Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **4**, July **2018**

Violent Scene Detection based on Multiple Instance Learning and 3D Histogram of Oriented Gradients

Jing Yu

School of Electronic Information and Engineering,
Beijing Jiaotong University
Beijing 100044, China
13111001@bjtu.edu.cn

Wei Song

School of Information and Engineering,
Minzu University of China
Beijing 100081, China
songwei@muc.edu.cn

Guo-zhu Zhou

Beijing polytechnic University,
Beijing 100176, China

Jian-jun Hou

School of Electronic Information and Engineering,
Beijing Jiaotong University
Beijing 100044, China

Received February, 2018; revised June, 2018

Abstract. *In order to take advantage of the logical structure of video sequences and improve the recognition accuracy of the violence videos, a novel violent scene detection method based on 3D Histogram of Oriented Gradients (HOG3D) and Multi-Instance Learning (MIL) is proposed. Firstly, HOG3D was extracted from video clips. Then, K-means clustering algorithm was implemented to generate visual words vocabulary, and Bag of Visual Words (BoVW) model was used to construct the final feature for video frame sequence. Next, the violent scene detection issue was formulated as a MIL problem, and an instance clean method was proposed to remove noise in sample data. With experimental evaluations on the well-known Hockey and Movies benchmarks, the results demonstrate that Citation-kNN is more suitable than the other two tested MIL methods named Axis-Parallel hyper Rectangle (APR) and mi-SVM for violent scene detection. And the proposed scheme obtains very competitive results: 92.7% on Hockey and 98.8% on Movies respectively in terms of mean average precision. And it outperforms the state-of-the-art on Hockey dataset and matches the best known accuracy on Movies, achieving the best balanced accuracy compared with all other methods.*
Keywords: Violent scene detection; Histogram of oriented gradients (HOG3D); Multi-instance learning (MIL); Instance clean.

1. **Introduction.** Internet video has grown quite fast in recent years with the success of multimedia social network as well as low cost of smart devices, accounting for 90% of the Internet traffic [1]. Videos with harmful content, such as pornographic videos, horror videos and violent videos, are flooding. Therefore, special video content detection is essential to maintain the Internet video ecosystem, especially as the number of young Internet users is increasing rapidly.

Most efforts have been taken to recognize pornographic films [2] and horror videos [3]. Violence detection, however, has attracted less attention until recently. Nam et al.,[4] exploited multiple audio-visual signatures to create perceptual relations for violent scene detection, violent events like fire and bleeding could be identified to help characterize and index violent scenes in TV drama and movies. Cheng et al.,[5] tried to detect audio events like explosions, gunshots and brakes using Hidden Markov Model (HMM), and Gaussian Mixture Model (GMM) was implemented as well to help detecting the semantic context. Gong et al.,[6] presented a three-stage method integrating both visual and audio cues. They identified shots with potential violent content, detected their typical violence-related audio effects, and finally generated a boosting inference using the output of the first two stages. Giannakopoulos et al.,[7] extracted low level audio-visual features such as Mel Frequency Cepstrum Coefficient (MFCC), Zero-Crossing Rate (ZCR) and motion features, and used Bayesian network and weighted k-Nearest Neighbors (kNN) algorithm separately to detect audio and visual events. A kNN binary classifier was trained to make a binary decision.

Recent studies about violent scene detection are focusing on spatial-temporal interest points (STIP), which are widely used in human action recognition. A typical way is to combine local or global STIP with Bag of Visual Words (BoVW) framework to construct a statistical feature for video content, which serves as the input of classifiers such as Support Vector Machine (SVM) [8,9]. Nievas et al., [8] used Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF) and Motion Scale Invariant Feature Transform (MoSIFT) [10] respectively as word space to construct a vocabulary using K-means clustering algorithm. The output histogram which reflected word distribution was then input into a SVM, using which the video was finally classified as violent or normal. When using Hockey as the test dataset, results showed that the classification accuracy using HOG feature (91.7%) is better than that using HOF (88.6%) or MoSIFT (90.9%). However, when the dataset named Movies was used, the classification accuracy using MoSIFT outperformed the best STIP under all conditions.

More recently, Deniz et al., [11] proposed a novel method using extreme acceleration pattern as the main feature, which was estimated by applying Radon transform to the power spectrum of consecutive video frames. SVM and Adaptive Boosting (Adaboost) were separately trained, and 10-fold cross-validation was performed on both Hockey and Movies. Results showed a significant accuracy improvement of up to 12% on Movies compared with known state-of-the-art techniques. Nowadays, deep learning has been demonstrated as an effective model in various kinds of fields such as face recognition, image content understanding, human action recognition, and so on [12-14,22]. One key step to apply deep learning is to construct appropriate architectures for specific problems. To the best of our knowledge, Ding et al., [15] is the first group who constructed a 3D Convolutional Neural Network (CNN) for violent scene detection. There were 9 layers including the input and output in their architecture. The input of the network was a video clip consisting of a sequence of consecutive frames. The input layer had a dimension of , and the output layer was a binary number which output 1 if violence was detected. Results showed that the classification accuracy using CNN on Hockey dataset was up to 91%.

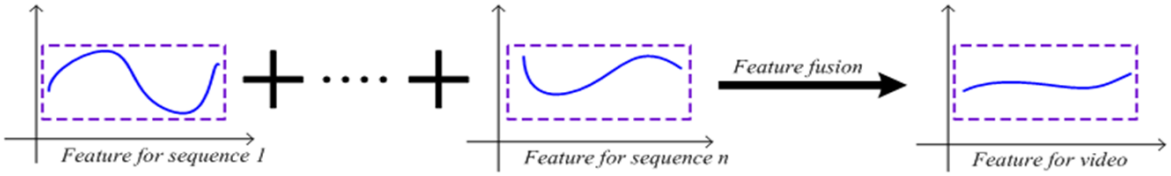

FIGURE 1. Alleviation of discrimination ability by feature fusion

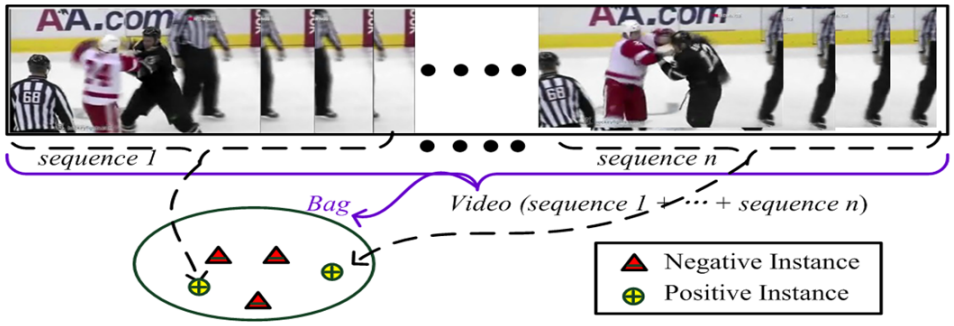

FIGURE 2. The correspondence between video logic structure and MIL

Most traditional methods for violent scene detection are based on either audio or visual features. Different from static images, video sequences contain more context information. HOG3D is a local descriptor for video sequences, which is based on histograms of oriented 3D spatial-temporal gradients [16]. The experiments on action recognition indicate that HOG3D descriptor outperforms the state-of-the-art. On the other hand, extracted features are used to train a classifier, but typical classifiers require a fixed-dimensional input. In order to obtain a fixed-dimensional feature for a video clip, a feature fusion step is usually adopted. However, although feature fusion can represent video clip to some extent, its shortcoming is quite obvious. A video clip usually contains more than one shot, and not every shot contains violent content. In fact, even in a violent video, only a few shots can be labeled as violent. In this case, as described in Fig.1, feature fusion will alleviate the discrimination ability of features. To solve this problem, we propose to use Multi-Instance Learning (MIL) method rather than feature fusion for violent video detection. The logical structure of video can well match the concepts of bag and instances in MIL. As shown in Fig.2, the whole video and video frame sequences correspond to bag and instances in MIL respectively. With the advantages of HOG3D and MIL, we propose a hybrid method combining both HOG3D and MIL for violent scene detection.

2. **Proposed Approach.** Fig. 3 demonstrates the framework of our approach. It consists of two flows, which are training and testing.

The basic procedure of training is as follows:

Step 1. Block level HOG3D is extracted from training video sequences;

Step 2. K-means clustering algorithm is implemented on feature vectors obtained from Step 1 to generate visual words for BoVW;

Step 3. Word frequency vectors are calculated using BoVW model, and the output vectors are the final features for video sequences;

Step 4. MIL problem is constructed, and the features in Step 3 are used to represent instances;

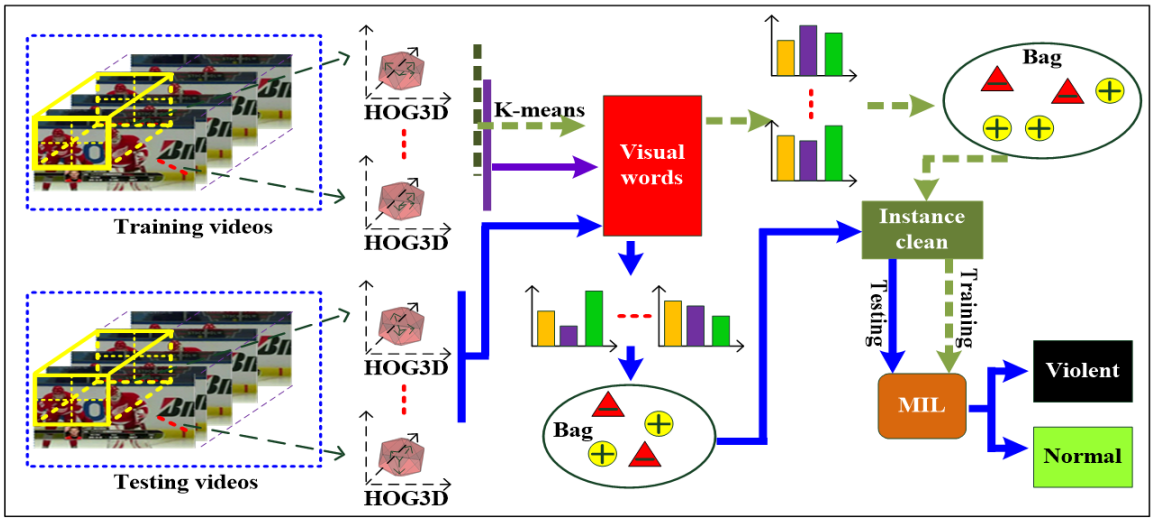

FIGURE 3. Scheme of MIL-based violence detection

Step 5. An instance clean algorithm is implemented to reduce noise in data;
Step 6. A MIL model is then trained.

The basic procedure of testing is as follows:
Step 1. Block level HOG3D is extracted from testing video sequences;
Step 2. Word frequency vectors are calculated for testing video sequences using BoVW model with the visual vocabulary obtained in training Step 2;
Step 3. Violent scene detection is formulated as a MIL problem as Step 4 in the training procedure;
Step 4. Implementing instance clean algorithm;
Step 5. The trained MIL model obtained in training procedure is used for violent scene detection.

In the rest of this section, more details of HOG3D descriptor, MIL algorithms and the proposed instance clean method are introduced.

2.1. **HOG3D descriptor.** HOG3D is an extension form of HOG in 3D space. HOG descriptor is created by Dalal et al., for human detection in static images [17], and it is widely used in human action recognition [18]. For calculating the HOG descriptor, static image is divided into small spatial regions called cells, and for each cell a local histogram of gradient directions or edge orientations over pixels is accumulated. In order to achieve a better invariance to illumination, shadowing, etc., contrast-normalization is achieved by accumulating a measure of local histogram energy over larger spatial regions named blocks, which can normalize all cells in a block. To get HOG3D, both blocks and cells are extended to 3D, and gradient orientation is quantized using regular polyhedrons, which means there are only 5 kinds of quantization strategies (4-, 6-, 8-, 12-, 20-sided polygon). Take regular icosahedron (20-sided polygon) for example, let its center of gravity lie at the origin of a three-dimensional Euclidean coordinate system. Let $\bar{g}_b$ a 3D gradient vector, in order to quantize $\bar{g}_b$, we first project it on the axes running through the origin of the coordinate system and the center positions of all faces. Suppose $\mathbf{P}$ that is the matrix of center positions of the 20 faces, then $\mathbf{P} = (p_1, \cdots, p_{20})^{\mathrm{T}}$ and p_i stands for vectors defined below.

$$(\pm 1, \pm 1, \pm 1), (0, \pm\frac{1}{\varphi}, \pm\varphi), (\pm\frac{1}{\varphi}, \pm\varphi, 0), (\pm\varphi, 0, \pm\frac{1}{\varphi}) \tag{1}$$

Where, $\varphi = \frac{1+\sqrt{5}}{2}$ is the golden ratio. The steps for calculating HOG3D (for each block) are as follows [16]:

1) Calculate the projection $\hat{q}_b$ of $\bar{g}_b$ as follows:

$$\hat{q}_b = \frac{P \cdot \bar{g}_b}{\|\bar{g}_b\|_2} \tag{2}$$

2) $\hat{q}_b$ is processed as Formula 3 using threshold $t = p_i^{\mathrm{T}} \cdot p_j$ to vote $\bar{g}_b$ into only one bin. p_i , p_j are two neighboring axes, and when using a regular icosahedrons for quantization.

$$\hat{q}'_{bi} = \begin{cases} \hat{q}_{bi} - t, if \hat{q}_{bi} - t > 0 \\ 0, others \end{cases} \tag{3}$$

where, $\hat{q}'_{bi}$ is the ith component of $\hat{q}'_b$.

3) Calculate the gradient magnitude distribution as follows:

$$q_b = \frac{\|\bar{g}_b\|_2 \cdot \hat{q}'_b}{\|\hat{q}'\|_2} \tag{4}$$

4) The whole video is divided into many blocks, and the histogram for each block is obtained by summing the quantized mean gradients of cells it contains. The mean gradient for each cell is computed by Formula 4.

In order to enhance statistical character, K-means clustering is implemented on histograms extracted from blocks instead of the original HOG3D descriptor. It splices all histograms extracted from blocks to make a large histogram vector, and the acquired clustering centers are used to generate the vocabulary. The advantages of combining K-means with block level HOG3D descriptor are as follows. On one hand, it is convenient for transferring the violent scene detection issue into a MIL problem. On the other hand, this strategy will enhance the global statistical character for a larger scale of samples.

2.2. **MIL algorithms.** MIL was proposed by Dietterich et al., [19] to solve the drug activity prediction issue. In MIL, the training set is composed of bags and each bag contains many instances. Different to supervised learning where each training instance owns a known label, in MIL, labels are known only on bag level, and labels for the training instances are unknown. There is an assumption of MIL: a bag is labeled as positive if it contains at least one positive instance; otherwise, it is labeled as negative. As described in Section 1, we form violent scene detection as a MIL issue by treating video clips as bags in MIL and frames in the same unit for HOG3D extraction as instances for bags. The most significant advantage of detecting violent scenes using MIL as a classifier is that feature fusion for the whole video can be avoided, and fixed-dimensional feature is no longer a required input. In our work, simulation experiments were done using three typical MIL algorithms named Citation-kNN, Axis-Parallel hyper Rectangle (APR) and mi-SVM respectively. More details about them are described below. Citation-kNN is a nearest neighbor algorithm that not only takes a bags neighboring bags into account when making a label for it but also analyzes the labels of bags which regard it as a neighbor [20]. The main idea of Citation-kNN is that an improved Hausdorff distance is used instead of traditional Euclidean distance. Suppose $A = \{a_1, a_2, ..., a_n\}$ and $B = \{b_1, b_2, ..., b_m\}$ are two bags, $a_i, b_j \in R^d$, $i = (1, 2, ..., n), j = (1, 2, .., m)$ stand for instances in bags A and B respectively, then the distance $minHD(A, B)$ between bags A and B is defined as follows [20]:

$$minHD(A, B) = \max\{h(A, B), h(B, A)\} \tag{5}$$

where $h(A,B) = \min_{a\in A}\min_{b\in B}\|a-b\|$. Dietterich et al. [19] proposed Axis-Parallel hyper Rectangle (APR) algorithms when raising multi-instance learning problem, and developed three design strategies for APR which are standard APR, outside-in algorithm and inside-out algorithm. The standard APR algorithm just forms the smallest APR that bounds the positive instances. The outside-in algorithm is called GFS elim-kde APR. It constructs the smallest APR that bounds all the positive instances and shrinks the APR to exclude false positives. The inside-out algorithm first finds a seed point and makes a rectangle growing with the goal of finding the smallest rectangle that covers at least one instance of each positive example and no instance of negative examples using greedy back-fitting strategy. Experimental results showed that inside-out algorithm was better than the standard APR and the outside-in algorithm on drug activity problem [19]. Andrew et al. [21] improved SVM to make it fit for MIL. One of SVM based MIL algorithm is mi-SVM and it uses the assumption that all instances in a negative bag are negative. The goal of mi-SVM is to find an optimal hyper plane to make the margin maximum. Bag label constraints are added into the SVM model, and the optimal model for mi-SVM is defined as Formula 6 [21].

$$\begin{aligned}
&\min_{y_i}\min_{w,b,\varepsilon}\frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i \\
&\forall i : y_i(< w, x_i > +b) \geq 1-\varepsilon_i, \varepsilon_i \geq 0, y_i \in \{-1,1\} \\
&\sum_{i\in I}\frac{y_i+1}{2} \geq 1, \forall I \ s.t.\ Y_I = 1 \ and \ y_i = -1, \ \forall I \ s.t.\ Y_I = -1
\end{aligned} \tag{6}$$

where, w, b are coefficients and intercept for hyper plane respectively, ε_i is slack variable, x_i, y_i are instance and its corresponding label, I, Y_I indicate bag and its label.

```
Init:
Maxloop = 10;
ISet (set of instance);
M = total number of instance in ISet;
N = 0;loop = 1;
While M != N
    If loop > Maxloop
      Break;
    End
    loop = loop +1;
    For inst=ISet(1):ISet(M)
      If label(inst) equals negative
        target = argmin |inst-target|
        If label(target) != label(inst)
          Remove inst from ISet;
        End
      End
    End
    N = number of instance in ISet;
End
Where label(inst) is a function return the label of bag contains instance inst.
```

2.3. **Instance clean method.** The basic assumption of MIL is that a bag is labeled positive if it contains at least one positive instance; otherwise it is labeled as a negative bag [19]. However, it is hard to satisfy this assumption because the sample data usually have noise. On the other hand, the contribution of negative and positive instance to a bag is asymmetric, which means the label of a bag remains unchanged when removing or adding one negative instance from it. In addition, video is composed of frames, and

FIGURE 4. Samples of violent (top) and normal (bottom) video frames from Hockey

FIGURE 5. Samples of violent (top) and normal (bottom) video frames from Movies.

neighboring frames more likely have much visual similarity. As a result, instances from neighboring frame sequences more likely own the same label. Therefore, an instance clean approach is proposed in our algorithm to remove mostly fake negative instances from negative bag, and to remove instances whose nearest neighbor belongs to a negative bag from positive bag. The key idea of the proposed instance clean approach is to remove negative instance whose hidden label is incorrect with large probability. The side effect of the proposed method is that some helpful information may also be removed while removing potential false negative instances, especially when the input video clips are not long enough. The pseudo-code for instance clean method is as follows:

3. **Experimental Results and Analysis.**

3.1. **Datasets.** The performance evaluation of this method was done in Matlab R2011b on two datasets: Hockey [8, 11, 15] and Movies [8, 11]. The Hockey video set is composed of 1000 clips from Hockey competition. 500 of the videos are labeled as violent (contains fighting) and the others are normal. Each video contains 41 frames and the resolution of video frame is 360×288 . The frame rate is 25f/s. Fig.4 shows samples of violent and normal frames from Hockey videos. The Movies dataset contains 100 violent video clips and 100 normal video clips which are collected from action movies and non-action movies. Resolution of most videos in Movies is 720×576 and the others are 720×480 . The frame rate is 25 f/s and total frame number for each video ranges from 10 to 60. Some sample frames are shown in Fig. 5. The first rows of Fig.4 and Fig. 5 indicate violence video frames and the second rows are normal ones.

TABLE 1. Accuracy of Citation-kNN on Hockey and Movies datasets with or without data clean processing (%)

Word Number	Hockey	Cleaned Hockey	Movies	Cleaned Movies
50	0.883	0.921	0.988	0.988
100	0.895	0.922	0.989	0.984
200	0.905	0.927	0.982	0.983
500	0.906	0.920	0.988	0.986
1000	0.885	0.913	0.989	0.979

TABLE 2. Accuracy of APR on Hockey and Movies datasets with or without data clean processing (%)

Word Number	Hockey	Cleaned Hockey	Movies	Cleaned Movies
50	0.559	0.587	0.900	0.890
100	0.576	0.623	0.850	0.855
200	0.639	0.681	0.820	0.825
500	0.653	0.671	0.735	0.725
1000	0.622	0.642	0.68	0.685

TABLE 3. Accuracy of mi-SVM on Hockey and Movies datasets with or without data clean processing (%)

Word Number	Hockey	Cleaned Hockey	Movies	Cleaned Movies
50	0.741	0.778	0.900	0.900
100	0.732	0.773	0.910	0.910
200	0.771	0.804	0.915	0.905
500	0.769	0.798	0.910	0.905
1000	0.776	0.780	0.905	0.905

3.2. **Performance analysis of instance clean approach.** Table 1-3 show the comparison experimental results on Hockey and Movies with or without instance clean step using Citation-kNN, APR and mi-SVM, respectively. The effect of instance clean processing was studied when the vocabulary size equals 50, 100, 200, 500 and 1000 respectively, and the table described the detection accuracy on Hockey and Movies datasets with or without instance clean step. The experimental results showed that instance clean could universally improve detection accuracy on Hockey among all three MIL methods. A maximum improvement of 4% (from about 88% to 92%) was achieved when vocabulary was 50 and classifier was Citation-kNN.

On Movies dataset, however, the detection accuracy slightly declined after instance clean strategy was applied. As described above, video clips in Movies dataset do not have a fixed length. In fact, some video clips only contain 10 frames, which probably mean that their corresponding bag may contain only 1 instance. In this case, the probability of incorrect instance removal could be increased, leading to a loss of useful information. From the results on both datasets, the proposed instance clean approach was effective, especially when the video clips were long enough.

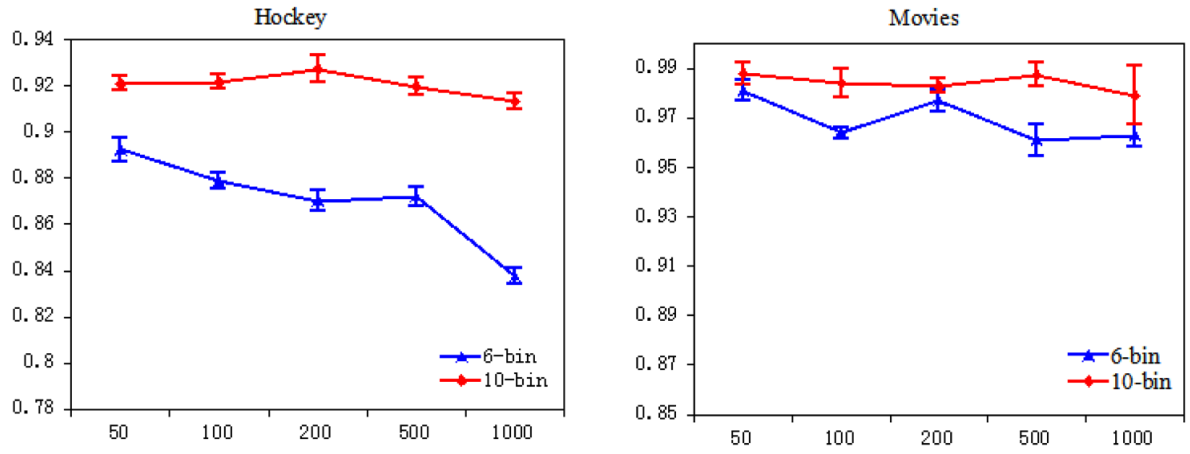

FIGURE 6. Accuracy of Citation-kNN on Hockey and Movies using different bin value for HOG3D

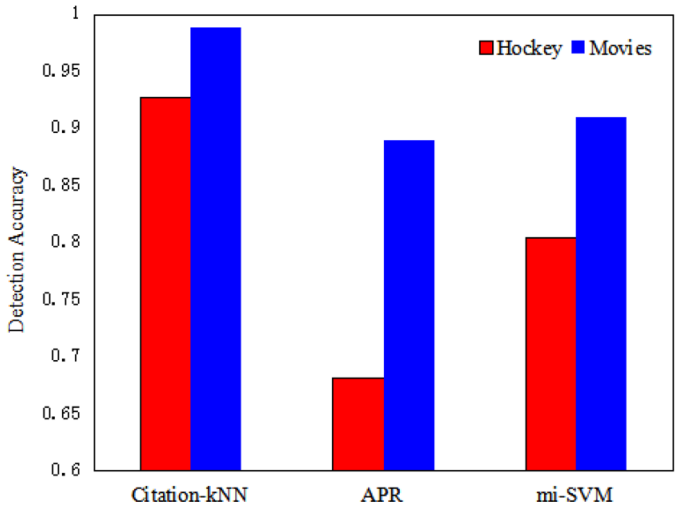

FIGURE 7. Accuracy of Citation-kNN, APR, mi-SVM on Hockey and Movies datasets

3.3. **Evaluation of parameter and experimental results.** In the proposed approach, HOG3D descriptor and BoVW framework were used to construct the final feature. We first set the parameters for number of words in vocabulary and bin number for gradient quantization in HOG3D. We chose dodecahedron and icosahedrons which corresponded to 6 and 10 bins respectively for gradient quantization, and generated vocabularies of 50, 100, 200, 500 and 1000 words. Accuracy curves obtained by Citation-kNN are shown in Fig. 6. For both Hockey and Movies, higher detection accuracy was obtained by using 10-bin for gradient quantization.

Especially on Hockey, the accuracies were 92.12%, 92.20%, 92.73%, 91.99%, 91.35% corresponding to 50, 100, 200, 500 and 1000 words in vocabulary when bin number was 10. And the accuracies were 89.26%, 87.89%, 87.03%, 87.20%, 83.76% when bin number was 6. Fig. 7 is a performance comparison of Citation-kNN, APR and mi-SVM. The accuracies of Citation-kNN, APR and mi-SVM were 92.73%, 68.15% and 80.44% respectively on Hockey dataset, and 98.8%, 89% and 91.5% respectively on Movies. Therefore, Citation-kNN achieved the best detection accuracy compared with the other two MIL methods both on Hockey and Movies. Fig. 8 is the ROC curve of Citation-kNN, which shows the effectiveness of the proposed approach on the two datasets.

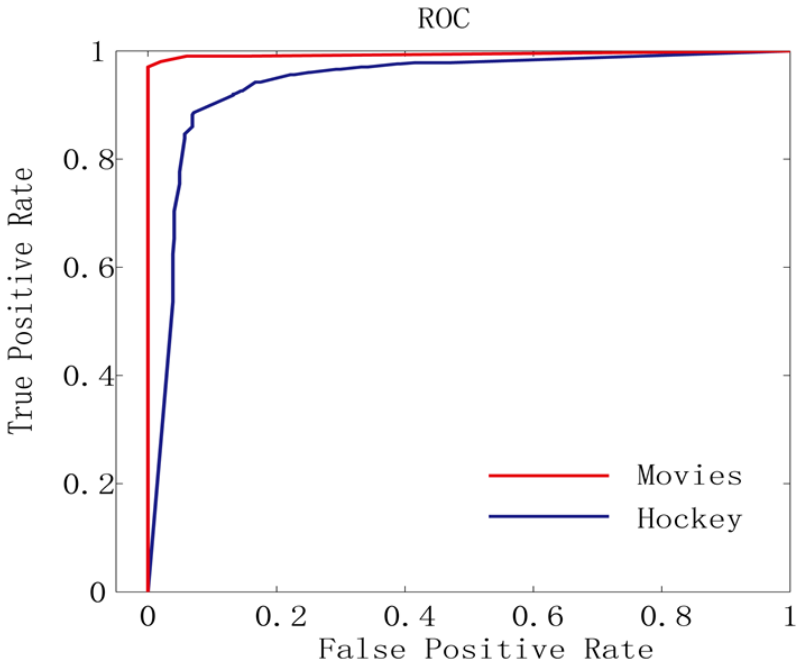

FIGURE 8. ROC curve of Citation-kNN

3.4. **Comparison to state-of-the-art.** Further comparison to state-of-the-art was done to evaluate the performance of the proposed method. The average precision on Hockey and Movies using various methods is shown in Table 4. The result of our hybrid method (92.73%, average of 5 runs of 10-fold cross-validation) on Hockey outperformed the state-of-the-art (91.7%) which used HOG descriptor and SVM with a histogram intersection kernel (HIK) [8]. The method using 3D CNN construction was slightly underperformed compared with the former two, which was 91% [15]. On Movies dataset, Deniz et al. proposed extreme acceleration patterns and Adaboost for violence detection [11], and obtained the best accuracy (98.9%) among all listed methods. Our hybrid method matched this performance with an accuracy of 98.8% (average of 5 runs of 10-fold cross-validation), which was much higher than all the other listed methods. And the average accuracy on two dataset is 95.75%, which is the highest one among these schemes, and this demonstrate that the proposed algorithm has good universality. Together, we can come to the conclusion that our proposed method using HOG3G and MIL is most effective for violent scene detection with a balanced accuracy on Hockey and Movies compared with all state-of-the-art methods.

4. **Conclusion.** In this paper, after synthesizing the good characteristics of multiple instance learning and 3D histogram of oriented gradients, a novel violent scene detection algorithm is put forward. In order to reduce noise in training and testing data, an instance clean approach is implemented before training the MIL classifier, and the experiments showed that it is effective. Three different MIL algorithms were studied on two datasets, and the results demonstrate that Citation-kNN is more suitable than the other two tested MIL methods. Performance comparisons with the existing schemes further demonstrate the effectiveness of the proposed approach. In the future, we will extend our algorithm to construct better descriptor for video by concentrating other features to improve algorithm's performance.

TABLE 4. Average precision on Hockey and Movies datasets(%)

Methods	Hockey	Movies	Average
HOG + SVM(HIK)[8]	91.7	49	70.35
HOF + SVM(HIK)[8]	88.6	59	73.8
HOG + SVM[11]	88.5	82.5	85.5
HOG + Adaboost[11]	86.5	74.5	80.5
MoSIFT + SVM[11]	91.2	84.2	87.7
MoSIFT + Adaboost[11]	89.5	86.5	88
Extreme acceleration +SVM [11]	90.1	85.4	87.75
Extreme acceleration +Adaboost [11]	90.1	**98.9**	94.5
3D-CNN[15]	91	----	----
Proposed Method	**92.7**	98.8	**95.75**

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China project (61503424); 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(YZK2015013, CJGX2016.7.8-SZ-04.5.6, CJGX2016.7.8-SZJC-03.4.5, 02362050301).

REFERENCES

[1] A. Balachandran, V. Sekar, A. Akella, et al., Developing a predictive model of quality of experience for internet video, *ACM SIGCOMM Computer Communication Review. ACM*, vol. 43, no.4, pp.339-350, 2013.

[2] S. W. Zhao, L. Zhuo, S. Y. Wang, et al. Research on key technologies of pornographic image/video recognition in compressed domain, *Journal of Electronics (China)*, vol.26, no.5, pp. 687-691, 2009.

[3] J. C. wang, B Li, W. M. Hu, et al. Horror movie scene recognition based on emotional perception, *Image Processing (ICIP)*, 2010 17th IEEE International Conference on. IEEE, pp.1489-1492, 2010.

[4] J. Nam, M. Alghoniemy, A. H. Tewfik, Audio-visual content-based violent scene characterization, *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on. IEEE*, vol.1, pp. 353-357, 1998.

[5] W. H. Cheng, W. T. Chu, J. L. Wu. Semantic context detection based on hierarchical audio models, *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval. ACM*, pp. 109-115, 2003.

[6] Y. Gong, W. Q. Wang, S. Q. Jiang, et al. Detecting violent scenes in movies by auditory and visual cues, *Advances in Multimedia Information Processing-PCM 2008. Springer Berlin Heidelberg*, pp. 317-326, 2008.

[7] T. Giannakopoulos, A. Makris, et al. Audio-visual fusion for detecting violent scenes in videos, *Artificial Intelligence: Theories, Models and Applications. Springer Berlin Heidelberg,* pp. 91-100, 2010.

[8] E. B. Nievas , O. D. Suraez , G. B. Garcia, et al. Violence detection in video using computer vision techniques. *Computer Analysis of Images and Patterns. Springer Berlin Heidelberg*, pp. 332-339, 2011.

[9] J. Yang, Y. G. Jiang, A. G. Hauptmann, et al. Evaluating bag-of-visual-words representations in scene classification, *Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM* , pp. 197-206, 2007.

[10] M. Y. Chen, A. Hauptmann MoSIFT: Recognizing Human Actions in Surveillance Videos, CMU-CS-09-161, Carnegie Mellon University, http://www.cs.cmu.edu/ mychen/publication/ChenMoSIFTCMU09.pdf, 2009.

[11] O. Denz , I. SerranoO ,G. Bueno, et al., Fast Violence Detection in Video. The 9th *International Conference on Computer Vision Theory and Applications* (VISAPP), pp. 478-485, 2014.

[12] G. E. Hinton, R. R. Salakhutdinoy, Reducing the dimensionality of data with neural networks. *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[13] X. W. Chen, X. T. Lin, Big data deep learning: Challenges and perspectives, *Access, IEEE*, vol. 2, pp. 514-525, 2014.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning, *Nature*, vol. 518, no.7540, pp. 529-533, 2015.

[15] C H Ding, S K Fan, M Zhu, et al. Violence Detection in Video by using 3D Convolutional Neural Networks, *Advances in Visual Computing. Springer International Publishing*, pp. 551-558, 2014.

[16] A. Klaser , M. Marszalek ,C. SCHMID, A spatio-temporal descriptor based on 3d-gradients, *BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association*, vol.275, pp. 1-10, 2008.

[17] N. Dalal , B. Triggs, Histograms of oriented gradients for human detection, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE*, vol. 1, pp. 886-893, 2005.

[18] Z. Gao, H. Zhang, A. A. Liu, et al., Human action recognition using pyramid histograms of oriented gradients and collaborative multi-task learning, *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 8, no. 2. Pp. 483-503.

[19] T. G. Dietterich, R. H. Lathrop , et al., Solving the multiple instance problem with axis-parallel rectangle, *Artificial intelligence*, vol. 89, no.1, pp. 31-71, 1997.

[20] J. Wang, J. D. Zucker, Solving multiple-instance problem: A lazy learning approach, *in Proc. of international Conference on Machine Learning*, pp. 1119-1125, 2000.

[21] S. Andrews , T.Hofmann, I. Tsochantaridis, Multiple instance learning with generalized support vector machines, *AAAI/IAAI,* pp. 943-944, 2002.

[22] H. Shim, S. Lee, Multi-channel electromyography pattern classification using deep belief networks for enhanced user experience, *Journal of Central South University*, vol. 22, no. 5, pp. 1801-1808, 2015.