# Online Anomaly Detection in Crowd Scenes via Structure Analysis

Yuan Yuan, *Senior Member, IEEE*, Jianwu Fang, and Qi Wang

***Abstract*—Abnormal behavior detection in crowd scenes is continuously a challenge in the field of computer vision. For tackling this problem, this paper starts from a novel structure modeling of crowd behavior. We first propose an informative structural context descriptor (SCD) for describing the crowd individual, which originally introduces the potential energy function of particle's interforce in solid-state physics to intuitively conduct vision contextual cueing. For computing the crowd SCD variation effectively, we then design a robust multi-object tracker to associate the targets in different frames, which employs the incremental analytical ability of the 3-D discrete cosine transform (DCT). By online spatial-temporal analyzing the SCD variation of the crowd, the abnormality is finally localized. Our contribution mainly lies on three aspects: 1) the new exploration of abnormal detection from structure modeling where the motion difference between individuals is computed by a novel selective histogram of optical flow that makes the proposed method can deal with more kinds of anomalies; 2) the SCD description that can effectively represent the relationship among the individuals; and 3) the 3-D DCT multi-object tracker that can robustly associate the limited number of (instead of all) targets which makes the tracking analysis in high density crowd situation feasible. Experimental results on several publicly available crowd video datasets verify the effectiveness of the proposed method.**

***Index Terms*—Anomaly detection, computer vision, machine learning, object tracking, structure analysis, video analysis.**

## I. Introduction

ABNORMAL behavior detection in crowd scenes is one of the hottest applications in computer vision field. The reason is that abnormal behavior in crowd scenes translates invaluable informative clues for various promising applications, such as intelligent surveillance [1], safety evaluation [2], behavior analysis [3], [4], etc. As a consequence, a surge of models [5]–[8] are motivated to this end. But as mentioned in [9], the exact notion of abnormality is hard to define because of distinctive applications involved. Therefore, it is still very challenging to design a general method for crowd abnormality detection [10], [11].

Based on the above consideration, this paper mainly focuses on the abnormal detection of crowd behavior. Among techniques toward this direction, some ones extract dramatic motion [7], [12], [13] of individuals to localize the abnormality, and others [8], [14], [15] advocate the object trajectories or paths which hardly appear to be the abnormality. For the first group, optical flow (OF) [6], [7], [13] or pixel/blob change [5], [16]–[18] is always utilized to extract the image motion clues. Through modeling the normal/abnormal crowd motion patterns, anomaly detection is conducted by pretrained classifiers. As for the second category, it always needs efficient trackers to obtain the trajectories. Through the classification or clustering, the trajectories hardly occur are associated with the abnormalities. Although trajectory-based methods have intuitional meaning for abnormal detection, they are difficult to be implemented in the crowd scenes with high density.

By psychological observations, crowd is arguably defined as a "collection of individuals who have relations to one another that make them interdependent to some significant degree" [19], and objects in the real world are almost always accompanied by other objects forming a global context or scene [20]. Inspired by that, this paper explores the crowd abnormal detection from the crowd structure modeling because visual contexts and scenes contain a rich, complex structure of covariation between visual objects and events [21]. In fact, there is a common sense that normal/abnormal clues of structural motion context are significantly different, which can be verified in Fig. 1. In spite of being intuitively reasonable, few work on anomaly detection has explored this information. One of the most related works, proposed by Ge *et al.* [22], argues that exploiting the crowd structure provides a basis for further mid-level analysis of events, and introduces this insight into the discovering of small groups in a crowd. But it does not reach scope of anomaly detection.

For crowd structure extraction, this paper proposes a novel SCD to intuitively exploit the context clues between individuals, which originally introduces the potential energy function of particle's interforce (PEF-PIF) [23] to model this relationship. For exploring the motion context, a novel selective

Manuscript received February 24, 2014; revised June 10, 2014; accepted June 11, 2014. Date of publication June 26, 2014; date of current version February 12, 2015. This work was supported in part by the National Basic Research Program of China (Youth 973 Program) under Grant 2013CB336500, in part by the State Key Program of National Natural Science of China under Grant 61232010, in part by the National Natural Science Foundation of China under Grant 61172143, Grant 61105012, and Grant 61379094, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102014JC02020G07. This paper was recommended by Associate Editor H. Qiao.

Y. Yuan is with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China (e-mail: yuany@opt.ac.cn).

J. Fang is with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China, and also with the University of the Chinese Academy of Science, Beijing 100049, China (e-mail: fangjianwu@opt.ac.cn).

Q. Wang is with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@nwpu.edu.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/TCYB.2014.2330853

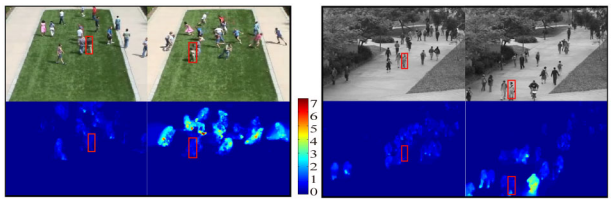


Fig. 1. Typical crowd scenes with changed structural context in different frames. The different brightness of the individuals in the flow field represents the different motion magnitude, one of the main attributes for determining anomaly. For the red rectangular target in each scene, the motion difference between it and its surrounding individuals changes obviously, which indicates the surrounding motion context changes and an abnormality occurs.

histogram of optical flow (SHOF) is proposed to pay more attention to different attributes of motion including direction and magnitude when dealing with different crowd anomalies. Then a robust multi-object tracker is designed to associate the targets in different frames, which employs the excellent ability of incremental analysis of the newly proposed 3-D discrete cosine transform (DCT) [24]. With the obtained targets' association, spatial-temporal analysis of the SCD variation is effectively computed, based on which the abnormality can be online detected.

### A. Overview of the Proposed Method

In this paper, an online anomaly crowd detection method is designed, and it is named as online anomaly detection for crowd via structure analysis (OADC-SA). The main steps are illustrated in Fig. 2, with a detailed description as follows.

*1) Pedestrian Detection:* Given a video sequence, in order to generate the target set of each frame, this paper utilizes the state-of-the-art pedestrian detection algorithm proposed by Dollár *et al.* [25] to extract each target of the crowd. The obtained targets are marked by rectangular regions with different sizes. To avoid the influence of the template scale for the future target association, every detected target sample is normalized into a general scale. Besides, to provide more clue for effective target association, a larger rectangle region centered in the target is simultaneously set to infer the neighborhood context.

*2) Structural Context Description:* For the individuals in the crowd, this paper proposes a SCD to exploit their valuable visual contextual information. The presented descriptor originally introduces the PEF-PIF in the solid-state physics [23] to model the relationship between the examined target and the other individuals. More specifically, this relationship is denoted by an inconsistency weight measuring the motion difference computed by a novel SHOF. The more different the descriptors, the larger the corresponding weight between them. This strategy is consistent with the human perception: "the individuals having large behavior difference to the observer are prone to violate the context structure and induce new configuration" [21].

*3) Multi-object Association:* For the target association, the difficult task is to explicitly model the appearance change [26]. One popular strategy is to learn a low-dimensional subspace, such as the incremental principle component analysis [27]. However, it has high computational complexity [24]. Considering this fact, an alternative object representation based on the newly proposed 3-D DCT [24] is utilized to accommodate to the appearance variation. With this representation, each target at different frames is associated by a context-aware multi-object tracker, which paves the way for the following SCD variation computation.

*4) Anomaly Detection:* In this step, the anomaly is online detected by temporal and spatial analysis of the SCD variation. Since the number of the targets in each frame may change with time, it causes different SCD dimensions in adjacent frames. Therefore, the SCD variation is computed by the Earth mover' distance (EMD) [28] which can analyze the similarity of two distributions with different dimensions.

### B. Contributions

Although many methods for crowd abnormality detection have been proposed recently, the method proposed in this paper is distinguished by the following aspects, which are also the main contributions of this paper.

1) Explore the anomaly detection from online crowd structure modeling. We exploit the visual structural context by directly treating each stable individual in the crowd as an observer. Through analyzing the context change of the observers, it is more efficient and effective to find anomaly. In addition, for constructing the observer's motion context, this paper proposes a novel SHOF to pay more attention to different motion property (including magnitude and direction) when facing different crowd anomalies. To the best knowledge of the authors, this kind of structure analysis method has not been exploited in the crowd abnormal detection literatures.
2) Propose a novel SCD for exploiting the contextual clues of the crowd. It originally introduces the PEF-PIF in the solid-state physics [23] to describe the relationship of the individuals. Then, the anomaly is detected by finding the large variation of SCD between newly observed frame and the previous ones. Through this comparison of context description, the proposed method is an online one, different from the traditional ones that need predefined normal/abnormal data for training.
3) Design a robust 3-D DCT multi-object tracker to associate the targets in different frames. This is inspired by the excellent ability of subspace learning to tackle the appearance change. More importantly, we only need to track the stable observers in the whole process, instead of analyzing the trajectories of every target. It makes the trajectory-based method feasible for anomaly detection in crowd scenes with high density.

The remainder of this paper is organized as follows. Section III presents the SCD for crowd structure modeling. Section IV proposes the 3-D DCT multi-object tracker to associate the targets among frames. Section V explains the criterion for determining the crowd abnormality. Experimental results and discussions are given in Section VI. The conclusion is summarized finally in Section VII.

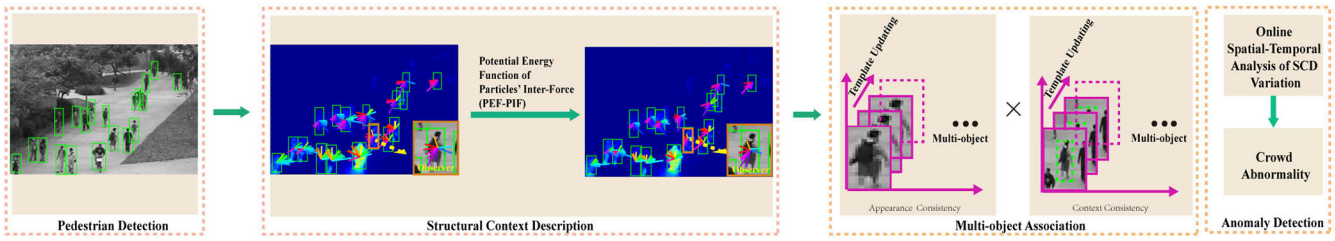



Fig. 2. Diagram of the proposed method. Given a video sequence, the individuals in every frame are first extracted by a pedestrian detection algorithm. Then their structural context descriptors (SCD) are computed for further visual contextual cueing of the crowd. After that, a robust 3-D DCT multi-object tracker is proposed to seek the stable individuals (named as observers) in every frame. Based on the spatial-temporal SCD variation analysis of these observers, the crowd abnormality is detected.

## II. Related Works

According to the clue types for defining crowd abnormality, recent approaches [8], [13]–[15], [18], [22], [29]–[34] of anomaly detection for crowd scenes can be categorized into two classes.

1) Trajectory-based techniques. The abnormal trajectories are prone to show much lower occurring frequency than the normal ones.
2) Motion-based techniques. The abnormal crowd has dramatic motion patterns compared with the normal one.

For the first category, by learning some knowledge of trajectories obtained from the normal situation, the abnormal trajectory is determined according to the learned rules [15], [29], [33]–[35]. For example, in Cheng and Hwang's work [15], by resolving the occlusion and object segmentation error via adaptive particle sampling and Kalman filtering, the reliable trajectory types were obtained. The abnormal event was then localized by the trajectory classification. In addition to the object-based trajectory extraction, some techniques exploit the trajectory from particle or feature point level. For example, Wu *et al.* [8] modeled the abnormal crowd patterns by utilizing chaotic invariant features of lagrangian particle trajectories. But it needs exhaust tracking for each representative particle. Cui *et al.* [14] extracted the normal/abnormal crowd patterns by tracking the interest points to calculate the interaction energy potentials (IEP), which explicitly exploited the relationships among a group of people. By analyzing the feature representation of different patterns, the abnormality was declared by SVM classifier. The most direct inspiration of this paper is the approach for crowd structure exploitation proposed by Ge *et al.* [22]. By tracking each individual robustly, the crowd groups were discovered by analyzing the relationships of trajectories. Although these trajectory-based methods have explicitly high-level semantics for defining abnormality, they are always infeasible and computationally expensive for tracking each individual.

As for the second category, motion patterns are usually explored by OF variation [4], [6], [7], [13], [32] or pixel/blob change [17], [18], [36]. Because of the high density of the crowd, motion pattern-based methods recently hold the main part in the crowd anomaly detection literatures. For example, Cong *et al.* [13] proposed a multiscale histogram of OF to represent the motion patterns for image sequences. By computing the reconstruction error with the trained sparse dictionary, the abnormality was detected by the motion patterns with large reconstruction cost. Mehran *et al.* [6] proposed a streakline technique to compute the crowd flow. By analyzing the obtained streak flows, the abnormal motion pattern was detected by a SVM predictor. Besides, the social force model (SF) is another hot technique proposed recently for motion modeling in abnormal crowd detection. Through the estimation of the particle OF with SF, the normal/abnormal motion patterns can be explicitly distinguished by Latent Dirichlet allocation (LDA) [7] or other analyzers [30]. Kim and Grauman [37] utilized the mixture of probabilistic principle component analysis (MPPCA) to model the local OF. Then, the modeled motion patterns were adopted to predict anomaly. Thida *et al.* [16] proposed a spatio-temporal *Laplacian eigenmap* to extract the crowd activities. It was achieved by learning a spatio-temporal variations of local motions in an embedded space. Li *et al.* [38] proposed an anomaly detection method which was constructed by a mixture of dynamic texture (MDT) model. MDT was subsequently updated to hierarchical MDT (H-MDT) [39] by combining spatial normalcy implemented by a center-surround discriminant saliency detector and a hierarchical model. Wu *et al.* [32] introduced the potential destinations and divergent centers to characterize the crowd motion in both the presence and absence of escape events. These motion-based anomaly detection methods usually need to exhaustively sample image patches in every frame. By analyzing the temporal appearance variation in these patches, the crowd context is extracted. Therefore, this procedure also include high computational cost.

While the main choice of crowd anomaly detection concentrates on modeling motion patterns, one universal limitation of the two categories is that labeled data should be available to train the normal/abnormal pattern. However, this assumption is difficult to be satisfied in practical applications.

## III. Structural Context Description

To detect crowd abnormality, the first important issue is crowd structure representation. Since the crowd consists of non-isolated individuals who have relations with each other [19], [40], its structure can be consequently explored from these connections. To this end, a novel SCD is proposed to represent the structure of each individual. In order to adapt to more kinds of anomalies, this paper also proposes a novel SHOF to adaptively select adequate motion property,

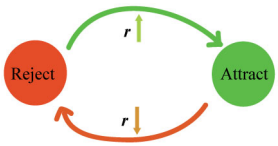



Fig. 3. State shifting of the potential energy between two particles.

such as motion magnitude and direction, when facing different crowded scenes.

Our assumption behind the SCD representation is that the individuals demonstrating large behavior difference with their surroundings are highly probable to be abnormal, and these large behavior difference should be manifested to make the anomaly detection easier. The discrepancy between the examined target and its surroundings is measured by an inconsistency weight. The larger the weight, the more distinct the target's behavior with its surroundings. However, according to the psychological experiments in [21], the layout of targets in an invariant configuration can be localized and discriminated more effectively than in a variable one, which is termed as the contextual cueing effect. Since the abnormality implies novel configuration, its detection is more difficult. For a more effective crowd contextual cueing, we thus lower the inconsistency weights of the individuals having little behavior difference and enlarge the ones with large behavior difference. Through this strategy, the large behavior differences are more salient and the irregularity of the crowd behavior is easier to be detected. For detailed implementation, the PEF-PIF in solid-state physics [23] is originally introduced. In the following, the PEF-PIF model is presented firstly, and then the generation of SHOF is given followed by the SCD computation inspired by PEF-PIF model.

### A. Potential Energy Function of Particle's InterForce

In solid physics, the potential energy of two particles represents their linking degree. The fundamental description can be expressed as

$$U(r) = \frac{a}{r^m} - \frac{b}{r^n} \tag{1}$$

where $U(r)$ denotes the potential energy between two particles, $r$ is their Euclidean distance, and $a$, $b$, $m$, and $n$ (generally, $m > n$) are the empirical constants. In (1), the first term represents the rejecting potential energy field, and the second term is the attracting potential energy field. The two particles demonstrate rejecting state when $r$ is small, and attracting state when $r$ is large. This kind of characteristic is visualized in Fig. 3.

For computing the force of two particles, combining power (CP) is defined. The farther the two particles are from each other, the weaker CP is. In fact, CP is the negative deviation of $U(r)$

$$f(r) = -\frac{dU(r)}{dr} = \frac{ma}{r^{m+1}} - \frac{nb}{r^{n+1}}. \tag{2}$$

The relations between $r$ and CP are illustrated in Fig. 4, as well as the variants of CP, where the constants $a$, $b$, $m$, $n$

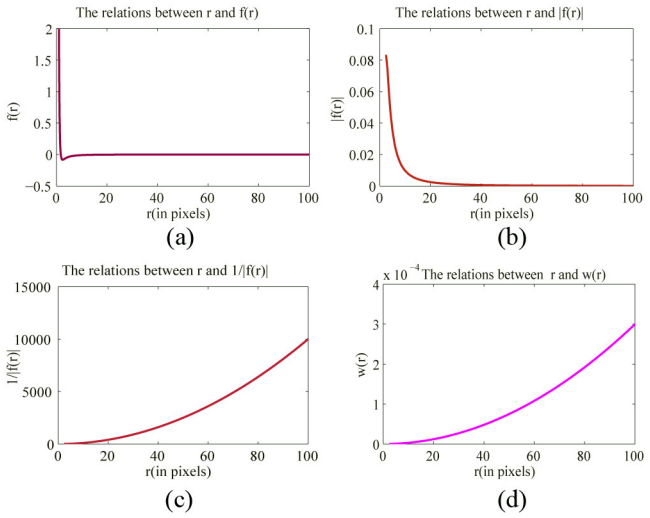



Fig. 4. (a) Relation curve of $r$ and $f(r)$ within the interval [0,100]. (b) Relation curve of $r$ and $f(r)$ within the interval $[\sqrt{2}, 100]$. (c) Relation curve of $r$ and $1/|f(r)|$ within the interval $[\sqrt{2}, 100]$. (d) Relation curve of $r$ and $w(r)$ within the interval $[\sqrt{2}, 100]$. The constants $a$, $b$, $m$, and $n$ in (1) are experimentally set as 1, 1, 3, and 1.

are experimentally set as 1, 1, 3, and 1. In Fig. 4(a), it can be seen that the curve achieves the minimum when $r = \sqrt{2}$ pixels. With the increase of $r$ ($r \geq \sqrt{2}$), the absolute value of $f(r)$ decreases gradually. This can be seen clearly from the replotted curve shown in Fig. 4(b). However, in practical applications, the distance between two individuals is obviously larger than $\sqrt{2}$ and the corresponding trend of $f(r)$ cannot meet the needs. This is because in the abnormal detection, we hope the increase of the variable will result in a proportional output. Therefore, we replace the function with the reciprocal of $f(r)$. It can be validated by the characteristic shown in Fig. 4(c). Based on it, we further normalize the reciprocal of $|f(r)|$ by

$$w(r) = \frac{1}{|f(r)|} \Big/ \left( \int_{\sqrt{2}}^{r} \frac{1}{|f(r)|} dr \right), \quad r \in \left[ \sqrt{2}, \infty \right] \tag{3}$$

where $w(r)$ specifies the linking weight of two particles with the distance $r$. After the weighting for different $r$, the relation curve of $w(r)$ and $r$ is visualized in Fig. 4(d).

### B. SHOF Generation

In order to construct the SCD, the motion difference of individuals should be efficiently computed. As for the motion property, OF [41] is utilized to characterize the motion of each individual. Since the output of pedestrian detection is a rectangle bounded region, the histogram of flow (HOF) [42] is calculated as the motion statistics where each bin of HOF represents the direction of OF and the value in each bin is proportional to the magnitude of OF. We find that the motion magnitude and the direction maintain consistent for a specific individual as shown in Fig. 5(a), which makes that the maximum of HOF can represent the magnitude property of individual. Considering that the anomaly definition in different crowded scenes may be different, such as magnitude inconsistency and direction inconsistency, for measuring the

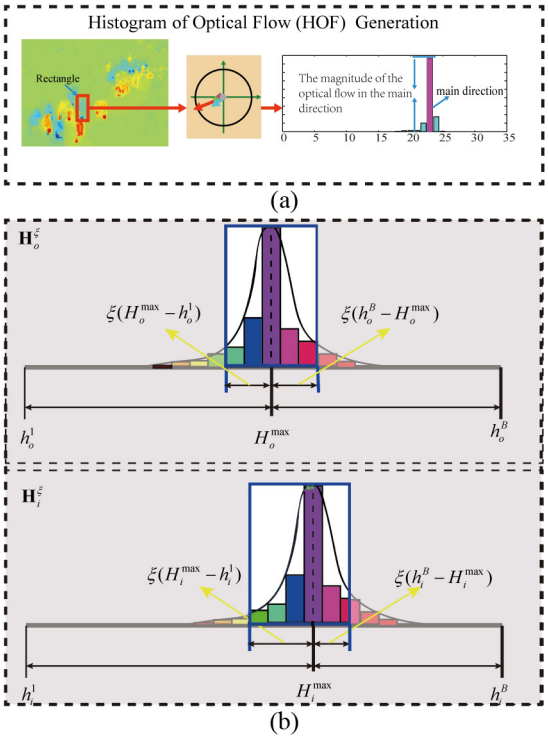



Fig. 5. (a) Generation of the histogram of optical flow (HOF). From the figure, it is clear that the pixels's motion directions are highly consistent within a single individual. (b) Diagram of the selective histogram of optical flow (SHOF). The parameter $\xi$ determines the range of HOF selected to conduct difference computation. Note that $\xi$ narrows or widens the range of SHOF in the same ratio, which can guarantee the shape of the SHOF be most similar to original HOF. When $\xi = 0$, only the bin with maximum is selected, and $\xi = 1$ for a whole HOF selection.

magnitude inconsistency of crowd, it needs to shield the influence of direction, and for direction inconsistency, the influence of magnitude of crowd should be avoided in some times.

To this end, this paper proposes a SHOF to represent the motion property of individuals. Actually, SHOF is a limited HOF by a parameter $\xi$ which determines a range of the HOF that needs to be used for motion difference computation, where $\xi$ is learned by several normal frames in the crowded scenes. The diagram of SHOF is demonstrated in Fig. 5(b). For each individual, we calculate its motion difference with the surroundings according to their SHOF. Assume the SHOF of the examined target to be $\mathbf{H}_o^\xi$ and $\mathbf{H}_i^\xi$ the SHOF of the $i^{th}$ surrounding individuals. The difference $\Delta f$ is computed by the $\chi 2$ distance, which is denoted as

$$\Delta f_i = \chi^2\left(\mathbf{H}_o^\xi, \mathbf{H}_i^\xi\right) \tag{4}$$

where $\chi^2(\mathbf{h}_1, \mathbf{h}_2) = \frac{1}{2}\sum_{i=1}^{B}\frac{|\mathbf{h}_{1;i}-\mathbf{h}_{2;i}|^2}{\mathbf{h}_{1;i}+\mathbf{h}_{2;i}}$ and $B$ denotes the number of histogram bins. All the motion difference between the examined individual and others is denoted as $\Delta\mathbf{f} = \{\Delta f_1, \Delta f_2, \ldots, \Delta f_M\}$ which indicates different motion property (including magnitude and direction) difference when giving different $\xi$, where $M$ indicates the number of individuals around the examined one.

It can be seen from Fig. 5(b) that the smaller $\xi$ is, the narrower the range of SHOF becomes, while the larger $\xi$ is, the wider the SHOF range is. With the limitation of $\xi$, when the magnitude inconsistency of individual is more important for anomaly detection, $\xi$ can be set as small as possible to narrow the range of SHOF to the maximum of HOF because many bins except for the one with maximum are abandoned. On the contrary, when moving direction of individual is the main aspect to determine the anomaly, $\xi$ can be set large to expand the range of SHOF to contain more bins which represents the direction of OF. By this strategy, this paper can be applied for more kinds of crowded scenes with different anomalies by adaptively abandoning or containing more bins of HOF.

Based on the assumption that, in normal situations, the motion property difference between individuals should be consistent, not only for the crowd with magnitude anomaly, but also the crowd with direction anomaly. The optimal $\hat{\xi}$ is learned by the motion difference of individuals in the normal video frame, and computed as

$$\hat{\xi} = \arg\min_{\xi} Var(\Delta\mathbf{f}) \tag{5}$$

where $Var(\cdot)$ is the variance calculation. The solving of (5) can be fulfilled by searching the whole range [0,1] of $\xi$ from 1 to 0 with 0.1 interval. Because the beginning of the video sequence is usually with a normal situation, $\hat{\xi}$ can be learned with several frames with these normal frames, and when $\hat{\xi}$ is generated, it can maintain unchanged for the subsequent video frames.

### C. SCD Computation

With this motion difference, the inconsistency weight can be obtained according to (3). However, this equation has to be modified to adapt to the actual problem, because if the linking weight is 1, $\Delta f$ tends to be $\infty$, which is practically impossible in this paper. When computing the surrounding structure of the examined individual, the reasonable consideration is that all of the other individuals in together contribute to its whole contextual structure. Therefore, the normalization process should be built on the number of the surrounding individuals, and the detailed expression is revised to

$$w(\Delta f_i) = \frac{1}{f\left(Z\Delta f_i\right)} \Big/ \left(\sum_{i=1}^{M}\frac{1}{f\left(Z\Delta f_i\right)}\right) \tag{6}$$

where $Z$ is a constant to enlarge $\Delta f$ to an adequate range that (3) can be used [$\Delta f$ acts like $r$ in (3)], $w(\Delta f_i)$ denotes the linking weight between the examined individual and the $i^{th}$ neighbor, and $M$ is the number of neighbors around it.

After computing every individual's behavior difference with its surroundings, the proposed SCD is constructed and denoted as $\{\mathcal{W}, \mathcal{F}\}$, where $\mathcal{W} = \{\mathbf{W}_k\}_{k=1}^{M+1}$, $\mathbf{W}_k \in \mathbb{R}^{1\times M}$ is the weight vector of the $k^{th}$ individual which can be computed by (5), and $\mathcal{F} = \{\mathbf{F}_k\}_{k=1}^{M+1}$, $\mathbf{F}_k \in \mathbb{R}^{4\times M}$ specifies the corresponding behavior feature vector of the $M$ neighbors, whose column components represent the max, min, mean, and the variance of motion energy in the individual bounding box. The efficiency of PEF-PIF is visualized in Fig. 6.

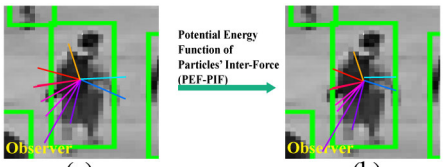



Fig. 6. Visualization of structural context descriptor (SCD). Each target's connection with the surrounding individuals is represented by a clutter of vectors with different color. The length is proportional to the inconsistency weight and the direction complies with the neighbor's location. (a) Weight of each connection is computed by the motion difference. (b) Weight of each connection is calculated by the introduced PEF-PIF. From the figure, it is clear that the weights are adjusted by making the large weights more manifest and small weights less effective.

## IV. 3-D DCT-Based Multiple Object Tracker

With the computed SCD for each individual in the crowd, the following work is to detect anomalies by analyzing the spatial-temporal SCD variation. However, the SCD variation between two frames needs to be built on the same target. For this purpose, the multiple targets in different frames should be associated. A straightforward strategy is to design a multi-object tracker. However, although many multi-object trackers [43], [44] are proposed, they are difficult to be implemented in the crowd anomaly detection because of the high density of the crowd, frequent occlusion and appearance/illumination change. Besides, these multi-object trackers are computationally expensive.

Fortunately, we find that the normal/abnormal can be judged alternatively by single individuals in the crowd. Because our purpose is to identify the abnormality, instead of tracking every target, we can fulfill this task by only employing the stable ones called observers. For these individuals, occlusion and appearance/illumination change hardly occur and they can be well tracked. By analyzing the observers' temporal SCD variations, the abnormality can be robustly detected. Therefore, different from the conventional multi-object trackers, this paper designs a new and efficient multi-object tracker to only seek the stable observers.

In order to design the robust multi-object tracker, we employ a newly proposed 3-D DCT model [24] with an excellent ability of incremental analysis. The designed 3-D DCT multi-object tracker has three components: compact 3-D DCT template representation, multi-target association, and incremental template updating. Each component is described sequentially as follows.

### A. Compact 3-D DCT-Based Object Representation

For the target association in a video sequence, we adopt to treat the frames as a 3-D volume by concatenating them. Then the self-correlation of the newly observed target sample with the previously collected target sample set is incrementally evaluated. To this end, the 3-D DCT [24] is utilized as a tool to fulfill this task.

Given a video sequence, assume the previously collected target sample set can be denoted as $(s_{\mathbf{III}}(x,y,z))_{N_1\times N_2\times N_3}$, where $N_1$, $N_2$ are the sample's width and height, and $N_3$ is the number of samples in the target sample set. The new target sample in the next frame is denoted as $(n(x,y))_{N_1\times N_2}$ (abbreviated as $n$ for short in the following description). The concatenated target sample set is specified as $(s'_{\mathbf{III}}(x,y,z))_{N_1\times N_2\times(N_3+1)}$, where the $(N_3+1)^{th}$ frame is concatenated to the end of the previous target sample set. According to the 3-D DCT [24], $(s'_{\mathbf{III}}(x,y,z))_{N_1\times N_2\times(N_3+1)}$ can be represented as

$$\begin{aligned}\mathcal{S}' &= \mathbf{C_{III}} \times_1 \mathbf{D}_1^T \times_2 \mathbf{D}_2^T \times_3 (\mathbf{D}')_3^T,\\ \mathbf{C_{III}} &= \mathcal{S}' \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 (\mathbf{D}')_3\end{aligned} \tag{7}$$

where $\mathcal{S}'=(s'_{\mathbf{III}}(x,y,z))_{N_1\times N_2\times(N_3+1)}$, $\mathbf{C_{III}} \in \mathbf{R}^{N_1\times N_2\times(N_3+1)}$ represents the 3-D DCT coefficient matrix, and $\times_m$ is the mode-$m$ product defined in tensor algebra [45]. $\mathbf{D}_1 = (a_1(o,x))_{N_1\times N_1}$ is a cosine basis matrix whose entries are represented as

$$a_1(o,x) = a_1(o)\cos\left(\frac{\pi(2x+1)o}{2N_1}\right). \tag{8}$$

$\mathbf{D}_2 = (a_2(p,y))_{N_2\times N_2}$ is a similar cosine basis matrix whose entries are specified as

$$a_2(p,y) = a_2(p)\cos\left(\frac{\pi(2y+1)p}{2N_2}\right) \tag{9}$$

and $(\mathbf{D}')_3 = (a'_3(q,z))_{(N_3+1)\times(N_3+1)}$ is a different cosine basis matrix whose entries are denoted as

$$a'_3(q,z) = \begin{cases}\sqrt{\frac{1}{N_3+1}}, & \text{if } q=0;\\ \sqrt{\frac{2}{N_3+1}}\cos\left(\frac{\pi(2z+1)q}{2(N_3+1)}\right), & \text{otherwise}\end{cases} \tag{10}$$

where $o \in \{0,1,\ldots,N_1-1\}$, $p \in \{0,1,\ldots,N_2-1\}$, $q \in \{0,1,\ldots,N_3-1\}$, and $a_k\,(o/p/q, x/y/z)$ is defined as

$$a_k(o/p/q, x/y/z) = \begin{cases}\sqrt{\frac{1}{N_k}}, & \text{if } o/p/q=0;\\ \sqrt{\frac{2}{N_k}}, & \text{otherwise.}\end{cases} \tag{11}$$

According to the properties of 3-D DCT, the larger the values $(o,p,q)$ are, the higher frequency the corresponding component of $\mathbf{C_{III}}$ encodes. Depending on the values of $(o,p,q)$, the work [24] compresses $\mathbf{C_{III}}$ by removing the high-frequency coefficients usually sparse (e.g., texture clue) and maintaining the low-frequency ones that are relatively dense (e.g., mean value). Therefore, the compact 3-D DCT object representation $\mathbf{C_{III}}$ is modified as

$$\mathbf{C}^*_{\mathbf{III}} = \mathcal{S}^* \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 (\mathbf{D}')_3 \tag{12}$$

where $\mathcal{S}^* = (s^*_{\mathbf{III}}(x,y,z))_{N_1\times N_2\times(N_3+1)}$ is the approximation of $\mathcal{S}'$, representing the corresponding reconstructed image sequence of $\mathcal{S}'$. Based on it, a reconstruction error representing the loss of low-frequency components is introduced, which is defined as

$$e = ||n - s^*_{\mathbf{III}}(:,:,N_3+1)||^2. \tag{13}$$

For a new target sample, its consistency likelihood with the target sample set can be measured by

$$\mathcal{L} = \exp\left(-\frac{1}{2\lambda}e\right) \tag{14}$$

where $\lambda$ is the scaling factor, and is set as 0.1 in this paper. This likelihood measurement is utilized both for target association in Section IV-B and the modeling of incremental template updating described in Section IV-C.

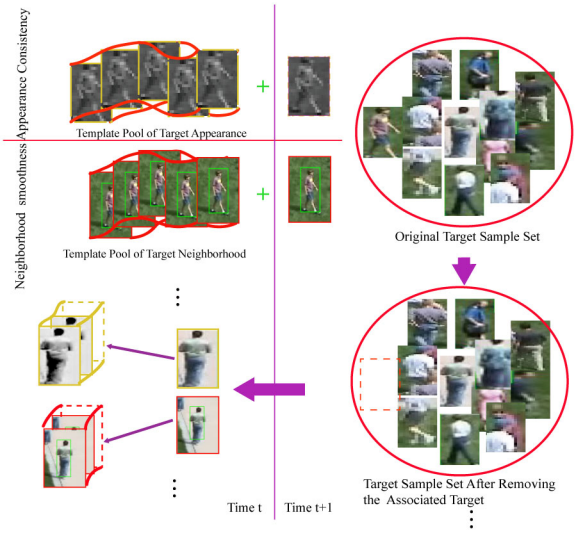



Fig. 7. Illustration of multi-target association. For each template pool, the most related target sample is associated according to the appearance consistency and neighborhood smoothness. Then the associated target sample is removed from the original target sample set to avoid repeated association. After that, the next template pool is associated similarly.

TABLE I
SOME MATHEMATICAL SYMBOLS IN THE TARGET ASSOCIATION

| Symbols | Meaning |
|---|---|
| $t$ | time index |
| $M$ | Number of target samples in each frame |
| $\{\mathcal{T}_T^f\}_{f=1}^F$ | $F$ template pools of target appearance |
| $\{\mathcal{T}_C^f\}_{f=1}^F$ | $F$ template pools of target neighborhood |
| $\{n_{t+1}^i\}_{i=1}^M$ | $M$ new target samples at time $t$+1 |
| $\{nc_{t+1}^i\}_{i=1}^M$ | $M$ new neighborhood samples at time $t$+1 |

### B. 3-D DCT-Based Multi-Target Association

Since the 3-DDCT can effectively represent the video sequence, it is utilized to associate the targets in different frames. Our task now becomes to correspond each newly detected pedestrian with a previously constructed template pool (target set). However, the targets detected by pedestrian detection algorithm [25] might change in different frames. Even the same target may have different scale and appearance. These factors increase the difficulty of effective association. In order to strengthen the accuracy of the target association, we consider not only the appearance consistency, but also the target neighborhood. The flowchart is illustrated in Fig. 7, which mainly contains two constraints: the appearance consistency and the neighborhood smoothness. Before detailed description, some mathematical notations are firstly presented in Table I.

*1) Appearance Consistency:* The assumption is that the new target should resemble its corresponding template pool in appearance. Each target appearance $n_{t+1}^i$ is compared with each template pool of appearance $\mathcal{T}_T^f$. Then the reconstruction error $e_{t+1}^{i,f}$ is computed by (13). The consistency of the $i^{th}$ target appearance with $f^{th}$ template pool is denoted as $\mathcal{L}_{T_{t+1}}^{i,f} = \exp(-\frac{1}{2\lambda} e_{t+1}^{i,f})$. The larger the value is, the more consistent the appearance is with the template pool.

*2) Neighborhood Smoothness:* Apart from the appearance consistency, the neighborhood smoothness is also a constraint for target association. The neighborhood of the target is represented by a surrounding rectangular region centered in the target. The smoothness of the $i^{th}$ target neighborhood with that of the $f^{th}$ template pool is denoted as $\mathcal{L}_{C_{t+1}}^{i,f}$. The inferring strategy is the same as the appearance consistency.

With the above two constraints, the target sample most related to the $f^{th}$ template pool is defined as

$$\overline{n}_{t+1}^f = \arg\max_{n_{t+1}^i} Normalize\left(\mathcal{L}_{T_{t+1}}^i \cdot \mathcal{L}_{\mathcal{C}_{t+1}}^i\right) \tag{15}$$

where $Normalize(\cdot)$ is the function normalizing the $\{\mathcal{L}_{T_{t+1}}^i \cdot \mathcal{L}_{\mathcal{C}_{t+1}}^i\}_{i=1}^M$ into [0, 1]. If $Normalize(\cdot) > 0.8$, the examined target is defined as an observer; otherwise, it is discarded and fails in the target association. At the same time, the corresponding template pool is updated by

$$\begin{aligned} \mathcal{T}_T^f &= Concatenate\left(\mathcal{T}_T^f, \overline{n}_{t+1}^f\right) \\ \mathcal{T}_C^f &= Concatenate\left(\mathcal{T}_C^f, \overline{nc}_{t+1}^f\right) \end{aligned} \tag{16}$$

where $Concatenate(\cdot)$ is the function concatenating the obtained new target sample with its related template pool.

After association, each target sample and its contextual sample are added to its corresponding template pool and the template pool will be updated at the same time. But for computational efficiency, the size of the template pool cannot increase without a limit. We therefore set a maximum threshold so that the redundant one will be ruled out if necessary. The detailed screening strategy is discussed in Section IV-C.

### C. Incremental Template Updating

As for the template updating (including template pools of target appearance and neighborhood), there are two aspects to be balanced.

1) The reliability with the previously constructed template pool. It hopes that the updated template maintains more information contained in the previous frames.
2) The adaptability for the dynamic scene. It on the contrary desires the updated template changes adaptively with the dynamic scene. In fact, there are no perfect criterions to balance them.

But in general, the templates in the beginning of the tracking have no dramatic changes in appearance or with surrounding environment. But in later times, the targets are prone to change. Therefore, this paper designs a two-stage updating strategy. In the beginning of the tracking, reliability is selected as the main criterion, and then the adaptability is advocated in the subsequent tracking.

*1) Reliability Preservation:* Assume the capacity of each template pool is fixed as $K$. For the new $f^{th}$ template pool $\mathcal{T}_f$ obtained by (16), if the size is smaller than $K$, $\mathcal{T}_f$ should not be changed. But if its size reaches $K$+1, the most dissimilar target should be removed from the template pool. For this purpose, we iteratively evaluate the similarity between each target sample $n_k^f, k \in [1, K]$ within $\mathcal{T}_f$ and the remaining $s_{\mathbf{III}}^f(:, :, 1{:}K)$ (which is actually $\{\mathcal{T}_f - n_k^f\}$) according to the reconstruction

error in Section IV-B. The removed target $\overline{n}^f$ is chosen by

$$\overline{n}^f = \arg\min_k \exp\left(-\frac{1}{2\lambda}||n_k^f - (s_{\mathbf{III}}^f)^*(:,:,K)||^2\right) \quad (17)$$

where $\left(s_{\mathbf{III}}^f\right)^*(:,:,1{:}K)$ is the obtained 3-DDCT model.

*2) Adaptability Preservation:* As for the adaptability preservation, it hopes that the template pool puts more emphasis on the newly observed sample than the historical ones. This makes the template pool adapt to the dynamic scene efficiently. The similarity measurement between each target sample and the remainder ones of the template pool is the same to the reliability preservation. However, the target sample $\overline{n}^f$ need to be removed is on the contrary defined as

$$\overline{n}^f = \arg\max_k \exp\left(-\frac{1}{2\lambda}||n_k^f - \left(s_{\mathbf{III}}^f\right)^*(:,:,K)||^2\right). \quad (18)$$

It is worth noting that the strategy of updating target contextual template pool is the same as the one mentioned above.

## V. Online Anomaly Detection

Most methods proposed recently for anomaly detection need labeled data for normal/abnormal definition. However, this requirement is difficult to be satisfied in actual applications. Considering the abnormal frame usually presents a very different state with the previous motion context, such as the "evacuation" in global abnormality and "sudden motion change" in local abnormality, this paper designs an online anomaly detection strategy by self-learning the normal motion patterns represented by proposed SCD. In order to efficiently detect anomalies, it is conducted from frame-level and pixel-level.

### A. Frame-Level Anomaly Detection

The frame-level abnormality is defined by comparing the corresponding observers between frames. The detection results label the abnormal frames as output. Suppose the number of observers is $\beta$. The SCD of the $\gamma^{th}(\gamma = 1, 2, \ldots, \beta)$ observer at time $t$ is $\{\mathbf{W}_t^\gamma, \mathbf{F}_t^\gamma\} \in \{\mathcal{W}_t, \mathcal{F}_t\}$ and $\{\mathbf{W}_{t+1}^\gamma, \mathbf{F}_{t+1}^\gamma\} \in \{\mathcal{W}_{t+1}, \mathcal{F}_{t+1}\}$ at time $t+1$. The SCD variation of the observers is computed by the EMD [28] which can adapt to the distributions with different dimensions. The difference between adjacent frames is defined as

$$\begin{aligned} d_\gamma^{t+1} &= 1 - \mathrm{EMD}\left(\mathbf{W}_t^\gamma, \mathbf{W}_{t+1}^\gamma, \mathbf{F}_t^\gamma, \mathbf{F}_{t+1}^\gamma\right) \\ AD_{frame}^{t+1} &= 1 - \left(\sum_{\gamma=1}^{\beta} d_\gamma^{t+1}\right)/\beta. \end{aligned} \quad (19)$$

By this mean, the anomaly is online declared by the average value $AD_{frame}^{t+1} > 0.5$. The threshold 0.5 is a reasonable choice for distinguishing normal/abnormal crowd behavior according to the following experiments. It is worth noting that if $AD_{frame}^{t+1}$ of the $\gamma^{th}$ observer is smaller that 0.5, $\mathbf{W}_t^\gamma \leftarrow \mathbf{W}_{t+1}^\gamma$, $\mathbf{F}_t^\gamma \leftarrow \mathbf{F}_{t+1}^\gamma$.

### B. Pixel-Level Anomaly Detection

Pixel-level abnormality is defined on the detected pedestrians of rectangular pixels at current frame. The output is the labeled abnormal regions in each frame. By analyzing its motion difference from the surrounding ones, abnormality can be identified on each individual. The pixel-level anomaly is conducted when the examined frame is judged as an abnormal frame. For computing the motion difference, $\chi^2$ distance is utilized. To be specific, assume the number of the individuals at time $t+1$ be $M+1$. The motion difference of the $m^{th}$ individual $d_m^{t+1}$ with other $M$ individuals is defined as

$$d_m^{t+1} = \frac{1}{M^2}\sum_j^M \sum_{k=1}^M \frac{\left(\mathbf{W}_{t+1}^m(k) - \mathbf{W}_{t+1}^j(k)\right)^2}{\mathbf{W}_{t+1}^m(k) + \mathbf{W}_{t+1}^j(k)} \quad (20)$$

where $k$ is the component index of the weight vector $W$, and $j$ is the index of the surrounding individuals. The abnormal degree is defined by

$$AD_{pixel}^{t+1} = d_m^{t+1} / \sum_{m=1}^{M+1} d_m^{t+1}. \quad (21)$$

The abnormal individual is declared by $AD_{pixel}^{t+1} > 0.5$. It means that if no less than half of the individuals think the $m^{th}$ individual is an abnormal, the $m^{th}$ individual is abnormal. With the detected abnormal individual, the pixel-level abnormal result is obtained by filling the rectangle bounding box of the abnormal individual. Fig. 8 gives an example of pixel-level abnormality detection result.

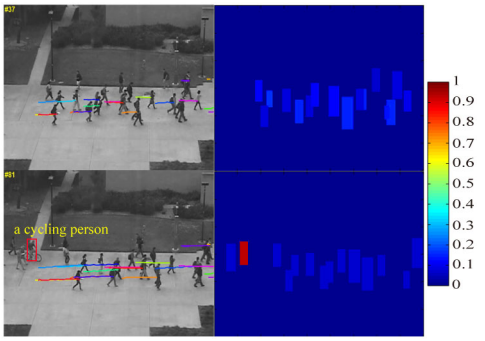



Fig. 8. Example of pixel-level abnormality detection result. Top row: normal crowd; bottom row: abnormal crowd (a cycling person appears). For each row, the right image demonstrates the abnormal degree of each individual. The motion differences between the normal individuals and the cycling person are larger than the ones between the normal individuals themselves. The abnormal cycling person is marked by red color.

With the above introduction, the proposed anomaly detection algorithm is finally summarized in Algorithm 1.

## VI. Experiments and Discussion

### A. Datasets

To test the performance of the proposed method, this paper applies several publicly available datasets to evaluate it. They are explained as follows.

*1) UMN Dataset:* UMN dataset [46] is recorded and labeled by University of Minnesota to evaluate the ability of detecting global abnormality, such as sudden crowd evacuation. It contains 11 video sequences representing three different crowd scenarios, each of which begins with a normal activity behavior. The total frame number is 7740 and the resolution of each frame is 320×240. The original ground truths are given

**Algorithm 1** OADC-SA

**Parameter setting**.

**Input**: Video sequence.

**Method**:

1: Detect each individual in each frame.
2: Compute the SCD for each individual in different frame.
3: Track multiple objects to seek $\beta$ observers.
4: Online compute the SCD variation $d_{\gamma}^{t+1}$ of the $\gamma^{th}$ observer by (19), and obtain the frame-level abnormal degree $AD_{frame}^{t+1}$.
5: Compute motion pattern difference $d_{m}^{t+1}$ of the $m^{th}$ individual and its abnormal degree $AD_{pixel}^{t+1}$ respectively by (20) and (21).

**Output**: $AD_{frame}^{t+1}$, the frame-level abnormality is determined by $AD_{frame}^{t+1} > 0.5$. $AD_{pixel}^{t+1}$, the pixel-level abnormality is declared by $AD_{pixel}^{t+1} > 0.5$.

with the dataset. In addition, for efficiently detecting pedestrian in this paper, the resolution of each frame is enlarged to $640 \times 480$. Besides, we also generate our $\xi$ to judge the main attribute for determining anomaly for UMN dataset. Interestingly, the optimal $\xi$ is set to 0, which indicates that the magnitude inconsistency is the main attribute for anomaly determination.

*2) USCD Dataset:* USCD dataset [47] is adopted to test the ability of detecting local abnormality. The local abnormalities display: 1) unusual individuals in crowd (e.g., individuals on wheelchair) and 2) irregular behaviors in the present surroundings (e.g., people cycling or skating across walkways). USCD dataset contains two different scenarios named as *ped1* and *ped2*. In ped1 video set, there are 34 normal video sequences for training and 36 abnormal video sequences for testing. Only frame-level ground truth is available for ped1. As for ped2 video set, 16 normal video sequences and 14 abnormal video clips are contained within it. There are both frame-level and pixel-level ground truths for ped2. For each video sequence, it has 200 frames with a resolution $158 \times 238$ of ped1, and $360 \times 240$ of ped2. Because the proposed method is an online one, only the abnormal testing video sequences are adopted. Besides, since the anomaly detection in this paper focuses on the pedestrian detection, the video sequences with abnormalities of unusual car or small carts are rule out of the testing dataset. For detecting pedestrian effectively, each frame in USCD dataset is enlarged twice the size of its original scale. In this dataset, the optimal $\xi$ for ped1 dataset is the same as UMN, and $\xi = 0.9$ for ped2 dataset, which indicates the magnitude plays more important role in ped1 dataset. As for ped2 dataset, it also depends on the magnitude inconsistently, but the learned $\xi$ equals 0.9. It seems against to our underlining meaning of SHOF, that when $\xi$ is large, the main attribute for anomaly detection is motion direction. However, from the observation on the first frame of video sequences in ped2, all the individuals move at the same direction in the normal frames, and the motion difference between them maintain consistent no matter what $\xi$ is. In other words, whatever $\xi$ is, the experimental results of ped2 dataset will not be affected by $\xi$.

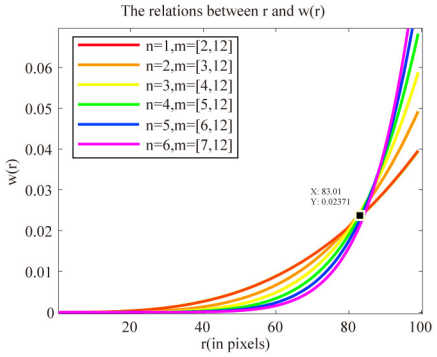



Fig. 9. Relations between $r$ and $w(r)$. Different $m$ and $n$ lead to a similar characteristics for weighting.

### B. Implementation Details

*1) Parameter Setting:* There are a few parameters to be set. The first one is the size of the template pool $K$ and the second one is the constants in (2). Once they are set, they maintain unchanged for all the sequences. For computational efficiency and robustness of the multiple object tracker, $K$ is empirically set as 6. The constants $a$, $b$, $m$, and $n$ in (2) are set as 1, 1, 3, and 1, respectively. The reason for this configuration is described as follows.

Based on the empirical consideration in physics, both $a$ and $b$ are set as 1. For a fixed $n$, different $m$ generates nearly the same characteristic. This is shown in Fig. 9. From the figure, the larger $n$ takes, the stronger the weighting effect (the curve is more steep). But for a larger $n$, the weights of $r < 83.01$ increase rather slow and the weights of $r > 83.01$ increase rather fast, which may cause the undistinguished normalized weights. Therefore, the constants $m$ and $n$ are set as 3 and 1, respectively.

Besides, the parameters in the state-of-the-art pedestrian detection algorithm [25] are set as default.

*2) Evaluation Criterions:* To evaluate the efficiency of the proposed method, both qualitative and quantitative criterions are utilized. For the qualitative evaluation, the detected frame shots are presented and analyzed subjectively. As for the quantitative criterions, there are four indexes. The first two are receiver operating characteristic (ROC) and area under ROC (AUC). For a better understanding of ROC, two terms should be introduced because ROC reflects the relationship of them.

1) *True positive rate (TPR):* the rate of correctly labeled frames.
2) *False positive rate (FPR):* the rate of incorrectly labeled frames.

They are defined as

$$TPR = \frac{TruePositive}{TruePositive+FalseNegative}$$
$$FPR = \frac{FalsePositive}{TrueNegative+FalsePositive}. \quad (22)$$

The third one is equal error rate (EER) for frame-level evaluation, which reports the frame percentage with the abnormal likelihood equal to 0.5. The fourth one is rate of detection (RD) for pixel-level evaluation, representing the detection rate at equal error point [38].


Fig. 10. Qualitative evaluation of the proposed method for the 11 sequences in UMN dataset. Top row represents the snapshots of the detected first abnormal frame in each sequence by the proposed OADC-SA. Bottom row shows the labels of the ground truth and the detection results respectively by the proposed method with (OADC-SA) and without (OADC) PEF-PIF. Among them, green represents the normal situation and the red specifies the abnormal situation.

TABLE II
FRAME-LEVEL AUC COMPARISON FOR ABNORMAL DETECTION IN UMN DATASET

| Method | AUC |
|---|---|
| OF [7] | 0.8400 |
| SF [7] | 0.9600 |
| CI [8] | 0.9900 |
| IEP [14] | 0.9892 |
| LA [18] | 0.9850 |
| SRC [13] | 0.9955 |
| H-MDT-CRF [39] | 0.9950 |
| OADC-SA | **0.9967** |

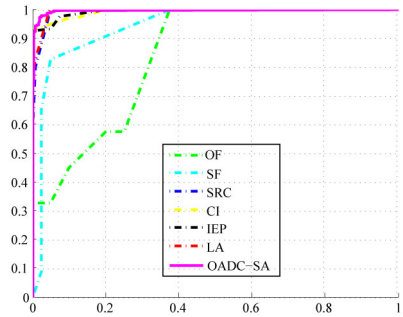



Fig. 11. Frame-level ROC comparison in UMN dataset.

*3) Competitors:* For UMN dataset, several methods representing the state-of-the-arts are selected. They are respectively the pure OF[7], SF [7], chaotic invariants (CI) [8], IEP [14], local aggregates (LA) [18], and sparse reconstruction cost (SRC) [13].

For USCD dataset, to the best knowledge of the authors, there are not trajectory-based anomaly detection methods implemented in this dataset. Therefore, the methods chosen here are all motion-based ones which represent the state-of-the-arts. They are the MDT model [38], SF [7], the MPPCA [37], Adam *et al.*'s work [48], SRC [13], spatial-temporal motion context (STMC) [49], and newly proposed Hierarchical MDTs with CRF filter (H-MDT-CRF) [39], respectively.

### C. Performance Analysis for Anomaly Detection

*1) Performance on UMN Dataset:* UMN dataset is collected to validate the performance of sudden evacuation event detection. According to Fig. 10, the detected results generated by OADC-SA are almost the same as the ground truth. This is also true for the quantitative evaluation of the frame-level AUC values in Table II and the ROC curves in Fig. 11. The reason of the superiority of the OADC-SA is that the other competitors all need labeled motion patterns for defining normal/abnormal prototypes. The procedure needs significantly different motion patterns, which may not detect the margin representing the transition between them. Therefore, they all may cause a delayed abnormal detection. As for OADC-SA, it is an online type which does not need any training data and is sensitive to the gentle structure change. In addition, through the structural context description, the abnormal individual with larger motion behavior difference to the surroundings is enlarged, which makes the motion behavior of normal and abnormal individual more distinctive, and anomaly detection easier. Thus, it can capture the tiny structure variation of the crowd, by which the abnormality can be predicted immediately.

*2) Performance on USCD Dataset:* Fig. 12 demonstrates the frame-level ROC comparisons for USCD dataset, and the AUC comparisons are shown in Table III. It can be seen that the proposed method is superior to the other state-of-the-arts. This is because the proposed method does not need any training data. For a certain video sequence, the normal motion patterns learned from itself are more adequate for predicting anomaly. However, the other methods depending on the motion patterns in the training data more or less are vulnerable to the margin between the normal and abnormal.

As ped1 dataset has not only the frame-level but also the pixel-level ground truth, we further present more pixel-level evaluation of ped1 dataset with other popular competitors. In Fig. 14 and Table IV, the AUC value (0.75) of our method demonstrates a little smaller than the newly proposed H-MDT-CRF (0.827). The reason is that H-MDT-CRF utilized a CRF filter to refine the detected pixel-level results, which rules out the false pixel around the individuals. However, our method takes all the pixels belonging to the marked rectangle region into computation. Therefore, the ROC curve shows a little weaker. But the proposed method is a more decisive one (EER = 0.09) and obtains a comparative *RD* = 0.74 with H-MDT-CRF (*RD* = 0.745). Meanwhile, from the typical abnormal frame shots in Fig. 13, our method outperforms the other competitive ones visually.

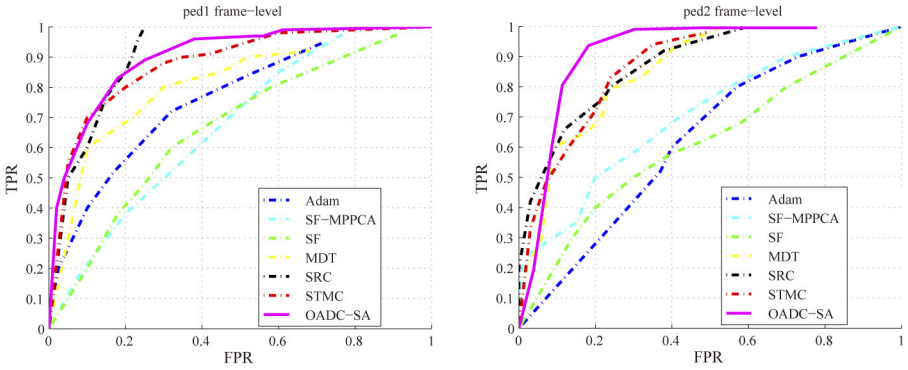



Fig. 12. Frame-level ROC comparisons in USCD ped1 and ped2 dataset.

TABLE III
FRAME-LEVEL AUC COMPARISON FOR ABNORMAL DETECTION IN USCD DATASET

| Method | ped1 | ped2 |
|---|---|---|
| Adam [48] | 0.650 | 0.630 |
| SF-MPPCA [37] | 0.590 | 0.710 |
| SF [7] | 0.670 | 0.630 |
| MDT [38] | 0.818 | 0.850 |
| SRC [13] | 0.860 | 0.861 |
| STMC [49] | 0.880 | 0.868 |
| OADC-SA | **0.910** | **0.925** |

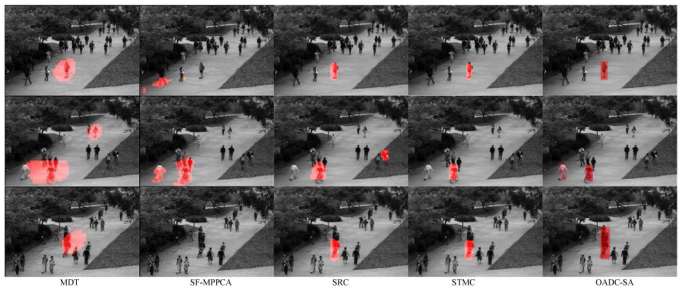



Fig. 13. Typical abnormal detection results generated by MDT [38], SF-MPPCA [37], SRC [13], STMC [49], and our OADC-SA in ped1 dataset. The detected pixel-level abnormality is marked by red color.

## D. Discussion

*1) Computational Complexity:* For the proposed method, the time is proportional to the crowd density, which can be proved from Fig. 15. The main time consumption contains two parts: SCD computation for every frame and target association via 3-D DCT. Assume the number of targets in the newly observed frame is $M$, and the target number in the previous frame is N. The computational complexity of SCD computation is $\mathcal{O}(M^2)$. The target association via 3-D DCT is $\mathcal{O}(N(N-1)/2)$. Therefore, the total computational cost in the approach is $\mathcal{O}(M^2)+\mathcal{O}(N(N-1)/2)$. Because the number of the targets in every frame is about 20 in this paper, the method is very fast.

To give a more fair comparison, the average time cost for judging a frame (to be normal/abnormal) is utilized. For MDT [38], the time cost is 25 s/frame on a standard platform with 3 GHz CPU and 2 GB RAM. With respect to the SRC [13], the average time cost for USCD dataset is 3.8 s and 0.8 s for UMN, executed on the same platform with MDT.

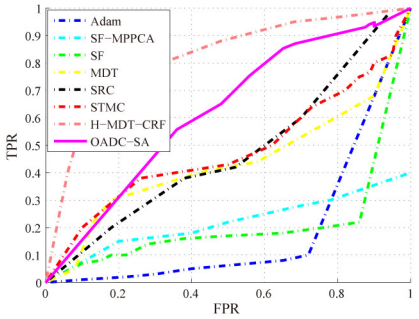



Fig. 14. Pixel-level ROC curves for ped1 in USCD dataset.

TABLE IV
PIXEL-LEVEL COMPARISON IN USCD PED1 DATASET

| Method | EER | RD | AUC |
|---|---|---|---|
| SF [7] | 0.36 | 0.41 | - |
| SF-MPPCA [37] | 0.32 | 0.18 | 0.213 |
| Adam [48] | 0.38 | 0.24 | 0.133 |
| SRC [13] | 0.19 | 0.46 | 0.461 |
| MDT [38] | 0.25 | 0.45 | 0.441 |
| STMC [49] | 0.23 | 0.47 | 0.471 |
| H-MDT-CRF [39] | 0.18 | **0.745** | **0.827** |
| OADC-SA | **0.09** | 0.74 | 0.75 |


Fig. 15. Efficiency evaluation of the proposed multi-object tracker. From the results, the time cost depends on the density of the crowd.

TABLE V
FRAME-LEVEL AUC COMPARISON BETWEEN OADC AND THE PROPOSED OADC-SA FOR UMN, PED1 AND PED2 DATASETS

| Method | UMN | ped1 | ped2 |
|---|---|---|---|
| OADC | 0.9661 | 0.65 | 0.71 |
| OADC-SA | 0.9967 | 0.91 | 0.925 |

STMC [49] takes about 1.2 s to judge a frame of USCD dataset on a platform with 4 GB RAM and 3 GHz CPU. H-MDT-CRF [39] only reports the average time cost 0.67 s on USCD ped1 dataset. As for Adam *et al.* [48], the authors claim that their method can run in real-time. For our method, it needs about 0.6 s to judge a frame of UMN and USCD datasets on a platform with 2 GB RAM and 2.93 GHz CPU without any code optimization. Therefore, from the efficiency mentioned above, the proposed method in this paper is the best one. In the future, through code optimization or GPU acceleration, the approach is probable to perform in real-time.

*2) Influence of PEF-PIF:* In this paper, we propose to use PEF-PIF model to adjust the inconsistency weight between individuals. Larger weight is magnified and smaller one is weakened. In order to prove the effectiveness of this principle, we conduct the experiments with and without this PEF-PIF adjustment (OADC and OADC-SA, respectively). The frame-level AUC comparative results are shown in Table V. It can be seen that the introduction of PEF-PIF makes the proposed method more feasible for anomaly detection. The physical meaning behind the success of introducing PEF-PIF is that PEF-PIF makes the large motion difference more salient and the small motion difference more uninteresting. Inspired by that, the anomaly with large motion context change is easier to be detected after introducing PEF-PIF.

*3) Difference With the Single 3-D DCT Tracker:* In order to show the originality of the proposed tracker, the difference between the single 3-D DCT tracker [24] and the proposed multi-object tracker is described as follows.

First, the candidate samples in this paper are generated by state-of-the-art pedestrian detection algorithm [25], instead of particle filter [24] which is computational expensive for multi-object tracker. Second, the target association strategy in [24] is based on discriminative classification which needs negative samples. However, the detected samples in this paper are all positive, which makes the discriminative framework infeasible in this paper. Therefore, we paper designs a simple yet efficient target association strategy considering both the appearance consistency and context consistence. Finally, the template updating strategy in [24] only considers the adaptability of the template pool. The template reliability is not treated. However, reliability can make the appearance of each target more discriminative in the crowd.

To further show the effectiveness and robustness of the proposed tracker, several more typical tracking results are displayed in Fig. 16. For each scenarios, two frames near the beginning and the end of the sequence are respectively shown with the observers' trajectories. From the figure, it is obvious that the computed trajectories are smooth and can adapt to the change of the scene. This greatly facilitates the abnormal detection.

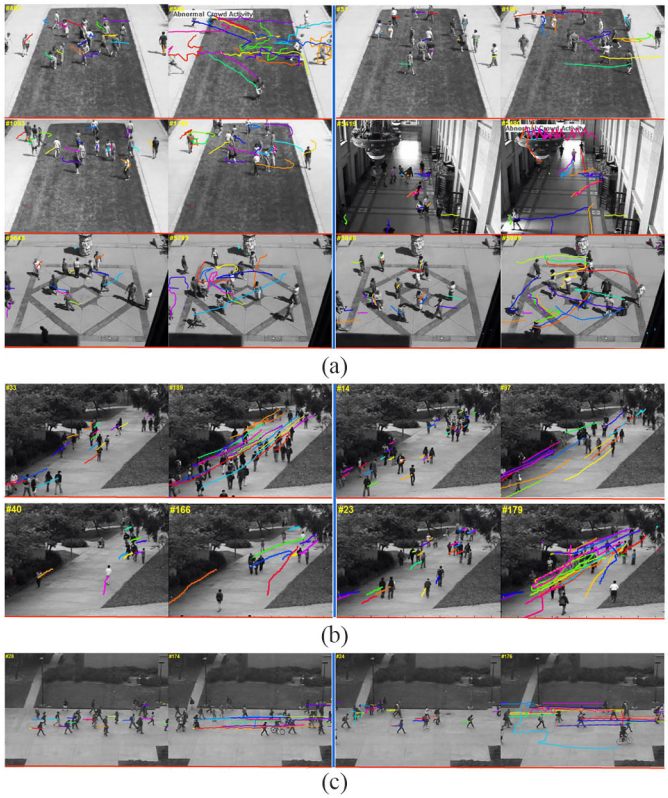



Fig. 16. Illustration of more typical tracking results. For each scenario, two frames near the beginning and end of the sequence are displayed as a group. The observers' trajectories are also demonstrated at the same time. From the tracking results, the trajectories are all smooth, which indicates that the designed multi-object tracker is robust to occlusion, appearance, and illumination change. These reliable trajectories make the further anomaly detection feasible. (a) Typical tracking results of sequences in UMN dataset. (b) Typical tracking results of sequences in USCD ped1 dataset. (c) Typical tracking results of sequences in USCD ped2 dataset.

*4) Adaptation and Potential Applications of the Approach:* This paper addresses the problem of anomaly detection in crowded scenes (frame-level anomaly, pixel-level anomaly). As for the crowded scenes with heavy occlusion, the success of the proposed method relies on two hypotheses: 1) it is impossible that every individual is occluded by the others and 2) the abnormal individuals should be visible or partially visible at least for pixel-level anomaly detection. Through observation, the former one is often the case in the crowd. The latter one is the fundamental condition for the pixel-level anomaly detection, not only for this paper, but also for the competitive methods in the literatures. If these two hypotheses can be established, we can say that the proposed method is effective for the crowded scenes with heavy occlusion. The reasons are described as follows.

1) In this paper, we find that the frame-level normal/abnormal can be judged alternatively by a single individual in the crowd by treating the individual as an observer. If some other individuals appear inconsistently with its historical observation, an abnormality occurs. Although the success of this approach depends on the accurate association of the observers, as long as at least

one observer is accurately matched at every association, the frame-level abnormality can also be detected by its SCD variation analysis.

2) With respect to the pixel-level abnormality, it only relies on the motion differences between each individual and others at current frame. If the abnormal individual is visible or partially visible, it can be detected by its significantly different motion pattern from others.

Since this paper exploits the crowd structure variations, the frames with severe structure motion context changes will be efficiently extracted. Therefore, the structure context analysis of this paper might provide some guidance for the video key frame extraction [50], [51] and video coding [52]. In addition, the selected key frames can also be helpful for video annotation [53]–[55].

## VII. Conclusion

This paper proposes an online anomaly detection method from crowd structure modeling. The visual structural context of the individuals is for the first time explored in this field. For this purpose, we originally introduces the PEF-PIF to construct the SCD contributed by a novel SHOF, which can effectively represent the relationship among individuals. In order to compute the SCD variation efficiently, a robust multi-object tracker is then designed to associate the targets in different frames. The proposed tracker introduces the excellent incremental ability of 3-D DCT and only limited number of targets need to be stably tracked. This makes the tracking method feasible for anomaly detection in crowd scenes with high density. By online spatial-temporal analysis of the SCD variation, the crowd abnormality is detected in the end. From the testing results on several popular datasets, the proposed method is superior to others representing the state-of-the-arts.

Our paper is tested only in the visible video sequences containing RGB channels. However, in severe weather conditions like foggy and rainy days, it takes difficulties to the proposed method, as well as the other competitive ones. We think one possible solution to this problem is to incorporate the multispectral clues that have different properties with the traditional visible spectrums. Besides, because of the individuals in the normal crowd always demonstrate consistent motion patterns, salient object in motion indicates that the object may be in exceptional. Therefore, in the future, we also want to introduce saliency detection method for RGB [56] or multispectral data [57], [58] into anomaly detection. This may compensate the limitations of existing methods. The future work is mainly toward these direction.

## References


[1] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1257–1272, Nov. 2012.

[2] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[3] H. Zhou and H. Hu, "Human motion tracking for rehabilitation—A survey," *Biomed. Signal Process. Control*, vol. 3, no. 1, pp. 1–18, 2008.

[4] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognit. Lett.*, vol. 33, no. 4, pp. 438–445, 2012.

[5] Y. Shi, Y. Gao, and R. Wang, "Real-time abnormal event detection in complicated scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Istanbul, Turkey, 2010, pp. 3653–3656.

[6] R. Mehran, B. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Greece, 2010, pp. 439–452.

[7] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 935–942.

[8] S. Wu, B. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 2054–2060.

[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–72, 2009.

[10] O. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.

[11] A. Patcha and J. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, 2007.

[12] P. Ronald, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Underst.*, vol. 108, no. 1, pp. 4–18, 2007.

[13] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, no. 1, pp. 1851–1864, 2013.

[14] X. Cui, Q. Liu, M. Gao, and D. Metaxas, "Abnormal detection using interaction energy potentials," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2011, pp. 3161–3167.

[15] H. Cheng and J. Hwang, "Integrated video object tracking with applications in trajectory-based event detection," *J. Vis. Commun. Image Represent.*, vol. 22, no. 7, pp. 673–685, 2011.

[16] 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. 215–230, Dec. 2013.

[17] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 2458–2465.

[18] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2112–2119.

[19] G. Mark, "Threshold models of collective behavior," *Amer. J. Sociol.*, vol. 83, no. 6, pp. 1420–1443, 1978.

[20] I. Biederman, "Perceiving real-world scenes," *Science*, vol. 177, no. 4043, pp. 77–80, 1972.

[21] M. M. Chun and Y. Jiang, "Contextual cueing: Implicit learning and memory of visual context guides spatial attention," *Cognit. Psychol.*, vol. 36, no. 1, pp. 28–71, 1998.

[22] W. Ge, R. Collins, and R. Ruback, "Vision-based analysis of small groups in pedestrain crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[23] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*. Orlando, FL, USA: Harcourt, 1976.

[24] X. Li, A. Dick, C. Shen, A. Hengel, and H. Wang, "Incremental learning of 3D-DCT compact representations for robust visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 863–881, Apr. 2013.

[25] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[26] J. Fang, Q. Wang, and Y. Yuan, "Part-based online tracking with geometry constraint and attention selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 854–864, May 2014.

[27] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, May 2008.

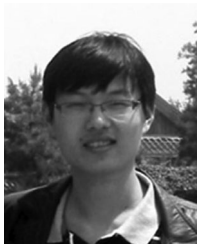
[28] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proc. IEEE Conf. Comput. Vis.*, Vancouver, BC, Canada, 2001, pp. 251–256.

[29] T. Xiang and S. Gong, "Incremental and adaptive abnormal behaviour detection," *Comput. Vis. Image Underst.*, vol. 111, no. 1, pp. 59–73, 2008.

[30] Y. Zhang, L. Qin, H. Yao, and Q. Huang, "Abnormal crowd behavior detection based on social attribute-aware force model," in *Proc. IEEE Conf. Image Process.*, Orlando, FL, USA, 2012, pp. 2689–2692.

[31] V. Jagannadan and J. M. Odobez, "Topic models for scene analysis and abnormality detection," in *Proc. IEEE Conf. Comput. Vis. Workshops*, Kyoto, Japan, 2009, pp. 1338–1345.
[32] S. 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] N. Anjum and A. Cavallaro, "Multi-feature object trajectory clustering for video analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1555–1564, Nov. 2008.
[34] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1544–1554, Nov. 2008.
[35] Y. Zhang, L. Qin, H. Yao, and Q. Huang, "Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition," in *Proc. IEEE Conf. Image Process.*, Melbourne, VIC, Australia, 2013, pp. 3572–3576.
[36] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1446–1453.
[37] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 2921–2928.
[38] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 1975–1981.
[39] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.
[40] D. E. Cartwright and A. E. Zander, *Group Dynamics Research and Theory*. Evanston, IL, USA: Row, Peterson, 1953.
[41] C. Liu *et al.*, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Electrical Engineering and Computer Science in Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.
[42] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1932–1939.
[43] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.
[44] C. Huang, Y. Li, and R. Nevatia, "Multiple target tracking by learning-based hierarchical association of detection responses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 898–910, Apr. 2013.
[45] N. Bourbaki, *Algebra, Chapter 3–5. Algebra I. Chapters 1–3. Elements of Mathematics*. Berlin, Germany: Springer-Verlag, 1989.
[46] (2006). *Unusual Crowd Activity Dataset of University of Minnesota* [Online]. Available: http://mha.cs.umn.edu/movies/crowd-activity-all.avi
[47] (2013). *UCSD Anomaly Detection Dataset* [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm
[48] 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.
[49] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1590–1599, Oct. 2013.
[50] L. Liu, L. Shao, and P. Rockett, "Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition," *Pattern Recognit.*, vol. 46, no. 7, pp. 1810–1818, 2013.
[51] L. Shao, S. Jones, and X. Li, "Efficient search and localization of human actions in video databases," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 504–512, Mar. 2014.
[52] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.
[53] M. Wang *et al.*, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.
[54] M. Wang, X.-S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.
[55] M. Wang, R. Hong, X. Yuan, and S. Yan, "Movie2comics: Towards a lively video content presentation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 858–870, Jun. 2012.
[56] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.
[57] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 34–41, 2013.
[58] Q. Wang, G. Zhu, and Y. Yuan, "Multi-spectral dataset and its application in saliency detection," *Comput. Vis. Image Underst.*, vol. 117, no. 12, pp. 1748–1754, 2013.

**Yuan Yuan** (M'05–SM'09) is a Full Professor with the Chinese Academy of Sciences, Xi'an, China. Her current research interests include visual information processing and image/video content analysis. She has published over 150 papers, including about 100 in reputable journals such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as conferences papers in CVPR, BMVC, ICIP, and ICASSP.



**Jianwu Fang** received the B.E. degree in automation and the M.E. degree in traffic information engineering and control from the Chang'an University, Xi'an, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree from the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

His current research interests include computer vision and pattern recognition.

**Qi Wang** received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently an Associate Professor with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His current research interests include computer vision and pattern recognition.