Multimedia Systems (2016) 22:275–285
DOI 10.1007/s00530-015-0456-7

REGULAR PAPER

A fast recognition algorithm for suspicious behavior in high definition videos

Chundi Mu · Jianbin Xie · Wei Yan · Tong Liu · Peiqin Li

Received: 11 March 2014 / Accepted: 14 January 2015 / Published online: 14 February 2015

Abstract Detecting suspicious behavior from high definition (HD) videos is always a complex and time-consuming process. To solve that problem, a fast suspicious behavior recognition method is proposed based on motion vectors. In this paper, the data format and decoding features of HD videos are analyzed. Then, the characteristics of suspicious activities and the ways of obtaining motion vectors directly from the video stream are concluded. Besides, the motion vectors are normalized by taking the reference frames into account. The feature vectors that display the inter-frame and intra-frame information of the region of interest are extracted. Gaussian radial basis function is employed as the kernel function of the support vector machines (SVM). It also realizes the detection and classification of suspicious behavior in HD videos. Finally, an extensive set of experiments are performed and this method is compared with some of the most recent approaches in the field using publicly available datasets as well as a new annotated human action dataset including actions performed in complex scenarios.

1 Introduction

Suspicious behavior is an anomalous behavior which is likely to threaten human life, health, and freedom. Suspicious behavior is not a fixed action (e.g., hands raised up, standing up, sitting down, etc.), nor a simple behavior (e.g., running, jumping, cycling, etc.). It is a complex behavior including a series of actions and simple behaviors. It is difficult to be recognized accurately because such behavior does not have fixed pattern or type.

Artificial intelligence-based analysis of suspicious behavior refers to the automatic detection of suspicious behavior through a computer or other hardware devices. In a video surveillance system, the earlier those suspicious behaviors are found, the less harmful they will be. Therefore, how to find a method that can automatically analyze suspicious behavior and can fully or partially replace monitors has become a hot topic in the field of computer vision.

All suspicious behavior recognition methods can be classified into two categories: single-layered approaches and hierarchical approaches [1]. Single-layered approaches can be classified into two types depending on how they model human activities: space–time approaches [2–5] and sequential approaches [6–8]. Space–time approaches view an input video as a 3-D (XYT) volume and recognize it by extracting trajectories or local interest points, whereas sequential approaches interpret an input video as a sequence of observations and recognize it by exemplar-based methodologies or model-based methodologies. Hierarchical approaches are classified on the basis of the recognition methodologies they use: statistical approaches [9], syntactic approaches [10], and description-based approaches [11, 12]. Single-layered approaches are suitable for gesture recognition. Hierarchical approaches is trying to describe high-level human activities in terms of simpler activities, that is to say, such approaches are suitable for the recognition of complex activities.

To accurately extract the features of moving targets, conventional video analysis algorithms often involves a

Communicated by T. Mei.

C. Mu · J. Xie · W. Yan (✉) · T. Liu · P. Li
College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China
e-mail: 15660565@qq.com

J. Xie
e-mail: xiejianbin@nudt.edu.cn

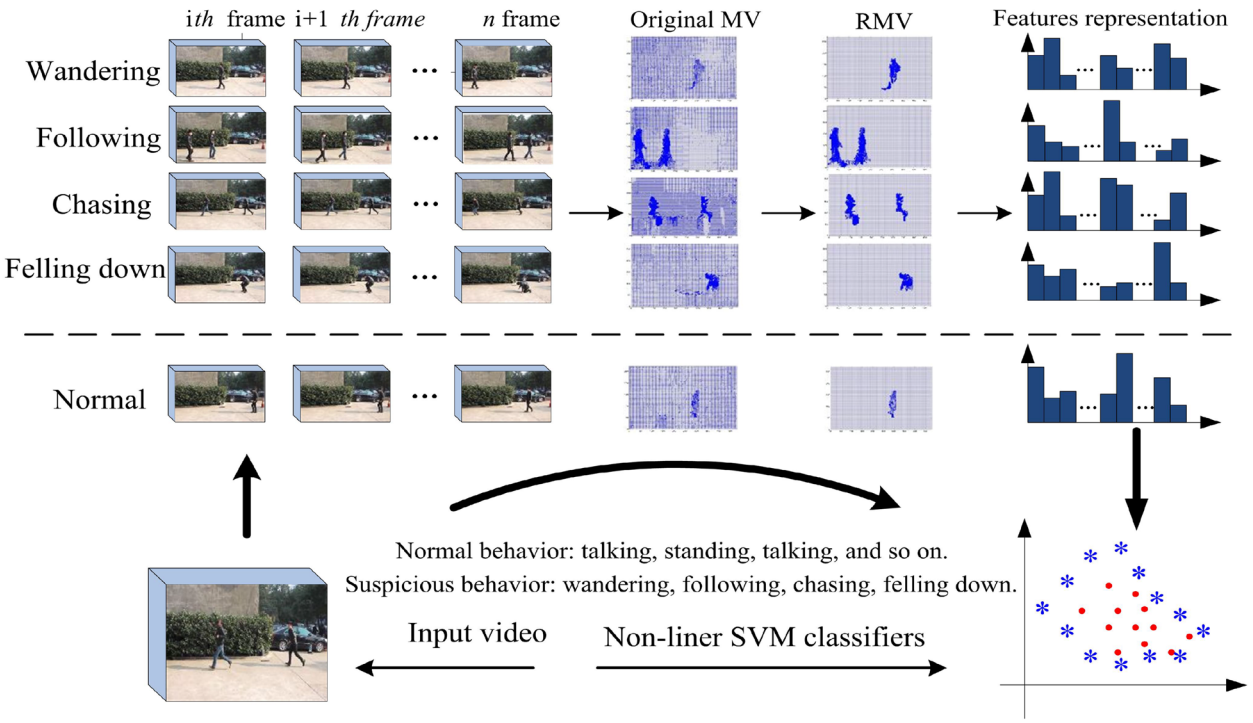


Fig. 1 Framework of our algorithm

complex process for preprocessing and segmenting the targets. This process requires enormous amount of computation. Thus, these algorithms cannot be applied in real-time video surveillance systems effectively.

To overcome the abovementioned problems, we propose a fast suspicious behavior recognition method for HD videos. In this method, the motion vectors in the video streaming are taken into account. The motion vectors are extracted from the video streaming directly. Then, the moving target region is extracted by optimal threshold method. The direction and velocity parameters are then extracted from the motion vectors' modulus. A support vector machine (SVM) is used to learn and classify the input videos. The framework of our method is shown in Fig. 1.

The contributions of our method are as follows:

(i) The motion vectors are extracted from HD videos directly, which takes full advantages of the compressed data stream and solve the time-related problems of traditional algorithms.
(ii) A moving object segmentation algorithm is proposed based on the modulus of motion vectors. This method uses velocity to distinguish interesting targets from interference targets (e.g., ambient noise). In this way, it solves noise problems faced by traditional algorithms.
(iii) Seven features are proposed to describe each moving target, which makes it easier to judge the state of the target.
(iv) This algorithm is fully automated and does not rely on manual selection. It also meets the requirements of real-time applications.

2 Related work

Suspicious behavior recognition is an important aspect of human activity recognition due to its potential for a multitude of applications. An increasing number of researchers are conducting studies on this issue. Suspicious behavior generally refers to abnormal activities between human and things or among human. Therefore, human behavioral characteristics should be studied before we analysis suspicious behavior. Previous works on human activity recognition focus on classifying video sequences of a single person in controlled environments. In this scenario, the background is simple and uniform [15]. However, with the development of activity recognition technologies, researchers are constantly attempting to introduce a greater number of natural and unconstrained videos. For instance, Laptev et al. [16] studied sequences from feature films, whereas other researchers focused on the recognition of "wild videos," such as videos from YouTube [17].

Porikli et al. [18] presented a compressed domain video object segmentation method based on the DCT coefficients and motion vectors in the video streams, however, it does not

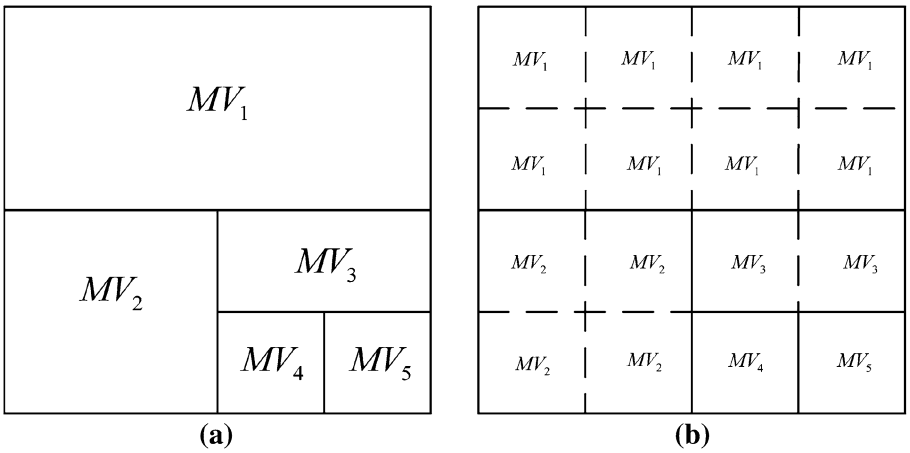


Fig. 2 **a** Original macroblock; **b** Normalized macroblock

suit for the high moving targets. Messing et al. [19] presented an activity recognition feature inspired by human psychophysical performance which is based on the velocity history of tracked key points. Schuldt et al. [13] constructed video representations in terms of local space–time features and integrated such representations with SVM classification schemes for recognition. A novel descriptor based on motion boundary histograms was introduced by Wang et al. [20] which have shown good performance by classifying many kinds of actions. Moreover, Sadanand et al. [21] presented Action Bank, a new high-level representation of the video, which is comprised of many individual action detectors sampled broadly in the semantic space as well as the viewpoint space.

For the recognition of suspicious activities, Lavee et al. [22] took advantage of the fact that the same event captured by different camera configurations looks identical when projected on to the temporal dimension. With this characteristic, they clustered a video into event-specific segments, devised an event comparison method, and defined "suspicious" activity by examples. Mecocci et al. [23] also introduced architecture of automated real-time video surveillance capable of detecting anomalous events. Suspicious behaviors were detected whenever the observed trajectories deviated from the typical learned prototypes. Similar to Mecocci's method, Daniel Barbará put forward a method which was capable of flagging anomalous images, where anomalies were defined as situations that were not encountered in a baseline of images used to train the technique [24]. Furthermore, Wiliem et al. [25] proposed a new clustering method to obtain the trajectory of human action, and then used neural networks to analyze the characteristics of the track. Syed Ahmar Qamar used temporal methodology for the history of motion in sequence of images, and then by analyzing the motion vectors got by the history images to recognize suspicious behavior [26]. Kaluža et al. [27] advanced a two-step detection system, where it first detects trigger events from multiagent interactions, and then combined the evidence to provide a degree of suspicion.

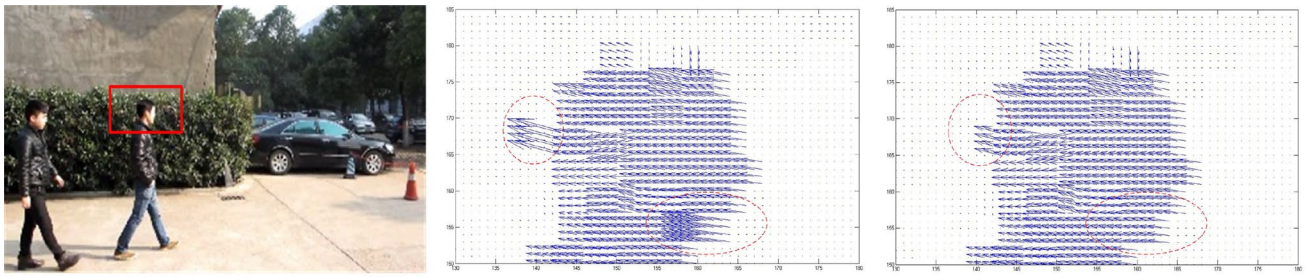
In summary, we find that the method for detecting suspicious behavior is always developed for specific types of behaviors, such as suspicious behavior in supermarkets, suspicious behavior at airport, and so on. To improve the accuracy of the experiment, many methods require a complex feature extraction algorithm, which indicates that they cannot be applied in a real-time system. To solve all these problems, we present a fast method for detecting wandering, trailing, chasing, and falling down. Through learning and analyzing the vectors from the data stream, the proposed method can identify the abovementioned suspicious behavior effectively in a short time.

3 Extracting motion vector

Nowadays, motion vectors can be extracted from videos in two ways: extracting from gray images and extracting from video streaming. The advantage of extracting from gray images is that each macroblock has its own motion vector. In this case, we could obtain a comprehensive characterization of the target. However, its shortcomings lie in that the algorithms are too complex to be realized and they are unable to meet the requirements of real-time applications. Extracting motion vectors from video streaming can directly obtain the information. But the characteristics of multiple reference frames (i.e., different macroblocks refer to different reference frames) and the multi-macroblocks mode (I, P and B macroblocks can coexist in one frame) make them difficult to be analyzed.

In this paper, combining the advantages of these two methods, an improved method for extracting motion vectors from video streaming is recommended. The proposed method can improve the efficiency of the system.

As we all know, for the state-of-the-art coding methods such as H.264 and HEVC, each macroblock may include MVs with different resolutions, so we must normalize the size of macroblock before future calculating. In this paper,



(a) **(b)** **(c)**

Fig. 3 **a** Frames of HD video; **b** Local motion vectors extracted directly from the bit stream (**a**); **c** Local motion vectors extracted by our method

we normalize all the macroblock as 4 × 4. The sample is presented in Fig. 2.

As shown in Fig. 2, each macroblock is divided into 4 × 4, and the motion vectors for each block are the same as before.

To convert all the reference frames of motion vectors to the previous frames, we processed I macroblock, P macroblock, P_Skip macroblock, and B macroblock as follows.

(i) For all I macroblocks, whether they are from I frame, B frame, or P frame, we set the motion vector to (0, 0).

(ii) For P macroblocks and P_Skip macroblocks in P frame or B frame, the motion vectors are based on the position of their reference frames. Assuming that the motion vector of one P macroblock is (MV_{xj}, MV_{yj}), j is the number of frames between the current frame and the reference frame. Considering the motion continuity of the target, we processed the motion vectors as follows:

$$(MV_{x1}, MV_{y1}) = \left(\frac{MV_{xj}}{j}, \frac{MV_{yj}}{j}\right) \tag{1}$$

(iii) For B macroblock in B frame, we assume that the motion vectors of the current macroblock are (MV_{fxk}, MV_{fyk}) and (MV_{bxl}, MV_{byl}), where f and b are abbreviations of forward and backward, and k and l are the distance from the reference frames. Then:

$$(MV_{fx1}, MV_{fy1}) = \left(\frac{MV_{fxk}}{k}, \frac{MV_{fxk}}{k}\right) \tag{2}$$

$$(MV_{bx1}, MV_{by1}) = \left(\frac{MV_{bxl}}{l}, \frac{MV_{byl}}{l}\right) \tag{3}$$

$$(MV_{x1}, MV_{y1}) = (MV_{fx1}, MV_{fy1}) - (MV_{bx1}, MV_{by1}) \tag{4}$$

We experimented on a HD surveillance video, and the results are presented in Fig. 3.

As shown in Fig. 3, at the upper left and bottom right side of Fig. 3b there are some motion vectors which are extremely large. After analyzing the bit stream, we found that the reference frame of those singular motion vectors is not the previous one. That is to say $j > 1$. So the refined motions can be obtained as Fig. 3c. Through observing Fig. 3c, we can easily find that this method avoids the emergence of the singular motion vectors and makes the distribution of the motion vectors more reasonable, which can be helpful to achieve a better segmentation results later.

4 Proposed framework for suspicious activity recognition

4.1 Extracting and segmenting the moving object

In HD videos, the suspicious target may be a single person or multiple people. When a suspicious behavior occurs, such as chasing, each simple action of the behavior would be difficult to be detected separately. To solve this problem, we considered them as one target and labeled it motion region (MR). Thus, we were able to effectively solve the problem of moving target occlusion. Strong robustness can be achieved on the detection of suspicious behavior, such as chasing and following. Meanwhile, to save computing time and improve recognition efficiency, we propose the concept of regional motion vectors. All of the vector features are extracted from regional motion vectors.

Major moving object segmentation algorithms are as follows: threshold, region growing, clustering, relaxation and others. Among which, threshold method is widely used in moving target detection because of its small computing requirement, high speed, and easy implementation. This paper used an optimized adaptive threshold segmentation algorithm, which has strong anti-noise performance, high speed, and high efficiency. This algorithm effectively reduces the false alarm rate. The process is as follows:

***Step* 1** The modulus of motion vectors is extracted.

$$MV(m,n) = \sqrt{MV_{x1}(m,n)^2 + MV_{y1}(m,n)^2} \tag{5}$$

m and n denote the horizontal and vertical coordinates of the current macroblock.

Assuming that $\{MV_1, MV_2, \ldots, MV_s\}$ is the unique values of motion vectors modulus, an adaptive threshold TH is gained by iterating all the values.

***Step* 2** For MV_i $(0 < i < s + 1)$, the probabilities of the two types of motion vectors are obtained as follows:

$$p = \frac{U}{N}, \quad q = \frac{V}{N} \tag{6}$$

where U is the number of motion vectors whose modulus are smaller than MV_i, and V is the number of motion vectors whose modulus are larger than MV_i, W is the total number of motion vectors, that is to say, $W = U + V$.

***Step* 3** The mean of two types of motion vectors is calculated.

$$\mu_0 = \frac{sum_1}{U}, \quad \mu_1 = \frac{sum_2}{V} \tag{7}$$

where sum_1 is the sum of the motion vector values which are smaller than MV_i, and sum_2 is the sum of the motion vectors' values which are larger than MV_i.

***Step* 4** The between-class variances are calculated.

$$\sigma^2 = p \times q \times (\mu_0 - \mu_1)^2 \tag{8}$$

***Step* 5** All the unique modulus of the motion vectors are iterated to find the optimal threshold which could make σ^2 maximum.

$$TH = \arg \max_{MV_i}(\sigma^2) \tag{9}$$

***Step* 6** The moduli of motion vectors are segmented by TH.

$$MV(m,n) = \begin{cases} 1 & MV(m,n) > TH \\ 0 & MV(m,n) < TH \end{cases} \tag{10}$$

***Step* 7** Closing operation is performed on the segmentation results. And then we calculate a scale coefficient which is affected by the initial conditions.

$$TH_s = K \times \frac{h_{\text{mean}}}{H_c \times L} \tag{11}$$

h_{mean} is the general height of target. H_c is the height of camera. L is related to the field of view of the camera. The above three parameters are all measured in meters. K is a scale coefficient.

We take the regions which size is bigger than $K_s \times N$ as the interesting region, where N is the total marcoblocks in current frame.

Figure 4 presents the result of background subtraction method (Barnich [31]), frame subtraction method (Jianbin [32]) and this algorithm. The programs of Olivier Barnich and Zheng Jiangbin come from the website provided in their paper. As shown in Fig. 4, this algorithm can segment the background and foreground very well, with excellent robustness about the camera moving and background noise. Compared with background subtraction method and frame subtraction method, the merits of this method can be concluded as follows.

(i) There is no need to construct a background model for our algorithm. That is the most intractable issue of background subtraction method. Such as in Fig. 4e–f, we cannot reconstruct the background effectively due to the shaking of the camera. The results of background subtraction are unreliable.
(ii) Targets extracted by this method were of high-quality and unbroken, which is the most intractable issue of frame subtraction method. Such as in Fig. 4i–k, because of the background noise and camera moving, the threshold of frame subtraction is not always effective, so the targets extracted by that method are not.
(iii) The minimum unit of this process is a block with 4×4 pixels. Therefore, when we consider the same segmentation, this algorithm is easier to be implemented and requires less computation.

4.2 Features of motion vectors

If there exists a single target suspicious behavior (e.g., wandering and falling down), the direction and trajectory of motion vectors show certain regularity. When there existing a wandering suspicious behavior in the video, the motion vectors of the target are periodically changing. For a man falling down, the speed and direction of the target would change. For suspicious behavior of multiple targets, such as following and chasing, direction and velocity have different characteristics. For following, the direction and velocity of targets are always the same. For chasing, the direction of the targets is the same, and the targets always have a great velocity. Therefore, we focused on analyzing the direction and velocity of the target. To analyze the features of the target more precisely, a new concept called motion vectors in region (MVR) is defined. Each MVR corresponds to a moving region (e.g., Fig. 5f).

4.2.1 Direction feature of motion vector

Abnormal movement of single target and relative motion between multiple targets are always accompanying with suspicious activities, the regularity of which has been reflected on the motion vectors. For each motion vector, we get the direction of it is gained by formula (12).

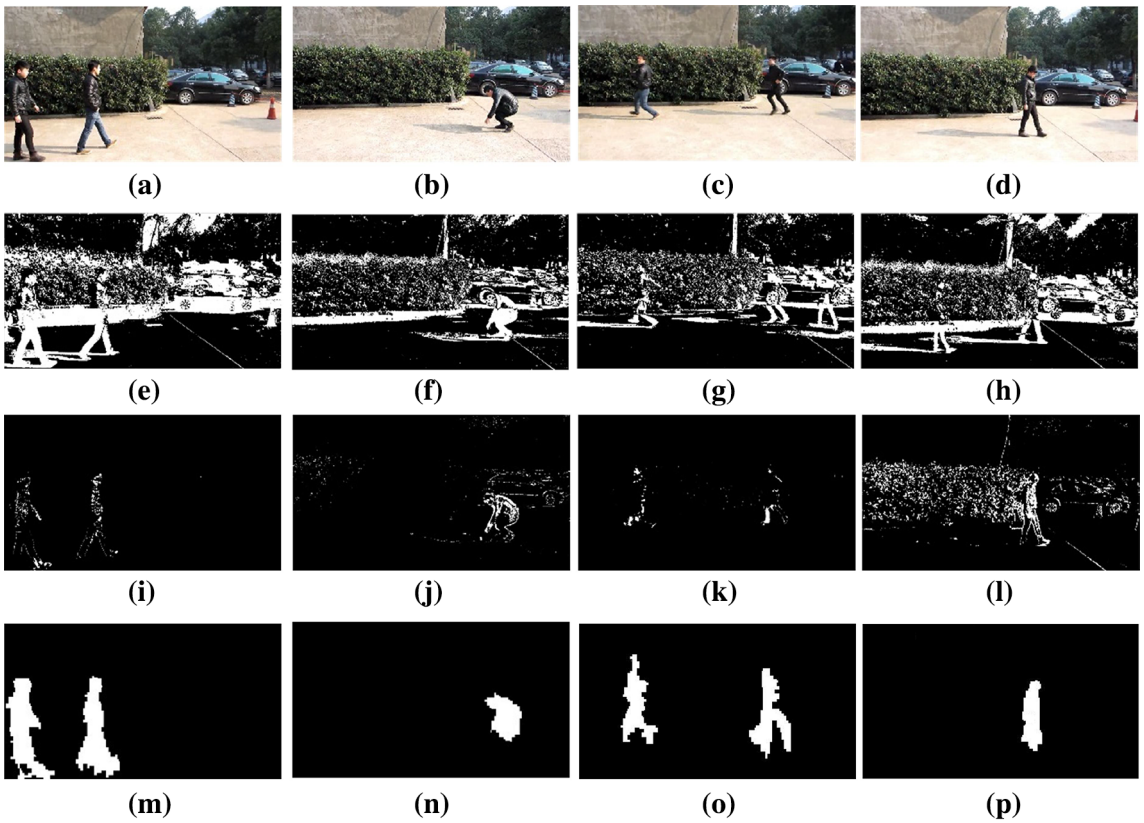


Fig. 4 **a–d** Original images; **e–h** Targets extracted by background subtraction method; **i–l** Targets extracted by frame subtraction method; **m–p** Targets extracted by our method

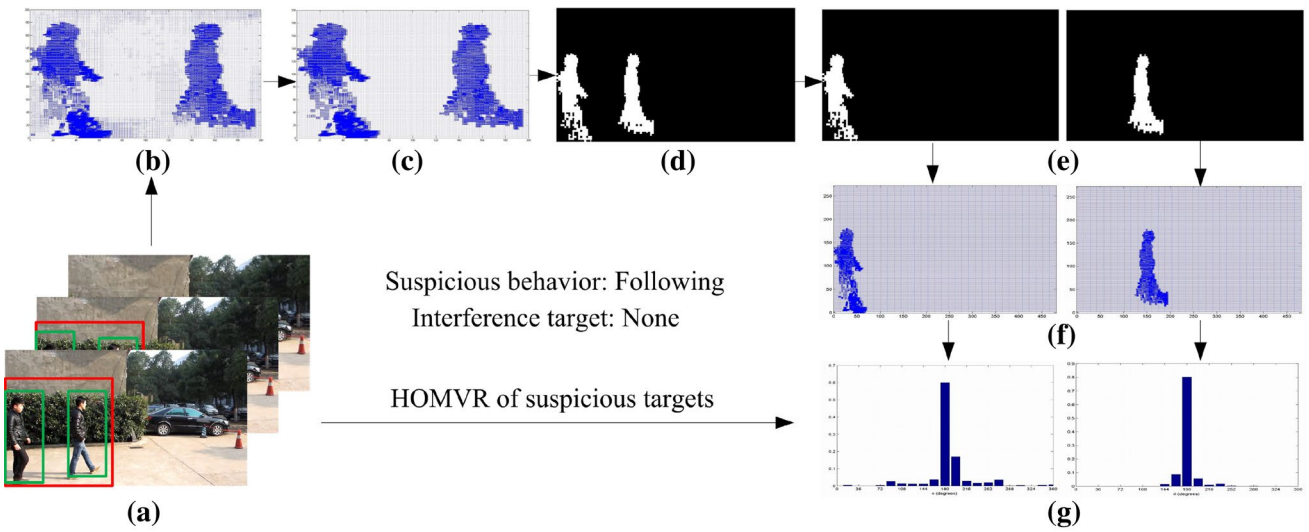


Fig. 5 **a** Original video image; **b** Local motion vectors extracted from video streaming; **c** Local motion vectors optimized by our algorithm; **d** Target extracted from motion vectors; **e** Separated targets; **f** Motion vectors for separated targets; **g** Histogram of motion vectors

$$\theta = \begin{cases} \arctan \dfrac{MV_{y1}}{MV_{x1}} & (MV_{x1} > 0, MV_{y1} > 0) \\ \arctan \dfrac{MV_{y1}}{MV_{x1}} + \pi & (MV_{x1} < 0) \\ \arctan \dfrac{MV_{y1}}{MV_{x1}} + 2 * \pi & (MV_{x1} > 0, MV_{y1} < 0) \end{cases} \tag{12}$$

4.2.2 Velocity feature of motion vector

Motion vector (MV_{x1}, MV_{y1}) represents the motion of macroblock and its modulus represents relative motion distance. The frame rate of a video is constant; therefore, we can obtain the velocity of macroblock as follows:

$$V \propto |MV| = \sqrt{MV_{x1}^2 + MV_{y1}^2} \tag{13}$$

4.2.3 Velocity and direction features of MVR

Based on the concept of histogram of gradient (HOG), we propose the histogram of MVR (HOMVR) and obtain the direction of motion vectors as Fig. 5.

The other features of MVR can be obtained except for HOMVR. If the number of motion vectors of j-th MVR is N_j, then we can obtain the following:

$$\theta_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \theta_{ji} \tag{14}$$

$$V_j = \frac{1}{N_j \cdot R} \sum_{i=1}^{N_j} V_{ji} = \frac{1}{N_j \cdot R} \sum_{i=1}^{N_j} \left| MV_{ji} \right| \tag{15}$$

$$\sigma_{\theta_j} = \frac{1}{N_j} \sqrt{\sum_{i=1}^{N_j} (\theta_{ji} - \theta_j)^2} \tag{16}$$

$$\sigma_{V_j} = \frac{1}{N_j} \sqrt{\sum_{i=1}^{N_j} (V_{ji} - V_j)^2} = \frac{1}{N_j} \sqrt{\sum_{i=1}^{N_j} (\frac{MV_{ji}}{K} - V_j)^2} \tag{17}$$

$$E_{\theta_j} = \sum_{i=0}^{N_{\theta_j}-1} p_{\theta_j i} \log p_{\theta_j i} \tag{18}$$

$$E_{Vj} = \sum_{i=0}^{N_{v_j}-1} p_{V_j i} \log p_{v_j i} \tag{19}$$

θ_{ji}, V_{ji} represent the i-th motion vector's direction and velocity. θ_j, V_j represent the average direction and velocity of the target, R is the frame rate. σ_{θ_j}, σ_{V_j} represent the direction variance and the velocity variance. N_{θ_j}, N_{V_j} are the number of θ_{ji},V_{ji} classifications. $P_{\theta_j i}$, $P_{V_j i}$ are the probability of each classification. E_{θ_j}, E_{V_j} represent entropy of direction and velocity.

4.2.4 Interesting degree of inter-frame $Inter_D$

All the features that we have discussed so far are intra-frame features. Inter-frame features are also important for us to detect suspicious behavior. Therefore, we extract them by χ^2 histogram method [33].

The definition of the interesting degree of inter-frame is as follows:

$$Inter_{D_j} = \begin{cases} \sum_{i=1}^{n} \frac{(H_{k,i,j} - H_{k+1,i,j})^2}{\max(H_{k,i,j}, H_{k+1,i,j})} & (H_{k,i,j} \neq 0 \quad or\ H_{k+1,i,j} \neq 0) \\ 0 & (H_{k,i,j} = 0\ and\ H_{k+1,i,j} = 0) \end{cases} \tag{20}$$

where n is the total number of histogram classification, $H_{k,j,i}$ means the i-th part of the j-th MVR's histogram of the k-th frame. If the histogram of j-th MVR of k-th frame is similar to the j-th MVR of $(k+1)$-th frame, $Inter_{D_j}$ is small, or else, it is large. That is to say that $Intra_{D_j}$ can display the moving change of the target.

Through the abovementioned method, 7-D features $\{\theta, V, \sigma_\theta, \sigma_V, E_\theta, E_V, Inter_{D_j}\}$ can be employed to describe the targets we have extracted from the frame (except the first frame). With these features, the direction and velocity by θ and V and the level of dispersion of θ and V by σ_{θ_j} and σ_{V_j} can be acquired, as well as the randomness of θ and V by E_{θ_j} and E_{V_j}. For these suspicious behaviors we have detected in our paper, we found that the direction, velocity and inter-frame difference are the most important features of them. For example, as in following and chasing, the targets which have the same motion direction and the inter-frame difference are always small. But they differ in the velocity, which we can get by parameter V. Thus, through the extracted parameters, we can get the moving states of the targets, effectively.

4.3 Support vector machines

For suspicious behavior, high-dimension features can hardly be classified by methods based on minimum distance or template matching. To solve this problem, SVM is used as the classifier. A brief review of the theory behind this type of algorithm is offered in this section.

SVM is a machine learning method based on statistical learning. The complexity of SVM does not only depend on the dimension of sample features. It is only influenced by the number of support vectors. In general, only a small number of support vectors are needed to determine the final result. SVM has a good robustness.

For non-liner SVM classifiers, the decision function is as follows:

Fig. 6 A montage of entries is in the action bank, 10 of the 300 are in the bank. Faces are redacted for presentation only

Table 1 Effect of various algorithms (experiment I)

	Motion vectors		Targets segmenting		Kernel function		Result	
	Original MV	Optimized MV	Frame Subtraction	MV Segmentation	Sigmoid	Gaussian radial basis	LAR (Leak alarm rate)	FAR (False alarm rate)
1	✓		✓		✓		20.8%	12.5%
2	✓		✓			✓	19.2%	15.0%
3	✓			✓	✓		17.5%	12.5%
4	✓			✓		✓	16.7%	10.0%
5		✓	✓		✓		17.5%	12.5%
6		✓	✓			✓	18.3%	12.5%
7		✓		✓	✓		10.0%	7.5%
8		✓		✓		✓	8.3%	7.5%

$$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{n} y_i \alpha_i^* K(x_i \times x) + b^*\right\} \tag{21}$$

where $x \in R^N$ is a feature vector, n is the number of support vectors, x_i is the support vectors, $y_i \in \{-1, 1\}$ is class label (-1 means negative sample and 1 means positive sample). α_i^* and b^* are found using a SVC learning algorithm. $K(\cdot)$ is kernel function, which can be used to change the computation in high dimension to low dimension. In this paper, Gauss radial basis function is used, which is the most popular kernel function and is very suitable for non-linear classification application problems.

$$K(x_i \cdot x) = \exp\left\{-\frac{|x_i - x|^2}{2\sigma^2}\right\} \tag{22}$$

The type of *subevents* [1] is obtained through classifying the features bank extracted from the video. We first study a threshold of $Intra_D$ by the suspicious videos which we have labeled before, and then take the frame which $Intra_D$ is bigger than the threshold as the trigger frame. For each trigger frame, the features of current frame and the next four frames are combined as the features bank. And if there is no trigger frames in the video, the features banks are taken by any sequential five frames.

At last, the type of the activity in the video can be acquired by classifying the *subevents* bank.

5 Experiments

Existing databases, such as KTH actions database [13], UCF sport actions database [4] and Hollywood2 actions database [14], are widely used for the recognition of individual behavior, sports activity, and activity in movies. Suspicious behavior databases ESCAPES and CACIAR consist of abnormal behavior at airports and in supermarkets. Observing these databases, it is obvious that all of them do not include those kinds of suspicious behavior we are studying. Thus, we build a suspicious behavior database for this paper in particular. This database consists of five human action classes: wandering, trailing, chasing, falling down, and normal activity. A total of 300 HD video samples in a resolution of 1,920 × 1,080, 25 fps are included in this database. All of these videos have 200 frames. Samples of this dataset are shown in Fig. 6.

Table 2 Suspicious behavior recognition performance (experiment II)

Result	Wandering	Following	Chasing	Felling down	Normal
Wandering	**28**	1	0	0	1
Following	0	**24**	4	0	2
Chasing	0	5	**25**	0	0
Felling down	0	1	1	**23**	5
Normal	3	9	2	0	**66**

Bold values indicate the right classification times

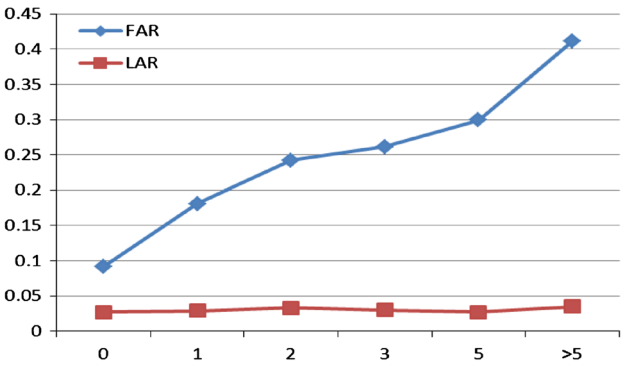


Fig. 7 Interference target influence on LAR and FAR

5.1 Detecting suspicious behaviors

For detecting suspicious behavior, controlling variable method is employed in performing the first experiment to prove the effectiveness of our algorithm. In this trial, 140 videos were used as training samples (including 80 suspicious behavior videos and 60 normal behavior videos), and 160 videos were used as testing samples. Results are shown in Table 1.

From experiment I, the proposed algorithm has the best result.

Comparing experiment 1 with experiment 5, experiment 2 with experiment 6, experiment 3 with experiment 7, experiment 4 with experiment 8, we found that the results of these methods taking the motion vectors extracted in our method are better than these methods taking the original motion vectors. These pairs of experiments have the same targets segmenting method and Kernel function except the way of extracting motion vectors. The reasons of the above-finding have been explained in chapter 3.

Comparing experiment 1 with experiment 3, experiment 2 with experiment 4, experiment 5 with experiment 7, experiment 6 with experiment 8, we found that if we segment the targets in this method better detection results will be gained. These pairs of experiments have the same method in the extraction of motion vectors and Kernel function except the way of segmenting targets. In experiment I, background subtraction method is not taken into account. Because this method was not suitable for the moving of the camera and complex background, and the results of different targets segmentation are shown in Fig. 4.

The difference between Sigmoid kernel and Gaussian radial basis kernel is unclear. However, considering the complexity of these two functions, Gaussian radial basis function is chosen as the kernel function.

In experiment I, we generate one-to-one classifiers, that is, they can only classify input videos as suspicious or not. However, in the intelligent monitoring system, the type of suspicious behavior must be identified. Therefore, a second experiment is conducted, which uses voting method to determine the category of the suspicious behavior.

In experiment II, 100 samples for training SVM classifiers (every category has 20 samples) are selected. The one-to-one classifiers were generated by choosing two kinds of behaviors arbitrarily. In this way, we can obtain ten classifiers (C_5^2) when five kinds of behaviors need to be classified. When a new video is input, it would be classified by these ten classifiers separately, and the one with the most votes is selected as the result. If one input video has more than one category with the same number of votes, a false one is taken as the result. The remaining 200 video clips were as experiment data to be dealt with, and the results are shown in Table 2.

After analyzing the misclassified video clips, the main reasons for classification errors are as follows:

- For following and chasing, all the features have the same expected velocity.
- Interference targets have a great effect on this system. If there are more than two targets in the video, it is hard to distinguish the real interesting targets from the interference target even though a series of rules are made to avoid such interference.

In this experiment, if multi-targets are detected in the video, the one with the largest $Intra_{Dj}$ is chosen as the interesting target, and whether the target is wandering or felling down is detected. If the target is wandering or felling down, we decide the result directly. Or the one with the second largest $Intra_{Dj}$ will be chosen as another interesting target, and whether the targets are following or chasing will be detected. To study the influence of interference target, another experiment on all of the videos that involve interference target is conducted. Results are shown in Fig. 7.

Table 3 Recognition accuracies on the KTH and Weizmann datasets (experiment III)

Method	Year	KTH (%)	Weizmann (%)
Schindler et al. [28]	2008	92.7	100
Gilbert et al. [29]	2009	94.5	–
Angela et al. [30]	2010	93.0	89.0
Sadanand et al. [21]	2012	98.0	–
Our method		97.5	100

Table 4 Recognition accuracies on our dataset

Method	Accuracy rate (%)
Schindler et al. [28]	77.2
Gilbert et al. [29]	76.3
Angela et al. [30]	74.2
Sadanand et al. [21]	80.0
Our method	91.7

Table 5 A break down of the average frame per second

Method	Frames per second
Schindler et al. [28]	<5
Gilbert et al. [29]	24
Angela et al. [30]	10
Sadanand et al. [21]	<5
Our method	45

Analyzing Fig. 7, we can see that with an increase of interference target, the leakage alarm rate of our system remains stable, but the false alarm rate deteriorates rapidly.

5.2 Benchmark action recognition datasets

Recent studies on suspicious behavior always focused on a specific situation because such behavior cannot be defined accurately. However, taking four kinds of suspicious behaviors into account, we found that all of the behaviors consist of basic actions, such as walking, running, bending, and so on. To test the effectiveness of the proposed feature, two benchmark action recognition datasets are used to test the recognition rate of fundamental actions (i.e., walking, running and bending) of this method. Experimental results are compared in Table 3.

From Table 3, this method has a preferable recognition results on benchmark action recognition datasets. Even though that result cannot directly reflect the ability of recognizing suspicious activities of this method, it can prove that the features of this method can describe basic actions very well, which is the basic of recognition activity.

5.3 Our action recognition dataset

In Sect. 5.2, the outstanding performance of this method is proven on the benchmark action recognition datasets. To further prove the effectiveness of this algorithm, we will take the four methods which we compared in Table 3 on our dataset. The results are shown in Table 4.

5.4 Computational cost

Most of the calculated amount of this method is spent on normalizing the motion vectors and extracting motion features from motion vectors. At the first step, the sum of motion vectors of a video frame with $M \times N$ pixels is no more than $M \times N/16$, and no more than $M \times N/16$ times addition and divisions are used. Then, we segment the targets from the background by a threshold TH, which is also needed to be done in the other algorithms. And following that step, 7-D features need to be calculated, in which calculated amount is $o\ (N_j)$, and N_j is the number of motion vectors of j-th MVR. While segmenting the target from the background, the calculated amount is less than traditional space–time approaches and sequential approaches, and the features we need to extract is also less than others. For instance, Sadanand et al. [21] need to take $N_a \times N_s \times 73$ features for each video, where N_a is the number of detectors and N_s is the scales (spatiotemporal) of each detector, but under the same conditions, we just need $N_s \times 7$ features. The other steps of this method such as getting motion vectors form video streams and using the trained SVM to classify feature banks need little calculation. To illustrate the computation directly, Table 5 shows the average frame rate of different approach on KTH benchmark action recognition dataset.

Table 5 illustrates the features extracting speed of different approach. According to this table, this method can handle 45 frames per second. It indicates this algorithm has the potential to be used in real-time video surveillance system, as in real-time video surveillance system the frame rate is always no more than 25.

6 Conclusions and future work

This paper proposes a fast suspicious behavior recognition method for HD videos based on motion vectors. The motion vectors are extracted from the video streaming. Then, the moving target region is extracted by optimal threshold method. The direction and velocity parameters are then extracted from the motion vectors' modulus. Finally, the support vector machine (SVM) is used to learn and classify the input videos. The results show that this approach can achieve high recognition rate and low false alarm rate.

There are also some limitations in this algorithm. First, this algorithm is only suit for the high-resolution videos, because for the low-resolution videos we cannot get sufficient information to segment the targets and to recognize the suspicious activity. Second, this approach is only suitable for static cameras, that is to say that only slight shaking can be handled. Finally, the effect of our algorithm is seriously affected by the encoding method, that is to say that it is not suit for all the situations.

Our further research would focus on how to handle these drawbacks we put forward above.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. In: ACM (2011)
2. Rao C., Shah, M.: View-invariance in action recognition. In: CVPR (2001)
3. Savarese, S., Delpozo, A., Niebles, J., Fei-Fei, L.: Spatial-temporal correlations for unsupervised action classification. In: WMVC (2008)
4. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR (2008)
5. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV (2009)
6. Jiang, H., Drew, M., Li, Z.: Successive convex matching for action detection. In: CVPR (2006)
7. Veeraraghavan, A., Chellappa, R., Roy-Chowdhury, A.: The function space of an activity. In: CVPR (2006)
8. Natarajan, P., Nevatia, R.: Coupled hidden semi-markov models for activity recognition. In: WMVC (2007)
9. Damen, D., Hogg, D.: Recognizing linked events: searching the space of feasible explanations. In: CVPR (2009)
10. Joo, S.W., Chellappa, R.: Attribute grammar-based event recognition and anomaly detection. In: CVPR (2006)
11. Ryoo, M.S., Aggarwal, J.K.: Semantic representation and recognition of continued and recursive human activities. In: IJCV (2009)
12. Ryoo, M.S., Aggarwal, J.K.: Recognition of composite human activities through context-free grammar based representation. In: CVPR (2006)
13. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR (2004)
14. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
15. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)
16. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
17. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos'in the wild'. In: CVPR (2009)
18. Porikli, F., Bashir, F., Sun, H.: Compressed domain video object segmentation, IEEE Trans. Circuits Syst. Video Technol. **20**(1), 2–14 (2010)
19. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: ICCV (2009)
20. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR (2011)
21. Sadanand, S., Corso, J.J.: Action Bank: a high-level representation of activity in video. In: CVPR (2012)
22. Lavee, G., Khan, L., Thuraisingham, B.: A framework for a video analysis tool for suspicious event detection. In: MDM (2005)
23. Mecocci, A., Pannozzo, M., Fumarola, A.: Automatic Detection of Anomalous Behavioural Events for Advanced Real-Time Video Surveilance, International Symposium on Computational Intelligence for Measurement Systems and Applications, pp. 187–192. Lugano (2003)
24. Barbará, D., Filippone, M.: Detecting suspicious behavior in surveillance images. In: ICDMW (2008)
25. Wiliem, A., Madasu, V., Boles, W., Yarlagadda, P.: A context-based approach for detecting suspicious behaviours. In: DICTA (2009)
26. Qamar, S.A., Jaffar, M.A., Habib, H.A.: A supervisory system to detect suspicious behavior in online testing system. In: ICMLC (2009)
27. Kaluža, B., Kaminka, G.A., Tambe, M.: Detection of suspicious behavior from a sparse set of multiagent interactions. In: AAMAS (2012)
28. Schindler, K., Gool, L.J.V.: Action snippets: how many frames does human action recognition require?. In: CVPR (2008)
29. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. In: ICCV (2009)
30. Yao, A., Gall, J., Van Gool, L.: A hough transform-based voting framework for action recognition. In: CVPR (2010)
31. Barnich, Olivier, Van Droogenbroeck, Marc: ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans. Image Process **20**(6), 1709–1724 (2011)
32. Jiangbin, Zheng, Xiuxiu, Li, Yanning, Zhang: Novel tacking algorithm for video surveillance. Syst. Eng. Electron. **29**(11), 191–193 (2007)
33. Nagasaka, A., Tanaka, Y.: Automatic video indexing and full video search for object appearances. Visual Database Systems II, pp. 113–127 (1992)