Understanding Collective Crowd Behaviors: Learning a Mixture Model of Dynamic Pedestrian-Agents

Bolei Zhou[1], Xiaogang Wang[2,3], and Xiaoou Tang[1,3]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
zhoubolei@gmail.com, xgwang@ee.cuhk.edu.hk, xtang@ie.cuhk.edu.hk

Abstract

In this paper, a new Mixture model of Dynamic pedestrian-Agents (MDA) is proposed to learn the collective behavior patterns of pedestrians in crowded scenes. Collective behaviors characterize the intrinsic dynamics of the crowd. From the agent-based modeling, each pedestrian in the crowd is driven by a dynamic pedestrian-agent, which is a linear dynamic system with its initial and termination states reflecting a pedestrian's belief of the starting point and the destination. Then the whole crowd is modeled as a mixture of dynamic pedestrian-agents. Once the model is unsupervisedly learned from real data, MDA can simulate the crowd behaviors. Furthermore, MDA can well infer the past behaviors and predict the future behaviors of pedestrians given their trajectories only partially observed, and classify different pedestrian behaviors in the scene. The effectiveness of MDA and its applications are demonstrated by qualitative and quantitative experiments on the video surveillance dataset collected from the New York Grand Central Station.

1. Introduction

Automatically understanding the behaviors of pedestrians in crowd is of great interest to video surveillance, and has drawn more and more attentions in recent years [26]. It has important applications, such as event recognition [12], traffic flow estimation [23], behavior prediction [2], and crowd simulation [20]. One of the underlying challenges of these problems is to model and learn the collective dynamics of pedestrian behaviors in crowded scenes.

Crowd behavior analysis has been studied in social science with a long history. French sociologist Le Bon (1841∼1931) described collective crowd behaviors in his book The Crowd: A Study of the Popular Mind as, "*the crowd, an agglomeration of people, presents new characteristics very different from those of the individuals composing it, the sentiments and ideas of all the persons in the gathering take one and the same direction, and their conscious personality vanishes.* " It leads to the motivation of this work: the crowd has its intrinsic collective dynamics. Although individuals in crowd might not acquaint with each other, their shared movements and destinations make them coordinate collectively and follow the paths commonly taken by others [13]. An illustrative example is shown in Figure 1A.

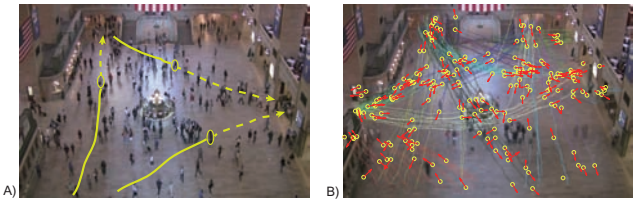

Figure 1. A) The crowd of pedestrians walking in a train station. Pedestrians have clear beliefs of the starting points and the destinations in mind. These beliefs and scene structures (e.g. the border of walls) influence their past behaviors (indicated as solid green lines) as well as the future behaviors (indicated as dashed green lines). The shared beliefs and dynamics of movements generate several dominant collective dynamic patterns in the scene. B) MDA learns the collective dynamic patterns of the crowd from fragmented trajectories and simulates the collective behaviors of the crowd. Yellow circles and red arrows represent the current positions of the simulated pedestrians and their velocities, along with their past trajectories in different colors.

In this paper, a new **M**ixture model of **D**ynamic pedestrian-**A**gents **(MDA)** is proposed to learn the collective dynamics of pedestrians from a large amount of observations without supervision. Observations are trajectories of feature points on pedestrians obtained by a KLT tracker [19]. Because of frequent occlusions in crowded scenes,

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

there are many tracking failures, and most trajectories are highly fragmented with large portions of missing observations. The movement of a pedestrian is driven by one of the pedestrian-agents, which are modeled as linear dynamic systems with initial and termination states (reflecting pedestrians' beliefs of the starting points and the destinations). Furthermore the timings of pedestrians entering the scene with different dynamic patterns are modeled as Poisson processes. Then, the collective dynamics of the whole crowd are modeled as a mixture dynamic system. The effectiveness of MDA is demonstrated by three applications: simulating collective crowd behaviors, clustering trajectories into different collective behaviors, and predicting the behaviors of pedestrians. Both qualitative and quantitative experimental evaluations are conducted on data collected from the New York Grand Central Station.

The novelty and contributions of this work are summarized as follows. 1) Although there exist some approaches [6, 23, 10, 25] to learn motion patterns in crowded scenes, they do not explicitly model the dynamics of pedestrians. Many of them only took local location-velocity pairs as input, while discarding the temporal order of trajectories, which is important for both classification and simulation. Instead, MDA takes trajectories as input, and models the temporal generative process of trajectories. Compared with those approaches, it is much more natural for MDA to simulate collective crowd behaviors and predict pedestrians' future behaviors, once its parameters are learned from real data. 2) Under MDA, pedestrians' beliefs, which strongly regularize their behaviors, are explicitly modeled and inferred from observations. In order to be robust to tracking failures, the states of missing observations on trajectories are modeled and inferred. Because of these two facts, MDA can well infer the past behaviors and predict the future behaviors of pedestrians given their trajectories only partially observed. They also lead to better accuracy of recognizing the behaviors of pedestrians. 3) To the best of our knowledge, MDA is the first agent-based model to learn collective dynamics from the crowd videos. Besides the collective dynamics, the behavior of a pedestrian is also driven by the interactions with his/her neighbors. In the future work, it would be much easier for MDA to integrate with the module of interactive dynamics such as the social force model [5, 15], which is also an agent-based model.

1.1. Related Works

In recent years, there has been significant amount of work on learning the motion patterns in crowded scenes due to growing interest in crowd behavior analysis and crowd management. For example, Ali *et al.* [1] and Lin *et al.* [10] computed the flow fields and segmented the patterns of crowd flows using Lagrangian coherent structures or Lie algebra. Wang *et al.* [23] explored the co-occurrence of moving pixels without tracking objects to learn the motion patterns in crowded scenes. These approaches took the local location-velocity pairs as input while ignoring the temporal order of observations in order to be robust to tracking failures. The beliefs of pedestrians were not considered either. Some approaches learned the motion patterns through clustering trajectories [11, 21, 22], and faced the challenge of fragmentation of trajectories in crowded scenes. None of the above methods used agent-based models, which could model the process of a pedestrian making decisions based on the current states. It is difficult for them to simulate or predict collective crowd behaviors.

To analyze the interaction between pedestrians, the social force model, first proposed by Helbing *et al.* [5, 4] for crowd simulation, was introduced to the computer vision community recently and was applied to multi-target tracking [15], abnormality detection [12], and interaction analysis [16]. The social force model is also an agent-based model and assumes that pedestrians' movements for the next step are affected by their destinations, the states of their neighbors, and the borders of buildings, walls, streets, and obstacles. It is complementary to MDA, since it models the interactive dynamics among pedestrians but requires the scene structures and the beliefs of pedestrians to be known in advance. MDA better models the collective dynamics, automatically learns the regularization added by scene structures, and infers the beliefs of pedestrians. Both MDA and the social force model are agent-based models and have the potential to be well combined. Therefore it would be very interesting to integrate both collective dynamics and interactive dynamics which characterize the crowd behaviors from different perspectives into a single model in the future work.

A number of pedestrian models for crowd simulation were proposed in computer graphics. Continuum-based pedestrian models [8, 20] treated the crowd motion as fluid with manually assigned parameters. Agent-based pedestrian models [3] treated pedestrians as autonomous agents based on a set of defined rules and known scene structures. Differently under MDA the collective dynamics for crowd behavior simulation are automatically learned from real videos without any prior knowledge about scene structures.

2. MDA Model

The crowd is an agglomeration of pedestrians. Although every pedestrian has his own movement dynamics and belief of the starting point and the destination, some statistical dynamic patterns would appear when enough pedestrians' behaviors are observed over time, because pedestrians in a specific scene share common movement dynamics and beliefs. These shared dynamic patterns could be abstracted as different pedestrian-agents with various *dynamics* and *be-*

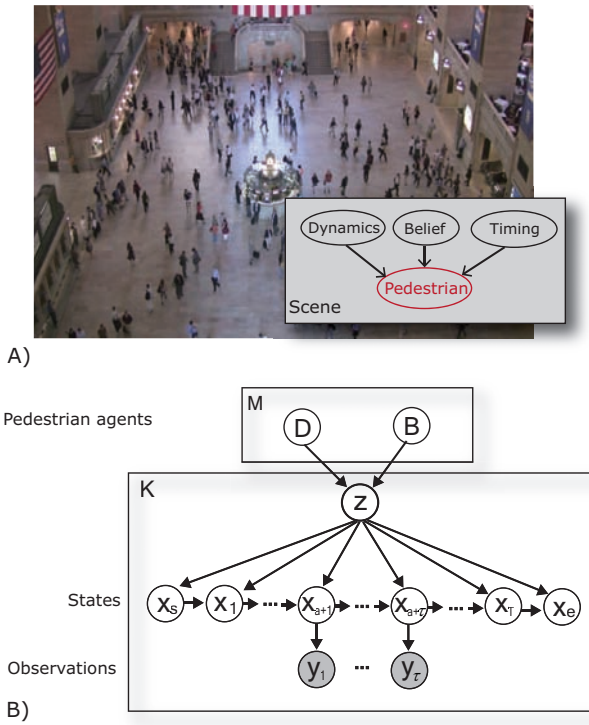

Figure 2. A) The behavior of a pedestrian in the crowd is influenced by three key factors, the dynamics of movements, the belief of starting point and destination, and the timing of entering in the scene. B) Graphical representation of the Mixture model of Dynamic pedestrian-Agents. The shadowed variables are partial observations of the hidden states due to frequent tracking failures in crowded environment.

liefs. In our model, *dynamics* and *beliefs* of pedestrians are modeled as two key modules D and B in the agent system. Meanwhile, the timings of the event that a pedestrian enters in the scene vary, because each pedestrian-agent emerges at different frequency from the entry in the scene. We augment MDA with another module, *timing* of emerging, for the dynamic pedestrian-agent. Thus, the crowd in the scene is formulated as a mixture model of dynamic pedestrian-agents as shown in Figure 2. In the following sections, each module will be explained in details.

2.1. Modeling Pedestrian Dynamics

Trajectories extracted in the scene are time-series observations of pedestrian dynamics. If we treat a pedestrian as a dynamic agent system which actively senses the environment and makes decisions, the trajectory of the pedestrian is a set of observations of the hidden dynamic states of this system. We model the dynamics of a pedestrian-agent as a linear dynamic system defined by

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \omega_t, \tag{1}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \varepsilon_t. \tag{2}$$

$\mathbf{x}_t = [x_t^1, x_t^2, 1]^\top$ is the current state of the agent system and represents the position of the agent in homogeneous coordinates. $\mathbf{y}_t \in \mathcal{R}^m$ is the observation of $\mathbf{x}_t$. $\mathbf{A} \in \mathcal{R}^{3\times3}$ is the state transition matrix and $\mathbf{C} \in \mathcal{R}^{m\times3}$ is the observation matrix. ω is the system noise, and ε is the observation noise. Since the observations of the agent system are its position, m is 3 and $\mathbf{C}$ is simplified as a 3×3 identity matrix. The conditional distributions of the state and the observation are

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{A}\mathbf{x}_{t-1}, \Gamma), \tag{3}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{y}_t|\mathbf{x}_t, \Sigma), \tag{4}$$

where $\mathcal{N}$ is the 3-dimensional multivariate Gaussian distribution, Γ and Σ are covariance matrices. Σ is assumed to be a known diagonal matrix. We denote $D = (\mathbf{A}, \Gamma)$ as the *dynamics* parameters to be learned for the agent system.

2.2. Modeling Pedestrian Beliefs

A pedestrian normally has a clear belief of the starting point and the destination when walking in a scene. This *belief* is a key factor driving the overall behavior of the pedestrian, and it is also considered as the source and sink of the scene [18, 25]. We model it as the initial state $\mathbf{x}_s$ and the termination state $\mathbf{x}_e$ of the agent system. $\mathbf{x}_s$ and $\mathbf{x}_e$ are sampled from Gaussian distributions,

$$\begin{aligned} p(\mathbf{x}_s) &= \mathcal{N}(\mathbf{x}_s|\mu^s, \Phi^s), \\ p(\mathbf{x}_e) &= \mathcal{N}(\mathbf{x}_e|\mu^e, \Phi^e). \end{aligned} \tag{5}$$

μ^s and μ^e are the means of the initial states and termination states. Φ^s and Φ^e are the corresponding covariance matrices. We denote $B = (\mu^s, \Phi^s, \mu^e, \Phi^e)$ as the *belief* parameters for the agent system.

For a trajectory k, the joint distribution of the system states and observations is

$$\begin{aligned} p(\mathbf{x}^k, \mathbf{y}^k, \mathbf{x}_s^k, \mathbf{x}_e^k) = p(\mathbf{x}_s^k)p(\mathbf{x}_e^k)p(\mathbf{x}_1^k|\mathbf{x}_s^k)p(\mathbf{x}_e^k|\mathbf{x}_{T_k}^k) \\ \prod_{t=2}^{T_k} p(\mathbf{x}_t^k|\mathbf{x}_{t-1}^k) \prod_{t=1}^{\tau_k} p(\mathbf{y}_t^k|\mathbf{x}_{a_k+t}^k), \end{aligned} \tag{6}$$

where $\mathbf{x}^k = \{\mathbf{x}_t^k\}_{t=1}^{T_k}$ and $\mathbf{y}^k = \{\mathbf{y}_t^k\}_{t=1}^{\tau_k}$. $\mathbf{y}^k$ is the partial observation of the whole state $\mathbf{x}^k$. In crowded environments, the trajectories of objects are highly fragmented due to the frequent occlusions among objects. Therefore, most trajectories are only partially observed. We assume that trajectory k is only observed from step $a_k + 1$ to $a_k + \tau_k$. If $a_k = 0$ and $\tau_k = T_k$, the complete trajectory is observed. The initial/termination states as well as the states of missing observations have to be estimated from the model.

2.3. Mixture of Dynamic Pedestrian-Agents

There are numerous pedestrians with various *dynamics* and *beliefs* in a scene. To model the diversity of pedestrian patterns, we extend the single agent system described above to a mixture system of agents, with M possible dynamics and beliefs $(D_1, B_1), ..., (D_M, B_M)$. A hidden variable $z^k = 1, \ldots, M$ indicates the mixture component, *i.e.* one pedestrian-agent system from which a trajectory k is sampled. z^k is sampled from a discrete prior distribution parameterized by $(\pi_1, \ldots, \pi_M)$. The joint distribution is

$$\begin{aligned} &p(\mathbf{x}^k, \mathbf{y}^k, \mathbf{x}_s^k, \mathbf{x}_e^k, z^k) \\ =&p(z^k)p(\mathbf{x}_s^k|z^k)p(\mathbf{x}_e^k|z^k)p(\mathbf{x}_1^k|\mathbf{x}_s^k, z^k)p(\mathbf{x}_e^k|\mathbf{x}_{T_k}^k, z^k) \\ &\prod_{t=2}^{T_k} p(\mathbf{x}_t^k|\mathbf{x}_{t-1}^k, z^k) \prod_{t=1}^{\tau_k} p(\mathbf{y}_t^k|\mathbf{x}_{a+t}^k, z^k). \end{aligned} \tag{7}$$

2.4. Model Learning and Inference

Given the trajectories $\{\mathbf{y}^k\}_{k=1}^K$, we would like to learn the model parameters $\Theta = \{(D_1, B_1), ..., (D_M, B_M)\}$ by maximizing the likelihood of observations,

$$\Theta^* = \arg\max_{\Theta} \sum_{k=1}^{K} \log p(\mathbf{y}^k; \Theta). \tag{8}$$

Since there are three kinds of hidden variables in the graphical model, 1) the index z^k of assigning a trajectory k to a mixture component, 2) the complete sequence of states $\mathbf{x}^k$ that produce the partial observation $\mathbf{y}^k$, and 3) the number t_k^e of steps with missing observations between $\mathbf{x}_{a+\tau}^k$ and the termination state $\mathbf{x}_e^k$, and the number t_k^s of steps with missing observations between the initial state $\mathbf{x}_s^k$ and $\mathbf{x}_{a+1}$ ($T_k = t_k^e + t_k^s + \tau_k$, τ_k is the length of the fragmented trajectory k). We apply the EM algorithm to estimate parameters. Each iteration of EM consists of

$$\begin{aligned} &\textbf{E-step:} \mathcal{Q} = E_{\mathbf{X},\mathbf{T},\mathbf{Z}|\mathbf{Y};\hat{\Theta}}(\log p(\mathbf{X}, \mathbf{Y}, \mathbf{T}, \mathbf{Z}; \Theta)), \\ &\textbf{M-step:} \hat{\Theta}^* = \arg\max_{\Theta} \mathcal{Q}(\Theta; \hat{\Theta}). \end{aligned}$$

where $p(\mathbf{X}, \mathbf{Y}, \mathbf{T}, \mathbf{Z}; \Theta)$ is the complete-data likelihood of the partial observations $\mathbf{Y}$, complete hidden states $\mathbf{X}$(including the initial states and termination states), the numbers of steps with missing observations $\mathbf{T}$, and hidden assignment variables $\mathbf{Z}$.

To initialize the estimation of the belief parameters, we first roughly draw the boundaries of entry/exit regions in the scene as shown in Figure 3A. For trajectories which start or end within these boundaries, their starting points or ending points are used to estimate the belief parameters.

We summarize the derived EM algorithm on MDA as follows. In the E-step, the posterior probabilities and the expectation of complete-data likelihood are,

$$\begin{aligned} \mathcal{Q} =&E_{\mathbf{X},\mathbf{T},\mathbf{Z}|\mathbf{Y};\hat{\Theta}}(\log p(\mathbf{X}, \mathbf{Y}, \mathbf{T}, \mathbf{Z}; \Theta)) \\ =&E_{\mathbf{Z},\mathbf{T}|\mathbf{Y}}(E_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}(\log p(\mathbf{X}, \mathbf{Y}, \mathbf{T}, \mathbf{Z}; \Theta))) \\ =&\sum_{k,m,g,h} \gamma_k(m, g, h) E_{\mathbf{x}^k|\mathbf{y}^k, z^k=m, t_k^s=g, t_k^e=h}(p(\mathbf{x}^k, \mathbf{y}^k, \mathbf{x}_s^k, \mathbf{x}_e^k, z^k)) \end{aligned}$$

where $\gamma_k(m, g, h)$ is defined as

$$\begin{aligned} \gamma_k(m, g, h) =&p(z^k = m, t_k^s = g, t_k^e = h \mid \mathbf{y}^k) \\ =&\frac{\pi_m p(\mathbf{y}^k|z^k = m, t_k^s = g, t_k^e = h)}{\sum_{m'=1}^{M} \sum_{g',h'} \pi_{m'} p(\mathbf{y}^k|z^k = m', t_k^s = g', t_k^e = h')}. \end{aligned}$$

Here we assume the priors for $p(t^s)$ and $p(t^e)$ are uniform distributions, and they are independent with label z^k.

In the M-step, the model parameters are updated as

$$\mathbf{A}_m^{new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h) \sum_{t=2}^{T_k} P_{t,t-1}^k}{\sum_{k,g,h} \gamma_k(m, g, h) \sum_{t=2}^{T_k} P_{t-1,t-1}^k}, \tag{9}$$

$$\Gamma_m^{new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h)(\sum_{t=2}^{T_k} P_{t,t}^k - \mathbf{A}_m^{new} \sum_{t=2}^{T_k} P_{t,t-1}^k)}{\sum_{k,g,h} \gamma_k(m, g, h)(T_k + 1)}, \tag{10}$$

$$\mu_m^{s,new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h) \hat{\mathbf{x}}_s^k}{\sum_{k,g,h} \gamma_k(m, g, h)}, \tag{11}$$

$$\Phi_m^{s,new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h)(\hat{\mathbf{x}}_s^k - \mu_m^s)(\hat{\mathbf{x}}_s^k - \mu_m^s)^\top}{\sum_{k,g,h} \gamma_k(m, g, h)}, \tag{12}$$

$$\mu_m^{e,new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h) \hat{\mathbf{x}}_e^k}{\sum_{k,g,h} \gamma_k(m, g, h)}, \tag{13}$$

$$\Phi_m^{e,new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h)(\hat{\mathbf{x}}_e^k - \mu_m^e)(\hat{\mathbf{x}}_e^k - \mu_m^e)^\top}{\sum_{k,g,h} \gamma_k(m, g, h)}, \tag{14}$$

$$\pi_m^{new} = \frac{\sum_{k,g,h} \gamma_k(m, g, h)}{\sum_{m'=1}^{M} \sum_{k,g,h} \gamma_k(m', g, h)}. \tag{15}$$

τ_k is the length of the trajectory k.

$$\begin{aligned} \hat{\mathbf{x}}^k =&E_{\mathbf{x}^k|\mathbf{y}^k, z^k=m, t_k^s=g, t_k^e=h}(\mathbf{x}^k), \\ P_{t,t}^k =&E_{\mathbf{x}^k|\mathbf{y}^k, z^k=m, t_k^s=g, t_k^e=h}(\mathbf{x}_t \mathbf{x}_t^\top), \\ P_{t,t-1}^k =&E_{\mathbf{x}^k|\mathbf{y}^k, z^k=m, t_k^s=g, t_k^e=h}(\mathbf{x}_t \mathbf{x}_{t-1}^\top), \end{aligned}$$

and $\gamma_k(m, g, h)$ are all computed efficiently by modified Kalman smoothing filter [14, 17], which can recursively estimate the hidden states given the partial observations. Note that $\gamma_k(m, g, h)$ has three discrete variables, it is time consuming to enumerate and compute all their possible combinations. However, for most (g, h), $\gamma_k(m, g, h)$ are approximately to 0. We first get the most plausible $\hat{h} = \arg\min_t \parallel \mu_m^e - \mathbf{A}_m^t \mathbf{y}_\tau^k \parallel, \hat{g} = \arg\min_t \parallel \mu_m^s - \mathbf{A}_m^{-t} \mathbf{y}_1^k \parallel$ by gradient descent. Then we limit the plausible range of t_k^s as $[\hat{g}-\Delta, \hat{g}-\Delta+1, ..., \hat{g}, ..., \hat{g}+\Delta-1, \hat{g}+\Delta]$, and the plausible range of t_k^e as $[\hat{h}-\Delta, \hat{h}-\Delta+1, ..., \hat{h}, ..., \hat{h}+\Delta-1, \hat{h}+\Delta]$, where Δ is an integer and empirically determined. When it is out of the plausible range, $\gamma_k(m, g, h)$ is approximated as 0. For each combination, the total step of all states $\hat{T}_k = \tau_k + t_k^e + t_k^s$.

Table 1. Algorithm for fitting a dynamic pedestrian-agent.

INPUT: trajectory k from any tracker.
OUTPUT: the optimal fitted z^*.
01: **for** $m = 1 : M$ **do**
02: compute $\gamma(z^k = m) = \sum_{g,h} \gamma_k(m, g, h)$
03: **end for**
04:$z^* = \arg\max_m \gamma(z^k = m)$
05:compute the future state or past state with $\mathbf{A}_{z^*}$.
predict its belief with B_{z^*}.

Table 2. Algorithm for sampling a dynamic pedestrian-agent.

INPUT: time length T, pedestrian-agent m
OUTPUT: simulated trajectories.
01:sample temporal order $\delta_{1\sim T}$ from $PoissonP(\lambda_m)$
02:**for** ω = 1 : T
03: **if** δ_ω == 1
04: sample $\mathbf{x}_s$ from $p_m(\mathbf{x}_s)$
05: $\tau = \arg\min_t \parallel \mu_m^e - \mathbf{A}_m^t \mathbf{x}_s \parallel$.
06: generate trajectory $\{\mathbf{y}_t\}_{t=1}^{\tau}$ by sequentially sampling $p_m(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p_m(\mathbf{y}_t|\mathbf{x}_t)$.
07: **end if**
08:**end for**

2.5. Algorithms for Model Fitting and Sampling

After the parameters of MDA are learned, given the fragmented trajectory of a pedestrian in the scene, our model can fit it to the optimal pedestrian-agent and predict the pedestrian's past and future paths, as well as the belief of the starting point and the destination. Meanwhile, by sampling from the pedestrian-agent model we can generate the trajectories characterized by this pedestrian-agent. These two important properties of MDA model will be used in the following experiments. The algorithms of fitting a dynamic pedestrian-agent and sampling trajectories from it are listed in Table 1 and 2.

3. Modeling Pedestrian Timing of Emerging

To fully capture the dynamics of pedestrians in the scene, we model pedestrian timings of emerging, $i.e.$ the frequency of new pedestrians entering in the scene over time, and integrate this module into MDA.

Considering the event that a pedestrian emerges in an entry region, we assume the timing of that event follows a homogeneous Poisson process $PoissonP(\lambda)$, whose underlying distribution is a Poisson distribution

$$p(n; \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \tag{16}$$

where n is the number of events that occur during an unit time interval. λ is the rate parameter of the Poisson process, and indicates the expected number of events that occur per unit time interval.

After $\{(D_1, B_1), ..., (D_M, B_M)\}$ being learned by the EM algorithm, every trajectory k has the most likely z^k, and its emerging time can also be estimated. Thus we can count the number of emerging pedestrians in each time interval (here we use 5 seconds), and estimate the rate parameter λ_m for each pedestrian-agent m by maximum likelihood estimation,

$$\hat{\lambda}_m = \frac{1}{L} \sum_{i=1}^{L} n_i^m, \tag{17}$$

where L is the number of time intervals over the whole video sequence, and n_i^m is the number of emerging pedestrians generated from the dynamic pedestrian-agent m in time interval i.

4. Experiments and Applications

Experiments are conducted on a 15 minute long video sequence collected from the New York Grand Central Station. The video is 24fps with a resolution of 480×720[1]. A KLT keypoint tracker [19] is used to extract trajectories. Tracking terminates when ambiguities caused by occlusions and scene clutters arise, and new tracks will be initialized later. After filtering some short or stationary trajectories, around 20,000 trajectories are extracted and shown in Figure 3A. Figure 3B plots the histogram of the lengths of trajectories. It shows that most trajectories are highly fragmented, and exist only for short periods.

4.1. Model Learning

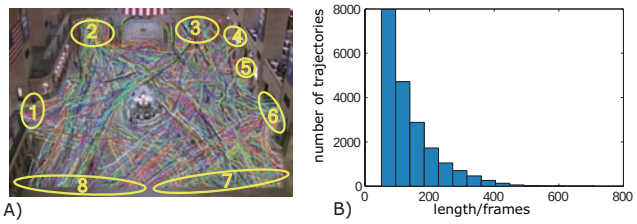

Figure 3. A) Extracted trajectories and entry/exit regions indicated by yellow ellipses. The colors of trajectories are randomly assigned. B) Histogram of the lengths of trajectories. Most of them are short and fragmented.

To initialize the belief parameters of MDA, we first roughly label 8 entry/exit regions with ellipses indexed by 1∼8 in Figure 3A. The parameters will be updated at the learning stage. Trajectories which start/end within these regions have observed initial/termination states. Their starting/ending points are used to initialize the estimation of parameters $(\mu_m^s, \Phi_m^s, \mu_m^e, \Phi_m^e)$. After initialization, all the

[1]Data is available at http://www.ee.cuhk.edu.hk/∼xgwang/grandcentral.html

Figure 4. A) Illustration of eight representative dynamic pedestrian-agents through sampling pedestrians from them. Green and red circles indicate the distributions of initial/termination states for each pedestrian-agent. Yellow circles indicate the current positions of sampled pedestrians along their trajectories, and red arrows indicate current velocities. The timings of pedestrians entering the scene sampled from the Poisson process are shown below. One impulse indicates a new pedestrian entering the scene driven by the corresponding pedestrian-agent. B) Flow fields generated from dynamic pedestrian-agents. C) Flow fields learned by LAB-FM [10].

parameters of MDA are automatically learned from the observations. It takes around one hour for the EM algorithm to converge, running on a computer with 3GHz Core Quad CPU and 4GB RAM with Matlab implementation. Totally $M = 20$ agent components are learned. In this work, M is chosen empirically, but it also could be estimated with Dirichlet process [23].

Figure 4A illustrates eight representative dynamic pedestrian-agents. Trajectories are sampled from each pedestrian-agent using the algorithm in Table 2. Results show that the learned dynamic pedestrian-agents have different dynamics, beliefs and timings of emerging, and they characterize various collective behaviors. By densely sampling, MDA also can estimate the velocity flow field for each pedestrian agent as shown in Figure 4B. For comparison, the representative flow fields learned by LAB-FM [10], which tried to learn motion patterns using Lie algebra, are shown in Figure 4C. MDA performs better in terms of capturing long-range collective behaviors and separating different collective behaviors. For example, some flow fields learned by LAB-FM are locally distributed, without covering the complete paths. The upper parts of the first two flow fields in Figure 4B, which represent two different collective behaviors, are merged by LAB-FM as shown in the first flow field in Figure 4C. This is due to the facts that 1) MDA better models the shared beliefs of pedestrians and states of missing observations, and takes the whole trajectories instead of local position-velocity pairs as input, and also that 2) LAB-FM assumes that the spatial distributions of the flow fields are Gaussian (indicated by cyan ellipses).

4.2. Collective Crowd Behavior Simulation

Compared with other approaches [6, 23, 25] of modeling global motion patterns in crowded scenes, one of the distinctive features of MDA is to simulate collective crowd behaviors once it is learned from observations. According to the superposition property of Poisson process [9], the timings of overall pedestrians entering the scene also follow a Poisson distribution with $\lambda = \sum_{m=1}^{M} \lambda_m$. To simulate a trajectory, its pedestrian-agent index is first sampled from the discrete distribution $(\pi_1, ..., \pi_M)$ then its trajectory is sampled from the pedestrian-agent using the algorithm in Table 2.

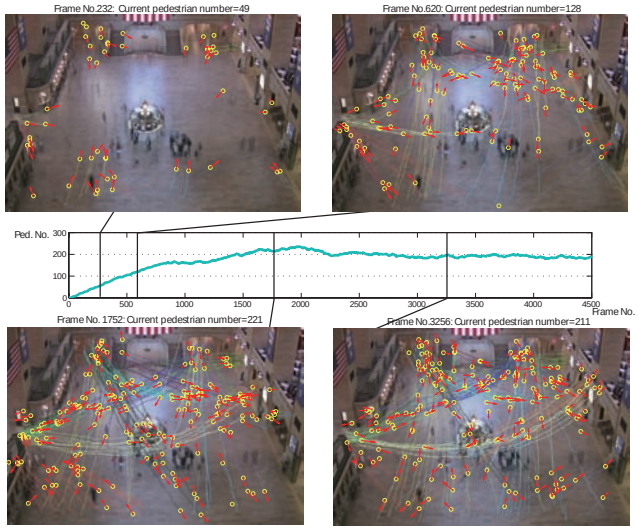

Figure 5. Four exemplar frames from the crowd behavior simulation. Simulated trajectories are colored according to the indices of their dynamic pedestrian-agents. The middle plots the population of pedestrians over time.

Figure 5 shows four exemplar frames of the simulated crowd behaviors. At the first frame pedestrians begin to enter the empty scene. After 1500 frames the crowd reaches the equilibrium population with around 200 pedestrians. Our model well learns the dynamics of the crowd, and the simulated pedestrian behaviors are similar to those observed in the real data.

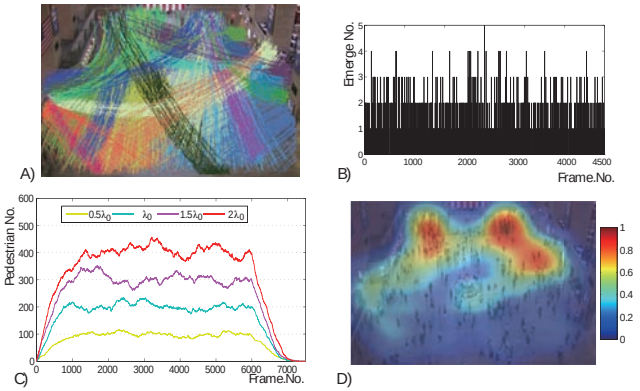

Figure 6. A) The plot of all the simulated trajectories. Colors of trajectories are assigned according to pedestrian-agent indices. B) The number of pedestrians entering the scene at different frames. C) The capacity of the train station with $\lambda = 0.5\lambda_0, \lambda_0, 1.5\lambda_0, 2\lambda_0$ in simulation, where λ_0 is the value learned from data. D) The population density map of the train station computed from the simulation. Color measures the relatively populated area.

Figure 6A plots all the simulated trajectories over 4500 frames. Figure 6B shows the timings of emerging of the crowd, *i.e.* the numbers of new pedestrians entering the scene over time. The crowd simulation with MDA can provide some valuable information about the dynamics of the crowd in the scene. For example, in Figure 6C, we investigate the relationship between the different rate parameter λ and the capacity of the train station, where pedestrians begin and stop to enter the scene at the Frame 1 and 6000 respectively. As pedestrians keep entering the scene with a constant birth rate, the scene will reach its capacity, which is the equilibrium state of the system. When $\lambda = \lambda_0$, which is learned from data, the system reaches its equilibrium state after 1500 frames with around 200 pedestrians in the scene. So the capacity of the scene could be measured as 200. And the equilibrium state will change with different birth rates as shown in Figure 6C. In Figure 6D we compute the averaged population density map when $\lambda = \lambda_0$, the populated areas of the scene are detected. These areas should deserve high attention of security since accidents would most likely happen there when panic or abnormal event strikes. These types of information are very useful for the crowd management and the public facility optimization.

4.3. Collective Behavior Classification

Once MDA is learned from observations without supervision, it can be used to cluster the trajectories of pedestrians into different collective dynamics. We simply take the inferred index z^k of every trajectory as its cluster index. A lot of works have been done on trajectory clustering in video surveillance. This problem is especially challenging in crowded scenes because trajectories are highly fragmented with many missing observations. Generally speaking, existing approaches are in two categories: distance-based [24, 7] and model-based [21]. We choose one representative approach from each category for comparison: Hausdorff distance-based spectral clustering [24] and hierarchical Dirichlet processes (HDP) [21].

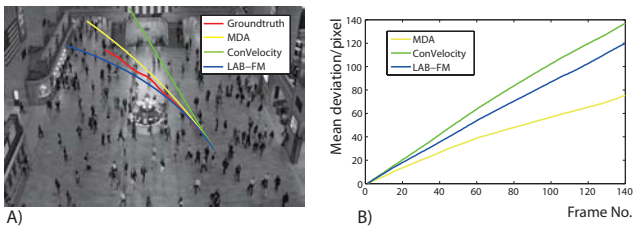

Figure 8. A) An example of predicting behaviors with different methods. B) The averaged prediction errors with different methods tested on 30 trajectories.

Figure 7A shows some representative clusters of trajectories obtained by MDA. Even though most trajectories are fragmented and are far away from each other in space, they are still well grouped into one cluster because they share the same collective dynamics. For example, the first cluster in Figure 7A explains the collective behavior of "pedestrians walking from entry 7 to exit 2". Figure 7B and Figure 7C show the representative clusters obtained by spectral clustering [24] and HDP [21]. They are all in short spatial range and it is hard to interpret their semantic meanings, because they cannot well handle the fragmentation of trajectories.

4.4. Behavior Prediction

MDA can predict pedestrians' behaviors given that their trajectories are only partially observed. We manually label 30 trajectories of pedestrians as ground-truth. For each ground-truth trajectory, we use the observations of the first 20 frames to estimate its pedestrian-agent index z with the algorithm in Table 1. Then, the model of the selected pedestrian-agent is used to recursively generate the following states as the predicted future trajectory. The performance is measured by the averaged prediction error, *i.e.* deviation between the predicted trajectories and the ground-truth trajectories.

Two baseline methods are used for comparison. In the first comparison method(referred as ConVelocity), a constant velocity which is estimated as the averaged velocity of the past observations, is used to predict the future positions. In the second comparison method LAB-FM [10], the learned flow field which best fit the first 20 frame observations, is used to predict future positions. The results in Figure 8 show that MDA has better prediction performance.

5. Concluding Remarks

In this paper, we propose a Mixture model of Dynamic Pedestrian-Agents to learn the collective dynamics from

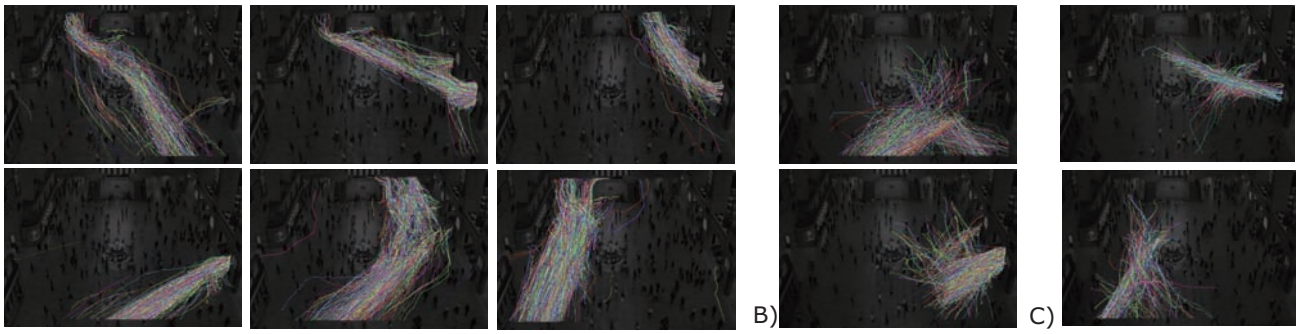

Figure 7. Representative clusters of trajectories by A)MDA model, B)Spectral Clustering [24] and C)HDP [21]. Colors of trajectories are randomly assigned.

video sequences in crowded scenes. Through modeling the beliefs of pedestrians and the missing states of observations, it can be well learned from highly fragmented trajectories caused by frequent tracking failures. It can not only classify collective behaviors, but also simulate and predict collective crowd behaviors.

This model has various potential applications and extensions to be explored in the future work. It can be integrated with the social force model to characterize both the collective dynamics and interactive dynamics of crowd behaviors at both the macroscopic and microscopic levels. It will lead to better accuracies on object tracking, behavior classification, simulation, and prediction. The extended model also has the potential to simulate other interesting crowd behaviors such as panic rising and evacuation.

6. Acknowledgement

This work is partially supported by the Research Grants Council of Hong Kong (RGC project No. CUHK417110 and CUHK417011) and National Natural Science Foundation of China (project no. 61005057), and by Guangdong Province through Introduced Innovative R&D Team of Guangdong Province 201001D0104648280. The first author would like to thank Deli Zhao and Wei Zhang for their insightful discussions.

References

[1] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *Proc. ECCV*, 2008.

[2] G. Antonini, S. Martinez, M. Bierlaire, and J. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *Int'l Journal of Computer Vision*, 2006.

[3] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 2002.

[4] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 2000.

[5] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 1995.

[6] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *Proc. ICCV*, 2009.

[7] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *IEEE Trans. on Image Processing*, 2007.

[8] R. Hughes. The flow of human crowds. *Annual Review of Fluid Mechanics*, 2003.

[9] J. Kingman. Poisson processes. *Oxford University Press*, 1993.

[10] D. Lin, E. Grimson, and J. Fisher. Learning visual flows: A Lie algebraic approach. In *Proc.CVPR*, 2009.

[11] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Trans. on SMC*, 2005.

[12] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Proc.CVPR*, 2009.

[13] M. Moussaid, S. Garnier, G. Theraulaz, and D. Helbing. Collective information processing and pattern formation in swarms, flocks, and crowds. *Topics in Cognitive Science*, 2009.

[14] W. Palma. *Long-memory time series: theory and methods*. Wiley-Blackwell, 2007.

[15] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. ICCV*, 2009.

[16] P. Scovanner and M. Tappen. Learning pedestrian dynamics from the real world. In *Proc. ICCV*, 2009.

[17] R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, 1982.

[18] C. Stauffer. Estimating tracking sources and sinks. In *Proc.CVPR Workshop*, 2003.

[19] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. In *Int'l Journal of Computer Vision*, 1991.

[20] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. In *ACM SIGGRAPH*, 2006.

[21] X. Wang, K. Ma, G. Ng, and W. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *Proc.CVPR*, 2008.

[22] X. Wang, K. Ma, G. Ng, and W. Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *Int'l Journal of Computer Vision*, 2011.

[23] X. Wang, X. Ma, and W. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 2008.

[24] X. Wang, K. Tieu, and W. Grimson. Learning semantic scene models by trajectory analysis. *Proc. ECCV*, 2006.

[25] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Proc.CVPR*, 2011.

[26] S. Zhou, D. Chen, W. Cai, L. Lyo, M. Yoke, L. Hean, F. Tian, D. Wee Sze Ong, V. Su-Han Tay, and B. Hamilton. Crowd modeling and simulation technologies. *ACM Transactions on Modeling and Computer Simulation*, 2009.