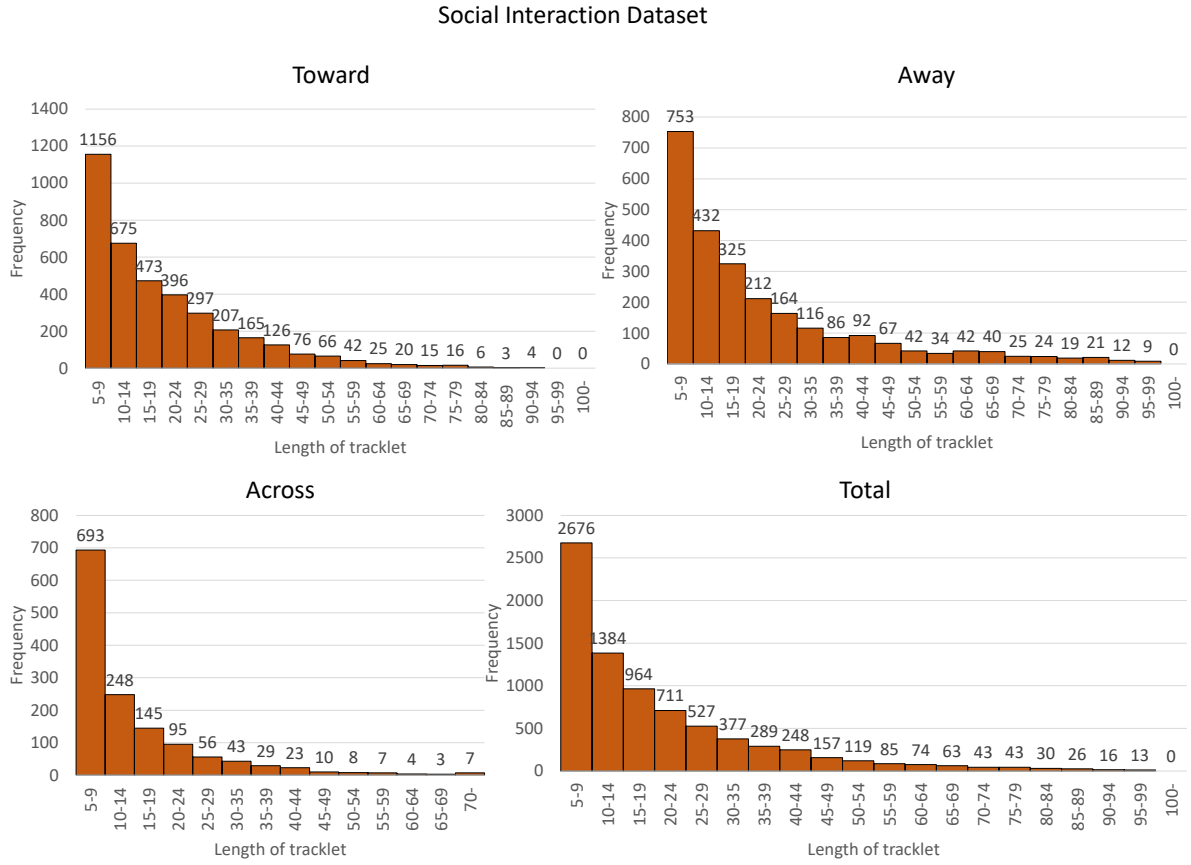

Figure A.2.: Distributions of tracklet lengths (Social Interaction). Frequency distributions of various lengths of tracklets extracted from Social Interaction Dataset [20] for three walking directions and the entire database, respectively.

B. Additional Results

B.1. Other Choices of Input/Output Lengths

In our main experiments, we fixed the input and output lengths $T_{\text{prev}}, T_{\text{future}}$ to be $T_{\text{prev}} = T_{\text{future}} = 10$. Table B.1 shows how performances changed for other choices of T_{prev} and T_{future}. Overall, longer input lengths led to better performance ($T_{\text{prev}} = 6$ vs. 10). Also, predicting more distant futures becomes more difficult ($T_{\text{future}} = 10$ 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.*, $T_{\text{future}} = 20$) in Table B.2. To overcome the lack of training examples, we set $T_{\text{prev}} = 6$ for this experiment. We confirmed that our method still worked well on this challenging condition. To generate a 20 frame prediction, we changed the kernel size of the deconvolution layers of 3, 3, 3, 3 to 3, 5, 7, 7.

B.2. Other Visual Examples

Figure B.1 shows additional visual examples of how our method, as well as several baselines, predicted future locations of people.

B.3. Ablation Study on Social Interaction Dataset

We performed an ablation study on Social Interaction Dataset [20] in Table B.3. While we computed ego-motion based on optical flows, the combination of ego-motion and pose cues contributed to performance improvements in a complementary manner. This result indicates that our method's effectiveness holds across different datasets.

T_{prev}	T_{future}	Walking direction			
		Toward	Away	Across	Average
6	10	8.01	5.82	7.06	6.08
10	10	8.06	5.76	7.03	6.04
6	6	4.04	3.54	4.00	3.61
10	6	4.00	3.52	3.87	3.59

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

Method	Walking direction			
	Toward	Away	Across	Average
Social LSTM [1]	22.12	17.56	19.69	17.75
Ours	**13.68**	**9.54**	**12.81**	**9.75**

Table B.2.: Predicting two-second futures. Final Displacement Error (FDE) where T_{prev} and T_{future} was set to 6 and 20, respectively.

Method	Walking direction			
	Toward	Away	Across	Average
$\boldsymbol{X}_{\text{in}}$	10.10	9.25	8.31	9.60
$\boldsymbol{X}_{\text{in}} + \boldsymbol{E}_{\text{in}}$	9.86	9.14	8.13	9.41
$\boldsymbol{X}_{\text{in}} + \boldsymbol{P}_{\text{in}}$	9.68	9.19	8.10	9.35
Ours ($\boldsymbol{X}_{\text{in}} + \boldsymbol{E}_{\text{in}} + \boldsymbol{P}_{\text{in}}$)	**9.58**	**9.12**	**7.90**	**9.25**

Table B.3.: Ablation study on Social Interactions Dataset [20]. Final displacement error (FDE) for various combination of input features. Notations were the same as those of Table B.2.

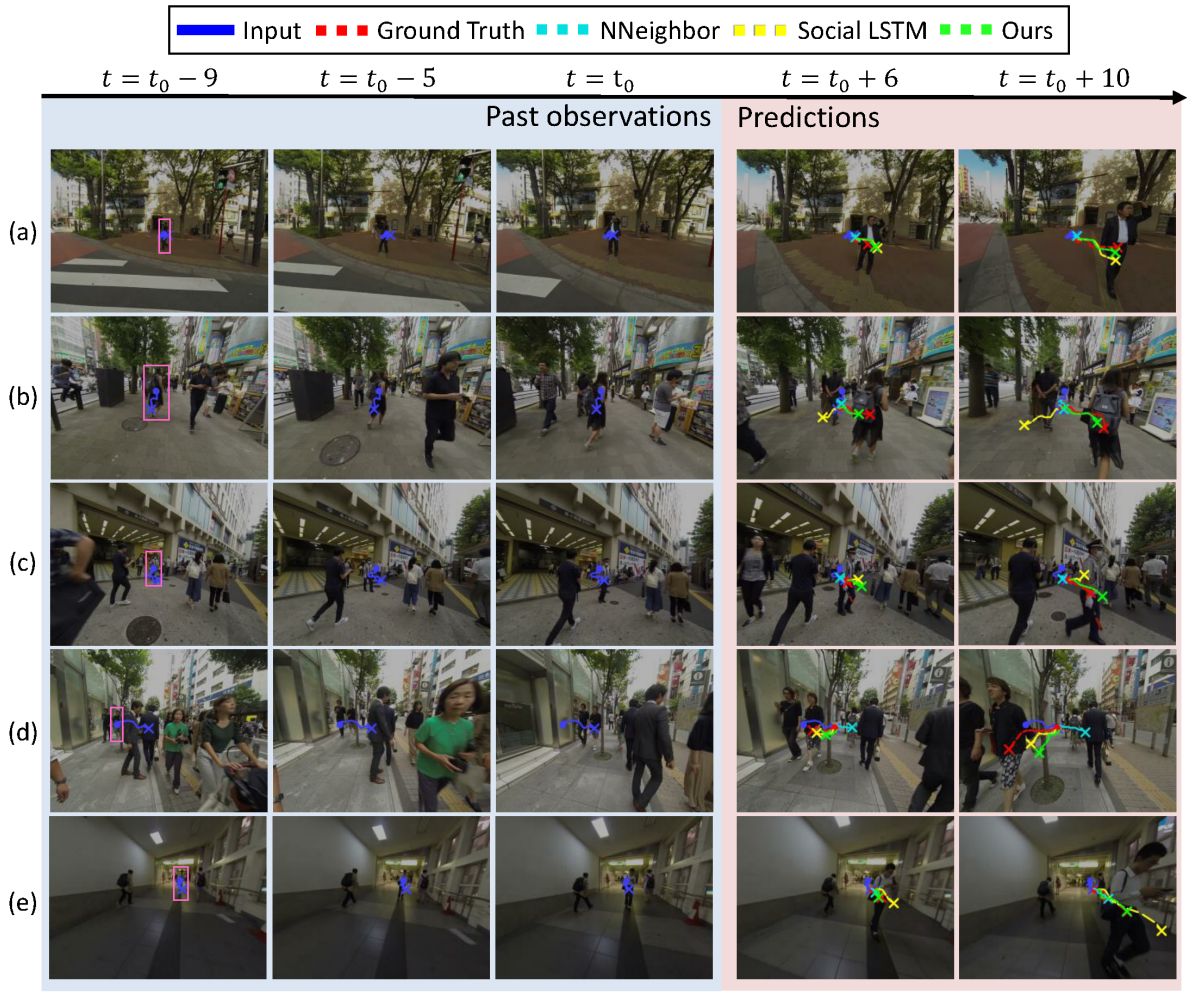

Figure B.1.: Additional prediction examples on First Person Locomotion Dataset. (Row 1) Even though the input sequence is almost static, our model is able to capture the left turn caused by the wearer's ego-motion. (Row 2, 3) In the input sequence, the target is changing the pose to move right. While the 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.

C. Runtime Analysis

As of now, we have designed the algorithm as an offline algorithm, ignoring the demand for real-time execution. Our proposed algorithm mainly consists of four processes: pose estimation, tracking, ego-motion estimation, and prediction. Since pose estimation and ego-motion calculation requires running a deep neural network model, it requires heavy computation to accomplish it. Concretely, in our environment, pose estimation, tracking, ego-motion estimation and prediction took 125ms, 200ms, 125ms, 100ms per frame, respectively. This means that it will take more than four seconds if we would like to process one-second input and make a prediction. One interesting direction is to implement the real-time version of our algorithm.

Bibliography

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio 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] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[3] Stefano Alletto, Giuseppe Serra, Simone Calderara, and Rita Cucchiara. Understanding social relationships in egocentric vision. *Pattern Recognition*, 48(12):4082–4096, 2015.

[4] Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, and Rita Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 580–585, 2014.

[5] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.

[6] Lamberto Ballan, Francesco Castaldo, Alexandre Alahi, Francesco Palmieri, and Silvio Savarese. Knowledge transfer for scene-specific motion prediction. In *European Conference on Computer Vision*, pages 697–713, 2016.

[7] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015.

[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision*, pages 404–417. Springer, 2006.

[9] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3457–3464, 2011.

[10] Gedas Bertasius, Aaron Chan, and Jianbo Shi. Egocentric basketball motion planning from a single first-person image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5889–5898, 2018.

[11] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4194–4202, 2018.

[12] Syed Zahir Bokhari and Kris M Kitani. Long-term activity forecasting using first-person vision. In *Proceedings of the Asian Conference on Computer Vision*, pages 346–360, 2016.

[13] Sarah Bonnin, Thomas H Weisswange, Franz Kummert, and Jens Schmüdderich. Pedestrian crossing prediction using multiple context-based models. In *Proceedings of the IEEE International Conference on Intelligent Transportation System*, pages 378–385, 2014.

[14] Minjie Cai, K. M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1360–1366, 2015.

[15] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291 – 7299, 2017.

[16] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[17] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *International Conference on Computer Vision Workshops*, pages 1282–1289, 2009.

[18] David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *IEEE International Conference on Computer Vision Workshops*, pages 1229–1234, 2009.

[19] Chenyou Fan, Jangwon Lee, and Michael S. Ryoo. Forecasting hand and object locations in future frames. *CoRR*, abs/1705.07328, 2017.

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

[21] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding Egocentric Activities. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 407–414, 2011.

[22] Amalia F Foka and Panos E Trahanias. Predictive autonomous robot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 490–495, 2002.

[23] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[24] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49(C):401–411, 2017.

[25] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[26] Tian Gan, Yongkang Wong, Bappaditya Mandal, Vijay Chandrasekhar, Liyuan Li, Joo-Hwee Lim, and Mohan S Kankanhalli. Recovering social interaction spatial structure from multiple first-person views. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 7–12, 2014.

[27] Weina Ge, Robert T Collins, and Barry Ruback. Automatically detecting the small group structure of a crowd. In *Workshop on Applications of Computer Vision*, pages 1–8, 2009.

[28] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[29] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical Review E*, pages 4282–4286, 1995.

[31] Leroy F Henderson. On the fluid mechanics of human crowd motion. *Transportation research*, 8(6):509–515, 1974.

[32] LF Henderson. The statistics of crowd fluids. *Nature*, 229(5284):381–383, 1971.

[33] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[34] Y. Hoshen and S. Peleg. An egocentric look at video photographer identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4284–4292, 2016.

[35] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proceedings of the European Conference on Computer Vision*, pages 788–801, 2008.

[36] Siyu Huang, Xi Li, Zhongfei Zhang, Zhouzhou He, Fei Wu, Wei Liu, Jinhui Tang, and Yueting Zhuang. Deep learning driven visual path prediction from a single image. *IEEE Transactions on Image Processing*, 25(12):5892–5904, 2016.

[37] Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Temporal localization and spatial segmentation of joint attention in multiple first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[38] Eddy Ilg, Nikolaus Mayer, T Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462 – 2470, 2017.

[39] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[40] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 5839–5849, 2017.

[41] Takeo Kanade and Martial Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[42] Vasiliy Karasev, Alper Ayvaci, Bernd Heisele, and Stefano Soatto. Intent-aware long-term prediction of pedestrian motion. In *the IEEE International Conference on Robotics and Automation*, pages 2543–2549, 2016.

[43] Christoph G Keller and Dariu M Gavrila. Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506, 2014.

[44] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[45] Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Discovering objects of joint attention via first-person sensing. In *the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[47] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Proceedings of the European Conference on Computer Vision*, pages 201–214, 2012.

[48] Julian Francisco Pieter Kooij, Nicolas Schneider, Fabian Flohr, and Dariu M Gavrila. Context-based pedestrian path prediction. In *Proceedings of the European Conference on Computer Vision*, pages 618–633, 2014.

[49] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming mul-

tiple people tracker. In *IEEE International Conference on Computer Vision Workshops*, pages 120–127, 2011.

[50] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

[51] Namhoon Lee and Kris M Kitani. Predicting wide receiver trajectories in american football. In *the IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, 2016.

[52] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[53] Tung-Sing Leung and Gerard Medioni. Visual navigation aid for the blind in dynamic environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 153 – 158, 2014.

[54] Cheng Li and Kris M Kitani. Pixel-level hand detection in ego-centric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2013.

[55] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[56] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *IEEE International Conference on Robotics and Automation*, pages 464–469, 2010.

[57] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[58] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[59] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 774 – 782, 2017.

[60] Barbara Majecka. Statistical models of pedestrian behaviour in the forum. *Master's thesis, School of Informatics, University of Edinburgh*, 2009.

[61] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009.

[62] Mathew Monfort, Anqi Liu, and Brian D Ziebart. Intent prediction and trajectory forecasting via predictive inverse linear-quadratic regulation. 2015.

[63] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[64] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, pages 807–814, 2010.

[65] Yun Seok Nam, Bum Hee Lee, and Moon Sang Kim. View-time based moving obstacle avoidance using stochastic prediction of obstacle motion. In *Proceeding of the IEEE International Conference on Robotics and Automation*, volume 2, pages 1081–1086, 1996.

[66] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[67] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 663–670, 2000.

[68] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3153–3160, 2011.

[69] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016.

[70] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. 3D Social Saliency from Head-mounted Cameras. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1–9, 2012.

[71] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the International Conference on Computer Vision*, pages 261–268, 2009.

[72] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European Conference on Computer Vision*, pages 452–465, 2010.

[73] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, 2012.

[74] Yair Poleg, Chetan Arora, and Shmuel Peleg. Head motion signatures from egocentric videos. In *Proceedings of the Asian Conference on Computer Vision*, pages 1–15, 2014.

[75] Eike Rehder, Florian Wirth, Martin Lauer, and Christoph Stiller. Pedestrian prediction by planning using deep neural networks. In *IEEE International Conference on Robotics and Automation*, pages 1–5, 2018.

[76] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 1–9, 2015.

[77] Nicholas Rhinehart and Kris M. Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017.

[78] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision*, pages 549–565, 2016.

[79] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2564–2571, 2011.

[80] M. S. Ryoo, Thomas J. Fuchs, Lu Xia, J.K. Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 295–302, 2015.

[81] Michael S. Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2737, 2013.

[82] Akanksha Saran, Damien Teney, and Kris M Kitani. Hand parsing for fine-grained recognition of human grasps in monocular images. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–7, 2015.

[83] Nicolas Schneider and Dariu M. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *Proceedings of the German Conference on Pattern Recognition*, pages 174–183, 2013.

[84] Paul Scovanner and Marshall F Tappen. Learning pedestrian dynamics from the real world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 381–388, 2009.

[85] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1501–1510, 2017.

[86] Satoshi Tadokoro, Masaki Hayashi, Yasuhiro Manabe, Yoshihiro Nakami, and Toshi Takamori. On motion planning of mobile robots which coexist and cooperate with human. In *Proceeding of the IEEE International Conference on Intelligent Robots and Systems*, page 2518. IEEE, 1995.

[87] Titus J. J. Tang and Wai Ho Li. An assistive eyewear prototype that interactively converts 3d object locations into spatial audio. In *Proceedings of the ACM International Symposium on Wearable Computers*, pages 119–126, 2014.

[88] Meng Keat Christopher Tay and Christian Laugier. Modelling smooth paths using gaussian processes. In *Field and Service Robotics*, pages 381–390. Springer, 2008.

[89] Ya Tian, Yong Liu, and Jindong Tan. Wearable navigation system for the blind people in dynamic environments. In *Proceedings of the Cyber Technology in Automation, Control and Intelligent Systems*, pages 153 – 158, 2013.

[90] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems*, pages 1–6, 2015.

[91] Peter Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 797–803, 2010.

[92] Peter Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. Robot navigation in dense human crowds: the case for cooperation. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 2153–2160, 2013.

[93] Adrien Treuille, Seth Cooper, and Zoran Popović. Continuum crowds. *ACM Transactions on Graphics*, 25(3):1160–1168, 2006.

[94] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2008.

[95] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2295–2304, 2016.

[96] D. Xie, S. Todorovic, and S. C. Zhu. Inferring dark matter and dark energy from videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2224–2231, 2013.

[97] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1352, 2011.

[98] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[99] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Pedestrian behavior understanding and prediction with deep neural networks. In *Proceedings of the European Conference on Computer Vision*, pages 263–279, 2016.

[100] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2629–2638, 2016.

[101] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[102] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.

[103] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851 – 1860, 2017.

[104] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 3*, pages 1433–1438. AAAI Press, 2008.

List of Publications

1. Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani and Yoichi Sato, Future Person Localization in First-Person Videos, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7593–7602, 2018.
2. 八木拓真, マンガラムカーティケヤ, 米谷竜, 佐藤洋一, 一人称視点映像における人物位置予測, 第 21 回画像の認識・理解シンポジウム (MIRU), 2018.
3. 八木拓真, マンガラムカーティケヤ, 米谷竜, 佐藤洋一, 一人称視点映像における人物位置予測, 第 211 回 CVIM 研究会, 2018.