Int. J. Mach. Learn. & Cyber. (2018) 9:1999–2010
DOI 10.1007/s13042-017-0682-8

ORIGINAL ARTICLE

Detection and localization of crowd behavior using a novel tracklet-based model

Hamidreza Rabiee[1] · Hossein Mousavi[2] · Moin Nabi[3] · Mahdyar Ravanbakhsh[4]

Received: 30 June 2016 / Accepted: 10 April 2017 / Published online: 18 April 2017

Abstract In this paper, two novel descriptors are introduced to detect and localize abnormal behaviors in crowded scenes. The first proposed descriptor is based on the orientation and magnitude of short trajectories extracted by tracking interest points in spatio-temporal 3D patches. The proposed descriptor employs a novel simplified Histogram of Oriented Tracklets (sHOT), which is shown to be very effective in the task of crowd abnormal behavior detection. In this scheme, abnormal behaviors are detected at different levels, namely spatio-temporal level and frame level. By combining the first proposed descriptor and the dense optical flow model, we propose our second framework which is able to localize the abnormal behavior areas in video sequences. The evaluation of our simple but yet effective descriptors on different state-of-the-art datasets, namely UCSD, UMN and Violence in Crowds yields very promising results in abnormality detection and outperforming different former state-of-the-art descriptors.

Keywords Crowd behavior analysis · Simplified histogram of oriented tracklets · Spatio-temporal features

✉ Hamidreza Rabiee
hr.rabiee@gmail.com

[1] Department of Electrical Engineering, Faculty of Mechatronics, Karaj Branch, Islamic Azad University, Karaj, Iran

[2] Polytechnique Montréal, Montréal, QC H3C 3A7, Canada

[3] DISI, University of Trento, Trento, Italy

[4] DITEN, University of Genoa, Genova, Italy

1 Introduction

In recent years, the field of crowd behavior analysis had an outstanding evolution in computer vision for several problems such as density estimation [1], motion detection [2], tracking [3] and crowd behavior analysis [2, 4–13]. However, crowd behavior analysis is still the topic of many studies in computer vision communities. This is mainly because of both inherent complexity and vast diversity in the crowd scene understanding. In contrast to a low-density crowd, the behavior of each individual in a dense crowd might be affected by different factors such as goals, dynamics, environment, etc. In other words, a dense crowd goes beyond a set of individuals who act independently and show their personal behavioral patterns [14].

Another major challenge in abnormality detection is that there is no absolute definition of abnormalities as they are context dependent. However, abnormalities are usually considered as outliers of normal distributions. Under this hypothesis, abnormalities are rare observations, which contrast highly with the normalcy. Sudden changes in pedestrian directions and their high speed in the presence of non-pedestrian moving objects could be considered as abnormal behaviors. The field of view is another effective parameter in video recording. However, the perspective geometry which introduces distortions and apparent motions do not correspond to the actual movements (an actual constant speed will not correspond to a constant arbitrary motion) and are ignored. Another point to consider is the number of individuals in a crowd scene, which may affect the quality of abnormality detection and localization. On the other hand, the noises, such as occlusion and clutter are substantial in very crowded scenes and have to be handled.

It is also important to detect abnormal behaviors both in space and time domains. This refers to isolate the abnormal

frames in a crowd video (we call it as *frame-level* abnormality detection) and to localize the abnormal areas in identified abnormal frames (we call it as *pixel-level* abnormality detection). With unanimous approval, the existing approaches for abnormal behavior detection in a crowd are mainly divided to two main categories, namely object-based approach and holistic approach. A typical object-based approach treats a crowd as a set of different objects. The segmented objects are then tracked through the video frames and their behaviors are then inferred. Despite compelling results in several crowd behavior problems [15, 16], these approaches are limited to low-density crowd scenarios as they rely on detection and tracking of each individual and object in a crowd and are not capable of handling severe occlusion and clutter in high-density crowd scenes. The holistic approaches, on the other hand, do not aim to separately detect and track each individual/object in a scene. Instead, they treat a crowd as a single entity and try to employ low/medium level visual features extracted from video frames to analyze the crowd scene as a whole [17–21].

1.1 Method overview

In this paper, we propose two novel video descriptors for abnormal behavior detection and localization. In our first model, spatio-temporal abnormalities in densely crowded scenes are detected by a new tracklet based descriptor. In this scheme, we first divide the video sequence to saptio-temporal 3D patches in order to derive more detailed motion information. The short trajectories (tracklets) are then extracted by tracking randomly selected points in video frames within a short period of time. Using the orientation and magnitude of extracted tracklets, which are two most important features used in the task of abnormality detection, we compute our proposed one-dimensional descriptor.

In a nutshell, a video sequence is segmented to spatio-temporal patches. Then, using motion trajectories represented by a set of tracklets [22], each patch is being described. Unlike most of the standard approaches which describe frames with dense descriptors such as optical flow [23] and interaction force [20], we define spatio-temporal patches and gather the statistics on trajectories that intersect them. More specifically, the orientation and magnitude of such intersecting tracklets are encoded in a histogram that we called *simplified histogram of oriented tracklets* (SHOT). In our first method, no clustering is needed to create a codebook and the histogram itself can describe a video frame. As a result, the proposed descriptor is built much simpler than the other state-of-the-art frameworks and is shown to have better results compared to them. Under the assumption that abnormalities are outliers of normal situations and considering the fact that we have only access to normal samples for training, which is seemed to be a realistic assumption in the real world, we employed one-class SVM generative model for behavior classification. By combining our proposed descriptor with the dense optical flow [24] (we call it as sHOT+DOF), we also propose our second novel framework to accurately localize the abnormal behavior areas in abnormal frames.

The objects of the study, in our case, are pedestrians: people walking with almost constant speed depending on the available space, following the curvature of the street or could be even circular (religious festivals). People usually move at a constant speed (an oscillator, typically an inverted pendulum with a fixed frequency) alone or together with other people, forming groups with similar dynamics. Our proposed models are evaluated on some state-of-the-art crowd datasets, namely UCSD [25], UMN [20] and Violence in Crowd [26]. We compared the proposed frameworks to several descriptors such as the SFM [20] and histogram of optical flow (HOF) [27], etc. Our method reached very competitive accuracy while being much simpler that of other techniques in the literature.

1.2 Contributions

In general, our paper has four contributions: (1) we introduce a novel descriptor for abnormality detection, namely *simplified Histogram of Oriented Tracklest* (*sHOT*), which is much simpler than other state-of-the-art models and is shown to have better results. (2) We present a novel framework by combining sHOT with Dense Optical Flow (DOF) [24] model which can localize the area of abnormal behavior occurrence in a frame. (3) Since abnormal behavior samples are hardly accessible in real-world crowd scenarios and are not sufficiently available at training time, we evaluate the proposed models using one-class SVM generative model which only needs normal samples at training time. (4) The proposed methods are validated on challenging abnormality detection datasets and the results show the superiority of our method compared to the state-of-the-art methods.

2 Related works

Considering object-based methods, some proposed works made substantial efforts to improve robustness issues. In [28], Zhao and Nevatia employed 3D models to detect persons in the observed scene and then applied a probabilistic framework to track extracted features from the individuals. On the contrary, some other approaches have adopted well-known KLT algorithm for tracking feature points in the observed scene. In these models, after clustering extracted

Fig. 1 *Left* Crowded street in China, *middle* Stockholm Marathon, and *right* Mecca (Saudi Arabia)

trajectories using space proximity, it is more simple to reach one-to-one association between individuals and trajectory clusters [15, 29]. However, this is a strong assumption hardly verified in a crowd scene.

In holistic approaches, on the other hand, spatio-temporal gradients and optical flows are employed as typical features. In [19, 30], Krausz and Bauckhage used optical flow histograms to demonstrate the global movements in a crowd. Both extracted histograms and some heuristic rules were then adopted to detect some dangerous crowd behaviors. More advanced methods, on the other hand, have adopted models extracted from fluid dynamics in order to model a crowd as a group of moving particles. Along with Social Force Model (SFM) [31], it is possible to elaborate the behavior of a crowd as a result of interaction of moving particles. In [16], the SFM is adopted to detect global abnormal behaviors and estimate local abnormal behaviors. Shah et al. in [32] proposes a method to classify the critical points of a continuous dynamical system for abnormality detection, which is applicable for high-density crowds such as religious festivals and marathons [33]. Figure 1 shows a few samples of these cases.

Besides, several approaches made noticeable efforts to decrease the complexity of crowd behavior analysis by partitioning a given video in spatio-temporal patches. For instance, In [17, 18] spatio-temporal gradients are derived from pixels of a frame. Then, the gradients of a 3D patch are modeled by spatio-temporal Motion Pattern Models, which are basically 3D gaussian clusters of gradients [34] and are used to group observed gradients at training time in separate cluster centers. The Kullback–Leibler distance [34, 35] is then used to choose the training cluster centers with the closest gradient distribution.

Since PCA spaces only can model the appearance of a given patch texture, an extension of PCA-based representations is introduced in [25] to model the observed motion in each spatio-temporal patch using dynamic textures. Dynamic textures are also able to show the statistically valid transitions between textures in a patch. In this model, all the possible dynamic textures in each patch are demonstrated with a Mixture of Dynamic Textures model, which gives the probability of a test patch to be abnormal. By applying this framework, it was shown that not only temporal abnormalities but also pure appearance abnormalities can be detected. In the same work, the authors introduced an interesting definition of spatial saliency based on mutual information [36] between features and foreground/background classes. In recent years, some deep learning techniques, attribute-based models and measure-based frameworks have been proposed for abnormal behavior detection [4, 5, 7, 8, 13, 37, 38]. Rabiee et al. in [37, 38] used crowd emotions as mid-level information to fill the semantic gap between low-level motion/appearance features and high-level concept of crowd behaviors and improved the crowd behavior classification results compared to works in [1, 17, 19–21, 25]. Mousavi et al. in [7] introduced a measure to capture the commotion of a crowd motion for the task of abnormality detection. On the other hand, deep learning techniques [4, 5, 8, 13] usually employ learning networks such as PCAnet, D-IncSFA, CNN, etc. to extract semantic information from crowd scene. By combining the semantic information with different low-level visual features such as optical flows and oriented gradients, these methods can detect abnormal behaviors more accurate. However, since these techniques need a large amount of training data, they are really time-consuming and can hardly be considered as realistic approaches for modeling and detecting crowd abnormal behaviors in real-time scenarios.

The rest of the paper is organized as follows. In Sect. 3 we describe the proposed *simplified Histogram of Oriented Tracklets* (sHOT) model for abnormality detection. In Sect. 4 abnormal behavior detection schemes are elaborated using sHOT and sHOT-*Dense Optical Flow* (DOF) models. The experiments regarding our proposed approaches and a discussion on the obtained results are presented in Sect. 5.

3 Simplified histogram of oriented tracklet (sHOT)

Tracklets are compact spatio-temporal representations of moving rigid objects [22]. They demonstrate fragments of

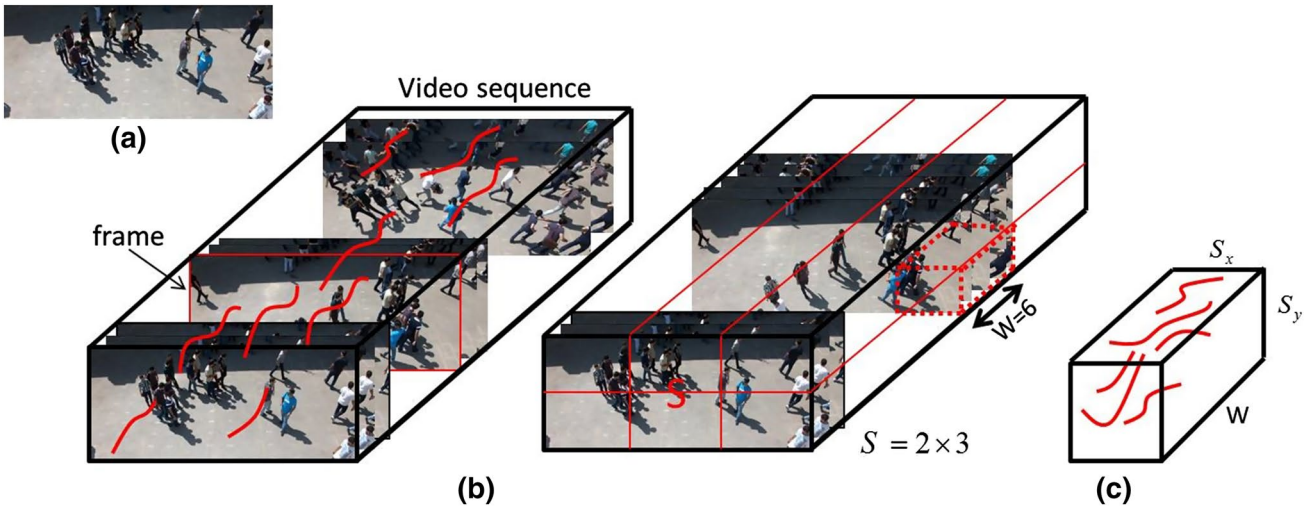

Fig. 2 **a** Interest points are detected and tracked through T frames [42] to form tracklets. **b** The video sequence is spatially divided in non-overlapping 3D patches to compute sHOT, in the figure $S = 2 \times 3$. For each frame a temporal window stride is then considered. **c** The sHOT descriptor is computed from portions of tracklet in each 3D patch of size $S_x \times S_y \times W$. In a sense it represents the expected motion patterns in 3D patches of a video

an entire trajectory corresponding to the motion pattern of an individual point, generated by the frame-wise association procedure between point localization results in the neighbor frames. Tracklets capture the evolution of patches and were originally developed to model the human movements for the task of action recognition in video sequences [22]. More specifically, a tracklet is represented as a sequence of points in the spatio-temporal space as:

$$tr = (p_1, \ldots, p_t, \ldots p_T) \tag{1}$$

where each p_t represents two-dimensional coordinates (x_t, y_t) of the tth point of the tracklet in the tth frame and T shows the length of each tracklet. Tracklets are formed by choosing areas (or points) of interest via a feature detector and by tracking them over a short period of time. So, it can be said that tracklets represent a trade-off between optical flow and object tracking schemes.

Since different regions usually indicate different patterns of motion, we present a histogram-based descriptor that captures the statistics on trajectories of objects/individuals passing through a spatio-temporal 3D patch. We call this new descriptor as simplified Histogram of Oriented Tracklets (sHOT), which is clearly inspired by the recent success of histogram of features in crowd behavior analysis [39, 40].

3.1 Tracklet extraction

In a given video sequence, all the tracklets are derived using standard OpenCV code.[1] More specifically, the SIFT algorithm is adopted to detect possible salient points in each frame [41]. Then, the KLT algorithm is employed to track the salient points for T frames. Finally, the spatial coordinates of the tracked points are used to form the tracklets set$\tau = \{tr^n\}_{n=1}^{N}$, where N denotes the number of all extracted tracklets and tr^n refers to the nth tracklet in the video sequence. The length of the tracklets T depends on the frame-rate of the video sequence, the camera position and the intensity of the motion-patterns available in the crowd scene.

3.2 sHOT computation

As aforementioned, tracklets are short sequences of two-dimensional points represented as$tr = \{(x_1, y_1) \ldots (x_T, y_T)\}$. For tth point in a tracklet, the local magnitude can be computed as:

$$m_t = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \tag{2}$$

The process of sHOT computation starts by splitting the video sequence into spatio-temporal 3D patches of size $S_x \times S_y \times W$ with overlapping 3D patches in the spatial domain; this is demonstrated in Fig. 2b. From now on, we will use the apex (i,s) to address the portion of the tracklet i that intersects the 3D patch S.

Under the hypothesis that sudden changes in pedestrian directions and their high speed in the presence of non-pedestrian moving objects are considered as abnormal behaviors, we extract two important parameters for each

[1] http://www.ces.clemson.edu/stb/klt/.

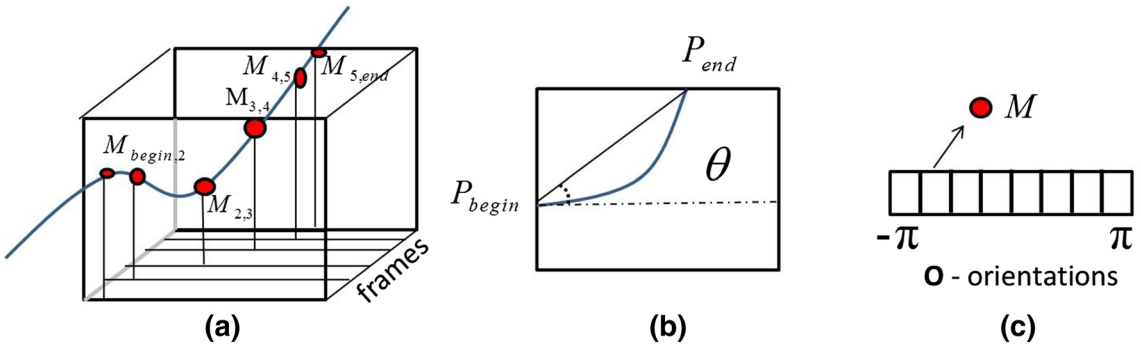

Fig. 3 The procedure of simplified histogram of oriented tracklet computation. **a** *Red circles* are represented as the magnitudes of the portion of a tracklet in a 3D patch and are summed up to form the corresponding magnitude. **b** The entry and exit point of a tracklet are used to compute the orientation. **c** Each tracklet presents a contribution. Specifically, the summation value of magnitudes of each tracklet is located in the appropriate orientation bin

tracklet, namely orientation and magnitude for abnormality detection. The orientation and magnitude of all the portions (fragment) of tracklets that intersect 3D patch s are computed by Eqs. (3) and (4) respectively.

$$\theta^{i,s} = \arctan \frac{\left(y^{i,s}_{end} - y^{i,s}_{begin}\right)}{\left(x^{i,s}_{end} - x^{i,s}_{begin}\right)} \tag{3}$$

$$M^{i,s} = sum\left\{ m^{(i,s)}_{t} \right\}_{t \in W} \tag{4}$$

Using the entry and exit points of the tracklet i in/from 3D patch S which are respectively indexed by $\left(x^{i,s}_{begin}, y^{i,s}_{begin}\right)$ and $\left(x^{i,s}_{end}, y^{i,s}_{end}\right)$, the orientation of each intersecting tracklet is achieved by Eq. 3. Also, the magnitudes of all consecutive points in a tracklet i which intersect 3D patch S (i.e.,$m^{(i,s)}_{t}$) are summed up to form the corresponding magnitude using Eq. 4. Since the magnitudes extracted from the portion of a tracklet can be noisy, we sum up them to compute the histogram of magnitudes rather than choose the noisy maximum one. The process of computing the magnitude and orientation of a tracklet within a 3D patch and creating a 1D histogram of a tracklet is illustrated in Fig. 3.

A sHOT is the one-dimensional version of HOT [39]. Given a set of magnitude-orientation pairs $\{\theta^n, M^n\}$, a sHOT descriptor is computed by the summation of magnitudes whose corresponding orientations fall in orientation bins. The bins of histogram $H^{s}_{\theta,m}$ are populated by simply counting how many times a magnitude-orientation pair $\{\theta, m\}$ is observed. This process is followed by a normalization to form a non-biased oriented histogram. Similar to HOT, sHOT is computed for each 3D patch S, which is temporally centered at each frame f in the form of $h^{s,f}_{\theta,m}$.

In a normal crowd, there are no significant changes in direction and speed of individuals and the sHOT descriptors are similar to each other. However, when there are sudden changes in directions and speeds of individuals, the corresponding magnitudes and orientations are unusually high and the corresponding sHOT descriptors are different from the normal ones.

For objects that occupy larger regions than pedestrians, a different mechanism should be adopted in order to reflect the presence of coherent tracklets (maybe by working on multi-resolution images). In this situation, the low-resolution sHOT should have a higher number of tracklets. The representation of sHOT on UCSD crowd dataset is demonstrated in Fig. 4.

4 Abnormal behavior detection

In this section, abnormal behavior detection is accomplished in two forms, namely frame level and pixel level. In the following, we elaborate these forms precisely.

4.1 Frame-level abnormal behavior detection

Unlike the usual classification tasks in computer vision, for crowd abnormal behavior detection we are not capable of collecting enough abnormal samples from movies, web pages, etc. In other words, although it is always feasible to gather a lot of normal behavior samples from different sources, it is time-consuming, very costly and even impossible to collect different abnormal behavior cases in real-world crowds. Since we aim to detect abnormal behavior samples rather than normal behavior ones, we learn a generative model for normal behavior samples and classify a given video frame as abnormal one if it largely deviates from the learned model.

In this framework, we employ the one-class SVM to determine what is normal in terms of co-occurrences

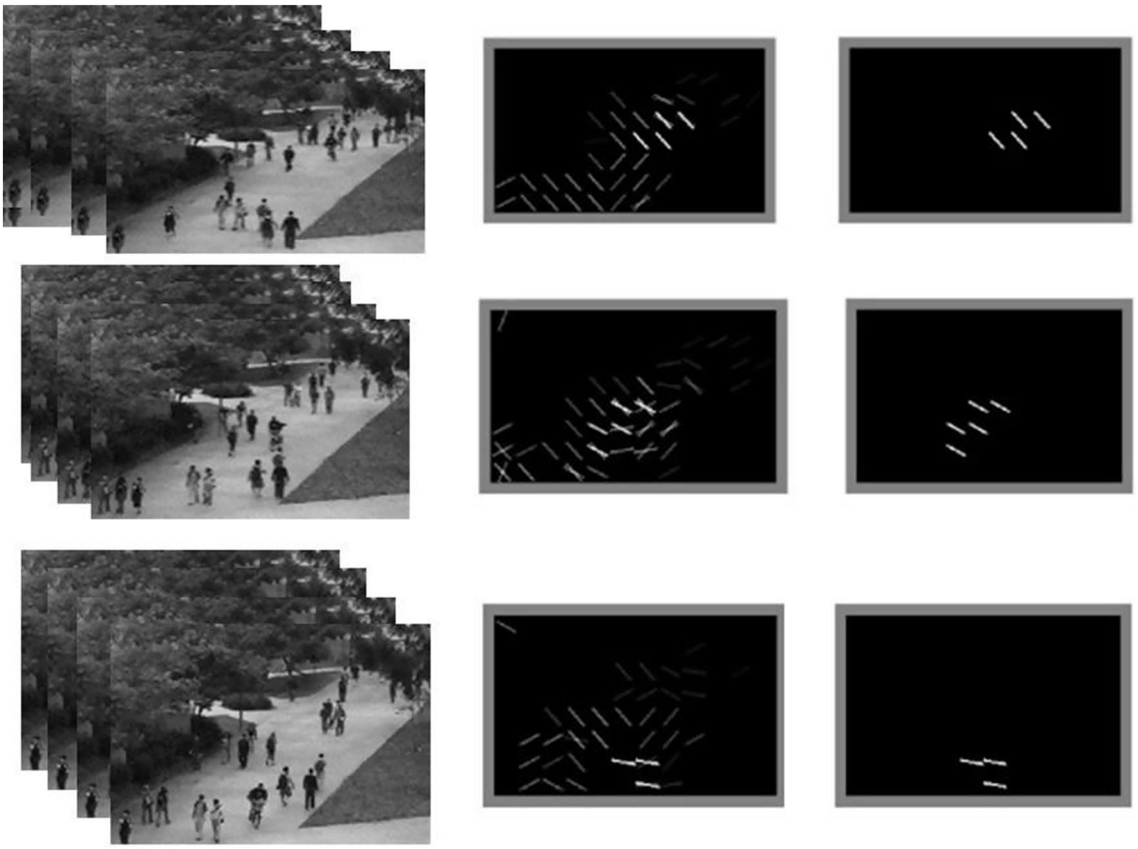

Fig. 4 Representation of sHOT on UCSD dataset (saliency points show abnormal event in the video sequence)

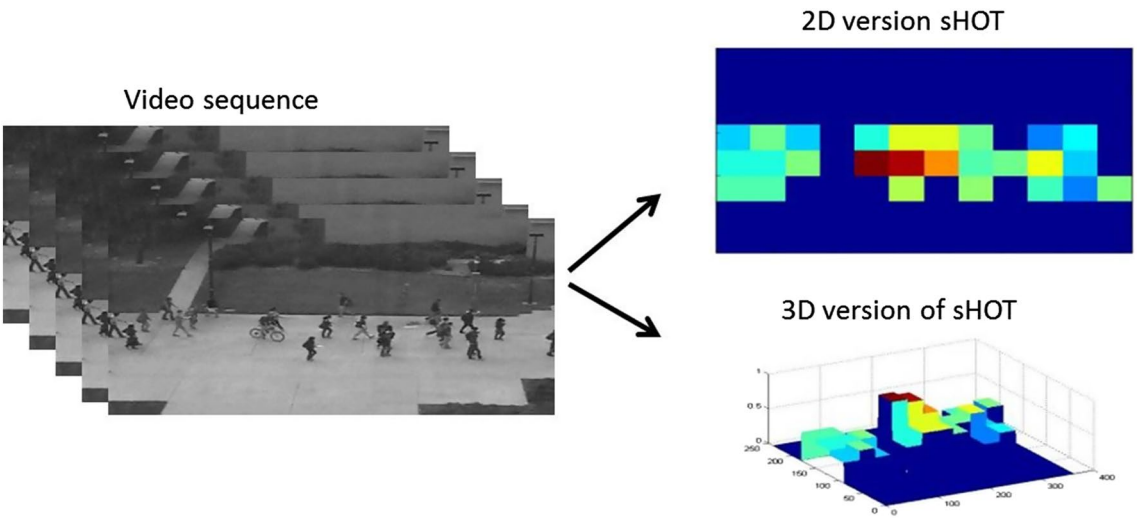

Fig. 5 Representation of 2D and 3D version of sHOT on UCSD dataset

between the motion pattern features. Given a set of one-dimensional histogram $h^{s,f}_{\theta,m}$ for each frame $f = 1, 2, \ldots, F$, a one-class SVM training corpus D is built using the strategy in Eq. 5. In this strategy, sHOTs derived from all the different 3D patches are concatenated in a single descriptor to preserve the spatial information of each frame:

$$D^f = \left\{ H^{1,f}_{o,m} \middle| H^{2,f}_{o,m} \middle| \ldots \middle| H^{s,f}_{o,m} \right\} \text{ and } \in D = \left\{ D^f \right\}_{f=1}^{F} \quad (5)$$

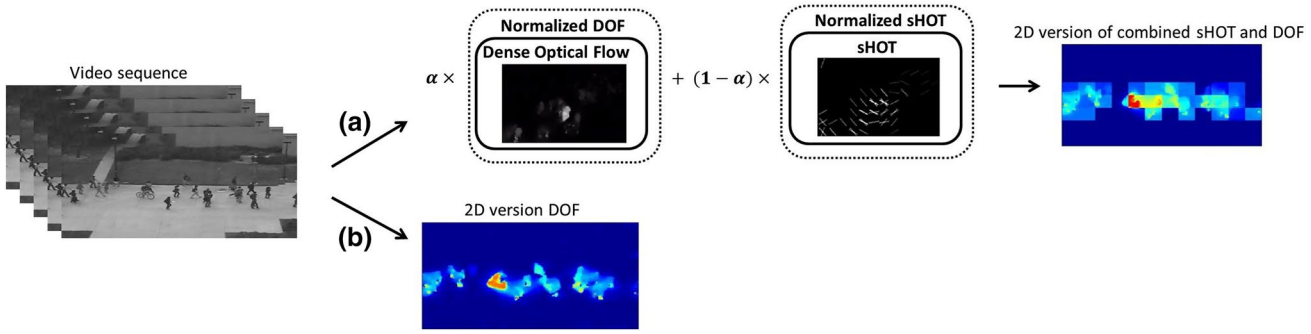

Fig. 6 **a** Representation of 2D version of combined sHOT-DOF for applied α value and threshold value β on UCSD dataset, **b** representation of 2D version of DOF

In this case, one-class SVM captures correlations between motion patterns that occur in different 3D patches of the scene. The representation of 2D version and 3D version of sHOT on sample frames of UCSD crowd dataset is illustrated in Fig. 5.

4.2 Pixel-level abnormal behavior detection (localization)

Saliency detection in computer vision was first introduced in [17, 33] for the center-surround manner and spatial abnormality detection. Salient locations are described as those attributes which make them remarkable from their surround. After acquiring an appropriate feature, saliency represents an objective definition of specific abnormality. In the pixel-level abnormality detection scheme, we localize the abnormal saliency areas in a frame using our proposed combined dense optical flow (DOF) and sHOT (sHOT + DOF) model. More specifically, after obtaining feature f (saliency score), the noisy features caused by fast movements of human body parts are removed in each frame. Then, the expected center for maximal saliency is distributed between features. Although sHOT gives us the saliency in each 3D patch of a video sequence, we apply dense optical flow algorithm to find exact abnormal pixels.

Since dense optical flow is capable of describing crowd motions between two frames and sHOT is useful for detecting abnormal part of each spatio-temporal window, the best approach to detect and localize the abnormality in a crowd is our sHOT + DOF model. To combine these two models, we need to first normalize dense DOF and sHOT. We define a variable $0 \leqslant \alpha \leqslant 1$, which trades-off between dense optical flows and sHOT as follow:

$$\begin{aligned} &(\alpha)sHOT(i,j) + (1-\alpha)DOF(i,j) = f(i,j) \\ &0 \leqslant \alpha \leqslant 1 \end{aligned} \tag{6}$$

We can find the best value for α by cross validation to create the matrices. This method will be continued for all videos. Finally, we find the threshold value β of each matrix to detect the saliency part of abnormality as best as possible.

$$\begin{aligned} &f(i,j) \geqslant \beta \\ &0 \leqslant \beta \leqslant 1 \end{aligned} \tag{7}$$

The representation of the 2D version of DOF and combined sHOT-DOF on UCSD crowd dataset are illustrated in Fig. 6.

5 Experimental results and discussion

In this section, we first introduce the crowd benchmarks which are used to evaluate our proposed models. Then, we compare our models with state-of-the-art methods in the literature [15, 20, 21, 25–27, 36, 43–46] on aforementioned datasets.

5.1 Crowd datasets

Three publicly available dataset are used for our evaluation, including UCSD [25], UMN [20] and Violence-in-crowds [26]. As aforementioned, we extract tracklets using KLT algorithm [41]. More specifically, interest points are selected at first frame and are tracked over T equals to 11 frames. Interest point re-initialization procedure is done in case of tracking failures.

UCSD Dataset consists of two subsets. The first subset, which is denoted by "PED1", contains clips of 158×238 pixels and indicates groups of individuals walking toward and away from the camera, with a certain degree of perspective distortion. The second subset, denoted by "PED2", has a spatial resolution of 240×360 pixels and indicates a scene where most pedestrians move horizontally. The video footage of each scene is divided into clips

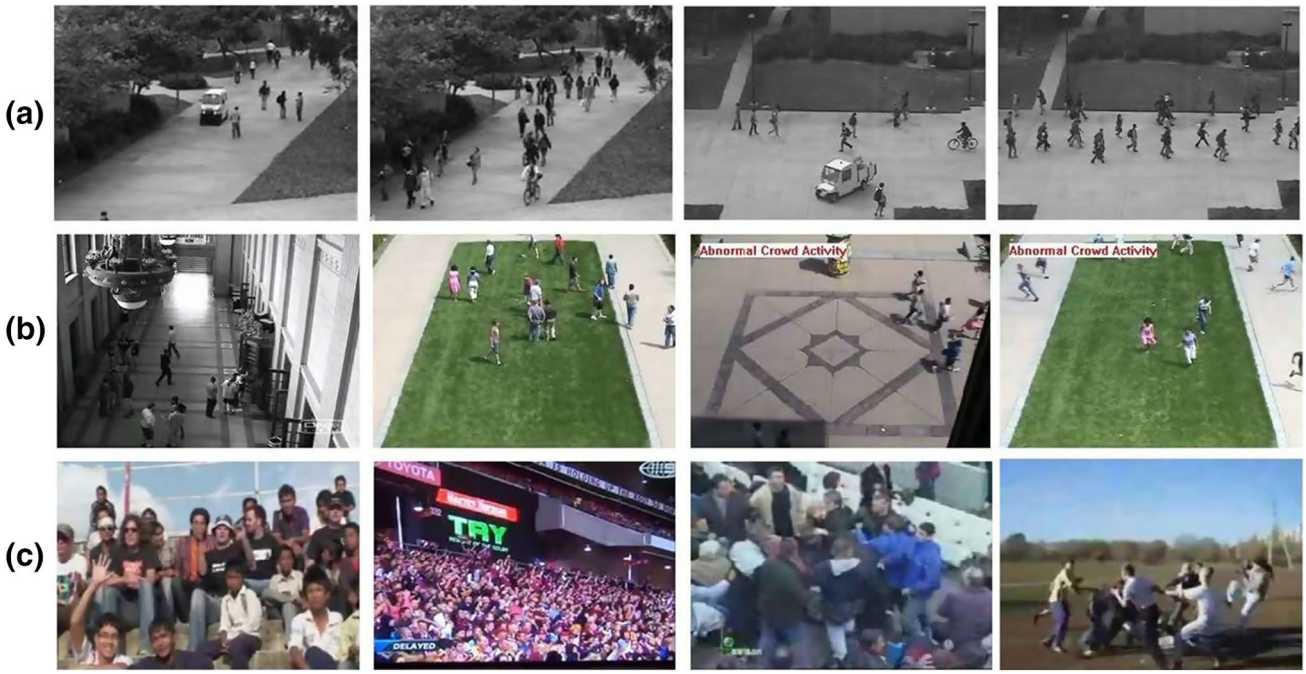

Fig. 7 **a** Normal and abnormal frames from UCSD dataset. **b** Normal and abnormal frames from the three different scenarios of the UMN dataset. **c** Normal and violent behaviors from the violence-in-crowds dataset captured in different scenes

of 120–200 frames. Some of these (34 in Ped1 and 16 in Ped2) are to be used as the training set for the normal condition. The test set, on the other hand, contains clips (36 for Ped1 and 12 for Ped2) with both normal (around 5500) and abnormal (around 3400) frames (see Fig. 7a). We only used the frame-level abnormality detection parts of this dataset.

UMN dataset contains 11 different scenarios of panic and normal conditions in three different indoor and outdoor situations. Figure 7b shows some samples of UMN dataset.

Violent-Flows dataset (see Fig. 7c) is composed of real-world scenarios and video sequences of crowd violence, along with standard benchmark protocols designed to test violent and non-violent classification. It is divided into five subsets: half violent crowd behavior and half non-violent crowd behavior, which are available at training time.

5.2 Proposed model setting

Like other frameworks, there are few parameters and constants to tweak, some of them are fixed and others depend on the monitored scene. These parameters are demonstrated in Fig. 2, containing temporal window W, length of the tracklet T, tessellation of the frame S and the quantization bins O. We quantized tracklets orientation in $O = 8$ like in [39]. Each bin corresponds to a range of orientation angles which fall in $[-\pi, \pi]$. We equally divide the $[-\pi, \pi]$ interval into eight ranges. Unlike [39], we changed the temporal window to $W = \{5, 11\}$ frames. We also considered three different spatial tessellation, namely $S = 2 \times 3$, $S = 4 \times 6$ and $S = 8 \times 12$.We obtained magnitudes and orientations of all the tracklets which intersect each spatio-temporal 3D patch with the size of $S_x \times S_y \times W$ and create corresponding 1D sHOT. The value of each bin of histogram is determined by the sum of magnitudes of tracklets that their orientation falls into the range of that bin. Using the extracted histograms for each 3D patch, we finally obtained sHOT and combined sHOT and DOF by employing the methods mentioned in Sect. 3.

The sHOT model shows the changes of each orientation bin with respect to the dominant motions of spatio-temporal window. As aforementioned, it would be hard and costly to collect a huge training set containing all possible abnormal behaviors. As a result, unlike in [6, 47], a one-class SVM classifier is employed as a generative model for the task of abnormality detection on crowd datasets.

5.3 Detection performance

In the following, we evaluate the detection and localization performance of our approaches comparing with the state-of-the-arts. For UCSD dataset, we applied its standard train-test partition for this experiment. We reported the best performance (smallest EER). For UMN and

Table 1 Equal error rates (EER) and accuracy on UCSD dataset (ped1 and ped2) using standard testing protocol

Method	Ped1 (frame)		Ped1 (pixel)		Ped2 (pixel)	
	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)
MDT [25]	25	81.8	58	44.1	25	82.9
SFM [20]	31	67.5	79	19.7	42	55.6
LMH [43]	38.9	70.1	80	37	40	64.4
AMDN [48]	22	84.9	47.1	57	24	81.5
sHOT	21.3	82.8	49	51	20.9	82.9

Violence-in-Crowds datasets, the parameters are fixed to $W=11$ and $O=8$.

5.3.1 Evaluation on UCSD dataset

For the UCSD dataset, considering its standard train/test partitioning in [39], the EERs are computed on the frame-level by our method and ones in [20, 25, 43, 48] using the extracted one-class SVM confidence scores. Moreover, for derived 3D patches, the sHOTs are computed based on confidence scores and after defining appropriate thresholds, the best results for AUCs are reported at the pixel-level (localization). The results on ped1 and ped2 are presented in Table 1. Results show that in most of the cases the proposed method hit the abnormality correctly in terms of detection and localization. Only in the case of AMDN [48], our measure achieved lower accuracy in abnormality detection for ped1(frame), while the abnormality detection performance always does better than all state-of-the-art methods. Note that the proposed method is not taking advantage of any kind of learning in comparison with the others. Moreover, the EERs of other methods are reported as the best results across all the possible experimental arrangements.

In Fig. 8, the qualitative results of sHOT for abnormal object localization are presented on a few sample frames of UCSD dataset (ped1 and ped2). As can be seen, abnormal objects are detected accurately in each frame.

According to what was mentioned in Sect. 3, we use combined sHOT and DOF descriptors to exploit the benefits of each of them. By extracting the best threshold β, some of the best results for UCSD dataset (ped1 and ped2) are visualized in Fig. 9. As can be seen, this method can accurately localize the abnormal objects and is an efficient scheme for abnormality localization.

5.3.2 Evaluation on UMN dataset

In this experiment, we compared the proposed sHOT with social force model (SFM) [20], sparse reconstruction (SR) [44], optical flow (OF) [27], PSO-SFM [15], Commotion [7], and using aforementioned setup. To have a consistent evaluation, we exploited a protocol by separately considering UMN three scenes. The results on each scene and complete dataset are presented in Table 2 in terms of area under the ROC curve (AUC). On the contrary to earlier approaches which used latent Dirichlet allocation (LDA) [39, 47], we used one-class SVM which is a generative model for abnormal behavior classification. The results show that our model reports much better accuracy on this dataset for both scene-based and all-scenes evaluations.

5.3.3 Evaluation on violent in crowd dataset

In the final experiment, we trained a one-class SVM with the linear kernel on a set of normal and abnormal video sequences from Violence-in-Crowd dataset across a five-fold cross-validation. Here, we aim to assign a normal or abnormal label to an input video rather a frame. The video level descriptor $D^v$of an input video V is simply computed by:

$$D^v = \sum_{f \in V} D^f \tag{8}$$

To train the one-class SVM, the training set as $D = \{D^v\}_{v=1}^{N}$ are formed where N is the number of positive and negative training videos. The result accuracies of this experiment are reported for sHOT and state-of-the-art methods [21, 26, 27, 45, 46, 49] in Table 3. Results show that the best performance belongs to our approach with 82.2%.

6 Discussion

By applying the proposed interest point tracking methods in crowded and semi-crowded environments, where camera calibration is not feasible and there is severe occlusions and background clutter, crowd abnormality detection is shown to be superior in comparison with other individual-based detector schemes. In this paper, we show that our method can deal with abnormality detection in various crowd densities, and evaluate our method on medium-level density crowd and dense crowd

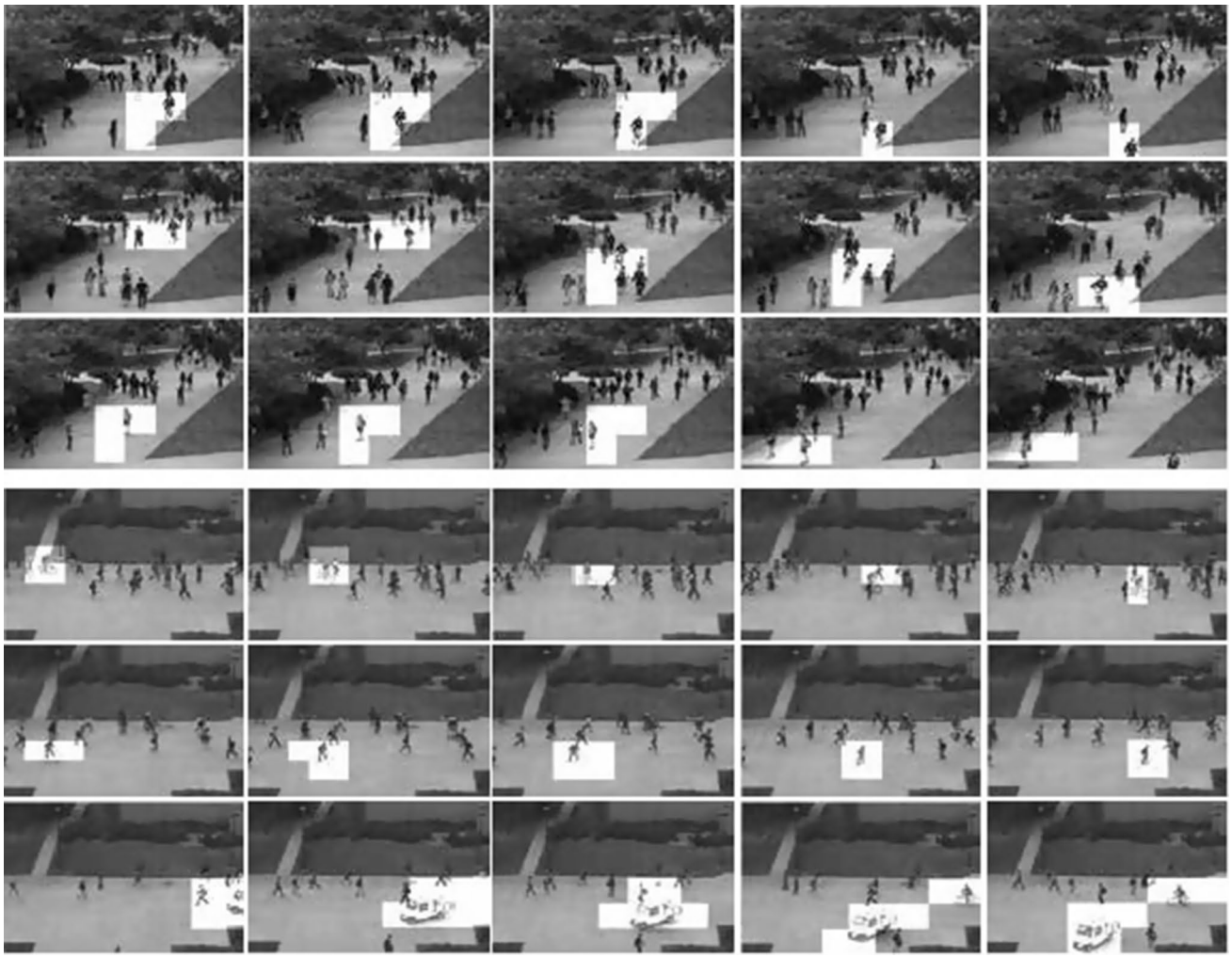

Fig. 8 The qualitative results of abnormal object localization on sample frames using sHOT on UCSD dataset (ped1 and ped2)

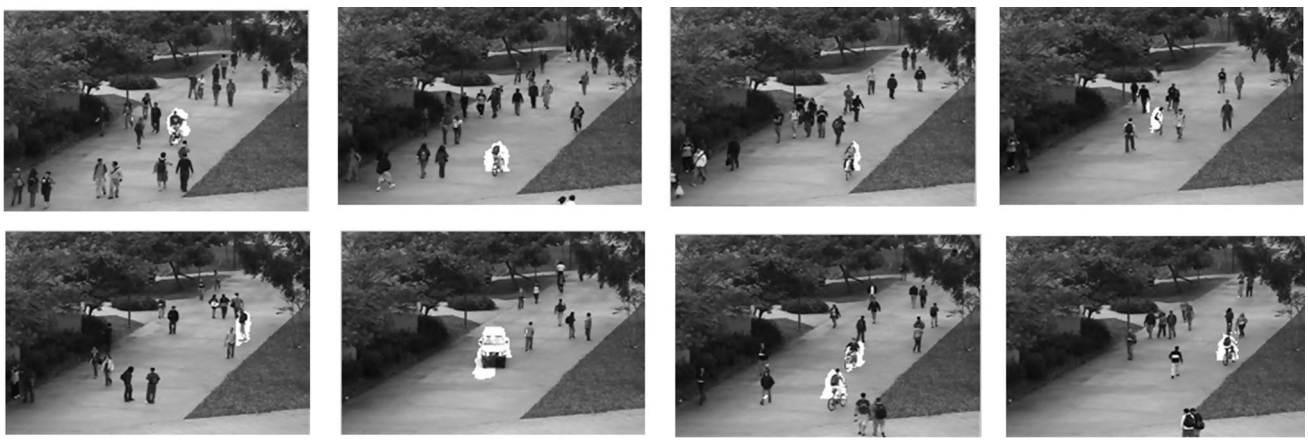

Fig. 9 The visualization of object localization using sHOT + DOF method under the best-defined threshold on some sample frames of UCSD dataset (ped1 and ped2)

Table 2 AUCs of different methods on UMN dataset in scene-1, scene-2, scene-3 and all the scenes

Dataset	SFM [20]	OF [27]	SR [44]	PSO-SFM [15]	Commotion [7]	sHOT
Scene-1	0.990	0.964	0.995	0.996	–	0.998
Scene-2	0.949	0.906	0.975	0.993	–	0.995
Scene-3	0.989	0.967	0.964	0.999	–	0.997
All scenes	0.960	0.840	0.978	0.996	0.988	0.996

Table 3 Classification results on Violence Crowd dataset for different methods

Method	Accuracy (%)
LTP [46]	71.53
HOG [49]	57.43
HOF [27]	58.53
HNF [45]	56.52
ViF [26]	81.30
DT [21]	78.21
sHOT	82.2

scenarios. However, in the scenarios which are capable of using person detection techniques, our method may not be satisfying as person-detector-based methods, since we are detecting the crowd behaviors rather than individual behaviors.

7 Conclusion

In this paper, we first introduced the simplified Histogram of Oriented Tracklet (sHOT) model for the task of abnormality detection in crowded scenes. This new descriptor contains both orientation and magnitude information in a single feature, which is often reached by combining multiple descriptors. We also combined sHOT and Dense Optical Flow (DOF) to form a novel abnormal behavior descriptor in order to localize abnormalities in a crowd. Since the abnormal samples are hardly accessible in real-worlds, we employed the generative one-class SVM to learn our models which is a more realistic approach. We showed that our proposed feature descriptors can detect behavior abnormalities much better than state-of-the-arts. Our models are very simple and efficient and can easily be reproduced. We plan to combine our models with other descriptors to reach more promising results as future work.

References

1. Chen K, Gong S, Xiang T, Change Loy C (2013) Cumulative attribute space for age and crowd density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2467–2474
2. Wu S, San Wong H (2012) Crowd motion partitioning in a scattered motion field. IEEE Trans Syst Man Cybern Part B Cybern 42:1443–1454
3. Bera A, Manocha D (2014) Realtime multilevel crowd tracking using reciprocal velocity obstacles. arXiv preprint arXiv:1402.2826
4. Fang Z, Fei F, Fang Y, Lee C, Xiong N, Shu L et al (2016) Abnormal event detection in crowded scenes based on deep learning. Multimed Tools Appl, 1–23
5. Hu X, Hu S, Huang Y, Zhang H, Wu H (2016) Video anomaly detection using deep incremental slow feature analysis network. IET Comput Vis 10:265
6. Mousavi H, Nabi M, Galoogahi HK, Perina A, Murino V (2015) Abnormality detection with improved histogram of oriented tracklets. In: International Conference on Image Analysis and Processing, pp 722–732
7. Mousavi H, Nabi M, Kiani H, Perina A, Murino V (2015) Crowd motion monitoring using tracklet-based commotion measure. In: Image Processing (ICIP), 2015 IEEE International Conference, pp 2354–2358
8. Ravanbakhsh M, Nabi M, Mousavi H, Sangineto E, Sebe N (2016) Plug-and-play CNN for crowd motion analysis: an application in abnormal event detection. arXiv preprint arXiv:1610.00307
9. Sabokrou M, Fathy M, Hoseini M (2016) Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electron Lett 52:1122–1124
10. Sabokrou M, Fathy M, Hoseini M, Klette R (2015) Real-time anomaly detection and localization in crowded scenes. Proceedings of the IEEE Conference on Computer Vision Pattern Recognition Workshops, pp 56–62
11. Wang B, Ye M, Li X, Zhao F, Ding J (2012) Abnormal crowd behavior detection using high-frequency and spatio-temporal features. Mach Vis Appl 23:501–511
12. Wu S, Wong H-S, Yu Z (2014) A Bayesian model for crowd escape behavior detection. IEEE Trans Circuits Syst Video Technol 24:85–98
13. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z (2016) Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process Image Commun 47:358–368
14. Wijermans A, Jorna R, Jager E, Van Vliet T (2007) Modelling crowd dynamics influence factors related to the probability of a riot
15. Rabaud V, Belongie S (2006) Counting crowded moving objects. 2006 IEEE Computer Society Conference on Computer Vision Pattern Recognition (CVPR'06), pp 705–711
16. Rittscher J, Tu PH, Krahnstoever N (2005) Simultaneous estimation of segmentation and shape. 2005 IEEE Computer Society Conference on Computer Vision Pattern Recognition (CVPR'05), pp 486–493
17. Kratz L, Nishino K (2009) Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models.

Computer Vision Pattern Recognition CVPR 2009 IEEE Conference, pp 1446–1453
18. Kratz L, Nishino K (2010) Tracking with local spatio-temporal motion patterns in extremely crowded scenes. Computer Vision Pattern Recognition (CVPR) 2010 IEEE Conference, pp 693–700
19. Krausz B, Bauckhage C (2011) Analyzing pedestrian behavior in crowds for automatic detection of congestions. Computer vision workshops (ICCV workshops) 2011 IEEE international conference, pp 144–149
20. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. Computer Vision Pattern Recognition CVPR 2009 IEEE Conference, pp 935–942
21. Wang H, Kläser A, Schmid C, Liu C-L (2011) Action recognition by dense trajectories. Computer Vision Pattern Recognition (CVPR) 2011 IEEE Conference, pp 3169–3176
22. Raptis M, Soatto S (2010) Tracklet descriptors for action modeling and video analysis. In: European conference on computer vision, pp 577–590
23. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. IJCAI, 674–679
24. Alvarez L, Weickert J, Sánchez J (2000) Reliable estimation of dense optical flow fields with large displacements. Int J Comput Vis 39:41–56
25. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. CVPR, p 250
26. Hassner T, Itcher Y, Kliper-Gross O (2012) Violent flows: Real-time detection of violent crowd behavior. 2012 IEEE Computer Society Conference on Computer Vision Pattern Recognition Workshops, pp 1–6
27. Wang T, Snoussi H (2012) Histograms of optical flow orientation for visual abnormal events detection. Advanced video signal-based surveillance (AVSS) 2012 IEEE Ninth International Conference, pp 13–18
28. Zhao T, Nevatia R (2003) Bayesian human segmentation in crowded situations, vol 2. Computer Vision Pattern Recognition 2003 Proceedings 2003 IEEE Computer Society Conference, pp 459–466
29. Shi J, Tomasi C (1994) Good features to track. Computer Vision Pattern Recognition 1994 Proceedings CVPR'94 1994 IEEE Computer Society Conference, pp 593–600
30. Krausz B, Bauckhage C (2012) Loveparade 2010: Automatic video analysis of a crowd disaster. Comput Vis Image Underst 116:307–319
31. Helbing D, Molnar P (1995) Social force model for pedestrian dynamics. Phys Rev E 51:4282
32. Solmaz B, Moore BE, Shah M (2012) Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Trans Pattern Anal Mach Intell 34:2064–2070
33. Hu M, Ali S, Shah M (2008) Detecting global motion patterns in complex videos. Pattern Recognition ICPR 2008 19th International Conference, pp 1–5
34. Duda RO, Hart PE, Stork DG (2012) Pattern classification. Wiley, Hoboken
35. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
36. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36:18–32
37. Rabiee H, Haddadnia J, Mousavi H, Nabi M, Murino V, Sebe N (2016) Emotion-based crowd representation for abnormality detection. arXiv preprint: arXiv:1607.07646
38. Rabiee H, Haddadnia J, Mousavi H (2016) Crowd behavior representation: an attribute-based approach. SpringerPlus 5:1179
39. Mousavi H, Mohammadi S, Perina A, Chellali R, Murino V (2015) Analyzing tracklets for the detection of abnormal crowd behavior. 2015 IEEE Winter Conference on Applications of Computer Vision, pp 148–155
40. Ravanbakhsh M, Mousavi H, Rastegari M, Murino V, Davis LS (2015) Action recognition with image based CNN features. arXiv preprint arXiv:1512.03980
41. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
42. Rabiee H, Haddadnia J, Mousavi H, Kalantarzadeh M, Nabi M, Murino V (2016) Novel dataset for fine-grained abnormal behavior understanding in crowd. Advanced video signal based surveillance (AVSS) 13th IEEE International Conference, pp 95–101
43. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell 30:555–560
44. Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. Computer Vision Pattern Recognition (CVPR) 2011 IEEE Conference, pp 3449–3456
45. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. Computer Vision Pattern Recognition CVPR 2008 IEEE Conference, pp 1–8
46. Yeffet L, Wolf L (2009) Local trinary patterns for human action recognition. 2009 IEEE 12th International Conference on Computer Vision, pp 492–497
47. Mousavi H, Galoogahi HK, Perina A, Murino V (2016) Detecting abnormal behavioral patterns in crowd scenarios. In: Toward robotic socially believable behaving systems-volume II., Springer, pp 185–205
48. Xu D, Ricci E, Yan Y, Song J, Sebe N (2015) Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553
49. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision Pattern Recognition (CVPR'05), pp 886–893