# Forecast the Plausible Paths in Crowd Scenes *

**Hang Su, Jun Zhu, Yinpeng Dong, Bo Zhang**
Tsinghua National Lab for Information Science and Technology
State Key Lab of Intelligent Technology and Systems
Center for Bio-Inspired Computing Research
Department of Computer Science and Technology, Tsinghua University, Beijing, China
{suhangss,dcszj,dongyp13,dcszb}@tsinghua.edu.cn

## Abstract

Forecasting the future plausible paths of pedestrians in crowd scenes is of wide applications, but it still remains as a challenging task due to the complexities and uncertainties of crowd motions. To address these issues, we propose to explore the inherent crowd dynamics via a social-aware recurrent Gaussian process model, which facilitates the path prediction by taking advantages of the interplay between the rich prior knowledge and motion uncertainties. Specifically, we derive a social-aware LSTM to explore the crowd dynamic, resulting in a hidden feature embedding the rich prior in massive data. Afterwards, we integrate the descriptor into deep Gaussian processes with motion uncertainties appropriately harnessed. Crowd motion forecasting is implemented by regressing relative motion against the current positions, yielding the predicted paths based on a functional object associated with a distribution. Extensive experiments on public datasets demonstrate that our method obtains the state-of-the-art performance in both structured and unstructured scenes by exploring the complex and uncertain motion patterns, even if the occlusion is serious or the observed trajectories are noisy.

## 1 Introduction

Tracking pedestrians in crowd draws significant attentions recently because of its vital and wide applications, e.g., profiling group behaviors [Zhou *et al.*, 2015], and detecting abnormal crowd behaviors [Li *et al.*, 2014]. Besides, it is valuable to consider the problem beyond the current scope of activity analysis by inferring crowd dynamics about the future. It is extremely attractive to forecast the plausible paths in a crowd scene in numerous applications, e.g., navigation of autonomous vehicles, recognition of potential risks in video surveillance, etc.

*The work was supported by the National Basic Research Program (973 Program) of China (No. 2013CB329403), National NSF of China Projects (Nos. 61571261, 61620106010, U1611461), the Youth Top-notch Talent Support Program, China Postdoctoral Science Foundation (No. 2015M580099) and Tiangong Institute for Intelligent Computing.

Previously, crowd motion prediction is studied in the areas of crowd simulation [Long *et al.*, 2016; Godoy *et al.*, 2016], but few attempts have been made in predicting the future paths of crowd in real world scenarios from the vision view. In this paper, we investigate the problem of crowd motion prediction by taking advantages of the inherent crowd dynamics and history observations. Compared with the tasks of understanding or recognizing the present crowd activities, prediction is more challenging since we do not have complete observations. Therefore, it requires a deeper understanding in determining not only what the activity is but how the activity should be unfolded.

Despite of its importance, motion prediction is a new topic and not well-studied in the vision discipline. Ever since the seminal works [Kitani *et al.*, 2012], it has witnessed some recent efforts in predicting the unobserved future action by exploring the spatio-temporal contextual information [Li and Fu, 2014; Robicquet *et al.*, 2016]. Despite this, motion prediction of a crowd is still not well addressed due to the *complex and non-linear patterns* in spatio-temporal varying structures [Yi *et al.*, 2016; Alahi *et al.*, 2016]. Human motions in a crowd are of unavoidable uncertainties which are produced by the physical and social environments [Helbing and Johansson, 2011], e.g., the scene structures, the presence of other people and the time of day. Researchers have proposed to cope with the *motion uncertainties* in a probabilistic manner including reinforcement learning [Ziebart *et al.*, 2008] and Gaussian Processes (GP) [Ellis *et al.*, 2009a]. However, most of the existing approaches operate under scenarios where motion prediction is contrived to a few individuals in simple scenes.

### 1.1 Our Proposal

To address the aforementioned challenges, we formulate the crowd motion anticipation as a spatio-temporal sequence forecasting problem based on a proposed *social-aware recurrent Gaussian processes* model. The model fully encapsulates the social-aware long short-term memory (LSTM) networks while retains the non-parametric probabilistic advantages of Gaussian processes (GP) to address the complex nonlinear transitions and uncertainties of crowd dynamics, respectively.

It is believed that a human's amazing ability to visualize the future is primarily driven by the *rich prior knowledge* about the visual world. The idea is also feasible for

computer-aided motion prediction with visual data because of the self-organizing phenomena for a moving crowd [Helbing and Johansson, 2011]. Individuals in crowd usually do not take complicated decisions between various possible alternative behaviors, but learn from the neighbors "automatically" and try the strategies. In this case, the prior knowledge learnt from other pedestrians is informative in predicting the invisible motion for a stranger.

Recently, the LSTM model [Lipton, 2015] is proved successful in sequence prediction [Srivastava *et al.*, 2015; Su *et al.*, 2016] partly due to its powerful capabilities in capturing the information of massive data. It inspires us to explore the inherent crowd dynamics with LSTM and implement motion prediction by interpreting the descriptor. To particularly address the social effects in crowd motion [Helbing and Johansson, 2011], we introduce an additional social-aware gate to LSTM, endowing the history motions into a hidden feature.

Despite its success, the standard LSTM state transition structure is entirely deterministic, thus lacks the capabilities in modeling uncertainties. However, crowd motion is always with varying degrees of uncertainties due to various factors. In this case, we propose a novel representation of crowd trajectories built on recent developments in deep Gaussian processes (DGP) [Dai *et al.*, 2016], which quantifies uncertainty and extracts full predictive distributions from latent variables. In this paper, we propose to drive the deep Gaussian processes with the latent features learnt form the social-aware LSTM. The predicted path is obtained by a underlying stochastic process with its own probability of being the "real" path. The problem is solved by joint optimizing a functional composition of Gaussian processes along with the recurrent model (i.e., social-aware LSTM) via a variational inference procedure. In summary, our model can therefore be interpreted as a deep Gaussian process driven by dynamical inputs, or as a deep recurrent network with probabilistic outputs, which naturally fits into the standard computational framework of deep learning models.

To the best of our knowledge, this is the first attempt to address the crowd motion prediction by exploring the dynamic uncertainties in big data scenarios. Therefore, our main contribution is to emphasize the connection of the social-aware LSTM and Gaussian processes. Compared with motion prediction with LSTM model [Su *et al.*, 2016; Alahi *et al.*, 2016], we address the implicit data uncertainties by associating the predicted trajectories with a distribution of the Gaussian processes. Motion prediction can also be implemented when missing values occur, therefore, the partially observed instances are also correctly accounted, such that all available information is exploited to increase the strength of the inference model. Besides, motion prediction based on stand-alone Gaussian processes [Ellis *et al.*, 2009a] learns only pairwise correlations, and is unable to account for long-term dependencies. Our method provides a direction to address the issue in a recurrent fashion, and explores rich prior knowledge embedded in massive data with the social-aware LSTM. Extensive experiments demonstrate that the algorithm supports effective anticipation of the motion tendencies in both structured and unstructured scenes.

## 2 Methodology

Considering the smoothness and continuity of motions in crowds, we propose to generate the future trajectories with social-aware recurrent Gaussian processes. In this section, we first overview the framework, and then elaborate the social-aware LSTM to explore the crowd dynamics, resulting a hidden feature vector; moreover, we present the deep Gaussian processes to conduct motion forecasting, which is of good capability in modeling the uncertainty.

### 2.1 Overview

In this paper, we aim to generate the future paths based on a set of history trajectories by training a deep architecture. Different from the previous works based on crowd tracklets [Su *et al.*, 2016; Alahi *et al.*, 2016], we advocate to anticipate the future paths in the velocity field $\{\mathbf{v}_t\}$ since the velocity remains stable unless obstacles occur or significant interaction force exists [Helbing and Johansson, 2011], yielding an easier estimation compared with the trajectories.

For generating the future sequence, we compute the Bayesian predictive distribution as $p(\hat{\mathbf{v}}|\mathbf{v})$, where $\mathbf{v}$ and $\hat{\mathbf{v}}$ indicate the present (i.e., observed) and future crowd velocity, respectively. Rather than generating the future path directly [Ellis *et al.*, 2009a], we introduce a latent variable $\mathbf{z}_t$ to address the varying degrees of uncertainty for crowd motions. Motion anticipation can be implemented as an integral of conditional probabilities over the latent variable as

$$p(\hat{\mathbf{v}}_{t+1}|\mathbf{v}_1, \cdots, \mathbf{v}_t) = \int p(\hat{\mathbf{v}}_{t+1}|\mathbf{Z}, \mathbf{v}_1, \cdots, \mathbf{v}_t) p(\mathbf{Z}) d\mathbf{Z}, \tag{1}$$

where $\{\mathbf{v}_t\}_{t=1}^{T}$ is the observed velocity; $\mathbf{Z} = \{\mathbf{z}_t^{(j)}\}$ is the hidden variable; and $\hat{\mathbf{v}}_{t+1}$ is the forecasted invisible velocity for the next timestep. To obtain a long term prediction, the above integral can be evaluated recursively. The plausible paths can be obtained by integrating the velocity over time.

In Fig. 1, we outline the general framework of the proposed social-aware recurrent Gaussian processes to anticipate the invisible paths in crowd scenes. At the core of the model is the use of deep Gaussian processes to anticipate future velocities of pedestrians given the current velocity. We then anticipate the crowd velocity by employing the Gaussian Processes (GPs) as nonparametric prior distributions over the latent variable $\mathbf{z}_t$. The model defines a forward (or generative) mapping from the latent space to observation space that is governed by Gaussian processes.

To further improve the capability in capturing the temporal structures of Gaussian processes, we propose to constrain the latent space of DGP via a dynamical prior with

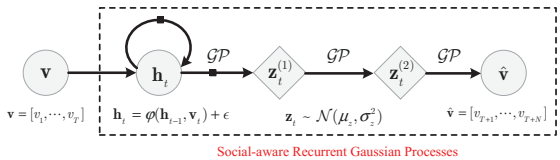


Figure 1: Outline of the social-aware recurrent Gaussian processes for forecasting the plausible paths in crowd scenes.

an LSMT network, based on which the rich prior knowledge in massive data is also well-explored. Specifically, we construct a recurrent structure by transforming the original input velocities to latent feature space with LSTM. Therefore, the inherent crowd dynamics is explored by the social-aware LSTM, yielding hidden features to drive the subsequential deep Gaussian processes as $f_{\mathbf{Z}} \sim \mathcal{GP}\big(\mathbf{m}, K(\mathbf{h}, \mathbf{h}')\big)$.

## 2.2 Social-Aware LSTM

In this section, we propose to learn a representation of the crowd dynamics with a recurrent LSTM by incorporating the nearby trajectories. Although human behaviors often seem to be "chaotic", irregular and unpredictable, crowd behaviors always exhibit systematic motions, which are more significant compared with the unknown fluctuations. The main reason is that pedestrians tend to learn from others, and obey the physical constraints and social rules [Helbing and Johansson, 2011]. In this case, we propose to learn a hidden feature that effects a pedestrian's decision making in a data-driven manner with LSTM [Lipton, 2015].

To this end, we train an LSTM architecture similar to the encoder-decoder framework [Sutskever *et al.*, 2014], which consists of two networks, an encoding network and a forecasting network. Specifically, the encoder LSTM recursively processes the sequence of the past velocity to come up with an inherent hidden feature; and the forecasting LSTM then computes the output sequence to extrapolate the motions beyond what has been observed ($\hat{\mathbf{v}}_t$ with $t > T$). In particular, the hidden feature vector $\mathbf{h}_t$ from the LSTM unit is computed by multiplying the updated cell state that passed through a tanh non-linearity with an output gate's activation as

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{2}$$

where $\mathbf{o}_t$ is the output gate, and the cell state $\mathbf{c}_t$ is a memory unit as

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{vc}\mathbf{v}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \tag{3}$$

where $\mathbf{f}_t$ and $\mathbf{i}_t$ are the forget and input gates, respectively; $\mathbf{v}_t$ is the current velocity; $\mathbf{W}_{vc}$ and $\mathbf{W}_{hc}$ are the weight parameters with proper size; and $\mathbf{b}_c$ is the bias.

Different from the individual behaviors, pedestrians in crowds always adjust their paths by implicitly reasoning about the motions of their surrounding neighbors [Helbing and Johansson, 2011]. Therefore, we introduce a social-aware gate to address the interaction among neighboring pedestrians. The hidden features of neighboring agents are shared across the memory unit by controlling the activation of a social-aware gate. The memory state is updated as

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{vc}\mathbf{v}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) + \mathbf{s}_t \odot \tanh(\mathbf{W}_{vs}\tilde{\mathbf{v}}_t + \mathbf{W}_{hs}\tilde{\mathbf{h}}_{t-1} + \mathbf{b}_s), \tag{4}$$

where $\mathbf{s}_t$ is the social-aware gate that controls the information flow from the surrounding neighbors; and $\tilde{\mathbf{v}}_t$ and $\tilde{\mathbf{h}}_{t-1}$ are, respectively, the velocity and hidden states by pooling the corresponding parameters of neighboring pedestrians that are detected using the coherent filter [Zhou *et al.*, 2012].

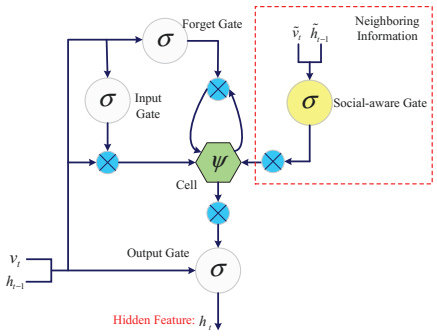


Figure 2: A social-aware LSTM unit.

The social-aware LSTM unit implicitly reflects hidden factors that effect a pedestrian's motion of a social force model [Helbing and Johansson, 2011], which is proved particularly successful in describing the social behaviors. The first and second terms are corresponding to a pedestrian's inner factor; the third term is corresponding to the group factor in a pedestrian's vicinity, e.g., the attractive and repulsive forces in a crowd. In this case, we propose to share information between neighboring pedestrians by integrating the states across each LSTM, as is illustrated in Fig. 2. Furthermore, we form both the encoding and forecasting networks by stacking several social-aware LSTM units.

## 2.3 Crowd Motion Modeling with DGPs

Although the social-aware LSTM is successful in modeling the crowd dynamics, the transition is entirely deterministic and thus fail to model motion uncertainties. To address this issue, we propose to incorporate deep Gaussian processes with the LSTM model for motion representation in a probabilistic manner, as shown in Fig. 1. In this case, the prediction is operated on a distribution rather than a single point, which facilitates the uncertain quantification or data imputation. In this paper, we use the latent features learnt from the social-aware LSTM to drive the deep Gaussian processes.

For generating the future sequence, we compute the Bayesian predictive distribution as $p(\hat{\mathbf{v}}|\mathbf{v})$, where $\mathbf{v}$ and $\hat{\mathbf{v}}$ indicate the present (i.e., observed) and predicted crowd velocities, respectively. Specifically, our algorithm conducts motion prediction with a Gaussian process to explicitly model motion uncertainties. In order to define a generative mapping from the latent space to the observation space, we extend the Gaussian process regression to a deep architecture [Damianou and Lawrence, 2013]. It consists of a cascade of $L$ hidden layers of latent variables, an intermediate layer contributes the output of that layer and acts as an input for the generative procedure of the subsequent layer.

Therefore, the mapping between layers is governed by a Gaussian process, which corresponds to a separate GP with mean and covariance functions. Moreover, we employ the hidden features derived from the social-aware LSTM to drive the Gaussian processes as

$$f_{\hat{\mathbf{v}}} \sim \mathcal{GP}(\boldsymbol{\mu}_{\hat{\mathbf{v}}}, k(\mathbf{z}, \mathbf{z}')),\ f_{\mathbf{z}} \sim \mathcal{GP}(\boldsymbol{\mu}_{\mathbf{z}}, k(\mathbf{h}, \mathbf{h}')), \tag{5}$$

where $\mathbf{z}$ is the latent variable corresponding to the deep Gaussian processes; $\mathbf{h}$ is the hidden variable derived from the

social-aware LSTM, which is the unobserved input for the generative procedure of the subsequent layer; and we use RBF kernels for the Gaussian processes.

### 2.4 Optimization

In this section, we introduce a variational inference framework for training the social-aware recurrent Gaussian processes model which allows joint training of the model from scratch. A Bayesian training procedure requires optimization of the predictive density as

$$\log p(\hat{\mathbf{v}}|\mathbf{v}) = \log \int_{\mathbf{H},\mathbf{Z}} p(\hat{\mathbf{v}}|\mathbf{Z})p(\mathbf{Z}|\mathbf{H})p(\mathbf{H}|\mathbf{v}), \quad (6)$$

where $\mathbf{H} = [\mathbf{h}_t]$ is the latent variables derived from the social-aware LSTM to drive the deep Gaussian processes; and $\mathbf{Z} = [\mathbf{z}]$ are the latent variables involved in a Gaussian process as $f \sim \mathcal{GP}(\mathbf{m}, k(\mathbf{z}, \mathbf{z}'))$ with $\mathbf{m}$ and $k(\cdot)$ being the mean and covariance matrix, respectively. However, the integral is intractable as $\mathbf{Z}$ and $\mathbf{H}$ appear nonlinearly inside the inverse of the covariance matrix.

To address this issue, we invoke the variational Bayesian methodology to derive a variational lower bound. Similarly as the variational inference method in [Titsias and Lawrence, 2010], we introduce underlying latent function values $\mathbf{F}_{\hat{\mathbf{v}}} = [f(\mathbf{z})]$ (noisy-free version of $\hat{\mathbf{v}}$ in Eq. (5)) and $\mathbf{F}_{\mathbf{z}} = [f(\mathbf{h})]$ (noisy-free version of $\mathbf{z}$ in Eq. (5)), and a variational distribution $q(\Theta)$ ($\Theta$ is a set of random variables upon which the models depend). We then derive a variational lower bound via Jensen's inequality as

$$\log p(\hat{\mathbf{v}}|\mathbf{v}) \geq \int_{\mathbf{H},\mathbf{Z}} q(\Theta) \log \frac{p(\hat{\mathbf{v}}, \mathbf{F}_{\hat{\mathbf{v}}}, \mathbf{F}_{\mathbf{z}}, \mathbf{Z}, \mathbf{H}|\mathbf{v})}{q(\Theta)}, \quad (7)$$

where the joint distribution can be expanded as

$$p(\hat{\mathbf{v}}, \mathbf{F}_{\hat{\mathbf{v}}}, \mathbf{F}_{\mathbf{z}}, \mathbf{Z}, \mathbf{H}|\mathbf{v}) = p(\hat{\mathbf{v}}|\mathbf{F}_{\hat{\mathbf{v}}})p(\mathbf{F}_{\hat{\mathbf{v}}}|\mathbf{Z})p(\mathbf{Z}|\mathbf{F}_{\mathbf{z}})p(\mathbf{F}_{\mathbf{z}}|\mathbf{H})p(\mathbf{H}|\mathbf{v}), \quad (8)$$

where $p(\hat{\mathbf{v}}|\mathbf{F}_{\hat{\mathbf{v}}})$ and $p(\mathbf{Z}|\mathbf{F}_{\mathbf{z}})$ are Gaussian distribution with zero means. Note that the lower bound is still intractable since $\mathbf{Z}$ and $\hat{\mathbf{v}}$ still appears nonlinearly inside $p(\mathbf{F}_{\mathbf{z}}|\mathbf{H})$ and $p(\mathbf{F}_{\hat{\mathbf{v}}}|\mathbf{Z})$, respectively.

To address this issue, we propose to apply variational inference after expanding the GP prior with data augmentation. More formally, we introduce a separate set of $M$ inducing variables $\mathbf{U}_{\hat{\mathbf{v}}}$ and $\mathbf{U}_{\mathbf{z}}$ evaluated at a set of inducing input locations, which are drawn from the GP prior. Using these inducing variables, the augmented joint probability density takes the form as

$$\begin{aligned} p(\hat{\mathbf{v}}, \mathbf{F}_{\hat{\mathbf{v}}}, \mathbf{F}_{\mathbf{z}}, \mathbf{U}_{\hat{\mathbf{v}}}, \mathbf{U}_{\mathbf{z}}, \mathbf{Z}, \mathbf{H}|\mathbf{v}) = \\ p(\hat{\mathbf{v}}|\mathbf{F}_{\hat{\mathbf{v}}})p(\mathbf{F}_{\hat{\mathbf{v}}}|\mathbf{U}_{\hat{\mathbf{v}}}, \mathbf{Z})p(\mathbf{U}_{\hat{\mathbf{v}}}) \cdot \\ p(\mathbf{Z}|\mathbf{F}_{\mathbf{z}})p(\mathbf{F}_{\mathbf{z}}|\mathbf{U}_{\mathbf{z}}, \mathbf{H})p(\mathbf{U}_{\mathbf{z}})p(\mathbf{H}|\mathbf{v}). \end{aligned} \quad (9)$$

We drop the auxiliary inducing inputs for simplicity since they are not random variables but variational parameters [Titsias, 2009]. Note that $\mathbf{F}_{\hat{\mathbf{v}}}$ and $\mathbf{U}_{\hat{\mathbf{v}}}$ are draws from the same Gaussian Processes, so that $p(\mathbf{U}_{\hat{\mathbf{v}}})$ and $p(\mathbf{F}_{\hat{\mathbf{v}}}|\mathbf{U}_{\hat{\mathbf{v}}}, \mathbf{Z})$ are also Gaussian distributions, which is similarly for $p(\mathbf{U}_{\mathbf{z}})$ and $p(\mathbf{F}_{\mathbf{z}}|\mathbf{U}_{\mathbf{z}}, \mathbf{H})$.

In the following, we derive the lower bound of the observed and hidden layers based on variational inference [Dai *et al.*, 2016], respectively. First, we derive the variational bound of the observed layer $p(\hat{\mathbf{v}})$. Assuming a particular form of the variational distribution of $\mathbf{F}_{\hat{\mathbf{v}}}$ and $\mathbf{U}_{\hat{\mathbf{v}}}$, we have $q(\mathbf{F}_{\hat{\mathbf{v}}}, \mathbf{U}_{\hat{\mathbf{v}}}|\mathbf{Z}) = p(\mathbf{F}_{\hat{\mathbf{v}}}|\mathbf{U}_{\hat{\mathbf{v}}}, \mathbf{Z})q(\mathbf{U}_{\hat{\mathbf{v}}})$. Using this factorization, the free energy of the observed layer in Eq. (9) can be lower bounded by

$$\mathcal{L}^{(o)} \geq \left\langle \log p(\hat{\mathbf{v}}|\mathbf{F}_{\hat{\mathbf{v}}}) - KL\big(q(\mathbf{U}_{\hat{\mathbf{v}}})\|p(\mathbf{U}_{\hat{\mathbf{v}}})\big)\right\rangle_{p(\mathbf{F}_{\hat{\mathbf{v}}}|\mathbf{U}_{\hat{\mathbf{v}}},\mathbf{Z})q(\mathbf{U}_{\hat{\mathbf{v}}})q(\mathbf{Z})}, \quad (10)$$

where the first term represents the entropy with respect to a distribution, and the $KL$ denotes the Kullback-Leibler divergence [Titsias and Lawrence, 2010].

Next, we derive the variational posterior distributions of the hidden layers, which is different from the variational bound of the observed layer since it also depends on the output variables. For the hidden layers, the variational posterior distribution is defined as $q(\mathbf{F}_{\mathbf{z}}, \mathbf{U}_{\mathbf{z}}|\mathbf{Z}, \mathbf{H}) = p(\mathbf{F}_{\mathbf{z}}|\mathbf{U}_{\mathbf{z}}, \mathbf{H})q(\mathbf{U}_{\mathbf{z}}|\mathbf{Z})$. Similar to the observed layer, a lower bound of the free energy for the hidden layer can be derived as

$$\mathcal{L}^{(h)} \geq \left\langle \log p(\mathbf{Z}|\mathbf{F}_{\mathbf{z}}) - KL\big(q(\mathbf{U}_{\mathbf{z}}|\mathbf{Z})\|p(\mathbf{U}_{\mathbf{z}})\big)\right\rangle_{Q(\mathbf{Z},\mathbf{H})}. \quad (11)$$

where $Q(\mathbf{Z}, \mathbf{H}) = p(\mathbf{F}_{\mathbf{z}}|\mathbf{U}_{\mathbf{z}}, \mathbf{Z})q(\mathbf{U}_{\mathbf{z}}|\mathbf{Z})q(\mathbf{Z})q(\mathbf{H})$.

In summary, the lower bound in Eq. (7) is as

$$\mathcal{L} \geq \mathcal{L}^{(o)} + \mathcal{L}^{(h)} + \mathcal{H}_{q(\mathbf{Z})} - KL(q(\mathbf{Z})\|p(\mathbf{Z})), \quad (12)$$

where $\mathcal{H}_{q(\mathbf{Z})}$ denotes the entropy of the variational distribution. Note that the corresponding terms in Eq. (12) involve only known Gaussian distribution and are therefore tractable [Titsias and Lawrence, 2010].

The cost of generating that hidden feature of social-aware LSTM is then defined as the negative logarithm of the likelihood derived from the variational distribution of $\mathbf{H}$. We use Backpropagation Through Time (BPTT) [Lipton, 2015] to compute the gradients of the parameters and Stochastic Gradient Descent (SGD) to optimize the model parameters of social-aware LSTM. Additionally, by transforming the original velocity space to latent variables with an recurrent LSTM, the parameter size does not increase along with the sample size, which alleviates optimization burden significantly.

## 3 Experiments

In this section, we demonstrate the effectiveness of our proposed algorithm in crowd motion prediction in both *structured* and *unstructured scenes*. In a structured crowd scene, crowds move coherently in common directions by obeying social rules or scene constraints, and the motion direction does not vary frequently, e.g., crowd motions in marathon races or along aisles. Whereas, the unstructured crowd scenes represent the scenarios with chaotic or random crowd motions where the participants move freely, e.g., crowd motions in railway stations or open squares. In this case, the uncertainty of crowd motions in unstructured scenes is more significant than in structured scenes.

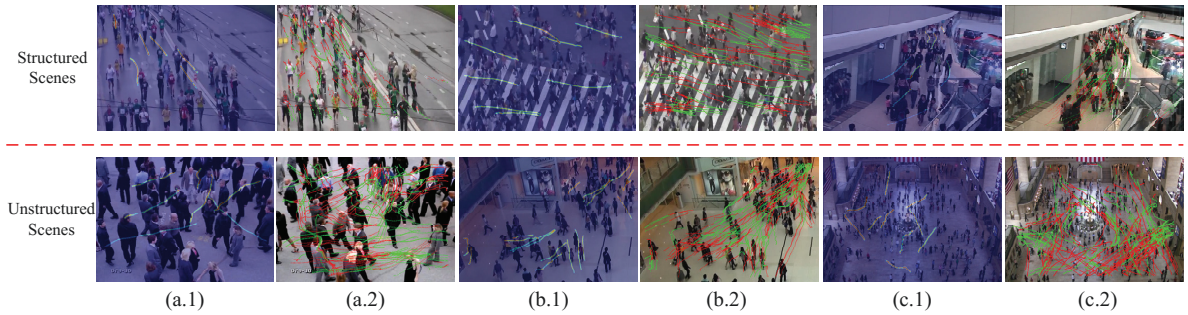


Figure 3: Sample results of trajectory prediction, where the top and bottom rows are corresponding to the results in structured and unstructured scenes, respectively. The figure pairs (a), (b), (c) are corresponding to crowd motions in different scenes. In the first figure of each pair (a.1), (b.1) and (c.1), we demonstrate the probability distribution of a few predicted paths; and in the second figure (a.2), (b.2) and (c.2), we show samples of trajectory prediction, in which the red fragments are corresponding to the observed trajectories, and the green fragments are generated via our proposed algorithm. (Best view in color.)

Table 1: Error of Path Prediction in Structured Scenes

| | LSTM | cLSTM | GP | DGP | SRGP | SRGP without social gate | SRGP without GP | SRGP with Trajectory |
|---|---|---|---|---|---|---|---|---|
| ADE | 4.17 ± 0.96 | 3.28±0.63 | 6.53 ± 1.32 | 4.23 ± 0.98 | **2.88± 0.59** | 3.91± 0.83 | 3.35± 0.71 | 3.45± 1.08 |
| FDE | 6.71 ± 1.56 | 5.33±1.39 | 9.45 ± 2.71 | 7.08 ± 1.94 | **4.97± 1.31** | 6.03± 1.63 | 5.69± 1.33 | 5.36± 1.71 |

### 3.1 Dataset and Experiment Setting

**Datasets.** Experiments are conducted on two public datasets: the CUHK Crowd Dataset [Shao *et al.*, 2014] that includes hundreds of crowd videos with different densities and perspective scales in many environments with each containing thousands of key point trajectories; and the subway station dataset [Zhou *et al.*, 2011], which is a 30-minute sequence collected in the New York Grand Central Station, resulting in more than 40,000 keypoint trajectories in total. The ground truth keypoint trajectories are available for both datasets. In our experiments, we randomly select a half of the trajectories to train the model, and keep the rest for testing.

**Setting.** Note that evaluating the true accuracy of future trajectories is difficult since we do not have access to the "ground truth" of the future. As a proxy, we evaluate the performance on the well-described trajectories, i.e., we take a fragment of tracklets as the input (e.g., 1/2 of each trajectory in this paper), and generate the rest of them for prediction (e.g., 1/2 of each trajectory). In our experiments, we use a social-aware LSTM with 128 hidden units, i.e., the input trajectories are mapped to a 128-dimensional hidden feature vector ($\mathbf{h}_t$); moreover, we set the latent variational variable in deep Gaussian processes as 8-dimensional vectors ($\mathbf{z}_t$). Moreover, we use one LSTM layer, and a two-layer Gaussian process model in the social-aware LSTM and deep Gaussian processes modules, respectively.

**Comparison methods.** We conduct the crowd velocity anticipation based on our proposed Social-aware Recurrent Gaussian Processes (SRGP), and then obtain the future trajectories by integrating the velocities. In order to demonstrate the contribution of each part of the proposed architecture, we compare variants of the proposed algorithms by removing the social gate of SRGP (SRGP without social gate), removing the module of Gaussian processes (SRGP without Gaussian processes), and predicting the future paths with SRGP over trajectories, respectively. The results for SRGP without GP came from predicting velocities from social-aware LSTM hidden features directly by adding an output layer.

Apart from these methods, we also conduct the trajectories prediction based on the alternative methods including Long Short-term Memory predictor (LSTM) [Srivastava *et al.*, 2015], Coherent LSTM (cLSTM) predictor [Su *et al.*, 2016], Gaussian Processes (GP) [Ellis *et al.*, 2009b], Deep Gaussian Processes (DGP) [Damianou and Lawrence, 2013]. The results for LSTM and cLSTM are obtained by incorporating an output layer.

**Metrics.** In each experiment, we evaluate the performance with two metrics including the Average Displacement Error (**ADE**) and the Final Displacement Error (**FDE**) in terms of the pixel displacement between the estimated trajectories and the ground truth [Alahi *et al.*, 2016].

### 3.2 Motion Prediction in Structured Scenes

In this section, we demonstrate the performance of our algorithm in structured scenes, which are collected from the CUHK Crowd Dataset. Qualitative results of motion forecasting are depicted in the top row of Fig. 3, in which the sub-figures (a), (b), (c) are corresponding to crowd motions in a marathon race, people cross the road via zebra crossings, and crowd motions in a plaza aisle. In the first figures of each pair, we demonstrate the probability distribution of the predicted paths; and we show more samples of trajectory prediction in the second figures, in which the red fragments are corresponding to the observed trajectories, and the green fragments are generated via our proposed algorithm. The results demonstrate that our algorithm is capable of capturing the evolution characteristics of crowd motions as well as the uncertainties in motions. Note that the environment or social

Table 2: Error of Path Prediction in Unstructured Scenes

| | LSTM | cLSTM | GP | DGP | SRGP | SRGP without social gate | SRGP without GP | SRGP with Trajectory |
|---|---|---|---|---|---|---|---|---|
| ADE | 6.37 ± 1.83 | 5.23±1.56 | 7.13 ± 1.44 | 4.89 ± 1.07 | **3.12± 0.78** | 4.56± 0.98 | 5.11± 0.78 | 3.98± 1.03 |
| FDE | 9.83 ± 3.12 | 8.53±2.46 | 10.65 ± 2.85 | 7.88± 1.93 | **5.62± 1.88** | 6.77± 1.73 | 8.49± 2.01 | 6.31± 2.07 |

constraints are restricted in a structured scene, and humans are willing to follow the rules when making their decisions, e.g., participants have a common destination in marathon, follow the zebra crossings, or move along the aisle. In this case, the tendency of pedestrians in structured scenes are well explored via the rich knowledge embedded in the massive training data, such as the physical scene characteristics and neighboring information.

The quantitative evaluation is reported in Table. 1 in terms of the average displacement error (ADE) and final displacement error (FDE). Obviously, our proposed algorithm outperforms the alternative methods since the *prior knowledge* is explored in a data-driven manner and the unavoidable *uncertainties* are also well addressed. The performance for *LSTM* [Srivastava *et al.*, 2015] and *coherent LSMT (cLSTM)* [Su *et al.*, 2016], although inferior to our proposed algorithm, excels the methods based on Gaussian processes. The main reason is because human motion in a structured scene is confined by the physical restriction, thus with a self-organizing characteristics. In this case, the prior knowledge learnt from the training procedure, e.g., semantic of the scenes or social rules, is essential in exploring the crowd tendency. The coherent motion is also critical in structured scene, since pedestrians learn from neighbors easily in such scenario, which results in an improved performance of *cLSTM* compared with *LSTM*. The performance degrades significantly when removing the social gate of SRGP due to the same reason.

Additionally, it is observed that motion prediction based on *Gaussian processes (GP)* [Ellis *et al.*, 2009b] and *deep Gaussian processes (DGP)* [Damianou and Lawrence, 2013] is not desirable since they fail to explore the prior knowledge whereas the motion uncertainty is not significant. Nevertheless, the Gaussian Processes module also contributes to the final performance because of the implicit motion uncertainties. The performance of *SRGP without GP* is obviously inferior to the approach when considering of the GP module. Finally, we implement motion prediction using the *SRGP with trajectory*, which is not comparable to our proposed algorithm. The main reason is that the velocity remains smooth when no significant force acts on a pedestrian.

### 3.3 Motion Prediction in Unstructured Scenes

In this section, we demonstrate the performance of our algorithm in unstructured scenes, which are collected from the CUHK Crowd Dataset and the subway station dataset. We demonstrate sample results of motion prediction in the bottom row of Fig. 3, in which the figure pairs (a), (b), (c) are corresponding to crowd motions in a square, crowd motions in a shopping mall, and crowd motions in New York Grand Central Station. Similar as the setting in structured scenes, we show the probability distribution maps and the predicted trajectories in the first and second sub-figures, respectively. The results demonstrate that the crowd motion patterns are well explored via our proposed algorithm. Crowd motions are of more significant uncertainties compared with the structured scenes. Even so, we can still discover scene semantic or social rules, e.g., pedestrians in crowd engage with coherent group while keep a certain distance from others. In this case, the prior knowledge leant in the training procedure still facilitates to explore the systematic or deterministic component in crowd motions. Moreover, the performance is still benefit from sharing information across neighboring pedestrians via the social-aware gate.

We report the quantitative evaluation in Table. 2 for the unstructured scenes. Similar as the results in the structured scenes, our proposed algorithm outperforms alternative methods. The main reason is because the prior knowledge in the massive data is well explored via the social-aware LSTM, and the motion uncertainty is modeled via the Gaussian processes. In general, the LSTM-based algorithms degenerates significantly compared with the structured scenes, since the motion uncertainty in crowd motions is more significant and there also exist missing data points in the observed trajectories due to the occlusions. The neighboring information is helpful, therefore, the performance of *coherent LSMT (cLSTM)* [Su *et al.*, 2016] and *SRGP without GP* outperform that based on *LSTM* [Srivastava *et al.*, 2015].

Moreover, the results demonstrate that the performance based on *Gaussian processes (GP)* [Ellis *et al.*, 2009b] is also inferior to our proposed algorithm, since the information embedded in massive data is not well exploited. Note that the *deep Gaussian processes (DGP)* [Damianou and Lawrence, 2013] is superior to the LSTM-based algorithms, since uncertainty in the unstructured scenes is much more significant, e.g., the occlusions or missing points of trajectories.

## 4 Conclusions

In this paper, we propose an effective framework for forecasting the invisible paths in crowd scenes via a social-aware recurrent Gaussian process. Crowd motion prediction is implemented by taking advantages of the interplay between the deterministic component embedded in the rich prior and uncertainties due to occlusions or neighboring interactions. Our approach is able to model the rich navigation patterns by encoding the prior knowledge embedded in massive data via a social-aware LSTM. The algorithm is also benefit from the use of Gaussian processes to explicitly model uncertainties in predictions such that the characteristic unpredictability can be accurately represented. By anticipating the near future paths, it provides a good probability to prevent the potential risks in video surveillance and autonomous vehicles before they would emerge.

## References


[Alahi *et al.*, 2016] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–7, 2016.

[Dai *et al.*, 2016] Zhenwen Dai, Andreas Damianou, Javier González, and Neil Lawrence. Variational auto-encoded deep Gaussian processes. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2016.

[Damianou and Lawrence, 2013] Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 207–215, 2013.

[Ellis *et al.*, 2009a] David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *Computer Vision Workshops (ICCVW), IEEE 12th International Conference on*, pages 1229–1234. IEEE, 2009.

[Ellis *et al.*, 2009b] David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *Computer Vision Workshops (ICCVW), 2009 IEEE 12th International Conference on*, pages 1229–1234. IEEE, 2009.

[Godoy *et al.*, 2016] Julio Erasmo Godoy, Ioannis Karamouzas, Stephen J Guy, and Maria L Gini. Implicit coordination in crowded multi-agent navigation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*, pages 2487–2493, 2016.

[Helbing and Johansson, 2011] Dirk Helbing and Anders Johansson. *Pedestrian, crowd and evacuation dynamics*. Springer, 2011.

[Kitani *et al.*, 2012] Kris Kitani, Brian Ziebart, James Bagnell, and Martial Hebert. Activity forecasting. *European Conference on Computer Vision (ECCV)*, pages 201–214, 2012.

[Li and Fu, 2014] Kang Li and Yun Fu. Prediction of human activity by discovering temporal sequence patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (TPAMI)*, 36(8):1644–1657, 2014.

[Li *et al.*, 2014] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (TPAMI)*, 36(1):18–32, 2014.

[Lipton, 2015] Zachary C Lipton. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[Long *et al.*, 2016] Pinxin Long, Wenxi Liu, and Jia Pan. Deep-learned collision avoidance policy for distributed multi-agent navigation. *arXiv preprint arXiv:1609.06838*, 2016.

[Robicquet *et al.*, 2016] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision (ECCV)*, pages 549–565, 2016.

[Shao *et al.*, 2014] Jing Shao, Chen Change Loy, and Xiaogang Wang. Scene-independent group profiling in crowd. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2227–2234, 2014.

[Srivastava *et al.*, 2015] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 843–852, 2015.

[Su *et al.*, 2016] Hang Su, Yinpeng Dong, Jun Zhu, Haibin Ling, and Bo Zhang. Crowd scene understanding with coherent recurrent neural networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3469–3476, 2016.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pages 3104–3112, 2014.

[Titsias and Lawrence, 2010] Michalis K Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *The 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 844–851, 2010.

[Titsias, 2009] Michalis K Titsias. Variational learning of inducing variables in sparse gaussian processes. In *The 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5, pages 567–574, 2009.

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

[Zhou *et al.*, 2011] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3441–3448, 2011.

[Zhou *et al.*, 2012] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang. Coherent filtering: detecting coherent motions from crowd clutters. In *European Conference on Computer Vision (ECCV)*, pages 857–871, 2012.

[Zhou *et al.*, 2015] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang. Learning collective crowd behaviors with dynamic pedestrian-agents. *International Journal of Computer Vision (IJCV)*, 111(1):50–68, 2015.

[Ziebart *et al.*, 2008] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *The Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 1433–1438, 2008.