Pattern Analysis and Applications (2020) 23:611–623
https://doi.org/10.1007/s10044-019-00821-3

THEORETICAL ADVANCES

Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm

I. P. Febin[1] · K. Jayasree[1] · Preetha Theresa Joy[2]

Received: 16 June 2017 / Accepted: 22 April 2019 / Published online: 4 May 2019

Abstract
Action recognition is an active research area in computer vision as it has enormous applications in today's world, out of which, recognizing violent action is of great importance since it is closely related to our safety and security. An intelligent surveillance system is the idea of automatically recognizing suspicious activities in surveillance videos and thereby supporting security personals to take up right action on the right time. Under this area, most of the researchers were focused on people detection and tracking, loitering, etc., whereas detecting violent actions or fights is comparatively a less studied area. Previous works considered the local spatiotemporal feature extractors; however, it accompanies the overhead of complex optical flow estimation. Even though the temporal derivative is a fast alternative to optical flow, it alone gives very low accuracy and scales-dependent result. Hence, here we propose a cascaded method of violence detection based on motion boundary SIFT (MoBSIFT) and movement filtering. In this method, the surveillance videos are checked through a movement filtering algorithm based on temporal derivative and avoid most of the nonviolent actions from going through feature extraction. Only the filtered frames may allow going through feature extraction. In addition to scale-invariant feature transform (SIFT) and histogram of optical flow feature, motion boundary histogram is also extracted and combined to form MoBSIFT descriptor. The experimental results show that the proposed MoBSIFT outperforms the existing methods in accuracy by its high tolerance to camera movements. Time complexity has also proved to be reduced by the use of movement filtering along with MoBSIFT.

Keywords Violence detection · Abnormal activity detection · Action recognition · Video content analysis · Video event detection

1 Introduction

Video content analysis (VCA) is the process of analyzing videos to detect spatiotemporal events present in it. Nowadays, it has been used in a wide range of application domains like healthcare, entertainment and security. Different functionalities that come under VCA are video motion detection, video tracking, human action recognition, behavior analysis, etc. Human action is not merely the pattern of motion of various body parts but is the real-world depiction of the person's intentions and thoughts. Hence, action recognition has high importance in designing many intelligent systems.

With the Internet revolution and high use of surveillance cameras, today we have access to huge amount of videos. Surveillance systems are very common in today's society but most of the existing systems rely on human observers for detecting activities from these videos. Human capability to monitor simultaneous events is very limited which usually leads to serious losses. Hence, automated video surveillance or intelligent surveillance system has become an important idea to focus on.

It is difficult to define violence as it is a complex action. Most of the previous action recognition works were focused on detecting simple actions like clapping, walking or jogging. Fillipe et al. [1] first introduced the concept of violence detection using spatiotemporal features for making public spaces safer and also for unwanted content filtering. He simplified the concept of violence by labeling scenes containing fights (aggressive human actions) as violence. In

✉ I. P. Febin
febimolu@gmail.com

[1] Department of Computer Engineering, Govt. Model Engineering College Thrikkakara, Ernakulam, Kerala 682021, India

[2] Department of Computer Science and Engineering, College of Engineering Cherthala, Alappuzha, Kerala 688541, India

real-life videos, common violent behaviors can be considered as aggressive fights between two or more people where we cannot predict the pattern of motion. Most fights will be vigorous wrestling fights where it is not possible to detect individual poses. Continuous kicking, punching or hitting with an object is very common in fights. In general, violent action recognition [1–3] can be considered as recognition of an aggressive actions by using an algorithm that analyzes the video sequence to learn about the actions and uses the learned knowledge to identify similar violent actions. Automatic fight or aggressive behavior detection capability can be extremely useful in some video surveillance scenarios like psychiatric centers, elderly centers, prisons, ATM, parking areas, elevators, etc.

The videos obtained from surveillance cameras are of low resolution, and the objects in the outdoor surveillance are often detected in the far field and the front view would not be available always. As a result, action detection methods that depend on person's appearance will not give good results. According to the survey [4], global features are sensitive to noise, occlusion and variation of viewpoint; hence, a promising result can be obtained only when we use local spatiotemporal features. Along with this, we take two more ideas into consideration: (1) Movements will be fast in a violent action; hence, it can be considered as a primary clue in distinguishing violence from other activities; (2) in surveillance scenarios, violence is a rare action compared to other common activities like walking, handshaking or running. Hence, in this paper, we introduce a new method based on local spatiotemporal feature MoBSIFT. In this method, the MoSIFT (Motion SIFT) descriptor [5] is improved both in accuracy and complexity by adding the motion boundary histogram (MBH) [6] and movement filtering algorithm. As motion information is considered as an important cue in action recognition, camera motion is a big challenge to every system. MBH is considered as a good feature to avoid the effect of camera motion. Movement filtering is proposed to reduce the complexity by bypassing most of the nonviolent videos from complex feature extraction.

2 Related work

Earlier works defined violence as explosions and blood flow; hence, most of the previous works were based on audio cues like screams, gunshots [7–9] or relied on color to detect cues such as blood or flame. Nam et al. [10] used flame, blood, sound and the degree of motion as the key features for detecting violence. Cheng et al. [11] used Gaussian mixture models and hidden Markov models (HMM) to recognize gunshots, explosions and car breaking in audio. Giannakopoulos et al. [12] used only audiovisual features to classify violence in movies. Chen et al. [13] used the face, blood and motion information to determine whether the action scene has violent content. Clarin et al. [14] presented a system based on Kohonen self-organizing map to detect violent actions involving blood. Zajdel et al. [15] introduced the Cassandra system to detect aggression in surveillance videos which has used motion features related to articulation in video and scream-like cues in audio.

Later researchers focused on detecting fights between people by identifying key actions like kicking and punching. In [16], Datta et al. proposed an in-depth hierarchical approach for detecting distinct violent events involving two people, namely: fist fighting, hitting with objects, kicking, among others. They have computed information (acceleration measure vector and its jerk) regarding the motion trajectory of image structures. However, this method presents some limitations; for example, it fails when the fighters fall down, or when it involves more than two people. In [17], Yun et al. proposed an interaction detection method to detect kicking, pushing, hugging, etc., and used body pose estimation method. From each frame of the observed video stream, the pose of a human body is recovered using a variety of image features, and action recognition is performed based on such pose estimates. Gao et al. in [18] introduced the use of a dictionary-based sparse feature representation for action recognition. However, it works in situations where multiple views of action being detected are available.

Fillipe et al. [1] presented a violence detector based on local spatiotemporal features with bag of visual words (BoVW). It compared the usage of STIP (space-time interest point) and SIFT (scale-invariant feature transform) along with BoVW and proved STIP as a better method. Later, Bermejo et al. [3] compared the STIP with the usage of MoSIFT (Motion SIFT) [5] and proved its high accuracy and adaptability to different datasets. In [19], authors improved the existing MoSIFT with a sparse coding technique. Many other descriptors also got introduced in the field of violence detection such as violent flows (ViF) descriptor based on the optical flow magnitude to detect violent crowd behavior [20], histogram of oriented tracklets (HOT) [21], histogram of optical flow orientation and magnitude (HOFM) [22], oriented ViF (OviF) [23], oriented histogram of optical flow (OHOF) [24], improved Weber local descriptor (IWLD) [25], Lagrangian local feature with bag of word [26] and magnitude and orientation of local interest frame (DiMOLIF) [27]. However, the high computational cost of extracting such features leads into the proposal of several fast violent action recognition systems.

In 2014, Deniz et al. [2] proposed a method based on extreme acceleration patterns. This method uses only motion information to classify video and avoids the appearance details. Later, Serrano et al. [28] also proposed a violence detection algorithm which was mainly based on temporal derivative and blob feature detection. Compared to the fast

methods, MoSIFT descriptor provides the scale-, rotation- and other deformation-invariant feature extraction. Hence, improving the MoSIFT for high accuracy and low complexity will provide a better alternative to violence detection.

Most of the works in violence recognition were on artificially created datasets. Currently, there is a lack of realistic data in action recognition. The three popular datasets, which are currently used by most of the approaches, are KTH [29], Weizmann [30] and IXMAS [31]. They all contain around 6–11 actions performed by various actors. They are all not very realistic and share strong simplifying assumptions, such as static background, no occlusions, given temporal segmentation and only a single actor. Hence, good results obtained on these datasets do not promise its accuracy in real-life actions. Working with true surveillance footage, sports recordings, movies and video data from the Internet can help shift focus to the important open issues mentioned above. In order to overcome these issues, we have used 'Hockey' and 'Movies' datasets, introduced by Bermejo et al. [3].

3 Proposed method

This paper proposes a cascaded method of feature extraction. The main architecture of the system is provided in Fig. 1. Consecutive frames always bring redundancy of data as it occurs in a very short period of time for any change to take place. Hence, it initially treats the video clip using a frame skipping algorithm to extract only few frames and thereby reduces the complexity of the entire system. After the frame skipping, it converts the frames to gray scale for further processing. Movement filtering is applied on each frame to find whether it has enough motion. Videos with weak motions will easily get rejected as nonviolent videos by using movement filtering, and only the filtered frames with enough motion will go to feature extraction.

3.1 Frame skipping

The popular violence detection methods proposed till now have the complexity issue due to the optical flow estimation between all frames. However, finding optical flow between all consecutive frames does not add much to the accuracy as the data present in these frames are always redundant. Hockey fight dataset which we process here has 40 frames per second (fps), which means actions and movements happening in one second are represented by 40 consecutive pictures. As a result, these videos include pictures or frames in every 0.025 s which is a very short duration for any meaningful action to take place. Hence, complexity of the entire system can be reduced by using a proper frame skipping method.

While computing motion between the frames, the frame reduction method should be bound to time, and it should be comparable in all the datasets. We can reduce the number of frames in second such that it holds information in every 0.1 s that will reduce time complexity to a high extent without compromising accuracy. Here, frame skipping is done by selecting a step size dynamically for each video such that it reduces the frame rate of the video.

If 'N' is the frame rate of a video, then step size (n) is defined as

$$n = \mathrm{abs}(N \times 1/h) \tag{1}$$

where 'h' can be any arbitrary number less than N according to the number of frames needed in one second. In case of Hockey fight dataset, $N=40$ which can be changed to 10 (i.e., $h=10$, $n=4$) such that each frame holds information in 0.1 s.

3.2 Movement filtering

Movement filtering is an initial preprocessing that can be done to evaluate the amount of motions present in a video. This method is only based on temporal differencing to reduce the complexity and to get an instant result. After frame skipping and gray-level conversion, we process the frames to find out whether it contains enough motion using movement filtering. Violent videos always have the high movements of body parts, and other main activities that come under this category are dancing and sports. However, all other actions with very less body movements can be easily identified with the help of this simple method, and hence, it can be used along with a good violence classifiers to alleviate the complexity of the entire system. Movement filtering alone is not capable of categorizing actions but it can make use as a filtering to eliminate those surveillance video frames which do not have enough motion.

Adjacent frames are always similar except for the points where motion occurs, and in order to get an estimate of motion, temporal derivative is the fastest method. According to the studies conducted on different violent and nonviolent videos, violent videos show a pattern on temporal difference binary image with big blobs at the area of motion compared to scattered small blobs on other nonviolent actions. Hence, in order to eliminate these scattered small blobs, we have used a morphological opening on binary image. Opening smoothens the inside of the contour, breaks narrow strips and eliminates thin portions of the binary image. It is done by applying erosion followed by dilation on image B. The resultant images have shown blobs only in the areas where high amount of motion is present like in violent videos, and hence, the estimate of these blobs gives an indication of the movement present in video.

The different steps in movement filtering are as follows:

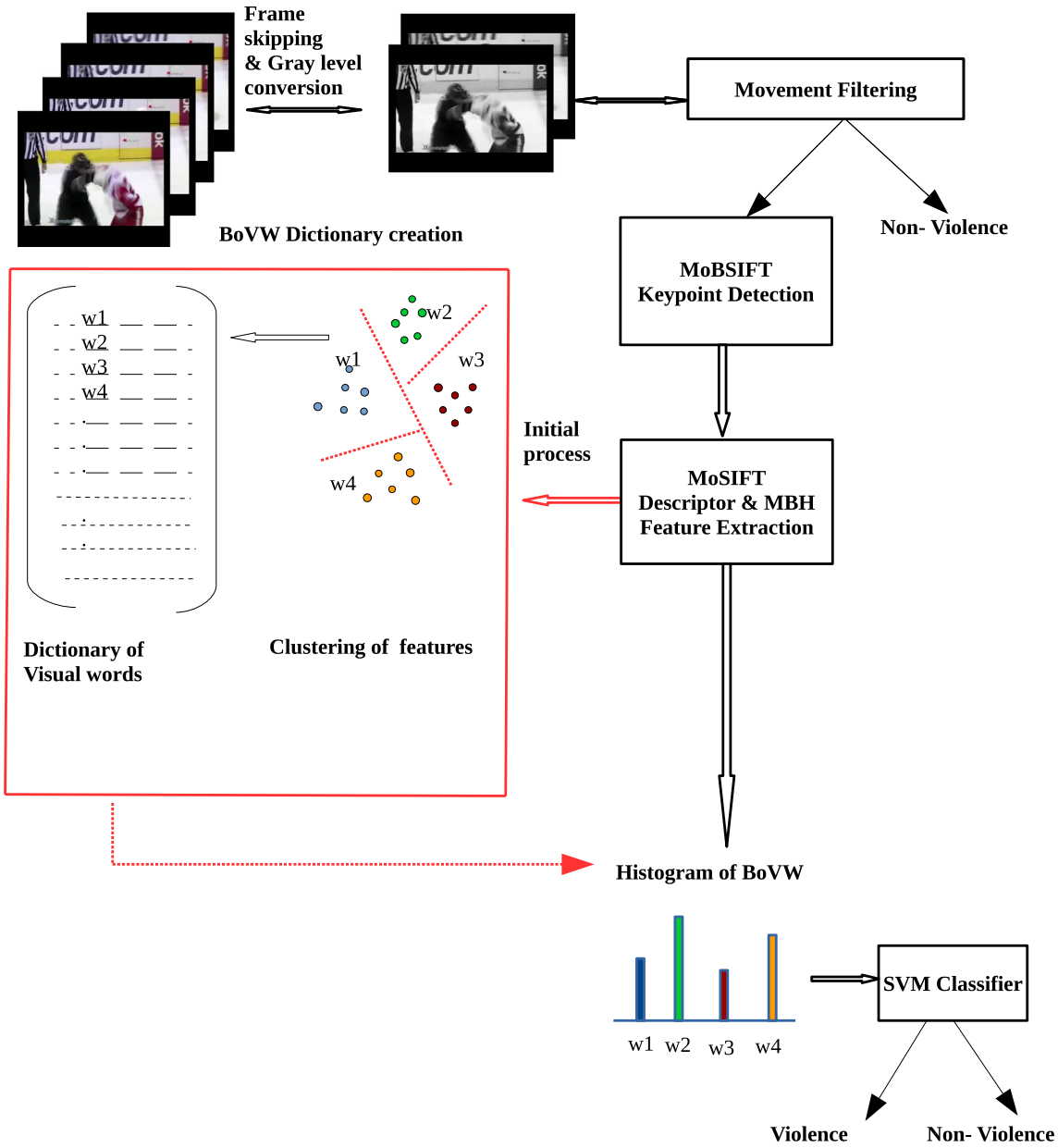

Fig. 1 Main architecture of the system

If I_t (x,y) represents a pixel of the frame at time 't' and I_{t-1} (x,y) represents the corresponding pixel in frame at time '$t-1$,' then the absolute difference of all such pixels gives us D, temporal difference image which is obtained by applying Eq. (2) on all pixels in the frames.

$$D(x,y) = |I_t(x,y) - I_{t-1}(x,y)| \tag{2}$$

Figure 2 shows the temporal derivative of frames.

Thresholding done on difference image D converts it into a binary image B. Thresholding is done by comparing the intensity values in each (x,y) location of the frame with a preset threshold value 'th' as in Eq. 3.

$$B(x,y) = \begin{cases} 1, & \text{If } D(x,y) > \text{th} \\ 0, & \text{otherwise} \end{cases}$$

$$0 < \text{th} < 255 \tag{3}$$

The threshold 'th' has been selected based on experimental analysis. Many violent and nonviolent videos are tested with different threshold values, and in order to avoid the situation of eliminating any short-duration violences, the threshold has set low in our experiment. If we are

Fig. 2 Frames of Hockey fight [3] (top) and temporal derivative of frames (bottom)

focusing only on long-duration violences, the threshold 'th' can be set moderately high to select only those areas with high amount of motion.

Morphological opening on binary image $B(x,y)$ helps to avoid irrelevant motion areas. Structuring element (S) can be selected by experimenting with different shapes. The opening can be mathematically expressed as below:

$$B \circ S = (B \ominus S) \oplus S \tag{4}$$

Figure 3 shows the area of high motion in white color.

Fig. 3 Frames after thresholding (top) frames after opening (Bottom)

The Hockey fight dataset is a sports dataset, and it includes more motion even in the nonviolent videos; hence, movement filtering does not play an important role here. However, Movies dataset includes different nonviolent actions like running, jumping, walking, etc.; hence, movement filtering algorithm improves its accuracy and complexity. In case of real surveillance videos, movement filtering will be beneficial to filter out irrelevant frames in a fastest way.

3.3 MoBSIFT (motion boundary SIFT)

This paper proposes an improved feature extraction based on MoBSIFT which can be considered as a combination of MoSIFT (motion SIFT) and MBH (motion boundary histogram). MoBSIFT algorithm includes two major steps: interest point detection and feature description. Detecting interest points converts the video into few interest points, and feature description is done locally around these interest points.

3.3.1 MoBSIFT interest point detection

MoBSIFT interest point detection is similar to the working of MoSIFT [5] detector. It acts as a temporal extension of the popular scale-invariant feature transform (SIFT) [32] algorithm. SIFT developed by Lowe is the most popular method for finding interest points (keypoints) and feature descriptors. In SIFT, keypoints are the spatial interest points identified by constructing difference of Gaussian (DoG) pyramids and then finding local extremes of the DoG images across adjacent scales.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{5}$$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{6}$$

where $L(x, y, \sigma)$ is the scale space of an input image $I(x, y)$ obtained by convolving it with variable-scale Gaussian, $G(x, y, \sigma)$. $D(x, y, \sigma)$ is considered as DoG of the input image. After constructing DoG pyramid, each pixel will be examined to detect the scale space extrema. Each pixel gets compared with eight neighbors in the same scale and nine neighbors in adjacent scales as shown in Fig. 4.

Keypoints detected by SIFT and dense optical flow [33] estimated on each frame are shown in Fig. 5. In MoBSIFT, the keypoints identified by SIFT method are further processed by finding derivative of optical flow and checking whether these points have sufficient variation in motion. Only spatial interest points with a considerable amount of motion variation will be selected as MoBSIFT interest points; hence, it can be considered as a spatiotemporal interest point detector. Using this method, interest points are selected only in motion boundaries and rest of the points get rejected.

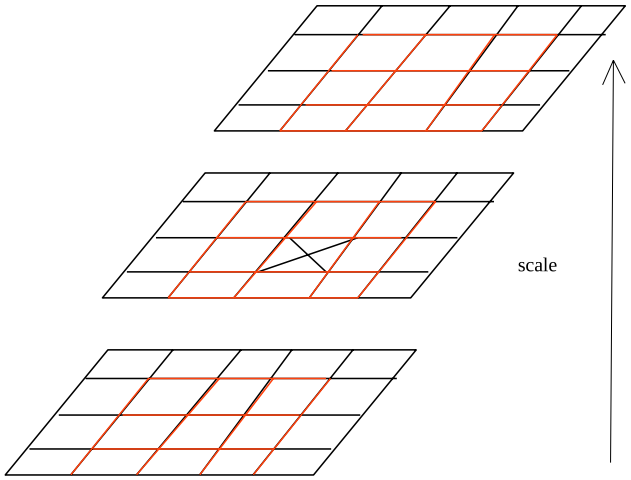

Fig. 4 For finding local extrema, the crossed pixel in the middle octave is compared with its 26 neighbors in adjacent DOG octaves, which includes the eight neighbors at the local scale and the nine neighbors at adjacent scales (up and down)

3.3.2 MoBSIFT feature description

Other than interest point detection, MoBSIFT also includes a feature description method. The standard SIFT extracts histograms of oriented gradients from the 16×16 neighborhood around each interest point. MoBSIFT extracts histogram of optical flow (HoF) feature and motion boundary histogram feature (MBH) along with the SIFT descriptor and combines HoG, HoF, MBHx and MBHy into one vector, which is known as 'early fusion.' It can be done as SIFT and HoF feature extraction and by fusing it with MBH feature (MBHx and MBHy) extraction to form a 320-dimensional feature vector.

3.3.2.1 SIFT and HOF feature extraction SIFT descriptor shows better tolerance to partial occlusion and deformation. For obtaining rotation invariance in SIFT, a dominant orientation is calculated after detecting interest points, and all gradients in the neighborhood are rotated according to the dominant orientation to achieve rotation invariance. Gradient magnitude and direction are calculated for every pixel in a region around the interest point to get the SIFT descriptor. Pixels in the neighboring region are fixed as 256 (16×16) elements. Elements are grouped as 16 (4×4) grids around the interest point. Each grid is described with its own orientation histogram. Orientation histograms are of eight bins, such that each bin covers 45 degrees. All histogram bins get weighted by its gradient magnitude and its distance from the interest point. This leads to a SIFT feature vector with 128 dimensions (4×4×8=128). Each vector is normalized to improve the tolerance to changes in illumination. Figure 6 illustrates the SIFT descriptor grid aggregation idea.

Fig. 5 Optical flow obtained on frames (top) SIFT Keypoints identified (Bottom)

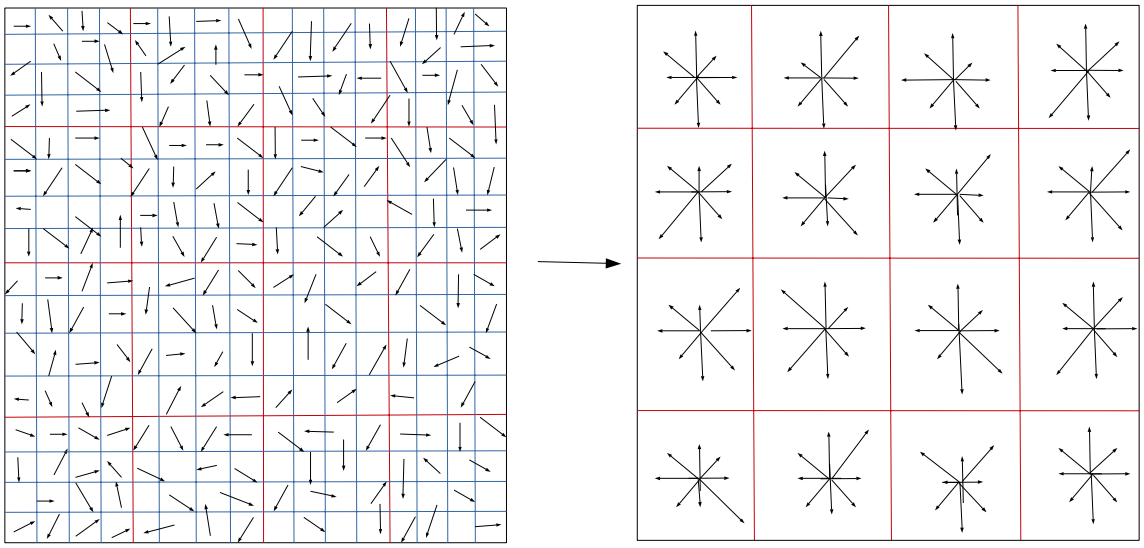

Fig. 6 16×16 region around the keypoint is divided into 4×4 blocks such that each block gives an orientation histogram of eight bins weighted by its magnitude. Both SIFT descriptor and HOF are of 4×4×8 (128) dimensions which result in a 256-dimensional MoSIFT descriptor

The same aggregation mentioned above can be applied to the optical flow in the neighborhood of interest points to form eight-bin histograms of optical flow also. In order to reduce the complexity, here we follow a dense optical flow estimation for entire frame. As the original MoSIFT algorithm, optical flow estimation based on DoG (difference of Gaussian) pyramid is eliminated here to improve working time of the algorithm. Finally, MoSIFT descriptor of 256 dimensions is formed by combining the histograms of appearance and optical flow.

3.3.2.2 MBH feature extraction The motion boundary histograms (MBH) measure the relative motion between pixels instead of absolute motion between frames as optical flow. Dalal et al. [6] proposed the MBH for the more realistic measure of motion by avoiding camera movements. Usually, camera motion will be smooth across the frames and it will get counted if we are measuring optical flow and this will reduce the accuracy of our action classifier. However, MBH represents the gradient of the optical flow or change of the optical flow and hence locally constant

camera motion is removed and only variations (i.e., motion boundaries) get stored [34]. Optical flow is considered as a combination of flow in x- and y-directions, which can be represented as $F(u,v)$. MBH extracts both horizontal and vertical components of flow and finds the gradient of flow separately. Orientation information of these spatial derivatives is quantized into histograms of eight bins weighted by its magnitude. MBHx and MBHy are considered as two separate features. These features are extracted locally from the neighborhood of MoSIFT keypoints. In order to reduce complexity, the region is divided into 2×2 and it results in a descriptor of 32 dimensions ($2\times2\times8$). Figure 7 shows the MBHx and MBHy in comparison with corresponding optical flow. It proves that the motion boundary histogram reduces irrelevant motions in both violent and nonviolent videos. Compared to Hockey fight videos, Movie dataset shows more positive response by eliminating most of the background motions. Same optical flow estimation will be used to extract both HoF and MBH estimation; hence, it does not add computational complexity.

3.4 Bag of visual words (BoVW)

Bag of visual words (bag of visual features) is a concept borrowed from the field of textual information retrieval, which has been successfully applied to a large range of image processing applications. In this approach, the feature domain is sliced into discriminative subspaces. Bag of visual words (BoVW) is the way of constructing a feature vector based on the number of occurrences of word for classification. Each visual word is just a feature vector of patch. It uses k-means algorithm to quantize feature vectors. BoVW representation translates a (usually very large) set of high-dimensional local image descriptors into a single fixed dimensionality vector across all images. By encoding only the occurrence of the appearance of the local patches, not their relative geometry, we get significant flexibility to viewpoint and pose changes.

BoVW method includes two steps: dictionary creation and histogram of BoVW formation based on the dictionary. For creating the dictionary, we extract features from several input videos and cluster them using popular K-means clustering algorithm. Later by applying distance measures, finds similar features to form individual clusters. Cluster heads of these different clusters will be considered as separate

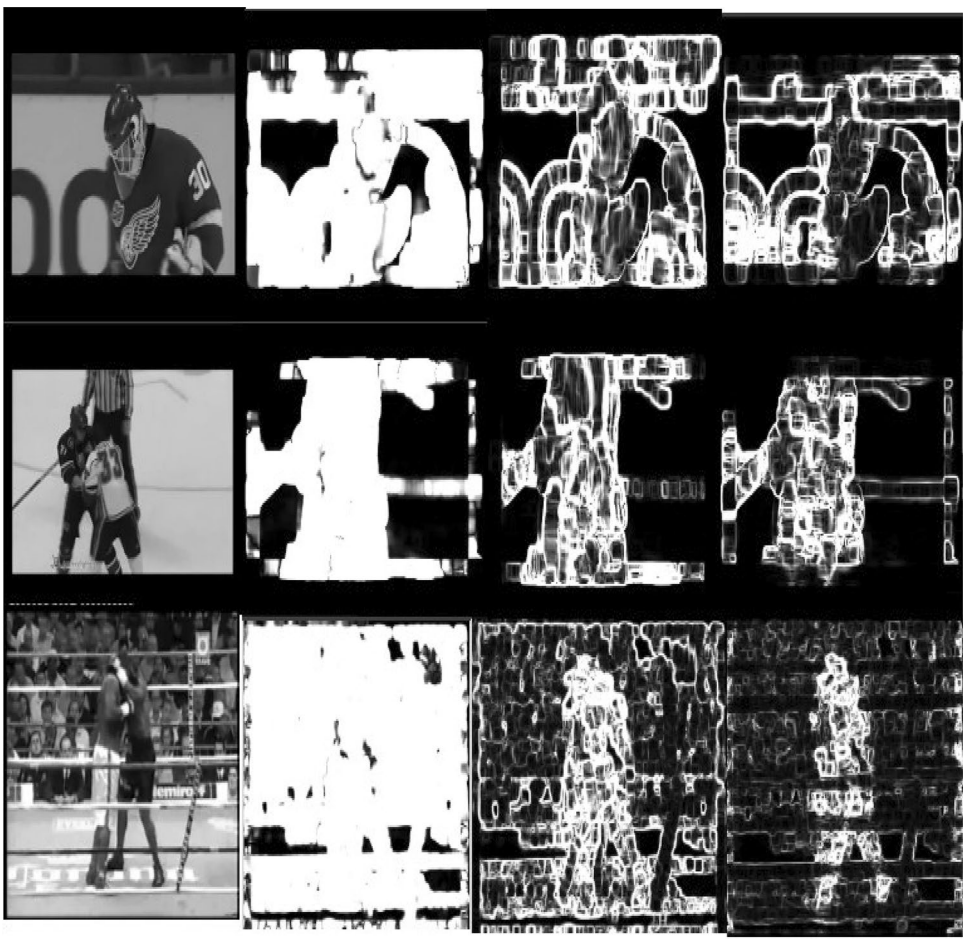

Fig. 7 Frame, optical flow, MBHx and MBHy are shown horizontally for three different videos (Hockey nonviolent, violent and Movie violent videos)

dictionary entries. After the dictionary creation, all the features extracted from input videos get mapped to any of the matching cluster centers and will be finally represented as a frequency of occurrence of these cluster centers.

4 Experimental results

The datasets used in this experiment are Hockey dataset and Movies dataset [3]. Hockey dataset includes 1000 video clips collected from National Hockey League ((NHL). This dataset is equally divided and labeled into two groups: 500 fights and 500 non-fights. Each video clip contains 50 frames with a resolution of 360×288 pixels. The second dataset 'Movies' includes 200 videos, divided into 100 fights and 100 non-fights. Fight video clips of this are extracted from action movies, and non-fight videos are from other action recognition datasets. Unlike Hockey dataset, Movies dataset includes video clips with varying resolution and illumination.

MoBSIFT uses the MBH along with SIFT and HoF features to eliminate camera movements present in videos. Experimental results obtained while comparing the individual features like SIFT, HoF, MBH and other combinations SIFT + HoF (MoSIFT), SIFT + MBH, SIFT + MBH + HoF (MoBSIFT) have proved the significance of MBH feature in violence recognition. Figure 8 shows the performance of different features on Movies dataset. SIFT feature has shown highest accuracy than using HoF or MBH individually. This points to the fact that optical flow alone does not contain enough discriminative power for correctly classifying violence. Even though HoF feature has shown slightly higher accuracy than MBH when used along with SIFT, the better accuracy has been provided by SIFT + MBH than SIFT + HoF. It shows the high discriminative power MBH holds when combined with the spatial feature. Figure 9 shows the performance of different features on Hockey dataset. When comparing Figs. 8 and 9, we can conclude that the MBH feature plays a significant role in Movies dataset than Hockey dataset.

Accuracy of the proposed method has been evaluated using five runs of tenfold cross-validation, and dictionary size used here is 1000. All popular violence detection methods are using SVM for classification; however, some comparative studies in [35] and [36] indicated the higher performance of random forest classifier. Hence, we are comparing SVM, random forest, and AdaBoost classifiers. For Hockey dataset and Movies dataset, we have used random 800 and 120 videos, respectively, to create the dictionary.

The proposed method MoBSIFT (motion boundary SIFT) is an extension of popular SIFT descriptor MoSIFT, and hence, it gives a better scale- and rotation-invariant solution than other fast methods proposed. In Hockey dataset, MoBSIFT with random forest classifier outperformed all the existing methods. Table 1 shows the accuracy obtained while using different popular algorithms on Hockey dataset, and Fig. 10 shows the corresponding ROC curve for the same. MoBSIFT with MF has been also proved to have comparable accuracy as the popular MoSIFT with an advantage

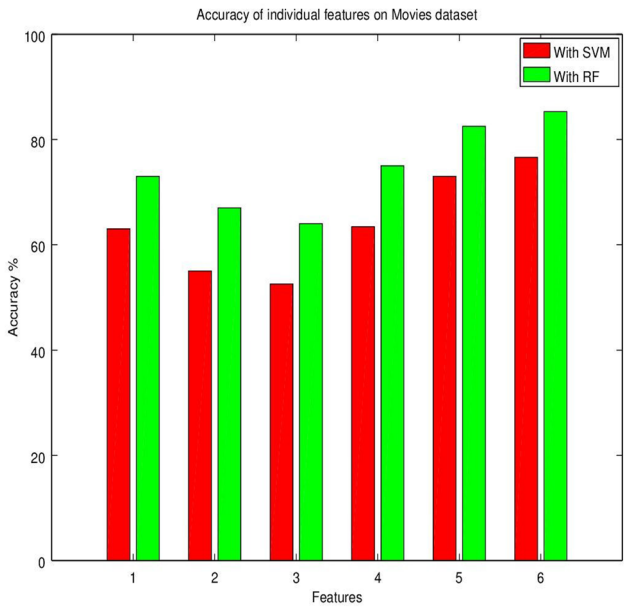

Fig. 8 Accuracy obtained by using individual features and combinations on Movies dataset 1. SIFT 2. HoF 3.MBH 4. SIFT + HoF (MoSIFT) 5. SIFT + MBH 6. SIFT + HoF + MBH (MoBSIFT)

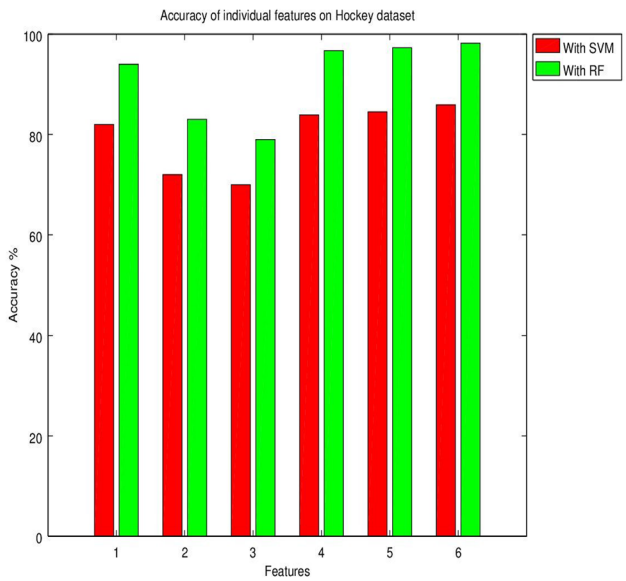

Fig. 9 Accuracy obtained by using individual features and combinations on Hockey dataset 1. SIFT 2. HoF 3.MBH 4. SIFT + HoF (MoSIFT) 5. SIFT + MBH 6. SIFT + HoF + MBH (MoBSIFT)

Table 1 Results obtained on Hockey dataset

Method	Classifier	Hockey
BoW (MoSIFT)	SVM	83.9 ± 0.6
	Random forest	96.7 ± 0.7
	AdaBoost	89.5 ± 0.40
Deniz et al. [2]	SVM	90.1 ± 0
	Random forest	61.5 ± 6.8
	AdaBoost	90.1 ± 0
Serrano et al. [28]	SVM	72.5 ± 0.5
	Random forest	82.4 ± 0.6
	AdaBoost	71.7 ± 0.3
BoW (MoBSIFT)	SVM	86.5 ± 0.6
	Random forest	98.2 ± 0.5
	AdaBoost	92.6 ± 0.4
BoW (MoBSIFT) + MF	SVM	85.0 ± 0.3
	Random Forest	96.5 ± 0.8
	AdaBoost	90.3 ± 0.3

of reduction in time complexity. According to the studies, the method proposed by Deniz et al. [2] has shown slightly higher performance when used with SVM classifier but with the Random Forest and AdaBoost classifiers MoBSIFT as well as MoSIFT methods exceeded in accuracy. This throws light into the fact that the scale-invariant features like MoSIFT, MoBSIFT, etc work well with tree-based ensemble classifiers.

In Movies dataset, MoBSIFT with MF outperformed the other popular methods by providing an accuracy of 98.9%. Deniz et al. [2] and Serrano et al. [28] methods also provided comparable accuracy in this dataset. However, these methods are proposed as the fast methods for violence detection; hence, they only consider motion feature. In Movies dataset, distinction between violence and nonviolence is very clear based on motion clue as most of the nonviolent actions in this are the slow actions; hence, methods based only on motion clue can perform well here. However, in more general way compared to other fast methods, the proposed method has the advantage of detecting objects involved in motion by extracting shape feature and consequently it has the capability to distinguish fights between people from other accelerated motions. Table 2 shows the accuracy provided by various popular methods on Movies dataset, and ROC curve obtained is given in Fig. 11.

The time complexity of MoSIFT is comparatively very high which has been reduced in the proposed MoBSIFT by changing the optical flow estimation to dense optical flow estimation for entire frame once and eliminating the DoG (difference of Gaussian) pyramid-based flow estimation. As we see in Table 1, these changes in estimation have not shown much influence on the accuracy of the system. Use of movement filtering has allowed the system to process nonviolent videos with very less complexity, which ensures that most of the nonviolent videos without sudden motions will be processed in a time 0.032 s/frame which is almost equal to the working of all fast methods proposed. The comparison of time complexity is given in Table 3.

Table 4 shows the result obtained after applying Movement Filtering on both the datasets and it had proved that the working of MF in both the dataset is different. However, the percentage of false negative (rejecting violence as nonviolence) is very less in both the datasets. In Hockey datasets around half of the nonviolent videos were easily filtered out

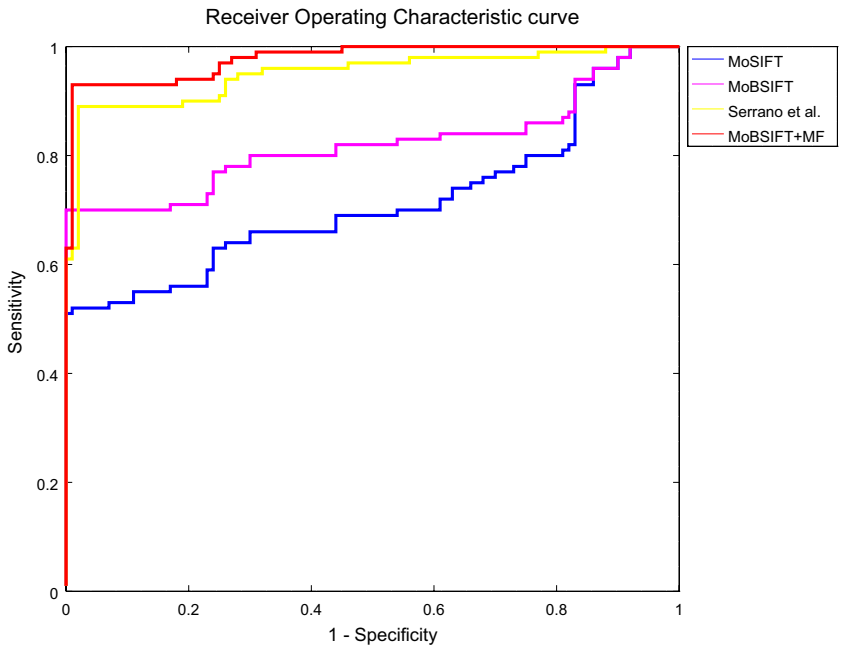

Fig. 10 ROCs on Movies dataset: ROC curves for the related methods with random forest classifier

Table 2 Results obtained on Movies dataset

Method	Classifier	Movies
BoW (MoSIFT)	SVM	63.4 ± 1.6
	Random forest	75.1 ± 1.6
	AdaBoost	86.5 ± 1.58
Deniz et al. [2]	SVM	85.4 ± 9.3
	Random forest	90.4 ± 3.1
	AdaBoost	98.9 ± 0.22
Serrano et al. [28]	SVM	87.9 ± 1
	Random forest	97.8 ± 0.4
	AdaBoost	81.8 ± 0.5
BoW (MoBSIFT)	SVM	76.6 ± 0.3
	Random forest	85.3 ± 0.3
	AdaBoost	88.2 ± 0.6
BoW (MoBSIFT) + MF	SVM	89.3 ± 1.5
	Random forest	98.9 ± 1.3
	AdaBoost	98.9 ± 0.10

Table 3 Time taken to process frames

Method	Time taken (sec/frame)	
	Violent	Nonviolent
BoW (STIP)	0.293	0.293
BoW (MoSIFT)	0.661	0.661
Deniz et al. [2]	0.0419	0.0419
Serrano et al. [28]	0.0225	0.0225
BoW (MoBSIFT)	0.257	0.257
BoW (MoBSIFT) + MF	0.257	0.032 (avg. case)
		0.257 (worst case)

Table 4 Performance result of movement filtering (MF) on two datasets

	Non-fight	Fight
Hockey dataset	247/500 (50.6%)	451/500 (90.2%)
Movies dataset	77/100 (77%)	91/100 (91%)

and only the rest processed with MoBSIFT. However, in Movies, 77% of nonviolent videos were filtered out by using movement filtering and only 23% passed to complex MoBSIFT feature extraction.

5 Conclusion

In this paper, we introduced an improved scale- and rotation-invariant detection method, called MoBSIFT, for violence detection which has shown higher performance both in accuracy and complexity. Experiments show that in Hockey dataset, this method has outperformed most of the existing methods. This paper also introduced a new movement filtering method to identify nonviolence easily and with very less complexity. MoBSIFT with movement filtering has shown highest accuracy on Movies dataset. Experiments done on different individual features also proved the significant role MBH played in both the datasets. However, further research can be conducted to use a feature selection method to reduce the complexity further without compromising accuracy.

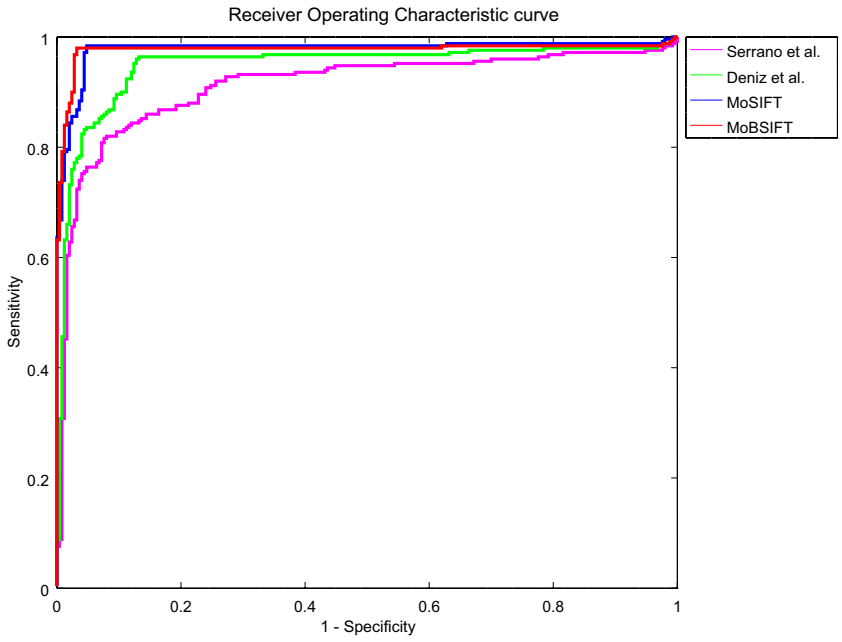

Fig. 11 ROCs on Hockey dataset: ROC curves for the related methods with random forest classifier

Acknowledgements I extent my gratitude toward Govt. Model Engineering College for providing all support for this work. I also appreciate the support provided by Bermejo et al. [3] by making Movies and Hockey dataset freely available to access.

References

1. de Souza FD, Chavez GC, do Valle EA, de A Araujo A (2010) Violence detection in video using spatio-temporal features. In: 23rd SIBGRAPI conference on graphics, patterns and images, pp 224–230
2. Deniz O, Serrano I, Bueno G, Tae-Tyun K (2014) Fast violence detection in video. In: VISAPP 2014 proceedings of the 9th international conference on computer vision theory and applications, pp 478–485
3. Bermejo E, Deni O, Bueno G, Sukthankar R. (2011) Violence detection in video using computer vision techniques. In: Proceedings of the 14th international conference on computer analysis of images and patterns. Springer, pp 332–339
4. Ke S-R, Thuc H, Lee Y-J et al (2013) A review on video-based human activity recognition. Computers 2:88–131. https://doi.org/10.3390/computers2020088
5. Chen M, Hauptmann A (2009) MoSIFT: recognizing human actions in surveillance videos. Technical report, Carnegie Mellon University, Pittsburgh, USA
6. Dalal N, Triggs B, Schmid C (2006) Human Detection using oriented histograms of flow and appearance. In: Proceedings of 9th ECCV, pp 428–441
7. Giannakopoulos T, Kosmopoulos D, Aristidou A, Theodoridis S (2006) Violence content classification using audio features. In: Proceedings of the 4th helenic conference on advances in artificial intelligence. Springer, pp 502–507
8. Gong Y, Wang W, Jiang S, Huang Q, Gao W (2008) Detecting violent scenes in movies by auditory and visual cues. In: Proceedings of the 9th Pacific Rim conference on multimedia. Springer, Berlin, Heidelberg, pp 317–326
9. Lin J, Wang W (2009) Weakly-supervised violence detection in movies with audio and video based cotraining. In: Proceedings of the 10th Pacific Rim conference on multimedia. Springer, Berlin, Heidelberg, pp 930–935
10. Nam J, Alghoniemy M, Tewfik AH (1998) Audio-visual content-based violent scene characterization. In: Proceedings 1998 international conference on image processing. ICIP98 (Cat. No. 98CB36269). IEEE Comput. Soc, Chicago, USA, pp 353–357
11. Cheng W, Chu W, Ling J (2003) Semantic context detection based on hierarchical audio models. In: Proceedings of the ACM SIGMM workshop on multimedia information retrieval, pp. 109–115
12. Giannakopoulos T, Makris A, Kosmopoulos D, Perantonis S, Theodoridis S (2010) Audio-visual fusion for detecting violent scenes in videos. In: Artificial intelligence: theories, models and applications, pp 91–100
13. Chen L-H, Hsu H-W, Wang L-Y, Su C-W (2011) Violence detection in movies. In: 2011 Eighth international conference computer graphics, imaging and visualization. IEEE Comput. Soc, Washington, DC, USA, pp 119–124
14. Clarin C, Dionisio J, Echavez M, Naval P (2005) DOVE: Detection of movie violence using motion intensity analysis on skin and blood. Technical report, University of the Philippines
15. Zajdel W, Krijnders JD, Andringa T, Gavrila DM (2007) CASSANDRA: audio-video sensor fusion for aggression detection. In: 2007 IEEE conference on advanced video and signal based surveillance, pp 200–205
16. Datta A, Shah M, Da Vitoria Lobo N (2002) Person-on-person violence detection in video data. In: 16th international conference on pattern recognition, pp 433–438
17. Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 28–35
18. Gao Z, Nie W, Liu A, Zhang H (2016) Evaluation of local spatial–temporal features for cross-view action recognition. Neurocomputing 173:110–117
19. Xu L, Gong C, Yang J, Wu Q, Yao L (2014) Violent video detection based on MoSIFT feature and sparse coding. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3538–3542
20. Hassner T, Itcher Y, Kliper-Gross O (2012) Violent flows: real-time detection of violent crowd behavior. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, Providence, USA, pp 1–6
21. Mousavi H, Mohammadi S, Perina A, Chellali R, Murino V (2015) Analyzing tracklets for the detection of abnormal crowd behavior. In: IEEE winter conference on applications of computer vision, pp 148–15
22. Colque RVHM, Junior CAC, Schwartz WR (2015) Histograms of optical flow orientation and magnitude to detect anomalous events in videos. In: 28th SIBGRAPI conference on graphics, patterns and images, pp 126–133
23. Gao Y, Liu H, Sun X, Wang C, Liu Y (2016) Violence detection using oriented violent flows. Image Vis Comput 48–49:37–41
24. Zhang T, Yang Z, Jia W, Yang B, Yang J, He X (2016) A new method for violence detection in surveillance scenes. Multimed Tools Appl 75:7327–7349
25. Zhang T, Jia W, He X, Yang J (2017) Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans Circuits Syst Video Technol 27(3):696–709
26. Senst T, Eiselein V, Kuhn A, Sikora T (2017) Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation. IEEE Trans Inf Forensics Secur 12(12):2945–2956
27. Mabrouk AB, Zagrouba E (2017) Spatio-temporal feature using optical flow based distribution for violence detection. Pattern Recognit Lett 92:62–67
28. Gracia IS, Suarez OD, Garcia GB, Kim T-K (2015) Fast fight detection. PLoS ONE 10(4):e0120448. https://doi.org/10.1371/journal
29. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: 17th international conference on pattern recognition (ICPR'04), IEEE Comp. Soc. Washington, DC, USA, vol 3, pp 32–36
30. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
31. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 104:249–257
32. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
33. Paul M, Haque SME, Chakraborty S (2013) Human detection in surveillance videos and its applications a review. EURASIP J Adv Signal Process 2013:176
34. Wang H, Klaser A, Schmid C, Liu C-L (2011) Action recognition by dense trajectories. In: 2011 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Colorado Springs, USA, pp 3169–3176
35. Liu M, Wang M, Wang J, Li D (2013) Comparison of random forest, support vector machine and back propagation neural network

for electronic tongue data classification: application to the recognition of orange beverage and Chinese vinegar. Sens Actuators B 177:970–980

36. Lorena AC, Jacintho Luis FO, Siqueira MF, De Giovanni R, Lohmann LG, de André CPLF, Carvalho MY (2011) Comparing machine learning classifiers in potential distribution modelling. Expert Syst Appl 38:5268–5275