A Novel Dynamic Model for Multiple Pedestrians Tracking in Extremely Crowded Scenarios

Xuan Song[†], Xiaowei Shao[†], Quanshi Zhang[†], Ryosuke Shibasaki[†], Huijing Zhao[‡] and Hongbin Zha[‡]

[†]*Center for Spatial Information Science, The University of Tokyo, Japan*
[‡]*Key Laboratory of Machine Perception (MoE), Peking University, China*

Abstract

Tracking hundreds of persons in the large and high density scenarios is a particularly challenging task due to the frequent occlusions and merged measurements. In such circumstances, a stronger dynamic model for prediction usually plays a more important role in the overall tracking process. In this paper, we propose an elaborate dynamic model for multiple pedestrians tracking in the extremely crowded environments. The novelty of this tracking model is that: the global semantic scene structure, local instantaneous crowd flow and the social interactions among persons are taken into account together and combined into an unified approach, which can make the prediction for persons' motion more powerful and accurate. We apply the proposed model by using an online "tracking-learning" framework, which can not only perform the robust tracking in the extremely crowded scenarios, but also ensures that the entire process is fully automatic and online. The testing is conducted on the JR subway station of Tokyo, and the experimental results show that the system with our tracking model can robustly track more than 180 targets at the same time while the occlusions and merge/split frequently occur.

Keywords: Multi-target Tracking, Motion Model, Laser-based Surveillance

1. Introduction

Multiple targets tracking plays a crucial role in various applications, such as surveillance, sports video analysis, human motion analysis and many others. Typically, a multi-target tracking algorithm can be improved with the two following ways: (1) a stronger dynamic model for prediction. (2) a stronger observation

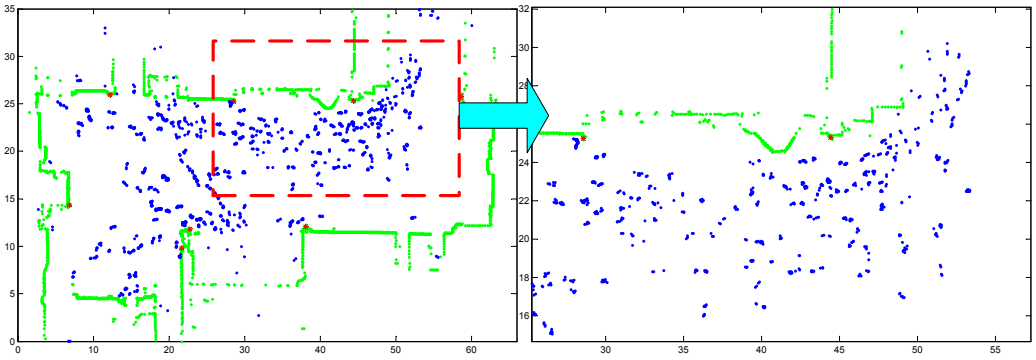
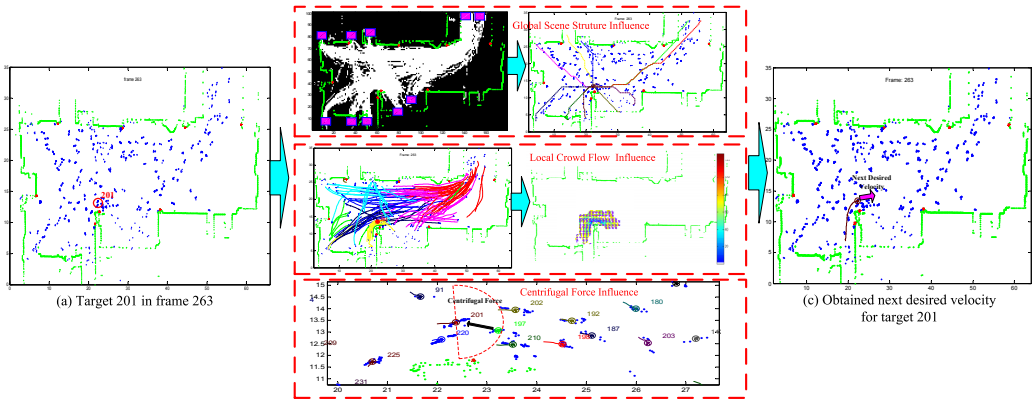


Figure 1: How to maintain the robust tracking in the large and high density scenarios? This is the JR subway station of Tokyo, and the data was obtained by eight single-row laser scanners. The green points are the background, the blue ones are the foreground, and the red ones show the position of single-row laser scanner. In this case, each person is represented by several points. For more details about the experimental site, please refer [1]. We can see that the occlusions and merged measurements frequently take place in such circumstances, and maintaining the robust tracking becomes quite a challenging task.

model (*e.g.* reliable measurements or detections, better data association algorithms) for updating. However, for tracking hundreds of persons in the extremely crowded environments (such as subway station, public square and etc.), it is usually difficult to obtain reliable measurements due to the frequent occlusions and merged measurements, and making the correct data association becomes significantly challenging (as shown in Fig. 1). Hence, a good dynamic model usually plays a more important role in the overall tracking process in such circumstances. *Therefore, the purpose of this paper is to develop a strong dynamic model that can help the tracking algorithm to robustly track hundreds of persons in the large and high density scenarios.*

While the pedestrians are walking in a specific scenarios, their tracking results can be significantly improved with the semantic scene knowledge (e.g., dominant paths, entry or exit, crowd flow and etc.). For instance, "persons usually walk from entrance to exit", "persons have to walk in the dominant paths and avoid static obstacles", "persons who are in a crowd flow can only follow the other people in it." A statistical scene model can provide a priori knowledge on where, when and what types of activities occur. Therefore, in this paper, *we intensively investigate the relationship between pedestrians' social behaviors and their walking scenarios, and propose a novel dynamic model for tracking hundreds of persons in the extremely crowded environments*. Our model considers various factors that



(a) Target 201 in frame 263

(b) Three factors that can influence human motion

(c) Obtained next desired velocity for target 201

Figure 2: Overview of the tracking model. While the pedestrian 201 is walking in the large and high density scene (Fig. a), three factors will influence its short-term path planning (Fig. b): (1) Global scene structure: person should consider the scene structure, move from entrance to exit, walk on the dominant paths, avoid obstacles and find the shortest path. (2) Local crowd flow: persons who are in a specific crowd flow have to follow other persons in it. (3) Centrifugal force: persons usually want to keep a comfortable distance from others. Based on the three factors, our model compute the *next desired velocity* of person 201 in frame 263 (Fig. c).

will influence human motion in short-term (as shown in Fig. 2), such as global semantic scene structure (paths, exit/entrance), local instantaneous crowd flow, centrifugal force among pedestrians. In addition, we apply this model by using an online "tracking-learning" framework, which can not only dynamically reflect the change of scene structure, but also make the overall system fully automatic and online.

The remainder of this paper is structured as follows: In the following section, related work is briefly reviewed. Section 3 and 4 provide the details about the proposed tracking model and the application of this model. Experiments and results are presented in Section 5 and the paper is finally summarized in Section 6.

2. Related Work

Multiple target tracking (MTT) has been studied extensively and an in-depth review of tracking literature can be found in a recent survey by Yilmaz *et al.* [2]. Typically, the tracking algorithm can be improved by a stronger observation model. In this aspect, data association [3] becomes an very important issue. The nearest neighbor standard filter (NNSF) [3] associates each target with the closest measurement in the target state space. However, this simple procedure prunes away

many feasible hypotheses and cannot solve "labeling" problems when the targets are in close proximity. In this respect, a widely used approach to multi-target tracking is achieved by exploiting a joint state space representation which concatenates all of the targets' states together or inferring this joint data association problem by characterization of all possible associations between the targets and observations, such as Joint Probabilistic Data Association Filter (JPDAF) [3, 4], Monte Carlo technique based JPDA algorithms (MC-JPDAF) [5, 6] and Markov chain Monte Carlo data association (MCMC-DA) [7, 8]. Moreover, researchers also propose some global optimization strategies [9, 10, 11] to reduce the complexity and optimize data association algorithms. On the other hand, some researchers also try to explore the stronger appearance model or detection to improve the observation model, and representative publications include [12, 13, 14, 15, 16]. However, most of these methods mentioned above are difficult to be applied to track hundreds of targets in the extremely crowded scene.

Recently, researchers are aware that a stronger dynamic model can significantly improve the tracking results in crowded environments where the measurements are unreliable. Pellegrini *et al.* [17] propose a Linear Trajectory Avoid (LTA) model for human motion prediction, which takes into account the social interactions between persons as well as their orientation towards a destination. Kratz *et al.* [18] predict human movements by capturing the spatial and temporal variations in the crowd. Ziebart *et al.* [19] propose a planning-based motion model, which can model the goal-directed trajectories of pedestrians by using maximum entropy inverse optimal control. Wang *et al.* [20] propose a novel tracking approach which incorporates the scene interaction model and a neighboring object interaction model to respectively perform the long-term and short-term persons' movements prediction. Ali *et al.* [21] propose a floor fields based motion model for tracking persons in crowded scene, and it also utilizes the information of scene structure to assist in tracking. In contrast, Rodriguez *et al.* [22] extend this work, and propose a Correlated Topic Model (CTM) for tracking persons in unstructured crowded scenes. On the other hand, researchers also focus on modeling the social behavior of pedestrians, and representative publications include [23, 24, 25, 26, 27].

In this paper, our model shares some characteristics with the works [21, 17], but differs in two crucial aspects: Firstly, although both the two works both utilize the information of scene structure, the two approaches can be only applied for some simple scene (e.g. single sink/source, single crowd flow). In contrast, our model are suitable to any crowded and complex scene (e.g. structured and unstructured, single and multiple crowd flow or sink/source). Secondly, Ali *et al.*

[21] use the cellular automaton model atop a set of scene-specific "floor fields" to make tracking in extremely crowded situations tractable, which can be seen as "group behavior modeling". On the contrary, Pellegrini *et al.* [17] model single pedestrian in the world coordinates. Compared to the two approaches, our modeling can be seen as "group behavior modeling"+"single pedestrian modeling". *To the best of our knowledge, the proposed model is the first tracking model that intensively explores the relationship between pedestrians' social behaviors and the complex scene in which they are walking.*

3. Tracking Model

3.1. Problem Formulation and Overview

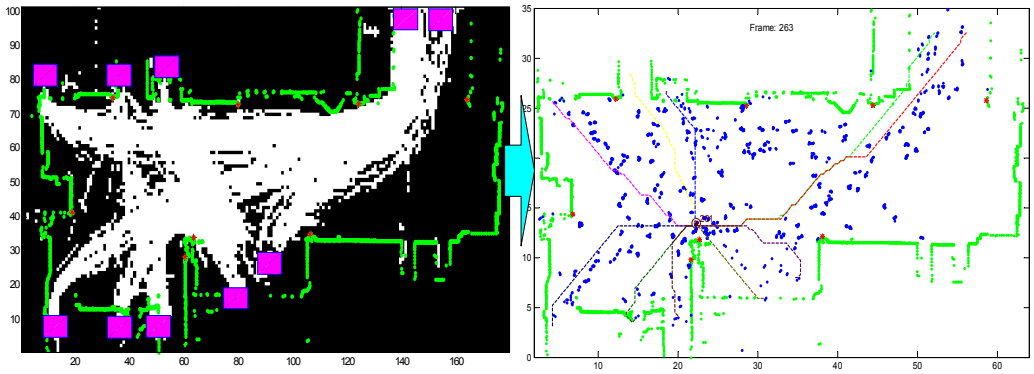
We begin by introducing notations to formulize our problem. At time t, pedestrian i is represented by $\mathbf{x}_{i,t} = (\mathbf{p}_{i,t}, \mathbf{v}_{i,t})$, where $\mathbf{p}_{i,t} = (x, y)$ denotes its 2D position on the ground plane and $\mathbf{v}_{i,t}$ its velocity vector at time t. Based on the current states $\mathcal{X} = \{\mathbf{x}_{i,t}\}$ of n pedestrians $(i = 1...n)$, our model should estimate the state $\mathbf{x}_{i,t+1}$ of pedestrian i in time $t + 1$, especially for its *desired velocity* $\widehat{\mathbf{v}}_{i,t}$ in next step.

While the pedestrians are walking in the large and high density scene, three factors will influence their short-term path planning (as shown in Fig. 2): (1) *Global scene structure*. A person usually plans to go to a specific exit of the scene, walks on the common road, avoids the obstacles and finds the shortest path. (2) *Local instantaneous crowd flow*. At some specific time, some local areas will be quite crowded and become a crowd flow. A person in a particular crowd flow will be greatly influenced by it because he must follow other persons in it. For instance, as shown in Fig. 4-b, at a specific time in a subway station, a large number of persons were just getting off from a train and walking together to catch another train, which were becoming a crowd flow. (3) *Repulsive force of pedestrians*. The motion of a pedestrian is also influenced by the *centrifugal force* [24] from its neighboring persons, as he wants to keep a comfortable distance from others, and he will feel increasing discomfort as he gets closer to a stranger.

Therefore, we propose an energy function, which takes into account these factors:

$$\begin{aligned} &E(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}) = E_{global}(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}) + \\ &\alpha_{i,t} E_{local}(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}) + (1 - \alpha_{i,t}) F_{cent}(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}), \end{aligned} \tag{1}$$

where E_{global} is the global scene structure energy, E_{local} the local crowd flow energy and F_{cent} the norm of repulsive force factor of pedestrians, and $\alpha_{i,t} \in (0, 1)$ is



(a) Online learned scene struture (b) Planned path of target 201

Figure 3: Global scene structure energy. Given the online learned scene structure (Fig. a), we can find possible planned paths of person 201 with A* algorithm (Fig. b), and these paths can be utilized to compute global scene structure energy in Eq.(3).

utilized to control the influence of E_{local} and F_{cent}, which depends on the density of crowd flow $S_p(t)$, where $\mathbf{x}_{i,t} \in S_p(t)$. This could be easily understood: while the density of some local area is quite low, there would be little crowd flow or the persons' number in this crowd flow is limited, and the pedestrians' motion will be greatly influenced by its nearby persons, not by the crowd flow. In contrast, the motion of persons will be more influenced by the crowd flow while the local area density becomes specially high. The details about how to compute this parameter will be discussed in Section 4.

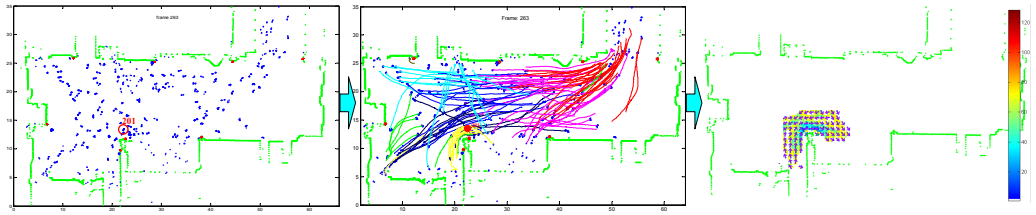
Hence, the *next desired velocity* $\widehat{\mathbf{v}}_{i,t}$ for pedestrian i can be computed by minimization of the energy function $E(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t})$:

$$\widehat{\mathbf{v}}^*_{i,t} = \arg\min_{\widehat{\mathbf{v}}_{i,t}} E(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}). \tag{2}$$

In the next subsections, we will provide the details about how to compute this energy function.

3.2. Global Scene Structure Energy and Local Crowd Flow Energy

Given the current position $\mathbf{p}_{i,t}$ of pedestrian i, online learned scene structure map and N exits/entrances (as shown in Fig. 3-a), it is easy for us to obtain N planned trajectories $\{L^l_{i,t}(x,y)\}_{l=1}^N$ for pedestrian i at time t with *A Star search* algorithm (as shown in Fig. 3-b). Hence, a person would like to make its motion



(a) Target 201 in frame 263 (b) Crowd flow (c) Motion distribution

Figure 4: Crowd flow energy. Given the online learned crowd flow (Fig. b), we can compute its motion distribution (Fig. c). With the help of motion distribution, crowd flow energy can be easily computed by Eq.(6).

be more like to its planned path, and the E_{global} can be computed by:

$$E_{global}(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}) = \sum_{l=1}^{N} w_l \times \exp(-||\frac{\widehat{\mathbf{v}}_{i,t}}{||\widehat{\mathbf{v}}_{i,t}||} - \frac{\partial L^l_{i,t}(\mathbf{p}_{i,t})}{\partial x \partial y}||^2 / 2\sigma_1^2), \tag{3}$$

where $\frac{\partial L^l_{i,t}(\mathbf{p}_{i,t})}{\partial x \partial y}$ is the tangent vector of $L^l_{i,t}(x, y)$, and it denotes the velocity vector of $L^l_{i,t}(x, y)$ at position $\mathbf{p}_{i,t}$. σ_1 is a constant parameter and w_l is the weight of the possible planned trajectories.

We need to compute pedestrians's planned paths in each iteration, and these paths are able to be updated per frame based on pedestrians' movements. Besides, many path planning algorithms are able to be used here, and we select the *A Star search* algorithm due to its fast and effectiveness. On the other hand, we note that not all the planned paths will significantly influence the computation of global scene structure energy because pedestrians usually follow only one path to move. Hence, the computation of planned path weight becomes a very important problem. In this research, the w_l is depend on the similarity between person's current trajectory and the planned ones. Here, we utilize the approach of Wang *et al.* [28] to measure the similarity of two trajectories. In each iteration, the weights of each planned trajectories will be automatically computed; once the weights of planned trajectories are very small, we throw them and stop making new path planning for these exits/entrances.

Meanwhile, the local crowd flow energy should be also computed. A person in a particular crowd flow has to make its motion be close to the crowd flow. Hence, the motion distribution for a specific crowd flow should be extracted firstly. Given

the the crowd flow $S_p(t)$, where $\mathbf{x}_{i,t} \in S_p(t)$ (as shown in Fig. 4-b), its motion distribution can be computed by:

$$\begin{aligned}\mathfrak{V}_{S_p(t)}(\mathbf{p}_{i,t}) = \exp(- < (\widetilde{v_x}(\mathbf{p}_{i,t}), \widetilde{v_y}(\mathbf{p}_{i,t})),\\ (\cos(\alpha^*_{S_p(t)}(\mathbf{p}_{i,t}), \sin(\alpha^*_{S_p(t)}(\mathbf{p}_{i,t})) > /\eta_v),\end{aligned} \tag{4}$$

here $<,>$ stands for dot product, η_v is a constant parameter, $(\widetilde{v_x}(\mathbf{p}_{i,t}), \widetilde{v_y}(\mathbf{p}_{i,t}))$ is the velocity expectation of cluster $S_p(t)$ at a particular position, and $\alpha^*_{S_p(t)}(\mathbf{p}_{i,t})$ is the principal component in the distribution of flow orientation:

$$\mathbb{Q}_{S_p(t)}(\mathbf{p}_{i,t}) = \sum_{m=1}^{M} \pi_m N(\alpha_{S_p(t)}(\mathbf{p}_{i,t}); \mu, \sigma), \tag{5}$$

where Eq.(5) is the GMM model and its parameters π_m can be obtained through EM iteration. An example of Eq.(4) is shown in Fig. 4-c, the color denotes the speed, and the arrows display the principal orientation.

Hence, the local crowd flow energy E_{local} can be computed by:

$$E_{local}(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}) = \exp(-||\widehat{\mathbf{v}}_{i,t} - \mathfrak{V}_{S_p(t)}(\mathbf{p}_{i,t})||^2/2\sigma_2^2), \tag{6}$$

where σ_2 is a constant parameter.

3.3. Repulsive Force Factor

For the pedestrian i, the repulsive effects from pedestrian j depend not only on the relative velocity between them, but also on the distance of them (i.e. the headway), and hence these effects can be expressed by a force term with the following form:

$$\mathbf{F}_{ij} = m_i \mathbf{a}_{ij} = -m_i f(\mathbf{v}_{ij}, ||\mathbf{p}_{ij}||)\mathbf{e}_{ij}, \tag{7}$$

where $\mathbf{a}_{ij}$ is the acceleration of pedestrian i caused by pedestrian j, m_i the mass of pedestrian i; $f(\mathbf{v}_{ij}, ||\mathbf{p}_{ij}||)$ is the function of $\mathbf{v}_{ij}$, and $||\mathbf{p}_{ij}||$ to be determined. $\mathbf{p}_{ij}$ is the distance between pedestrian i and j, and it should be:

$$\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i. \tag{8}$$

$\mathbf{v}_{ij}$ denotes the projection of the relative velocity of of pedestrian i and j in the direction $\mathbf{e}_{ij}$, and can be computed by:

$$\mathbf{v}_{ij} = \frac{1}{2}[(\widehat{\mathbf{v}}_{i,t} - \mathbf{v}_{j,t}) \cdot \mathbf{e}_{ij} + ||(\widehat{\mathbf{v}}_{i,t} - \mathbf{v}_{j,t}) \cdot \mathbf{e}_{ij}||], \tag{9}$$

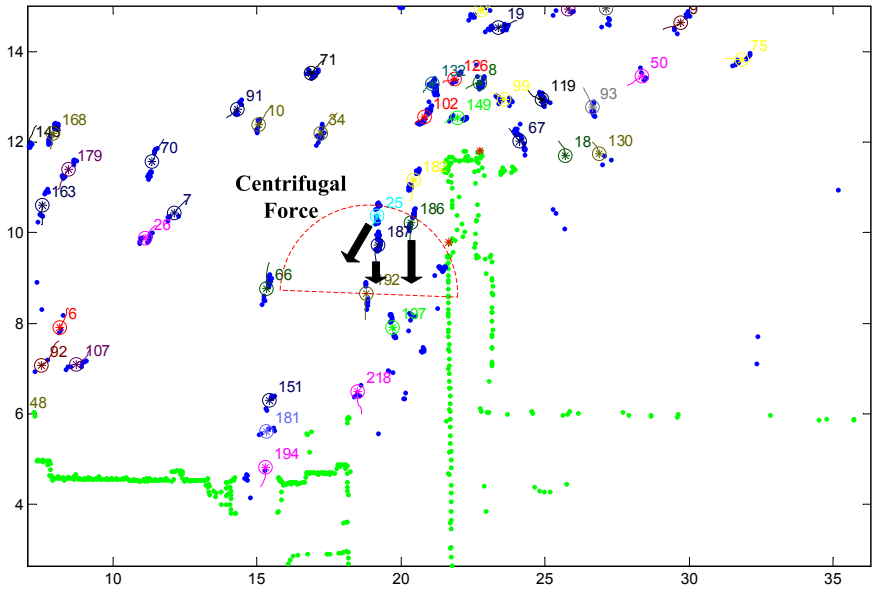


Figure 5: Repulsive force of pedestrians. As shown in this figure, person 92 was subjecting to the repulsive effects from person 25, 187 and 186. The red half-circle show the angle of view for person 92 as discussed in Eq. (14), the black arrows show the repulsive force from person 25, 187, and 186, and the color lines show the pedestrians' trajectories.

$$\mathbf{e}_{ij} = \frac{\mathbf{p}_{ij}}{||\mathbf{p}_{ij}||}. \tag{10}$$

From Eq.(7), we can obtain

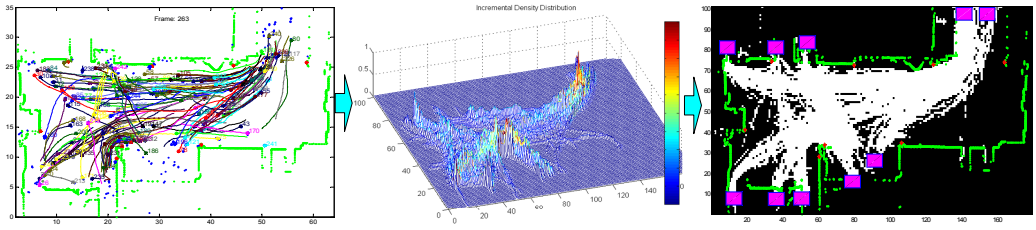
$$||\mathbf{a}_{ij}|| = f(\mathbf{v}_{ij}, ||\mathbf{p}_{ij}||). \tag{11}$$

According to the proof of [24], there should be

$$\frac{|\mathbf{a}_{ij}|||\mathbf{p}_{ij}||}{\mathbf{v}_{ij}^2} = C, \tag{12}$$

where C is a constant depending on the pedestrian's character. For simplicity, we assume $C = 1$, and obtain

$$\mathbf{F}_{ij} = -m_i \frac{\mathbf{v}_{ij}^2}{||\mathbf{p}_{ij}||}\mathbf{e}_{ij}. \tag{13}$$



(a) Tracking results (b) Density distribution (c) Learned scene structure

Figure 6: Global scene structure learning. With the help of tracking results (Fig. a), we can compute incremental density distribution (Fig. b), and the global scene structure can be obtained by thresholding the incremental density distribution and the gradient searching.

If $(\widehat{\mathbf{v}}_{i,t}-\mathbf{v}_{j,t})\cdot\mathbf{e}_{ij} > 0$, i.e., pedestrian i gets close to pedestrian j, the repulsive effects occur. However, if $(\widehat{\mathbf{v}}_{i,t}-\mathbf{v}_{j,t})\cdot\mathbf{e}_{ij} < 0$, i.e., pedestrian j walks faster than pedestrian i, there are no repulsive effects. Larger $\mathbf{v}_{ij}$ creates greater repulsive effects in the former case. We assume that pedestrians react to those who are within their angle of view and the field of vision is $180°$ (as shown in Fig. 5), this situation can be characterized by the following coefficient:

$$K_{ij} = \frac{1}{2} \times \frac{\widehat{\mathbf{v}}_{i,t}\cdot\mathbf{e}_{ij} + ||\widehat{\mathbf{v}}_{i,t}\cdot\mathbf{e}_{ij}||}{||\widehat{\mathbf{v}}_{i,t}||}. \tag{14}$$

Hence, the *Centrifugal Force* between pedestrian i and j should be given in the form

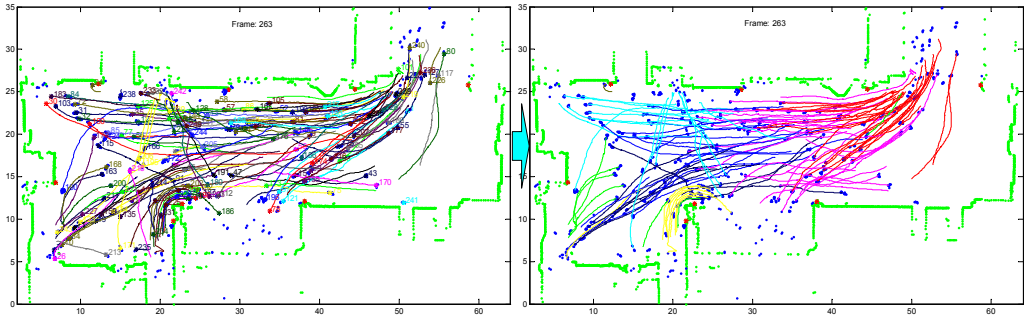
$$\mathbf{F}_{ij} = -m_i K_{ij} \frac{\mathbf{v}_{ij}^2}{||\mathbf{p}_{ij}||}\mathbf{e}_{ij}. \tag{15}$$

Therefore, for pedestrian i, its norm of repulsive force factor of J neighboring persons (as shown in Fig. 5) can be computed by:

$$F_{cent}(\widehat{\mathbf{v}}_{i,t}, \mathbf{p}_{i,t}) = -||\sum_{j=1}^{J} m_i K_{ij} \frac{\mathbf{v}_{ij}^2(\widehat{\mathbf{v}}_{i,t})}{||\mathbf{p}_{ij}||}\mathbf{e}_{ij}||. \tag{16}$$

4. Application of Model by Online Semantic Scene Learning

In order to implement the proposed tracking model, we need the global scene structure to compute Eq.(3), and the local instantaneous crowd flow to compute Eq.(6). In this research, we combine our model into a "tracking-learning loop" [29], which can make the overall process can be fully online and automatic: once the tracking results are obtained with our model, they are collected and can be



(a) Tracking results (b) Crowd flow

Figure 7: Crowd flow learning. With the help of tracking results (Fig. a), the crowd flow (Fig. b) can be easily obtained by online clustering.

in turn utilized for semantic scene learning. In this section, we will provide the details about them.

4.1. Global Scene Structure Learning

With the proceeding of tracking, it is easy for us to obtain a large number of trajectories of pedestrians, and we can utilize them to learn the global scene structure. Firstly, we should estimate the spatial extent of these trajectories and it can be described by the density distribution. Given all the trajectories $L_{i,t}(x, y)$ we obtained at time t, the density distribution at position (x, y) is estimated as:

$$\mathfrak{D}_{global}(x, y, t) = \sum_{(x_i, y_i) \in L_{i,t}} \sum_{L_{i,t} \in \Omega} \exp(-\|(x - x_i, y - y_i)\|^2 / \eta_d), \tag{17}$$

where Ω is all the trajectories we have obtained at time t, η_d a constant parameter. Therefore, the dominant paths of the scene were easily extracted by thresholding the incremental global density distribution (as shown in Fig. 6).

On the other hand, the exit and entrance of the scene are two important scene properties, which are also called sources/sinks. The sinks/sources can be easily detected from the global density distribution $\mathfrak{D}_{global}$. As shown in Fig. 6-c, the sinks/sources usually occur at the region of great change of the global density distribution after thresholding. Moreover, the changed direction must follow the principal orientation of the crowd flow. Hence, the location $\mathbf{p}^*_{sink}$ of sinks/sources

can be easily detected by a gradient searching at the principal orientation of density distribution:

$$\mathbf{p}^*_{sink} = \arg\max_{\mathbf{p}_{sink}}(< \nabla \mathfrak{D}_{S_p(t)}(\mathbf{p}_{sink}, t), \vec{v}_p/|\vec{v}_p| >), \tag{18}$$

where where $\vec{v}_p$ is the principal direction of the crowd flow $S_p(t)$, and $\mathfrak{D}_{S_p(t)}$ is its density distribution.

An example is illustrated in Fig. 6, as the 1090 frames proceeded, the dominant paths and sinks/sources of the scene were obtained, and this scene structure map can be utilized to find persons's planned path to compute E_{global} in Eq.(3).

4.2. Crowd Flow Learning

The crowd flow $S_p(t)$ can be seen as a group of persons who have similar motion and spatial information, they are usually going to the same destination with close velocity. Hence, once we obtain the trajectories of pedestrians at time t, they should be clustered into n clusters $\{S_p(t)\}_{p=1}^n$. In order to dynamically reflect the change of crowd flow, the clustering must be online and can be seen as a function of time t. We consider each cluster $S_j(t)$ as a moving hyperplane. Thus, we can model a union of n hyperplane in $\mathbb{R}^D$, where $S_p(t) = \{\mathbf{x} \in \mathbb{R}^D : \mathbf{b}_p^\top(t)\mathbf{x} = 0\}, p = 1, ..., n$, where $\mathbf{x}$ is the person state, $\mathbf{b}(t) \in \mathbb{R}^D$, as the zero set of a polynomial with time varying coefficients using normalized gradient descent. Then the hyperplane normals are estimated from the derivatives of the new polynomial at each trajectory. Lastly, the trajectories are grouped by clustering their associated normal vectors. An example is illustrated in Fig. 7, for more details about this part, please refer [30, 29].

On the other hand, we have to control the influence of the crowd flow in the energy function. Let us come back to Eq.(1), for the pedestrian i, the influence from crowd flow $S_p(t)$ will be depend on the density distribution of this crowd flow, and it can be computed by:

$$\alpha_{i,t} = 1 - \exp(-\mathfrak{D}_{S_p(t)}(\mathbf{p}_{i,t}, t)), \tag{19}$$

where pedestrian $\mathbf{x}_{i,t} \in S_p(t)$, $\alpha_{i,t} \in (0, 1)$, and this equation can be understood like this: at a specific time t, a person $\mathbf{x}_{i,t}$ in crowd flow $S_p(t)$ were walking in position $\mathbf{p}_{i,t}$. If this area in this crowd flow is quite crowded, this person's motion will be greatly influenced by this crowd flow. In contrast, this influence will be little.

In summary, we can utilize the online learned scene structure and crowd flow to compute Eq.(3) and Eq.(6). Then, the *next desired velocity* $\widehat{\mathbf{v}}_{i,t}$ for pedestrian i

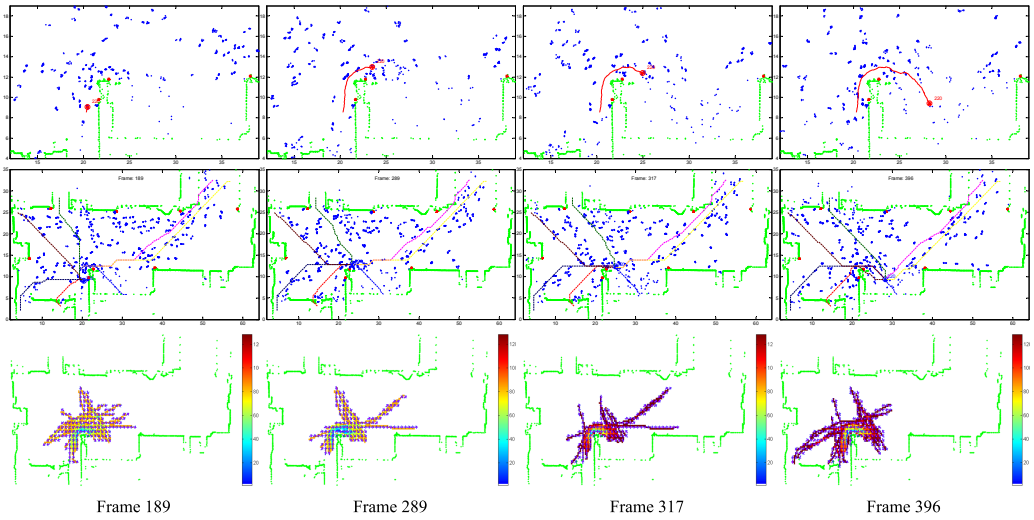


Figure 8: Prediction results without updating. The first row is the prediction results of target 220, the second row is its global planned path, and the third row is the motion distribution of its crowd flow. Please see our supplementary video for more details.

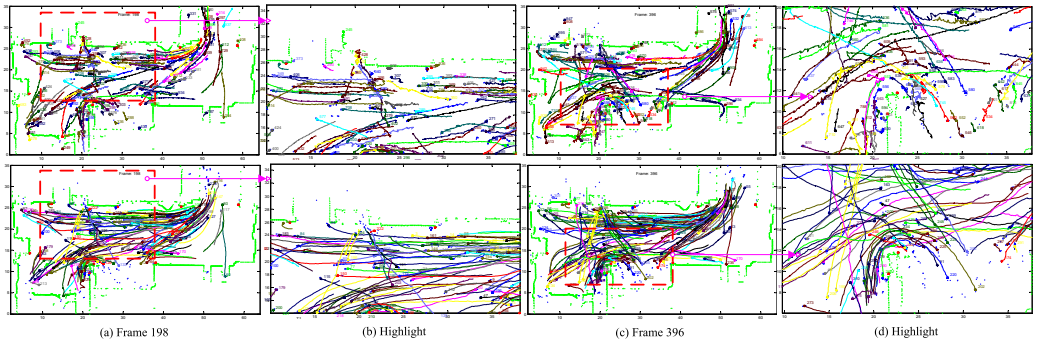
can be computed by Eq.(1) and Eq.(2). Lastly, the state $\mathbf{x}_{i,t}$ of pedestrian i can be easily obtained by any Bayesian filter, such as *Kalman filter* or *particle filter*.

5. Experiments and Results

We utilized the proposed tracking model to track hundreds of pedestrians in the lobby of JR subway station (about 60m$\times$35m). Eight single-row laser scanners (LMS291) produced by SICK were utilized. They were set above 10cm on the ground surface and performing the horizontal scanning with a frequency of 37 fps. We utilized a time server to deal with time synchronization problem between different sensors and the calibration was conducted by several control points in a box. For more details about this part, please refer [1]. The selected data used for evaluation was from 7:00 am to 8:30 am when was a very busy time in Tokyo. We utilized genetic algorithms (GA) to optimize Eq.(2), and σ_1, σ_2 in Eq.(3) and Eq.(6) was set to 0.3. In this section, we will present our experimental results and perform the quantitative evaluation and comparison.

5.1. Tracking Results

In order to test the prediction performance of our model, we utilized it to track single target in crowded environments without any observation updating, and Fig.



(a) Frame 198 (b) Highlight (c) Frame 396 (d) Highlight

Figure 9: Results of multi-target tracking. The first row is the results with second-order motion model, and the second row is our results. We can see that the trajectories with second-order motion model was quite short and frequently broke. In contrast, our model provided more accurate results. Please see our supplementary video for more details.

8 shows the details about this experiment. The first row is the prediction results of target 220, the second row is its global planned path, and the third row is the velocity distribution of its crowd flow. In frame 289 and 317, we can see that the observation is terrible, which is quite difficult for some detection-based trackers. However, our model still maintained the correct tracking of target 220 (as shown in frame 396).

Then, we implemented our model with the observation updating in a particle filter framework [31], and utilized it to track multiple targets in the high density scenarios. The tracking results are shown in Fig. 9, where the first row is the results with second-order motion model, and the second row is our results. From this figure, we can see that the trajectories with second-order motion model is quite short and frequently broke. In contrast, our model provide more accurate results.

5.2. Performance Evaluation

To evaluate the performance of the proposed model, we selected 5000 continuous frames which occlusions or merged measurements frequently took place and made a statistical survey about how many challenging situations (such as occlusions or merged measurements condition) we could deal with. The ground truth was obtained by a semi-automatic way: trackers+manual labeling. We firstly utilized the particle filter-based trackers [31] to obtain the rough results of pedestrians, and then manually edited or labeled some incorrect tracking results. Moreover, we set up two cameras in the experimental site (as shown in Fig. 10) to help

Figure 10: Recorded video for ground truth labeling. We set up two cameras in the major exsit/entrance of station where large numbers of pedestrians would pass. With the help of these recorded videos, some confusing trajectories were able to be easily recognized and correctly labeled.

us label the ground truth. For data synchronization, each laser scan and video stream was stamped with a time log at the moment it was captured, or started to be captured, using the client computers local clock, which was corrected periodically according to that of the server computer. For more details about it, please refer [32]. With the help of these recorded videos (as shown in Fig. 10), some confusing trajectories were able to be easily recognized and correctly labeled.

Based on the ground truth, the occlusions or merged measurements conditions were able to be automatically found. Here, we assumed that if the distance between two pedestrians was smaller than 0.3m, a merged measurements condition was counted. Meanwhile, for a pedestrian's trajectory, if it cannot find any laser measurements at a specific time, a occlusion condition was as well counted. In addition, by comparing the tracking results with the ground truth, it is easy for us to automatically recognize different failed tracking situations, including missed targets, false locations, and identity switches. For more details about this part, please refer to our previous work [33]. Once one of the occlusions or merged measurements condition occurred, but no failed tracking was caused by it, a successful disposal was counted. The details for this are shown in Table 1. From this table, we can see that the trackers with the proposed model are able to easily deal with most occlusions or merged measurements conditions, which is difficult for the second-order motion model.

In addition, we also evaluated the tracking accuracy under different crowd density conditions in these 5000 frames, and the results are shown in Fig. 11. The bottom figure shows the normalized pedestrian density in 5000 frames, and the top

Table 1: Disposal of Challenging Situations

	Occlusion Situation		Merged Measurements	
	Total/Disposal	Disposal Rate	Total/Disposal	Disposal Rate
Tracker with proposed model	8932/7365	**82.46**%	28671/23293	**81.24**%
Tracker with second-order model	8932/3851	43.11%	28671/10137	35.36%

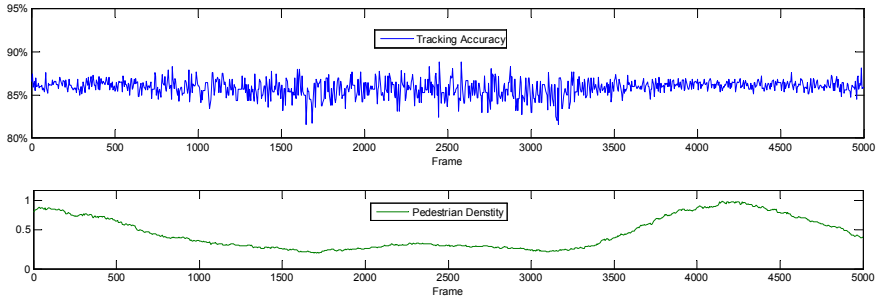


Figure 11: Tracking accuracy under different crowd density conditions. The bottom figure shows the normalized pedestrian density in 5000 continuous frames, and the top figure shows the tracking accuracy of our model in these frames.

figure shows the tracking accuracy in these frames. In this evaluation, we found that there were approximately 1367 pedestrians passing through exists/entrances in these frames, and the proposed tracker obtained 86% accuracy in average under various crowd density situations.

5.3. Quantitative Comparison

A quantitative comparison was conducted among four methods: Song *et al.* [29], Cui *et al.* [34], second-order motion model [13]+particle filter (PF) tracker and our model+ PF tracker.

We made a statistical survey of 5000 continuous frames to evaluate the tracking performance of these methods in the high density scene. The ground truth was obtained by the same way as discussed in the last subsection. The failed tracking included missed targets, false location and identity switch, which was able to be automatically computed from the ground truth. The details about this are illustrated in Fig. 12 and the overall success rate of these methods is shown in Table 2. From Fig. 12, we can see that our model has the best performance among all the methods in the high density scene and the proposed dynamic model provides

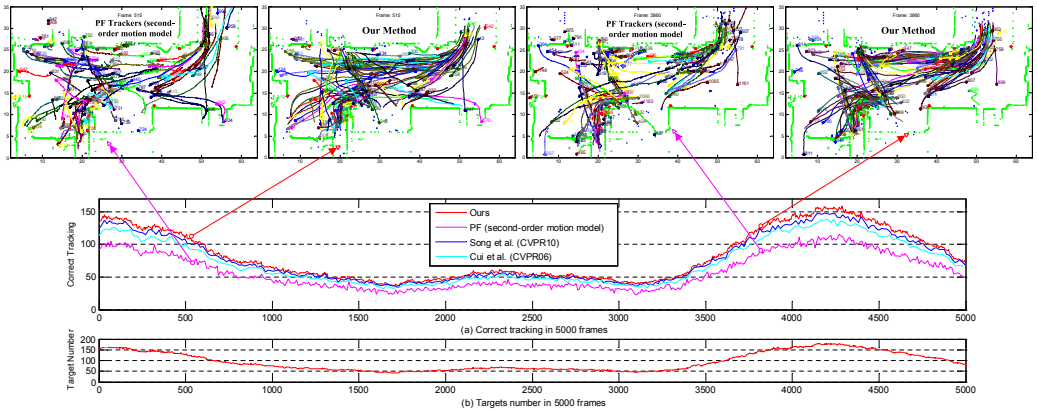


Figure 12: Quantitative comparison among four methods. (a) shows the correct tracking of four methods in 5000 continuous frames. (b) shows the target number in these frames.

Table 2: Success rate among four methods

Algorithm	Success Rate	Missed Targets	False Location	Identity Switch
Song *et. al* (CVPR10)	82.3%	52.3%	26.3%	21.4%
Cui *et. al* (CVPR06)	77.6%	43.7%	22.5%	33.8%
PF (second-order motion model)	62.7%	39.2%	28.3%	32.5%
Our Method	**86.8%**	31.6%	32.7%	35.7%

approximately 24% performance improvements over second-order motion model. As the illustrated in frame 515 and 3860, the trajectories obtained by the trackers with second-order motion model were inaccurate and frequently broken. In contrast, with the help of our model, the trackers were able to easily maintain the long time and robust tracking.

5.4. Computational Cost Evaluation

The computational cost evaluation was conducted on the same 5000 continuous frames. The core algorithms of the proposed model were implemented by the non-optimized MATLAB code, and the evaluations were run on a PC with the Core i7-2620M/S2 Intel-processor. The computational cost of proposed model mainly contained five parts: global scene structure energy computation (it did not contain the scene structure learning, we utilized the off-line learned one in this experiment), local crowd flow energy computation, repulsive force factor computation, GA optimization computation, and some other trivial computation. Hence,

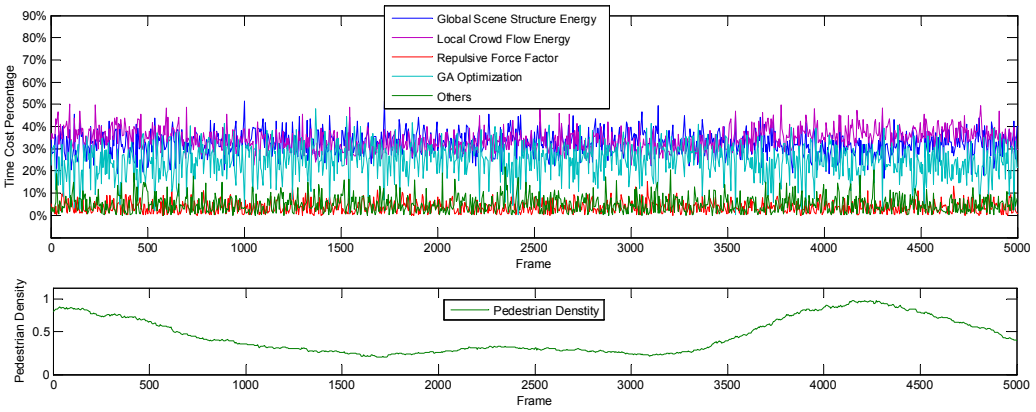


Figure 13: Computational cost evaluation. The bottom figure shows the normalized pedestrian density in continuous 5000 frames, and the top figure shows the computational cost percentage of different parts for average person in these frames.

in this evaluation, we computed the computational cost percentage of these parts for average person in these frames. The details of this evaluation are shown in Fig. 13, the bottom figure shows the normalized pedestrian density in 5000 frames, and the top figure shows the computational cost percentage in these frames.

From this figure, we can see that the global scene structure energy and local crowd flow energy computation were usually in a large proportion, and they were mainly caused by the path planning and online crowd flow learning. Meanwhile, the computational cost of GA Optimization is very unstable: in some cases, several iteration round after, GA algorithm was able to find local optimum, and only took approximately 7% of total computational cost. But in some bad cases, the optimization did not execute successfully, and it almost took more than 45% of total computational cost. In contrast, the computational cost of repulsive force factor was usually in a very small proportion. At present, because all of the algorithms are implemented by the non-optimized MATLAB code, the computational cost of proposed model is a bit expensive, and it usually needs 3-4s per frame for single pedestrian in average. However, it will have a much lower computational cost if it is implemented by the well-written C++ code, and it is able to be applied into more real-world applications.

6. Conclusion and Discussion

In this paper, we present a novel dynamic model for tracking hundreds of persons in the extremely crowded environments. This model incorporates the inten-

sively explored semantic scene knowledge and social interactions among persons, and the experimental results on laser-based system have demonstrated its feasibility and robustness.

In the future, this research should be also improved and extended in the following ways: (1) At present, we have not applied the proposed model to the video data because the measurements from the two sensors are quite different [32]. From the application view, that is meaningful and necessary because the camera sensors are much cheaper and more common than the laser scanners. However, compared to the laser-based tracking system, the facing challenges of camera-based ones are also obvious: (a) The computational cost of image processing will be much huger than the laser scanner data. (b) It is usually difficult to obtain the robust detections in real environment. (c) It is sensitive to the change of lighting and weather conditions. In the future, we will try to implement and test the proposed model on the video data. In such cases, the comparisons or evaluations with the state of the art tracking model [21, 17] for video are able to be conducted. (2) We find that our method sometimes does not perform well for events in which a person goes against the dominant flow of a dense crowd, and some strategies [35] should be applied to deal with them. (3) At present, the computation cost of our method is a bit expensive, some optimization strategy should be conducted. (4) Based on the proposed model, some abnormal detection approaches are able to be explored in future.

7. Acknowledgement

This work was supported by a Grant-in-Aid for Young Scientists (23700192), GRENE (Environmental Information) project, and "Strategic Project to Support the Formation of Research Bases at Private Universities": Matching Fund Subsidy by Japan's Ministry of Education, Culture, Sports, Science, and Technology (MEST). This work was also partially funded by NSFC Grants (No.60975061) of China, East Japan Railway Company and Microsoft Research. We thank all anonymous reviewers for their helpful comments on this work.

References

[1] H. Zhao, R.Shibasaki, A novel system for tracking pedestrians using multiple single-row laser range scanners, IEEE Transactions on Systems, Man and Cybernetics, part A (2005) 283–291.

[2] A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, ACM Computing Surveys (2006) 3–47.

[3] Y. Bar-Shalom, T. E. Fortmann, Tracking and data association, New York: Academic Press.

[4] G. Gennari, G. Hager, Probabilistic data association methods in visual tracking of groups, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2004) 876–881.

[5] J. Vermaak, S. Godsill, P. Perez, Monte carlo filtering for multi target tracking and data association, IEEE Trans. Aerospace and Electronic Systems 41 (1) (2005) 309–332.

[6] D. Schulz, W. Burgard, D. Fox, A. Cremers, People tracking with a mobile robot using sample-based joint probabilistic data association filters, International Journal of Robotics Research 22 (2) (2003) 99–116.

[7] Z. Khan, T. Balch, F. Dellaert, Mcmc data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements, IEEE Trans. on Pattern Analysis and Machine Intelligence 28 (1) (2006) 1960–1972.

[8] Q. Yu, G. Medioni, I. Cohen, Multiple target tracking using spatio-temporal markov chain monte carlo data association, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2007) 642–649.

[9] H. Jiang, S. Fels, J. Little, A linear programming approach for multiple object tracking, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2007) 1380–1387.

[10] B. Leibe, K. Schindler, N. Cornelis, L. V. Gool, Coupled detection and tracking from static cameras and moving vehicles, IEEE Trans. on Pattern Analysis and Machine Intelligence (2008) 1683–1698.

[11] L. Zhang, Y. Li, R. Nevatia, Global data association for multi-object tracking using network flows, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2008) 1–8.

[12] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. V. Gool, Robust tracking-by-detection using a detector confidence particle filter, Proc. IEEE International Conference on Computer Vision (2009) 1515 – 1522.

[13] X. Song, J. Cui, H. Zha, H. Zhao, Vision-based multiple interacting targets tracking via on-line supervised learning, Proc. European Conference on Computer Vision (2008) 642–655.

[14] M. Andriluka, S. Roth, B. Schiele, People-tracking-bydetection and people-detection-by-tracking, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2008) 1555–1562.

[15] C. Kuo, C. Huang, R. Nevatia, Multi-target tracking by on-line learned discriminative appearance models, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2010) 685–692.

[16] J. Fan, X. Shen, Y. Wu, Closed-loop adaptation for robust tracking, Proc. European Conference on Computer Vision (2010) 411–424.

[17] S. Pellegrini, A. Ess, K. Schindler, L. van Gool, You will never walk alone: Modeling social behavior for multi-target tracking, Proc. IEEE International Conference on Computer Vision (2009) 261 – 268.

[18] L. Kratz, K. Nishino, Tracking with local spatio-temporal motion patterns in extremely crowded scenes, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2010) 693 – 700.

[19] B. Ziebart, N. D. Ratliff, G. Gallagher, C. Mertz, K. M. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, S. S. Srinivasa, Planning-based prediction for pedestrians, Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (2009) 3931–3936.

[20] C. Wang, T. Lo, S. Yang, Interacting object tracking in crowded urban areas, Proc. IEEE International Conference on Robotics and Automation (2007) 4626–4632.

[21] S. Ali, M. Shah, Floor fields for tracking in high density crowd scenes, Proc. European Conference on Computer Vision (2008) 1–14.

[22] M. Rodriguez, S. Ali, T. Kanade, Tracking in unstructured crowded scenes, Proc. IEEE International Conference on Computer Vision (2009) 1389 – 1396.

[23] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, Physical Review E 77 (5) (1995) 4282C4286.

[24] W. Yu, R. Chen, L. Dong, S. Dai, Centrifugal force model for pedestrian dynamics, Physical Review E 72 (2) (2005) 112C119.

[25] R. Mehran, A. Oyama, M. Shah., Abnormal crowd behavior detection using social force model, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2008) 935–942.

[26] G. Antonini, S. V. Martinez, M. Bierlaire, J. Thiran, Behavioral priors for detection and tracking of pedestrians in video sequences, International Journal of Computer Vision 69 (2006) 159–180.

[27] M. Luber, J. A. Stork, G. D. Tipaldi, K. O. Arras, People tracking with human motion predictions from social forces, Proc. IEEE International Conference on Robotics and Automation (2010) 3264–3269.

[28] X. Wang, K. Tieu, E. Grimson, Learning semantic scene models by trajectory analysis, Proc. European Conference on Computer Vision (2006) 110–123.

[29] X. Song, X. Shao, H. Zhao, J. Cui, R. Shibasaki, H. Zha, An online approach: Learning-semantic-scene-by-tracking and tracking-by-learning-semantic-scene, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2010) 739–746.

[30] R. Vidal, Online clustering of moving hyperplanes, Proc. Neural Information Processing Systems (2006) 1433–1440.

[31] X. Shao, H. Zhao, K. Nakamura, K. Katabira, R. Shibasaki, Y. Nakagawa, Detection and tracking of multiple pedestrians by using laser range scanners, Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (2007) 2174–2179.

[32] X. Song, H. Zhao, J. Cui, X. Shao, R. Shibasaki, H. Zha, Fusion of laser and vision for multi-target tracking via on-line learning, Proc. IEEE International Conference on Robotics and Automation (2010) 406–411.

[33] X. Song, J. Cui, X. Wang, H. Zhao, H. Zha, Tracking interacting targets with laser scanner via on-line supervised learning, Proc. IEEE International Conference on Robotics and Automation (2008) 2271–2276.

[34] J. Cui, H. Zha, H. Zhao, R.Shibasaki, Fusion of detection and matching based approaches for laser based multiple people tracking, Proc. IEEE International Conference on Computer Vision and Pattern Recognition (2006) 642–649.

[35] M. Rodriguez, J. Sivic, I. Laptev, J. Audibert, Data-driven crowd analysis in videos, Proc. IEEE International Conference on Computer Vision (2011) 1235–1242.