A

A Fully Online and Unsupervised System for Large and High Density Area Surveillance: Tracking, Semantic Scene Learning and Abnormality Detection

XUAN SONG, Center for Spatial Information Science, The University of Tokyo
XIAOWEI SHAO, Center for Spatial Information Science, The University of Tokyo
QUANSHI ZHANG, Center for Spatial Information Science, The University of Tokyo
RYOSUKE SHIBASAKI, Center for Spatial Information Science, The University of Tokyo
HUIJING ZHAO, Key Laboratory of Machine Perception (MoE), Peking University
JINSHI CUI, Key Laboratory of Machine Perception (MoE), Peking University
HONGBIN ZHA, Key Laboratory of Machine Perception (MoE), Peking University

For reasons of public security, an intelligent surveillance system that can cover a large, crowded public area has become an urgent need. In this paper, we propose a novel laser-based system that can simultaneously perform tracking, semantic scene learning, and abnormality detection in a fully online and unsupervised way. Furthermore, these three tasks co-operate with each other in one framework to improve their respective performances. The proposed system has the following key advantages over previous ones: (1) It can cover quite a large area (more than 60×35m), and simultaneously perform robust tracking, semantic scene learning, and abnormality detection in a high-density situation. (2) The overall system can vary with time, incrementally learn the structure of the scene, and perform fully online abnormal activity detection and tracking. This feature makes our system suitable for real-time applications. (3) The surveillance tasks are carried out in a fully unsupervised manner, so that there is no need for manual labeling and the construction of huge training datasets. We successfully apply the proposed system to the JR subway station in Tokyo, and demonstrate that it can cover an area of 60×35m, robustly track more than 150 targets at the same time, and simultaneously perform online semantic scene learning and abnormality detection with no human intervention.

Categories and Subject Descriptors: I.2.10 [**Vision and Scene Understanding**]: Perceptual reasoning

General Terms: Measurement, Algorithms

Additional Key Words and Phrases: Surveillance, Multi-target Tracking, Abnormality Detection, Semantic Scene Learning

ACM Reference Format:
Song, X., Shao, X., Zhang, Q., Shibasaki, R., Zhao, H., Cui, J., Zha, H. 2012. A Fully Online and Unsupervised System for Large and High Density Area Surveillance: Tracking, Semantic Scene Learning and Abnormality Detection. ACM Trans. Intell. Syst. Technol. V, N, Article A (January YYYY), 20 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The task of surveillance is to monitor the activities of people in a scene. This requires low-level detection, tracking, and classification, as well as high-level activity analysis and abnormality detection. Both the low-level and high-level tasks can be improved by knowledge of the scene structure (e.g., crowd flow, dominant paths, entry or exit). For instance, "people and cars are moving on d-

Authors' addresses: X. Song (corresponding author), X. Shao, Q. Zhang and R. Shibasaki, Center for Spatial Information Science, The University of Tokyo, Japan. Tel.:+81-04-7136-4307. Fax:+81-04-7136-4292. E-mail:{songxuan,shaoxw,zqs1022,shiba}@csis.u-tokyo.ac.jp; H. Zhao, J. Cui and H. Zha, Key Laboratory of Machine Perception (MoE), Peking University, China. Tel.:+86-10-6275-5569. Fax:+86-10-6275-5654 E-mail:{zhaohj,cjs,zha}@cis.pku.edu.cn.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

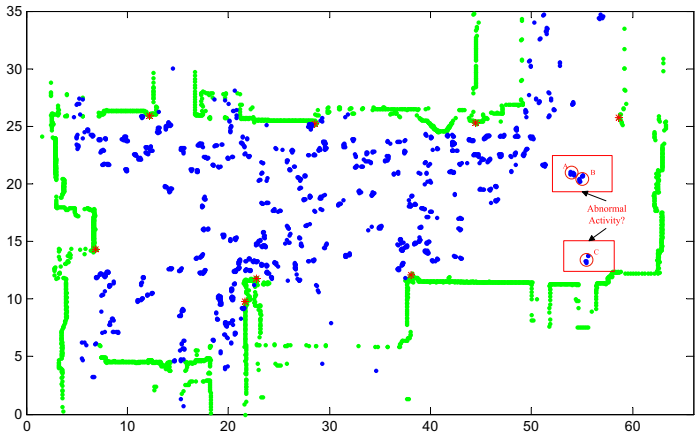


Fig. 1. How can accurate tracking and abnormal activity detection be maintained in a high density scene? This is the JR subway station in Tokyo, and the data was obtained by eight single-row laser scanners. The green points are the background, the blue points are the foreground, and the red points show the position of each single-row laser scanner. In this case, each person is represented by several 2D points. Persons A, B, and C were walking on a closed road, how can we detect their activities?

ifferent roads," "people only appear/disappear at an entry/exit," "people 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. However, scene knowledge cannot be dynamically learned and updated under existing methods. Furthermore, tracking is the basis of surveillance, playing a crucial role in any kind of monitoring task, but this becomes especially challenging in a high-density, crowded scene, as shown in Figure 1.

In addition, abnormal activity detection is a key component of a surveillance system, and there is an increasing demand for a robust system to detect abnormal activity in a large and crowded area, such as a subway station, public square, or intersection (as shown in Figure 1). However, existing systems are usually based on cameras, which can only cover a small area, suffer from changes in weather conditions, or require the unreliable process of data labeling, and need huge training datasets. Moreover, most of these systems usually face difficulties in performing fully online abnormal activity detection. Due to the real-time nature of many surveillance applications, it is very desirable to have a completely online and automatic system that runs robustly and requires little human intervention. *Thus, the purpose of this paper is to develop such a fully online system that can not only cover a large and high-density area, but also simultaneously perform the tracking, semantic scene learning, and abnormality detection robustly in an unsupervised way.*

The novelty of the proposed system is that tracking, semantic scene learning, and abnormality detection are integrated and can supplement each other in one framework. The key idea of this research and the proposed system is depicted by Figure 2: multiple single-row laser scanners are integrated and cover a large area, providing robust measurements of pedestrians. The tracking module then provides the system with a large number of trajectories. Thus, knowledge of the scene structure (e.g., dynamic vs. static properties) can be dynamically learned and updated via an online unsupervised learning method. In the meantime, the learned statistical scene model supervises the tracking module to make the results increasingly accurate. Finally, abnormal activity can be detected with the tracking results and the learned semantic scene, both globally and locally. Therefore, this mode of co-operation among tracking, semantic scene learning, and abnormality detection becomes "an adaptive loop," not only dynamically reflecting the change of scene structure and detecting abnormal activity, but also solving the tough problems encountered during tracking. Moreover, the entire process is completely online and automatic, thus requiring no human intervention.

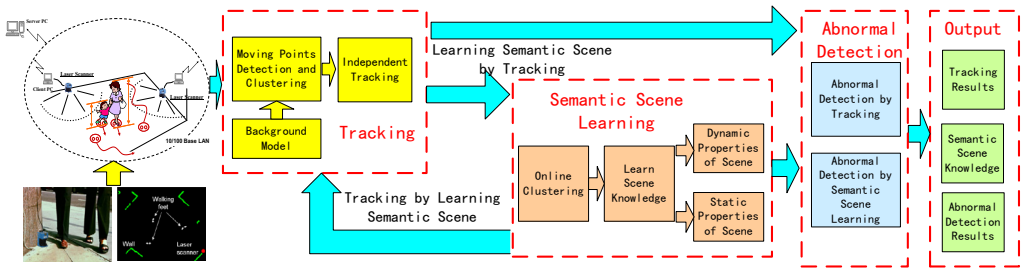


Fig. 2. Overview of the proposed system. There are four main modules: sensor module, tracking, semantic scene learning, and abnormality detection. They supplement each other in the overall system to improve their respective performances. The final output of our system consists of the trajectories of pedestrians, learned scene knowledge, and detected abnormal activities.

The main contributions of this paper can be summarized as follows: *(1) We develop a unified framework that couples the tracking, semantic scene learning, and abnormality detection, which then supplement one another. (2) We apply an online learning algorithm [Vidal 2006] to the trajectories analysis, and make this task can be firstly online. (3) We develop a fully online abnormality detection method, with the help of tracking and semantic scene learning, which can perform robustly in high-density areas in an unsupervised manner. (4) We present the first application of an online system that can robustly track more than 150 targets concurrently while performing abnormality detection in a real-life scene (JR subway station, Tokyo).*

The remainder of this paper is structured as follows. In the next section, we briefly review related work, and provide a system overview in Section 3. Section 4, Section 5, and Section 6 present details about the semantic scene learning by tracking, tracking by semantic scene learning, and abnormality detection by tracking and semantic scene learning. Experiments and results are presented in Section 7, and the paper is summarized in Section 8.

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.* [Yilmaz et al. 2006]. Typically, multi-target tracking can be solved through data association [Bar-Shalom and Fortmann 1998], which consists of the linear complexity-based methods [Jiang et al. 2007] and exponential complexity-based: Multi-Hypothesis Tracker (MHT) [Read 1979], Joint Probabilistic Data Association Filter (JPDAF) [Bar-Shalom and Fortmann 1998; Gennari and Hager 2004; Rasmussen and Hager 2001], Monte Carlo technique based JPDA algorithms (MC-JPDAF) [Schulz et al. 2003; Vermaak et al. 2005], and Markov chain Monte Carlo data association (MCMC-DA) [Khan et al. 2006; Yu et al. 2007]. In addition, MTT will encounter incredible difficulties when interactions or occlusions frequently take place among targets. Thus, research has also focused on how to model these interactions and solve the "merge/split" problem. Representative publications include [Bose et al. 2007; J.Sullivan and S.Carlsson 2006; Lanz and Manduchi 2005; Leibe et al. 2007; P.Nillius et al. 2006; Qu et al. 2005; Yang et al. 2007; Zhao and Nevatia 2004]. More recently, there has been a trend of introducing online learning techniques into target tracking problems, and representative publications include [Babenko et al. 2009; Ross et al. ; Song et al. 2008; Avidan 2007]. However, most of the methods mentioned above are difficult to apply to the tracking of hundreds of targets in a high-density, crowded scene. To track a large number of targets in crowded environments, Betke *et al.* [Betke et al. 2007] proposed a two cluster-based data association approaches that are linear in the number of detections and tracked objects. However, this method is difficult to use in human-based surveillance applications. Ali *et al.* [Ali and Shah 2008] proposed a floor fields-based method for tracking people in a crowded scene, which is very related to our work. The main difference between their method and ours is that: the computation of a dynamic floor field in [Ali and Shah 2008] at a particular time

period should use future information, which is not a completely online approach. Moreover, our proposed method is not only a tracking system, but also a semantic scene learning and abnormality detection system, which is quite different from [Ali and Shah 2008].

Abnormal activity detection has, however, been an active area of research over the years, and an in-depth review of relevant literature can be found in a recent survey by Chandola *et al.* [Chandola et al. 2009]. Typically, abnormality detection approaches can be broadly categorized as supervised learning-based and unsupervised learning-based. The supervised learning-based methods must usually pre-define the known a priori behavior classes (normal activities and abnormal ones), and utilize the supervised learning model to detect the abnormal ones. Representative publications include [Mahadevan et al. 2010; Wu et al. 2010; Adam et al. 2008; Mehran et al. 2009; Patino et al. 2010]. Hence, these methods usually need huge training datasets, which require significant human resources and efforts. In addition, the process of manual labeling is usually unreliable, and sometimes it is quite difficult to obtain enough abnormal training samples.

In addition, researchers have also proposed some methods to detect abnormal activity in an unsupervised way [Hu et al. 2007; Pusiol et al. 2010; Fu et al. 2005; Junejo and Foroosh 2007; Junejo et al. 2004; Makris and Ellis 2003]. These methods are usually based on trajectory analysis with the help of tracking. The trajectories obtained are clustered, and some small clusters are labeled as abnormal. Moreover, the semantic scene can be easily learned via the statistical trajectories model [Zhang et al. 2009; Wang et al. 2008; Wang et al. 2006]. However, most of these methods are batch, i.e., the clustering of trajectories is obtained after all the data has been collected and their cluster structure cannot change as a function of time. Hence, they are difficult to apply in online and real-time applications. Recently, some promising camera-based systems have been proposed [Xiang and Gong 2008; Wang et al. 2010; Saleemi et al. 2009; Loy et al. 2010] that can perform intelligent surveillance and understand human activities. Due to some essential camera defects, it is difficult for these systems to perform robustly while the weather or light conditions are frequently changing. Moreover, they cannot be applied in some extremely crowded public spaces. Hence, in this paper, we firstly propose a novel system (the earlier version is [Song et al. 2010]) that can cover a large and crowded area, simultaneously outputing robust tracked trajectories of pedestrians, semantic scenes in both dynamic (crowd flow) and static (paths, exit/entrance) aspects, and abnormality detection results.

3. SYSTEM OVERVIEW

In this research, we couple the tracking, semantic scene learning, and abnormality detection in a unified framework, and design the overall system as illustrated in Figure 2. The proposed system consists of four main components: sensor module, tracking module, semantic scene learning module, and abnormality detection module. In the sensor part (as shown in Figure 2), a number of laser scanners are exploited so that a relatively large area can be covered while occlusions can, to an extent, be solved. Each laser scanner is located at a separate position and controlled by a client computer. All client computers are connected through a network to a server computer, which synchronizes and integrates all of the data from the client computers. For data synchronization, each laser scan stream is stamped with a time log at the moment it is captured, or starts to be captured, using the client computer's local clock, which is synchronized periodically with that of the server computer. Data measured by different client computers that is stamped with the same time log is aligned to make up an integrated frame. For registration, a degree of overlay between different laser scans is retained. Relative transformations between the local coordinate systems of neighboring laser scanners are calculated by the pair-wise matching of background images using the distances to common objects. Based on the registration, the range data can be easily converted into rectangular coordinates (2D laser points) in the sensors' local coordinate system, and then the laser points from multiple laser scanners are temporally and spatially integrated into a global coordinate system. The final sensor measurements of our system are illustrated in Figure 1. For more details about the system platform, please refer to our previous work [Zhao and R.Shibasaki 2005].

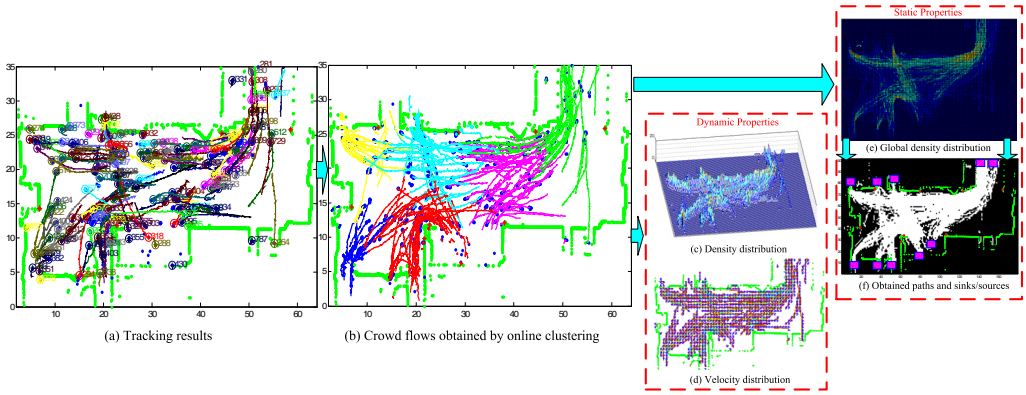


Fig. 3. Semantic scene learning by tracking. Once we obtain the tracking results (a), these trajectories are clustered into different crowd flows via an online unsupervised learning (b). We could then learn scene knowledge: (1) Dynamic properties: density distribution (c) and velocity distribution (d) of the crowd flow. (2) Static properties: walk paths and sinks/sources (f). Note that the arrows in (d) show the principal orientation at each position. Please refer to the text for more details.

The tracking module, semantic scene learning module, and abnormality detection module closely co-operate with and supplement each other in the overall system: at first, the tracking module provides the preliminary tracking results; based on these results, the semantic scene knowledge can be automatically learned online using the learning module. Meanwhile, the learned statistical scene model is used in turn to supervise and improve the tracking results. As time proceeds, with the increasing number and accuracy of tracked trajectories, the semantic scene model is increasingly accurate and can provide better supervision for the tracking module, which makes the tracking results more robust. Hence, we call this mode of co-operation between the tracking and learning modules "learning by tracking" and "tracking by learning." Furthermore, the trajectories provided by the tracking module and the semantic scene knowledge learned by the learning module are combined and simultaneously utilized for the detection of abnormalities. Thus, the proposed system can simultaneously output tracking trajectories of people, semantic scene knowledge of environments, and some abnormal activities. In the following sections, we provide details on how to learn semantic scene knowledge with the tracking results, how to use the statistical scene model to improve tracking, and how to detect some abnormal activities.

4. SEMANTIC SCENE LEARNING BY TRACKING

As illustrated in Figure 3, with the help of tracking, it is easy for us to obtain a large number of trajectories. We can use these trajectories to explore knowledge of a scene at a specific time. Firstly, we should cluster these trajectories online based on different activity types. This is very easy to understand. As shown in Figure 3-(a), at a specific time in a subway station, a large number of people get off a train and walk together to catch another train. This would become a crowd flow, and can be seen as a type of activity. Secondly, we extract knowledge of this scene from these clusters. Scene knowledge contains two properties: dynamic (e.g., information about the crowd flow) and static (e.g., walk paths and sinks/sources). This scene knowledge can provide great assistance to the tracking, and the overall pipeline is illustrated in Figure 3. In this section, we will provide details about these items.

4.1. Online clustering for crowd flows

Problem Formulation:

At time t, person i is represented by $\mathbf{x}_i(t) = (\mathbf{p}_i, \mathbf{v}_i)$, where $\mathbf{p}_i = (x_i, y_i)$ is the persons' position, $\mathbf{v}_i = (v_i^x, v_i^y)$ is its velocity. With the help of the trackers, we obtain the N trajectories $\{L_i(t)\}_{i=1}^N$, where $L_i(t) = \{\mathbf{x}_i(t) : t = 1, ..., T\}$. We need to cluster these trajectories into n

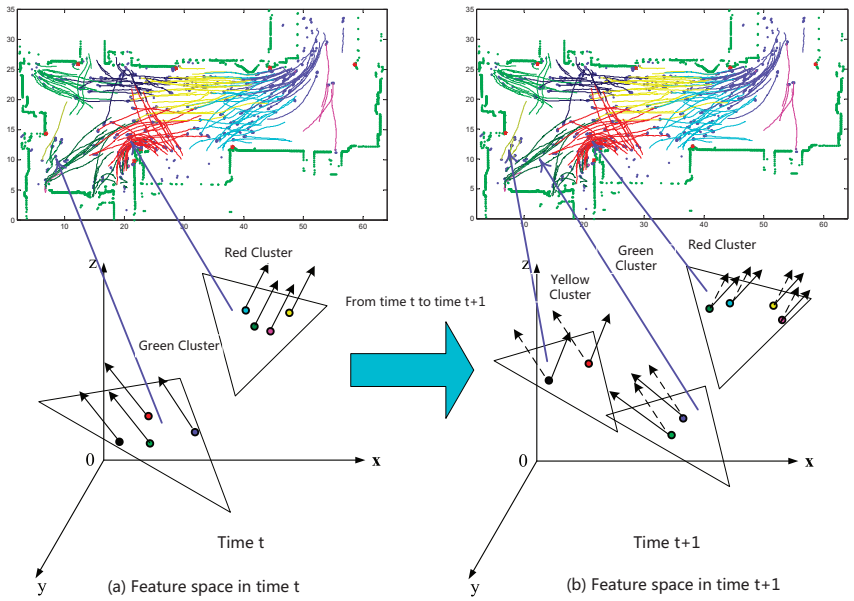


Fig. 4. Key idea of online clustering. We consider each cluster as a moving hyperplane, and there are some vectors (solid line arrows) normal to these hyperplanes. These moving hyperplanes can be seem as the zero set of a polynomial with time varying coefficients. Starting from an initial polynomial at time t, the updated polynomial coefficients are computed using normalized gradient descent, and the hyperplane normals in time $t+1$ are then estimated and updated from the derivatives of the new polynomial based on its previous normal vectors (dotted line arrows). Finally, the trajectories are grouped by clustering their associated normal vectors. Thus, the overall clustering will vary as a function of time, and some hyperplanes will change based on the newly obtained data. For instance, the black and red points in green cluster change to a new cluster (yellow cluster) in time $t+1$. Please note that the illustrated feature space is not the real feature space in this research, and this is just for visualization.

clusters $\{S_j(t)\}_{j=1}^n$ at a specific time t based on their activities, and each cluster can be seem as a crowd flow (as shown in Figure 3-(b)). As time proceeds, some persons in one crowd flow may change their direction of movement, becoming a new crowd flow or associate to another crowd flow. Hence, in order to dynamically reflect such situations and the change of scene information, all of the clustering must be online and and vary as a function of time. Therefore, in this research, we apply the online clustering algorithm [Vidal 2006] developed by *Vidal* to our problem, and deal with online crowd flow learning.

The key idea of overall algorithm is illustrated in Figure 4, and we consider each cluster $S_j(t)$ as a moving hyperplane. Thus, we can model a union of n moving hyperplane in $\mathbb{R}^D$, where $S_j(t) = \{\mathbf{x} \in \mathbb{R}^D : \mathbf{b}_j^\top(t)\mathbf{x} = 0\}, j = 1, ..., n$, where $\mathbf{b}(t) \in \mathbb{R}^D$, as the zero set of a polynomial with time varying coefficients. Starting from an initial polynomial at time t, the updated polynomial coefficients are computed using normalized gradient descent. The hyperplane normals are then estimated and updated from the derivatives of the new polynomial at each trajectory. Finally, the trajectories are grouped by clustering their associated normal vectors. As time proceeds, new data are added, and the estimates of the polynomial coefficients become more accurate. In addition, the hyperplane will change based on the newly obtained data. Hence, the overall clustering will vary as a function of time, and is fully online and automatic.

Algorithm Detail:

In this part, we briefly review the theory of [Vidal 2006], and provide the final clustering algorithm for our problem.

Given a point $\mathbf{x}(t)$ in one of hyperplane $S_j(t)$, there is a vector $\mathbf{b}_j(t)$ normal to it such that $\mathbf{b}_j^\top(t)\mathbf{x}(t) = 0$. In principle, for the N trajectories $\{L_i(t)\}_{i=1}^N$ lying in n hyperplanes $S_j(t)$, we could directly estimate and update their normal vectors through *normalized gradient recursive identifier*. However, due to the uncertain laser measurements and frequently appearing/disappearing of persons from our tracking area, sometimes the tracking results are not reliable. Thus, we do not know which data to use for updating each one of the n identifiers in these cases. Therefore, in this research, the n hyperplanes are represented with a single polynomial whose coefficients do not depend on the segmentation of the data. By updating the coefficients of this polynomial, the normal vectors of all hyperplanes can be simultaneously estimated and updated without first clustering the point trajectories.

Thus, the following homogeneous polynomial of degree n in D variables must vanish at $\mathbf{x}(t)$:

$$p_n(\mathbf{x}(t), t) = (\mathbf{b}_1^\top(t)\mathbf{x}(t))(\mathbf{b}_2^\top(t)\mathbf{x}(t))...(\mathbf{b}_n^\top(t)\mathbf{x}(t)) = 0. \tag{1}$$

This homogeneous polynomial can be written as a linear combination of all the monomials of degree n in $\mathbf{x}$, $\mathbf{x}^I = x_1^{n_1}x_2^{n_2}...x_D^{n_D}$ with $0 \leq n_k \leq n$ for $k = 1, ..., D$, and $n_1 + n_2 + ...n_D = n$, as

$$p_n(\mathbf{x}, t) \doteq \sum e_{n_1,...,n_D}(t)x_1^{n_1}...x_D^{n_D} = \mathbf{e}(t)^\top \mu_n(\mathbf{x}) = 0, \tag{2}$$

where $e_I(t) \in \mathbb{R}$ represents the coefficient of the monomial $\mathbf{x}^I$. The map $\mu_n : \mathbb{R}^D \to \mathbb{R}^{M_n(D)}$ is known as *Veronese map* [Harris 1992] of degree n, which can be defined as:

$$\mu_n : [x_1, ..., x_D]^\top \longmapsto [..., \mathbf{x}^I, ...], \tag{3}$$

where I is chosen in the degree-lexicographic order and $M_n(D) = \binom{n+D-1}{n}$ is the total number of independent monomials.

As a result of polynomial Eq. (2), we can perform the online hyperplane clustering by operating on the polynomial coefficients $\mathbf{e}(t)$ rather than on the normal vectors $\{\mathbf{b}_j(t)\}_{i=1}^n$. That is because $\mathbf{e}(t)$ does not depend on which hyperplane the $\mathbf{x}(t)$ belong to. Thus, at each time t, the estimate $\hat{\mathbf{e}}(t)$ of $\mathbf{e}(t)$ can be computed as

$$\hat{\mathbf{e}}(t) = \arg\min_{\mathbf{e}(t)} f(\mathbf{e}(t)), \tag{4}$$

where the objective function is

$$f(\mathbf{e}(t)) = \frac{1}{N}\sum_{\kappa=1}^{t}\sum_{i=1}^{N}(\mathbf{e}(\kappa)^\top \mu_n(\mathbf{x}_i(\kappa)))^2. \tag{5}$$

By using normalized gradient descent, the following recursive identifier can be updated by

$$\hat{\mathbf{e}}(t+1) = \hat{\mathbf{e}}(t)\cos(\|\mathbf{v}(t)\|) + \frac{\mathbf{v}(t)}{\|\mathbf{v}(t)\|}\sin(\|\mathbf{v}(t)\|), \tag{6}$$

where the negative normalized gradient is computed as

$$\begin{aligned}\mathbf{v}(t) = &-\beta(I_{M_n(D)} - \hat{\mathbf{e}}(t)\hat{\mathbf{e}}^\top(t)) \\ &\times \frac{\sum_{i=1}^N(\hat{\mathbf{e}}^\top(t)\mu_n(\mathbf{x}_i(t)))\mu_n(\mathbf{x}_i(t))/N}{1 + \beta\sum_{i=1}^N \|\mu_n(\mathbf{x}_i(t))\|^2/N},\end{aligned} \tag{7}$$

and where $\beta > 0$ is a fixed parameter.

Once we obtain the estimate of $\mathbf{e}(t)$, it is easy to estimate the normal vector to the hyperplane containing a trajectory $\mathbf{x}(t)$ as

$$\hat{\mathbf{b}}_{vec}(\mathbf{x}(t)) = \frac{D\mu_n^\top(\mathbf{x}(t))\hat{\mathbf{e}}(t)}{\|D\mu_n^\top(\mathbf{x}(t))\hat{\mathbf{e}}(t)\|}, \tag{8}$$

where $D\mu_n(\mathbf{x})$ is the Jacobian of μ_n at $\mathbf{x}$.

Thus, we have obtained the estimate $\hat{\mathbf{b}}_{vec}(\mathbf{x}_i(t))$ for the normal to the hyperplane passing through each one of the N trajectories $\{\mathbf{x}_i(t) \in \mathbb{R}^D\}_{i=1}^N$ at each time instant. The next step is to cluster these normals into n groups. This is done by a recursive K-means algorithm. Essentially, we can seek the normal vectors $\hat{\mathbf{b}}_j(t) \in \mathbb{R}^{D-1}$ and the group indicator $\phi_{ij}(t) \in \{0, 1\}$ of trajectory i to hyperplane j by maximizing

$$f(\{\phi_{ij}(t)\}, \{\hat{\mathbf{b}}_j(t)\}) = \sum_{i=1}^{N} \sum_{j=1}^{n} \phi_{ij}(t)(\hat{\mathbf{b}}_j^{\top}(t)\hat{\mathbf{b}}_{vec}(\mathbf{x}_i(t)))^2. \tag{9}$$

Vidal has proved that the recursive identifier (6)-(8) provide L_2-stable estimates of the parameters and n can be a variable number. For more details about this proof, please refer to [Vidal 2006]. Thus, the overall online clustering algorithm for our problem can be summarized as follows:

Online Clustering Algorithm

Input: N persons' state $\{\mathbf{x}_i(t)\}_{i=1}^N$ at time t.

Output: The group indicator $\{\phi_{ij}(t) : i = 1, ..., N, j = 1, ..., n\}$ at time t, where n is a variable group number.

Initialization:

(1) Randomly choose $\{\hat{\mathbf{b}}_j(1)\}_{j=1}^n$ and $\hat{\mathbf{e}}(1)$.

For each time $t \geq 1$

(1) Update the coefficients of $\hat{p}_n(\mathbf{x}_i(t), t) = \hat{\mathbf{e}}(t)^{\top}\mu_n(\mathbf{x}_i(t))$ by
$\hat{\mathbf{e}}(t+1) = \hat{\mathbf{e}}(t)\cos(\|\mathbf{v}(t)\|) + \frac{\mathbf{v}(t)}{\|\mathbf{v}(t)\|}\sin(\|\mathbf{v}(t)\|)$,
$\mathbf{v}(t) = -\beta(I_{M_n(D)} - \hat{\mathbf{e}}(t)\hat{\mathbf{e}}^{\top}(t))$
$\times \frac{\sum_{i=1}^N(\hat{\mathbf{e}}^{\top}(t)\mu_n(\mathbf{x}_i(t)))\mu_n(\mathbf{x}_i(t))/N}{1+\beta\sum_{i=1}^N\|\mu_n(\mathbf{x}_i(t))\|^2/N}$.

(2) Solve for the normal vectors at the given trajectories
$\hat{\mathbf{b}}_{vec}(\mathbf{x}_i(t)) = \frac{D\mu_n^{\top}(\mathbf{x}_i(t))\hat{\mathbf{e}}(t)}{\|D\mu_n^{\top}(\mathbf{x}_i(t))\hat{\mathbf{e}}(t)\|}, \qquad i = 1, ..., N.$

(3) Clustering the normal vectors using K-means
While ($\phi_{ij}(t)$ has not converged)
(a) Set $\phi_{ij}(t) = \begin{cases} 1, & \text{if} j = \arg\max_{k=1,...,n}(\hat{\mathbf{b}}_k^{\top}(t)\hat{\mathbf{b}}_{vec}(\mathbf{x}_i(t)))^2 \\ 0, & \text{otherwise} \end{cases}, i = 1, ..., N, j = 1, ..., n$
(b) Set $\hat{\mathbf{b}}_j(t) = PCA([\phi_{1j}(t)\hat{\mathbf{b}}_{vec}(\mathbf{x}_1(t))...\phi_{Nj}(t)\hat{\mathbf{b}}_{vec}(\mathbf{x}_N(t))]), j = 1, ..., n.$
End
Set $\hat{\mathbf{b}}_j(t+1) = \hat{\mathbf{b}}_j(t)$.

End

4.2. Learning scene knowledge

Dynamic Properties:

Once we obtain the clustering results, it is easy to estimate the spatial extent of each clusters. Each cluster $S_j(t)$ can be seen as a crowd flow, and we should estimate its density and velocity distribution in the region. For cluster $S_j(t)$, the density distribution at position $\mathbf{p}_i = (x_i, y_i)$ at time t is estimated as

$$\begin{aligned} \mathfrak{D}_{S_j(t)}(x_i, y_i, t) = & \sum_{(x_i^*, y_i^*) \in L_i(t)} \\ & \sum_{L_i(t) \in S_j(t)} \exp(-\|(x_i - x_i^*, y_i - y_i^*)\|^2/\eta_d), \end{aligned} \tag{10}$$

where η_d is a constant parameter. An example of density distribution is shown in Figure3-(c), and the color denotes the density value.

The velocity distribution is based on the principal component of flow orientation, and can be built as:

$$\begin{aligned}\mathfrak{V}_{S_j(t)}(\mathbf{p}_i,t) = \exp(- < (\widetilde{v_x}(\mathbf{p}_i,t), \widetilde{v_y}(\mathbf{p}_i,t)),\\ (\cos(\alpha^*_{S_j(t)}(\mathbf{p}_i,t)), \sin(\alpha^*_{S_j(t)}(\mathbf{p}_i,t))) > /\eta_v),\end{aligned} \tag{11}$$

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 $\alpha^*_{S_j(t)}(\mathbf{p}_i,t)$ is the principal component in the distribution of flow orientation:

$$\mathfrak{a}_{S_j(t)}(\mathbf{p}_i,t) = \sum_{m=1}^{M} \pi_m N(\alpha_{S_p(t)}(\mathbf{p}_i,t); \mu, \sigma), \tag{12}$$

where Eq.(12) is the GMM model, and the parameters π_m can be obtained through EM iteration. An example of Eq.(11) is shown in Figure 3-(d), where the color denotes the speed, and the arrows display the principal orientation.

The density and velocity distributions for each crowd flow reflect the dynamic properties of the scene. They can provide a motion prior to the targets that are in a particular crowd flow.

Static properties (semantic words):

For a particular scene, there are some constant properties, such as dominant paths, exits, and entrances. We should output these semantic words. They can easily be obtained from the global density distribution (as shown in Figure 3-(e). The global density distribution is similar to the crowd flow density, but instead reflects properties of the whole scene; it can be computed as:

$$\begin{aligned}\mathfrak{D}_{global}(x_i,y_i,t) = \sum_{(x_i^*,y_i^*)\in L_i(t)}\\ \sum_{L_i(t)\in\Omega} \exp(-\|(x_i-x_i^*, y_i-y_i^*)\|^2/\eta_d),\end{aligned} \tag{13}$$

where Ω is the set of trajectories we have obtained at time t. As shown in Figure 3-(e)(f), after a long time period, the dominant paths of the scene were easily extracted by thresholding the global density distribution.

Additionally, the exits and entrances to the scene are two very interesting properties, known as sinks and sources, respectively. Such scene knowledge can powerfully assist the tracking to deal with the appearance or disappearance of targets. The sinks/sources can be easily detected from the global density distribution $\mathfrak{D}_{global}$. As shown in Figure 3-(f), the sinks/sources usually occur at regions of great change in the global density distribution after thresholding. Moreover, the changed direction must follow the principal orientation of the crowd flow. Hence, the sinks/sources can be easily found by a gradient search along the principal orientation of each crowd flow, and the results obtained by this method are shown in Figure 3-(f).

In summary, with the help of tracking, we are able to learn the following information online: (1) Dynamic properties: density distribution $\mathfrak{D}$ and velocity distribution $\mathfrak{V}$ of crowd flows. (2) Static properties (semantic words): dominant paths, sinks/sources. Such knowledge is very helpful for high-level activity analysis and low-level tracking or classification. In the next section, we will utilize these informations to dynamically supervise and improve the tracking results.

5. TRACKING BY SEMANTIC SCENE LEARNING

Knowledge of a scene structure can be of great help to the tracking. Firstly, a person in a particular crowd flow will be greatly influenced by it, because he must follow other persons in it. The density and velocity distributions can be used to describe this influence and supervise independent

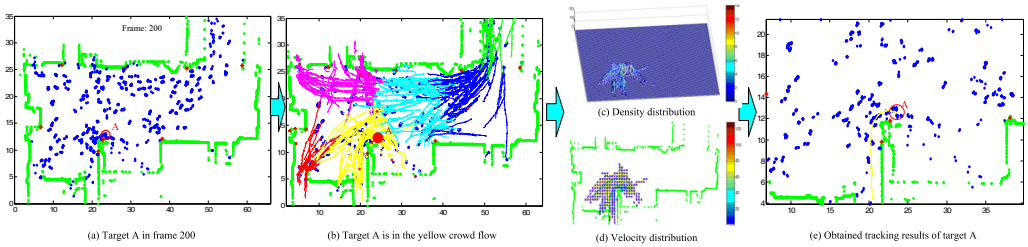


Fig. 5. Tracking by semantic scene learning. We wish to track target A in frame 200 (a). We detect that A is in the yellow crowd flow (b). The density distribution (c) and velocity distribution (d) of this crowd flow are then computed. The two distributions give us prior knowledge about the motion information of A. Hence, we can easily obtain tracking results for A with their help (e).

tracking. Secondly, a birth/death probability, which depends on the gradient of the density distribution, is assigned to the targets or measurements. This probability can help the tracking deal with the appearance/disappearance of targets, allowing us to maintain correct tracking even though frequent uncertain measurements take place.

5.1. Tracking based on crowd flow

Consider the state $\mathbf{x}_{i,t} = (x_{i,t}, y_{i,t}, v^x_{i,t}, v^y_{i,t})$ of person i at time t with its measurement $\mathbf{z}_{i,t} = \{(x^*_{j,t}, y^*_{j,t}), j = 1, ..., P\}$ which is the set of foreground laser points set after Mean-shift clustering [Comaniciu and Meer 1999], we estimate its state as

$$\hat{\mathbf{x}}_{i,t} = \arg\max_{\mathbf{x}_{i,t}} p(\mathbf{x}_{i,t}|\mathbf{z}_{i,t}). \tag{14}$$

The posterior probability $p(\mathbf{x}_{i,t}|\mathbf{z}_{i,t})$ can be computed by a Bayesian recursion as

$$p(\mathbf{x}_{i,t}|\mathbf{z}_{i,t}) = \gamma p(\mathbf{z}_{i,t}|\mathbf{x}_{i,t}) \int p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) p(\mathbf{x}_{i,t-1}|\mathbf{z}_{i,t-1}) d\mathbf{x}_{t-1}, \tag{15}$$

where γ is the normalization constant, $p(\mathbf{z}_{i,t}|\mathbf{x}_{i,t})$ is the similarity between the target's state and their measurement (observation model), and $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ is the transition probability.

Obviously, the crowd flow will have a great influence on the people who are in it. We can use the density and velocity distribution to describe this influence, and the transition probability of person i in the crowd flow $S_j(t)$ can be computed as

$$p_{S_j(t)}(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) = \mathfrak{D}_{S_j(t)} \mathfrak{V}_{S_j(t)} \mathfrak{W}, \tag{16}$$

where $\mathfrak{W}$ is the walking model, which can be a constant velocity model, a second order autoregressive model [Song et al. 2008] or the "two feet model" [Zhao and R.Shibasaki 2005]. In this research, we utilize a simple second-order autoregressive walking model as follows:

$$\mathbf{x}_{i,t} = \mathbb{A}\mathbf{x}_{i,t-1} + \mathbb{B}\mathbf{x}_{i,t-2} + \mathbb{C}N(0, \Sigma), \tag{17}$$

where matrices $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$ are of constant ratio, and are obtained by regression with 150 representative sequences. N is the normal distributed random noise. The observation model $p(\mathbf{z}_{i,t}|\mathbf{x}_{i,t})$ in Eq. (15) is the similarity between people's positions and their nearest cluster of laser points:

$$p(\mathbf{z}_{i,t}|\mathbf{x}_{i,t}) = \frac{1}{\sqrt{2\pi}\delta_p} \times \exp \sum_{j=1}^{P} (-\frac{(x_{i,t} - x^*_{j,t})^2 + (y_{i,t} - y^*_{j,t})^2}{2\delta_p^2}), \tag{18}$$

where δ_p is the constant ratio.

Eq. (16) can be understood very easily. As shown in Figure 5, if an immediate crowd behavior moves in a particular direction, the individuals in it will favor this direction with a high transition

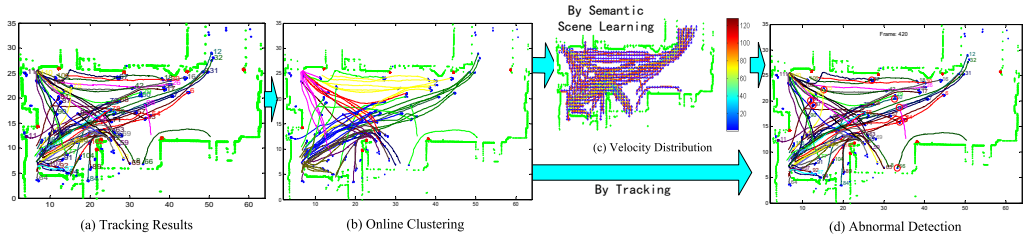


Fig. 6. Online abnormal activity detection. Once accurate tracking results and scene knowledge have been obtained, they can be used to perform both global and local online abnormality detection. (1) Global abnormality detection is based on tracking. Once we obtain the tracking results (a), these trajectories are grouped based on different activity types via an online unsupervised learning, as discussed in Section 4.1 (b). Abnormal activities can be thought of as outliers of the clustering algorithm (such as the green lines shown in (b)). (2) Local abnormality detection can be detected via the learned semantic scene (velocity distribution) in each inter-cluster. As shown in (d), some trajectories (such as person 54) with a large bias from their velocity distribution (c) in the inter-clusters are detected as abnormal. Note that the arrows in (c) show the principal orientation at that specific position, and the color denotes its value.

probability, and the motion transition will also follow the density and velocity distributions of this crowd flow.

Hence, we can utilize a particle filter (PF) [Doucet et al. 2000] to compute Eqs. (14) and (15) and obtain the targets' state at each time.

5.2. Uncertain measurements

Most failed tracking in our application was actually caused by uncertain measurements. For instance, a new target is often incorrectly initialized due to a false alarm. In addition, merge/split measurements and non-detections due to occlusion often result in breaks in trajectories, because the targets cannot be matched to their measurements. We can easily ameliorate these problems with the help of source/sink knowledge in a scene.

We assigned a death probability P^i_{death} to each target and a birth probability P^i_{birth} to each measurement. If a target with a high death probability cannot find any matching measurement, it can be considered a disappearing target in the scene. Otherwise, it should still be tracked whether or not it has a suitable matching measurement. Similarly, if a measurement with a high birth probability does not associate with any target, we initialize and track it as a new appearing target. Otherwise, it is considered a false alarm. As shown in Figure 3-(c), the birth and death probabilities depend on the gradient of the density distribution, and the direction of gradient descent/ascent must follow the principal direction of the crowd flow. Hence, the two probabilities can be computed as

$$P^i_{birth} \propto \exp(< \nabla \mathfrak{D}_{S_j(t)}(x, y, t), \vec{v}_j/|\vec{v}_j| > /\xi_1), \tag{19}$$

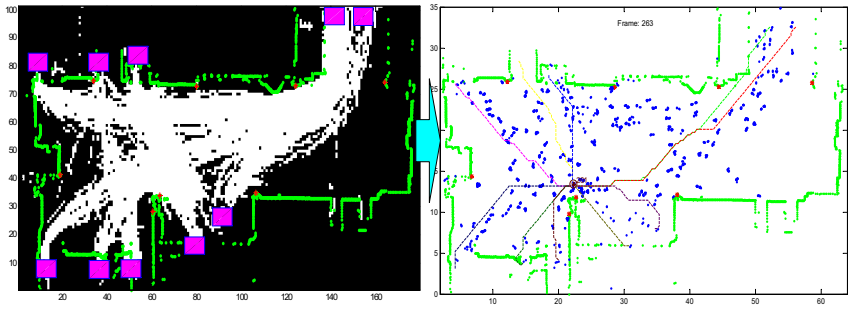
$$P^i_{death} \propto \exp(< \nabla \mathfrak{D}_{S_j(t)}(x, y, t), \vec{v}_j/|\vec{v}_j| > /\xi_2), \tag{20}$$

where $\vec{v}_j$ is the principal direction of the crowd flow, and ξ_1, ξ_2 are the constant parameters.

In summary, the tracking benefits from the learned scene knowledge and also provides new results to update this knowledge. Therefore, this mode of co-operation between tracking and learning not only obtains accurate tracking results and scene knowledge, but also ensures that the entire process is completely automatic and online.

6. ONLINE ABNORMALITY DETECTION BY TRACKING AND SEMANTIC SCENE LEARNING

As illustrated in Figure 2, once accurate tracking results and scene knowledge are obtained, they can be used to perform both global and local online abnormality detection. Global abnormality detection can be addressed by tracking, whereas local detection is performed using the learned scene knowledge (dynamic properties). Furthermore, we also consider the learned global scene structure to improve the abnormality detection model in further.



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

Fig. 7. Abnormality detection by using learned scene structure. Given the online learned scene structure (a), we can find possible planned paths of person 201 with A* algorithm (b), and a normal person would like to make its motion be more like to its planned path.

6.1. Global abnormality detection

We utilize the algorithm discussed in Section 4.1 to cluster the obtained tracking results online based on different activity types. The first type of abnormal activity (global detection) can be detected from outlier trajectories of the online clustering. If some clusters contain very few trajectories, their owners are considered to be performing some abnormal activity. An example is shown in Figure 6: at a specific time in a subway station, a large number of people get off a train. Most of them walk along the common path of this subway station, but, in contrast, some people (such as persons 66 and 98) walk along an uncommon path. Their trajectories are difficult to group into any cluster, and these activities are detected by our system as abnormal activity.

6.2. Local abnormality detection

Another type of abnormal activity (local detection) can be detected via the learned semantic scene (velocity distribution) in each inter-cluster. We define the velocity abnormal energy $E_{velocity}(\mathbf{p}_{i,t}, \mathbf{v}_{i,t})$ of each person to measure its activity in the inter-cluster, where a higher energy $E_{velocity}$ is more likely to be an abnormal activity. Hence, given the velocity distribution $\mathfrak{V}_{S_j(t)}$ in cluster $S_j(t)$ computed by Eq. (11), for person i in position $\mathbf{p}_{i,t}$ with velocity $\mathbf{v}_{i,t}$, this energy can be computed by

$$E_{velocity}(\mathbf{v}_{i,t}, \mathbf{p}_{i,t}) = \exp(-||\mathbf{v}_{i,t} - \mathfrak{V}_{S_p(t)}(\mathbf{p}_{i,t})||^2/2\sigma_1^2), \tag{21}$$

where σ_1 is a constant parameter. Once this energy is larger than the threshold, their activities are detected as abnormal.

This can be easily understood. As shown in Figure 6, at a specific time, many people were walking together and going to the same destination, which would become a crowd flow. Usually, people in this crowd flow should follow one another. But in contrast, some individuals (such as person 54) performed quite different motions from others in the crowd, and this activity could be detected as abnormal.

6.3. Abnormality detection by learned scene structure

Besides, while the normal people are walking in the large and crowded environment, they usually plan to go to a specific exit of the scene, walk on the common road, avoid the obstacles and finds the shortest and comfortable path (as shown in Figure 7). Thus, the learned scene structure (e.g., dominant paths, exits, and entrances) can also help us to detect some abnormal activities.

We assign the scene abnormal energy E_{scene} for each pedestrian to measure its activity, where a higher energy E_{scene} is more likely to be an abnormal one. Given the current position $\mathbf{p}_{i,t}$ of pedestrian i by tracking, online learned scene structure map and Q exits/entrances (as shown in

Figure 7-(a)), it is easy for us to obtain Q planned trajectories $\{L^l_{i,t}(x,y)\}^Q_{l=1}$ for pedestrian i at time t with *A Star search* algorithm (as shown in Figure 7-(b)). Hence, a normal person would like to make its motion be more like to its planned path, and the E_{scene} can be computed by:

$$E_{scene}(\mathbf{v}_{i,t}, \mathbf{p}_{i,t}) = \sum_{l=1}^{Q} w_l \times \exp(-||\frac{\mathbf{v}_{i,t}}{||\mathbf{v}_{i,t}||} - \frac{\partial L^l_{i,t}(\mathbf{p}_{i,t})}{\partial x \partial y}||^2 / 2\sigma_2^2), \tag{22}$$

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_2 is a constant parameter and w_l is the weight of the possible planned trajectories, which is depend on the similarity between person's current trajectory and this planned one. In this research, we utilize the approach of Wang *et al.* [Wang et al. 2006] to measure the similarity of two trajectories, please refer it for more details. Once the weights of planned trajectories are very small, we throw them and stop making new path planning for these exits/entrances. Obviously, if this energy is larger than the threshold, their activities are detected as abnormal.

7. EXPERIMENTS AND RESULTS

We applied our system to a real scene: the lobby of JR subway station, Tokyo (an area of about 60×35m). Eight single-row laser scanners (LMS291) produced by SICK were utilized. They were set 10 cm above the ground surface and performed horizontal scanning at a frequency of 37 fps. We utilized a time server to deal with the problem of time synchronization between different sensors, and the calibration was conducted by several control points in a box. For more details about the experimental setting, please refer to [Zhao and R.Shibasaki 2005]. Selected data from between 7:00 am and 8:30 am, which is quite a busy time in Tokyo, were used for the evaluation, and this data contains five clips. In this section, we will present our experimental results and perform a quantitative evaluation and comparison.

7.1. Tracking results and learned semantic scene

Figure 8 shows some selected tracking results of our system. The first row contains tracking results, the second contains clustering results, and the third and fourth show the incrementally learned density and velocity distribution maps. From this figure, we can see that people were clearly clustered based on the different crowd flows. In addition, the density distribution map became increasingly clear so as to reveal knowledge of the scene. Furthermore, from the velocity distribution, we can see that people in a crowded place usually walk quite slowly.

As the 1090 frames proceeded, dominant paths and sinks/sources of the scene were obtained, as illustrated in Figure 13. Actually, with an increase in the number of tracking frames, these results can become more accurate.

In some cases, some failed tracking of our system was caused by the highly unreliable measurements (e.g. more than six people were walking together, and the measurements were completely merged.). In such circumstances, a stronger motion model (e.g. pedestrians' walking or movement model) for prediction usually plays a more important role in the overall tracking process. In the future, we will try to enhance the motion model, and consider the social interactions among walking pedestrians to deal with this problem.

7.2. Abnormality detection results

Selected abnormality detection results (without using scene structure information) from our system are shown in Figure 9. The first row shows the tracking results, the second shows the online clustering results, the third contains the incrementally learned motion distribution, and the fourth row shows the abnormality detection results. From this figure, we can see that some suspicious people

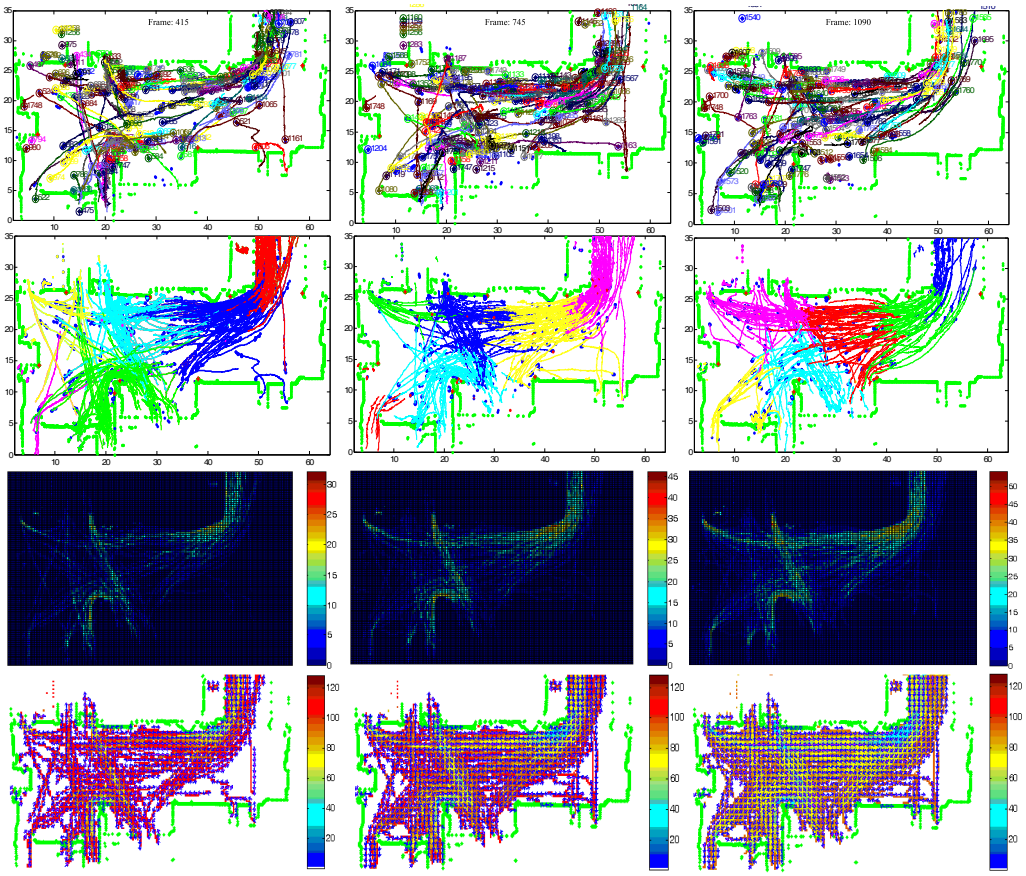


Fig. 8. Tracking results from the proposed system. The first row shows the tracking results, the second row shows the clustering results, and the third and fourth rows show the incrementally learned density and velocity distributions. Please see our supplementary video for more details.

could be easily detected by our system, such as person 11 and person 111 in frame 600 (what were they doing?), person 67 in frame 762 (walking to an uncommon exit), and persons 263 and 277 in frame 2005 (appearing at an uncommon entrance).

Besides, we also present the abnormality detection results with scene structure information as discussed in Section 6.3, and they are shown in Figure 10. The first row shows the abnormal energy, where the color shows abnormal energy value, the darker the greater value (as shown in the color bar). The abnormality detection results are shown in the second row. From this figure, we can see that some suspicious persons also could be easily detected, such as person 131, 117, and 60 in frame 50(walking on the closed road), person 24 in frame 111 (what was it doing?), person 9 in frame 151 (it was not following other persons and walking in a strange path) and etc.

7.3. Quantitative comparison of tracking

In order to evaluate the tracking performance of the proposed system, a quantitative comparison was conducted between four methods: those of Song *et al.* [Song et al. 2008] and Cui *et al.* [Cui et al. 2006], a PF-based tracker (no scene learning), and our method.

We made a statistical survey of 3000 continuous frames to evaluate the performance of these methods in a high-density scene. The ground truth was obtained in a semi-automatic manner (track-

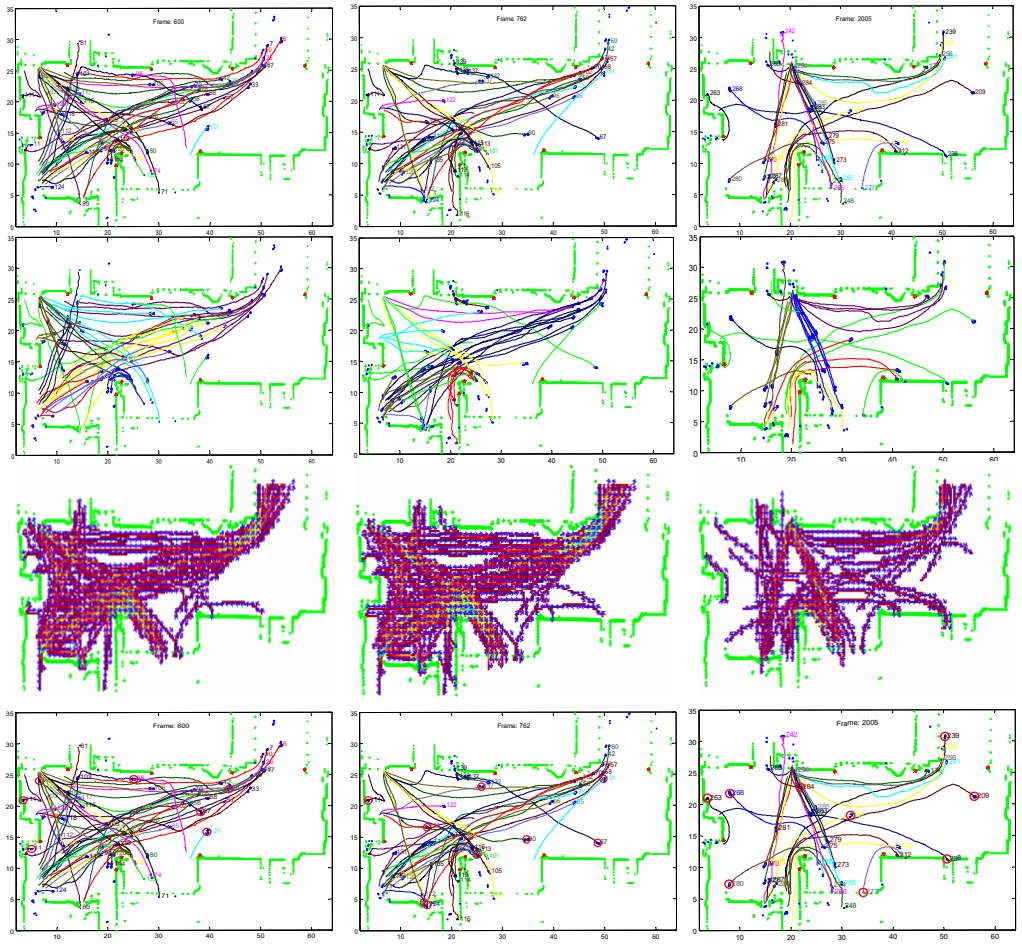


Fig. 9. Abnormality detection results without using learned scene structure. The first row shows the tracking results, the second row shows the online clustering results, the third row shows the incrementally learned velocity distribution, and the fourth shows the abnormality detection results. Some suspicious people were easily detected by our system: persons 11 and 111 in frame 600, person 67 in frame 762, and persons 263 and 277 in frame 2005. For more results from our system, please see our supplementary video.

ers and manual labeling). Tracking failures included those where the target was missed, a false location was given, or the targets' identity was switched, which can be automatically computed from the ground truth. Details of this comparison are illustrated in Figure 11, and the overall success rate of each method is listed in Table 1. From Figure 11, we can see that our method exhibits the best performance of the four in high-density scenes, with the scene knowledge providing a performance improvement of about 16%. As illustrated by the tracking results in frame 990, the trajectories obtained by the trackers with no scene learning were quite short and frequently broken. By contrast, with the help of the scene knowledge, our method can easily maintain long duration, robust tracking.

7.4. Quantitative comparison of abnormality detection

In order to perform a quantitative evaluation of the abnormality detection, we would ordinarily need the ground truth. Obviously, it is hard to say whether a person's activity is abnormal. Thus, we

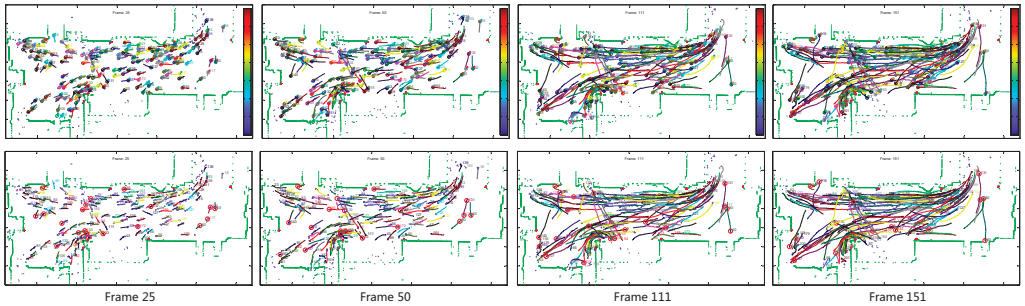


Fig. 10. Abnormality detection results using learned scene structure. The first row is the abnormal energy, where the color shows abnormal energy value, the darker the greater value (as shown in the color bar). The abnormality detection results are shown in the second row. Some suspicious persons could be easily detected by our system, please note that person 131, 117, and 60 in frame 50, person 24 in frame 111 and person 9 in frame 151. For more results about them, please see our supplementary video.

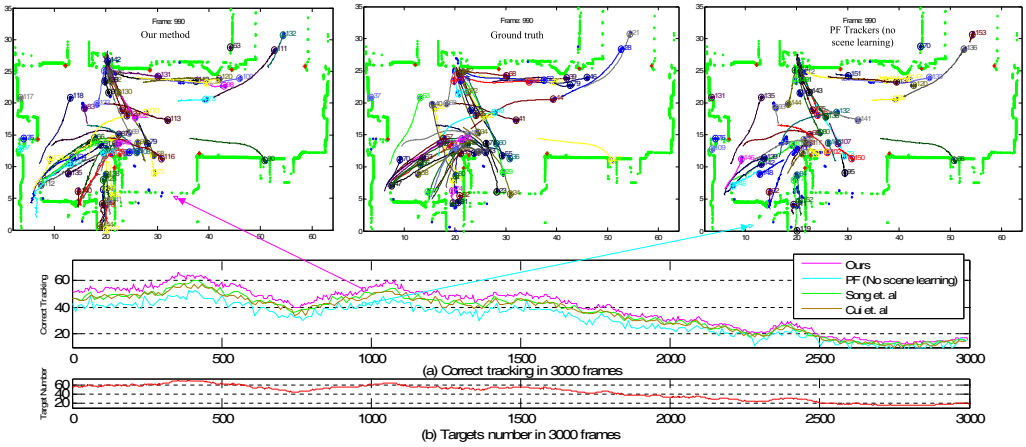


Fig. 11. Quantitative comparison between four methods showing (a) the correct tracking of the four methods over 3000 continuous frames, and (b) the target numbers in these frames.

Table I. Success rate among the four tracking methods

Algorithm	Success Rate
Song *et. al*	85.5%
Cui *et. al*	83.2%
PF only	75.2%
Our Method	**91.3%**

invited three people to label the activity in our data as abnormal or not. One of them has an academic background, and the other two do not have any academic background in this field. At a specific time, if the invited people think the activity of a person is abnormal, based on the trajectories, they will label that person's activity state as 1, and as 0 otherwise. The final ground truth was based on the union of the three people's opinions with its confidence.

We tested our system over 3000 frames, and the ROC curve of our system is shown in Figure 12. In addition, a quantitative comparison was also conducted between our method (using scene structure information and without using it) and some batch-based clustering methods, such as Fuzzy

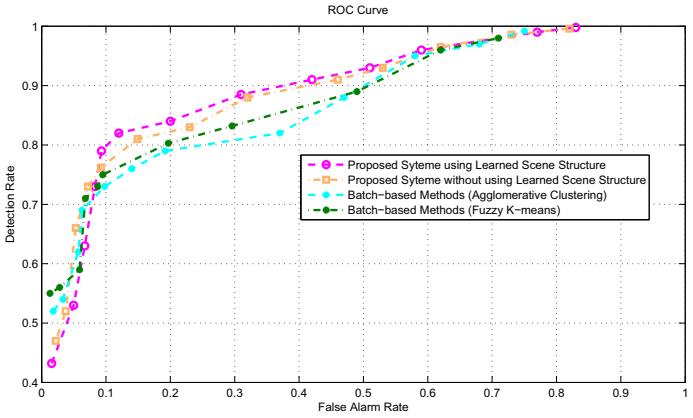


Fig. 12. Quantitative comparison with batch-based methods. This figure shows the ROC curves of four methods. From this figure, we can see that the proposed system exhibits better performance than the batch-based methods. This is because our system varies with changes in the situation and detects sudden abnormal activities in some specific frames, which is difficult for the batch-based methods. In addition, with the help of global scene structure information, the abnormality detection results are improved in further.

K-means and Agglomerative Clustering. For the two competing algorithms, abnormal activity was detected by finding the outliers of the clusters. The details of this comparison are also shown in Figure 12. From this figure, we can see that the proposed system performs better than the batch-based methods. This is because our system can vary with a change in the situation and detect any sudden abnormal activities in specific frames, which is difficult for the batch-based methods. Besides, with the help of global scene structure information, the abnormality detection results are improved in further.

8. CONCLUSION

In this paper, we presented a novel online system that can simultaneously perform semantic scene learning, tracking, and abnormality detection in a large, high-density area. The experimental results demonstrated our method's feasibility and robustness. In the future, this work can be extended and improved in the following aspects: (1) Some failed tracking of our system is usually caused by the highly unreliable measurements. In such circumstances, a stronger motion model usually plays a more important role in the overall tracking process. Thus, we will try to enhance the motion model, and consider the social interactions among walking pedestrians to deal with this problem. (2) We found that, with an increasing number of clusters, the computation became very large. Hence, decreasing the computational cost and optimizing our system is an important problem. (3) With the laser scanner data, it is difficult to give a single, clear definition of abnormal activity, especially for an unsupervised method. Therefore, a method to scientifically evaluate the detected results is another important issue. (4) The proposed system can easily be extended to intersections for traffic surveillance. In this case, we will encounter different objects, such as pedestrians, bicycles, and cars. Therefore, an object classification module can be added into our system. Obviously, this module can also benefit from tracking and semantic scene learning.

9. 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

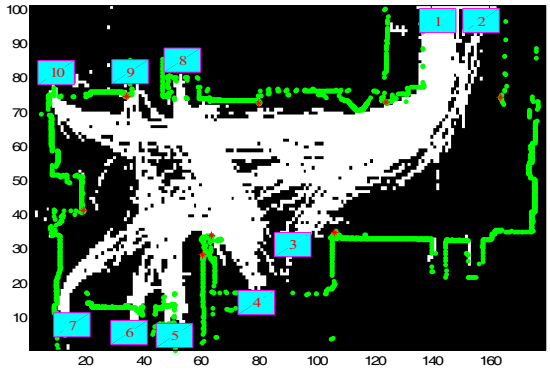


Fig. 13. Paths and sinks/sources obtained after 1090 frames. The white color shows the paths in this scene, and the rectangles show the sinks/sources.

(No.60975061) of China, East Japan Railway Company and Microsoft Research. We thank all anonymous reviewers for their helpful comments on this work.

REFERENCES

ADAM, A., RIVLIN, E., SHIMSHONI, I., AND REINITZ, D. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. on Pattern Analysis and Machine Intelligence 30,* 3, 555–560.

ALI, S. AND SHAH, M. 2008. Floor fields for tracking in high density crowd scenes. *Proc. European Conference on Computer Vision*, 1–14.

AVIDAN, S. 2007. Ensemble tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence 29,* 2, 261–271.

BABENKO, B., YANG, M., AND BELONGIE, S. 2009. Visual tracking with online multiple instance learning. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 983–990.

BAR-SHALOM, Y. AND FORTMANN, T. E. 1998. Tracking and data association. *New York: Academic Press*.

BETKE, M., HIRSH, D., BAGCHI, A., HRISTOV, N., AND MAKRIS, N. 2007. Tracking large variable numbers of objects in clutter. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1180–1187.

BOSE, B., WANG, X., AND GRIMSON, E. 2007. Multi-class object tracking algorithm that handles fragmentation and grouping. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1550–1557.

CHANDOLA, V., BANERJEE, A., AND KUMAR, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys 41,* 15, 1–72.

COMANICIU, D. AND MEER, P. 1999. Distribution free decomposition of multivariate data. *Pattern Analysis and Applications 38*, 22–30.

CUI, J., ZHA, H., ZHAO, H., AND R.SHIBASAKI. 2006. Fusion of detection and matching based approaches for laser based multiple people tracking. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 642–649.

DOUCET, A., GODSILL, S. J., AND ANDRIEU, C. 2000. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing 10,* 1, 197–208.

FU, Z., HU, W., AND TAN, T. 2005. Similarity based vehicle trajectory clustering and anomaly detection. *Proc. IEEE International Conference on Image Processing*, 1133–1136.

GENNARI, G. AND HAGER, G. 2004. Probabilistic data association methods in visual tracking of groups. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 876–881.

HARRIS, J. 1992. Algebraic geometry: A first course. *Springer-Verlag*.

HU, W., XIE, D., FU, Z., ZENG, W., AND MAYBANK, S. 2007. Semantic-based surveillance video retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence 16,* 4, 1168–1181.

JIANG, H., FELS, S., AND LITTLE, J. 2007. A linear programming approach for multiple object tracking. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1380–1387.

J.SULLIVAN AND S.CARLSSON. 2006. Tracking and labeling of interacting multiple targets. *Proc. European Conference on Computer Vision*, 661–675.

JUNEJO, I. AND FOROOSH, H. 2007. Trajectory rectification and path modeling for video surveillance. *Proc. IEEE International Conference on Computer Vision*, 230–237.

JUNEJO, I., JAVED, O., AND SHAH, M. 2004. Multi feature path modeling for video surveillance. *Proc. IEEE Conference on Pattern Recognition*, 383– 386.

KHAN, Z., BALCH, T., AND DELLAERT, F. 2006. 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, 1960–1972.

LANZ, O. AND MANDUCHI, R. 2005. Hybrid joint-separable multibody tracking. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 413–420.

LEIBE, B., SCHINDLER, K., AND GOOL, L. 2007. Coupled detection and trajectory estimation for multi-object tracking. *Proc. IEEE International Conference on Computer Vision*, 1110–1117.

LOY, C. C., XIANG, T., AND GONG, S. 2010. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision 90,* 1, 106–129.

MAHADEVAN, V., LI, W., BHALODIA, V., AND VASCONCELOS, N. 2010. Abnomaly detection in crowded scenes. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1975–1981.

MAKRIS, D. AND ELLIS, T. 2003. Automatic learning of an activitybased semantic scene model. *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance*, 183– 188.

MEHRAN, R., OYAMA, A., AND SHAH., M. 2009. Abnormal crowd behavior detection using social force model. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 935–942.

PATINO, L., BREMOND, F., EVANS, M., SHAHROKNI, A., AND FERRYMAN, J. 2010. Video activity extraction and reporting with incremental unsupervised learning. *Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 54–61.

P.NILLIUS, J.SULLIVAN, AND S.CARLSSON. 2006. Multi-target tracking - linking identities using bayesian network inference. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2187–2194.

PUSIOL, G., BREMOND, F., AND THONNAT, M. 2010. Trajectory based activity discovery. *Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 75–82.

QU, W., SCHONFELD, D., AND MOHAMED, M. 2005. Real-time interactively distributed multi-object tracking using a magnetic-inertia potential model. *Proc. IEEE International Conference on Computer Vision*, 535–540.

RASMUSSEN, C. AND HAGER, G. 2001. Probabilistic data association methods for tracking complex visual objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence 23,* 1, 560–576.

READ, D. 1979. An algorithm for tracking multiple targets. *IEEE Trans. Automation and Control 24,* 1, 84–90.

ROSS, D., LIM, J., LIN, R., AND YANG, M. Incremental learning for robust visual tracking. *International Journal of Computer Vision 77,* 1.

SALEEMI, I., SHAFIQUE, K., AND SHAH, M. 2009. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Trans. on Pattern Analysis and Machine Intelligence 31,* 8, 1472–1485.

SCHULZ, D., BURGARD, W., FOX, D., AND CREMERS, A. 2003. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research 22,* 2, 99–116.

SONG, X., CUI, J., WANG, X., ZHAO, H., AND ZHA, H. 2008. Tracking interacting targets with laser scanner via on-line supervised learning. *Proc. IEEE International Conference on Robotics and Automation*, 2271–2276.

SONG, X., CUI, J., ZHA, H., AND ZHAO, H. 2008. Vision-based multiple interacting targets tracking via on-line supervised learning. *Proc. European Conference on Computer Vision*, 642–655.

SONG, X., SHAO, X., ZHAO, H., CUI, J., SHIBASAKI, R., AND ZHA, H. 2010. An online approach: Learning-semantic-scene-by-tracking and tracking-by-learning-semantic-scene. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 739–746.

VERMAAK, J., GODSILL, S., AND PEREZ, P. 2005. Monte carlo filtering for multi target tracking and data association. *IEEE Trans. Aerospace and Electronic Systems 41,* 1, 309–332.

VIDAL, R. 2006. Online clustering of moving hyperplanes. *Proc. Neural Information Processing Systems*, 1433–1440.

WANG, X., MA, K., NG, G., AND GRIMSON, E. 2008. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1–8.

WANG, X., TIEU, K., AND GRIMSON, E. 2006. Learning semantic scene models by trajectory analysis. *Proc. European Conference on Computer Vision*, 110–123.

WANG, X., TIEU, K., AND GRIMSON, W. E. L. 2010. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Trans. on Pattern Analysis and Machine Intelligence 32,* 1, 56–71.

WU, S., MOORE, B. E., AND SHAH, M. 2010. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2054–2060.

XIANG, T. AND GONG, S. 2008. Video behavior profiling for anomaly detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence 30,* 5, 893–908.

YANG, M., YU, T., AND WU, Y. 2007. Game-theoretic multiple target tracking. *Proc. IEEE International Conference on Computer Vision*, 110–117.

YILMAZ, A., JAVED, O., AND SHAH, M. 2006. Object tracking: A survey. *ACM Computing Surveys*, 3–47.

YU, Q., MEDIONI, G., AND COHEN, I. 2007. Multiple target tracking using spatio-temporal markov chain monte carlo data association. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 642–649.

ZHANG, T., LU, H., AND LI, S. Z. 2009. Learning semantic scene models by object classification and trajectory clustering. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 1940–1947.

ZHAO, H. AND R.SHIBASAKI. 2005. A novel system for tracking pedestrians using multiple single-row laser range scanners. *IEEE Transactions on Systems, Man and Cybernetics, part A*, 283–291.

ZHAO, T. AND NEVATIA, R. 2004. Tracking multiple humans in complex situations. *IEEE Trans. on Pattern Analysis and Machine Intelligence 7,* 1, 1208–1221.