



# Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation

Jing Yu<sup>1</sup> · Wei Song<sup>2</sup> · Guozhu Zhou<sup>3</sup> · Jian-jun Hou<sup>1</sup>

Received: 29 May 2018 / Revised: 14 November 2018 / Accepted: 19 November 2018 /

Published online: 7 December 2018

© The Author(s) 2018

## Abstract

Most existing feature descriptors for video have limited representation ability. In order to improve the recognition accuracy of method for detecting the videos that include violent scenes and take advantage of the logical structure of video sequences, a novel feature constructing approach based on three dimensional histograms of gradient orientation (HOG3D), the Bag of Visual Words (BoVW) model, and feature pooling technology is proposed. This approach, combined with kernel extreme learning machine (KELM), can be used to detect violent scene. First, the HOG3D feature is extracted on the block level for video, and then the K-Means clustering algorithm is implemented to generate visual words. Then, the bag of visual words framework is used for the quantization of feature. And the feature pooling technology is operated to generate a feature vector for an entire video segment, and feature vectors of training data and testing data were used separately to train the model and evaluate the performance of the proposed approach. The experimental results showed that the proposed feature descriptor had good representation and generalization abilities. The proposed approach is efficient for violent scene detection, and the accuracy matches the best result on Hockey dataset, and it outperforms state-of-the-art on Movies.

**Keywords** Violent scene detection · HOG3D · Bag of visual words · Feature pooling · Kernel extreme learning machine

---

✉ Jing Yu  
ssoohay@139.com

<sup>1</sup> School of Electronic Information and Engineering, Beijing Jiaotong University, Beijing 100044, China

<sup>2</sup> School of Information and Engineering, Minzu University of China, Beijing 100081, China

<sup>3</sup> Beijing Polytechnic, Beijing 100176, China

## 1 Introduction

The rapid growth of the Internet has led to an increase in the number of user-generated videos (UGVs), and the need for filtering harmful content has augmented significantly. It is a challenging task to detect violent scenes, and it is important in different applications, such as intelligent surveillance, video retrieval, Internet filtering, and so on.

Many schemes have emerged in the last few years to detect the violent scene in number of videos effectively. Nam et al. [18] tried to detect violent scenes in TV drama and movies by detecting visual concepts, such as fire and blood. Gong et al. [8] proposed a three-stage approach that integrated low level features and high-level affective concepts. Candidate shots with potential violent content were detected based on the extracted visual-audio features, and typical violence-related audio effects were detected from candidate shots. The estimations from the previous steps were combined to make final decisions. Giannakopoulos et al. [7] presented an approach to detect violent scene in movies using a combination of low-level audio-visual features, such as the Mel Frequency Cepstrum Coefficient (MFCC), Zero-Crossing Rate (ZCR), motion features, and k-Nearest neighbor classifier.

Recently, techniques used for human action recognition have received significant attentions quickly and correctly [22], and a popular way of achieving such recognitions is to combine local or global spatial-temporal interest points (STIPs) with bag of visual word framework [19, 23]. Nievas et al. [19] used computer vision techniques to detect violent scenes. Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF) and Motion Scale Invariant Feature Transform (MoSIFT) [2] are combined with bag of visual word frame to construct video descriptors, respectively. A Support Vector Machine (SVM) classifier can be trained to detect violent scene. Experimental results on Hockey and Movies datasets indicate that, the method using HOG as feature gets the best accuracy(91.7%) on Hockey, MoSIFT performs well on both datasets, 90.9% on Hockey and 84.2% on Movies. Deniz et al. [5] proposed a novel video descriptor based on extreme acceleration patterns, and then SVM and Adaboost were used separately as classifiers for violent scene detection. They ran 10-fold cross-validation experiments on both the Hockey and Movies datasets. The accuracy on Movies has reached 98.9%, which outperformed the state-of-the-art approaches. However, the accuracy on Hockey was lower than the scheme proposed in [14], and it was 90.1%. In [14], Long et al. tried to apply the sparse coding technique to detect violent scene and obtained encouraging experimental results on Hockey. However, they did not give their result on the Movie dataset. To date, deep learning (DL) has been demonstrated as an effective model in computer science [3, 6, 9, 16, 20], and the state-of-the-art results have been promoted extensively in many fields, including face recognition. Ding et al. [6] constructed a Convolutional Neural Network (CNN) that contained nine layers, including the input and output layers. The dimensions of the input layer and the output layer were both  $60 \times 90 \times 40$ . Their experimental results on Hockey indicated that the proposed deep neural network was effective for detecting violent scene problem. In [21], Zhang proposed a Gaussian Model of Optical Flow (GMOF) in a fast and robust framework which is used to detect and localize violence in surveillance scenes. Then, the Support Vector Machine (SVM) is used to as a classifier, and the results showed the effectiveness of their method. In the algorithm for the violent scene detection [13], the authors evaluated many features, and, based on the performance of each feature, they combined some features for use in their system in a violent scene detection system. The low-level features and mid-level visual information also were evaluated. The author in [25] developed a new deep learning network named FightNet. Three kinds of input modalities were trained, and the first one was optical flow images, the other two were RGB images for spatial network and the acceleration images for temporal network. This method was not based on

the hand-crafted features. The author said the scheme was so good that it was better than other methods. The method proposed in [1] combined the spatio-temporal positions and the local features and proposed a novel algorithm. It was the extension of the improved Fisher Vectors, and the popular window approach was used. This method can detect when the violence scene occurs. Zhang [24] proposed a robust motion image descriptor to detect the violence scene. And the motion weber local descriptor (MoWLD) is used to be added the temporal component. The descriptor can combine the low-level image appear information and the local motion information. And the non-parametric Kernel Density Estimation (Ked) is used to eliminate the irrelevant features and the redundant information. In [17], the authors proposed a scheme that can detect the dynamics of pedestrians. And this method used the spatial information by the substantial derivative. It was effective in the violence detection, and it can also be used in the panic situation.

Most local or global features used in previous research were extracted from static video frames, and the problem of using these features for video representation was that the context information between frames cannot be defined. HOG3D is an extension of HOG that is used as a local descriptor for video frame sequences, and related studies have shown that it has better ability than 2DHOG to recognize in human actions in videos [12]. In addition, the training process of most classifiers, such as SVM, is complex and time consuming. The Extreme Learning Machine (ELM) is a single hidden layer feedforward network, and it is used in many tasks [10, 11, 15]. The Kernel ELM extends the kernel method to ELM, and an advantage of KELM is that there is no longer any need to specify the number of nodes in the hidden layer. Inspired by the representation ability of HOG3D and the performance of KELM, in this paper, we describe a video descriptor based on block level HOG3D that trains KELM as the classifier for violent scene detection.

The main contributions of this paper include:

- 1) A feature descriptor is constructed based on block scale HOG3D, bag of visual word and pooling technique;
- 2) The influences of different feature pooling techniques are compared through experiments on the well-known Hockey and Movies datasets;
- 3) The Kernel Extreme Learning Machine is used to solve the problem of detecting violent scene, and different types of kernel functions were studied;
- 4) Experiments using the proposed video descriptor and the different classifiers were conducted to check the validation of the descriptor and KELM.

The remainder of this paper is organized as follows. Section 2 discusses the basic theory of HOG3D and the Kernel Extreme Learning Machine. The proposed scheme, including procedure for constructing the video and setting up the parameters, is presented in Section 3. Section 4 discusses the experimental results, including the analysis of the influence of the pooling strategies, analysis of the rationality of choosing KELM as the classifier, and comparisons of performances. Our conclusions are presented in Section 5.

## 2 Basic theories

### 2.1 HOG3D descriptor

HOG3D is an extension of HOG, and it was created by Dalal et al. [4] for human detection in static images. Static images are divided into overlapped small spatial

regions called blocks, and each block is divided further into several sub-blocks that are called cells. A local 1D histogram of gradient directions or edge orientations over pixels is accumulated for each cell. Contrast-normalize are implemented by accumulating a measure of local histogram energy over a block to achieve less invariance in illumination, and shadowing. In HOG3D, the concepts block and cell are extended to 3D, and gradient orientation is quantized using a regular polyhedron, which means there are only five kinds of quantization strategies to choose from, i.e., 4-, 6-, 8-, 12-, and 20-sided polygon. For example, let the center of gravity of a regular icosahedron (a 20-sided polygon) be the origin of a three-dimensional Euclidean coordinate system, the centers of the surfaces match the following points:

$$(\pm 1, \pm 1, \pm 1), \left(0, \pm \frac{1}{\varphi}, \pm \varphi\right), \left(\pm \frac{1}{\varphi}, \pm \varphi, 0\right), \left(\pm \varphi, 0, \pm \frac{1}{\varphi}\right) \quad (1)$$

where  $\varphi = \frac{1+\sqrt{5}}{2}$  is the golden ratio, and the vectors from the origin to centers of the surfaces are the orientation vectors for quantization. Suppose  $\vec{g}_b$  is a 3D gradient vector, then the procedure for calculating HOG3D in block scale can be described as follow [15]:

- 1) Let  $\mathbf{P} = (p_1, \dots, p_{20})^T$ , where,  $p_i$  is the orientation vector for quantization defined in eq. 1.
- 2) Calculate the projection  $\hat{q}_b$  for  $\vec{g}_b$  as

$$\hat{q}_b = \frac{P \cdot \vec{g}_b}{\|\vec{g}_b\|_2} \quad (2)$$

Process  $\hat{q}_b$  as formula 3 using threshold  $t = p_i^T \cdot p_j$ , where  $p_i, p_j$  are the neighboring quantization vectors. When using regular icosahedrons for quantization,  $t \approx 1.29107$ .

$$\hat{q}'_{bi} = \begin{cases} \hat{q}_{bi} - t, & \text{if } \hat{q}_{bi} - t > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Where  $\hat{q}'_{bi}$  is the  $i^{\text{th}}$  component of the processed vector  $\hat{q}'_b$ .

- 3) Calculate the weight parameters as follows:

$$q_b = \frac{\|\vec{g}_b\|_2 \cdot \hat{q}'_b}{\|\hat{q}'_b\|_2} \quad (4)$$

- 4) Use Eq. 4 to calculate mean gradient for each cell, and get the sum of the quantized mean gradients of the cells in a block as its histogram.

In order to enhance the statistical character, the block level HOG3D descriptor rather than original HOG3D is used further to construct the video descriptor. The K-means clustering algorithm is used to generate vocabulary on the extracted block level descriptors, and the pooling technique is applied to make the final feature descriptor for the entire video clip.

## 2.2 Kernel extreme learning machine

The extreme learning machine (ELM) was proposed by Huang et al. [10, 12] for single layer feedforward networks, and it can be used to solve regression and classification problems. The input parameters for the hidden nodes can be assigned to randomly generated values, and the parameters do not have to be refreshed during the whole training procedure.

Let  $\mathbf{X} \in R^d$  be the input vector, and then the output function of ELM is defined as:

$$f_L(X) = \sum_{i=1}^L \beta_i h_i(X) = h(X)\beta \tag{5}$$

where  $\beta = [\beta_1, \beta_2, \dots, \beta_L]^T$  is the output weights vector between the hidden layer of  $L$  nodes and the output node, and  $h(X) = [h_1(X), h_2(X), \dots, h_L(X)]^T$  denotes the outputs of the hidden layer when the input vector is  $X$ . The mapping function  $h(X)$  maps the  $d$  dimension vector  $X$  into  $L$  dimension space. For binary classification problems, the decision function of ELM is:

$$f_L(X) = \text{sign}(h(X)\beta) \tag{6}$$

The aim of ELM is to minimize the training error and to get the minimum norm of the output weights. The mathematical model built for ELM is

$$\begin{aligned} \min L_{ELM} &= \frac{1}{2} \|\beta\|^2 + C \frac{1}{2} \sum_{i=1}^N \xi_i^2 \\ \text{s.t. } h(X_i)\beta &= t_i - \xi_i, \quad i = 1, \dots, N \end{aligned} \tag{7}$$

where  $\beta$  is the output weights vector between the hidden layer and the output node, and  $C$  is a parameter that provides a tradeoff between the distance of the separating margin and the training error. The corresponding dual optimization problem is:

$$L_{DEL M} = \frac{1}{2} \|\beta\|^2 + C \frac{1}{2} \sum_{i=1}^N \xi_i^2 - \sum_{i=1}^N \alpha_i (h(X_i)\beta - t_i + \xi_i) \tag{8}$$

The Lagrange multiplier  $\alpha_i$  is corresponds to the  $i^{\text{th}}$  training sample. The corresponding KKT conditions are:

$$\frac{\partial L_{DEL M}}{\partial \beta} = 0 \rightarrow \beta = \sum_{i=1}^N \alpha_i h(X_i)^T = H^T \alpha \tag{9}$$

$$\frac{\partial L_{DEL M}}{\partial \xi_i} = 0 \rightarrow \alpha_i = C \xi_i, \quad i = 1, \dots, N \tag{10}$$

$$\frac{\partial L_{DEL M}}{\partial \alpha_i} = 0 \rightarrow h(X_i)\beta - t_i + \xi_i = 0, \quad i = 1, \dots, N \tag{11}$$

where  $\alpha = [\alpha_1, \dots, \alpha_N]^T$ . Substituting formulas 9 and 10 into formula 11, the equations can be written as

$$\left( \frac{I}{C} + HH^T \right) \alpha = T \tag{12}$$

where  $I$  is identify matrix, and  $T = [t_1, \dots, t_N]^T$ . The following equation can be obtained from formulas 9 and 12:

$$\beta = H^T \left( \frac{I}{C} + HH^T \right)^{-1} T \quad (13)$$

Substituting formulas 13 and 5, the output function of ELM can be written as:

$$f_L(\mathbf{X}) = h(\mathbf{X})\beta = h(X)H^T \left( \frac{I}{C} + HH^T \right)^{-1} T \quad (14)$$

If  $h(X)$  is considered as an unknown feature mapping, and its kernel matrix is defined as:

$$\Omega_{KELM} = h(X) \cdot h(X') = K(X, X') \quad (15)$$

Then, the output function can be written equivalently as:

$$f_L(X) = \begin{bmatrix} K(X, X_1) \\ \vdots \\ K(X, X_N) \end{bmatrix}^T \left( \frac{I}{C} + \Omega_{KELM} \right)^{-1} T \quad (16)$$

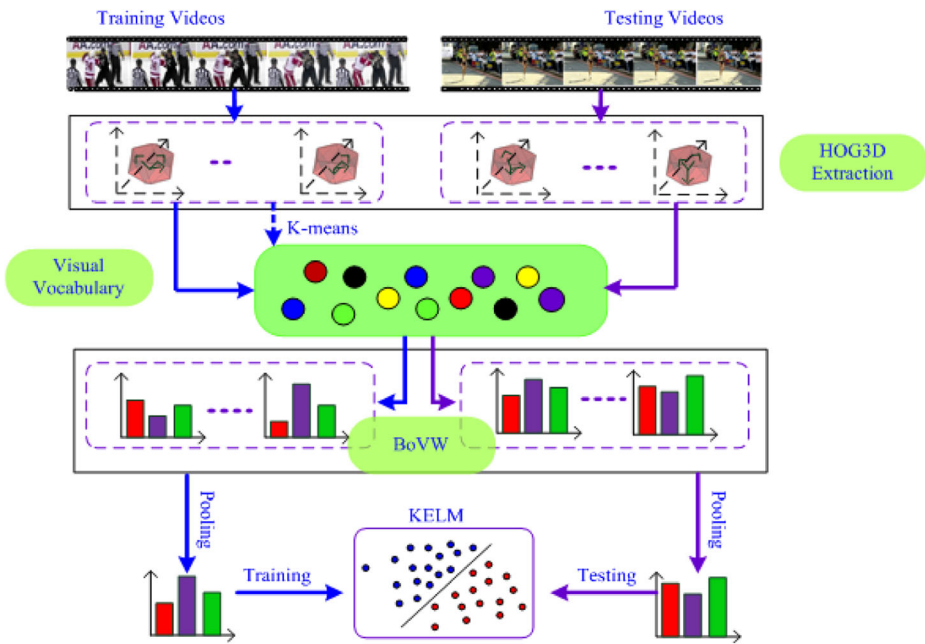
From the above output function, we found that the kernel ELM treats the output of the hidden layer as some unknown feature mapping, and the kernel method can be used in this situation. In this case, determining the proper kernel function is important and there is no need to specify the number of nodes in the hidden layer.

### 3 Proposed method

Fig. 1 shows a schematic diagram of the proposed approach. It consists of two flows, i.e., training and testing. The dataset of the videos is divided into two parts, i.e., training videos and testing videos. The training videos are used to train a binary classifier, and the testing videos are used to test the performance of the trained classifier.

The basic procedure of training is as follows:

- Step1: Extraction of the feature of: The feature of HOG3D are extracted on a block level for each video clip in training video sequences;
- Step2: Construction of the Cluster center: The K-means clustering algorithm is implemented on the extracted feature vectors, and then the cluster center is obtained;
- Step3: Calculation of the Word Frequency vector: The centers of the clusters cluster obtained in the step where the centers of the clusters were constructed are used to construct the visual vocabulary for the BoVW framework, and each video clip is represented by several vectors by using the BoVW framework as a quantization method;
- Step4: Feature pooling: In order to make the extracted features have the same dimensions, feature pooling techniques are implemented for each video clip;
- Step5: Classifier building: Due to the better performance of the extreme learning machine, the kernel extreme learning machine is used as the binary classifier to train the module.



**Fig. 1** Scheme of the proposed approach

The basic procedure of testing is as follows:

- Step1: Extracting the features of HOG3D on the block level for the testing video clips;
- Step2: With the same method, the cluster centers are calculated by the K-means method, and then the word frequency vectors are constructed for testing video sequences by using the BoVW framework with the visual vocabulary obtained in training Steps 2 and 3;
- Step3: Feature pooling techniques are used to generate a fixed dimension feature for the test video clip;
- Step4: The trained KELM model obtained in the training procedure is used to detect the test violence video and to attain the accurate detection.

This method is a hybrid method, and it takes the advantages of the many techniques, such as the HOD3D, K-means, BoVW, feature pooling and the KELM. Although each technique is existing, the combination of these methods is the first time to be proposed to the best of our knowledge.

In the HOG3D extraction procedure, first, a video clip is divided into some overlapped cubes along the time dimension. Each cube is divided into blocks, and each block consists of several cells. In this paper, each cube contains eight video frames, and the sizes of the blocks and the cells are  $8 \times 8 \times 8$  and  $2 \times 2 \times 2$ , respectively. As mentioned in section 2, there are five kinds of strategies for gradient quantization that were inspired by the research in [12]; we use regular icosahedrons to quantify the gradient vectors. Bag of Visual Word is an effective method for quantifying features, and K-means clustering is used to construct the visual vocabulary for BoVW. As was done in the literature [19], we ran experiments that corresponded to different visual vocabularies that contained 50, 100, 200, 500 and 1000 visual words, respectively.

After implementing BoVW in order to quantify the features, each cube used for the extraction of HOG3D is represented by a fixed dimensional vector. The feature pooling technique is need for each video clip which contains several cubes.

Assume that we have the matrix  $M = [Y^{(1)}, \dots, Y^{(k)}]_{k \times d}^T$ , where  $Y^{(i)}$  denotes vectors for pooling.  $Y^{(i)} \in R^d$ , ( $i = 1, \dots, k$ ), Then  $\tilde{Y} = f(M)$  is defined as pooling operation for feature matrix  $M$ .  $Y$  is the pooled feature and  $f(\cdot)$  is the pooling function. Typical pooling functions include sum pooling, average pooling and Max pooling which are defined as follows:

$$f_{sum}(\cdot) : \quad \tilde{Y}_j = \sum_{i=1}^k Y_j^{(i)} \quad (17)$$

$$f_{average}(\cdot) : \quad \tilde{Y}_j = \left( \sum_{i=1}^k Y_j^{(i)} \right) / k \quad (18)$$

$$f_{max}(\cdot) : \quad \tilde{Y}_j = \max(|Y_j^{(1)}|, \dots, |Y_j^{(k)}|) \quad (19)$$

where  $\tilde{Y}_j$  s the  $j^{\text{th}}$  component of  $Y$ , and  $Y_j^{(i)}$  is the  $j^{\text{th}}$  component of  $Y^{(i)}$ . In this paper, we have provided comparison studies on the performance of different pooling techniques using different size of visual vocabulary for BoVW and different classifiers on the Hockey and Movies datasets.

The pooled feature for video clip must be put into a classifier for classification, and compared to extreme learning machine, the kernel extreme learning machine usually has better performance with less limitation in the solving procedure. Histogram intersection kernel KELM is proposed as the classifier for the violent scene detection based on comparison experiments when using different kernel function for KELM and SVM.

## 4 Experimental results and discussion

### 4.1 Benchmark datasets

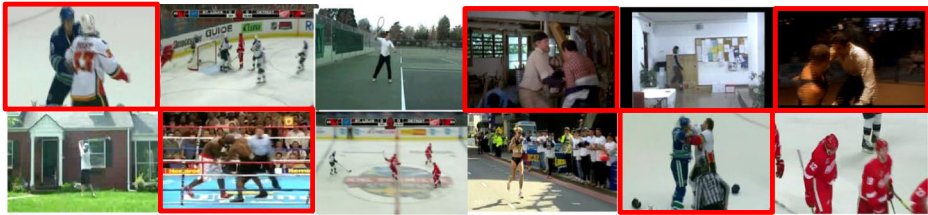
In order to evaluate the performance of the proposed method, two well-known datasets were used, i.e., Hockey [19] and Movies [19], which are described in the following.

**Hockey Dataset:** The Hockey dataset consists of 1000 video clips from Hockey competition, and 500 are labelled as violent and the others are labels as normal. There are 41 frames in each video clip. The frame rate is 25 frames per second and the resolution of each frame is  $360 \times 288$ .

**Movies Dataset:** The Movies dataset is composed of 100 violent video segments and 100 normal video clips. The resolution of the video clips is either  $720 \times 576$  or  $720 \times 480$ . The total number of each video is ranging from 10 to 60, and the frame rate is 25 frames per seconds. Figure 2 shows some sample frames from Hockey and Movies datasets, and the ones in red rectangle are from violent videos.

In the next part, we use average precision for the two datasets to measure the performance of the pooling strategy (part 4.2), the performance of the different kernel extreme learning machine (part 4.3), and the comparison of the performance of KELM





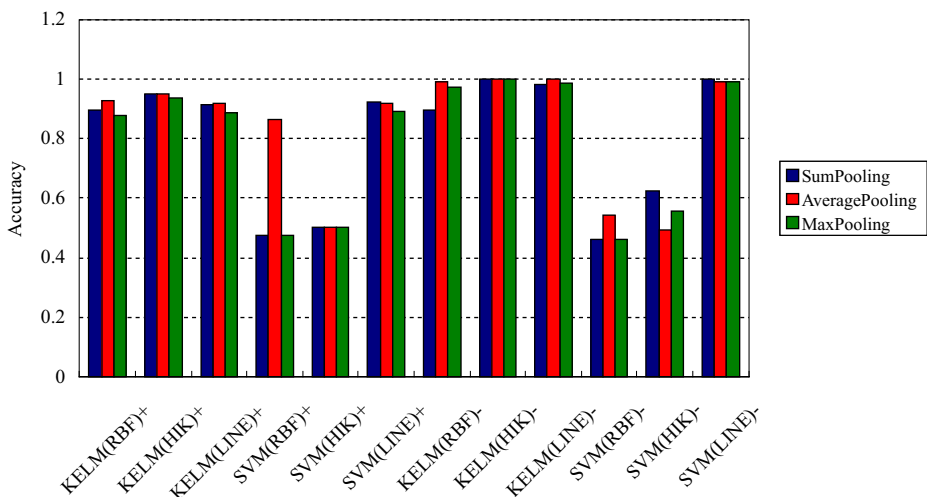
**Fig. 2** Sample frames from Hockey and Movies datasets (frames in red rectangles are violent video clips)

and SVM (part 4.4). We obtained a very superior result over several state-of-the art approaches in term of average precision, which demonstrates the advantage of the proposed (part 4.5).

### 4.2 Comparison of pooling strategy

As mentioned in section 3, three kinds of pooling techniques i.e., sum pooling, average pooling, and max pooling, were studied to evaluate their performance on violent scene detection. Figure 3 shows the results of the three kinds of pooling strategies on Hockey and Movies using KELM and SVM with RBF, HIK and Line as kernel. The accuracy values are the average result of 10 runs and 10-fold cross-validation.

From Fig. 3, we can conclude that the performances of the three pooling methods are similar under most conditions. On the Hockey dataset, when using KELM as the classifier, average pooling outperformed both sum pooling and max pooling. But on Movies, sum pooling with HIK KELM had an accuracy up to 99.95%. When using SVM as the classifier, the largest accuracy is obtained by using sum pooling and the line kernel on both Hockey and Movies datasets.



**Fig. 3** Comparison of the feature pooling strategies (\*+ represents the result obtained on Hockey, and \*- represents the results obtains on Movies)

The results describe above indicate that, under most conditions, sum pooling and average pooling tend to have better performances than the max pooling method. And for the task of violent scene detection, sum pooling was a better choice on both Hockey and Movies when using line kernel SVM as classifier. When using HIK KELM as the classifier, the accuracy of using sum pooling and average pooling on Hockey are almost the same, i.e., 95.02% and 95.05%, respectively.

### 4.3 Analysis of the performance of KELM

One important step for constructing the proposed video feature descriptor is using BoVW for feature quantization. Fig. 4 shows the relationship between accuracy and the size of the visual vocabulary on the Hockey dataset using the average pooling method and three kinds of kernel functions. With the increase in the number of words in the visual vocabulary, the accuracies were increased by using RBF or HIK as the kernel for KELM. For line kernel ELM, the accuracy increased when the number of words was less than 200, and it decreased when number of words was larger than 200. The best accuracy using KELM on Hockey, i.e., 95.05% was obtained when the number of words was 1000 and the kernel was HIK. Fig. 5 is the result of using Movies as the evaluation dataset, and the figure indicates that the HIK kernel was better for Movies than the other kernels. The best accuracy using HIK based KELM was 99.95%, and the corresponding number of words was 100 and pooling method was sum pooling.

Figure 6 describes the best performance using different kernel ELM on Hockey and Movies. HIK-based KELM provided the best accuracy on both datasets, and the accuracies were 95.05% for Hockey and 99.95% for Movies. On the Hockey dataset, the RBF kernel ELM outperformed the Line kernel ELM, but just be opposite was true on Movies. In addition, the performances of all three kinds of KELM on both datasets further supported the effectiveness of KELM for violent scene detection.

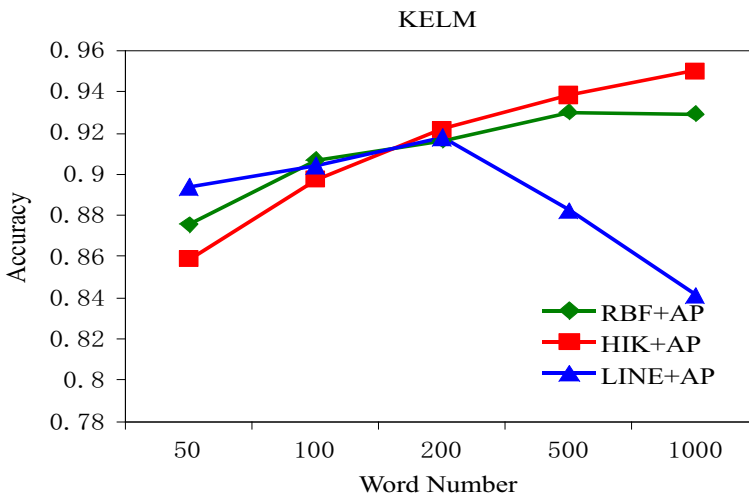


Fig. 4 Comparison of different kernel extreme learning machines on Hockey

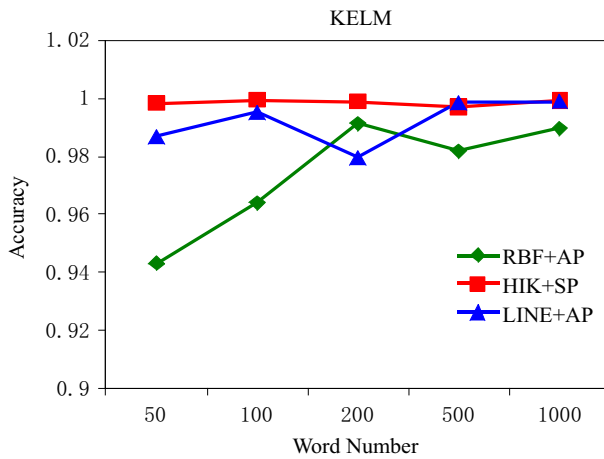


Fig. 5 Comparison of different kernel extreme learning machines on Movies

### 4.4 Comparison of the performances of KELM and SVM

To evaluate the performance of KELM further, in this section, comparison experiments using different classifiers have been implemented. Fig. 7 and Fig. 8 are comparison results using KELM and SVM on Hockey and Movies, respectively. In the Hockey dataset, HIK-based KELM always had higher accuracy than SVM, and the best results were 95.05% for KELM and 92.40% for SVM. On Movies, the performances of the two classifiers were extremely close, and the accuracy of KELM was as high as to 99.95%. According to the average performances of the two datasets that we have studied, we concluded that KELM was better than SVM for detecting to violent scenes.

### 4.5 Comparison with state-of-the-art algorithms

In order to evaluate the performance of the proposed approach, we compared of the existing and the proposed algorithms for violent scene detection, and the results are shown in Table 1. From Table 1, it is apparent that that using SVM as the classifier result in the accuracy obtained by the features in literatures [5, 19] was less than the accuracy obtained when the proposed descriptor was used. The Deep learning method is recognized as the popular scheme right now. And there are many different strutures and different networks, such as [6, 25]. The proposed

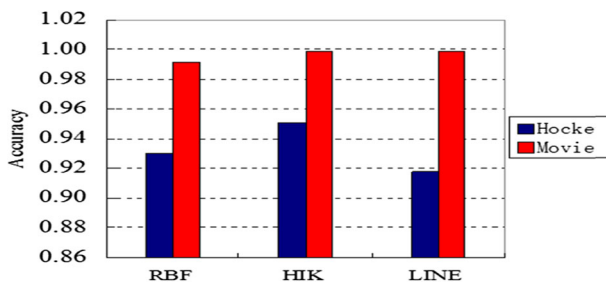
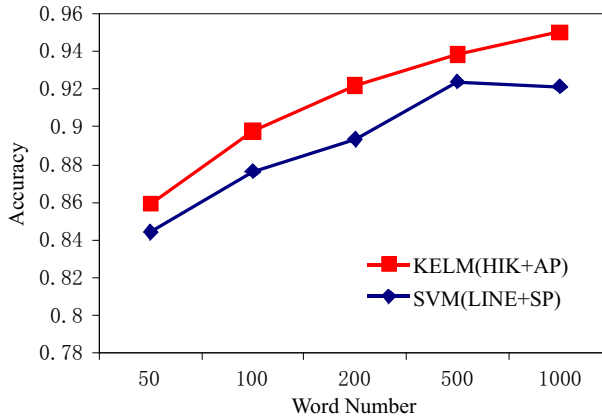
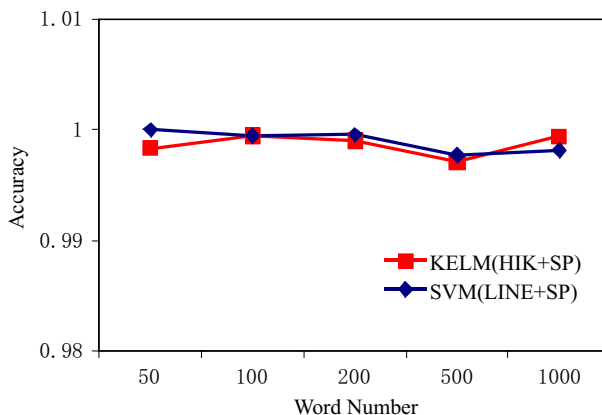


Fig. 6 Performance of each KEML on Hockey and Movies



**Fig. 7** Comparison of the performance of KEML and SVM on Hockey

descriptor is more effective than the 3DCNN [6], but the accuracy is less than the FightNet [25]. The reason is that the FightNet uses three features, and they are RGB, optical flow and temporal network. That is to say, the features are more than the proposed method. However, the method in [6] just used the convolution network to detect the violence scene. So the accuracy in this scheme is less than the proposed method. The Method in [1] used the local features to detect the violent scene. However, the accuracy are less than the proposed method, and the results showed that the descriptor in this method is more effective than these two methods. For the Hockey dataset, the accuracy was up to 95.05% when using the proposed descriptor and KELM, which was almost 1% higher than the state-of-the-art algorithms. When the proposed video descriptor is combined with different classifiers, much better results were obtained. The results we obtained by using KELM as the classifier on both datasets are better than the state-of-the-art SVM classifier. On Movies, the accuracy of KELM matched that of SVM. The results from the two datasets indicate that the proposed video descriptor has strong representation ability for violent videos and the KELM is effective for violent scene detection. The accuracy in FightNet [25] is 100%, and it is better than the proposed method, the reason is the same as the reason on the Hockey dataset. But the accuracy of the proposed method is



**Fig. 8** Comparison of the performance of KEML and SVM on Movies

**Table 1** Comparison of the performance with other algorithms (%)

	Hockey	Movies
HOG + SVM [19]	91.7	49
HOF + SVM [19]	88.6	59
MoSIFT + SVM [5]	91.2	84.2
Extreme acceleration + SVM [5]	90.1	98.9
MoSIFT + KDE + Sparse Coding [14]	94.3	–
3D-CNN [6]	91	–
FightNet [25]	97	100
FastFV-based Sliding Window [1]	93.7	99.5
MoWLD+BoW [24]	91.9	–
MoWLD+Sparce Coding [24]	93.7	–
MoWLD+KDE + Sparce Coding [24]	94.9	–
Local force- $F^L$ [17]	–	93.4±1.24
Convective force - $F^{Cv}$ [17]	–	92.16±1.13
$F^L F^{Cv}$ [17]	–	96.89±0.21
Proposed descriptor + ELM	94.08	99.43
Proposed descriptor + SVM	92.40	99.98
Proposed descriptor + KELM	95.05	99.95

better than the other methods, such as the method based the sliding window [1], and the  $F^L$  and  $F^{Cv}$  [17], and the result showed that the proposed descriptor is effective.

## 5 Conclusions

In this paper, we present a novel special video content detection algorithm based on the extreme learning machine and the spatio-temporal descriptor based on 3D Gradients. It is the combination of many techniques, and it has many advantages:

- 1) After synthesizing the good characteristics of HOG3D, BoVW model, and feature pooling technology, a novel video descriptor approach has been proposed. The external experimental results indicate that, the descriptor reveals has strong representation ability, and is stable.
- 2) Because of the good generalization performance and easy implementation of KELM, the classifier can be constructed based on KELM, and the analysis shows that KELM is more suitable for classifying special videos.
- 3) Comparisons of the performance of the new, innovative technique with the performances of existing schemes further demonstrated the effectiveness of the proposed scheme. In the future, we will try to develop a novel deep learning method based on multi-layer ELM to further improve the detection performance for the violence videos.

**Acknowledgements** This work was supported in part by the National Natural Science Foundation of China project (615034244,61331013); Promotion plan for young teachers' scientific research ability of Minzu University of China Project; Promotion Project for teachers' scientific research ability of the Beijing Polytechnic Project (YZK2015013 CJGX2016-SZ-05/008 CJGX2016[18]-SZJC-05/008 02362050301).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. Bilinski P, Bremond F (2016) Human violence recognition and detection in surveillance videos[C]// IEEE Int Conf Adv Video Sign Based Surveill. IEEE, Colorado, USA: 30–36
2. Chen MY, Hauptmann A (2009) Mosift: recognizing human actions in surveillance videos. CMU-CS-09-161, Carnegie Mellon University, 1-16
3. Chen XW, Lin XT (2014) Big data deep learning: challenges and perspectives. Access, IEEE 2:514–525
4. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. Comput Vision Pattern Recogn 2005. CVPR 2005. IEEE Comput Soc Conf IEEE, 2005 1:886–893
5. Deniz O, Serrano I, Bueno G, et al. (2014) Fast violence detection in video. 9th Int Conf Comput Vision Theory Appl (VISAPP): 478-485
6. Ding CH, Fan SK, ZHU M, et al. (2014) Violence detection in video by using 3D convolutional neural networks. Advances in visual computing. Springer international publishing, 551-558
7. Giannakopoulos T, Makris A, KOSMOPOULOS D et al (2010) Audio-visual fusion for detecting violent scenes in videos. artificial intelligence: Theories, models and applications. Springer, Berlin Heidelberg, pp 91–100
8. Gong Y, Wang WQ, Jiang SQ et al (2008) Detecting violent scenes in movies by auditory and visual cues. Advances in Multimedia Information Processing-PCM 2008. Springer, Berlin Heidelberg, pp 317–326
9. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
10. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1):489–501
11. Huang GB, Zhou HM, Ding XJ et al (2012) Extreme learning machine for regression and multiclass classification. Syst Man Cybernet Part B: Cybernet IEEE Trans 42(2):513–529
12. Klaser A, Marsezlek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. BMVC 2008-19th Brit Mach Vision Conf. Brit Mach Vision Assoc 275:1–10
13. Lam V, Phan S, Le DD, Duong DA, Satoh SI (2017) Evaluation of multiple features for violent scenes detection. Multimed Tools Applic 76(5):7041–7065
14. Long X, Chen G, Yang J, et al. (2014) Violent video detection based on MoSIFT feature and sparse coding. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE Int Conf IEEE: 3538–3542
15. Luo MX, Zhang K (2014) A hybrid approach combining extreme learning machine and sparse representation for image classification. Eng Applic Artif Intell 27:228–235
16. Mnih V, Kavukcuoglu K, Silver D et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
17. Mohammadi S, Kiani H, Perina A, Murino V (2015) Violence detection in crowded scenes using substantial derivative. 2015 12th IEEE Int Conf Adv Video Signal Based Surveill (AVSS), IEEE: 1–6
18. Nam J, Alghoniemy M, Tewfik AH (1998) Audio-visual content-based violent scene characterization. Image Process 1998. ICIP 98. Proc. 1998 Int Conf IEEE 1:353–357
19. Nieves EB, Suarez OD, Garcia GB, et al (2011) Violence detection in video using computer vision techniques. In: Computer analysis of images and patterns. Springer Berlin Heidelberg, 332–339
20. Shim H, Lee S (2015) Multi-channel electromyography pattern classification using deep belief networks for enhanced user experience. J Cent South Univ 22(5):1801–1808
21. Tao Z, Yang ZJ, Jia WJ, Yang BQ, Yang J, He XJ (2016) A new method for violence detection in surveillance scenes. Multimedia Tools and Applications 75(12):7327–7349
22. Xia LM, Huang JX, Tan LZ (2013) Human action recognition based on chaotic invariants. J Cent South Univ 20:3171–3179
23. Yang J, Jiang YG, Hauptmann AG, et al. (2007) Evaluating bag-of-visual-words representations in scene classification. Proc Int Workshop Workshop Multimed Inform Retrieval. ACM, 197–206
24. Zhang T, Jia W, Yang B, Yang J, He X, Zheng Z (2017) Mowld: a robust motion image descriptor for violence detection. Multimed Tools Applic 76(1):1419–1438
25. Zhou P, Ding Q, Luo H et al (2017) Violent interaction detection in video based on deep learning. Journal of physics conference series. J Phys Conf Ser 844(1):012044



**Yu Jing** received his master degree from University of science and technology of China. Now he is pursuing Ph. D in Beijing Jiaotong University. And he is now a professor in department of electronic technology, Beijing Polytechnic. His research interests include image processing and electronic communication.



**Song Wei** received his Ph.D. degree from Beijing Jiao tong University in 2010 in Electronics Engineering. He was an assistant professor in Research Institutes of Information Technology, Tsinghua University from 2011 to 2012. He was also a visiting scholar in New Jersey Institute and Technology form 2017–2018. Now he is an associate professor in school of Information Engineering, Minzu University of China. His research interests include multimedia network and image processing.



**Zhou Guozhu** is a professor in Beijing Polytechnic. His research interests include image processing and software design.



**Hou Jianjun** is now a professor in school of electronic engineering and information, Beijing Jiaotong University. His research interests include signal processing.