Pattern Recognition 105 (2020) 107394

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Video anomaly detection and localization using motion-field shape description and homogeneity testing

Xinfeng Zhang[a,b], Su Yang[a,c,*], Jiulong Zhang[c], Weishan Zhang[d]

[a] Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China
[b] College of Information Engineering, Yangzhou University, Yangzhou, China
[c] School of Computer Science, Xi'an University of Technology, Xi'an, China
[d] Department of Artificial Intelligence, China University of Petroleum, Qingdao, China

ARTICLE INFO

Article history:
Received 22 August 2018
Revised 19 March 2020
Accepted 21 April 2020
Available online 27 April 2020

Keywords:
Abnormal activity
Anomaly detection
Anomaly localization
Shape description
K-NN similarity-based outlier detection

ABSTRACT

Detection and localization of abnormal behaviors in surveillance videos of crowded scenes is challenging, where high-density people and various objects performing highly unpredictable motions lead to severe occlusions, making object segmentation and tracking extremely difficult. We associate the optical flows between multiple frames to capture short-term trajectories and introduce the histogram-based shape descriptor to describe such short-term trajectories, which reflects faithfully the motion trend and details in local patches. Furthermore, we propose a method to detect anomalies over time and space by judging whether the similarities between the testing sample and the retrieved *K*-NN samples follow the pattern distribution of homogeneous intra-class similarities, which is unsupervised one-class learning requiring no clustering nor prior assumption. Such a scheme can adapt to the whole scene, since the probability is used to judge and the calculation of probability is not affected by motion distortions arising from perspective distortion, which gains advantage over the existing solutions. We conduct experiments on real-world surveillance videos, and the results demonstrate that the proposed method can reliably detect and locate the abnormal events in video sequences, outperforming the state-of-the-art approaches.

1. Introduction

Due to the arising demand for public security issues and the widely equipped surveillance machines in public places, it is urgent to develop an automated system that can monitor and percept human activities to alarm abnormal events. In surveillance videos, the dominant activities occurring frequently are referred to as normal behaviors, which are in general not of concern. Apart from the normal activities, the most important and challenging task of an intelligent video surveillance system is to detect and localize anomalous events, which are defined as those to occur with a low probability [1]. In general, an abnormal event appears rarely and disappears in a short time. The goal of anomaly detection and localization is to identify the small time span and the spatial region covering the anomalous activities in an automatic manner [2,3].

In surveillance videos of public spaces, high-density people and various objects performing highly random motions [2] make anomaly detection especially challenging in crowded scenes. The traditional object-based approaches deem crowd as a collection of individuals. As this kind of methods conduct anomaly detection based on objects' appearances and trajectories, its performance is directly dependent on the accuracy of object extraction [4] and object tracking [5]. Unfortunately, capturing the single individuals is nearly impossible in crowded scenes, because of the high density of people and the various objects performing irregular motions to incur frequent and severe occlusions [2]. Aside from the aforementioned difficulties, tracking multiple objects is quite time-consuming [6].

To avoid the difficulty of segmenting individuals in crowded scenes, the latest trend in terms of anomaly detection is shifted to partition the surveillance videos into a couple of spatio-temporal volumes of a fixed size to focus on local scenes of a short time duration [7]. Then, the volume-based detection model in temporal and spatial contexts is established to discriminate whether the local scenes correspond to abnormal events or not, where the anomalies refer to such patterns that have never appeared at a specified site in contrast to the historical records or deviate remarkably from those of their neighborhoods at the same time [3]. In the literature, the unsupervised framework that makes use of normal volumes only for training has drawn considerable atten-

* Corresponding author at: Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China.
E-mail addresses: suyang@fudan.edu.cn, yangsu71@yeah.net (S. Yang).

https://doi.org/10.1016/j.patcog.2020.107394
0031-3203/

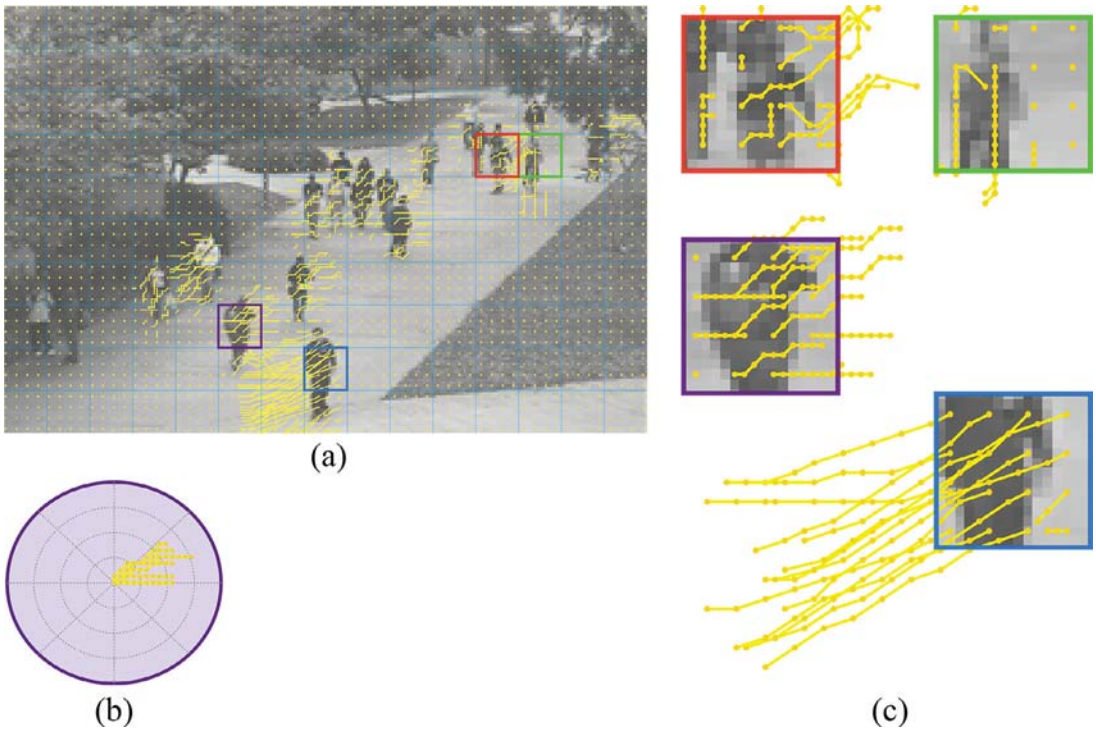

Fig. 1. An example of short-term trajectories and the histogram-based shape description for trajectories. (a) The patches with red, blue, and green borders correspond to anomalies, namely, skaters and one biker, while the purple region is a normal case with pedestrian only. (b) The histogram to figure out the short-term trajectories the purple patch in (a). (c) The enlarged view of the patches in (a) with the same colors and the corresponding trajectories. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tion, since anomalies are always rare and differ from one to another with unpredictable variations, making it almost impossible to model all the abnormal types [8].

We review two major categories of unsupervised approaches applied in anomaly detection in the following:

(1) A straightforward way to detect anomalous event is applying clustering methods to find outliers as anomalies [3]. In fact, such a scheme has been widely used in the existing works [2,7]. However, how to determine the number of clusters remains unsolved yet, which prevents its usage from being extended to a broad spectrum of practical applications. Classical clustering algorithms such as *k*-means and Gaussian mixture model (GMM) [3,7] require the number of prototypical patterns to be known a priori [2]. In crowded scenes, however, motion patterns are changing continuously and randomly such that some of them cannot be foreseen, which leads to uncertainty in regard to the number of prototypical patterns. Thus, it is impracticable to define the number of prototypical patterns in advance.

An alternative solution is to perform clustering based on a distance threshold so as to determine whether a sample belongs to an existing prototypical pattern or corresponds to a new prototype that should be created [9,10] as well as whether two clusters should be merged or not [8]. This kind of methods does not require the number of prototypical patterns to be known in advance but a specific distance threshold applicable to the whole scene to perform clustering does not exist due to the size variation of the object of interest, which is subject to the distance to the camera, namely, perspective distortion, which causes motion distortions. This gives rise to the same problem in defining the number of prototypes. For example, as shown in Fig. 1(a), the size of the skater in red color is much smaller than that of the one in blue color, in association with which the enlarged view of such objects is illustrated in Fig. 1(c) to enable an intuitive insight into the perspective distortion. In the case as shown in Fig. 1, it is impossible to define a uniform threshold to group the motion trajectories represented by any descriptor into reasonable clusters on account of the varying sizes of the objects caused by perspective distortion.

Due to the aforementioned object and motion distortion problem in surveillance scenarios, that is, the target size and motion step becomes larger when approaching more closely to the camera, some endeavors aiming to tackle such challenging issue have been made. Chen and Lai [11] use thermal diffusion processing and perspective transformation to construct a coherent motion flow field, and then establish a physical characteristic descriptor of crowd motion to model the crowd motion state of the flow field. However, the correction coefficient calculated for perspective transform requires manual selection of two parallel lines from each scene, which makes it difficult to deploy in practice. Leyva et al. [12] divide the scene into size-varying cells to adapt to the change of target size caused by scene' s perspective, and then extract foreground occupancy and optical flow features from these cells to detect abnormal events. However, the distortion extents are different for various scenes, so the setting of the changing rate of cells' sizes in a scene is difficult.

(2) The other category of methods is reconstruction-based approaches, for example, the method referred to as sparse representation cost [13]. Yang et al. [13] reconstruct testing samples from the normal samples of previous or surrounding volumes that act as the dictionary, and identify the samples with large reconstruction errors that exceed a predefined threshold as anomalies. However, once a very small number of abnormal samples are mixed into the dictionary, it will fail to detect the same kinds of abnormal behaviors due to the corruption on the dictionary. Besides, it is impossible to find a threshold applicable to the whole scene on account of the perspective distortion imposed inhomogeneity of the reconstruction errors for local regions of different positions.

In view of the weakness of the aforementioned approaches, we propose a motion-field shape descriptor along with a *K*-NN

(*K*-nearest neighbors) similarity-based statistical model to detect anomalies over time and space, where clustering or prior assumption are not needed. First, we associate the optical flows across multiple frames to capture the short-term trajectories in a video clip. The short-term trajectories characterize the motions in consecutive multi-frames and thus enhance motion pattern description. Hereafter, we introduce the histogram-based shape descriptor referred to as shape contexts [14] to figure out the short-term trajectories within each patch in a statistical sense, which reflects faithfully the motion trend and details in every local patch. To the best of our knowledge, this is the first attempt to apply shape description to quantize trajectories as motion features for anomaly detection in crowded scenes. Then, we propose to compute the *K*-NN similarity-based statistical model for anomaly detection as follows: First, we retrieve the *K*-NN samples from the training set in regard to the testing sample, and then use the similarities between every pair of the *K*-NN samples in the training set to construct a Gaussian model. Finally, the probabilities of the similarities from the testing sample to the*K*-NN samples under the Gaussian model are calculated in the form of a joint probability to check whether they are compatible with the Gaussian model. Abnormal events can be detected by judging whether the joint probability is below predefined thresholds in temporal and spatial contexts, separately. The major advantage is: The anomaly detection through probability can adapt to the whole scene, since the probability computed as such is not affected by the so-called perspective distortion. We carried out extensive experiments on three benchmarks with real-world scenes, UMN dataset [15], Subway dataset [16], and UCS-Dped1 dataset [3], for anomaly detection and localization, and the results validate the effectiveness and robustness of the proposed method.

The remainder of the paper is organized as follows: Section 2 reviews related work on anomaly detection. In Section 3, we introduce the histogram-based shape description method to characterize the short-term trajectories. Then, we propose the *K*-NN similarity-based model to detect anomalies in Section 4. In Section 5, we introduce the spatio-temporal anomaly detection scheme. We evaluate the performance of the proposed method in detecting and locating abnormal behaviors in Section 6. In Section 7, we draw conclusions.

2. Related work

Many methods detect anomalies by judging individual behaviors. For example, Hinami et al. [17] train a generic convolutional neural network (CNN) model on large datasets to learn individual objects' attributes and action features and then detect and recount abnormal events based on these features. For this kind of methods, the major challenge for abnormal event detection in crowded scenes is that the high density of the presence of objects makes detecting and tracking individual objects extremely difficult and thus inevitably unreliable. To tackle this problem, an emerging trend is to establish the detection model from local primitives such as pixels, image blocks/patches, and 3D cuboids/bricks [7] so as to avoid the error-prone object segmentation and tracking in crowded scenes. In the state-of-the-art works, the local features adopted for anomaly detection can be sorted into 3 classes, namely, the representations based on the properties of, interactions among, and trajectories of local primitives. (1) As for local-property based feature, Adam et al. [16] use histograms of optical flows (HOF) at specific regions to derive decision rules for anomaly detection. Giorno et al. [18] extract a set of features from a video, such as histogram of oriented gradient (HOG), HOF, and motion boundary histogram (MBH) descriptors, and then shuffle and split these features to find the most anomalous events in the contexts of the same video [19]. Uijlings et al. [20] design speed-ups for HOG and HOF descriptors. The HOG reflects the gradient magnitude responses, and HOF describes optical flow displacement vectors. However, in many scenarios, appearance features, such as HOG and scale-invariant feature transform (SIFT), are not suited for crowded scenes to distinguish normality and abnormality since the appearances are changing over time and the widespread artificial textures like clothing and car painting in arbitrary forms. Besides, as HOF captures motion clues between two successive frames only, it can only reflect the instantaneous motion, which is not enough to figure out the motion patterns of objects or object parts in terms of trajectory shapes across multiple frames. Mahadevan et al. [3] employ a mixture of dynamic textures (MDTs) to describe jointly the appearances and the dynamics of local portions of videos in crowded scenes. In order to further address the scale problem caused by different sizes of objects, they [21] train MDTs at multiple spatial scales, that is, a hierarchy of MDT model, and integrate anomaly scores across time, space, and scale with a conditional random field (CRF) for global consistency towards anomaly judgments. These approaches capture both temporal and spatial anomalies at the cost of highly intensive computations. Leyva et al. [22] present two binary-based video features, binary Wavelet differences (BWD), and binary dense trajectories (BDT), to describe motion information. The BWD and BDT descriptors are rotation and direction invariant. This means that the two features cannot distinguish different directions of movements, so they are not suitable for anomaly detection. (2) The well-known interaction based feature is the social force model (SFM) introduced by Mehran et al. [15], where crowd actions are modeled as interaction forces estimated from the corresponding optical flow field. However, SFM is not reliable and robust enough in the present of disturbance. (3) Wu et al. [10] make use of chaotic invariants as a trajectory feature, which is known as maximum Lyapunov exponent and correlation dimension, to measure how much neighboring particles deviate from their original closeness to each other after a certain steps of evolution. However, for the scenes that people' s movements are spatially constrained such as in corridors and underpasses, the evolution of trajectories might not follow the assumption of chaos, which assumes that neighboring trajectories will increasingly fall apart accompanying elapse of time. In such a case, Lyapunov exponent may not reflect exactly the chaotic degree of the collective human mobility corresponding to anomaly due to the constrained evolution of the trajectories. Furthermore, this feature is only applicable to the specific type of anomaly known as crowd chaos, which is the same as the entropy-based [23] and energy-based features [24].

Except for feature extraction, another important issue for anomaly detection is machine learning. The problem for supervised learning is that the annotations of normal and abnormal samples are difficult to be obtained since anomalies happen rarely and the diversity between each other with unpredictable variations makes modeling all the abnormal classes in advance impossible [8]. Consequently, recent works favor unsupervised machine learning. According to different unsupervised models applied, we broadly classify the anomaly detection approaches into three categories: Clustering-based approaches, reconstruction-based approaches, and relationship-based modeling. The details are presented below.

(1) For clustering-based approaches, Roshtkhari and Levine [9] construct a codebook by predefining a uniform Euclidean distance threshold to judge whether a observed volume should be used to update the existing codewords or treated as a new one. Then, they calculate the probability of the spatio-temporal collection of the volumes according to the codebook to detect anomalies, where predefining the max number of Gaussians used in the GMM is required. Ionescu et al. [25] introduce an unsupervised feature learning framework based on object-centric convolutional

auto-encoders to encode both motion and appearance information. Then, they cluster the training samples into normality clusters for training a one-against-rest classifier. During the inference, an object is labeled as abnormal if the highest classification score assigned by the classifier is negative. This approach needs to cluster the training samples before training the classifier. Therefore, its performance also depends on clustering. Cheng et al. [7] cluster the local features extracted around interesting points into a low-level visual vocabulary using the *k*-means algorithm in the sense of the Euclidean metric, and then measure the distances of a testing cuboid against the visual vocabulary to detect local anomalies. They cluster the collection of the features from nearby interesting points to build a high-level codebook of templates using a greedy clustering algorithm, and then construct a model for each template by fitting into a multivariate Gaussian distribution to detect global anomalies. However, using high-dimensional features to train a multivariate Gaussian distribution is subject to overfitting. Besides, a series of problems prevent this kind of approaches from being applied to broad-spectrum applications. For example, due to the uncertainty of the motions in crowded scenes, it is impracticable to define the number of prototypical patterns in advance [2] for classical clustering algorithms, such as *k*-means and GMM [3,7]. An alternative solution is to perform clustering based on a distance threshold, for instance, the greedy clustering algorithm [7]. For this kind of solutions, a specific distance threshold applicable to the whole scene does not exist due to perspective distortion.

(2) For reconstruction-based approaches, Yang et al. [13] identify the observed samples with large sparse reconstruction errors that exceed a predefined threshold as anomalies. Abati et al. [26] apply a deep autoencoder with a parametric density estimator to learn the probability distribution through an autoregressive procedure. The novelty of a sample is assessed in terms of the summation of the reconstruction errors and the log-likelihood. Luo et al. [27] optimize the reconstruction coefficients through mapping a temporally-coherent sparse coding to a stacked recurrent neural network, and perform detection based on reconstruction errors. For such anomaly detection approaches, once a very small number of abnormal samples are mixed into the dictionary, this kind of approaches will fail to detect the same kind of abnormal behaviors. Besides, a fixed threshold specified in advance also suffers from perspective distortion. Lu et al. [28] find a set of combinations of base vectors via sparse combination learning, and then, select the most suitable combination for each testing sample by evaluating the least square error in terms of fitting, and finally judge whether the testing sample is anomaly or not according to the fitting error. This approach learns sparse combinations in training phase, which increases the speed of the computation in testing. However, the extremely slow training speed limits its actual deployment as this approach needs to do specialized training for different scenes.

(3) Another category of approaches is based on modeling relationships among normal volumes. Kim and Grauman [29] utilize a space-time Markov random field (MRF) to detect abnormal activities in a video sequence, where each node in the MRF graph corresponds to a local region in the video frames and neighboring nodes in both space and time are associated with links. A MRF model is built for regular behaviors and the cases not compatible with the learned model are considered as anomalies. In crowded scenes, the volumes used for modeling may be taken from different parts of the same object or different objects. Moreover, due to dynamic occlusions, such volumes are dynamic changing. These result in very complex and uncertain relationships among volumes that go very easily beyond trained prototypes. Hu et al. [30] scan the video using a large number of windows, and then measure the abnormality of each scanning window for abnormal activity detection by computing a semiparametric density ratio between the observations inside and outside the window. However, the scan statistic method constructs a uniform codebook of optical flow-based features without taking into account the influence caused by perspective distortion, which leads to unsatisfactory results in terms of locating anomalies.

In this study, we focus on motion or behavior attributes for anomaly detection. We detect the correspondences among the optical flows across multiple frames to capture short-term trajectories and employ the histogram-based shape descriptor referred to as shape contexts [14] to characterize such short-term movements across a couple of consecutive frames. Then, we model the motion features using the proposed *K*-NN similarity-based statistical model to detect anomalies over time and space, which is an unsupervised one-class learning algorithm requiring no clustering nor prior assumption.

3. Short-term trajectory feature

Most of the existing motion-based approaches employ optical flow features [16,20], e.g., HOF, which capture motions between two successive frames only but fail to associate motions over multiple frames. In view of such limit, we associate the optical flows between multiple frames to capture short-term trajectories and employ a histogram-based shape descriptor, namely, shape contexts [14], to characterize such short-term trajectories.

3.1. Short-term trajectory

A given video of a crowd scene is divided into a series of non-overlapping clips, and each clip consists of a couple of frames streaming over a short time. Here, each clip is represented by a matrix of $W \times H \times T$ size, where $W \times H$ denotes the image resolution of every frame (width by height) and T is the number of sequential frames. We apply the general optical flow algorithm [31] to obtain the motion vectors denoted as follows:

$$\left\{\left(u_w^t, v_h^t\right) \mid w \in [1, W], h \in [1, H], t \in [1, T-1]\right\} \tag{1}$$

where u represents the horizontal velocity and v the vertical velocity [10]. We assume that the particles overlaying on pixels move with the optical flows to form particle trajectories in a video clip [32]. The position of a moving particle is formulated as follows:

$$\begin{cases} x_w^{t+1} = x_w^t + \left[u_w^t\right] \\ y_h^{t+1} = y_h^t + \left[v_h^t\right] \end{cases} \tag{2}$$

where $[\cdot]$ denote rounding operation and vector$(x_w^t, y_h^t)$the position of particle (w, h) at time t. Following [10], a particle trajectory is denoted as $\left\{(x_w^t, y_h^t) | t \in [1, T]\right\}$, and all the particle trajectories in a clip are denoted as

$$\left\{\left(x_w^t, y_h^t\right) \mid w \in [1, W], h \in [1, H], t \in [1, T]\right\} \tag{3}$$

As an example illustrated in Fig. 1(a), the yellow dot lines denote the short-term particle trajectories in a clip. Note that the positions of the particles are re-initialized for each clip, enabling the short-term trajectories to record only motions within each clip. Obviously, the nature of short-term particle trajectory is also optical flow, but it associates consecutive multi-frames and thus enhances motion pattern description. Also, it has been successfully applied to segment coherent crowd flows for video segmentation [32].

3.2. Shape histogram for short-term trajectory

We divide the starting frame of a clip into non-overlapping small patches $\{s(m) | m \in [1, M]\}$, where M denotes the total number of the patches and the frame partition should meet the condition that each patch does not involve too many objects to avoid

interference with each other. Then, the short-term trajectories staring in the same patch $\left\{(x_w^t, y_h^t)|(w,h)\in s(m), t\in[1,T]\right\}$ will undergo the histogram-based shape description as follows to characterize the corresponding motion patterns.

Firstly, perform translating on the particle trajectories in each patch to make the starting point of every trajectory locate at the origin of the polar coordinate, that is, $\left\{(x_w^t - x_w^1, y_h^t - y_h^1)|(w,h)\in s(m), t\in[1,T]\right\}$. Then, arrange the particle trajectories into $b_M \times b_A$ bins that are uniformly partitioned in terms of both magnitude and angle in the polar space, where b_M represents the number of the magnitude intervals and b_A that of the angle intervals. Finally, count the non-overlapping particles falling into each bin to obtain a histogram $\{h(n)|n \in [1, N]\}$, where $h(n)$ denotes the number of the particles falling in the nth bin and $N = b_M \times b_A$ the total number of the bins. As shown in Fig. 1, the starting point of every trajectory in the patch labeled with purple color is translated to the origin of the histogram in the Fig. 1(b) in order to figure out the distribution of the particles falling into each bin of the histogram.

3.3. Advantages over other motion features

HOG is originally developed for feature extraction on a single image, focused on gradient field. When HOG is applied to optical flow, its variant, HOF is developed, as follows [20]: First, a histogram with a couple of bins spreading over 0 to 360 degrees should be constructed for each image patch. Then, each pixel with optical flow (u_o, v_o) in the patch of interest votes to the corresponding histogram bin according to its angle $\theta = arctan\left(\frac{v_o}{u_o}\right)(1 \le o \le O)$ with the voting weight $\sqrt{u_o^2 + v_o^2}$, where O is the number of the pixels in the patch. Finally, the weights voted to each bin are accumulated to obtain the weight of each bin and the HOF of a patch is the vector of normalized weights of all histogram bins. In this sense, HOF describes only the instantaneous motions of optical flows between two successive frames and the continuous evolutions of the optical flows across multiple frames are missing.

Because short-term trajectory associates motions over multiple frames by tracking the particles following the motion vectors of optical flows from one position to the subsequent one, this leads to an essential difference from HOF. We illustrate the differences between HOF and short-term trajectory feature by the example of a skater and a pedestrian interlacing shown in Fig. 2, where the faster skater appearing on the walkway is abnormal. In this example, the calculated optical flows caused by the skater motion between two adjacent frames is not significantly longer than the normal ones, and in fact, there are often errors in the calculation of optical flow, such as in the first two frames of Fig. 2(b), the optical flow amplitudes caused by the crus movement of the left pedestrian are greater than those caused by the skater, and the directions of some optical flows are wrong. This directly leads to the degradation of HOF performance. The adjacent patches in time and space can be concatenated to construct 3D volumes of HOF. In terms of time, the blocks at the same position are abnormal only crossing 3 to 5 frames; in terms of space, the adjacent patches to an abnormal patch is likely to involve normal motions or background as shown in Fig. 2(b). Therefore, this 3D volumes of HOF, built by aggregating more patches, has limited description ability for abnormal movements.

The particles forming short-term trajectories move with the target, such as shown in Fig. 2(a), the yellow points on the skater move with the skater out of the observed patches. For complex interlacing, as illustrated in the patches from the second row to the fourth row of the two left columns of Fig. 2(c), at the beginning, the particles move with the pedestrian, and when the pedestrian is blocked by a skater, the particles move with the skater in reverse motion. This shows that there are setbacks on the corresponding short-term particle trajectories. It demonstrates that the short-term trajectories are able to record the motion trend and details of local motions. From Fig. 2(c), we can see that the short-term trajectories obtained by tracking particles are not sensitive to the calculation errors of optical flows, and the appearance of abnormal target can cause obvious change of trajectories, such as the skater appears in the patches in the two right columns earlier than in the patches in the two left columns, which result in the particle trajectories in the patches in the two right columns are obviously longer than those in the patches in the two left columns, because the particles in the patches in the two right columns move with the abnormal target for a longer time. In conclusion, the short-term trajectory feature has a good discrimination for anomaly description.

In contrast to the trajectory feature of the chaotic invariants [10], we apply the shape histogram, namely shape contexts [14], to describe short-term trajectories, which preserves the details and the trends of local motions. Even in the case that the movements are constrained in a narrow space, such as in corridors and underpasses, abnormal human mobility in terms of speed can also be reflected by the length of the trajectories in the shape histogram. Moreover, the histogram of short-term trajectories can also be spited into two histograms by accumulating the number of particles along the magnitude and angle dimension, respectively, which enable anomaly detection on speed and direction to be independent.

4. Statistical modeling of *K*-NN similarities

First, we use χ^2 test to measure the similarity between a testing sample and the normal training data, and retrieve the K-NN samples from the given training set in regard to the testing sample. Then, we establish a Gaussian model for the K retrieved samples to characterize the similarities between them in a statistical sense.

Note that there are two seemingly natural but in practice error-prone solutions in decision making: (1) Training the detection model such as the probabilistic model of multivariate Gaussian distribution [7,10] using directly high-dimensional feature, which is subject to overfitting and curse of dimensionality. (2) To avoid overfitting, reduce the dimension of the original feature by means of principal component analysis (PCA) [33]. Note that there is no anomalous samples in the training set and the whole training set is composed of one-class samples only, say, the normal samples. PCA may destroy the consistency of the training samples since its objective is to maximize the diversity of samples after projecting the data onto a low-dimensional space. Besides, it is not known a priori which dimensions of the feature contribute the majority of discriminant power in identifying abnormal and normal cases.

In view of the limits of the aforementioned decision-making methods, we establish a Gaussian model for the K retrieved samples based on the similarities computed from every pair of the K-NN samples, which can be regarded as statistical modeling of the intra-class similarities in the local manifold of the normal samples retrieved by the input testing sample. Here, the K retrieved samples are modeled through a one-dimensional Gaussian model over similarity in order to overcome the overfitting problem. Note that even if the samples to be compared in terms of similarity are high-dimensional data, the similarity itself is one-dimensional metric.

4.1. Gaussian distribution over K-NN similarities

Let H_p and H_q denote the shape contexts based motion features of two patches, respectively. Then, we use the χ^2 test to measure

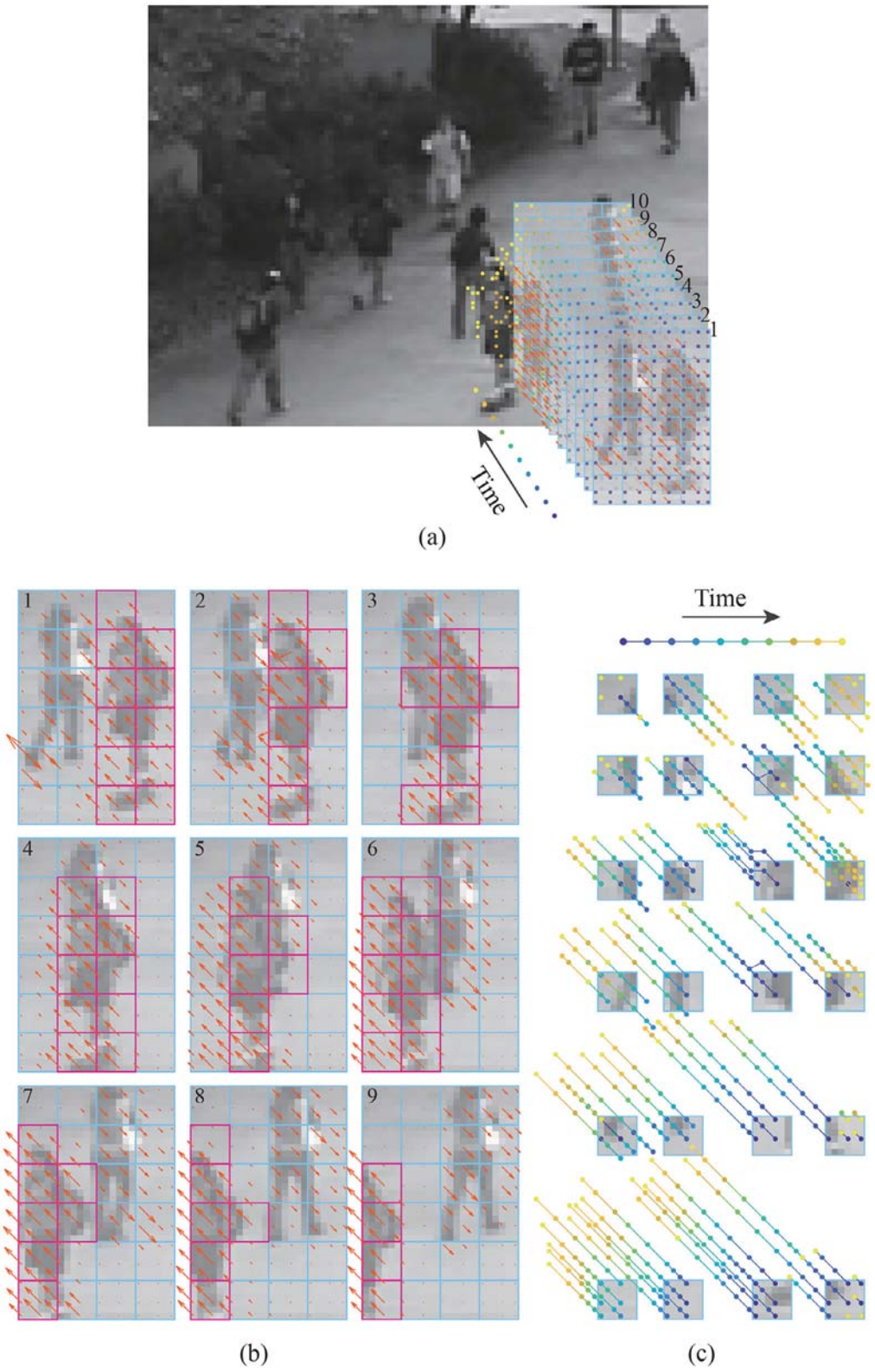

Fig. 2. An example of optical flows and short-term trajectories. (a) The orange arrows correspond to the optical flows aroused by a skater and a pedestrian within the observed patches, and the different color points correspond to the particles in each frame. (b) The enlarged view of the patches in (a), where the patches with red borders correspond to a skater. (c) The connecting lines of the enlarged view of different color points in (a) correspond to the particle trajectories. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the similarity between the two histograms:

$$\chi^2(H_p, H_q) = \frac{1}{2} \cdot \sum_{n=1}^{N} \frac{[h_p(n) - h_q(n)]^2}{h_p(n) + h_q(n)} \tag{4}$$

From the training set, we retrieve the K patches whose motion patterns are similar at most to that of the testing sample H_T in the sense of Eq. (4), which are denoted as $\{H_{NN}(K)|k \in [1, K]\}$. Other forms of distance functions can also be used here, such as L_2 norm. Then, we calculate the similarities between every pair of them to obtain $\{\chi^2(H_{NN}(i), H_{NN}(j))|i, j \in [1, K] \wedge i \neq j\}$. We model these intra-class similarities as a one-dimensional probabilistic model, that is, K-NN similarities-rendered Gaussian model $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 denote the mean and variance of the similarities $\{\chi^2(H_{NN}(i), H_{NN}(j))|i, j \in [1, K] \wedge i \neq j\}$ with the definitions as follows:

$$\begin{cases} \mu = \frac{2}{K \cdot (K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \chi^2\left(H_{NN}(i), H_{NN}(j)\right) \\ \sigma^2 = \frac{2}{K \cdot (K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \left[\chi^2\left(H_{NN}(i), H_{NN}(j)\right) - \mu\right]^2 \end{cases} \tag{5}$$

4.2. Similarity-rendered joint posterior probability

Once the similarities between the testing patch H_T and its K-NN patches $\{H_{NN}(K)|k \in [1, K]\}$ are obtained as $\{\chi^2(H_T, H_{NN}(k))|k \in [1, K]\}$, the fitness of H_T into the K-NN similarities-rendered Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is calculate as the joint posterior probability L_T defined below:

$$L_T = \sum_{k=1}^{K} \log\left\{\Pr\left[\chi^2\left(H_T, H_{NN}(k)\right) \in \mathcal{N}(\mu, \sigma^2)\right]\right\} \tag{6}$$

where Pr denote probability and the definition of the K-NN similarities-rendered Gaussian model $\mathcal{N}(\mu, \sigma^2)$ refer to Eq. (5). The joint posterior probability tends to be 0 as the number of K-NN patches increases, so we compute the sum of the logarithms of the probabilities instead to avoid this problem. Subsequently, we judge whether the testing sample is normal or abnormal by comparing its joint probability with a user-defined threshold T_P. If $L_T < T_P$, as a low-probability event, the corresponding testing sample is classified as anomaly.

The K-NN similarity-based statistical model for anomaly detection is reasonable in that: (1) For a normal testing patch, this patch and its K-NN patches are from the same normal class, so the similarities between this testing patch and its K-NN patches should follow the same distribution $\mathcal{N}(\mu, \sigma^2)$ spanned by the similarities between its K-NN patches. (2) For an abnormal testing patch, the similarities between it and the retrieved K-NN samples in the training data can be regarded as inter-class similarities. In such a case, their similarities deviate from the intra-class similarities-rendered Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ such that the possibility to follow the same distribution $\mathcal{N}(\mu, \sigma^2)$ is low. Note that the proposed K-NN similarity-based statistical model does not rely on any specific feature, so that it is a generic machine learning method applicable to a variety of features for anomaly detection.

5. Spatio-temporal anomaly detection

As stated previously, we are interested in detecting abnormal temporal and spatial activities. For temporal anomaly detection at a given location, the training data are composed of the patches from the same location of a long history. For spatial anomaly detection, the training samples are from the surrounding patches. The training data sets for temporal and spatial anomaly detection are used in the K-NN similarity-based statistical model to infer the occurrence probabilities of the testing patches over time and space, respectively. Two independent thresholds are used to detect the patches occurring with small probabilities in contrast to the temporal and spatial contexts, respectively, that is, those never seen before at a specified site or different from their neighborhoods at the same time span.

For the scenes containing active regions and non-active regions that are less visited, a lower probability threshold is usually set for non-active regions than that for active regions, since the anomalies appear with obviously lower probabilities in the rarely visited non-active regions.

In crowded scenes, motion clues such as optical flows are not stable in that one object often falls into different patches and the detected patches are in general only a part of the entire abnormal objects, so we will extend the detected abnormal patches to the surrounding regions to cover the most part of abnormal objects. Here, we apply multi-scale analysis with two thresholds, a lower one to location the anomalies with high certainty and a higher one to spread the detected areas to enclose the major portions of the objects corresponding to anomalies.

6. Experiments

The proposed method is tested on public real-world datasets: The UMN dataset [15], Subway dataset [16] and UCSDped1 dataset [3] with varying densities of people. The challenge is that the scenes in the datasets are not only crowded but also with some extent of perspective distortion.

6.1. Evaluation criteria

We evaluate different methods following the criteria widely used in previous works [3,8,29], which is as follows:

Frame-level evaluation: If any region in a frame is identified as anomaly to be consistent with the ground truth, such detection is granted to be correct regardless of the location and the size of the region.

Pixel-level evaluation: If over 40% portion of the ground truth are detected as anomalies in a frame, such detection is regarded as a right detection. So pixel-level evaluation is stricter than frame-level evaluation.

Event–level evaluation: If any position with true anomaly is detected and localized as abnormal, the detection is regarded as a correct hit. On the other hand, if any normal frame is detected as anomaly, it is counted as a false alarm in terms of event detection [29,34].

As for quantitative evaluation, receiver operating characteristic (**ROC**) reflects the relationship of true positive rate (TPR) against false positive rate (FPR), which are defined below:

True positive rate: The rate of correctly detected frames to all abnormal frames in ground truth.

$$\text{TPR} = \frac{\#\text{True Detection}}{\#\text{Abnormal Frames}} \tag{7}$$

False positive rate: The rate of incorrectly detected frames to all normal frames in ground truth.

$$\text{FPR} = \frac{\#\text{False Detection}}{\#\text{Normal Frames}} \tag{8}$$

ROC curve is plotted according to the detection results under different parameters. We quantify the performance in terms of the equal error rate (**EER**) and the area under ROC curve (**AUC**). The EER is the point on the ROC curve that FPR is equal to (1-TPR). A smaller EER corresponds to better performance. As for the AUC, a bigger value corresponds to better performance.

6.2. Crowd abnormal activity detection on UMN dataset

A crowd abnormal activity detection experiment is carried out to examine the performances of different methods. In the experi-

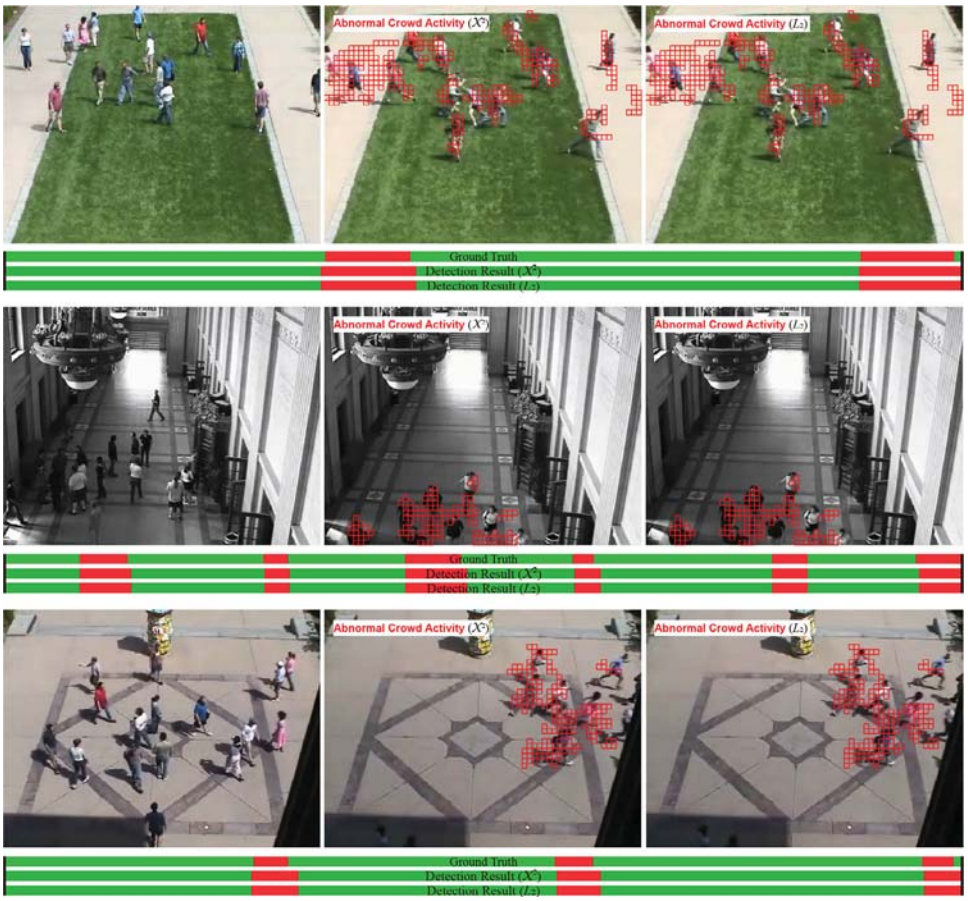

Fig. 3. Results of abnormal activity detection on the UMN dataset. Each row represents one scene. The three bars underlying each row are the detection results by using χ^2 test and L_2 norm distances against the ground truth, where the green color corresponds to the normal frames and the red color represents the abnormal frames. Above the bar, the left column is an example of the normal frames, and the middle column and the right column are examples of the detected results of our method by using χ^2 test and L_2 norm distances, respectively. In these examples, the abnormal regions are marked with red grids. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ment, we use the surveillance videos from the University of Minnesota (UMN) [15]. The UMN dataset consists of 11 videos of 3 different indoor and outdoor scenes, and the three scenes comprise 1453 frames, 4143 frames, and 2143 frames, respectively. The resolution of each video frame is 240×320 pixels. Each video begins with people walking around, followed by a portion of abnormal panic movements such as running and escaping. In the crowded scenes, lighting changes in a few of the videos bring some challenges to abnormal activity detection.

Results: We split the video into clips of 10 consecutive frames, and divide the starting frame of each clip into a couple of patches of 5×5 pixels without overlap. The optical flows in each video clip are linked into short-term trajectories. On the basis of such setting, our method is established to determine whether each patch is abnormal. If a certain range of patches are abnormal, it is thought that global anomaly has occurred. The detection results of crowd abnormal activity on the UMN dataset achieved by using the proposed method on χ^2 test and L_2 norm distances are shown in Fig. 3 against the ground truth. In Fig. 3, the 3 rows correspond to the 3 scenes composed of the 11 video sequences. Fig. 3 also illustrates some examples of the detection results. As can be seen from Fig. 3, our method can detect every abnormal event, and there is no false alarm. Using different distance functions, the proposed method implements basically the same performance as presented in Table 1.

Comparison with the existing approaches: In Table 1, we compare the proposed method with h-mixture of dynamic textures (H-MDT) [21], object-centric auto-encoders (OCAE) [25], divcurl characteristics (DCC) [11], scan statistic (SS) [30], sparse reconstruction cost (SRC) [13], and compact feature sets (CFS) [12]. Since the type of anomaly is known and consistent, crowd abnormal activity detection on this dataset is relatively easy. Almost all frame-level AUC scores are higher than 90% as listed in Table 1. The top two scores of 99.6% and 99.5% are reported by OCAE and H-MDT, and the proposed method attains a competitive performance on the UMN dataset. In addition, both of the best-performing meth-

Table 1
AUC by using different methods on the UMN dataset.

Methods	AUC on UMN dataset (%) (Scene 1/2/3)
Proposed (χ^2)	99.3 (99.4/99.1/99.4)
Proposed (L_2)	99.2 (99.4/99.1/99.3)
H-MDT	99.5
OCAE	99.6 (99.9/99.1/99.8)
DCC	(90.9/87.5/97.7)
SS	(99.1/95.1/99.0)
SRC	(99.5/97.5/96.4)
CFS	88.3

χ^2 means using χ^2 test; L_2 means using L_2 norm; "-" means the results are not provided.

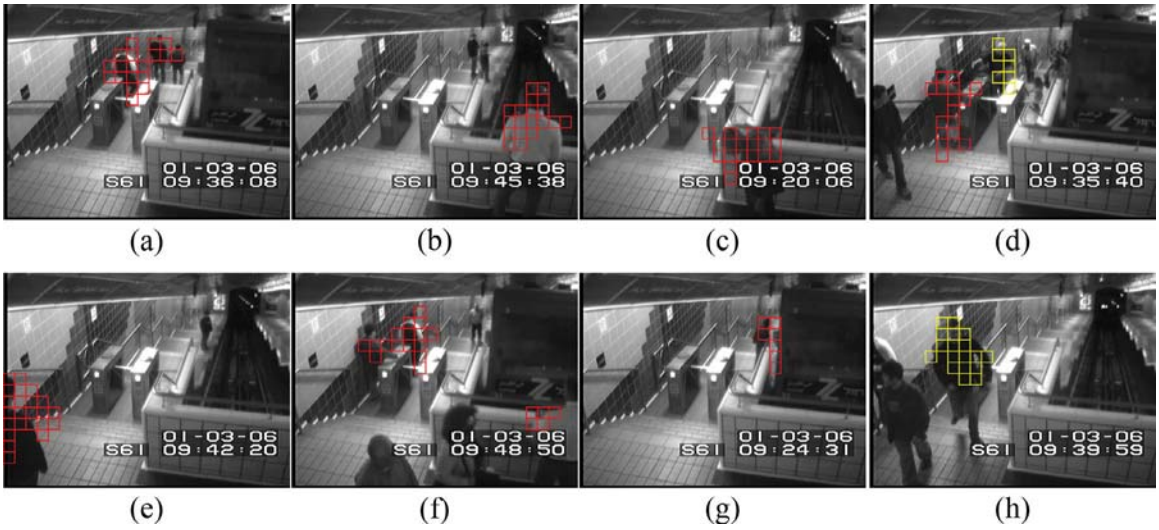

Fig. 4. Examples of the detection results on the Subway dataset. The red color indicates correct detection and the yellow color denotes false alarm. (a) and (e) Wrong direction: Some persons are entering through the exit gate. (b) and (f) Loitering: A person is wandering; Two persons are entering through the exit gate. (c) and (g) Miscellany: A person is cleaning the wall; A person gets off the train and then gets on the train again very soon. (d) and (h) False alarm: An adult is helping a child pass the turnstile; A person is coming up from the turnstile with an irregular jump; A correct detection: A person is entering through the exit gate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ods require more computational cost, such as the OCAE requires running on high-performance GPUs and the processing time per video frame of the H-MDT is about 1100 milliseconds. The comparison of the detection results also verifies the performance of the features. From the comparison, we can see that the proposed shape descriptor of short-term trajectories outperform the compact set of highly descriptive features based on foreground occupancy and optical flow information, such as optical flow energy and a histograms of optical flow (HOF) descriptor in the CFS [12]. For the UMN dataset, detection results are compared in terms of the frame-level AUC, that is, the whole scene is identified as normal or abnormal. Since this dataset provides no pixel-level ground truth [13], we will further accurately compare our method with others.

6.3. Anomalous event detection on subway dataset

The Subway dataset is provided by Adam et al. [16]. In the experiment, we use the surveillance videos at the exit of subway to examine the performances of different approaches. The camera is pointed toward the exit gate, where the dominant behaviors are exiting from the platform, coming up through the turnstiles, and turning to the left or right. The video lasts about 43 min with a resolution of 384×512 pixels, which contains 19 anomalous events, mainly involving walking in the wrong direction, loitering, and miscellany [29].

Results: The starting frame of each clip is divided into non-overlapping patches of 20×20 pixels, and the length of each video clip is set to be 10 consecutive frames (400 milliseconds), where no overlap exists between two continuous clips. The first 10 min of the video are used for training, while the rest of the frames are used for testing. Some examples of the detection results achieved by using the proposed method are shown in Fig. 4, where correct detections and false alarms are both included. The results validate that the proposed method with a fixed threshold can capture multiple abnormal objects simultaneously at different scale no matter whether the anomalies are close to or far from the camera. In fact, the regions marked with yellow grids in Fig. 4 to denote false alarms in accordance with the ground truth can also be true anomalies. For example, in the case shown in Fig. 4(d), an adult is helping a child pass the turnstile, and for the case shown in Fig. 4(h), a person is coming up from the turnstile with an irregular jump. Such unusual behaviors are missed in the annotations of the ground truth [29] but detected by the proposed method. Besides, it is apparent that the proposed method can accurately detect and localize anomalies in surveillance videos with perspective distortion.

Table 2
Comparison of abnormal event detection rate and false alarm rate on the Subway dataset.

Methods	TP	WD	LT	Misc.	Total	FA
Ground	-	9	3	7	19	0
Proposed	10	9	3	7	19	2
SCL	15	9	3	7	19	2
STC	15/CL	9	3	7	19	2
SS	CL	9	3	7	19	2
CFS	CL	6	3	2	11	7
SRC[c]	10	9	-	-	9	0
LOF[c]	5	9	-	-	9	2

TP: Training period; CL: Continuous learning; WD: Wrong direction; LT: Loitering; Misc.: Miscellany; FA: False alarm; "-" means the results are not provided. [c]Used annotation with a reduced number of abnormality types.

Comparison with the existing approaches: In Table 2, we compare quantitatively the proposed method with sparse combination learning (SCL) [28], spatio-temporal compositions (STC) [9], scan statistic (SS) [30], sparse reconstruction cost (SRC) [13], compact feature sets (CFS) [12] and local optical flow (LOF) [16] at event level. It can be seen that the proposed method requires the least training data to achieve the same level of high detection rate and low false alarm rate in comparison with SCL and STC.

6.4. Anomaly detection and localization on UCSDped1 dataset

We conduct anomaly detection and localization test on the UCSDped1 dataset [3], which records a large number of pedestrians walking on the walkway in a college campus to approach or move far away from the camera. Since the number as well as the density of the people appearing in the monitoring area varies

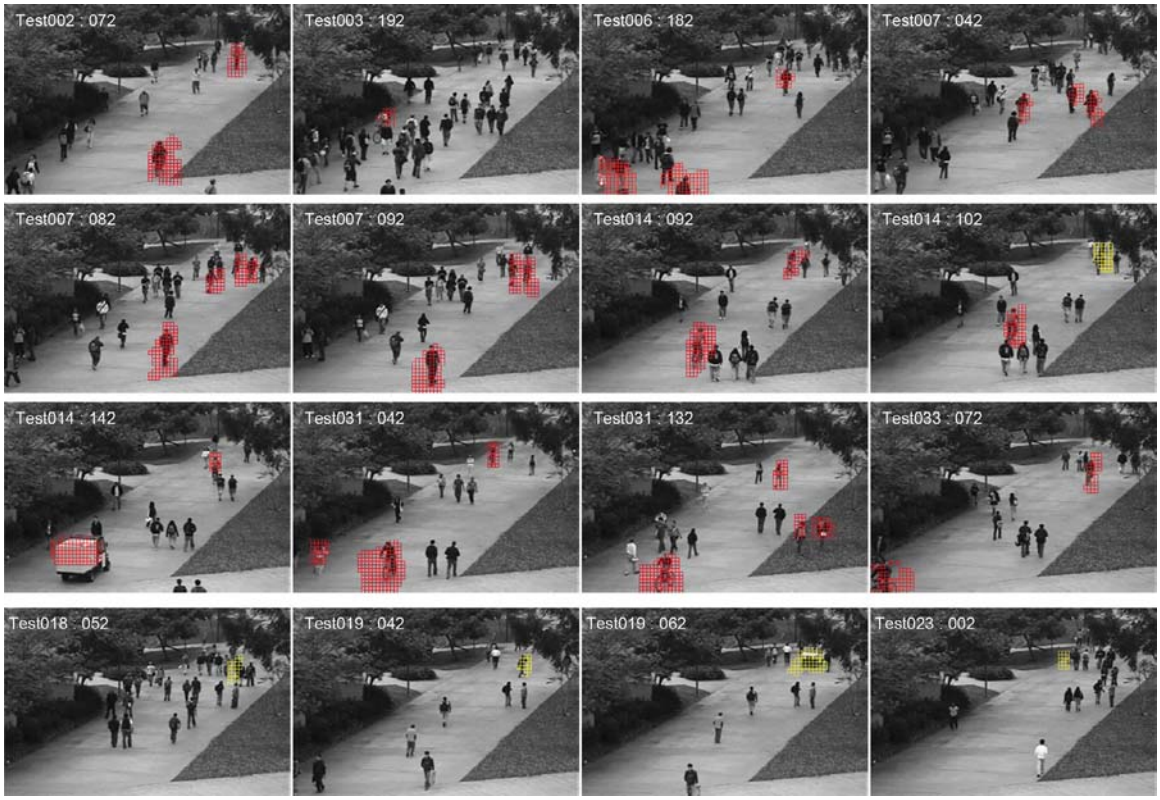

Fig. 5. Some examples of the detection results on the UCSDped1 dataset. The red color means correct detection and the yellow color denotes false alarm according to the ground truth. The last row shows some true anomalies missing annotations in the ground truth but captured by the proposed method, where a pedestrian crossing the walkway in abnormal direction, a skate, the frontal of a vehicle, and a person on wheelchair take places in the frames labeled "Test019 : 042", "Test018 : 052", "Test019 : 062", and "Test023 : 002", respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with time and the video sequences are captured with a certain degree of perspective distortion, this benchmark is quite challenging. There are 34 training video sequences of normal cases and 36 testing video sequences involving various abnormal events, which include non-pedestrians on the walkway (e.g., bikers, skaters, small carts, and people in wheelchairs etc.), as well as pedestrians with anomalous motion patterns or in non-walkway regions (e.g., people running or walking across the grass etc.). Each video sequence has 200 frames with 158×238 resolution.

Results: To enable detection of anomalies of small sizes, we uniformly divide the starting frame of every video clip into a couple of patches of 3×3 pixels without overlap, and then link the optical flows in the 10 consecutive frames of each video clip to obtain short-term trajectories. Some examples of the detection results achieved by using the proposed method are shown in Fig. 5, which should be the most difficult tasks in the literature. One challenging issue is that multiple objects may appear in one scene. For instance, 4 persons with abnormal behaviors appear in the frames labeled "Test007 : 082" and "Test031 : 132". Another difficulty is the occlusions caused by the high density of people. For example, a large portion of the biker in the frame labeled "Test003 : 192" is blocked and thus invisible due to the high density of people. In the aforementioned examples, the anomalies are detected correctly by using the proposed method, which are marked with red grids as shown in Fig. 5. In Fig. 5, we also illustrate some examples of false alarm, which are labeled in yellow color. To our surprise, some regions labeled in yellow color are indeed true anomalies such as the yellow region in the frame labeled "Test014 : 102", which corresponds to a partially visible biker blocked by the vegetation and pedestrians. In the previous frame labeled "Test014 : 092", the corresponding object is a true anomaly but it misses being annotated in the ground truth of the frame labeled "Test014 : 102" [3]. The results manifest that the proposed method is able to detect different types of anomalies and performs well on very small patches. In another words, the proposed method can adapt to the whole scene with perspective distortion as it can detect the anomalies wherever they are, close to or far from the camera.

Comparison with the existing approaches: In Fig. 6, we compare the ROC curves of the proposed method with Gaussian process regression (GPR) [7], sparse reconstruction cost (SRC) [13], sparse combination learning (SCL) [28], scan statistic (SS) [30], spatio-temporal compositions (STC) [9], compact feature sets (CFS) [12] and local optical flow (LOF) [16]. The ROC curves in Fig. 6 illustrate the True Positive Rate and False Positive Rate tradeoff. In order to quantitatively evaluate the performances of different approaches, EER and AUC of anomaly detection at both pixel level and frame level as suggested by [3] are listed in Table 3. The dense STC is implemented by [7] and the results of these methods are obtained from the above-mentioned papers. For the frame-level evaluation, EERs of GPR, CFS, SRC, SS, SCL, and STC are 23.7%, 21.2%, 19%, 17.5%, 17% and 16%, respectively. These baseline approaches achieve comparative performances in contrast to the proposed method with an EER of 19.5% (AUC of 86.9%).

However, frame-level evaluation does not consider whether the detection coincides with the actual location of the anomaly [3]. In contrast, pixel-level evaluation emphasizes the localization ability of an algorithm [9]. In surveillance videos, perspective distortion causes the motion vectors not consistent with each other to fall in diverse directions and scales. This will significantly affect anomaly detection in local regions. Since almost all approaches [7,9,13] treat

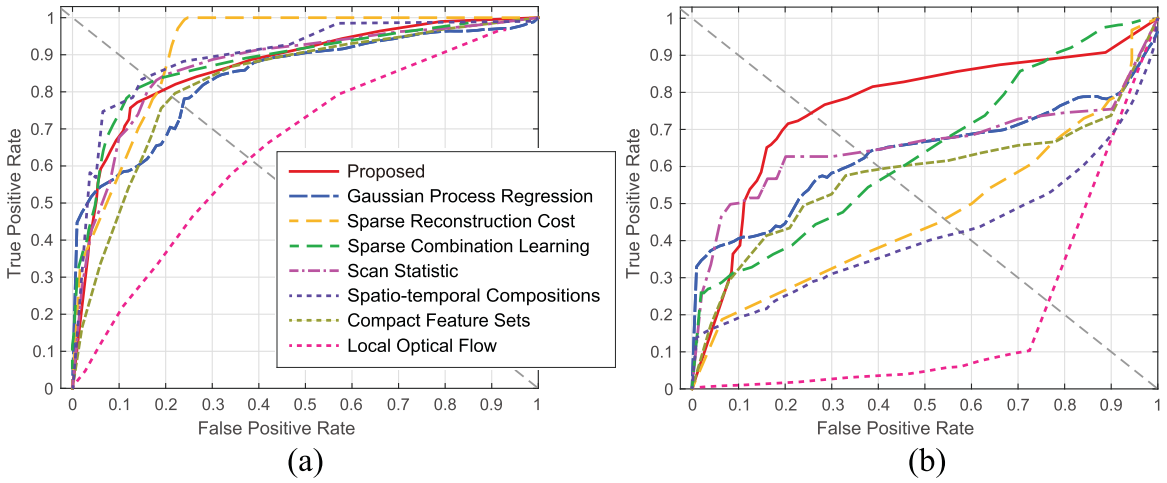

Fig. 6. The comparison of ROC curves on the UCSDped1 dataset using different approaches. The dashed diagonal is the EER line. (a) Frame-level evaluation. (b) Pixel-level evaluation.

Table 3
Comparison of EER and AUC using different approaches (%).

Methods	Frame Level		Pixel Level	
	AUC	EER	AUC	EER
Proposed	86.9	19.5	**76.2**	**25.6**
GPR	83.8	23.7	63.3	37.3
SRC	**91.4**[a]	19	47[a]	54
SCL	88.4[a]	17	64.3[a]	42
SS	87.6[a]	17.5	66[a]	36
STC	89.9	**16**	41.7	57.7
LOF	65.2[a]	38	17.3[a]	76
CFS	-	21.2	-	39.7

"-" means the results are not provided; [a]Estimated in terms of the corresponding ROC curves.

Table 4
Computational time of different approaches (processing time per frame in milliseconds).

Methods	Learning	Inferring	CPU (GHz)	RAM (GB)
Proposed	total: 175 ms		3.4	16
GPR	140.2 ms	515.3 ms	3.4	4
STC	2432.5 ms	2424.1 ms	3.4	4
SRC	-	3800 ms	2.6	2
SCL	37 mins	6.965 ms	3.4	8
CFS	total: 31 ms		2.7	8
SS	-	200 ms	3	4

"ms" and "mins" are short for millisecond and minute, respectively; "-" means the results are not provided.

all the volumes in a scene equally by applying uniform threshold or vocabulary without taking into account the aforementioned effect caused by perspective distortion, their performances degrade in locating anomalies due to perspective distortion. SRC, SCL, CFS, GPR and SS yield EERs of 54%, 42%, 39.7%, 37.3% and 36%, respectively. The values of EERs at pixel level localization are significantly larger than the values of EERs achieved by them at frame level detection. Our method identifies the testing sample to appear with a low probability as anomaly, which is calculated in the form of the joint probability of the similarities from the testing sample to the K-NN samples under the Gaussian model computed from the corresponding K-NN similarities. Since the probability threshold is capable of adapting to perspective distortion, the proposed method achieves an EER of 25.6% (AUC of 76.2%) as listed in Table 3, and performs much better than the other approaches at pixel level as shown in Fig. 6(b).

6.5. Influence of K

As the K training samples similar to the testing sample at most are retrieved to establish the K-NN similarity-based Gaussian model, it is necessary to further analyze the influence of K. EER for the UCSDped1 dataset at frame level and pixel level with different K value are plotted in Fig. 7. It can be seen that the EER reach minimum at the value of K close to 35. Overall, the EER varies little for different settings of the value of K from 15 to 70. This demonstrates that the proposed method is robust to the value of K. In contrast, for the clustering algorithm, the number of the prototypical patterns has a strong impact on the clustering result.

6.6. Computational complexity

We compare the computational time of the proposed method with the approaches that have comparative performances on the frame-level evaluation, including GPR [7], STC [9], SRC [13], SCL [28], and SS [30] on the UCSDped1 dataset. The speed of the approaches reported in the corresponding literatures are listed in Table 4.

The proposed method is implemented using MATLAB and the experiments are performed on a computer with Core i7-2600 3.4GHz CPU and 16GB RAM. Our method requires training and inferring time, approximately 175 milliseconds per frame, which contains the shape feature extraction, and the construction of the K-NN similarity-based model as well as the decision making. It is obviously that the proposed method is very efficient.

The average speed of SCL is 143.58 frames per second, which is the fastest testing speed in the literature so far. However, in actual deployment, this approach needs to do specialized training for different scenes. This restricts the practical application of this approach, because the training process is inefficient, that is, training 20K-sample data approximately needs 20 min. From the above experiments, we can see that the CFS is very fast, but its performance is inferior to that of our method. It can be drawn from Fig. 6(b) and Table 4 that our method leads to the least training and inferring time among the approaches for comparison while results in the highest accuracy in locating anomalies.

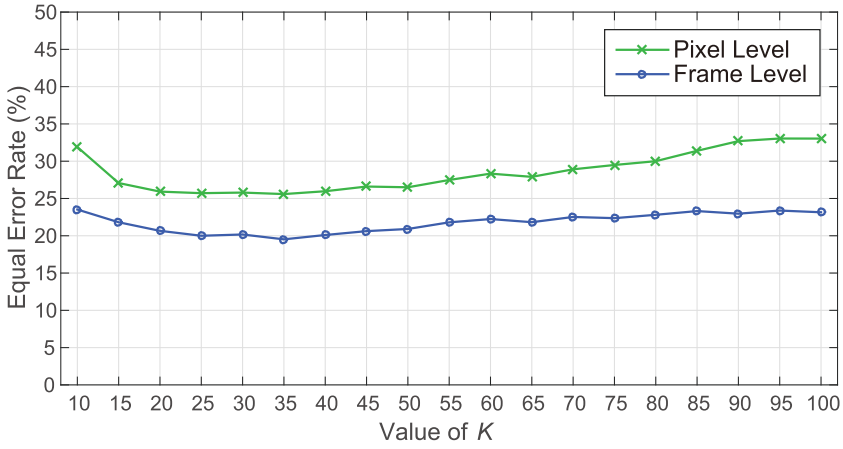

Fig. 7. Influence of the value of K for anomaly detection on the UCSDped1 dataset.

7. Conclusions

The contribution of this paper is two-fold: (1) The representation of video contents is one key issue for anomaly detection. We transfer the problem into shape description on short-term motion trajectories by associating the optical flows between multiple frames. The advantage is that the low-level short-term trajectory feature does not rely on unreliable object segment and tracking in crowded scenes while preserve the motion information of object parts, which is a promise of robustness. Moreover, the rich contexts of shape description enables discriminative representations of such short-terms trajectories for pattern analysis. (2) A new outlier detector is proposed. It is a homogeneity testing of the similarities. The mechanism is that if testing sample is compatible with its K-NN samples, the similarities between it and its K-NN samples should be compatible in a statistical sense with the similarities between its K-NN samples. We test the similarity homogeneity, rather than directly judge the homogeneity of samples in feature space, in order to avoid the overfitting problem caused by high-dimensional features. The proposed similarity-based statistical model for detecting anomalies over time and space is an unsupervised one-class learning algorithm, which does not require clustering or prior assumption in contrast to the existing solutions, for example, some approaches need to manually set distortion parameters for each scene. Compared with the error threshold and distance threshold that is not applicable to the whole scene due to the perspective distortion, the proposed statistical model determines anomaly according to the probability, which can adapt to the whole scene, since the probability of different position motion patterns is not affected by motion distortions arising from perspective distortion. We carried out experiments on three real-world surveillance videos, UMN dataset, Subway dataset and UCSDped1 dataset, for anomaly detection and localization, and the results demonstrate that our method is robust to the parameter variation, and promises competitive performance in terms of anomaly detection and better performance in the sense of localization compared with the state-of-the-art approaches.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by NSFC (grant No. 61801417), Natural Science Research of Jiangsu Higher Education Institutions of China (No. 18KJB520051), Shanghai Science and Technology Commission (grant No. 17511104203), National Key R&D Program (No. 2018YFE0116700), and the Shandong Provincial Natural Science Foundation (No. ZR2019MF049).

References

[1] T. Xiao, C. Zhang, H. Zha, Learning to detect anomalies in surveillance video, IEEE Signal Process. Lett. 22 (9) (2015) 1477–1481.
[2] L. Kratz, K. Nishino, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1446–1453.
[3] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981.
[4] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, T. Yu, Unified crowd segmentation, in: Computer Vision - ECCV 2008, European Conference on Computer Vision, Marseille, France, October 12–18, 2008, pp. 691–704.
[5] A. Basharat, A. Gritai, M. Shah, Learning Object Motion Patterns for Anomaly Detection and Improved Object Detection, vol. 0, IEEE Computer Society, Los Alamitos, CA, USA, 2008, pp. 1–8, doi:10.1109/CVPR.2008.4587510.
[6] Y. Yuan, J. Fang, Q. Wang, Online anomaly detection in crowd scenes via structure analysis, IEEE Trans. Cybern. 45 (3) (2015) 562–575.
[7] K.W. Cheng, Y. Chen, W.H. Fang, Video anomaly detection and localization using hierarchical feature representation and gaussian process regression, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[8] M. Bertini, A.D. Bimbo, L. Seidenari, Multi-scale and real-time non-parametric approach for anomaly detection and localization, Comput. Vis. Image Underst. 116 (3) (2012) 320–329.
[9] M.J. Roshtkhari, M.D. Levine, An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions, Comput. Vis. Image Underst. 117 (10) (2013) 1436–1452.
[10] S. Wu, B.E. Moore, M. Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2054–2060.
[11] X.-H. Chen, J.-H. Lai, Detecting abnormal crowd behaviors based on the div-curl characteristics of flow fields, Pattern Recognit. 88 (2019) 342–355.
[12] R. Leyva, V. Sanchez, C. Li, Video anomaly detection with compact feature sets for online performance, IEEE Trans. Image Process. 26 (7) (2017) 3463–3478, doi:10.1109/TIP.2017.2695105.
[13] C. Yang, J. Yuan, L. Ji, Sparse reconstruction cost for abnormal event detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3449–3456.
[14] S.J. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2010) 509–522.
[15] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.
[16] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 555–560.

[17] R. Hinami, T. Mei, S. Satoh, Joint detection and recounting of abnormal events by learning deep generic knowledge, in: IEEE International Conference on Computer Vision (ICCV), 2017.
[18] A. Del Giorno, J.A. Bagnell, M. Hebert, A discriminative framework for anomaly detection in large videos, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 334–349.
[19] P. Perera, R. Nallapati, B. Xiang, OCGAN: one-class novelty detection using GANs with constrained latent representations, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[20] J.R.R. Uijlings, I.C. Duta, N. Rostamzadeh, N. Sebe, Realtime video classification using dense HOF/HOG, in: Proceedings of International Conference on Multimedia Retrieval, in: ICMR'14, ACM, New York, NY, USA, 2014, pp. 145:145–145:152, doi:10.1145/2578726.2578744.
[21] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 18–32, doi:10.1109/TPAMI.2013.111.
[22] R. Leyva, V. Sanchez, T.-L. Chang, Fast binary-based video descriptors for action recognition, in: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2016, pp. 1–8, doi:10.1109/DICTA.2016.7797041.
[23] X. Zhang, S. Yang, Y.Y. Tang, W. Zhang, A thermodynamics-inspired feature for anomaly detection on crowd motions in surveillance videos, Multimed. Tools Appl. 75 (14) (2016) 8799–8826.
[24] G. Xiong, J. Cheng, X. Wu, Y.L. Chen, Y. Ou, Y. Xu, An energy model approach to people counting for abnormal crowd behavior detection, Neurocomputing 83 (7) (2012) 121–135.
[25] R.T. Ionescu, F.S. Khan, M.-I. Georgescu, L. Shao, Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[26] D. Abati, A. Porrello, S. Calderara, R. Cucchiara, Latent space autoregression for novelty detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[27] W. Luo, W. Liu, S. Gao, A revisit of sparse coding based anomaly detection in stacked RNN framework, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 341–349, doi:10.1109/ICCV.2017.45.
[28] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 FPS in MATLAB, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2013, pp. 2720–2727, doi:10.1109/ICCV.2013.338.
[29] J. Kim, K. Grauman, Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2921–2928.
[30] Y. Hu, Y. Zhang, L.S. Davis, Unsupervised abnormal crowd activity detection using semiparametric scan statistic, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, in: CVPRW '13, IEEE Computer Society, Washington, DC, USA, 2013, pp. 767–774, doi:10.1109/CVPRW.2013.115.
[31] T. Brox, N. Bruhn A.and Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: Computer Vision - ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004, pp. 25–36, doi:10.1007/978-3-540-24673-2_3.
[32] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.
[33] D. Du, H. Qi, Q. Huang, W. Zeng, Abnormal event detection in crowded scenes based on structural multi-scale motion interrelated patterns, in: IEEE International Conference on Multimedia and Expo, 2013, pp. 1–6.
[34] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Pattern Recognit. 46 (7) (2013) 1851–1864.

Xinfeng Zhang is a lecturer in College of Information Engineering at Yangzhou University. He received a Bachelor degree in electronic and information engineering from Hebei University, a Master degree in signal and information processing from Shantou University, and Ph.D. in computer science from Fudan University. His research interests are computer vision and multi-perception information processing.

Su Yang is a full professor in School of Computer Science at Fudan University. His main research interest is pattern recognition and its applications in media processing and smart cities. His works in symbol recognition and feature selection were widely cited. He received the best paper award from CPSCom 2010 and chaired the 7thSocialComin Beijing, 2014.

Jiulong Zhang is an associate professor in School of Computer Science at Xi'an University of Technology. His current research interests are computer vision, image processing, affective computing, and human computer interaction. He has published over 40 papers.

Weishan Zhang is a full professor, and deputy head for research of Department of Software Engineering, China University of Petroleum. His current research interests are big data processing, pervasive and service oriented computing. He has published over 100 papers. According to Google Scholar, his current total citations are over 1200, H-index is 19, and I10-index is 37.