Noname manuscript No.
(will be inserted by the editor)

MoWLD: A Robust Motion Image Descriptor for Violence Detection

Tao Zhang · Wenjing Jia · Baoqing Yang · Jie Yang · Xiangjian He · Zhonglong Zheng

Received: date / Accepted: date

Abstract Automatic violence detection from video is a hot topic for many video surveillance applications. However, there has been little success in designing an algorithm that can detect violence in surveillance videos with high performance. Existing methods typically apply the Bag-of-Words (BoW) model on local spatiotemporal descriptors. However, traditional spatiotemporal features are not discriminative enough, and also the BoW model roughly assigns each feature vector to only one visual word and therefore ignores the spatial relationships among the features. To tackle these problems, in this paper we propose a novel Motion Weber Local Descriptor (MoWLD) in the spirit of the well-known WLD and make it a powerful and robust descriptor for motion images. We extend the WLD spatial descriptions by adding a temporal component to the appearance descriptor, which implicitly captures local motion information as well as low-level image appear information. To eliminate redundant and irrelevant features, the non-parametric Kernel Density Estimation (KDE) is employed on the MoWLD descriptor. In order to obtain more discriminative features, we adopt the sparse coding and max pooling scheme to further process the selected MoWLDs. Experimental results on three benchmark datasets have demonstrated the superiority of the proposed approach over the state-of-the-arts.

Keywords Violence detection, surveillance systems, Motion Weber Local Descriptors (MoWLD), Kernel Density Estimation (KDE), sparse coding, max pooling

1 Introduction

Violent behavior seriously endangers social and personal security [20]. Analysis of crowd behavior is an area of increasing interest within the safety. Currently, millions of video surveillance equipment have been used in places such as streets, prisons and supermarkets (some sample frames from public datasets are shown in Fig. 1. If a surveillance system can detect these violent activities automatically and alarm correspondingly, it will greatly improve security. Therefore, it is highly necessary to investigate the problem of automatically identifying violent contents from

T. Zhang · B. Q. Yang · J. Yang (✉)
Institute of Image Processing and Pattern Recognition
Shanghai Jiaotong University
E-mail: zhb827@sjtu.edu.cn

B. Q. Yang
E-mail: yang_baoqing@sjtu.edu.cn

J. Yang
E-mail: jieyang@sjtu.edu.cn

W. J. Jia · X. J. He
Faculty of Engineering and Information Technology
University of Technology Sydney, Australia
E-mail: Wenjing.Jia@uts.edu.au E-mail: Xiangjian.He@uts.edu.au
Z. L. Zheng
Zhejiang Normal University
E-mail: zhonglong@zjnu.edu.cn

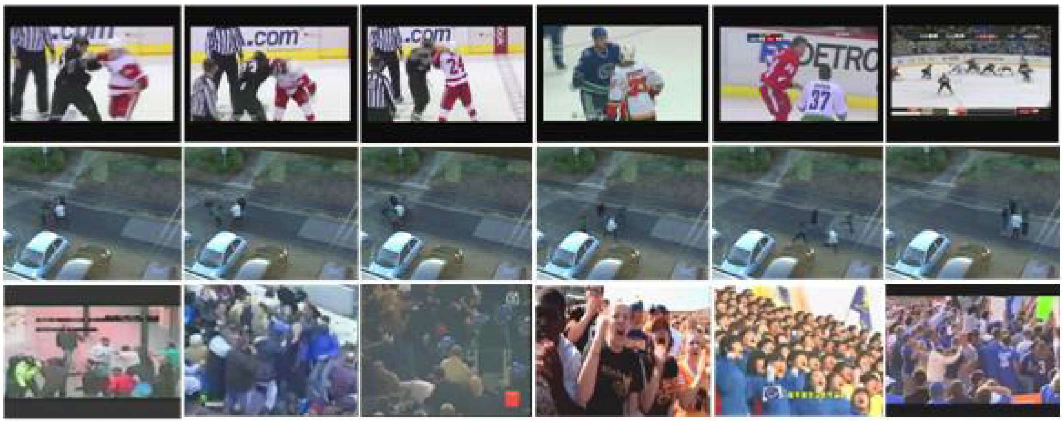

Fig. 1 Sample frames from the Hockey Fight dataset (first row), the BEHAVE dataset (second row) and the Crowd Violence dataset (third row). In each row, the left three columns are violent scene while the right three columns are non-violent scene.

surveillance video. Computer vision techniques are highly demanded for intelligent surveillance and automatic video annotation. However, complex background, variable illuminating conditions and different distances between the subjects and the camera have made this task very challenging. Compared with other related issues of action recognition, violence detection is less studied by now. The variations of body motion caused by scale, viewpoint, occlusion, and the clutter background have made violence detection very difficult. For this practical consideration, in this paper, we focus on the challenging work of detecting violence in surveillance videos and aim to develop a system to effectively detect violent behaviors using computer vision techniques.

Up to now, there have been some developmental systems about violence detection. In early research, Nam et al. [24] proposed to recognize violent scenes in videos by detecting flame and blood and capturing the degree of motion, as well as the characteristic sounds of violent events. Cheng et al. [10] recognized gunshots, explosions and car-braking using a hierarchical approach. [28] presented a novel technique to characterize and index violent scenes in general TV drama and movies. Unlike simple low-level video visual feature analysis, this kind of method allows searching and access to specific violent scenes. [12] presented a new method that was able to integrate audio and visual information for scene analysis in a typical surveillance scenario, using only one camera and one monaural microphone. This method allows one to detect separate audio and visual patterns representing unusual unimodal events in a scene. [23] presented a weakly-supervised method to detect violent shots in movies. This detection process is split into two aspects, i.e. audio and video. However, the above approaches require audio information, which is often not available in many surveillance scenes. [16] relied on motion trajectory information and orientation information of a person's limbs to detect violent behavior. This method requires foreground segmentation to extract the precise silhouettes, which is difficult in a real life environment. Clarin et al. [11] 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 method relies on the skin color information, which performance will degrade greatly when the color feature is not discriminating enough. In recent studies, some methods based on spatiotemporal interest points (STIPs [35], MoSIFT [4]) 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 compute only in the regions of interest (located around the 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 features. To solve the above problems, Zhou et al. [45] proposed a structured codebook construction method to encode spatial and temporal contextual information among local features for video representation. The method better suits for structured videos, rather than the more textural videos in our data set. Hassner et al. [36] detected crowd violence using the Violent Flow (ViF) descriptor formed from computing a magnitude-change map of optical flow over time. However, the performance of this method degrades significantly when dealing with faces with crowded scenes. Zhang et al. [44] proposed a fast and robust framework for detecting and localizing violence in surveillance scenes, and experimental results on several benchmark datasets have demonstrated the superiority of this method over the state-of-the-arts in terms of both detection accuracy and processing speed, even in crowded scenes. Ye et al. [22] proposed a physical bullying detection algorithm based on activity recognition. The algorithm is designed for smartphones requiring a 3D accelerometer and a 3D gyroscope to collect data and therefore is not suitable for general scenario.

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

- First, to detect sufficient number of interest points containing the necessary information to recognize a violent activity, we propose a novel image descriptor, i.e. Motion Weber Local Descriptor (MoWLD), to extract the low-level image and motion properties of a query video. The MoWLD algorithm detects spatially distinctive interest points with substantial motions. In a sense, this descriptor takes the advantages of both SIFT in terms of computing the histogram using the magnitude and orientation of gradient, and LBP in terms of computational efficiency.
- Secondly, to eliminate redundant and irrelevant features, Kernel Density Estimation (KDE) is employed on the MoWLD descriptor. This not only avoids unnecessary computation and speeds up the system but also contributes to a high detection rate.
- Lastly, sparse coding is adopted to transform the low-level descriptors into compact mid-level features. To obtain a highly discriminative representation of the extracted feature, the max pooling algorithm is employed over the whole sparse code set of the query video.

Experimental results on three challenging datasets have demonstrated the superiority of our proposed approach over the state-of-the-arts.

The remaining of this paper is organized as follows. Related work is discussed in Section 2. Section 3 introduces a novel violent detection method. In Section 4, experimental results and analysis are presented. Finally, conclusions are given in Section 5.

2 Related Work

Action recognition is a hot topic in computer vision research. We refer readers to a recent survey [33] and focus our discussion to single-camera methods. Most existing works in action recognition have used spatiotemporal features, trajectories, and set of features [1]. Sequential approaches represent human activities with a sequence of actions and recognize activities by analyzing a set of features extracted from the input video [23]. Zhang et al. [43] used two-layer Hidden Markov Models (HMMs) to recognize group actions, where one layer models the basic individual activities from audio-visual features, and the other models the interactions between the individual activities. However, the impact of the layered decomposition on the size of the parameter space was not given. Also, the effects of the inference on learning requirements and accuracy for different amounts of training were ambiguous. Nguyen et al. [29] presented an application of the Hierarchical Hidden Markov Model (HHMM) for activity recognition. Their main contributions lie in the application of the shared-structure HHMM and the estimation of the model's parameters at all levels simultaneously. However, it failed to recognize complex behaviors. Shi et al. [34] presented Propagation Networks (P-Nets) for representing and recognizing sequential activities that include parallel streams of action. Their work was focused on a common task for elderly people who have developed late stage diabetes. However, the performance strongly relied on the manually labeled training data. Dai et al. [13] introduced a novel event-based dynamic context model, where a multilevel dynamic Bayesian network (DBN) model was used to detect multilevel events. However, the applicable scenario was limited. Damen and Hogg [14] proposed to construct Bayesian networks using AND-OR grammars to encode pairwise event constraints. However, it failed to recognize complex and ambiguous events. Bobick and Davis [5] used two components, i.e. MEI and MHI, to represent and recognize human activities. It first constructed a vector image, which was 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. In [33], the optical flow based approach was used to represent apparent velocities of movement of brightness patterns in an image, which has been employed for modeling typical motion patterns [9,37]. However, this measure may also become invalid in extremely crowded scenes. A dense local sampling of optical flow has been proposed to solve this issue [26].

Baysal and Duygulu [3] utilized a line based pose representation to recognize human actions in videos. However, they used line-flow histograms, which can be easily effected by the performance of segmentation. Manifold learning is another efficient approach for recognizing human actions. Saghafi and Rajan [32] proposed a novel embedding which is optimum in the sequence recognition framework based on Spatiotemporal Correlation Distance (SCD) as the distance measure. However, its performance mainly relies on the key poses chosen equidistantly from one action period and works not very well in complex environment. Oikonomopoulous et al. [30] 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 [38] 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 works well in semantically varying events and is robust to scale and view changes. Yang et al. [42] proposed to use a scheme of multi-feature learning via hierarchical regression for multimedia semantics understanding. The algorithm can be applied to a wide range of multimedia applications and the performance of the proposed algorithm is remarkable when only a small amount of labeled training data are available. Gao et al. [18] proposed a semi-supervised annotation approach by learning an optimal graph (OGL) from multi-cues (i.e., partial tags and multiple features) which can more accurately embed the relationships among the data points.

Recently, more and more research attention is given to anomaly detection in video [31], i.e. detecting irregular patterns that are different from regular video events. Despite there are many existing works on video anomaly detection [26,31], few of them can work well in crowded scenes. Vijay Mahadevan et al. [24] used the mixture of dynamic textures (MDT) model to detect both temporal and spatial abnormality. Marco Bertini et al. [26] constructed a multi-scale local descriptor for anomaly detection and achieved real-time performance in video surveillance applications. Mehrsan Javan et al. [27] densely sampled the spatiotemporal information in videos for learning dominant and anomalous behaviors online. Xu et al. [15] presented a novel unsupervised learning approach for video anomaly detection based on deep representations. The proposed method is based on multiple stacked autoencoder networks for learning both appearance and motion representations of scene activities.

Following the above works, in this paper, to detect violent video we focus on interest point detection and feature representation. For interest point detection, a widely-adopted one is the scale invariant feature transform (SIFT), introduced by Lowe [17]. Many attempts to improve the SIFT descriptor have been reported [17,41]. However, the feature descriptor is an important step which is almost ignored. Chen et al. [9] proposed the motion SIFT (MoSIFT) to detect interest points, which not only encodes their local appearance but also explicitly models local motion. This MoSIFT descriptor consists of two main parts. The first part is an aggregated histogram of gradients (HoG) to describe the spatial appearance. The second part is an aggregated histogram of optical flow (HoF) to indicate the movement of the feature point. In the aspect of action recognition, they have also demonstrated the superiority of their MoSIFT over four different descriptors, i.e. 3D HoG, HoF, HoG and HoF, and grid aggregated HoG and HoF. However, as mentioned above, SIFT is a sparse descriptor, because it only considers the regions of interest.

Another simple, yet very powerful and robust local descriptor is Weber Local Descriptor (WLD), first proposed by Chen et al. in [8] for texture classification and face detection. Wang et al. [39] further exploited the illumination insensitive characteristics of the WLD and used it for face recognition. Li et al. [21] proposed multi-scale WLD and multi-level information fusion approaches for face recognition. It states that the change of a stimulus (such as sound, lighting) will be just noticeable when the change is smaller than the constant ratio of the original stimulus [8], i.e. the proportion of the change to the original stimulus value is a constant.

The highly successful illumination-invariant WLD for object recognition detects many interest points in an image and the descriptors of these points are used to match static objects. Since recognizing violent activities is more complicated than face recognition, violence detection requires enhanced local features that can provide sufficient motion information because only WLD interest points with sufficient motion provide the necessary information for action recognition. The widely used optical flow approach detects the movement of a region by calculating where a region moves by measuring temporal differences. In this paper a novel descriptor, i.e. Motion WLD (MoWLD), is proposed to represent the feature point. Our proposed MoWLD is composed

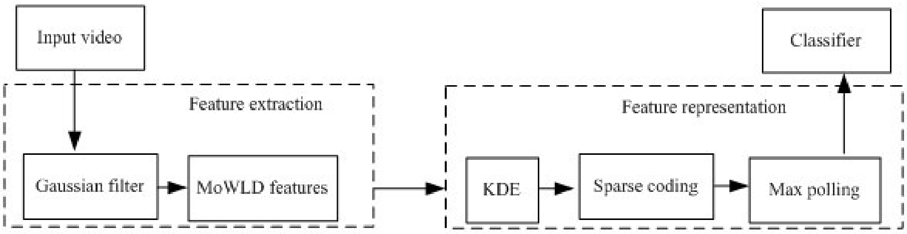

Fig. 2 The framework of the proposed method.

of two parts of information. The first part is an aggregated histogram of WLD describing the spatial appearance, and the second part is an aggregated histogram of optical flow (HoF) indicating the movement of feature points. Our MoWLD can therefore detect spatially distinctive interest points with substantial motions.

Approaches based on local spatiotemporal descriptors are traditionally combined with Bag-of-Words (BoWs) model and have achieved promising performance in violence detection [4,35]. However, the conventional BoW methods rely on the discriminative power of local spatiotemporal descriptors and how often they occur in the video. Moreover, the performance of BoW model can be degraded significantly due to high quantization error. Currently, methods based on sparse coding have been successfully utilized in action and image classification field [40,41,46]. The sparse coding method transforms each low-level descriptor to a linear combination of a few atoms in a well-trained dictionary. Compared with the BoW model, it generates less reconstruction errors and can achieve a more discriminative feature representation.

In this work, we combine our proposed MoWLD descriptor with the sparse coding method in order to generate a more discriminative representation of violent video. The framework of our approach is illustrated in Fig. 2. Firstly, we extract MoWLD features from the input video. Secondly, we employ the Kernel Density Estimation (KDE) based feature selection method to eliminate redundant and irrelevant features from the original MoWLD descriptor. Subsequently, sparse coding is adopted to transform the reduced low-level descriptors into mid-level features. To obtain a highly discriminative representation of the extracted feature, the max pooling algorithm is employed over the whole sparse code set of the query video. Finally, an SVM classifier is trained using these video level feature vectors.

3 Our Approach

Intuitively, effective features can reveal distinct visual patterns of query video. In this section, we propose a more discriminative and robust violence detection algorithm. Firstly, a MoWLD algorithm is proposed to extract low-level features of a query video. Then, to eliminate redundant and irrelevant features, KDE is applied. Lastly, to obtain a highly discriminative representation of the extracted features, sparse coding and max pooling are introduced.

3.1 MoWLD algorithm

This subsection presents our MoWLD algorithm to detect and describe interest points spatiotemporally. We aim to develop an effective feature representation method, which can detect a sufficient number of interest points containing the necessary information to recognize violent behavior.

3.1.1 The Original WLD

Weber's law describes a fact, that for a stimulus, the ratio between the smallest perceptual change and the background is a constant, which implies stimuli are not perceived in absolute terms but in relative terms.

Inspired from this law, Chen et al. [8] proposed a local image descriptor named Weber Local Descriptor (WLD) for the task of face recognition. Chen's WLD descriptor consists of two components, i.e. differential excitation (magnitude) and orientation, which are defined as below [8,39].

Weber Magnitude:

$$\xi_m(x_c) = \arctan(\alpha \sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c}), \tag{1}$$

where the arctangent function is used to prevent the output from being too large and thus can partially suppress the side-effect of noise, x_c denotes the center pixel, x_i $(i = 0, 1, \ldots, p-1)$ are the neighboring pixels, p is the number of neighbors, and α is a parameter used to adjust the intensity difference between neighboring pixels.

Weber Orientation:

$$\xi_o(x_c) = \arctan(\frac{x_1 - x_5}{x_3 - x_7}), \tag{2}$$

where $x_1 - x_5$ and $x_3 - x_7$ indicate the intensity difference of two neighboring pixels of x_c in vertical and horizontal direction, respectively.

According to [8], ξ_m and ξ_o are then linearly quantized into T (in our experiments, T is set to 12) dominant differential magnitudes and orientations respectively.

Chen's WLD uses the intensity differences between the current pixel and its neighbors as the changes of a current pixel. By this means, we can find the salient variations within an image to simulate the perception pattern of human beings. Both differential excitation and differential orientation have been proved to be illumination insensitive and computationally efficient [39]. The 2D concatenated histogram about the differential excitation and orientation can be constructed to represent the image. As is shown in [8] and [21], each row of the 2D WLD histogram corresponds to a dominant differential excitation $\xi_m(x_c)$, and each column corresponds to a dominant orientation $\xi_o(x_c)$. The original WLD histogram [8,21,39] denotes the frequencies of a certain dominant differential excitation on a certain dominant orientation.

The WLD descriptor employs the advantages of both the SIFT in terms of computing the histogram using the gradient and its orientation, and the LBP in terms of computational efficiency and smaller support regions. Different from the SIFT and LBP, the WLD is a dense descriptor computed for every pixel and depends on both the local intensity variation and the magnitude of the center pixel's intensity. Since the WLD is computed around a relatively small square region (e.g., 3×3), while SIFT is computed around a relatively large region (e.g., 16×16), the description granularity of the WLD is much smaller than that of the SIFT. That is to say, the WLD is computed in a finer granularity than SIFT. The smaller size of the support regions for WLD enables it capture more local salient patterns.

3.1.2 Modified WLD

The original WLD feature described as above is not rotation invariant, and is also sensitive to partial occlusion and deformation. Rotation invariance is important to appearance since it provides a standard to measure the similarity of two key points. Aiming to address the above problems, we propose to rebuild the WLD histogram by aggregating the WLD histograms of neighboring regions and also aligning the WLD histograms to their dominant orientation. In details, these are achieved with the following steps:

1. The Weber magnitude and orientation are calculated according to Eqs. 1 and 2 for every pixel in a region of a Gaussian-blurred image F.
2. The Weber orientation is quantified into 12 dominant bins by using the non-linear quantization method in [21], with each bin covering 30 degrees. An orientation histogram with 12 bins is then formed.
3. An aggregated histogram of Weber gradients from neighboring regions is captured as local appearance feature. This gives our WLD descriptor better tolerance to partial occlusion and deformation.
4. Each sample in the neighboring window is added to a histogram bin and weighted by its Weber magnitude and its distance from the current point. When a dominant orientation is calculated, all Weber magnitudes in the neighborhood are rotated according to the dominant orientation to achieve rotation invariance.

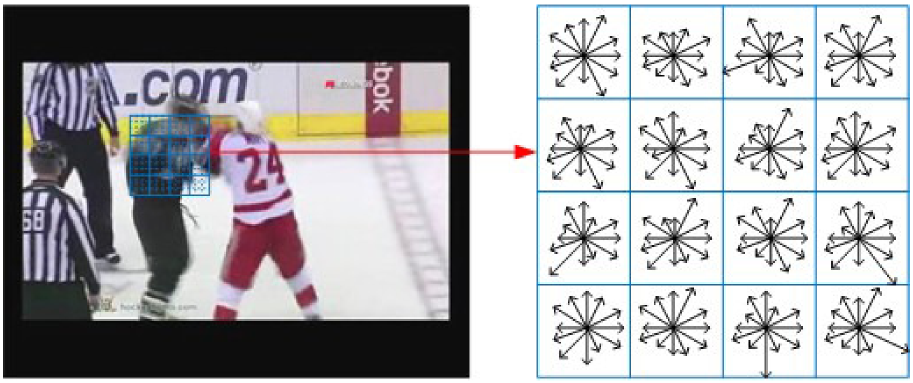

Fig. 3 Grid aggregation for WLD feature descriptors. Pixels in a neighborhood are grouped into 4 × 4 blocks, each containing 3 × 3 = 9 pixels. An orientation histogram with 12 bins is formed for each grid resulting into a 192-element vector for the neighborhood.

5. Pixels in the neighboring region are normalized into 144 (16×9) elements, which are grouped as 16 (4×4) grids. Each grid has its own Weber orientation histogram describing the orientation of the sub-region. This results in a WLD feature vector of 192 dimensions ($4 \times 4 \times 12 = 192$).

Fig. 3 illustrates the idea of the WLD histogram grid aggregation. Pixels in a neighborhood are grouped into 4×4 blocks, each containing 3×3 pixels. By constructing the WLD feature vector in this way, we can obtain a more discriminative descriptor of in total $16 \times 12 = 192$ dimensions.

The modified WLD feature constructed as above only describes the properties of still images and carries no motion information of video. Therefore, the detected candidate points are distinctive in appearance only, but are independent of the motions or actions in video. As its result, a cluttered background can produce many interest points which are unrelated to human actions. Clearly, motion information is essential for interest points to provide sufficient information for action recognition. In our MoWLD algorithm, we adopt the widely used optical flow approach to detect the movement within an image region. A local extreme from WLD feature points can only become an interest point if it has sufficient motion in the optical flow field.

Also, the WLD features described above are extracted from a small patch ('grid') of 3×3 pixels, which implies a single and fixed granularity. However, the sub-images with different sizes at the same location can result in different feature vectors, and multi-scale sub-images can be conducted to extract more discriminative and robust features of different human local structures. Thus, in this work we adopt the multi-scale WLD feature analysis approach [8], which is computed using a square symmetric neighbor set of p pixels placed on a square.

Next, we continue introducing our motion WLD, and then multi-scale MoWLD is introduced.

3.1.3 Motion WLD (MoWLD)

The optical flow approach detects the movement of a region by calculating the temporal differences of the region in image space between consecutive frames. Compared to video cuboids or volumes, optical flow explicitly captures the magnitude and direction of a motion, instead of implicitly modeling motion through appearance change over time. Explicitly measuring motion is beneficial for recognizing actions. In our work, to add motion information into our modified WLD feature, we apply the same aggregation idea to the optical flow of every grid in a region and propose our motion WLD, i.e. MoWLD.

Our MoWLD adopts the idea of grid aggregation in WLD into optical flow to describe motion information among frames. Optical flow detects the magnitude and direction of movement between frames, which produces the same properties as WLD and can be used to construct optical flow histograms. The dominant orientation feature is the main difference between the WLD and optical flow. Since surveillance video is typically captured by stationary cameras, the direction of movement generated by violent actions is typically irregular and variable, which can be used to distinguish them from normal actions. Therefore, for optical flow histograms, we omit the step of adjusting orientation invariance in the MoWLD motion descriptors.

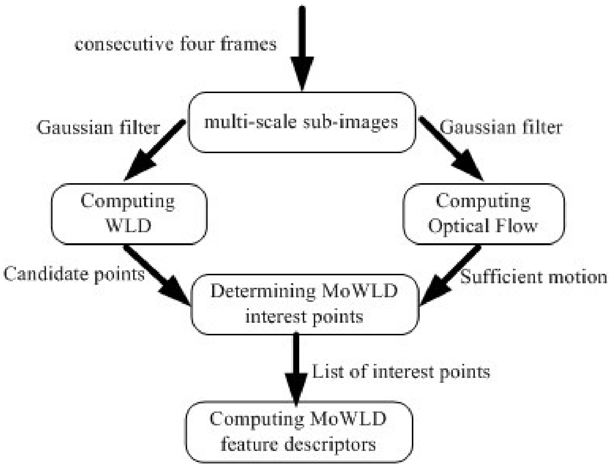

Fig. 4 System flowchart of our MoWLD algorithm. Four consecutive frames are input to compute the WLD and optical flow. Candidate points with sufficient motion are determined as the MoWLD interest points, for which MoWLD features are extracted.

Similarly as in the WLD, the orientations of optical flow in each grid is normalized into 12 directions and an optical flow histogram of 12 bins is constructed for each grid. For a 4×4 grid neighborhood, this results in an aggregated optical flow histogram of a dimension of $4 \times 4 \times 12 = 192$. Also, we consider temporal contextual information for better robustness and add the WLD and optical flow histograms of three previous frames into the descriptor.

Thus, all of the aggregated histograms (WLD and optical flow) are concatenated into the MoWLD descriptor (as shown in Fig. 4), which now has 1536 ($4 \times 2 \times 192 = 1536$) dimensions.

Fig. 4 illustrates our MoWLD algorithm. The algorithm takes consecutive four frames to find spatiotemporal interest points at multiple scales. Two major computations are applied, i.e. WLD feature and optical flow computation, according to the scale of the WLD.

3.1.4 Multi-Scale MoWLD

In our MoWLD algorithm, we adopt the multi-scale WLD feature analysis approach [8] and calculate multi-scale optical flows according to the WLD scales.

Specifically, optical flow pyramids are constructed over two Gaussian pyramids. Multiple-scale optical flows are calculated according to the WLD scales. A local extreme from WLD feature points can only become an interest point if it also has sufficient motion in the optical flow pyramid. We assume that a complicated action can be represented by the combination of a reasonable number of interest points. Therefore, we do not assign strong constraints to spatio-temporal interest points. As long as candidate interest points contain a minimal amount of movement, the algorithm can extract them as MoWLD interest points. The extracted MoWLD interest points are scale and rotation invariant in spatial domain but they are not scale invariant in the temporal domain. The MoWLD can select distinctive interest points with sufficient motion where humans can ‘see’ the action based on these points and machines can learn an action model.

Since our MoWLD is based on the WLD and the optical flow, it is natural that our descriptor leverages the following advantages. Instead of combining a HoF classifier with a HoG classifier, we build a single feature descriptor, which concatenates both HoG and HoF into one vector, which is also called ‘early fusion’. We believe appearance and motion information together are the essential components for classifying actions. Since an action is only represented by a set of spatio-temporal point descriptors, the descriptor features critically determines the information used by later recognition steps. It can also be seen that our MoWLD descriptor captures local appearance with an aggregated histogram of gradients from neighboring regions. This gives our MoWLD descriptor better tolerance to partial occlusion and deformation. Also, when an interest point is detected, a dominant orientation is calculated and all gradients in the neighborhood are rotated according to the dominant orientation. This makes our MoWLD rotation invariant.

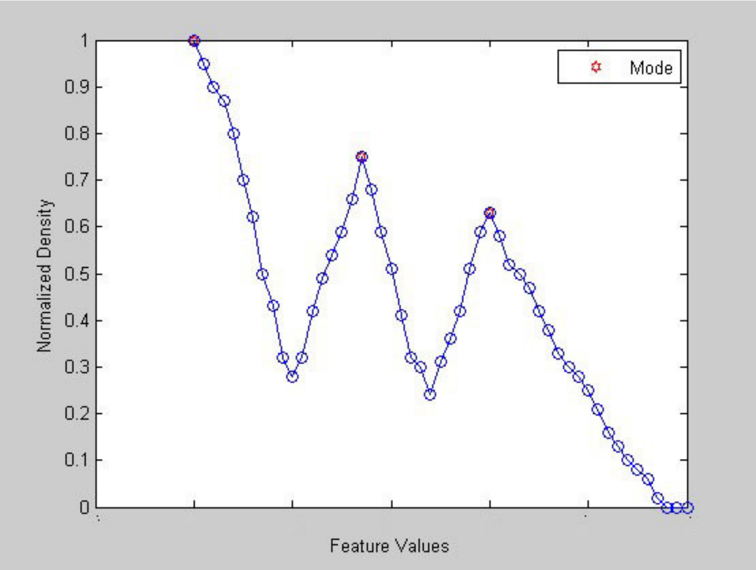

Fig. 5 The normalized probability density function estimated by the KDE method.

3.2 KDE-based Feature Selection

The extracted high-dimensional MoWLD descriptor may contain some irrelevant and redundant information. To improve both performance and computational efficiency, we employ the KDE-based feature selection method [6,19] to select the most representative features from the extracted MoWLD.

Suppose $x_1, x_2, \dots, x_N$ are N independent identically distributed data of a one-dimensional random variable x. KDE infers the probability density function of x by centering a kernel function $K(x)$ at each data point x_i as:

$$f_h(x) = \frac{1}{hN} \sum_{i=1}^{N} K(\frac{x - x_i}{h}), \tag{3}$$

where h is a smoothing parameter, named as bandwidth, which can be adaptively chosen using the method proposed in [19].

In order to reduce the dimension of the MoWLD feature, we use KDE to obtain a smooth probability density function based on our training data. However, the common Gaussian kernel density estimator [6] lacks local adaptivity, and this often results in a higher sensitivity to outliers. So, an adaptive kernel is chosen to be $K(\bullet)$, as discussed in [19], which can be used as a way to improve local adaptivity and reduce bias.

If the probability density function of a feature is bimodal or multimodal, this feature is considered to be more discriminative than those with only a single mode. Fig. 5 shows an example of our probability density function (PDF) with three modes. We estimate the PDF of each feature on the original 1536 features of MoWLD. According to the number of modes, we sort the 1536 MoWLD features in descending order. Finally, the first 850 features are selected to form the reduced MoWLD, which is more effective than the original ones.

3.3 Sparse Coding Scheme

In order to obtain more discriminative features, instead of the BoW model, we adopt sparse coding to further process the selected MoWLD features for violence detection.

Let X be a set of reduced MoWLD feature vectors extracted from a query video clip. $X = [x_1, x_2, \dots, x_N]$ ($x \in \mathbb{R}^{d\times\mathbb{N}}$), where x_i denotes ith vector of the total N data samples. The sparse coding problem can be formulated as:

$$Z = \arg \min_{Z \in \mathbb{R}^{k \times \mathbb{N}}} \frac{1}{2} \|X - DZ\|_{\ell_2}^2 + \lambda \|Z\|_{\ell_1}, \tag{4}$$

where $Z = [Z_1, Z_2, \ldots, Z_N]$ $(Z \in \mathbb{R})$ and Z_i is the corresponding sparse representation of vector y_i, $D = [d_1, d_2, \ldots, d_N]$ $(D \in \mathbb{R})$ is a pre-trained dictionary, which is an overcomplete basis set $(k > d)$, and λ is a positive regularization parameter to control the tradeoff between fitting degree and sparseness (according to [25], it is set to 0.069). The optimization over Z is convex when the dictionary D is constant. To seek a sparse Z, the LARS-lasso approach [41] is employed to solve Eq. 4. In this way, the original low-level descriptors are converted into compact mid-level features (corresponding spare code representation Z). Then, the violence analysis/recognition is carried out on Z domain.

The dictionary D contains atoms representing basic patterns of the specific data distribution in feature space. Given a large collection of the reduced MoWLD features (processed by KDE-based feature selection) extracted from training data $Y = [y_1, y_2, \ldots, y_N]$ $(y \in \mathbb{R}^{d \times \mathbb{M}})$, the dictionary learning problem in sparse coding scheme can be defined by:

$$\arg \min_{U \in \mathbb{R}^{k \times \mathbb{M}}, D \in C} \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|y_i - Du_i\|_{\ell_2}^2 + \lambda \|u_i\|_{\ell_1}, \tag{5}$$

where $U = [u_1, u_2, \ldots, u_N]$ $(U \in \mathbb{R})$ is the coefficients set and C is a convex set, and

$$C \triangleq \{D \in \mathbb{R}^{d \times k}, s.t. \|d_i\|_{\ell_1} \leq 1, i \in \{1, \ldots, k\}\}, \tag{6}$$

where the formulation is not convex with respect to D and U. Online dictionary learning algorithm [25] that has been proven to be more appropriate for large training sets was employed to solve this joint optimization problem.

3.4 Max Pooling Scheme

Pooling is used to achieve invariance to image transformations, more compact representations, and better robustness to noise and clutter. It has been stated that max pooling outperforms the average pooling [7,35]. In order to further capture globally optimized feature, max pooling is applied over the sparse code set Z as:

$$\beta = \digamma(Z), \tag{7}$$

where β is a vector with k dimensions and $\digamma$ is a pooling function defined on each row of $Z \in \mathbb{R}^{k \times \mathbb{N}}$. Different pooling functions construct different video statistics [40,41]. In our experiment, we adopt the max pooling function approach [7], which is defined as:

$$\beta_i = \max | Z_{i1} |, | Z_{i2} |, \ldots, | Z_{iN} |, \tag{8}$$

where β_i denotes the ith element of β, and Z_{ij} denotes the $(ij)th$ entry of the matrix Z.

Compared with the BoW model, the sparse coding method achieves a much lower reconstruction error and captures the salient properties of human actions. Then, with the help of max pooling approach, the irrelevant information is discarded, and only the strongest response to some certain atoms is preserved. It is spatially-temporally pooled, and generates a compact and discriminative video level feature β for our violence detection task. In classification stage, we employ the SVM with a RBF kernel to classify the input video as either violent or non-violent.

4 Experiments and Results

4.1 Dataset

Experiments of our method were conducted on three challenging datasets: the Hockey Fight dataset [36], the BEHAVE dataset [2], and the Crowd Violence dataset [36].

The Hockey Fight dataset contains 1000 video clips of action from hockey games of the National Hockey League (NHL). 500 videos in the dataset are manually labeled as fight and others are labeled as non-fight. Each clip consists of 50 frames (360×288 pixels image resolution).

The BEHAVE dataset contains more than $200,000$ frames (640×480 pixels image resolution) and 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 as violence or non-violence. Each clip consists of at least 100 frames. Finally, we picked 80 clips for violence detection, including 20 violence clips and 60 non-violence clips.

The 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 violent 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 presents violent crowd behavior and the other half presents non-violent crowd behavior.

4.2 Results and Discussion

We compare our proposed method against the state-of-the-art techniques including the BoW based methods, the violence detection method in [44], the Appearance and Motion DeepNet (AMDN) method in [15] and the ViF method in [36]. To evaluate the classification accuracy, we employed the 5-fold cross validation test on each dataset. Results are reported with mean prediction accuracy (ACC)± standard deviation (SD) as well as the area under the ROC curve (AUC). In our experiments, SVM is employed as classifier in all approaches compared and both the MoWLD feature and the final video level feature vector are ℓ_2 normalized. Also, to assess the impact of dictionary size on accuracy, we have run the experiments with dictionaries of different sizes being learned.

Results on Hockey Fight dataset. Table 1 shows the results of various methods on the Hockey Fight dataset. The results on this dataset using the BoW model paired with HOG, HOF and MoSIFT (i.e. “HOG+BoW”, “HOF+BoW” and “MoSIFT+BoW” respectively) are reported in [4]. As it can be seen from the table, MoSIFT and HOG based BoW models perform comparably, with a slight improvement achieved with MoSIFT over HOF. The proposed MoWLD (noted as “MoWLD+BoW” in the table) outperforms all above approaches, indicating our proposed MoWLD descriptor is more discriminative and effective. Also, with the increase of the dictionary size, the performance begins to rise and then stays stable. This phenomenon indicates that selection of an appropriate dictionary size is significant to both high accuracy and computational efficiency.

Table 2 shows the results obtained after adopting the sparse coding and KDE-based feature selection into our MoWLD and BoW approach. It can be seen that using the sparse coding based approach have resulted in higher accuracy than the BoW based approaches alone, due to the less quantization error of sparse coding. The performance is further improved after using the KDE-based feature selection. This is due to the fact that the irrelevant and redundant features of MoWLD are removed while leveraging feature selection, thus contributing to a more discriminative local descriptor. In this experiment, the number of words in the dictionary of BoW equals to the size of sparse dictionary in sparse coding.

Results on the BEHAVE dataset. 20 clips of this dataset are randomly picked for training. In order to demonstrate the superior performance of our algorithm, we compared our approaches with those of the state-of-the-art approaches implemented by us, including HOG, HOF, HNF (combination of HOG and HOF), ViF [36], the robust violence detection (RVD) [44], the

Table 1 Accuracy comparison of BoW-based violence detection methods on the Hockey Fight dataset.

Vocabulary	HOG+BoW [4]	HOF+BoW [4]	MoSIFT+BoW [4]	MoWLD+BoW
50 words	87.8%	83.5%	87.5%	88.1%
100 words	89.1%	84.3%	89.4%	90.4%
150 words	89.7%	85.9%	89.5%	90.7%
200 words	89.4%	87.5%	90.4%	91.3%
300 words	90.8%	87.2%	90.4%	91.3%
500 words	91.4%	87.4%	90.5%	91.5%
1000 words	**91.7%**	**88.6%**	**90.9%**	**91.9%**

Table 2 Detection results on the Hockey Fight dataset using sparse coding with and without using the KDE.

Vocabulary	MoWLD+SparseCoding		MoWLD+KDE+SparseCoding	
	ACC±SD	AUC	ACC±SD	AUC
50 words	$89.1 \pm 1.31\%$	0.9318	$91.4 \pm 1.78\%$	0.9597
100 words	$90.5 \pm 0.88\%$	0.9492	$92.9 \pm 2.18\%$	0.9615
150 words	$92.4 \pm 1.51\%$	0.9618	$93.9 \pm 1.84\%$	0.9695
200 words	$83.1 \pm 1.91\%$	0.9708	$94.7 \pm 1.62\%$	0.9715
300 words	$93.5 \pm 1.51\%$	0.9638	$94.6 \pm 1.71\%$	0.9708
500 words	$93.3 \pm 1.29\%$	0.9706	$\mathbf{94.9 \pm 1.68\%}$	**0.9789**
1000 words	$\mathbf{93.7 \pm 1.68\%}$	**0.9781**	$94.2 \pm 1.91\%$	0.9719

Table 3 Detection results on the BEHAVE dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [36]	$58.69 \pm 0.35\%$	0.6322
HOF+BoW [36]	$59.91 \pm 0.28\%$	0.5893
HNF+BoW [36]	$57.97 \pm 0.31\%$	0.6089
ViF [36]	$82.02 \pm 0.19\%$	0.8592
MoSIFT+BoW [4]	$62.02 \pm 0.23\%$	0.6578
MoWLD+BoW	$83.19 \pm 0.18\%$	0.8517
MoWLD+SparseCoding	$85.75 \pm 0.15\%$	0.8891
RVD [44]	$85.29 \pm 0.16\%$	0.8878
AMDN [15]	$84.22 \pm 0.17\%$	0.8562
MoWLD+KDE+SparseCoding	$\mathbf{87.17 \pm 0.13\%}$	**0.8993**

Appearance and Motion DeepNet (AMDN) [15] and MoSIFT [4]. Table 3 presents the results obtained with the above mentioned methods on the BEHAVE dataset. The dictionary size is fixed to 500 in this set of experiments. HOG, HOF and HNF are spatiotemporal descriptors with BoW model while ViF is a global representation based approach. As it can be seen from the table, our two sparse coding based methods (the bottom two in the table) outperform other approaches. This demonstrates that our MoWLD descriptor is significantly superior in performance to HOG, HOF and HNF. It proves that MoWLD is a more effective descriptor for describing action feature. The performance of the RVD method in [44] is close to ours, that is because this method has adapt a Gaussian Model of Optical Flow (GMOF) to extract candidate violence regions, which reduces many noise disturbances. The AMDN utilizes deep neural networks to automatically learn feature representations. Undeniably, its performance is very stable. However, it uses optical flow as the input image feature, and there exist many redundant and interference features, so its performance is not the best. Also, our MoWLD combined with the sparse coding method outperforms all BoW based methods and employing the KDE-based feature selection further improves the accuracy. Results on this dataset demonstrate that our algorithm is also effective for detecting violence in group fighting scene. False alarm only happened when a group of people get together to do some strenuous non-violence activities.

Results on the Crowd Violence dataset. This dataset is more challenging than the above two datasets because it contains many crowded scenes. The set contains 246 clips divided into five splits, each containing 123 violent and 123 non-violent scenes. In order to demonstrate the superior performance of our proposed approach, we compare our algorithm with those of the state-of-the-art approaches, including HOG, HOF, HNF (combination of HOG and HOF), ViF, and MoSIFT, which are reported in [4] and [36]. We also compared our approaches with those of the state-of-the-art approaches implemented by us, including the robust violence detection

Table 4 Detection results on the Crowd Violence dataset.

Algorithm	ACC±SD	AUC
HOG+BoW [36]	$57.43 \pm 0.37\%$	0.6182
HOF+BoW [36]	$58.53 \pm 0.32\%$	0.5760
HNF+BoW [36]	$56.52 \pm 0.33\%$	0.5994
ViF [36]	$81.30 \pm 0.21\%$	0.8500
MoSIFT+BoW [4]	$57.09 \pm 0.37\%$	0.6073
MoWLD+BoW	$82.56 \pm 0.19\%$	0.8651
MoWLD+SparseCoding	$86.39 \pm 0.15\%$	0.9018
RVD [44]	$82.79 \pm 0.19\%$	0.8659
AMDN [15]	$84.72 \pm 0.17\%$	0.8891
MoWLD+KDE+SparseCoding	$\mathbf{89.78 \pm 0.13\%}$	**0.9472**

(RVD) [44], the Appearance and Motion DeepNet (AMDN) [15]. Table 4 presents the results obtained with various methods on this dataset. Same as before, the dictionary size is fixed to 500 in this set of experiments. In this dataset, due to more crowded scenes, the detection rate of RVD method decreases. On the contrary, the performance of AMDN method is still very stable. However, because of the introduction of optical flow noise, its performance is not very good. Our sparse coding based methods still outperform other approaches. MoWLD descriptor is significantly superior in performance to HOG, HOF, HNF, RVD and AMDN. It proves that our proposed MoWLD is a more effective descriptor for describing action feature. Consistent with the results on the previous two datasets, our MoWLD combined with the sparse coding method outperforms the BoW based methods and employing the KDE-based feature selection has effectively improved the accuracy. Results on this dataset demonstrate that our algorithm is also effective for detecting violence in crowded scene. Some false alarms are caused by people's fast running.

By verifying the obtained results, we can find that our proposed system is effective and robust for correct detection of violence. Our algorithm is able to handle violence detection with complex scenarios, including different camera distance, severe occlusion between people and crowed scenes.

5 Conclusion

Aiming for a robust violence detection method in surveillance scenes, in this paper we proposed a novel violent video detection approach based on the MoWLD feature and sparse coding. Several popular approaches have been employed to generate a highly discriminative video feature: 1) MoWLD employs the advantages of SIFT in computing the histogram using the gradient and its orientation, and those of LBP in computational efficiency; 2) The KDE-based feature selection method eliminates some redundant and irrelevant features of the MoWLD; 3) Integrating the sparse coding method with max pooling generates a discriminative, high-level global video feature. Experimental results on three challenging datasets have demonstrated that the proposed method outperforms the state-of-art techniques for violence detection in both crowded and non-crowded scenes, which has shown the effectiveness of the proposed video representation.

Acknowledgments. This research was partly supported by NSFC, China (No: 61273258, 61375048, 61170109).

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys **43**(3), 1–43 (2011)
2. Andrade, E., Fisher, R.: Modelling crowd scenes for event detection. In: In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), vol. 01, pp. 175–178. IEEE (2006)
3. Baysal, S., Duygulu, P.: A line based pose representation for human action recognition. Signal Processing: Image Communication **28**(5), 458–471 (2013)
4. Bermejo, E., Deniz, O., Bueno, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: In: Proceedings of the 14th international conference on computer analysis of images and patterns, pp. 332–339. Springer (2011)
5. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23**(3), 257–267 (2001)

6. Botev, Z.I., Grotowski, J.F., Kroese, D.P.: Kernel density estimation via diffusion. The Annals of Statistics **38**(5), 2916–2957 (2010)
7. Boureau, Y.L., Ponce, J., Yann, L.: A theoretical analysis of feature pooling in visual recognition. in Proceedings of the 27th International Conference on Machine Learning (ICML-10) **31**(6), 111–118 (2010)
8. Chen, J., Shan, S., He, C., Zhao, G., Chen, X., Gao, W.: Wld: A robust local image descriptor. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9), 1705–1720 (2010)
9. Chen, M., Hauptmann, A.: Mosift: Recognizing human actions in surveillance videos. In: Tech. rep, Carnegie Mellon University, pp. 1–10. Carnegie Mellon University (2009)
10. Cheng, W., Chu, W., Wu, J.: Semantic context detection based on hierarchical audio models. In: Proceedings of the ACM SIGMM workshop on Multimedia information retrieval pp. 109–115 (2003)
11. Clarin, C., Dionisio, J., Echavez, M., Naval, P.: Detection of movie violence using motion intensity analysis on skin and blood. Tech. rep., University of the Philippines (2005)
12. Cristani, M., Bicego, M., Murino, V.: Audio-visual event recognition in surveillance video sequences. In: Multimedia, IEEE Transactions on, pp. 257–267. IEEE (2007)
13. Dai, P., Di, H., Dong, L., Tao, L., Xu, G.: Group interaction analysis in dynamic context. In: Systems, Man, and Cybernetics, IEEE Transactions on, pp. 275–282. IEEE (2008)
14. Damen, D., Hogg, D.: Recognizing linked events: searching the space of feasible explanations. In: Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on, pp. 927–934. IEEE (2009)
15. Dan, X., Elisa, R., Yan, Y., Jingkuan, S., Nicu, S.: Learning deep representations of appearance and motion for anomalous event detection. In: In: The British Machine Vision Conference (BMVC), pp. 1–12. BMVA Press (2015)
16. Datta, A., Shah, M., da Vitoria Lobo, N.: Person-on-person violence detection in video data. Proceedings of IEEE International Conference on Image Processing (ICIP2002) pp. 433–438 (2002)
17. David, G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
18. Gao, L., Song, J., Nie, F., Yan, Y., Sebe, N., Shen, H.T.: Optimal graph leaning with partial tags and multiple features for image and video annotation. IEEE Conference on Computer Vision and Pattern Recognition pp. 4371–4379 (2015)
19. Geng, X., Yu, C., Hu, G.: Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting. Biomedical Signal Processing and Control **7**(2), 112–117 (2012)
20. Huesmann, L., Moise-Titus, J., Podolski, C., Eron, L.: Longitudinal relations between childrens exposure to tv violence and their aggressive and violent behavior in young adulthood. Developmental Psychology **39**(2), 201–221 (2003)
21. Li, S., Gong, D., Yuan, Y.: Face recognition using weber local descriptors. In: Neurocomputing, vol. 122, pp. 272–283. Elsevier (2013)
22. Liang, Y., Hany, F., Tapio, S., Esko, A.: Physical violence detection for preventing school bullying. Advances in Artificial Intelligence pp. 1–9 (2014)
23. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: In the 10th IEEE Pacific-Rim Conference on Multimedia, Dec, pp. 990–935. ACM (2009)
24. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1975–1981. IEEE (2010)
25. Mairal, G., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: in Proceedings of the 26th Annual International Conference on Machine Learning (ICML-09), pp. 689–696. JMLR.org (2009)
26. Marco, B., Alberto, D.B., Lorenzo, S.: Multi-scale and real-time non-parametric approach for anomaly detection and localization. Computer Vision and Image Understanding **116**(3), 320–329 (2012)
27. Mehrsan, J.R., Martin, L.: Online dominant and anomalous behavior detection in videos. Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on pp. 2609–2616 (2013)
28. Nam, J., Alghoniemy, M., Tewfik, A.: Audio-visual content-based violent scene characterization. Proceedings of IEEE International Conference on Image Processing (ICIP1998) pp. 353–357 (1998)
29. Nguyen, N., Phung, D., Venkatesh, S., Bui, H.: Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. In: Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on, pp. 955–960. IEEE (2005)
30. Oikonomopoulos, A., Patras, I., Pantic, M., Paragios, N.: Trajectory-based representation of human actions. Artificial Intelligence for Human Computing **44**(51), 133–154 (2007)
31. Popoola, O.P., Wang, K.: Video-based abnormal human behavior recognition - a review. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **42**(6), 865–878 (2012)
32. Saghafi, B., Rajan, D.: Human action recognition using pose-based discriminant embedding. Signal Processing: Image Communication **27**(1), 96–111 (2012)
33. Sarvesh, V., Anupam, A.: A survey on activity recognition and behavior understanding in surveillance video. The Visual Computer **29**(10), 983–1009 (2013)
34. Shi, Y., Huang, Y., Minnen, D., Bobick, A., Essa, I.: Propagation networks for recognition of partially ordered sequential action. Computer Vision and Pattern Recognition (CVPR), 2004 IEEE Conference on pp. 862–869 (2004)
35. de Souza, F.D.M., Chavez, G.C., do Valle, E.A., de A. Araujo, A.: Violence detection in video using spatio-temporal features. In: SIBGRAPI 2010, Proceedings of the 23rd SIBGRAPI Conference on Graphics, Patterns and Images, pp. 224–230. IEEE (2010)
36. Tal, H., Yossi, I., Orit, K.G.: Violent flows: Real-time detection of violent crowd behavior. 3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) pp. 1–6 (2012)
37. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: European Conference on Computer Vision (ECCV), 2008, pp. 548–561. Springer (2008)
38. Vishwakarma, S., Sapre, A., Agrawal, A.: Action recognition using cuboids of interest points. In: In: IEEE Int. Conf. on Signal Processing, Communications and Computing (ICSPCC), pp. 1–6. IEEE (2011)

39. Wang, B., Li, W., Yang, W., Liao, Q.: Illumination normalization based on weber's law with application to face recognition. In: Signal Processing Letters, IEEE, vol. 18, pp. 462–465. IEEE (2011)
40. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on, pp. 1794–1801. IEEE (2009)
41. Yang, J., Yu, K., Huang, T.: Supervised translation-invariant sparse coding. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on pp. 3517–3524 (2010)
42. Yang, Y., Song, J., Huang, Z., Ma, Z., Sebe, N., Hauptmann, A.G.: Multi-feature fusion via hierarchical regression for multimedia analysis. IEEE Transactions on Multimedia pp. 572–581 (2013)
43. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Modeling individual and group actions in meetings with layered hmms. Multimedia, IEEE Transactions on **8**(3), 509–520 (2006)
44. Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., He, X.: A new method for violence detection in surveillance scenes. Multimedia Tools and Applications pp. 1–23 (2015)
45. Zhou, W., Wang, C., Xiao, B., Zhang, Z.: Action recognition via structured codebook construction. Signal Processing: Image Communication **29**(4), 546–555 (2014)
46. Zhu, Y., Zhao, X., Fu, Y., Liu, Y.: Sparse coding on local spatial-temporal volumes for human action recognition. 10th Asian Conference on Computer Vision, ACCV2010 pp. 660–671 (2011)