Multimed Tools Appl
DOI 10.1007/s11042-015-2648-8

A new method for violence detection in surveillance scenes

Tao Zhang[1] · **Zhijie Yang**[1] · **Wenjing Jia**[2] · **Baoqing Yang**[1] · **Jie Yang**[1] · **Xiangjian He**[2]

Received: 30 October 2014 / Revised: 4 March 2015 / Accepted: 20 April 2015

Abstract Violence detection is a hot topic for surveillance systems. However, it has not been studied as much as for action recognition. Existing vision-based methods mainly concentrate on violence detection and make little effort to determine the location of violence. In this paper, we propose a fast and robust framework for detecting and localizing violence in surveillance scenes. For this purpose, a Gaussian Model of Optical Flow (GMOF) is proposed to extract candidate violence regions, which are adaptively modeled as a deviation from the normal behavior of crowd observed in the scene. Violence detection is then performed on each video volume constructed by densely sampling the candidate violence regions. To distinguish violent events from nonviolent events, we also propose a novel descriptor, named as Orientation Histogram of Optical Flow (OHOF), which are fed into a linear SVM for classification. Experimental results on several benchmark datasets have demonstrated the superiority of our proposed method over the state-of-the-arts in terms of both detection accuracy and processing speed, even in crowded scenes.

Keywords Action recognition · Violence detection · Surveillance scenes · Gaussian model of optical flow (GMOF) · Orientation histogram of optical flow (OHOF)

1 Introduction

Violent behavior seriously endangers social and personal security. Currently, there are millions of video surveillance equipment used in public places, such as streets, prisons and

✉ Tao Zhang
zhb827@sjtu.edu.cn

✉ Jie Yang
jieyang@sjtu.edu.cn

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200000, China

[2] Faculty of Engineering and Information Technology, University of Technology, Sydney, PO Box 123, Sydney, Australia

Published online: 03 May 2015

supermarkets. It is highly necessary to investigate the recognition of harmful contents from surveillance video [21]. For this practical consideration, we focus on the recognition of violent behaviors and aim to find a method that can automatically detect violent behaviors using computer vision techniques. Violence detection aims to determine the presence of fighting within a video sequence, which involves similar techniques of many related computer vision applications, e.g., action recognition, object detection, surveillance, etc. [1, 3, 8, 11, 26, 36, 43, 47]. Complex background, illumination changes and different distances between the subjects and the camera have made this task very challenging, especially in the case of real-time applications.

During the last a few years, action recognition has been a hot topic. To some extent the problem has become tractable by using computer vision techniques. Based on the direct or indirect recognition methods of human activities, action recognition approaches can be classified into two categories: hierarchical and nonhierarchical [1, 36].

1) Hierarchical approaches described recognition methods for complex human activities such as human-object interactions and group activities, which can be classified into three categories, i.e., statistical, syntactic and description-based approaches [1].

   Statistical approaches use statistical models to recognize activities. Traditionally, some statistical models such as hidden Markov models (HMMs) and DBNs are used to recognize activities. A human action recognition method based on HMM was firstly proposed in [47], which is a feature-based, bottom-up approach characterized by its learning capability and time-scale invariability. Followed by this step, Oliver et al. [34] developed layered HMMs (LHMMs) to model a real-time activity with sequential structures. LHMMs used layered probabilistic representations to perform learning and inference at multiple levels of temporal granularity, and can be regarded as a cascade of HMMs. However, the influence of the layered decomposition on the size of the parameter space was not given, also the resulting effects on learning requirements and accuracy of inference for different amounts of training were ambiguous. Zhang et al. [51] used two-layer HMMs to recognize group actions, which were modeled as a two-layer process, with one layer modeling the basic individual activities from audio-visual features, and another modeling the interactions between individual activities. Nguyen et al. [32] presented an application of the hierarchical HMM for the problem of activity recognition. The main contributions of this method are in the application of the shared-structure HMM and the estimation of the model′s parameters at all levels simultaneously. Shi et al. [39] presented the Propagation Networks (P-Nets) for representing and recognizing sequential activities that included parallel streams of action. Their work focused on a common task of elderly people who have developed late stage diabetes and the performance strongly relied on manually labeled training data. Yu and Aggarwal [49] used a block-based discrete HMM to recognize multiple actions, where each block contained a subset of hidden states and was trained independently to improve the model estimation accuracy with a limited number of sequences. This method is not suitable to recognize multiple concurrent actions. Cupillard et al. [12] introduced a new approach for recognizing groups of people's behaviors using multiple cameras. This recognition process relied on a hierarchy of operators, each corresponding to a method to recognize behavior entities. When applied for the fighting scenario, the rate of false alarms is very high. Dai et al. [13] introduced a novel event-based dynamic context model, multilevel dynamic Bayesian network (DBN) model were used to detect multilevel events. But the applicable scenario was very limited. Damen and Hogg [14] proposed to construct Bayesian networks using AND-OR

grammars to encode pairwise event constraints, but it failed to recognize complex and ambiguous events. Gong and Xiang [17] developed a Dynamically Multi-Linked HMM (DML-HMM) to interpret group activities involving multiple objects captured in outdoor scene, where Dynamic Probabilistic Networks (DPNs) were exploited for modeling the temporal relationships among a set of different object temporal events in the scene. However, the uncertainty caused by occlusions and tracking errors were not given.

Syntactic approaches model human activities as a string of symbols, where each symbol corresponds to an atomic-level action [1]. Ivanov and Bobick [23] described a probabilistic syntactic approach to detect and recognize temporally extended activities and interactions between multiple agents. It worked very well for temporal behaviors and interactions between multiple objects. However, modeling temporally explicit behavior among more interactive objects is a difficult task. Moore and Essa [28] provided an endorsement for the use of hybrid model-exemplar approaches where flexible, SCFG (Stochastic Context-Free Grammar)-based models are applied for high-level recognition and feature-based exemplars for low-level detection. This work focused on multitask activities but failed to recognize complicated human activities. Also, it was not equally suited to many other modeling tasks, which involve non-sequential data. Minnen et al. [27] presented a system that used human-specified grammars to recognize a person performing the Towers of Hanoi task by analyzing object interaction events. However, there exist many high-level constraints, so it is not self-adaptive.

A description-based approach explicitly describe spatiotemporal structures for human activities, which represent a high-level human activity in terms of simpler activities composing the activity by describing their temporal, spatial, and logical relationships. Pinhanez and Bobick [35] developed a representation for the temporal structure inherent in human actions and demonstrated an effective method for using that representation to detect the occurrence of actions, which was computed by considering the minimal domain of its PNF-network. Intille and Bobick [22] designed a complex Bayesian network model to identify actions of a football player in a crowded scene. It can recognize highly structured or uncertain visual perception by representing them in a three-level hierarchy to deal with uncertain and incomplete nature of real world application. However, this method is based on the trajectory, while accurate trajectory is hard to get in real scene. Nevatia et al. [31] proposed Video Event Representation Language (VERL) to construct a heuristic algorithm for detecting ongoing human activities from input images. However, it failed to describe complex composition of activities. Gupta et al. [19] used a context-free (AND-OR) grammar to solve the problem of activity classification, which focused on recognition of atom level action. One of the main drawbacks is the use of 2D video leads to relatively low accuracy. Ryoo and Agrawal [37] described a method for recognizing complex human activities using a context-free grammar (CFG) based representation scheme. However, it is not able to learn representations of activities automatically in large training data.

2) Nonhierarchical approaches can deal with simple and short activities such as primitive action and periodic activities that they recognized from unknown image sequences. They can be further divided into space-time approaches and sequential approaches. The space-time approaches recognize activity with volume, trajectories, and set of features [11]. Bobick and Davis [5] used two components, i.e. MEI and MHI, for representation and recognition of human activities. It constructed a vector image, which was then matched against a stored representation of known movements. However, it is only applicable to these situations where the motion of object movement can be separated easily. Shechtman and Irani [38] introduced a behavior-based similarity measure that tells us whether two

different space-time intensity patterns of two different video segments could have resulted from a similar underlying motion field. By examining the intensity variations in video patches, the space of their possible motions can be implicitly characterized. Oikonomopoulous et al. [33] proposed a representation of human action as a collection of many short trajectories, which are extracted by a particle filtering tracking method. They used a longest common subsequence algorithm to verify different sets of trajectories. Vishwakarma and Agrawal [45] considered multiclass activities fused in a three-dimensional (spatial and time) coordinate activity recognition system to achieve maximum accuracy. They quantized feature vectors of interest points utilizing a histogram. This method worked well in semantically varying events and was robust to scale and view changes. Sequential approaches represent human activities with a sequence of actions and recognize activities by analyzing a sequence of features extracted from input video [1]. Motivated by psychological studies of human perception, Gupta and Davis [18] proposed a probabilistic model (a Bayesian approach which unifies the inference processes) that exploited the contextual information for visual action analysis to improve object recognition as well as activity recognition. By placing object classification and localization in this framework, they can detect and recognize activities that are hard to recognize due to lack of discriminative features. This method is based on object detection, while accurate detection can be hard to get in complex environment, and the computational complexity is high. Natarajan and Nevatia [30] proposed coupled Hidden Semi-Markov Model (CHSMM) to recognize human activity. They demonstrate the algorithm's effectiveness for representing the interaction between multiple people and show its utility by experiments with synthetic and real data. However, a major limitation of this model is its high complexity.

The goal of activity recognition is to recognize common human activities in real life settings. Common application that make use of activity recognition is health-assistive smart homes and smart environments, such as the Activities of Daily Living (ADLs) system [1] monitoring the functional health of a smart home resident [1, 3, 43], et al. The goal of this paper is to find the methods of recognizing fighting activities in videos through studying the performance of modern action recognition approaches. Most of previous works on action recognition focus on simple human actions like walking, jumping or hand waving [1, 3, 8, 11, 26, 36, 43, 47]. In this paper, we focus on the challenging task of detecting violence in videos. Despite its potential usefulness, violent action detection has been less studied compared to action recognition where there are many well-studied datasets available, but much less significant datasets for violent actions.

In this paper, we propose a fast and robust framework for violence detection in surveillance scenes. The primary contributions of this paper are two-fold. First, we propose a Gaussian Model of Optical Flow (GMOF) to extract candidate violence regions. Then, a novel descriptor, named as Orientation Histogram of Optical Flow (OHOF), is constructed on each video volume by densely sampling the candidate violence regions for violence classification.

The proposed method consists of two main modules: the training module and the detection module, as shown in Fig. 1. The training module consists of selecting representative training data and extracting the OHOF descriptor, and then obtaining the feature model using a linear SVM. The detection module mainly detects candidate violence regions in the video sequence, and determines the accurate regions of violent activities. In the detection module, firstly, GMOF is proposed to extract candidate violent regions. Then, our novel OHOF descriptor

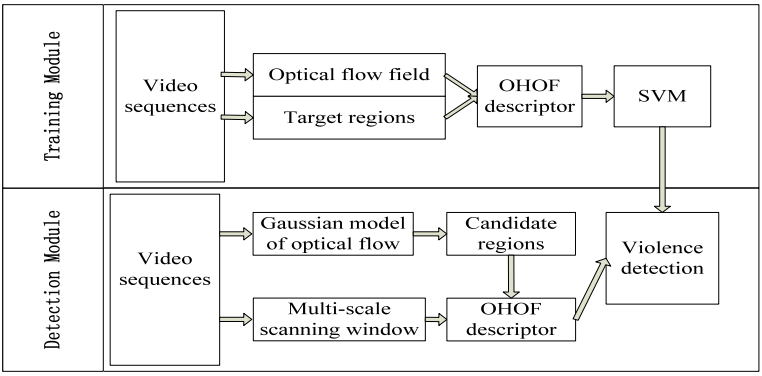

Fig. 1 The framework of the proposed method

is extracted using a multi-scale scanning window technique in densely sampled candidate violence regions. Lastly, it is matched against a trained SVM model of known violent activities. Experimental results on three challenging datasets have been conducted to demonstrate the superiority of our proposed approach over the state-of-the-arts.

The rest of this paper is organized as follows: Section 2 introduces related works on violence detection. Section 3 details our candidate violence regions detection algorithm, which is based on the newly proposed GMOF. Section 4 presents our OHOF descriptors and multi-scale scanning window technique that are used for violence detection. In Section 5, experimental results and analysis are presented. Finally, conclusions are drawn in Section 6.

2 Related work

Violence detection has been less studied compared to the related issue of action recognition. Up to now, there have been some developmental systems for violence detection. In early attempts, some violent behavior detection methods are based on audio feature [9, 11, 25, 29]. Nam et al. [29] proposed recognizing violent scenes in videos using flame and blood detection and capturing the degree of motion, as well as the characteristic sounds of violent events. It is the first attempt for violence recognition in video. Based on GMM and HMM, Cheng et al. [9] recognized gunshots, explosions and car braking using a hierarchical approach. However, this kind of method has many disadvantages, such as low detection rate, high false alarm etc. Datta et al. [15] relied on motion trajectory information and orientation information of a person's limbs to detect violent behavior. This method requires foreground segmentation to extract precise silhouettes, which is difficult in a real environment. Clarin et al. [10] presented a system that uses a Kohonen self-organizing map to detect skin and blood pixels in the video sequences and motion intensity analysis to detect violent actions involving blood. The use of skin color information has limited its applications.

As introduced above, we note that there are important applications, particularly in surveillance, where audio is not available and where the video comes in grey scale. In recent study, some methods based on spatiotemporal interest-point, e.g., STIP [16], MoSIFT [3], have been proposed for violence detection. After extracting interest points over the frames, the Bag-of-Words (BoW) framework is used for violence recognition. This kind of methods computes

only in the regions of interest (located around the limited detected interest points) and are not discriminative enough. Moreover the BoW model roughly assigns each feature vector to only one visual word and ignores the spatial relationships among the feature. Hassner et al. [20] detected crowd violence using the ViF descriptor formed from computing a magnitude-change map of optical flow over time. The performance of this method degrades significantly when dealing with faces with non-crowded scenes. Optical flow is used to represent apparent velocities of movement of brightness patterns in an image [1], which has been employed for modeling typical motion patterns [7, 44]. This measure may also become unreliable in extremely crowded scenes. A dense local sampling of optical flow has been adopted to solve this issue [4].

The above methods are effective on a close-up scene like the Hockey dataset [3], but they cannot handle scenes where human behaviors only occurs in a small region, like the BEHAVE dataset [41] and the CAVIAR dataset [42]. In such scenes, there may not be enough interest points for an accurate judgment and most cells of the ViF descriptor are actually nonviolent.

Targeting the above challenges, this paper proposes a simple but robust violence detection algorithm. Our contributions are mainly in the following three aspects:

1) We propose a GMOF method to extract candidate violence regions, which are adaptively modeled as a deviation from the normal behavior of crowd observed in the scene. Since our method was established in cells for anomaly motion detection, this not only avoids unnecessary computation and speeds up the system but also contributes to a high detection rate.
2) Multi-scale scanning window technique is used to search for violent events in the densely sampled candidate violence regions. This helps to overcome the problems of different sizes and different camera distances, also avoids the use of multi-scale spatiotemporal features with minimum computation.
3) We propose a novel OHOF descriptor to distinguish violence from these candidate violence regions. It is constructed by rearranging the histogram, adding contextual information and normalization. The descriptor is scale and rotation invariant, so the new OHOF descriptor is very effective.

It is worth mentioning that, the preliminary work of our idea has been published in [48]. In this paper, we further extended our idea and adopted multiple GMMs to determine the candidate violent regions and proposed multi-scale scan technology. Also, more details of our proposed approach are provided for clarity and integrity. More comparative experiments on more benchmarking data sets have also been conducted, all demonstrating the superiority of our simple but effective approach.

3 Candidate violence regions extraction

This section presents our candidate violence region detection algorithm using our new GMOF, i.e. Gaussian Model of Optical Flow.

Although many strategies have appeared, the estimation of human motion is still a challenging task [1]. Optical flow is a convenient and commonly used motion representation approach. The optical flow approach is based on unchanged gray gradient and constant brightness and can detect an object of independent motion without knowing any information of the scene in advance. It is the vector field that describes how the image changes with time

[1, 2, 20, 36]. Optical flow can be affected by illumination variation and view angle change. Also, when the contrast of the target and background image is low, or there exists noise in the image, this approach can lead to high false alarm rates. On the other hand, the size of persons and fighting regions in a scene varies in a large scale, so detecting by searching different scales requires a great amount of calculation, while this high computational cost is unacceptable for real-time applications.

Targeting the above challenges, we propose a simple, two-step strategy. First, we perform a rough searching by using the magnitude information of optical flow field, which generates candidate violence regions with motion. Secondly, for these candidate regions, we use a novel and more effective feature descriptor to distinguish violence from non-violence regions. Compared with the traditional mechanism, our strategy adopts the advantages of optical flow without suffering the high computational cost. In this section, we focus on the first step, i.e. candidate region detection, and will present the second step, i.e. violence classification, in Section 4.

Due to clutter and complicated occlusions, background is always dynamic and noisy. Among the most mature methods for statistically modeling an observed scene, GMM [40, 50] has been extensively adopted in surveillance applications for background/foreground separation. It can be regarded as a type of unsupervised learning methods, and more importantly, it can adapt to scene changes.

Inspired by the well-known GMM approaches, in our work, we adopt GMM to produce candidate violence regions of motion features extracted from the magnitude information of Optical Flow, and name this method as *Gaussian Model of Optical Flow*, i.e. GMOF. These motion features are exploited to learn repetitive variations of crowd scenes for GMM, which models the distribution of normal behaviors. We build a GMOF to obtain candidate violence regions. Different from the traditional GMM, our GMOF aims to detect anomaly in motion rather than in pixels' values. We build a background model of optical flow by using the magnitude information of optical flow in a defined image block. According to this model, the abnormal block of optical flow (greater than the values in the model) will be considered suspicious and marked as candidate violence regions. The candidate regions will be further verified as violence or nonviolence in Section 4.

As discussed before, the GMM is adopted to learn the behavior of motion features extracted from the optical flow. In order to reduce unnecessary computation, a grid of optical flow is disposed on the video frame, which is repeatedly initialized over a temporal window of a video sequence. Unlike other methods, our GMOF is established on cells rather than on each pixel to reduce unnecessary computation and improve the robustness. We partition a video frame into $n \times n$ grids (each cell with a size of 4×4) with an overlapping of 50 % for robustness. In each grid, we calculate the mean value of the magnitudes of the optical flow vectors, build and update the Gaussian model.

Motion features, defined in terms of velocity magnitudes, are extracted by tracking the grids using the Lucas-Kanade optical flow [6]. We do not consider the grids having motion features with very low magnitudes.

According to the GMM framework, every new motion feature is checked against the existing distributions for that cell, and is incorporated into the distribution if a match is found; otherwise, it forms a new distribution indicating a new cluster. This forms the basis of the adaptability of our GMOF.

At any time t, let us denote the history set of the motion features (in term of velocity magnitude m) of a cell as $\{m_1,\cdots,m_t\}$, which is modeled by a mixture of K Gaussian distributions. Given the optical flow field (u,v) of each pixel, the velocity magnitude m for a 4×4 cell can be calculated as:

$$m = \frac{1}{16}\sum \sqrt{u^2 + v^2} \tag{1}$$

The probability of observing the current cell p is calculated as:

$$p(m(p),t) = \sum_{k=1}^{K} w_k(t)g\left(m(p),\mu_k(t),\sigma_k^2(t)\right) \tag{2}$$

where K is the number of distributions (set as 3 in our case), $w_k(t)$ is the weight of the k^{th} Gaussian in the mixture at time t and $\sum_{k=1}^{K} w_k(t)=1$, $\mu_k(t)$ and $\sigma_k^2(t)$ are the mean and covariance of the k^{th} Gaussian in the mixture at time t respectively, and g is a Gaussian probability density function.

At time $t=1$, we start with initializing $\mu = m$ and $\sigma = \overline{m(p)}$, where $\overline{m(p)}$ is the mean value of Optical Flow magnitude of the whole image. Based on the persistence and the variance of each Gaussian distribution, we determine which Gaussians can be associated to the crowd model. This can be achieved by:

$$|m(p)-\mu| \le n\sigma \tag{3}$$

If Eq. (3) is satisfied, the parameters of the distribution that matches the new observation are updated as follows:

$$\begin{aligned} w_k(t+1) &= (1-\alpha)w_k(t) + \alpha \\ \mu_k(t+1) &= (1-\beta)\mu_k(t) + \beta m(p) \\ \sigma_k^2(t+1) &= (1-\beta)\sigma_k^2(t) + \beta(\mu_k(t)-m(p))^2 \end{aligned} \tag{4}$$

where α and β are the pre-set weight-learning and mean/variance-learning rate respectively. A low weight-learning rate indicates that the new motion feature will be incorporated slowly into the model.

For these unmatched distributions, the weight will be updated according to Eq. (5) and mean and variance remain unchanged.

$$w_k(t+1) = (1-\alpha)w_k(t) \tag{5}$$

If Eq. (3) is not satisfied, the Gaussians are ordered by the value of $w_k(t)/\sigma$, which increases both as a distribution gains more evidence and as the variance decreases. After re-calculating the parameters of the mixture, it is sufficient to sort from the matched distribution towards the most probable normal crowd distribution, because only the relative value of matched models will have changed. This way of updating not only ensures the effectiveness and robustness of our constructed GMOF, but also improves the speed of model updating and reduces the computation load.

At last, candidate violent regions are determined by the following rules: for a given $m(p)$, if Eq. (6) is satisfied, it will be regarded as a candidate violence region; otherwise, it will be abandoned.

$$\arg\min\left(\sum_{k=1}^{C} m(p) > T_{thresh}\right) \tag{6}$$

where C denotes the number of satisfied distributions, T_{thresh} is a measure of verifying candidate violence regions, which takes the best distributions until a certain portion. If a small value for T_{thresh} is chosen, the crowd model is usually unimodal; otherwise, if T_{thresh} is higher, a multi-modal distribution can be included in the crowd model.

In our experiment, based on the statistical results of a large amount of experimental data, T_{thresh} is set to $2.5 \times \mu$.

Compared with traditional pixel-based algorithms which update all pixels for every frame, our algorithm has the ability to selectively update region information within each frame, while offering the capability to real-time processing. Figure 2 presents results of the candidate violence regions detection by our GMOF model obtained on different image styles, where the localized candidate violence regions are represented in green and they need to be further verified in the following step. As it can be seen, our extraction results can capture all abnormal regions (including violence regions) in motions and help to avoid unnecessary computation in next violence detection stage.

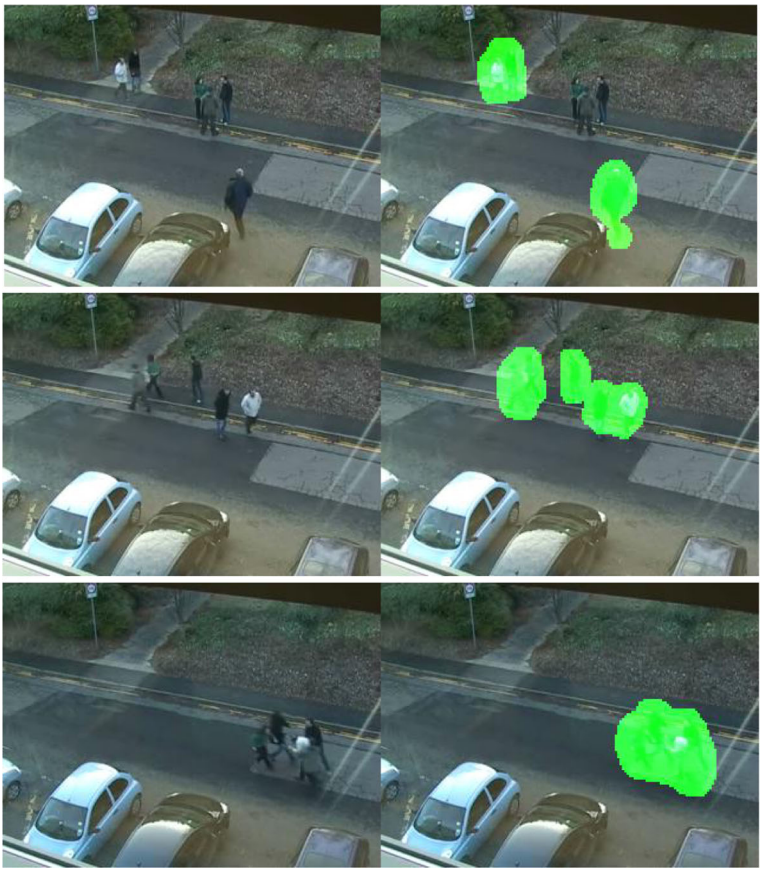

Fig. 2 Results on candidate violence regions detection (indicated in *green*) using our proposed GMOF model

4 Violence verification

In this section we present the details of our violence verification algorithm based on our newly proposed OHOF descriptor and the multi-scale scanning window technique.

In Section 3, we have obtained candidate violence regions by making use of the magnitude information of the optical flow vectors. To further determine whether a candidate region contains violent actions or not, we have observed that the orientation information of optical flow vectors can be a distinctive feature.

As shown in Fig. 3, the distributions of the orientations of optical flow in the regions of violence and non-violence are obviously different. In a fighting scene (see the bottom-right sub-figure in Fig. 3), due to the irregular motions of arms and legs, the distribution of orientations is very chaotic. However, in a normal scene, when people walk, run and ride, they are mostly distributed regularly in a certain orientation. Inspired by this phenomenon, we propose to use the distribution of orientations of optical flow to verify whether there presents violent behavior in a candidate region. For this purpose, we construct a new descriptor, named as OHOF – Orientation Histogram of Optical Flow. Here, it is worth mentioning that this orientation distribution feature will fail without first extracting the candidate violence regions. This is because the iterative process of optical flow computation will cause these non-motion regions also appear with chaotic orientations.

Our proposed violence verification algorithm contains two main aspects: (1) A multi-scale scanning window is used to search for violence events in the densely sampled candidate violence regions; (2) The OHOF descriptor is extracted for each image area covered by the scanning window from these candidate violence regions to distinguish violence for non-violent actions. The detailed process is described in the next sub-section.

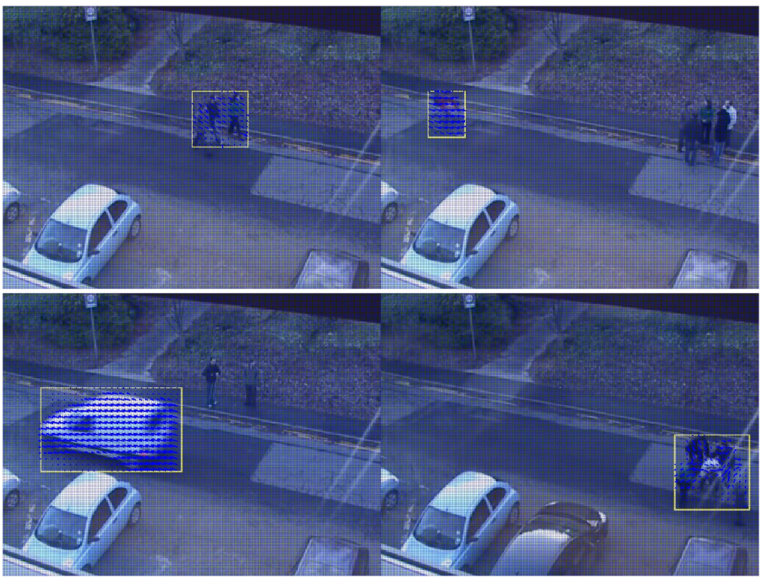

Fig. 3 Orientation distribution of optical flow in different scenes (denoted as a *yellow box*)

4.1 Multi-scale scan window

Fighting events are generated by human movements and can appear in different parts of the scene. Also, their sizes can vary significantly due to their different ranges to the camera. It is thus necessary to analyze video at multiple scales. Spatiotemporal features that are densely sampled on a grid of cuboids have been used for human action recognition [46]. This approach allows the localization of anomaly in terms of the position both in a frame and in time, with its precision depending on the size and overlap of the cuboids. It also states the fact that different parts of a scene may be subject to different anomalies, emergency cases, etc. Moreover, it can achieve real-time processing speed, since it does not require spatiotemporal interest point localization. This approach has therefore been widely used for typical surveillance scenes. Bertini et al. [4] proposed a multi-scale non-parametric anomaly detection approach that can be executed in real-time in a completely unsupervised manner. The approach is capable of localizing anomalies in spatiotemporal, and can be applied to complex scenes containing different sizes of pedestrians at different distances with the camera.

If we only consider local feature information, it can be impossible to differentiate violent actions (such as a person's hand or leg movement when fighting) from nonviolent actions (such as walking, running and jumping), because these two types of behaviors can appear very similar from their local properties. Modeling interaction behavior between people is a possible solution to the problem. The multi-scale and spatiotemporal feature approach in [4] can solve the above problem. However, when the feature descriptor is complicated, its high computation load becomes unacceptable, especially for our application that requires real-time processing.

In our paper, we propose to use the scanning window technique to address the high computation load problem of the multi-scale and spatiotemporal feature approach. In our approach, we limit the scales of scanning window to three levels. In order to detect fights in a scene, we propose to use a scalable searching window to traverse the video images in turn, where the searching window is scanned with patches of different size, with a scaling factor of *3*. In order to further reduce the computation load, our multi-scale scanning method takes the candidate violence regions obtained in previous steps (in Section 3) as input and process all these regions in order to extract more effective and representative features. This allows the system to filter spurious small false positives and increases the capability of the system to accurately localize even smaller subjects. In our implementation, in order to balance the execution speed and performance, three scales are used (i.e. 72 × 72, 24 × 24 and 8 × 8) with a 50 % overlap (i.e. 36 pixels, 12 pixels and 4 pixels overlap respectively). The detailed procedure of our multi-scale scanning technique is as follows:

```
Algorithm 1 Multi-scale Scanning
While there exits input frame in the last loop
    Step 1. Build scanning windows of three scales, 72×72, 24×24
and 8×8.
    Step 2. Traverse video sequence images in turn by multi-scale
searching in a step of eight pixels.
    Step 3. If the current scan window spans more than half of
candidate violence region, jump to Step 4; otherwise, go back to Step
2.
    Step 4. Update candidate violence regions with these scanning
window regions, and mark them as new candidate violence regions.
```

```
        Step 5. Densely sample new candidate violence regions using
the method in [43], and go back to Step 2.
  end
```

4.2 The OHOF descriptor

In this work, we propose a novel feature descriptor, i.e. Orientation Histogram of Optical Flow (OHOF), to distinguish violence regions from other non-violence regions. In order to be more discriminative, the orientation histogram of optical flow is constructed by rearranging the bins, adding contextual information and normalization, detailed in the following steps.

1) Optical flow field is firstly extracted using the Lucas-Kanade optical flow method [2]. For each pixel, two orientation magnitudes are obtained, denoted as F_x and F_y, which can be expressed using polar coordinates:

$$\begin{cases} \sqrt{F_x^2 + F_y^2} \\ \theta = \arctan\dfrac{F_x}{F_y} \end{cases} \tag{7}$$

2) Multi-scale scanning window technique is used in the candidate violence regions. Build a histogram with 16 orientations of candidate violence regions, as shown in Fig. 4b. The computation method of orientation histogram of optical flow is as follows:

$$\theta' = ang\left(F_x, F_y\right) + \pi \tag{8}$$

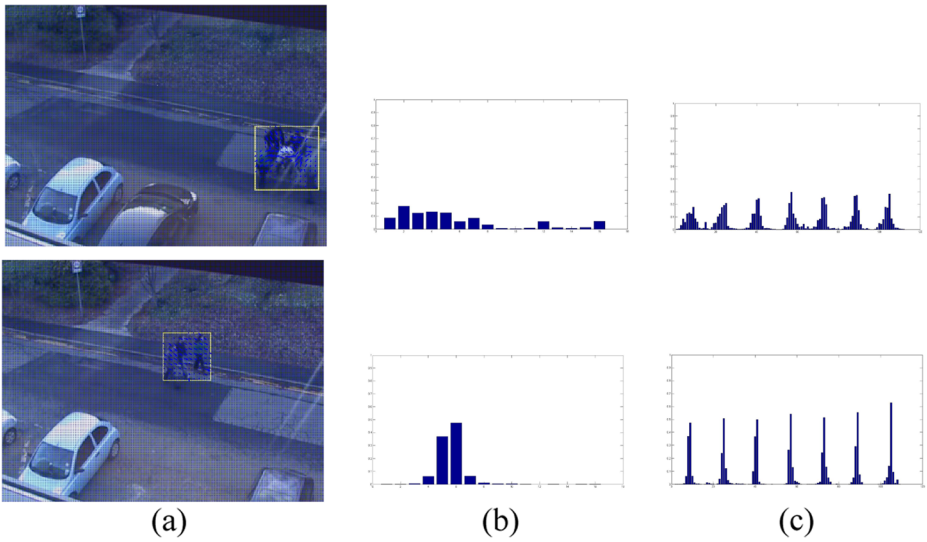

Fig. 4 The procedure of constructing the OHOF descriptor. **a** Optical Flow extracted from the candidate violence regions (*circled in yellow boxes*) **b** The orientation histogram on candidate violence regions **c** Context-based orientation histogram

$$ang(F_x, F_y) = \begin{cases} \theta, F_x > 0 \& F_y > 0 \\ \theta + \pi, F_x > 0 \& F_y < 0 \\ \theta - \pi, F_x < 0 \& F_y < 0 \\ \theta, F_x < 0 \& F_y > 0 \end{cases} \quad (9)$$

Note that, during the training, we extract the motion parts as our candidate violence regions using a background subtraction method, i.e. GMM in [50]. During the detection, each frame is partitioned into multi-scale overlapping grid and candidate violence regions are those with enough abnormal cells according to our GMOF method.

3) Rearrange the bins of the resultant histogram. Keep these orientations which gradient values are larger than T_{thresh} (see Section 3), and then sort the histogram bins in descending order according to their values. The purpose of this operation is to make the resultant OHOF feature descriptor rotation invariant.
4) Add context information. Considering that normally people's behavior is coherent, single frame alone is not sufficient to describe the feature. Histograms of the six previous continuous frames are added to the descriptor to achieve a robust performance. Therefore, the whole descriptor is 16×7=112 dimensions as shown in Fig. 4c.
5) Histogram normalization. In order to address the disturbance resulted from tiny light changes and eliminate the differences of region sizes, the above obtained histogram is normalized.

Clearly, the descriptor is scale invariant because it only measures orientation information, and it is also rotation invariant thanks to the rearrangement of histogram bins. The descriptors extracted from all windows are then used to train an SVM classifier.

5 Experimental results

5.1 Dataset

To demonstrate the performance of our proposed violence detection algorithm, experiments were conducted on three public datasets: the BEHAVE dataset, the CAVIAR dataset and Crowd Violence dataset.

The BEHAVE dataset The BEHAVE dataset contains more than 200,000 frames (image resolution: 640×480 pixels) with various scenarios, including walking, running, chasing, discussing in group, driving or cycling across the scene, fighting and so on. We partitioned the dataset into clips with various activities and manually labeled them as violence or non-violence. Each clip consists of at least one hundred frames. Finally, we randomly picked 80 clips for violence detection, including 20 violence clips and 60 non-violence clips.

The CAVIAR dataset The CAVIAR dataset (image resolution: 384×288 pixels) were recorded acting out the different scenarios of interest. These include people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place. It contains two sets of data filmed in two different scenes, i.e. the entrance lobby of INRIA Labs and a hallway in a shopping center in Portugal Clips taken from INRIA contain 4 clips of fighting and 24 clips with other activities like

slumping or fainting, shaking hands, people meeting, walking together and splitting. Finally, we picked 28 clips for violence detection.

Crowd violence dataset This dataset is assembled for testing violent crowd behavior detection. All video clips are collected from YouTube, presenting a wide range of scene types, video qualities and surveillance scenarios. The dataset consists of 246 video clips including 123 violence clips and 123 normal clips with a resolution of 320×240 pixels. The whole dataset is split into five sets for 5-fold cross-validation. Half of the footages in each set present violent crowd behavior and the other half presents non-violent crowd behavior.

5.2 Experiment settings

In our experiments, we have made few attempts to optimize the parameters used in our method, and so improved performance may be obtained by exploring other values. The GMOF is performed on each 4×4 cell and updated every 5 frames. The OHOF descriptor is classified by a linear SVM [24]. We selected 10 violence clips and 30 non-violence clips from the BEHAVE and Crowd Violence dataset, and labeled each frame manually. These clips are enough for training because each clip involves hundreds of frames and the whole 40 clips contains nearly 20,000 frames with various scenarios of human interactions. During the violence detection, the frames are again partitioned into multi-scale (three scales are chosen in our experiments, i.e. 72×72, 24×24 and 8×8, to balance the computation cost and the performance) and densely sampled grids. In each grid with enough labeled candidate violence cells, the OHOF descriptor is performed for violence detection.

5.3 Results and discussion

We compare the proposed method against the state-of-the-art techniques. Results are reported as both mean prediction accuracy (ACC)±standard deviation (SD) as well as the area under the ROC curve (AUC).

Results on the BEHAVE dataset 20 clips of this dataset were picked for training. In order to demonstrate the superior performance of our algorithm, we compare our algorithm with those of the state-of-the-art approaches implemented by us, including HOG, HOF, HNF (combination of HOG and HOF) [16], MoSIFT [3] and ViF [20]. For these spatiotemporal descriptors (HOG, HOF, HNF, MoSIFT), the dictionary size is fixed to 500 in our experiment. Table 1 shows the results of the 60 video clips. Our method is well performed over all group fighting clips. It has only failed to detect the scenes where two people are fighting with chasing, which is hard to be distinguished from running. Overall, all individual behaviors are well distinguished. And our method also does not make a decision of violence when the group of people standing, discussing with some kind of movement. False alarm only happened when a group of people got together to do some exaggerated activities.

As we can see, among the six different approaches, our method achieved the best performance, followed by ViF, MoSIFT, HOG and HNF. On the contrary, HOF turns out to be the least discriminative in classifying violence and non-violence. That is because this dataset contains a lot of walking, running, cycling behavior, and these approaches based on MoSIFT, HOG, HNF and HOF are less discriminative in differentiating these behavior.

Table 1 Classification results on BEHAVE database

Algorithms	Accuracy (±SD)	AUC
Ours	85.29±0.16 %	88.78 %
HOG [16]	58.69±0.35 %	63.22 %
HOF [16]	59.91±0.28 %	58.93 %
HNF [16]	57.97±0.31 %	60.89 %
ViF [20]	82.02±0.19 %	85.92 %
MoSIFT [3]	62.02±0.23 %	65.78 %

Results on the CAVIAR dataset 4 fighting clips and 24 non-fighting clips were tested, while the model was trained from the BEHAVE and Crowd Violence dataset. Scenes of these clips are completely different from the BEHAVE dataset, with different views, illumination and camera distance. The results of the state-of-the-art approaches, including HOG, HOF, HNF [16], MoSIFT [3] and ViF [20] are compared. For these spatio-temporal descriptors (HOG, HOF, HNF, MoSIFT), the dictionary size is also fixed to 500. Table 2 shows the evaluation results obtained on the 28 clips. Our method is well performed on all group fighting clips. Although no training data is selected from this dataset, our algorithm is still effective on this dataset. Some false alarms are caused by the camera shaking, which has affected the computation of Optical Flow.

As we can see, among the six different approaches, our method still achieved the best detection accuracy, followed by ViF. These approaches (MoSIFT, HOG, HNF and HOF) are also less discriminative. That is because this dataset is rich in texture and somehow difficult to classify due to the large within-class difference.

Results on the crowd violence dataset This dataset is more challenging than the above two dataset because it contains many crowded scenes. The set contains 246 clips divided into five splits, each containing 123 violent and 123 non-violent scenes. Methods are required to detect violence in a 5-folds cross validation test. HOG, HOF, HNF [16], MoSIFT [3] and ViF [20] are compared. In order to assess the impact of vocabulary size, we generated vocabularies of 100 and 500 words. Table 3 shows the comparison results.

It can be shown in Table 3, our algorithm outperforms the other methods compared, especially for these spatiotemporal descriptors (HOG, HOF, HNF, MoSIFT). Our algorithm and ViF [20] approach perform comparably. Note that increasing vocabulary size does not improve detection accuracy greatly and uniformly. These approaches based on MoSIFT, HOG,

Table 2 Classification results on CAVIAR dataset

Algorithms	Accuracy (±SD)	AUC
Ours	86.75±0.15 %	89.68 %
HOG [16]	59.37±0.33 %	64.12 %
HOF [16]	60.93±0.27 %	59.63 %
HNF [16]	59.05±0.30 %	62.19 %
ViF [20]	83.92±0.17 %	86.42 %
MoSIFT [3]	64.02±0.22 %	67.39 %

Table 3 Classification results on Crowd Violence dataset

Algorithms	Accuracy (±SD)	AUC
Ours	82.79±0.19 %	86.59 %
HOG Vac100 [16]	55.07±0.39 %	58.42 %
HOG Vac500 [16]	57.29±0.37 %	60.91 %
HOF Vac100 [16]	55.81±0.38 %	54.25 %
HOF Vac500 [16]	58.11±0.32 %	57.13 %
HNF Vac100 [16]	53.42±0.41 %	56.01 %
HNF Vac500 [16]	55.97±0.36 %	58.89 %
MoSIFT Vac100 [3]	56.11±0.35 %	59.17 %
MoSIFT Vac500 [3]	57.09±0.37 %	60.73 %
ViF [20]	81.30±0.21 %	85.00 %

HNF and HOF have performed the worst on this dataset. This is because these approaches all rely on magnitude information, which can be seriously affected by wider distance ranges.

Results on above three dataset demonstrate that our algorithm is very effective for detecting violence in surveillance scene.

Our computational complexity tests were performed on a 3Gb RAM, Intel core i5 computer running Windows 7. We compare only our algorithm to ViF, because these spatiotemporal approaches performed too slowly for real-time processing, requiring 0.31 s per frame just for STIP feature extraction. Table 4 summarizes the run-times for the two methods.

In order to further demonstrate the superior performance of our algorithm compared with the ViF method, to be intuitive, we created three new measures, i.e., well performed, less well performed and failed. For violence clips, well performed are those clips from which violence is well detected during the whole behavior (more than 50 % of the frames); less well performed denotes clips with only a few of the violence frames are detected. For non-violence clips, only clips without any false alarm will be regard as well performed; less well performed denotes the clips with less than 2 % false alarm. We evaluate these measures on the above three dataset (in total 334 clips), as shown in Table 5. As it shows, our algorithm owns obvious advantage compared to the ViF method.

Some examples of violence detection are shown in Fig 5, where the detected violence regions are marked in red. Our method can not only determine whether there is fighting behavior or not in a frame, but also accurately determine the position where the fighting happened. As it shows, our algorithm is able to handle heads with various scenes, including full 360-degree rotation and up to 90-degree tilting, different camera distance, severe occlusion between people and crowed cases. Figure 6 shows the ROC curve of different methods on the whole dataset.

Table 4 Run-time comparison of two methods

Algorithms	Processing time per frame (ms)
Ours	36
ViF [20]	43

Table 5 Detection results on three dataset

Algorithms	Well performed	Less well performed	Failed
Ours	276 clips	45 clips	13 clips
ViF [20]	224 clips	89 clips	21 clips

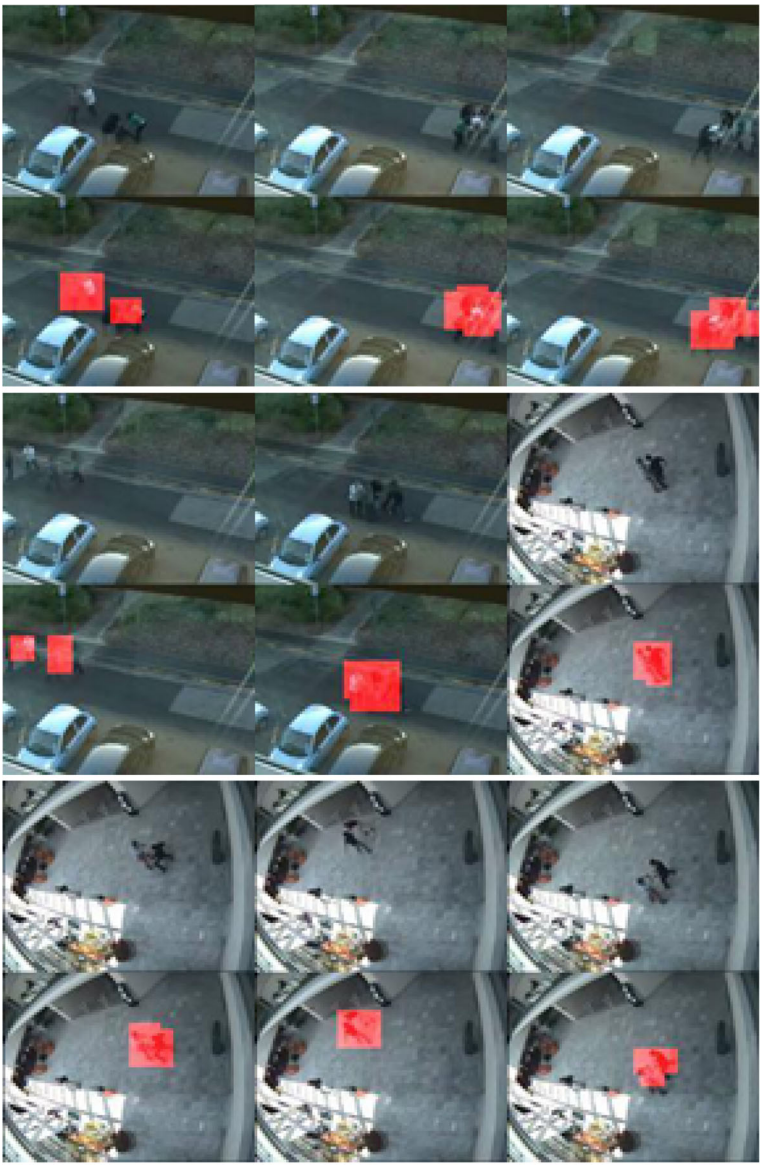

Fig. 5 Some examples of our violence detection results where the detected violence is marked in *red*

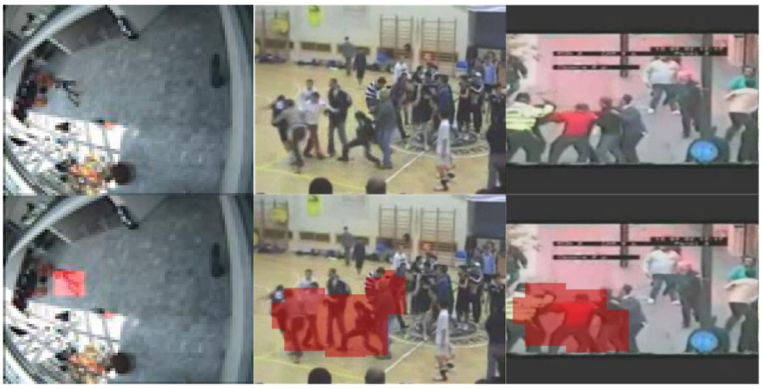

Fig. 5 (continued)

Unsurprisingly, the STIP representations are better suited for structured videos rather than the more textural videos in these datasets. When the distance between human and camera becomes larger, our method can still detect the violence behavior . This is due to the fact that our GMOF approach applied to candidate violence regions is very effective, and the proposed scale and rotation invariant OHOF descriptor is discriminative.

By verifying the obtained results, we can find that our proposed system is effective and robust for the correct detection of violence. This detector can handle human occlusion, arbitrary camera distance, textured foregrounds and backgrounds, and multiple moving people in the background, all simultaneously.

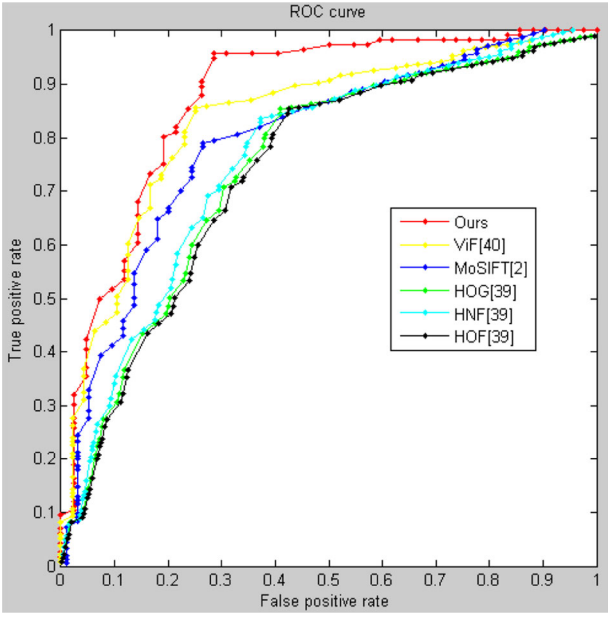

Fig. 6 ROC curves on fight detection using different methods on all datasets

6 Conclusion

Our work aims to establish a fast violence detection method in surveillance scenes. In this paper, a novel and fast violence detection method is proposed. The GMOF is firstly used to extract candidate violence regions. Then a novel OHOF descriptor is performed on each video volume constructed by densely sampling the candidate violence regions. Experimental results on three challenging datasets have demonstrated the superiority of our proposed method, which not only detects but also localizes violence in both crowded and non-crowded surveillance scenes. The merits of our work are: (1) GMOF is proposed to detect candidate violence regions. The method adopts the GMM to learn the behavior of the crowd. It is fast and also robust to complex background; (2) A multi-scale scanning window technique is proposed to capture the scene changes, which contributed to a high detection and precision rate; (3) A novel OHOF descriptor is proposed, which is scale and rotation invariant. Last but not least, our method has strong real time property.

Acknowledgments This research is partly supported by NSFC, China (No: 61273258, 61105001).

References

1. Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv 43:1–43
2. Beauchemin SS, Barron JL (1995) The computation of optical flow. ACM Comput Surv (CSUR) 27(3):433–466
3. Bermejo E, Deniz O, Bueno G, and Sukthankar R (2011) Violence detection in video using computer vision techniques. Proc. of the 14th Int Conf Comput Anal Images Patterns II: 332–339
4. Bertini M, Bimbo AD and Seidenari L (2012) Multi-scale and real-time non-parametric approach for anomaly detection and localization. CVIU 320–329
5. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
6. Bouguet JY (1999) Pyramidal implementation of the Lucas Kanade feature tracker. Microsoft Res Labs Tech Rep
7. Chen MY, Hauptmann A (2009) MoSIFT: recognizing human actions in surveillance videos. Tech Rep Carnegie Mellon University
8. Chen D, Wactlar H, Chen M, Gao C, Bharucha A, Hauptmann A (2008) Recognition of aggressive human behavior using binary local motion descriptors. Eng Med Biol Soc 20:5238–5241
9. Cheng WH, Chu WT, Wu JL (2003) Semantic context detection based on hierarchical audio models. In: Proc ACM SIGMM Work Multimedia Inf Retr 109–115
10. Clarin CT, Dionisio JAM, Echavez MT, Naval PCJ (2005) DOVE: detection of movie violence using motion intensity analysis on skin and blood. Tech Rep University of the Philippines
11. Cristani M, Bicego M, Murino V (2007) Audio-visual event recognition in surveillance video sequences. IEEE Trans Multimedia 257–267
12. Cupillard F, Bremond F, Thonnat M (2002) Group behavior recognition with multiple cameras. WACV 177–183
13. Dai P, Di H, Dong L, Tao L, Xu G (2008) Group interaction analysis in dynamic context. IEEE Trans Syst Man Cybern 38(1):275–282
14. Damen D, Hogg D (2009) Recognizing linked events: searching the space of feasible explanations. CVPR 927–934
15. Datta A, Shah M, Lobo NDV (2002) Person-on-person violence detection in video data. ICIP 433–438
16. de Souza FDM, Chavez GC, do Valle EA, de A Araujo A (2010) Violence detection in video using spatio-temporal features. SIBGRAPI 224–230

17. Gong S, Xiang T (2003) Recognition of group activities using dynamic probabilistic networks. ICCV 2:742–749
18. Gupta A, Davis LS (2007) Objects in action: an approach for combining action understanding and object perception. CVPR pp 1–8
19. Gupta A, Srinivasan P, Shi J, Davis LS (2009) Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. CVPR 2012–2019
20. Hassner T, Itcher Y, Kliper-Gross O (2012) Violent flows: real-time detection of violent crowd behavior. CVPRW 1–6
21. Huesmann LR, Moise-Titus J, Podolski CL, Eron LD (2003) Longitudinal relations between children's exposure to TV violence and their aggressive and violent behavior in young adulthood: 1977–1992. Dev Psychol 39:201–221
22. Intille SS, Bobick AF (1999) A framework for recognizing multiagent action from visual evidence, In: AAAI-99. AAAI Press, Menlo Park, pp 518–525
23. Ivanov YA, Bobick AF (2000) Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 22(8):852–872
24. LIB-SVM. http://www.csie.ntu.edu.tw/~cjlin/
25. Lin J, Wang WQ (2009) Weakly-supervised violence detection in movies with audio and video based co-training. PCM 990–935
26. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. CVPR 1975–1981
27. Minnen D, Essa I, Starner T (2003) Expectation grammars: leveraging high-level expectations for activity recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2:626–632
28. Moore D, Essa I (2002) Recognizing multitasked activities from video using stochastic context-free grammar. Proc AAAI Natl Conf AI 770–776
29. Nam JH, Alghoniemy M, Tewfik AH (1998) Audio-visual content-based violent scene characterization. ICIP 353–357
30. Natarajan P, Nevatia R (2007) Coupled hidden semi Markov models for activity recognition. IEEE Work Motion Video Comput pp 1–8
31. Nevatia R, Zhao T, Hongeng S (2003) Hierarchical language-based representation of events in video streams. CVPR Work 4:39–47
32. Nguyen NT, Phung DQ, Venkatesh S, Bui H (2005) Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. CVPR 2:955–960
33. Oikonomopoulos A, Patras I, Pantic M, Paragios N (2007) Trajectory-based representation of human actions. Artif Intell Hum Comput 4451:133–154
34. Oliver N, Horvitz E, Garg A (2002) Layered representations for human activity recognition. Proc. 4th IEEE Int Conf Multi-modal Inter faces 3–8
35. Pinhanez CS, Bobick AF (1998) Human action detection using pnf propagation of temporal constraints. Proc. IEEE Comput Soc Conf Comput Vis Pattern Recognit 898–904
36. Popoola Oluwatoyin P and Wang KJ (2012) Video-Based Abnormal Human Behavior recognition - a review. IEEE Trans. Cybernet 865–878
37. Ryoo MS, Aggarwal JK (2009) Semantic representation and recognition of continued and recursive human activities. Int J Comput Vis 82:1–24
38. Shechtman E, Irani M (2005) Space-time behavior based correlation. CVPR 1:405–412
39. Shi Y, Huang Y, Minnen D, Bobick A, Essa I (2004) Propagation networks for recognition of partially ordered sequential action. CVPR 2:862–869
40. Stauffer C, Grimson WEL (1999) Adaptive background mixture models for real-time tracking. CVPR
41. The BEHAVE dataset. http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/
42. The CAVIAR dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
43. Tran D, Alexander S (2008) Human activity recognition with metric learning. ECCV 548–561
44. Tran D, Sorokin A (2008) Human activity recognition with metric learning. ECCV 548–561
45. Vishwakarma S, Sapre A, Agrawal A (2011) Action recognition using cuboids of interest points. IEEE Int Conf Signal Process Commun Comput (ICSPCC) 1–6
46. Wang H, Ullah MM, Kläser A, Laptev I, Schmid C (2009) Evaluation of local spatiotemporal features for action recognition. BMVC 127–140
47. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden Markov model. CVPR 379–385
48. Yang ZJ, Zhang T, Yang J, Wu Q, Bai L, Yao LX (2013) violence detection based on histogram of optical flow orientation, in Proc. SPIE 9067, Sixth Int Conf Mach Vision 1–4
49. Yu E, Aggarwal JK (2006) Detection of fence climbing from monocular video. 18th Int Conf Pattern Recognit 1:375–378

50. Zhang J, Chen CH (2007) Moving object detection and segmentation in dynamic video backgrounds. IEEE Conf Technol Homeland Security 64–69
51. Zhang D, Gatica-Perez D, Bengio S, McCowan I (2006) Modeling individual and group actions in meetings with layered HMMs. IEEE Trans Multimed 8(3):509–520

Tao Zhang received his Bachelor degree from Henan Polytechnic University, China, in 2008. He is currently a PhD candidate at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. His major research interests include visual surveillance, object detection, and pattern analysis.

Zhijie Yang received his Bachelor and Master degrees from Shanghai Jiao Tong University, China, in 2011and 2014, respectively. His research interests include video surveillance, action recognition and behavior analysis.

Wenjing Jia is currently a Lecturer at the School of Computing and Communications, UTS. She received her PhD degree in Computing Sciences from University of Technology, Sydney (UTS) in 2007. Her research interests are mainly in the areas of image processing/analysis,computer vision and pattern recognition, particularly car license plate detection and text detection.

Baoqing Yang received his B.S., M.S., degrees in the School of Internet of Things Engineering, Jiangnan University, in 2005 and 2008, respectively. He is currently a PhD candidate at Department of Automation, Shanghai Jiao Tong University, China. His major research interests include visual surveillance, face recognition, and pattern analysis.

Jie Yang received his PhD from the Department of Computer Science, Hamburg University,Germany, in 1994. Currently, he is a professor at the Institute of Image Processing and Pattern recognition, Shanghai Jiao Tong University,China. He has led many research projects (e.g., National Science Foundation, 863 National High Tech. Plan), had one book published in Germany, and authored more than 200 journal papers. His major research interests are object detection and recognition, data fusion and data mining, and medical image processing.

Xiangjian He received the Bachelor of Science degree in Mathematics from Xiamen University in 1982, the Master of Science degree in Applied Mathematics from Fuzhou University in 1986, the Master of Science degree in Information Technology from the Flinders University of South Australia in 1995, and the PhD degree in Computing Sciences from the University of Technology, Sydney, Australia in 1999. From 1982 to 1985, he was with Fuzhou University. From 1991 to 1996, he was with the University of New England. Since 1999, he has been with the University of Technology, Sydney, Australia. He is a full professor and the Director of Computer Vision and Recognition Laboratory. He is also a Deputy Director of the Research Centre for Innovation in IT Services and Applications at the University of Technology, Sydney.