# Social Attribute-Aware Force Model: Exploiting Richness of Interaction for Abnormal Crowd Detection

Yanhao Zhang, Lei Qin, *Member, IEEE*, Rongrong Ji, *Senior Member, IEEE*, Hongxun Yao, *Member, IEEE*, and Qingming Huang, *Senior Member, IEEE*

***Abstract*— Interactions among pedestrians usually play an important role in understanding crowd behavior. However, there are great challenges, such as occlusions, motion, and appearance variance, on accurate analysis of pedestrian interactions. In this paper, we introduce a novel social attribute-aware force model (SAFM) for detection of abnormal crowd events. The proposed model incorporates social characteristics of crowd behaviors to improve the description of interactive behaviors. To this end, we first efficiently estimate the scene scale in an unsupervised manner. Then, we introduce the concepts of social disorder and congestion attributes to characterize the interaction of social behaviors, and construct our crowd interaction model on the basis of social force by an online fusion strategy. These attributes encode social interaction characteristics and offer robustness against motion pattern variance. Abnormal event detection is finally performed based on the proposed SAFM. In addition, the attribute-aware interaction force indicates the possible locations of anomalous interactions. We validate our method on the publicly available data sets for abnormal detection, and the experimental results show promising performance compared with alternative and state-of-the-art methods.**

***Index Terms*— Abnormal detection, crowd behaviors, social attributes, social force model.**

## I. Introduction

ANALYZING crowd behavior has become a salient research topic in video surveillance and beyond. In contrast to individual actions, crowd behaviors are far more challenging to analyze, because of the possibility for complex interactions among individuals.

Crowd behavior models aim to describe individuals and groups in crowded scenes. From a sociological viewpoint, crowd behaviors usually occur under the constraints of the sociologically inspired prior knowledge, and hence reflect high-level semantic interactions. With the underlying motion characteristics, midlevel visual representation has become increasingly popular, as it can extend the semantic description of visual content from feature level to object level, focusing on the specific connections and interactions among the objects. Therefore, constructing midlevel representations of crowds for exploiting the richness of interactions can lead to breakthroughs in a wide range of applications.

To achieve this goal of automatically identifying the interaction of crowd behaviors, one emerging and challenging task here is the detection of abnormal crowd behaviors. Under such a circumstance, the crowd behavior is usually modeled as a quantitative result of semantic representation, which is the basis of analyzing and understanding the abnormality. Given a video clip, the abnormal detection models motion consistency among individuals, labeling the ones that are significantly inconsistent with others as abnormal. To this end, extensive work has been proposed toward accurate and robust abnormality discovery, ranging from motion feature representation to inconsistency model definition, such as mixture of probabilistic principal component analyzers (MPPCA) [1], low-level statistics [2], dynamic texture [3], Markov random field (MRF) [4], and sparse representation [5]. Rather than classifying the overall inconsistency, [3] and [5]–[7] also focus on detecting and locating the local inconsistency or the selective mechanism of the crowded scene.

### A. Issues

Detecting abnormal crowd behaviors, by analyzing low-level semantics of crowds, is still far from real-world applications. Despite the complex models extensively studied in the literatures, key issues remain in the implementation of robust yet accurate feature representation specified for crowd behaviors, because of the following reasons.

1) The traditional motion features in the existing works [2], [8], such as optical flow, space-time interest point [9], as well as spatial–temporal volume [8], are

Manuscript received October 3, 2013; revised March 10, 2014, May 28, 2014, and July 12, 2014; accepted September 1, 2014. Date of publication September 8, 2014; date of current version June 30, 2015. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400 and in part by the National Natural Science Foundation of China under Grants 61025011, 61133003, 61332016, and 61035001. This paper was recommended by Associate Editor N. J. Sarhan.

Y. Zhang and H. Yao are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: yhzhang@hit.edu.cn; h.yao@hit.edu.cn).

L. Qin is with Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Institute of Computing Technology, Beijing 100089, China (e-mail: qinlei@ict.ac.cn).

R. Ji is with the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Xiamen 361000, China (e-mail: rrji@xmu.edu.cn).

Q. Huang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2014.2355711

1051-8215 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

incapable of characterizing motion inconsistency among individuals. For instance, basic representation elements are supposed to have the same status (such as scale and motion velocity) without considering dynamic changes and perspective effects of the scene.

2) Although a few recent works [10]–[12] attempt to model crowd interactions, such modeling still relies on the low-level visual and/or motion statistics among individuals, without regard to their higher level semantics and structures of crowds, such as whether the group behavior is actually inconsistent from the social or behavior point of view. In other words, the social and community semantics from both microscopic and macroscopic support [12] are missing in these works, which in turn are of high importance toward precise abnormity definition.
3) Although there are some recent works leveraging social interaction theory [13], the modeling of social behaviors is still far from satisfactory. The expressions of interactions among individuals lack intermediate features, which ignores the social characteristics of the crowd working on the individuals. In real-world scenarios, traditional approaches cannot efficiently give realistic interactive clues for complex scenes. The principle to overcome such issues may resort to boosting the feature representation from low-level to middle-level, for example, attributes or concepts [14]–[16], as recently emerging trend in object and event detection [17], [18] in the community.

In this paper, we investigate the feasibility of modeling the social interactions among individuals for abnormal crowd behavior detection. Our innovation depends on an attribute-level modeling for social interaction behaviors, by which two groups of social attributes are proposed: 1) social disorder and 2) congestion. Together, they form a novel social attribute-aware force model (SAFM) to exploit the richness of interactions. Such a model can accurately and robustly express complex interactions among individuals. Given such attributes at hand, abnormality detection can be achieved at either a global or local scale. In both cases, the proposed attributes can be integrated with most existing abnormal detection frameworks to achieve state-of-the-art performance, as we have extensively tested in our experiments.

### B. Related Work

Modeling social interaction plays an important role in describing group behavior, and contributes to abnormality modeling in crowded scenes. Social interaction modeling considers the important point that the individuals in crowded scenes move purposefully, forming interactions with others and affecting each other's trajectories. It also improves traditional dynamic models for crowd simulation, which considers that people have their future destinations and take actions to adjust their trajectories [12], [19]. There are three main perspectives for modeling and analyzing social interactions: 1) macroscopic models, which deal with the group density and velocity instead of determining the motion of individuals; 2) microscopic models, which concern the pedestrians' motivation in the movements and deal with the group density/velocity or the individual motion, for instance, the social force model proposed in [13] that investigates the motion dynamics; and 3) hybrid methods that take the above two types of models into a whole account. Both group and individual statistics are fused to analyze the crowd behaviors, in which external and self-organization factors are exploited [20].

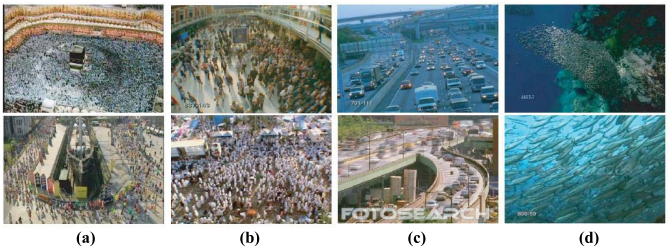



Fig. 1. External and self-organization factors of crowd behavior. The crowd behaviors are influenced by the collective and interactive effects. The collective effects are formed by individuals sharing similar moving patterns such as behaviors of organisms of school fish, flocking birds in biology (d). It usually characterizes crowd behaviors at a macroscopic level which are regularized by external environment like scene structures, obstacles as (a) and (c). The interactive effects characterize crowd behaviors at the microscopic level. The behaviors are caused by interactions among individuals within social groups which are closely relevant with crowd density (b).

As shown in Fig. 1, external organization aims to define the belief priors such as obstacles, positions, or scene structures. Self-organization refers to exploiting the collective and interactive effects, which have been widely used for understanding crowd behaviors [21], [22]. In the following, we give a brief review for modeling the social behavior from collective and interactive aspects for abnormal crowd behavior detection.

*1) Collective Effect:* Collective effect offers a macroscopic view in terms of modeling similar motion patterns in the crowd behavior, which can be treated as regular patterns influenced by scene structures and environmental conditions. For instance, many works on global motion pattern learning are capable of capturing the collective effect. Ali and Shah [23] proposed a method based on fluid dynamics to simulate crowd behaviors, which required precise segmentation of motion flow. Lin *et al.* [24] proposed extraction of robust flow patterns by using potential geometry migration process and algebraic expression model. Recent works in modeling global motion patterns use Hierarchical Bayesian models [25], [26] by mining the co-occurrences of moving pixels to learn behavior patterns. Bellomo *et al.* [27] propose a kinetic theory model to show how the dynamics at the microscale is transferred to collective behaviors. Zhou *et al.* [4], [28] use MRF as hypothesis to establish the connection model for track segments, which is then combined with topic model to cluster group behaviors.

*2) Interactive Effect:* Analyzing the interaction among individuals typically relates to the microscopic-level description on crowd behavior. It can be used to describe different interactions, social groups, or leadership for individuals. A social force model to simulate pedestrians that are forced by surrounding environment and people was proposed in [13]. This model is extended by Mehran *et al.* [10] for anomaly

detection in crowd scenes. References [12] and [29] used social behavior of individual interaction model to analyze the dynamic behaviors of pedestrians. Agent models influenced by personal, social, and environmental factors are constructed in [19], which effectively estimated pedestrians purposes and social relationships. Compared with the previous works, we encode collective effect into the local interaction of crowd behavior and integrate social attributes to reveal the richness of interactions from both the macroscopic and microscopic viewpoints.

*3) Abnormal Crowd Detection:* Detecting abnormal behaviors from the crowd is a longstanding research topic which has attracted extensive research in the past. With the application scenarios, the task of abnormal crowd behavior detection can be categorized into twofold, that is, local abnormal event (LAE) and global abnormal event (GAE) [5]. LAE aims at detecting the local behaviors which are different from the neighborhoods. For instance, Kim and Grauman [1] modeled the activity pattern with MP-PCA on the local optical flow and used MRF to localize anomalies. Adam *et al.* [2] employed histograms to measure the probability of local optical flow pattern. In [8], spatial–temporal gradients are extracted by Kratz and Nishino to fit a Gaussian model, and a coupled hidden Markov model is used to detect abnormalities in crowded event. By modeling motion patterns in local area, dense activity and intrinsic structure of crowd can be exploited. Mahadevan *et al.* [3] proposed a dynamic texture model to jointly model the appearance and dynamics of the crowded scene. In addition, detecting temporal and spatial anomalies can be formulated as detecting crowd saliency in mixtures of dynamic textures [3]. Thida *et al.* [30] employed spatiotemporal Laplacian Eigenmap by constructing a pairwise graph to detect and localize the abnormal regions considering the visual context of multiple local patches.

The goal of the GAE is to detect whether the whole scene is abnormal. Mehran *et al.* [10] adopted the social force model and particle advection scheme to detect the abnormal crowd behavior by analyzing the interaction force. With the Lagrangian framework of fluid dynamics, a streakline representation [31] is proposed to enhance the representation of abnormal event using the Helmholtz decomposition theorem. Cui *et al.* [11] explored the relationships between the states of the subjects using interaction energy potentials. The preceding approaches are successfully carried out in crowd behavior modeling for abnormal detection.

### C. Our Contributions

In this paper, we introduce a SAFM to achieve robust, efficient, and interactive crowd behavior modeling. Our first contribution is to consider the scene scale information by proposing a fast novel scene scale estimation scheme which captures the scene perspective to better infer crowd characteristics. With such a scale estimation scheme, it is then feasible to partition the foreground movements from the background, and extract scale attribute and density properties of a crowd.

Our second contribution is to revisit the abnormal detection task from both microscopic and macroscopic viewpoints by using a statistical motion analysis approach. Subsequently, we introduce two social attributes to formulate the interaction of crowd behavior. Both the attributes offer the richness of interaction which reflect the social evidence. Our attributes are constructed by quantitative measurement of the motion features, which provides socially motivated clues in expressing the interaction effect.

Our third contribution is to emphasize the social semantic influence on the crowd interaction behaviors by using the proposed SAFM, which can also be extended to an online process via an effective attribute weight fusion algorithm. This is achieved by representing each attribute as a grid-weighted map with adaptive pooling. We jointly integrate different weighted maps in an online manner, which ensures the processing efficiency.

*Justification to Existing Model:* In contrast to a recent Bayesian model [32], which also uses different flow-field attributes to characterize crowd motion, our model differs in the following aspects.

1) SAFM is embedded with completely different scale information from Bayesian model. We use the combination of local scale and global perspective, which naturally achieves more robustness to the scale changes of scene than the regular grid used by Bayesian models.
2) Our model copes with crowd interaction by social characteristics in an online fusion manner such that abnormal crowd patterns of various types and density in this case could be identified. In comparison, the recent Bayesian model updates the probability density of optical flow in a Bayesian framework, which only employs concepts of divergent centers to detect limited escape events. Additionally, Bayesian model is sensitive to optical flow estimation and works on low or medium crowd to identify escape patterns.
3) The two methods have different objectives, that is, our method aims at exploring the interaction of crowd motion for all types of abnormal detection, whereas the Bayesian model in [32] focuses on modeling the crowd motion only in the cases of escape.

The rest of this paper is organized as follows. In Section II, we describe the social force estimation and propose the social attribute hypotheses. Section III details SAFM. Experimental results on global and local abnormal detection are reported in Section IV. We conclude this paper in Section V.

## II. Estimating Crowd Interaction Force

Fig. 2 shows the flow chart of the proposed method. We first place a grid of particles in the frame as individuals and compute the conventional social force by particle advection scheme. Then, we develop a SAFM, which aims to enrich scene scale and social attributes on group interactions. We calculate the interaction force and obtain social attribute-aware force maps by an online fusion algorithm. Finally, we use bag-of-words representation for GAE detection and abnormal map approach for LAE detection as well as localization with corresponding classifier. We describe the social attributes to reflect the interaction characteristics, and to

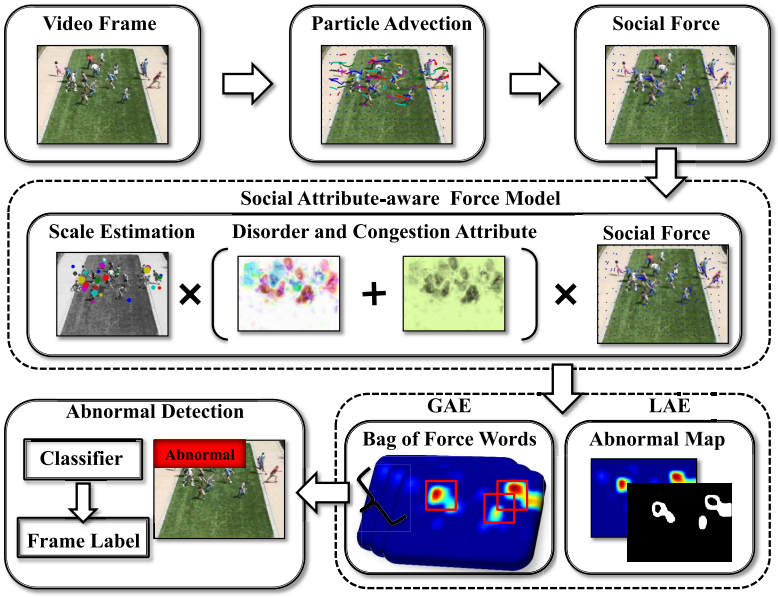



Fig. 2. Flow chart of proposed method. The input sequences are extracted social forces with particle trajectories in the first scheme. In the second scheme (main part of our method), we estimate scale and social attributes. Inspired by the social hypotheses, social attributes are modeled by the proposed statistical features. From an online fusion algorithm, we automatically obtain the SAFM map which covers the rich interactions in the social groups by integrating the attributes. We propose to detect abnormality by a bag of force for GAE in a supervised learning way. The abnormality map obtained by SAFM map acts as an effective tool for the detection and localization in the task of LAE (in the experimental scheme).

mode attractive and repulsive phenomenon under the social attribute hypotheses. Our SAFM is thus capable of describing the self-organized interactive effects of the pedestrian behaviors reliably.

### A. Particle Advection

For the ideal case, it is best to track and analyze the motion of all the individuals to describe the crowd. However, tracking all the pedestrians in a dense-crowded scene presents challenges in terms of both accuracy and efficiency. In this section, we detail an approach to simulate the trajectories of pedestrians in a crowded scene. In earlier works [23], [33], the particle advection scheme is carried out to simulate crowd motion behaviors instead of tracking the targets. Under such circumstances, the continuous evolution of the group motion will be captured as the particle trajectories [23]. To deal with that, we set up particles from the scene as an efficient approximation of the individual motion which are viewed as the representatives of the pedestrians. Therefore, crowd flow can be treated as the particle flow, which overcomes the difficulties in tracking. Then, we can analyze the crowd behaviors directly from the particle flow. The particles that propagate in the optical flow also address the issue of continuity in crowd motion, and capture more effective individual movements than the methods simply using optical flow.

To start the process, a homogeneous grid of particles is placed over the video frames with a set amount of particles. Then, the velocity for each particle is calculated using a fourth-order Runge–Kutta algorithm, along with the average velocity $OF_{\text{ave}}$ in optical flow field. Consequently, individual particles will follow the trajectories in a fluid flow guided by the average flow direction of their neighborhood.

### B. Social Attributes Hypotheses

The social force (SF) [10], which is computed for every particle, is formalized as follows. It describes crowd dynamics by considering personal desires and constraints of the environment as social force to reflect the various drives as individual pedestrian experiences. It is a quantity that describes the concrete motivation to act [13]. Typical social forces include things like reaching a preferred velocity, attraction to goals, and repulsion from obstacles. These forces can be typically combined in superposition as $F_a = F_p + F_{\text{int}}$, which is most commonly used in physical systems [13]. $F_p$, as the first part of the actual force $F_a$, is the desire force of the target. $F_{\text{int}}$, the second part of $F_a$ is the interaction force, which is influenced by the environment force $F_w$ and pedestrians force $F_{\text{ped}}$ around. Thus, the social force on a pedestrian is the sum of the individual forces from nearby pedestrian's goals and obstacles. It is made up of several components which act to change the generalized desired velocity $v_i^q$, of each pedestrian $P_i$. Note that if the agent's current velocity differs from the preferred velocity defined by the desired speed and direction, there will be a personal desire force $F_p$ which causes the agent to accelerate. This force $F_p$ takes the form as

$$F_p = \frac{1}{\tau}\left(v_i^q - v_i\right) \tag{1}$$

where $\tau$ is the relaxation parameter, $v_i^q$ is the ideal velocity of the pedestrians motion by the effect of $F_p$ (every pedestrian has a movement velocity toward to the target direction), and $v_i$ is the actual velocity which is influenced from the external factor, to maximum extent, is close to the ideal velocity under the effect of $F_p$. Overall, basic social force is summarized as

$$m_i \frac{dv_i}{dt} = F_a = \frac{1}{\tau}\left(v_i^q - v_i\right) + F_{\text{int}} \tag{2}$$

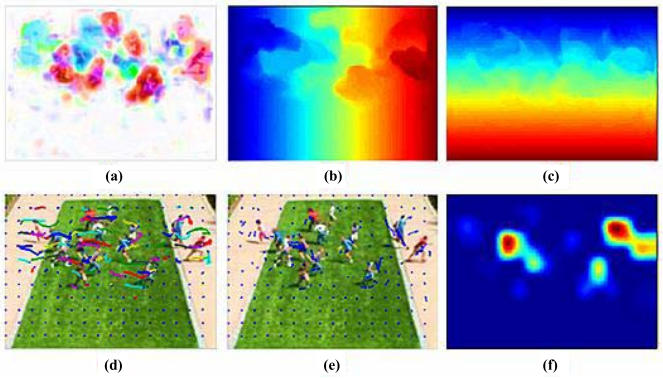



Fig. 3. Social force model computation process and results. (a) Average optical flow shown in HSV color space. (b) and (c) $x$-axis and $y$-axis optical flow cumulative results. (d) Particle trajectories in the particle advection scheme. (e) Calculated interaction force of the particles. (f) Results of mapping the interaction force to the image plane by bilinear interpolation.

where $m_i$ is the mass of the pedestrian. Fig. 3 shows the process of the social force computation in our experiment.

*1) Social Force Discussion:* In the social force formulation, $m_i$ denotes the mass of the individual, $F_{\text{int}}$ denotes the interaction force experienced by the individuals. Helbing's social force model [13], to some extent, can provide basic explanations to actual phenomenon like Arch-like exit effects, faster-is-slower effects, and so forth. However, the model has obvious deficiencies in representing social interactions.

For example, as the reflection of the attractive and repulsive force, $F_{\text{int}}$ indicates the interaction among individuals which is only influenced by velocities. In this case, masses of individuals are all treated the same, which means all pedestrians take the same effect to pedestrian $P_i$ and they follow the same rules changes. Nevertheless, in the realistic situation, it is easy to find different people and scenarios that create various influences on specific pedestrian $i$. In some cases, friends enjoy walking closely, only strangers need larger safe distance among them. In addition, the desired velocity in the social force is a constant value, which means the pedestrian keeps the same desired motion in different regions of the scene. However, an actual pedestrian's velocity varies as time and space changes. The velocity interaction which is given by the desire velocity $v_i^q$ and the actual velocity $v_i$ in the social force model is also too simple to represent the realistic interaction of crowd behaviors.

*2) Hypotheses Proposal:* To best mimic real situations, both contextual and social semantic effects should be considered in formulating crowd behaviors. Inspired by social behavior modeling [12], people are driven by their own destination and emotion which is influenced by interactions with other people. We can make three reasonable assumptions.

*a) Contact:* People can predict the motion intentions of others within a certain distance and have a general estimation about when they might meet each other. The self-driven contact interaction shows how the pedestrian moves, and how the velocity changes with the time. As shown in the green block of Fig. 4, the girl has a chat with the couple and contacts them for some purpose, which is the source of interaction.

Fig. 4. Social attribute hypotheses illustration. Green block: Chatting people in line with *Contact* hypothesis. Blue block: *Consistency* hypothesis between the friends. Red block: *Exclusion* hypothesis as chaotic and congested situation.

*b) Consistency:* People with similar destination or motion direction rarely repel each other (for instance, people attracted by the exit, in a marathon race game, group of friends, etc.). As we can see from the blue block of Fig. 4, the two friends with the same desired goal have no repulsive force at all.

*c) Exclusion:* Chaotic and congested situations make people repulsive to each other. Repulsive forces arise in situations shown in the red region of Fig. 4. We wish to model the attributes hypotheses based on the social behavior characteristics and prompt reliable social semantic responses of midlevel attributes.

## III. Social Attributes-Aware Force Model

Before explaining our model, we first formalize our definitions of the terms attribute and social attribute-aware force.

*Definition 1 (Attribute):* Attribute is a specific feature set of an object, which contains the semantic properties and reflects distinguishable characteristics of an object. In our context, social attribute is the feature set for social interaction within crowd motion.

*Definition 2 (Social Attribute-Aware Force):* Given a crowded scene, the social force aims to compute the effect force carried by the particles to simulate the interaction between pedestrians. For social attribute-aware force, it refers to the force with social semantic information, which can be carried out by construction of attributes in social context, with regard to the social behavior constraints.

SAFM constructs an intermediate representation with our proposed social attributes which is scale-aware and context-oriented extracted from the scene. By reflecting interaction of social behaviors and take the space-time difference into account. We formulate our model as

$$F_{\text{SAFM}}^{\text{int}} \propto \underbrace{W_{ij}^{S}}_{E_{\text{Scale}}} \times \Big( \underbrace{W_{ij}^{D}}_{E_{\text{Disorder}}} + \underbrace{W_{ij}^{C}}_{E_{\text{Congestion}}} \Big) \times F_{\text{int}}. \tag{3}$$

The interaction force $F_{\text{SAFM}}^{\text{int}}$ is proportional to attribute weights terms multiplied by interaction force $F_{\text{int}}$. The attribute weight affects the interaction behaviors, by considering geometrical scale variance, influence of the social characteristics, and contextual environment. We directly model intensity of interaction using social attributes constructed by statistical features. Details are provided below.

### A. Interaction Scale Estimation

Scene scale imposes a strong prior on the sizes and velocities of pedestrians which is consistent with the geometry of a scene. We seek to roughly estimate the scale changing, which in turn informs the approximation of scene geometry and density in crowds. In preliminary processing, we first adopt a background subtraction method to a video sequence based on frame differences. The scene scale can infer an individual region in the frame, which is also an indicator of individual mass. To distinguish the basic scene scale, a straightforward strategy is to get the camera parameters. However, most videos lack such information. We address this issue by using a distance transform (DT) map. In the image domain, Breu *et al.* [34] adopted DT to detect the boundary for medical image processing. Kim and Grauman [35] presented a DT process as part of a boundary preserving feature in object description. We present a DT map strategy to obtain the local peak value of the foreground movement, which yields the approximate scale of an individual.

*1) Distance Transform:* It is an operator normally applied to binary images as (4), which calculates the gray level intensities of points inside the foreground region $O$. We use DT to generate a map $D$ in each pixel $p$ which is the smallest distance from the binary mask $O^c$

$$\begin{aligned} D(p) &:= \min\{d(q,p)|q \in O^c\} \\ &= \min\{d(p,q)|I(q)=0\}. \end{aligned} \tag{4}$$

In our case, $d$ is defined by Manhattan distance $d(x,y) = |x_1 - x_2| + |y_1 - y_2|$. The DT map on the foreground binary mask is computed based on background subtraction, which makes an average scale estimation in each scene region. By obtaining the region scale, we can get the approximation of the individual scale. It also provides the density clues of the crowd by cumulative sum of the peak points for social attribute construction. The first term $E_{\text{scale}}$ in (3) ensures force of interaction is consistent with geometry of the scene. For each region of the foreground, we can find the local maxima $D_{\text{localmax}}(p)$ of the DT map, and set the maximal distance value as the region's scale. Fig. 5 shows the estimation details for the interaction scale. From video frame (a), the foreground movement (b) is obtained by background subtraction. The corresponding DT map (c) encodes the scene sale information. The highlight peak value can approximately represent the scale of the individual.

*2) Grid Cell Setting:* To capture the interaction variance, we adopt a grid based strategy to characterize spatial inconsistency. With the principle that interaction varies in different regions, each frame is divided into $i \times j$ cells, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, M$. Let $\Gamma_{ij}$ denote each cell. We use the maximum and minimum region average scale $S_{\Gamma_{\max}}$ and $S_{\Gamma_{\min}}$ in horizontal cells $\Gamma_i$ and the corresponding vertical coordinates $i_{\Gamma_{\max}}$ and $i_{\Gamma_{\min}}$ to compute the scale attribute weight.

The scale attribute weight is linearly extended to the whole height $H$. The associated formulation can be given by

$$W_{ij}^S = \frac{(H-i)}{H} \times \left( \frac{S_{\Gamma\max} - S_{\Gamma\min}}{i_{\Gamma\max} - i_{\Gamma\min}} \times \frac{i_{\Gamma\max}}{i_{\Gamma\min}} - 1 \right) + 1. \tag{5}$$

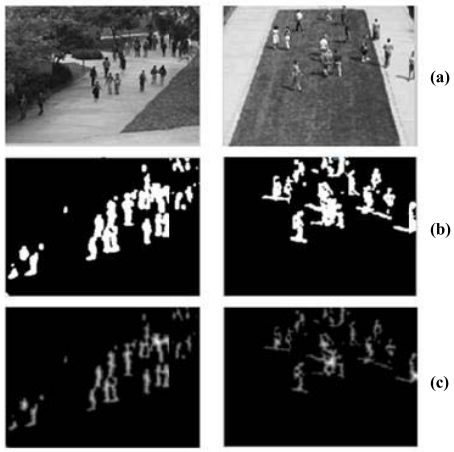



Fig. 5. Interaction scale estimation based on DT map. (a) Video frame. (b) Foreground movement by background subtraction. (c) Corresponding DT map.

*3) Mass and Scale Correlation:* If we rewrite the $F_{\text{int}}$ and extract the mass of individual $m_i$, and replace it with the scale attribute weight as

$$\begin{aligned} F_{\text{int}} &= m_i \frac{dv_i}{dt} - \frac{1}{\tau}\left(v_i^q - v_i\right) \\ &= W^S \left( \frac{dv_i}{dt} - \frac{1}{\tau'}\left(v_i^q - v_i\right) \right). \end{aligned} \tag{6}$$

We can see that substitution of individual mass with scale attribute weight yields a feasible measurement for generalized interaction. Therefore, our interaction force is able to model the relative size for the regions of the scene, which is addressed by constructing the grid-based scale attribute as (5). We can further enrich the interaction by the reinforcement of the social characteristics.

### B. Social Attributes

Social attributes offer a useful intermediate representation between low-level features and high-level social behavior. Our next goal is enable the interaction to cope with more realistic social semantics. In such cases, social attribute learning is not a good choice, because the social attributes are diverse for various level social definitions. Additionally, there are not enough labeled data sets, and the attributes can be difficult to get empirically. Obviously, there are many social attributes which indicate group distinct characteristics, such as environments, relationships, and so on. Such diversity is an important cue to discriminate the interaction of crowds. To handle this issue, our attributes achieve the goal via sociologically inspired responses built from low-level motion features under the social attribute hypotheses. To unify different attributes, we attach the attributes to the social force which provides carriers for the interaction description.

The second term $E_{\text{Disorder}}$ and the third term $E_{\text{Congestion}}$ in (3) stand for the social *disorder attribute* and *congestion attribute* generated from low-level motion features of groups, respectively. These are the typical attributes of social behavior to reflect social attractive and repulsive interactions.

*1) Disorder Attribute:* $E_{\text{Disorder}}$ expresses the weight of force representing the social attribute of disorder. The attribute can be regarded as a quantitative measurement for disorder degree of group behaviors. Disorder attribute is in line with the property of the consistency hypothesis. When particle $P_i$ and $P_j$ contact, the disorder motion indicates the inconsistency of the group behavior. In a disordered situation, the pedestrian wants to preserve their self-driven motion and keep a safe distance from the others. The interaction force is increasing in such case. We characterize statistical features of direction to define this crowd motion attribute as

$$\begin{aligned} W_{ij}^{D} &= A_{ij} \exp(\text{std}(\phi_{ij}) - \text{std}(\phi_T)) \\ A_{ij} &= \text{sgn}(\text{std}(\phi_{ij}) - \text{std}(\phi_T)) \\ \phi_{ij} &= \text{Hist}_{ij}\{O_n\}, \quad n \in 1 \ldots 8 \end{aligned} \tag{7}$$

where $\phi_{ij}$ is an orientation histogram. After the particle advection, we can get the $k$ frame particle position set $S = (x_1, y_1),(x_2, y_2), \ldots, (x_n, y_n)$ and corresponding $k+1$ frame particle position set $Q = (X_1, Y_1), (X_2, Y_2), \ldots,(X_n, Y_n)$. For the particle $P_i$, the motion phrase and orientation are

$$\begin{aligned} |V_i| &= \sqrt{(x_i - X_i)^2 + (y_i - Y_i)^2} \\ O(V_i) &= \arctan\left(\frac{|y_i - Y_i|}{|x_i - X_i|}\right). \end{aligned} \tag{8}$$

The histogram $\phi_{ij}$ with 8 bins is formed for each cell $\Gamma_{ij}$ by partitioning the inside particle orientation $O(.)$ into 8 parts with 45 angles intervals. std(.) denotes the standard deviation of the histogram, which is an efficient way to describe variance of group motion orientation. Compared with the entropy and other uncertainty measurement in information theory, it leads to better overall performance in describing diversity measurement and is simple to implement. std $(\phi_T)$ is set as a certain threshold. $A_{ij}$ is the symbol function. When the std $(\phi_T)$ is larger than a threshold, $A_{ij}$ is positive, otherwise, $A_{ij}$ is zero. We construct the disorder attribute weight $W_{ij}^{D}$ to measure motion orientation inconsistency and enhance the influence of contextual neighbor motion on the group interaction. The disorder attribute reflects the group's overall interaction intensity. It is easy to see that the value rises gently at the beginning, and then increases dramatically when it becomes larger in the exponential function. When the local chaos exceeds a certain level, the disorder attribute begins to have an effect on the interaction force. Further, the force grows as the existed disorder effect becomes larger. Higher values for the disorder attribute signifies increasing group chaos.

*2) Congestion Attribute:* Similar to the disorder attribute, $E_{\text{Congestion}}$ is defined to model the congestion attribute of crowds. To keep a safe distance and space, it generates great repulsive forces to keep the self-driven force which follows from the exclusion hypothesis. We use the linear form to formulate the attribute as shown

$$\begin{aligned} W_{ij}^{C} &= K_{ij} B_{ij} (\theta_{ij} - \theta_T) \\ B_{ij} &= \text{sgn}(\theta_{ij} - \theta_T) \\ K_{ij} &= \text{std}(\text{Hist}(V_{ij})). \end{aligned} \tag{9}$$

The congestion attribute is linearly related to the density of the cells. We count the number of DT map peak points $\theta_{ij}$ in each grid cell to make a coarse crowd density estimation

$$\theta_{ij} = \sum_{p \in \Gamma_{ij}} D_{\text{localmax}}(p) \tag{10}$$

where $\theta_T$ is a threshold which is a certain density of the cells and $K_{ij}$ is a friction coefficient that is computed by the standard deviation of the magnitude histogram of particle velocity $V_{ij}$. The congestion attribute weight $W_{ij}^{C}$ increases linearly as crowd density and diversity of group velocities grows. Thus, when the local pedestrians are congested, the attribute begins to react on the interaction force. Higher congestion values depend on the local pedestrian density, as well as speed differences between them. In real-world situation, pedestrians begin to change movements chaotically when encountering congested crowds. This occurs because of huge densities in the local region and big variations in velocities.

### C. Effects Underlining

Our aforementioned attributes are concerned with enriching the interaction of the social force. We account for the mass of individuals in the social force formulation and distinguish the surrounding interaction for the pedestrians as self-organization factor. Thresholds obtained by pooling have the adaptive and stationary properties that make the attributes more robust against the local motion noise. As a quantitative measure of crowd interaction, we make use of the several underlining effects for exhibiting the realistic behavior.

*1) Mass Effect:* The mass of a particle is mainly related to the foreground detection and parameters of the grid setting according to the scale variance. For the purpose of motion estimation and normalization of scale, we set the frame-difference ten-frame interval for background subtraction.

*2) Orientation Effect:* The disorder attribute draws practical implications from the observation that pedestrians are more strongly repulsed when they are moving in different directions than in the same direction. Note that the considerations lead to the case of orientation interactions. We use the max pooling to select the most salient disorder region as the pooled value for the latest five frames $[F_{t-5}, F_t]$, and then a center-surround normalization is computed for a eight neighborhood to get the threshold, which is inspired by the mechanism of the complex cells

$$\text{std}(\phi_T) = \frac{1}{8} \max \text{std}(\phi_{ij}), \quad \phi_{ij} \in [F_{t-5}, F_t]. \tag{11}$$

*3) Density Effect:* The congestion attribute concerns the dependence of the social force on the density of crowd in a region. We observe that the influence of dynamic density is sufficient by setting density threshold in a average pooling manner by adopting the elementwise average value over the cell for the latest five frames

$$\theta_T = \frac{1}{i \times j} \sum \theta_{ij}, \quad \theta_{ij} \in [F_{t-5}, F_t]. \tag{12}$$

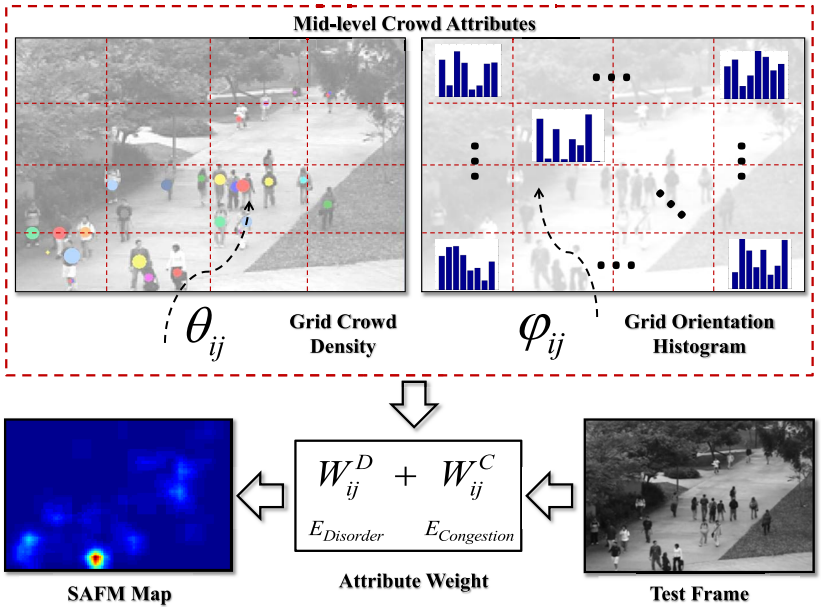



Fig. 6. Midlevel crowd attributes using statistical low-level features. They are generated by the grid crowd density and the grid orientation histogram. SAFM map finally acts as the midlevel representation for indicating abnormality.

**Algorithm 1** Online Attribute Fusion Algorithm

**Input:** Sequential input frame $F \in [F_1, F_2, \ldots, F_t]$; Initial particle $P \in [p_1, p_2, \ldots, p_N]$; Initialize the attribute weight maps;
**for** $t = 1$ ***to*** $T$ **do**
  Pursuit social force for each particle $F_{\text{int}}^{p_i}$ by 2;
  Extract the frame difference feature map $F_{\text{diff}}$;
  Compute the DT map $DT_{\text{diff}}$ ;
  Divide the $F$ into $i \times j$ cells;
  Weight the scale attribute of interaction force by 5;
  **for** *each cell* $\tau_{ij}$ **do**
    Threshold selection of the attributes in the past five frames: Compute the $\phi_T$ by max pooling as 11 ;
    Compute the $\theta_{ij}$ by average pooling as 12;
    Calculate the disorder $W_{ij}^D$ and congestion attribute $W_{ij}^C$;
    $F_{\text{int}}^{P(x,y)} = F_{\text{int}}^{p_i} \times (W_{ij}^D + W_{ij}^C) \times W_{ij}^S, (x, y) \in \tau_{ij}$;
  **end**
  Bilinear interpolation for $I_{\text{SAFM}} = Bi(F_{\text{int}}^{P(x,y)})$;
  Return $I_{\text{SAFM}}^i$.
**end**
**Output:** SAFMs $\{I_{\text{SAFM}}^i\}_{i=1}^T$.

### D. Online Attribute Fusion Strategy

In this section, we adopt an online attribute fusion strategy to efficiently obtain the SAFM. We represent each of the attributes by the weighted grid map of the particles. Specifically, we get the DT map $\text{DT}_{\text{diff}}$ for each ten-frame difference $F_{\text{diff}}$. We compute the attribute threshold for each cell during the latest five frames to update the attribute weight map. The above objective attribute fusion problem following (3) simplifies to a few matrix manipulations, which we superimpose the attribute weight maps accordingly. The detailed steps of the online attribute fusion is summarized in Algorithm 1. In this way, social attribute-aware force is computed with the corresponding attributes and the magnitude for every pixel is mapped onto image plane by bilinear interpolation of the resulting particle forces. Midlevel social attributes produced by statistical low-level features and the generated SAFM ($I_{\text{SAFM}}$) are shown in Fig. 6. Consequently, we construct the maps $\{I_{\text{SAFM}}^i\}_{i=1}^T$ to complete the following abnormal detection task. The higher value for interaction in the map indicates higher probability of abnormality may happen. By introducing social attributes of crowd behaviors, the model represents the crowd from midlevel viewpoints which encode more realistic interaction behavior.

## IV. Experimental Evaluation

To validate the effectiveness of our proposed model, we perform three groups of quantitative comparisons. The first group compares our approach with several state-of-the-art methods for the tasks of GAE and LAE detection with the quantized measurement. We evaluate it on three public available data sets, including the University of Minnesota (UMN) data set [36], University of California, San Diego (UCSD) ped1 data set [3], and University of Central Florida (UCF) Web data set [10]. The second group compares our approach with the alternative approaches of attributes combination and also investigates how the visual content, visual changes as well as parameters affect abnormal detection performance. The final group reports the detection and localization results on the UCSD ped1 data set.

### A. Comparisons With State-of-the-Arts

*1) UMN Data Set:* The UMN data set is used to test the GAE detection by event-level measurement [3]. It consists of 11 clips of crowded escape events in three different scenes. The resolution is $240 \times 320$. Each video begins with normal behaviors and ends with panicked escape.

In the particle advection step, we set a particle every 5 pixels in the optical flow field. The grid $\Gamma$ is built with $6 \times 8$ cells, which are $40 \times 40$ pixels in size. For the construction of the visual words, we randomly select 30 spatial–temporal

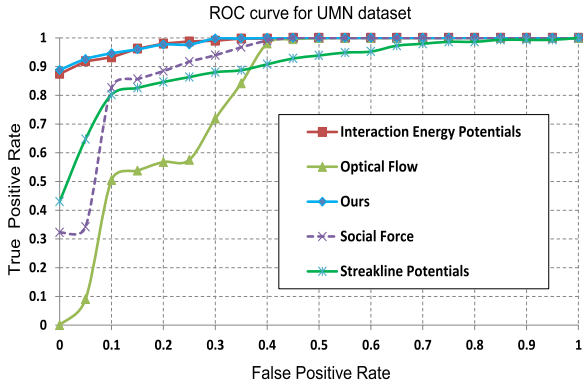



Fig. 7. ROC curves of GAE detection in UMN data set.

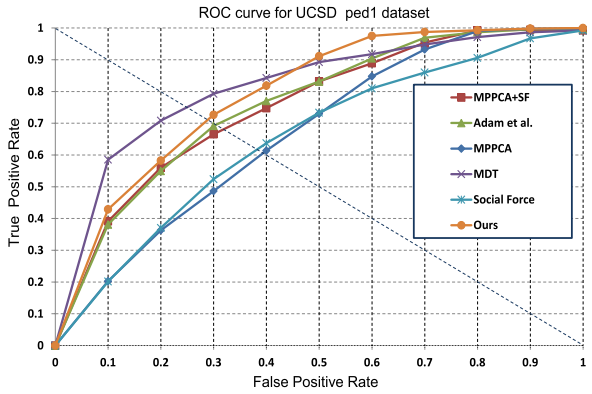



Fig. 8. Performance of the approaches for LAE detection on UCSD ped1.

TABLE I
COMPARISON OF DIFFERENT METHODS FOR ABNORMAL DETECTION IN UMN DATA SET

| Method | **Ours** | IEP [11] | SF [10] | SP [31] | OF |
|---|---|---|---|---|---|
| AUC | **0.986** | 0.985 | 0.96 | 0.90 | 0.84 |

TABLE II
COMPARISON OF DIFFERENT STATE-OF-THE-ART METHOD, EER AND RD FOR DETECTION AND LOCALIZATION IN UCSD PED1

| Method | AUC |
|---|---|
| **Ours** | **0.7821** |
| MDT [3] | 0.8141 |
| SF+MPPCA [10] | 0.7422 |
| Adam [2] | 0.7478 |
| MP-PCA [1] | 0.6663 |

volumes in the salient region of the frames (ensure the value in the volume is big enough) with the size $5 \times 5 \times 10$ following the setup in [10]. The codebook contains 30 cluster centers. In the experiment, we use SVM with RBF kernels to train the model on ten videos and compute the false positive rate and true positive rate in a leave one out manner. Fig. 7 shows the receiver operating characteristic curve (ROC) curves of methods, which are listed for comparison directly obtained from [10], [11], and [31].

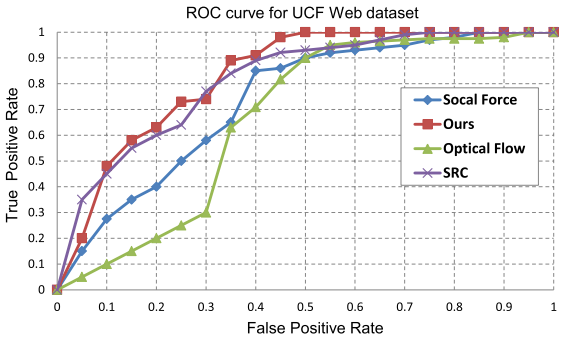
The result shows that our method outperforms other state of the art methods. The AUC of our method is 0.986, which is shown as quantitative results shown in Table I. Note that SAFM can achieve better performance over available state-of-the-art methods, including interaction energy potentials [11], SF [10], streakline potential [31], and optical flow. It also suggests that SAFM could achieve a better effect to represent the interaction patterns in such crowd activities. We attribute the performance to the fact that we consider the contextual interaction attribute in the group scale, as well as the crowd's motion pattern information in local scale, which is more effective in improving the performance.

*2) UCSD Ped1 Data Set:* For the UCSD ped1 data set, which is tested for LAE detection, we use the frame-level measurement defined in [5]. Our model is tested in the 36 clips, and each clip has 200 frames with a $158 \times 238$ resolution. Each frame is resized to $240 \times 320$ pixels. We obtain one abnormality map for each frame by using our model in unsupervised way. Because the ground truth annotation is given per frame, we use a continuous threshold to generate the ROC curve. We also compare our model with SFM [10], MPPCA [1], and so on, which are shown in Fig. 8. Our method is competitive with the other state-of-the-art methods. The quantitative results of AUC are presented in Table II.

We notice that mixture of dynamic texture (MDT) [3] obtains slightly better results than our method. The reasonable explanations may be that we focus on the issue of modeling the interaction context between groups of people, so the appearance or dynamics of visual processes are not considered as much. As an interactive description, the addressed scale and density of the local region may bring some false alarms and lead to unexpected errors. Nevertheless, the MDT method is a supervised method and is very time-consuming (about 2 h for training, 25 s/frame for testing). Our method, in contrast, can achieve online update with a testing time around 2 s/frame without training scheme (details in Section IV-C).

*3) UCF Web Data Set:* To further evaluate the effectiveness of our proposed model, we also conducted an experiment on a more challenging data set called UCF Web data set [10]. We used our algorithm in the real-life scenarios to see how it performs. The data set contains 12 sequences of normal real-world scenarios including pedestrian crossing, a marathon race and eight abnormal scenes including evacuation, fighting, and escaping. Some selected sample frames are shown in Fig. 9(a). Following a similar configuration as the UMN data set, SAFM maps were extracted from the frames and divided into $6 \times 8$ cells. We extracted $5 \times 5 \times 10$ volumes from a block of force flow to construct visual words. To learn the codebook, we randomly used the normal sequences in a twofold fashion and trained on the rest. In the testing phase, we added the excluded sequences to the test set. We repeated this experiment ten times and constructed the ROC by averaging the results of these experiments reported in Table III. In this experiment, our approach works well in various conditions such as extreme congestion in classifying the crowd fighting [Fig. 9(b)]. The reasonable explanation behind this performance is the congestion attribute which considers a neighbor velocity deviation upon the local density. This allows for identification of panicked fighting from the regular motion in extreme congestion.

(a)




(b)

Fig. 9. (a) Top row: samples from normal events. Bottom row: samples from abnormal events. (b) ROC curves for abnormal detection on the UCF Web data set.

TABLE III
COMPARISON OF DIFFERENT METHODS FOR ABNORMAL DETECTION IN UCF WEB DATA SET

| Method | **Ours** | SRC [5] | SF [10] | OF |
|---|---|---|---|---|
| AUC | **0.86** | 0.84 | 0.73 | 0.66 |

TABLE IV
QUANTITATIVE AUC COMPARISONS OF ALTERNATIVE APPROACHES FOR ABNORMAL DETECTION IN UMN, UCSD, AND UCF WEB DATA SET

| Dataset | UMN | UCSD ped1 | UCF web |
|---|---|---|---|
| Method | Area Under Curve | | |
| SAFM | 0.986 | 0.7821 | 0.86 |
| SAFM-Scale | 0.979 | 0.7561 | 0.78 |
| SAFM-Disorder | 0.967 | 0.7168 | 0.76 |
| SAFM-Congestion | 0.971 | 0.7298 | 0.80 |
| Social Force | 0.960 | 0.6505 | 0.73 |

### B. Comparisons Between Alternative Approaches

Our alternative approaches emphasize attributes that can be separately constructed, and validated independently or jointly with the algorithm. To explore the performance of different attributes, we conduct a series of experiments to study the effects of the terms in the attribute construction by using $W^S$, $W^D$, and $W^C$ separately in (3). SAFM with several attributes combinations among the attribute set $\Omega = \{\text{Scale, Disorder, Congestion}\}$ are designed in the experiment, called SAFM-Scale (with only scale), SAFM-Disorder (with only disorder), SAFM-Congestion (with only congestion), and SAFM-Combined (with both disorder and congestion attributes).

*1) Insights Into Effect of the Visual Content:* It is informative to look at how the visual content of different data set affects the alternative approaches with attributes combination. We investigate these two data sets (UMN, UCSD ped1) for distinct visual content. Note that, UMN and UCSD encompass typical global and local abnormal motion scenarios, which allows us to evaluate how our attributes react to specific crowd motion. Table IV shows the quantitative comparison of area under curves, which we can find from columns 1 and 2 that, although SAFM alternative approaches perform almost the same results, the attributes with contextual spatiotemporal information produces a significant improvement in both data sets compared to the baseline without attributes. Fig. 10 shows the following details of ROC comparisons.

1) *UMN Data Set*: As this data set is about the global escaping without large scale changes, we discover that our model ensures that motion patterns in spatial–temporal patches will be quantized into the attributes, such that this operation will significantly increase the performance. Disorder plays a more important role than congestion in such scenes. As shown in Fig. 10(a), the scale and congestion is not significant in such not very congested case. And, the performance degenerates with more background noise by comparing disorder. The AUC results (Table IV) show each attributes can make a contribution to our global abnormal detection improvement.
2) *UCSD ped1 Data Set*: Note that employing scale attribute improves the performance greatly and disorder attribute promotes capture of local changes as well, which is highly suitable for an overlooking perspective and the local motion detection in UCSD ped1. Congestion attribute also helps to distinguish the object from neighborhood movements which achieve higher results. Notice that congestion attribute results are positively correlated with scale [Fig. 10(b)]. This is due to our bottom-up density estimation in DT, which greatly affects the congestion attribute generation. Integrating all attributes into our model, SAFM achieves the best performance as shown in Table IV.

*2) Insights Into Effect of the Visual Changes:* We also investigate the robustness of different attributes for various conditions. Because the scenes of UMN and UCSD are fixed, we test our attributes against the visual changes across various scenarios in the UCF web data set. In Fig. 10(c), the ROC curves for alternative approaches are shown, from which we can find SAFM with scale attribute performs better than the basic social force. Table IV presents our detection performance in UCF web data set in column 3. The AUC of SAFM-Congestion is about 80% as the highest promotions among all the three attributes, while the one with Disorder is average 76%. The result underscores our contribution in employing congestion characteristics to capture the interaction

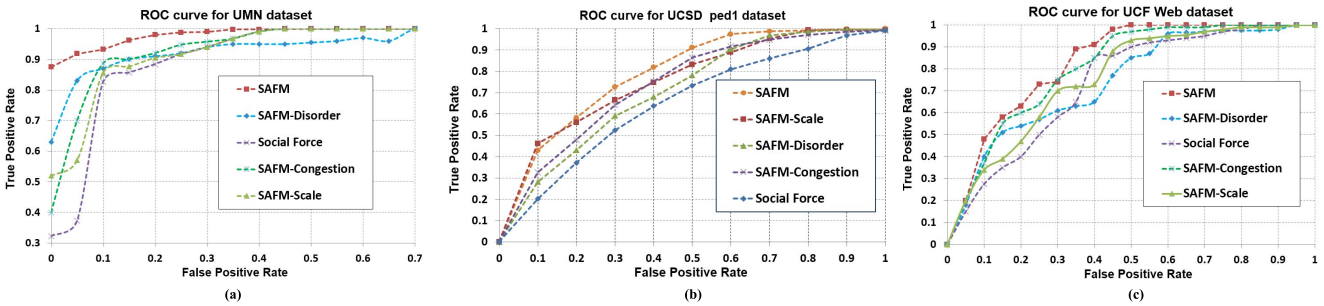



Fig. 10. ROC curves for alternative approaches of our proposed model. (a) UMN data set. (b) UCSD ped1 data set. (c) UCF Web data set.

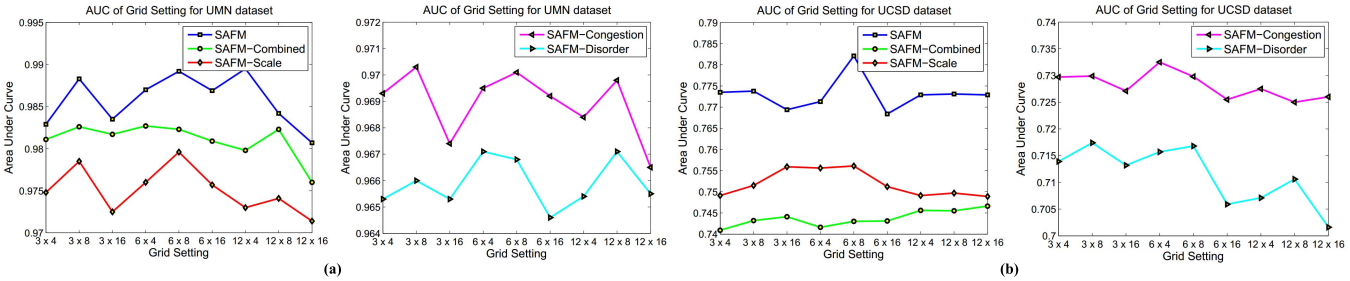



Fig. 11. Cross validation for parameters tuning. (a) Grid setting for alternative methods in UMN data set. (b) Grid setting for alternative methods in UCSD ped1 data set.

effect between local group patterns. SAFM-Scale outperforms SAFM-Disorder and is comparable to SAFM-Congestion, which demonstrates the robustness of our method against the resolution and perspectives. We can see that our model produces a more effective improvement by social attributes, due to: 1) congestion attribute reduces illumination influence by considering density and motion statistics instead of appearance and 2) disorder attribute compensates the occlusion errors in extreme congested scene by orientation quantization. Overall, the results reveal the high robustness of our attributes, which is tolerant of complex scenarios.

*3) Grid Setting and Weight Ratio Tuning:* Because the SAFM maps are constructed using several parameters and constants, we also evaluate the best tuning parameters of the Grid Setting and the Weight Ratio for alternative attributes. The cross-validation strategy is adopted to tune these parameters. Our validation set contains 11 videos in UMN data set and 36 videos in UCSD ped1 data set.

Fig. 11 shows the performance variations caused by different Grid Setting of the video frame. The grids are divided into $i \times j$ cells, $i = 3, 6, 12, j = 4, 8, 16$ to test the influence of cell size. Alternative approaches of SAFM with different attributes are tested. For the UMN data set, Fig. 11(a) shows that, by embedding identical cells into our attribute construction, SAFM-Combined works better than the SAFM-Scale. A sparser cells setting (with fewer cells for each frames in average) produces better performance in general. As can be seen in Fig. 11(b), the $6 \times 8$ offers the best ROC performance, which indicates this setting is the most suitable for the UCSD data set. It shows the scale attribute is more affected by the grid setting than social attributes. Grid Settings too fine or too coarse can both decrease performance. For the congestion and disorder, we can draw the conclusion that more detailed grids decreases performance. As such, we choose $6 \times 8$ as the best fitting grid setting in the all of our experiments. However, notice that the grid setting affects the disorder much more than congestion, it therefore proves that our congestion attribute is very effective to discriminate the local interaction pattern.

Fig. 13 shows the best fitting weight ratio assignment for disorder and congestion attributes in constructing our SAFM. For each assignment, the resulting point by cross validation of different weight distribution $\alpha$ on the social attributes ranged from 0.1 to 0.9. It is worth mentioning that, in the UMN data set, the balance of social attributes guarantees similar high performance. In comparison, the detection results are still largely affected by assigning less disorder attribute ratio, because such global abnormalities are based solely on the overall inconsistency of orientation. For most instances of UCSD data set, setting $\alpha$ around 0.4 for the disorder attribute and 0.6 for the congestion attribute yields the best performance. It indicates that congestion makes a more important contribution to the individual group interaction and disorder is a complement to correct unnecessary false alarm by balancing the orientation of crowd movements. There are two conclusions of tuning grid setting and weight ratio: 1) denser grids will bring the negative effect of representing the interactions in the local area and 2) balance of the social attributes and a slightly higher congestion ratio generally produces better performance.

## C. Detection and Localization

We further test our SAFM on the task of abnormity localization. SAFM maps are used as a reference for localizing and segmenting the abnormal area. As described in Section III-D,

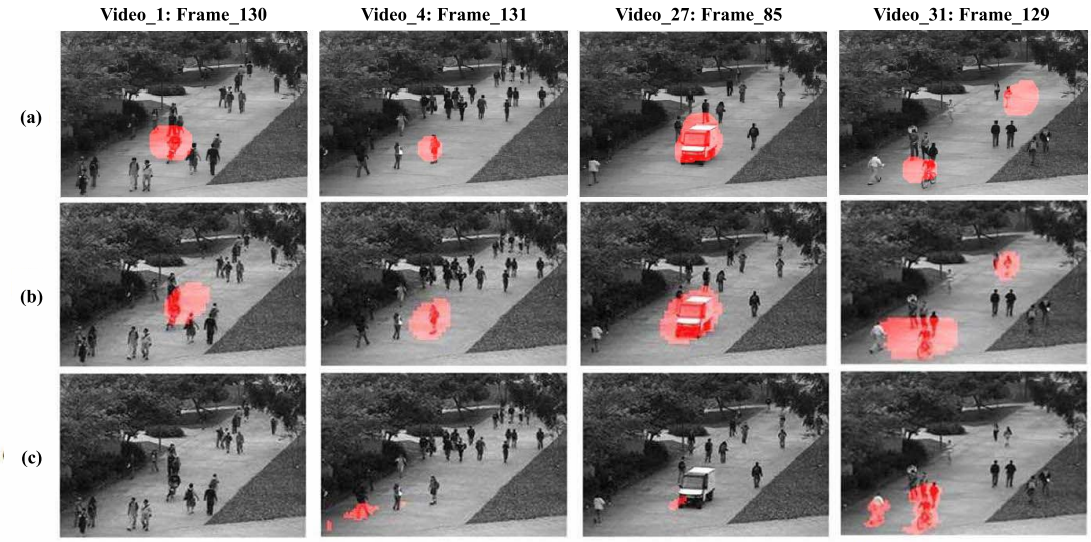



Fig. 12. Examples of the abnormal location results with comparisons of abnormal localization results from (a) our SAFM, (b) MDT, and (c) SF-MPPCA. The localization results of MDT and SF-MPPCA are directly provided by Mahadevan *et al.* [3].

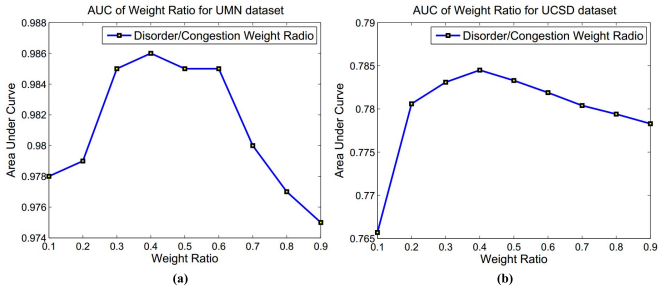



Fig. 13. Cross validation for parameters tuning. (a) Weight ratios $\alpha$ disorder/ $1-\alpha$ congestion attributes ($\alpha$ from 0.1 to 0.9) in UMN data set. (b) Weight ratios for the social attributes in UCSD data set.

the anomalous region are usually located in the high value of interaction force in the map. We adopt (13) to compute a localization protomask

$$O(x,y)\begin{cases}1 & \text{if } S > 5\times\bar{S}\\ 0 & \text{otherwise}\end{cases} \tag{13}$$

where $O$ is the protomask map and $S = \delta(I_{\text{SAFM}})$ is a refined SAFM map with Gaussian operator $\delta(.)$. We set the threshold to be $5\times\bar{S}$, $\bar{S}$ is the mean intensity of $S$. We further measure the performance by the following two criteria, which are commonly used in detection and classification.

1) Rate of detection (RD), which measures the percentage of the overlapping area between ground truth mask and our detection mask. It is

$$\text{Rate}_{\text{Overlapping}} = \frac{R^{\text{GroundTruth}} \bigcap R^{\text{Mask}}}{R^{\text{GroundTruth}}}. \tag{14}$$

To test anomaly localization accuracy, we take our detection result map compared with pixel-level ground truth masks on the ten clips of UCSD ped1 [3]. If at least 40% truly anomalous pixels are detected ($\text{Rate}_{\text{Overlapping}} \geq 40\%$), the frame is considered to have been correctly detected. The ROC curve is then generated by the true positive rate and false positive rate.

TABLE V

COMPARISONS OF STATE-OF-THE-ART METHODS, EER, AND RD FOR DETECTION AND LOCALIZATION IN UCSD PED1

| - | Detection and Localization in Ped1 | |
|---|---|---|
| Method | Equal Error Rate | Rate of Detection |
| SF [10] | 31% | 21% |
| MPPCA [1] | 40% | 18% |
| SF-MPPCA [10] | 32% | 28% |
| Adam [2] | 38% | 24% |
| MDT [3] | 25% | 44% |
| **Ours** | **26%** | **47%** |

2) Equal error rate (EER), which is the percentage of misclassified frames when the false positive rate is equal to the miss rate. For the localization component, we report detection EER for different methods.

Quantitative performance comparison as EER and RD values are shown in Table V on all clips of the USCD ped1 data set. All the methods are used to generate the candidate locations using the threshold abnormal map. The results are directly obtained from [1]–[3] and [10]. RD of SAFM increases by at least 3% using our localization method, while our EER measurement is also competitive with MDT. It is easy to see that our approach significantly improves performance by using online attribute fusion algorithm. Our midlevel representation produces a significant improvement in the pixel-level interaction description.

Some examples of video frames with abnormal detection and location by our SAFM approach and two other spatial temporal approaches (SF-MPPCA [1] and MDT [3]) for anomaly location are shown in Fig. 12. Visual results indicate that our method is capable of accurately localizing the anomalous interaction in the crowds and outputs better segmentation results with well-defined boundaries. We can see from that, our SAFM approach can detect and locate the bike in video 1, the skater in video 4, and the person running in video 31, which the SF-MPPCA completely misses. Compared with the

TABLE VI
COMPUTATIONAL COMPLEXITY COST EVALUATION OF DIFFERENT PART IN ONLINE FUSION ALGORITHM

| Time Cost Comparisons | |
|---|---|
| Steps/Cost | Time cost(s) |
| The Social Force process | 1.526s |
| The Attribute Computing process | 0.547s |
| The Online Fusion process | 0.021s |
| The Overall process | 2.094s |

complex MDT approach, ours provides more arcuate detailed locations. For video 31, we also detect the people gathering. This is mainly because the disorder and congestion attributes take effects in this region and raise the value of interaction force. Additionally, our localization is slightly in front of the detected motion. This is because the interactions for the desired moving goal are detected by our approach, which has a higher force value than other regions.

*Computational Efficiency Analysis*: The online fusion algorithm consists of three parts. The computation efficiency of each part is as follows.

1) The social force process, which (since we compute the social force with the particle advection) has a time complexity of approximately $O(N \times h^4)$, for $N$-particle locations in the flow field using fourth-order Runge–Kutta algorithm, with a adjust value $h$ of the advection process to compute the force.
2) The attribute computing process, which (as there is a DT map computation at the first step) has a time complexity of $O(m \times n)$ for an $m$-height by $n$-weight frame. The disorder and congestion attribute cost based on the histogram computation is $O(k)$, for $k$ is number of particles in each of $i \times j$ cells. The total cost is $O(m \times n) + O(i \times j \times k)$ for attribute construction.
3) The online fusion process, in which the time cost is linearly $O(k)$ to fuse the attributes and $O(N)$ in the framework with pooling strategy to get the SAFM maps. The total cost is $O(k) + O(N)$ for the online fusion processing.

Regarding the order of magnitudes, the overall cost for all steps can be represented as $O(N \times h^4) + O(m \times n)$. As further shown in Table VI, we list the comparisons of the computation complexity (measured by seconds of average running time) for the attribute construction and the different parts in our online fusion pipeline. We implement our approach with MATLAB on an Intel 2.4 G dual-core computer. The SAFM computation is around 2 s/frame and constructs a $240 \times 320$ map frame in about 0.02 s for the online process algorithm proposed in Section III-D. Therefore, abnormal detection can be performed for continuous video frames. In addition, we can easily extend our representation for the crowd interaction modeling problem, using techniques such as the localization with proto mask, classification using bag-of-words, crowd pattern matching, and retrieval.

## V. CONCLUSION

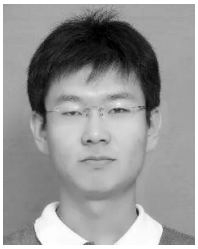
In this paper, we present a SAFM for discriminative visual social attribute construction. We aim to construct the midlevel quantitative representation by exploiting social attribute hypotheses, overcoming the drawbacks of the conventional SF model by associating with contextual attributes. With social attribute construction, the advantage of the proposed model lies in consideration of scene scale, disorder, and congestion attributes to effectively capture inherent interactions of crowd motion patterns. In other words, we represent the social behavior combining these statistical features to convey the richness of crowd interaction within the attribute construction procedure. We further propose an online attribute fusion algorithm, where we take adaptive pooling and map superposition focusing on threshold selection to reinforce the performance. The effectiveness of the experiments on the public data set indicates the method is very suitable for abnormal crowd behavior detection. Compared with the state-of-the-art and alternative methods, our SAFM is validated in global and local abnormal detection and localization tasks on benchmark data sets. We also report superior performances of our model which is robust to the various content and changes.

However, our work is validated only in the public video sequences containing basic interaction semantics. In more complex conditions like grouping and formation of crowds, it takes difficulties using the proposed attributes-based model without specific sociological priors. One possible solution is to incorporate the intelligent fusion of sensors, geospatial, and contextual information [37] that have different properties beyond the traditional interaction characteristics. Future work toward this direction may compensate for the limitations of existing methods.

## REFERENCES

[1] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2921–2928.
[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.
[3] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 1975–1981.
[4] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3441–3448.
[5] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3449–3456.
[6] Y. Tian, Y. Wang, Z. Hu, and T. Huang, "Selective eigenbackground for background modeling and subtraction in crowded scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1849–1864, Nov. 2013.
[7] Y. Zhang, L. Qin, H. Yao, P. Xu, and Q. Huang, "Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition," in *Proc. 20th IEEE ICIP*, Sep. 2013, pp. 3572–3576.
[8] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1446–1453.
[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[10] A. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 935–942.
[11] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3161–3167.
[12] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 261–268.
[13] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, p. 4282, 1995.
[14] R. Ji, L.-Y. Duan, J. Chen, L. Xie, H. Yao, and W. Gao, "Learning to distribute vocabulary indexing for scalable visual search," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 153–166, Jan. 2013.
[15] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE ICCV*, Nov. 2011, pp. 503–510.
[16] D. Mahajan, S. Sellamanickam, and V. Nair, "A joint learning framework for attribute models and object descriptions," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1227–1234.
[17] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.
[18] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1543–1550.
[19] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1345–1352.
[20] D. Helbing and P. Molnár, "Self-organization phenomena in pedestrian crowds," in *Self-organization of Complex Structures: From Individual to Collective Dynamics*. London, U.K.: Gordon and Breach Science Pub., 1997, pp. 569–577.
[21] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.
[22] J. C. Turner, P. J. Oakes, S. A. Haslam, and C. McGarty, "Self and collective: Cognition and social context," *Pers. Soc. Psychol. Bull.*, vol. 20, no. 5, pp. 454–463, 1994.
[23] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–6.
[24] D. Lin, E. Grimson, and J. Fisher, "Learning visual flows: A lie algebraic approach," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 747–754.
[25] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *Int. J. Comput. Vis.*, vol. 98, no. 3, pp. 303–323, 2012.
[26] R. Emonet, J. Varadarajan, and J. Odobez, "Extracting and locating temporal motifs in video scenes using a hierarchical non parametric Bayesian model," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3233–3240.
[27] N. Bellomo, A. Bellouquid, and D. Knopoff, "From the microscale to collective crowd dynamics," *Multiscale Model. Simul.*, vol. 11, no. 3, pp. 943–963, 2013.
[28] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 2871–2878.
[29] P. Scovanner and M. F. Tappen, "Learning pedestrian dynamics from the real world," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 381–388.
[30] M. Thida, H.-L. Eng, and P. Remagnino, "Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2147–2156, Dec. 2013.
[31] B. Mehran, B. E. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. 11th ECCV*, 2010, pp. 439–452.
[32] W. Wu, H.-S. Wong, and Z. Yu, "A Bayesian model for crowd escape behavior detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 85–98, Jan. 2014.
[33] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Proc. 10th ECCV*, 2008, pp. 1–14.
[34] H. Breu, J. Gil, D. Kirkpatrick, and M. Werman, "Linear time Euclidean distance transform algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 5, pp. 529–533, May 1995.
[35] J. Kim and K. Grauman, "Boundary preserving dense local regions," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1553–1560.
[36] *Unusual Crowd Activity Dataset of University of Minnesota*. [Online]. Available: http://mha.cs.umn.edu/movies/crowdactivity-all.avi
[37] *Evacuate Project*. [Online]. Available: http://www.evacuate.eu/



**Yanhao Zhang** is currently working toward the Ph.D. degree with Harbin Institute of Technology, Harbin, China.

His research interests include computer vision, multimedia, and machine learning, particularly crowd behavior analysis.

**Lei Qin** (M'06) received the B.S. and M.S. degrees in mathematics from Dalian University of Technology, Dalian, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008.

He is an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. He has authored or co-authored over 30 technical papers in computer vision. His research interests include image/video processing, computer vision, and pattern recognition.

Dr. Qin is a reviewer of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON CYBERNETICS. He has served as a TPC Member for various conferences, including the European Conference on Computer Vision, the International Conference on Pattern Recognition, the International Conference on Multimedia and Expo, the Pacific-Rim Symposium on Image and Video Technology, the International Conference on Internet Multimedia Computing and Service, and the Pacific-Rim Conference on Multimedia.

**Rongrong Ji** (SM'14) is a Professor, the Director of the Intelligent Multimedia Technology Laboratory, and the Dean Assistant of the School of Information Science and Engineering with Xiamen University, Xiamen, China. He has authored over 100 paper published in international journals and conferences. His research interests include innovative technologies for multimedia signal processing, computer vision, and pattern recognition.

Prof. Ji is a member of the Association for Computing Machinery (ACM). He is an Associate/Guest Editor of the international journals and magazines, such as *Neurocomputing*, *Signal Processing*, *Multimedia Tools and Applications*, *IEEE Multimedia Magazine* and *Multimedia Systems*. He also serves as a Program Committee Members for several tier-1 international conference. He was a recipient of the ACM Multimedia Best Paper Award and Best Thesis Award of Harbin Institute of Technology.

**Hongxun Yao** (M'03) received the B.S. and M.S. degrees in computer science from Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and 1990, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, in 2003.

She is a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. She has authored six books and has published over 200 scientific papers. Her research interests include computer vision, pattern recognition, multimedia computing, and human–computer interaction technology.

Prof. Yao received both the Honor Title of the New Century Excellent Talent in China and the Enjoy Special Government Allowances Expert in Heilongjiang, China.

**Qingming Huang** (SM'08) received the B.S. degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is a Professor with University of Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or co-authored over 200 academic papers in international journals and conferences. His research interests include multimedia computing, image processing, computer vision, and pattern recognition.

Dr. Huang has served as an Organization Committee Member and a TPC Member for various well-known conferences, including the Association for Computing Machinery's Annual Conference on Multimedia, the Computer Vision and Pattern Recognition Conference, the International Conference on Computer Vision, and the International Conference on Multimedia and Expo. He has been granted by the China National Funds for Distinguished Young Scientists.