International Journal of Computational Intelligence and Applications
Vol. 11, No. 2 (2012) 1250010 (21 pages)
© Imperial College Press
DOI: 10.1142/S1469026812500101

Imperial College Press
www.icpress.co.uk

A NOVEL VIOLENT VIDEOS CLASSIFICATION SCHEME BASED ON THE BAG OF AUDIO WORDS FEATURES

LEI LI

Computer Science Department
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
leili.cmu@gmail.com

Received 9 November 2011
Revised 20 February 2012
Published 28 July 2012

A novel method to identify the violent videos only with audio features is introduced. Most previous content-based image or video classification schemes apply the bag of words (BOW) or bag of visual words (BOVW), which employ multiple visual features to characterize image or video content. In our method, the bag of audio words (BOAW) is suggested to be built by effective audio features. Two reasons are considered here. First, audio features should have very special significance for violent videos. Second, the computational complexity of dealing with audio features is much lower than that of visual features. The MPEG-7 low level features such as AudioSpectrum-Centroid and AudioSpectrum-Spread, and the high level feature such as AudioSignature, are combined into one 44-dimensions vector in the BOAW model. The audio words are built from the vector by the clustering strategy, and support vector machine (SVM) with revised soft-weighting scheme is used to group the audio words features into two classes, i.e., the violent and nonviolent. Experiments demonstrate that the proposed method can achieve good recall accuracy and precision accuracy on detecting violent videos. The method also can be applied to classify other types of videos.

Keywords: Violent videos classification; MPEG-7 audio descriptor; multi-layer perceptron; bag of audio words; soft weighting scheme.

1. Introduction

With the explosively growing amount of multimedia information made available in digital form, many types of videos are available on the internet. The videos with some violent or sexual scenes are not suitable for children, and should be considered under some internet information control policies. The content-based video classification would play more and more important role in this field. Humans can easily figure out different genres of videos by just watching them. However, for the computer, it is quite complicated to automatically recognize the theme of a video. Many methods are proposed to make the compute recognize image automatically,[1] and some content-based image retrieval systems have been discussed.[2]

One common approach of video categorization is to use the global features extracted from the video. Many researches[3–10] show that the combination of various features is an effective way to describe the characters of a video. The basic idea is to extract the visual features, and it is obviously better than just utilizing text features.[11] Huang *et al.*[8] pay their effort on the contribution of low-level visual features combination for the film classification. Their research shows that combining low-level features such as average shot length, color variance, motion content and lighting key with cinematic principles can be powerful tools for video classification. Perronnin[12] combines the low-level descriptors such as the popular SIFT (Scale Invariant Feature Transform) and high-level histograms of word co-occurrences in a spatial neighborhood.

Audio features can also play an important role in video classification. In Ref. 13, the authors utilize the fusion of audio-visual features to characterize the video semantics according to human perceptual features. Cai *et al.*[14] focus on the problem of automatically categorizing audio scenes in unsupervised manner. They characterize the auditory scenes with either low-level acoustic features or some mid-level representations such as audio effects. Aucouturiera[15] presents the "bag of frames" approach for audio pattern recognition. They treat the total audio spectrum as a whole rather than extracting frame-based spectral feature, and find that this method performs well in urban soundscapes. In Ref. 15, some basic audio features such as audio power and MFCC (Mel Frequency Cepstrum Coefficients) are used to classify the audios. The average classification accuracy is about 80%.

Recently many papers[7,10,16–18] focus on the research of the bag of words (BOW) model in the image and video classification. Originating from the text categorization area, BOW has become a popular method for visual categorization because of its effectiveness and flexibility.[17] In the BOW method, the features are extracted from the video frames, and then are clustered into visual words first. Then the frequencies of visual words for an image or video can be used in the subsequent classification process.[10] Based on the traditional BOW method, Li *et al.*[17] propose a visual categorization approach with using a novel contextual bag of words (CBOW) which can model two kinds of typical contextual relations between local patches: a semantic conceptual relation and a spatial neighboring relation. Similarly, both in Refs. 16 and 18, the authors propose the methods for scene categorization by integrating contextual information into the popular BOW approach. Koike and Takagi[19] study the application of BOW in the biomedical figures. The classification accuracy of bag of key-points is only about 20%, which is better than that of field-level image descriptors. When BOW for legends is combined with bag of key-points, the prediction performance achieved 75.7% classification accuracy. Weizman and Goldberger[20] use the bag of visual words (BOVW) to detect the urban zones in satellite images. They build 60 words from 127,490 figures and get a reasonable result.

Supervised classifiers are utilized to categorize the videos based on various video features. Support vector machine (SVM) is the most popular classifier to learn the relationship of information. Some statistical models like Bayesian Network[9] and

Neural Network[21] have also been employed to learn the relationships between video features and semantics. Besides, combined AdaBoost classifier[4,5] which utilizes an ensemble of multiple weak learners to build a stronger classifier can be used to interactively exploit the intrinsic relationship of information. The supervised classification methods mentioned above have been proved to be effective and powerful in capturing the available helpful knowledge and modeling complex mappings between features and semantics. Clustering theory provides the most intuitive framework for grouping similar videos into a semantic category in an unsupervised manner.[14] K-means is a traditional one-way clustering algorithm in which similarities are estimated by measuring the distances among the relevant points in the feature space, assuming that each feature gives equal contribution to the distance measure. In Ref. 10, the authors compare the k-means visual vocabulary with Random Forest, and evaluate subsample, dimension reduction with PCA, and division strategies of the Spatial Pyramid in the step of Visual Word Assignment.

In the past five years, more and more researchers focus the research on violent content detection. Zajdel *et al.*[22] propose a smart surveillance system named CASSANDRA for detecting instances of aggressive human behavior in public environments. In Ref. 23, a multi-class classification algorithm for audio segments from movies has been proposed. In this paper Bayesian Networks and the One Versus All classification architecture are combined to classify the audio segments into six classes, in which three classes are related to the violent content. Mecocci and Micheli[24] propose an approach based on global chromatic features extracted from moving object in the video stream. Gong *et al.*[25] present a three-stage method with integrating visual and auditory cues. In Ref. 26, a violence detector built on the concept of visual codebooks is proposed, in which BOVW has been adopted to represent the spatio-temporal features and support vector machine (SVM) has been used to supervise learning. In Ref. 27, a fusion of three modalities: audio, moving image and text data is used to detect violent content in video sharing sites. As the methods mentioned above, visual features are usually used as the essential elements to represent the violent content, and audio features can be combined with visual features to enhance the detection performance.

In this paper, we propose to use only audio features to recognize the MPEG-7 violent videos. The bag of audio words (BOAW) is suggested to be built by effective audio features as a novel descriptor. Two reasons are considered here. First, audio features should have very special significance for violent video. Second, the computational complexity of dealing with audio features is much lower than that of visual features. In our method, the audio words are extracted from the typical violent audio scenes, including the MPEG-7 low level features such as AudioSpectrum-Centroid and AudioSpectrum-Spread, and the MPEG-7 high level feature such as AudioSignature. AudioSignature is considered as the "fingerprint" of an audio stream,[28] which can indentify an unknown piece of audio based on a database of registered audio items. SVM with revised soft weighting scheme is used to group the audio words feature vectors into two classes, i.e., the violent and nonviolent. The experiments have shown that the accuracy of the method is satisfactory and the computational time of dealing with audio features is shorter than that of visual features.

In Sec. 2, the MPEG-7 audio features used in our method are introduced briefly. The construction of BOAW and our classification scheme are investigated in Sec. 3. Section 4 describes the experiments and presents some classification experimental results. We conclude with a brief discussion of our work and some future work in Sec. 5.

2. MPEG-7 Audio Features

MPEG-7,[29,30] formally named "multimedia content description interface", is an ISO/IEC standard developed by MPEG Moving Picture Experts Group in September 2001. MPEG-7 is an audiovisual information representation standard for describing multimedia content data that supports some degree of interpretation of the information meaning. Besides support for metadata and text descriptions of the audiovisual content, the audio and video parts provide standardized "audio only" and "visual only" descriptors. The multimedia description schemes (MDS) part provides standardized description schemes involving both audio and visual descriptors. The description definition language (DDL) provides a standardized language to express description schemes, and the systems part provides the necessary glue that enables the use of the standard in practical environments.

2.1. *MPEG-7 audio low level features*

The MPEG-7 low level audio descriptors are of great importance in describing audio. There are 17 temporal and spectral descriptors that may be used in a variety of applications. These descriptors can be automatically extracted from audio and depict the variation of audio properties over time or frequency. Based on these descriptors, it is often feasible to analyze the similarity between various audio files. Thus it is possible to be used to identify identical, similar or dissimilar audio content. Some low level features used in our method are briefly introduced as follows: (1) AudioSpectrum-Centroid Type (ASC): ASC is defined as the center-of-gravity of a log-frequency power spectrum. This definition is adjusted in the extraction to take into account the fact that a nonzero DC component creates a singularity, and eventually very-low frequency components (possibly spurious) have a disproportionate weight; (2) AudioSpectrum-Spread Type (ASS): ASS is another simple measure of the spectral shape. The spectral spread, also called instantaneous bandwidth. In MPEG-7, it is defined as the second central moment of the log-frequency spectrum. For a given signal frame, the ASS feature is extracted by taking the root-mean-square (RMS) deviation of the spectrum from its centroid; (3) Audio-Fundamental-Frequency Type (AFF): AFF is a good predictor of musical pitch and speech intonation. Hence it is an important descriptor of an audio signal; (4) Audio-Harmonicity Type (AH): AH descriptor provides two measures of the harmonic properties of a spectrum: the harmonic ratio, i.e., the ratio of harmonic power to total power; and the upper limit of harmonicity, i.e., the frequency beyond which the spectrum cannot be considered as harmonic. They both rely on a standardized fundamental frequency estimation method, which is based on the local normalized autocorrelation function of the signal. This approach, widely used for local pitch estimation, is independent of the extraction of AFF presented above.

2.2. *High level feature*: *AudioSignature*

The AudioSignature high level feature is adopted as a representation of an audio signal and designed to provide a unique content identifier for robust automatic identification. It is a compact-sized audio signature which can be used as a "fingerprint".

The AudioSignature descriptor mainly consists of a statistical summarization of the low level features over a period of time, which can be extracted for the low level descriptor AudioSpectrum-Flatness (ASF) by using the tool "MPEG-7 audio encoder". The MPEG-7 ASF descriptor can be extracted as shown below:

For each bandwidth b, a spectral flatness coefficient is estimated as the ratio between the geometric mean and the arithmetic mean of the spectral power coefficients within this band:

$$\mathrm{ASF}(b) = \frac{hb - lb + 1\sqrt{\prod_{k=lb}^{hb} P(k)}}{\frac{1}{hb-lb+1}\sum_{k=lb}^{hb} P(k)}, \tag{1}$$

where hb and lb are the nearest integer of the bandwidth b, and $P(k)$ is the power spectrum extracted from the kth frame of the input digital audio signal.

3. Classification Scheme Based on BOAW

3.1. *Bag of audio words* (*BOAW*)

BOW was first applied in the Text Categorization (TC).[31] In TC, some words were built first. After that, the document can be represented by these words with their frequencies. So the document can be easily indexed by the words frequencies in the database.

Similar to the words in the text, an image has its local interest points or key-points defined as salient image patches (small regions) that contain rich local information of the image. By the Difference of Gaussian (DOG),[19] we can detect the key-points in the image. Using some feature extraction of the key-points (i.e., the SIFT method), we can build the BOVW.[20] An image can be represented by these words. While different genre of image has different words, we can classify the image by its own words.

We can also define the words in the audio spectrum. On the one hand, we should find the basic cell of the audio spectrum like the words in the text and the key-points in the image. The audio words should have some semantic meaning of the audio. On the other hand, it should be detected by the audio frame rather than the whole audio spectrum. Fortunately, the MEPG-7 high level feature, AudioSignature, can do this perfectly. It can be detected through the audio frame (often 10 ms or 30 ms as a frame), and it is considered as the fingerprint of the audio because of its good performance in audio classification.

The main difference between the words of the text and of the audio is: text words are sampled naturally according to language context, while audio words are the outputs of data clustering. Text words carry semantic meaning, while audio words infer statistical information.

In our method, it is not necessary to build audio words from all the audio features. The MPEG-7 low level features such as ASC, ASS, AFF and AH, and the high level feature such as AudioSignature, are combined into one vector in the BOAW model. These features are mentioned in Sec. 2. The audio words of the semantics are built from the vector by the clustering strategy. The K-means clustering algorithm as shown below[32]:

Given an initial set of k-means $m_1, m_2, \ldots, m_k$, which is specified randomly from the vectors. The algorithm proceeds by alternating between Assignment step and Update step, and the algorithm is deemed to have converged when the assignments no longer change.

Assignment step: Assign each observation x_j to the cluster with the closest mean.

$$S_i^{(t)} = \{x_j : \|x_j - m_i^{(t)}\| \leq \|x_j - m_{i^*}\|\}, \quad \text{for all } i^* = 1, \ldots, k. \tag{2}$$

Update step: Calculate the new means as the center of the observations in the cluster.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{X \in S^{(t)}} x_j. \tag{3}$$

In our method, we build the words through the typical violent audio scenes, including fight, gunplay, explosion, etc. So all the words we defined can represent some character of the violent scenes.

3.2. *Soft-weighting scheme*

Term weighting is known to have critical impact in text information retrieval. The weighting scheme is also considered in the image retrieval based on the visual words. Since the audio words are the result of the data clustering, the weighting scheme is very important for the whole classification scheme.

In our method, we use the soft-weighting scheme which is proposed by Jiang and Ngo.[33] The soft-weighting scheme as shown below:

For each audio shot, instead of searching only for the nearest visual word, we select the top-N nearest visual words. Suppose there is a visual vocabulary of K visual words, we use a K-dimensional vector $T = [t_1, t_2, \ldots, t_K]$ with each component

t_k representing the weight of a visual word k, such that

$$t_k = \sum_{i=1}^{N} \sum_{j=1}^{Mi} \frac{1}{2^{i-1}} \text{ similarity } (j, k), \tag{4}$$

where M_i represents the number of key-points whose ith nearest neighbor is visual word k. The measure similarity (j, k) represents the similarity between key-point j and visual word k. Note that in Eq. (4), the contribution of a key-point is dependent on its similarity to word k weighted by $1/2^{i-1}$, representing the word is its ith nearest neighbor. In our experiment, N is empirically set to 4.

Figure 1 shows the soft-weighting scheme of the words for one vector of features. The whole audio is composed by several audio shots, so the final feature defined as "bag of audio words" is the mixture of all the audio shots. The process is shown in Fig. 2.

In Fig. 2, the original audio is divided into several audio shots. For each shot, the feature is computed with the set of audio words. According to the "soft-weighting histogram", we can compute the BOAW feature by adding all the weights of each bin.

3.3. *Framework of our method*

In fact, it is very hard to establish the effective model of violent scenes. In Ref. 34, the author defines the violent videos by themselves. Since there is no uniform definition of the violent video, we have to try to define them by ourselves. According to the popular experience of violent videos, five typical scenes are chosen to define the scope of the violent videos in our experiments, such as screaming, fighting, explosion, gunplay and car-racing, as shown in Fig. 3. If the audio contains some of the five scenes, the audio would be classified as a violent video. The five scenes are maybe not enough to represent all the scope of violent videos, but it is still reasonable and useful to draw some positive conclusions.

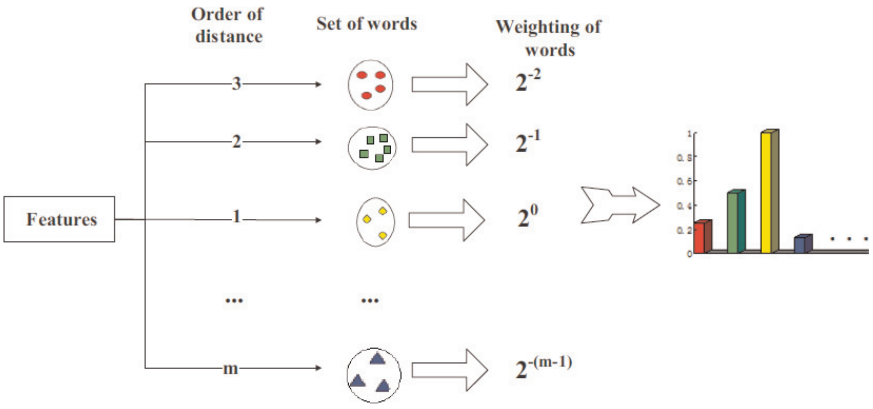


Fig. 1. The weighting scheme of the words.

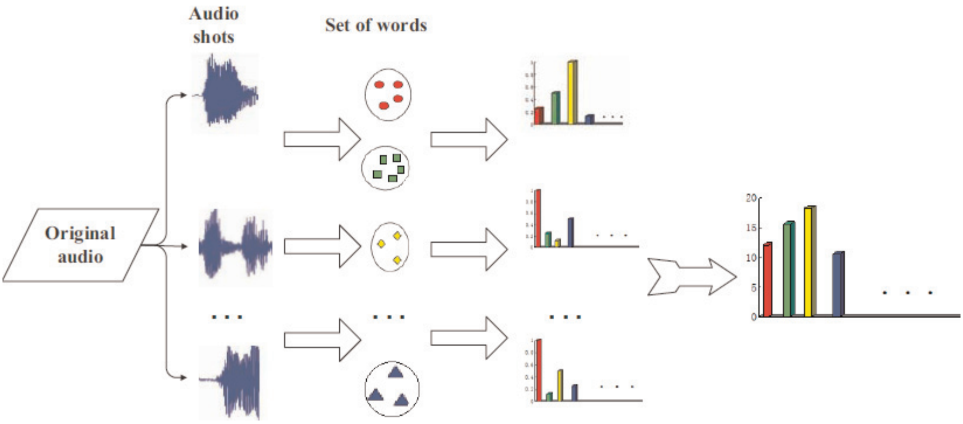


Fig. 2. Bag of audio words of the whole audio.

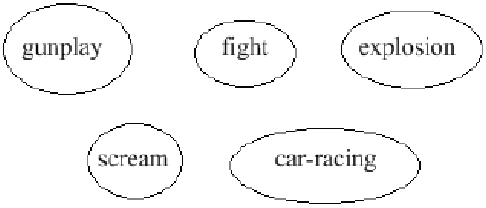


Fig. 3. Typical violent audio scenes.

In our experiments, the five typical violent audio scenes are chosen according to the definitions as follows: (1) *Gunplay*: when gun fires in the scene of war, polices, and robbers etc; (2) *Fight*: the voices appear when two or several conflict with each other by empty-hand or fighting tools such as stick, sword and harmer etc; (3) *Explosion*: a release of mechanical, chemical, or nuclear energy in a sudden and often violent manner with the generation of high temperature and usually with the release of gases; (4) *Scream*: a loud piercing sound coursed by horror, pain and fear etc; (5) *Car-racing*: when two or several cars race with each other, often going with noises of motors and brakes.

The whole framework of the proposed method as shown in Fig. 4, the testing flow is similar to the training flow. There is only one difference between them. In the training flow, we use the training features to make a cluster center file. This file is used to extract the BOAW in the training and testing process. While in the testing flow, we just only use the cluster center file extracted in the training process. More details are discussed below.

(1) Extract the audio stream from the original video: The original video has two streams: audio stream and video stream. Since our method is designed for the audio features, the audio stream is first extracted. There are a lot of tools that can do this task. In our method, we use "*ffmpeg*" tools to extract the audio stream.

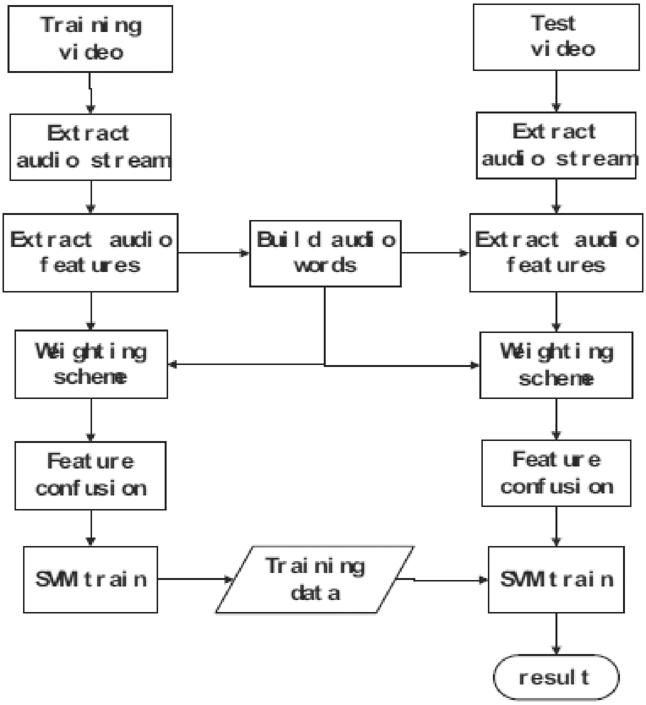


Fig. 4. The whole framework of our method.

(2) Extract the audio features of the audio stream: The tool named "*mpeg-7 audio encoder*" is used to extract the MPEG-7 audio features. Considering the complexity of the extraction, we specify some parameters for the process. For example, the low edge of the audio spectrum is set to 250 Hz while the high is 1600 Hz, and the decimation of the Audio flatness is set to 32.

After extracting the features of the audio stream, we need to reduce the dimension of the audio streams in order to reduce the computational complexity in the later process. The average of the vector of the ASC feature and the column average of the ASF are computed. The total dimension of the low level audio features is 12. For the high level feature, AudioSignature, we extract the 32-dimentions vector of each sample, with 16 dimensions for the mean of each sample and the other 16 dimensions for the variance.

(3) Fuse the features into one vector: The low level features and high level features are fused into one vector in the BOAW model. By the dimension reduction, the vector is of 44-dimensions. Then audio words are built from the vector by the clustering strategy and weighting scheme mentioned in Secs. 3.1 and 3.2.

The whole feature extraction module is shown in Fig. 5. The audio features are extracted from each audio frame, always a 10 ms audio frame. So an audio shot or sample includes many audio frames. Taking the time cost into consideration, we

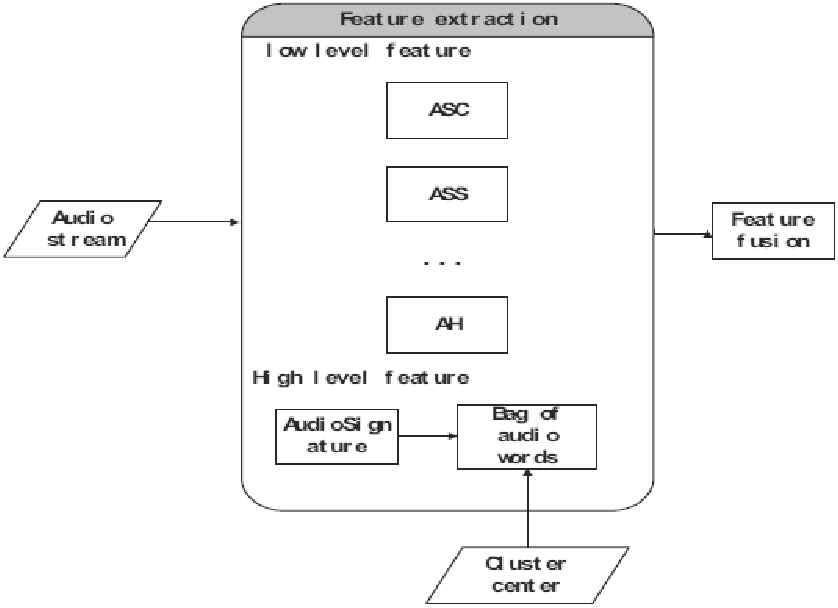


Fig. 5. The feature extraction module.

choose the mean of audio frames other than the cumulation of them. These features can represent the average of the whole sample.

(4) Using SVM classifier to do the classification: SVM is proved to be a good classifier used for classification and regression analysis. In our method, we adopt the 1-1 and 1-all SVM strategy to classify the different genres of videos. Furthermore, we adopt special SVM classification strategy as shown in Fig. 6. In this strategy, one video is divided into several shots, and audio features are extracted from each audio

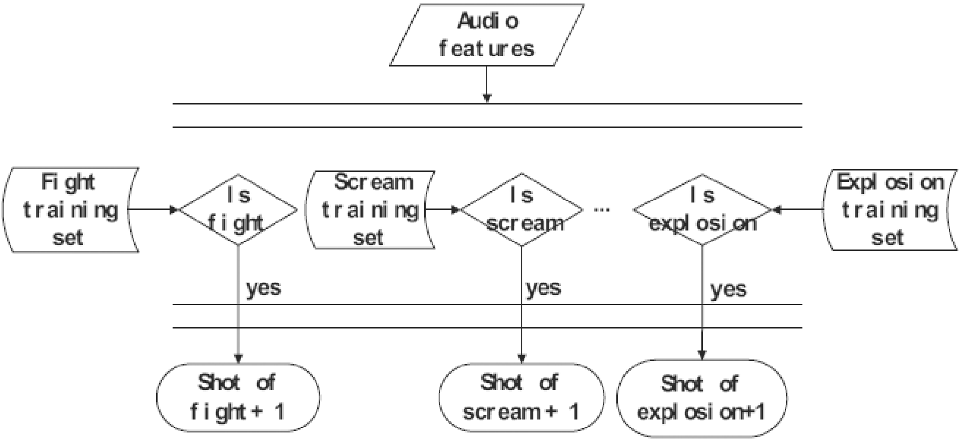


Fig. 6. SVM classification strategy.

shot. We classify each audio shot by 5 SVM training modules, i.e., fight, scream, explosion, gunplay and car-racing. Each shot can be classified into one or several of five types. The whole video is classified as a violent video if any of the violent shots exceeds a threshold of the whole shot numbers.

4. Experiments

Two classes of videos, including violent and nonviolent videos are collected in our experiment dataset. The nonviolent videos are collected from the TRECVID-2010 database.[35] Since there is no public violent video dataset built by the third party can be used, all the violent video samples are randomly chosen from our own violent video dataset. There are more than 100 famous violent videos (movies) in our dataset, such as "Terminator Salvation", "Kill Bill", etc. The total length of the videos is over 6 h. In our experiments, the videos are divided into samples, and the duration of each sample is about 1 min. There are totally 1342 samples in our test database in which 657 samples are nonviolent and the others are violent. We adopt the 1-all and 1-1 SVM strategy to make the classification.

Recall rate and precision rate can be used to analyze the total classification accuracy. For one certain category, the recall rate represents recognition accuracy of the positive samples which could be recognized as *True*. For one dataset, the precision rate represents recognition accuracy of the samples which could be recognized correctly. RECALL and PRECISION are defined as follow:

$$\mathrm{RECALL} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{5}$$

$$\mathrm{PRECISION} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}, \tag{6}$$

where TP is the number of positive samples which were recognized as *True* for one certain category; FP is the number of positive samples which were recognized as *False* for one certain category; TN is the number of samples which were recognized correctly from one dataset; FN is the number of samples which were recognized wrongly from one dataset.

The software tools we used in our experiments include *ffmpeg*, *MPEG-7 audio-encoder*, *python* and *libsvm*. We use the *ffmpeg* library to extract audio stream from the video, and *mpeg-7 audio encoder* to extract some MPEG-7 low level audio features. We deal with the xml files and fuse the features by *Python*, and classify the data through *libsvm* classifier.

4.1. *Classify five genres of videos by audio features*

Given five genres of videos are cartoon, entertainment, news, sports, and violent, we randomly choose 500 samples for training in which 300 are nonviolent and 200 are violent, and 842 samples in which 357 are nonviolent and 485 are violent for testing. In this experiment, the audio features extracted would be fused into one vector, and

Table 1. The result of classification for five genres of videos by SVM without weighting scheme.

	Results (Be Recognized as)						
Samples (Five Genres)	Cartoon	Entertainment	News	Sports	Violent	Total	Recall Rate
Cartoon	45	6	1	3	13	68	66.2%
Entertainment	10	84	2	21	30	147	57.1%
News	12	4	53	6	7	82	64.6%
Sports	4	6	9	23	18	60	38.3%
Violent	8	6	4	10	457	485	94.2%
Precision rate							
for each type	57.0%	79.2%	76.8%	36.5%	87.0%	—	—
for total	78.6%					—	—

the vector would be classified by SVM without weighting scheme. For the AudioSignature feature, we just compute the column average of each audio sample. Table 1 shows the performance result of video classification using only AudioSignature high level feature.

From Table 1, we can find that the total precision accuracy of the classification is 78.6%, but the violent has the highest classification accuracy. The major reason is that audio features of violent videos such as scream, gun fight and explosion have the very distinguishing characteristics of intense sound while the other categories do not have so much. Violent videos can be well recognized by audio features. For certain application purpose with information control policies, it should be more significant to group the samples into two types i.e., violent and nonviolent, so we can combine cartoon, entertainment, news, and sports into nonviolent type. The statistic result for only classifying videos into two types by SVM is shown in Table 2.

Contrast to Table 1, the total precision accuracy of the classification is 88.6%, which is satisfying for violent video detection. So we can conclude that our method is an effective way for violent video detection and MPEG-7 audio features make a significant contribution to violent content detecting.

In general five genres classification, we can find that a certain number of misclassification occurs between the sports and violent. As a result, the sports get the worst performance with the classification accuracy of 38.3%. The reason is that

Table 2. The result of classification for two types of videos by SVM without weighting scheme.

	Results (Be Recognized as)		
Sample (Two Types)	Violent	Nonviolent	Recall Rate
Violent	457	28	94.2%
Nonviolent	68	289	81.0%
Precision rate			
for each type	87.0%	91.2%	—
for total	88.6%		—

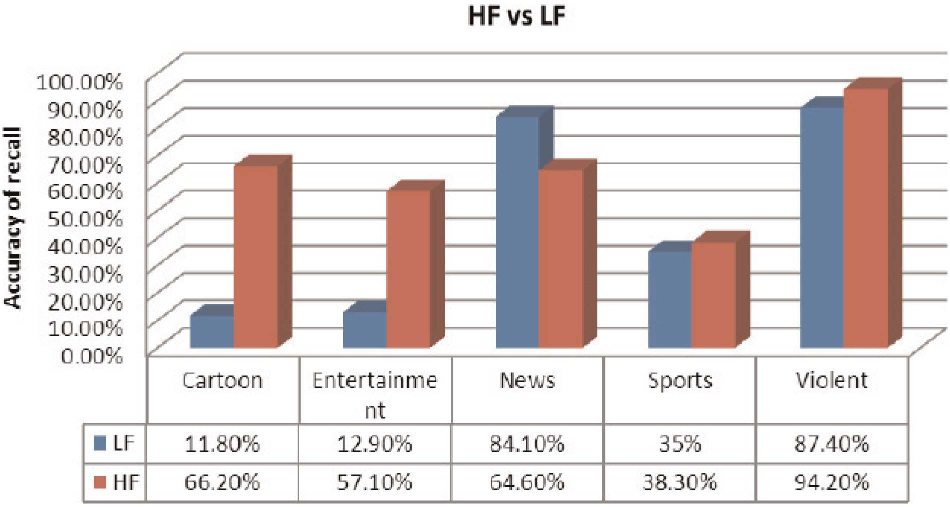
sports dataset includes Ball sports, Slide show, Bike tricks show, Horizontal bar shows, etc., which are always with the intense background sounds. The sounds mentioned above are similar to scream, gun fight and explosion scenes in violent videos. While for the nonviolent videos, the cartoon category performs the best. The reason is that in cartoon videos, the scenes are still, and the sound is always slow and soothing which is most different from the other four categories.

4.2. *Comparison between low level features and high level features*

This experiment is to evaluate the efficiency of the fused method with the low level audio features and the high level features. We classify the five genres of videos by two methods. One is to use only the low level features (LF) mentioned in Sec. 2.1, such as ASC, AFF and AH, etc. The other is to use the high level features (HF) based on AudioSignature mentioned in Sec. 2.2. We use the same training and testing samples to contrast the result, as shown in Fig. 7.

Figure 7 shows that for genres: cartoon, entertainment, sports and violent, the recall rate accuracy by high level features is higher than by low level features. For the genre of news, the recall rate accuracy by low level features is higher than by high level features. The total precision rate of high level features is 78.6%, while the low level features is 64.3%.

The low level features and high level features can be combined into one vector for classifying, which can be called the fused feature method mentioned in Sec. 3.3.3. For the classification of two types: violent and nonviolent, we use the three methods to take the experiment results, as shown in Table 3. The testing samples include 485 violent and 357 nonviolent videos all together.



	Cartoon	Entertainment	News	Sports	Violent
■ LF	11.80%	12.90%	84.10%	35%	87.40%
■ HF	66.20%	57.10%	64.60%	38.30%	94.20%

Fig. 7. The recall rate of five genres of videos detected by LF versus HF.

Table 3. The result of classification for two types of videos by three methods: LF, HF, the fused features, by SVM with soft-weighting scheme.

Classification Methods	Samples (Two Types)	Results (Be Recognized as)		Recall Rate
		Violent	Nonviolent	
LF	Violent	424	61	87.4%
	Nonviolent	65	292	81.8%
	Precision rate			
	for each type	86.7%	82.7%	—
	for total	85.0%		—
HF	Violent	457	28	94.2%
	Nonviolent	68	289	81.0%
	Precision rate			
	for each type	87.0%	91.2%	—
	for total	88.6%		—
Fused	Violent	449	36	92.6%
	Nonviolent	32	325	91.0%
	Precision rate			
	for each type	93.3%	90.0%	—
	for total	91.9%		—

From Table 3, the recall accuracy of HF for violent videos is 94.2%, which is the best of the three methods. The recall accuracy of the fused features for violent videos is 92.6%, and LF is 87.4%. It should be noted that the fused feature method performs the highest precision accuracy of detecting violent videos, while the HF method gets the highest recall accuracy. Of course, the recall accuracy of the fused features is at the same level with HF.

4.3. *The effect with different size of audio words*

In our method, we need to build the BOAW by clustering the AudioSignature feature which has been extracted with the weighting schemes mentioned in Sec. 3.2. The size of audio words should have a significant effect on the classification accuracy.

Yang and Jiang[36] set up an experiment to discuss the effect with the size of visual words. The conclusion is that with the increase of the number of words, the accuracy raises first and then peaks at certain points, and after that either levels off or drops mildly. In their experiment the peak point on the size of visual words is at about 5000. In Ref. 36, the key point is detected by 128-dimensions SIFT vectors for visual features with BOVW. In our method, we only focus on the audio features with BOAW, and the AudioSignature feature of each sample is represented only by 32-dimensions vector. It is obvious that the proper size of audio words should be much smaller than of video words. In this experiment, we discuss the effect on the classification accuracy with different size of audio words, the results as shown in Fig. 8.

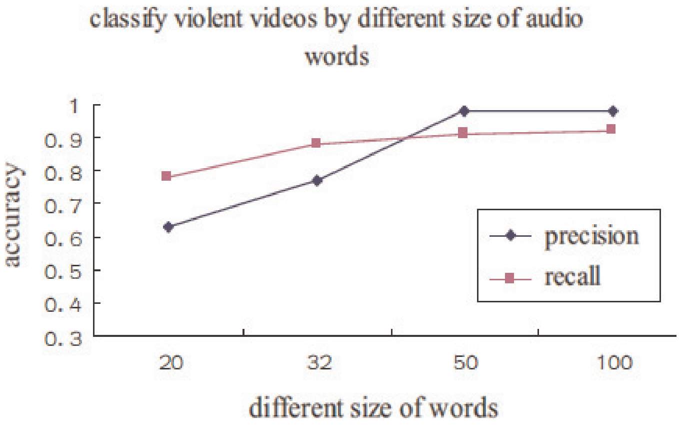


Fig. 8. The effect with different size of audio words in classifying the same dataset.

From Fig. 8, with the growth of the size, the precision and the recall are both growing, while the recall is higher than the precision. We can find that when the size of the words is larger than 50, the precision does not drop or rise obviously while the recall can get very high accuracy. Hence, in our method, the optimal number of the size of audio words is set to 50.

4.4. *Comparative experiment with other video classification method*

Li *et al.*[17] propose a novel video classification method recently. Li *et al.*'s method use a CBOW representation to model two kinds of typical contextual relations between local patches, a semantic conceptual relation and a spatial neighboring relation. To model the semantic conceptual relation, visual words are grouped on multiple semantic levels according to the similarity, accordingly local patches are encoded and images are represented. To explore the spatial neighboring relation, an automatic term extraction technique is adopted to measure the confidence that neighboring visual words are relevant. Word groups with high relevance are chosen and their statistics are incorporated into the BOW representation. Classification is taken using the SVM with an efficient kernel to incorporate the relational information. In this experiment, we use the same database with Experiment 4.1 to simulate Li *et al.*'s method, and of course we have the same result of our method as Experiment 4.1. Table 4 shows the simulation result of Li *et al.*'s method.

From Table 4, we can see that Li *et al.*'s method is a stable method for general video classification, it gives a similar classification accuracy rate among these five categories, with the total precision rate 62.6%, and the recall rate of concept sports about 61.7%. In Ref. 17, the authors only show the recall rate of concept sports about 66% and do not process the other four concepts. There is little difference on the experimental results between our simulation and in Ref. 17, it may be caused by the different datasets. From the recall rate result, we can find that the news category

Table 4. The result of classification for five types with method in Ref. 17.

	Results (Be Recognized as)						
Samples (Five Genres)	Cartoon	Entertainment	News	Sports	Violent	Total	Recall Rate
Cartoon	41	9	0	3	15	68	60.3%
Entertainment	15	88	12	13	19	147	59.9%
News	2	7	59	9	5	82	72.0%
Sports	1	8	10	37	4	60	61.7%
Violent	0	55	124	4	302	485	62.3%
Precision rate							
for each type	69.5%	52.7%	28.8%	56.1%	87.8%	—	—
for total		62.6%				—	—

performs the best. The reason is that in news videos, there are always speech and talking scenes, which have a higher appearance frequency than in the other four genres.

Table 5 shows the result of the simulation of Li *et al.*'s method for the combined two types: violent and nonviolent. It can be contrasted to the result of our method from Table 2. Our method has a total precision rate of 88.6% while Li *et al.*'s method is 73.2%. Thus for violent video detection, our method performs much better than Li *et al.*'s method. Since scream, gun fight and explosion scenes appear frequently in violent videos which always include much intense sound, so audio features are good choice for violent video detection.

From Table 5, we can make a conclusion that the performance of violent video recognition in our algorithm is well, because the average precision of our algorithm is 88.6% in Table 2 which is better than 73.2% of Li *et al.*'s method.

From Fig. 9, the graph on the right shows that the MPEG-7 features such as audio signature in our method can make a great contribution to violent video detection. From this graph we can see directly that our method has a good performance for violent videos detection, the recall rate value is higher than Li *et al.*'s method. For nonviolent, the recall rates are almost closer in both method. We still can find that for video classification, Li *et al.*'s method has a more stable performance than our

Table 5. The result of classification for two types of method in Ref. 17.

	Results (Be Recognized as)		
Samples (Two Types)	Violent	Non-Violent	Recall Rate
Violent	302	183	62.3%
Nonviolent	43	314	87.9%
Precision rate			
for each type	87.5%	63.2%	—
for total	73.2%		—

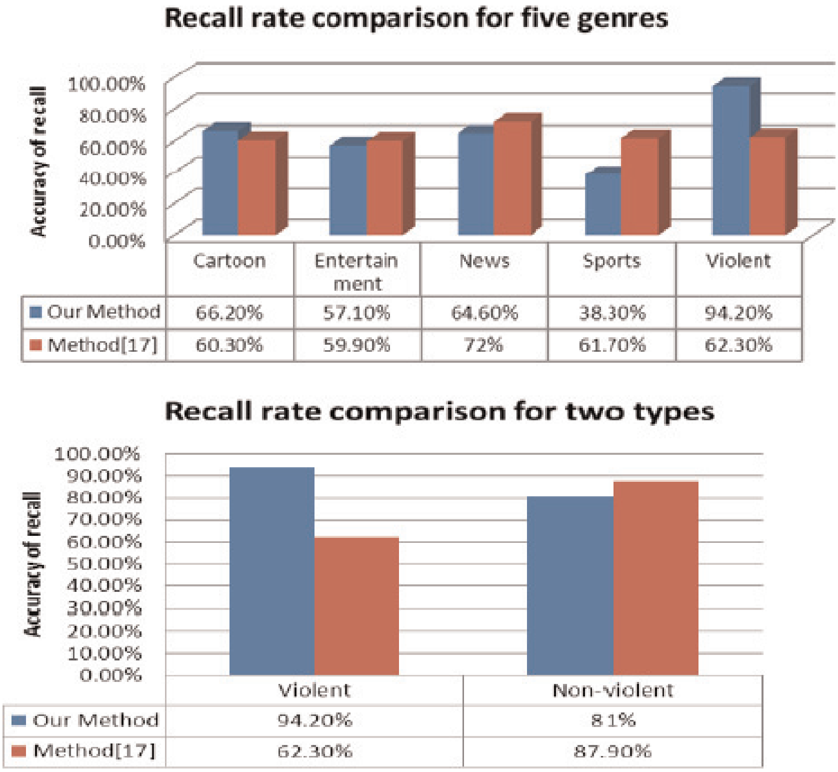


Fig. 9. The result of recall rate for both five genres and two types of our method versus the method in Ref. 17.

method since we only use the audio signature as the descriptor and it is more meaningful for violent videos. So Li *et al.*'s method has a better adaptability.

4.5. *Contrast experiments to classify the candidate shots*

Gong and Wang[34] propose a filter process of violent videos which run as follow: For each shot (video samples), low-level visual features are first extracted and a classifier is trained to identify the shots with potential violent content, which are called candidate shots. Then candidate shots are detected by some low level audio features, such as spectrum power, bandwidth, etc. In this experiment, we adopt the same filtering process and use our methods of BOAW to classify the candidate shots, contrast to Gong and Wang's method only with some low-level audio features. In this experiment, 100 violent videos and 200 nonviolent ones are evaluated. From Table 6, the BOAW is shown to perform better than the common low-level audio features in terms of accuracy. It gets 5% better in both recall and precision accuracy.

While taking the time cost into account, we make another comparison experiment. One is to filter the candidate violent videos by structure tensor in Ref. 34. The other is by our method with BOAW. We totally choose 100 violent videos and 100

Table 6. Classify the candidate shots with different methods.

	Recall	Precision
Our method (BOAW)	95%	85%
Method in Ref. 34	90%	80%

Table 7. Compare the time cost and the accuracy of different filtering methods.

Methods	Time Cost (s)	The Number of Candidates Shots	The Accuracy of Candidate Shots
Our method	3.5−5.6	130	94.4%
Method in Ref. 34	8.5−12.4	110	95.7%

nonviolent videos. The duration of each sample is 1 min. We make a comparison about the time cost and the filtering accuracy of candidate shots.

In Table 7, the time cost has a distribution scale of totally 200 videos. In our method, the shortest time cost is 3.5 s while the longest is 5.6 s. It is obvious that the time cost of audio filtering method is much lower than that of structure tensor method. The main time cost of structure tensor method is composed by the followings: compute every pixel of video frame, and compare the adjacent frame by some features. For example, if the resolution of a picture is 600 × 480, there are 307,200 pixels all together. If we represent each pixel by 8 bits (gray image), the total storage space is 2400 Kb. So for the visual features more time cost has to be spent to compute the relations between the pixels. On the other hand, the visual method has a higher filtering accuracy of candidate shots than the audio method. From this experiment, it shows that audio method can filter the violent videos with much higher speed while the visual method performs better accuracy.

5. Conclusion

The content-based video classification and recognition would play more and more important role on the internet. The videos with some violent scenes are not suitable for children and should be considered under some internet information control policies. Our research focuses on classifying the violent videos from nonviolent videos. Most previous content-based image or video classification schemes employ multiple visual features to characterize image or video content. In this paper, a novel method to identify the violent video only with audio features is introduced. The MPEG-7 low level features such as AudioSpectrum-Centroid and AudioSpectrum-Spread, and the high level feature such as AudioSignature, are combined into one 44-dimensions vector in the BOAW model. The audio words are built from the vector by the clustering strategy, and SVM with revised soft weighting scheme is used to group the audio words features into two classes, i.e., the violent and nonviolent. Experiments demonstrate that our method can achieve good recall and precision accuracy on violent videos detecting. Our method shows a better performance in time cost and can filter the violent videos with much higher speed. The method can also be applied to classify other types of videos.

Since there is no uniform definition of the violent video, five typical scenes are chosen to define the scope of the violent videos in our experiments. The five scenes may not be enough to represent the scope of all violent videos, but it is still reasonable and useful to draw some positive conclusions. Experiments 4.1 and 4.2 are performed to evaluate the efficiency of the fused method with the low level audio features and high level features. Experiment 4.3 is performed to choose the proper size of audio words. Experiment 4.4 is performed to compare the precision rates on violent videos detection between the general video classification method and our method. Experiment 4.5 is performed to compare the time cost and the filtering accuracy of our method with other method. Experiment results have shown that: (1) audio features should have very special significance for violent video; (2) the fused feature method performs the highest precision accuracy of detecting violent videos, while the HF method gets the highest recall accuracy; (3) the clustering method and soft weighting scheme can raise the detecting accuracy; (4) in our method, the optimal number of the size of audio words is set to 50, which is much smaller than the size of video words in other methods; (5) the computational complexity of dealing with audio features is much lower than that of visual features, the time cost of our method is half of the visual feature extraction.

The future work of our team is to combine the audio features with some visual features, such as Motion Vector Features, Motion 3D SURF Features, and so on. By using fusing matrix and other methods of feature fusion, we may classify the videos faster and more accurately.

References

1. H. D. Cheng, X. J. Shi and R. Min, Approaches for automated detection and classification of masses in mammograms, *Pattern Recogn.* **39**(4) (2006) 646–668.
2. J. Zhang and L. Ye, Series feature aggregation for content-based image retrieval, *Comput. Electr. Eng.* **36**(4) (2010) 691–701.
3. T. Giannakopoulos, A. Makris, D. Kosmopoulos *et al.*, Audio-visual fusion for detecting violent scenes in videos, artificial intelligence: Theories, models and applications, *6th Hellenic Conf. AI* (*SETN 2010*), Athens Greece (2010), pp. 91–100|ix+429.
4. L. Tan, Y. Cao, M. Yang *et al.*, A novel fusion method for semantic concept classification in video, *J. Software* (2009) 968–975.
5. N. Liu, Y. Zhao, Z. Zhu *et al.*, Commercial shot classification based on multiple features combination, *IEICE Trans. Inf. Syst.* (2010) 2651–2655.
6. L. Ballan, M. Bertini, A. Del Bimbo *et al.*, Video event classification using string kernels, *Multimedia Tools Appl.* (2010), 69–87.
7. H.-S. Min, Y. B. Lee, W. De Neve *et al.*, *Semantic Home Video Categorization*, *Image Processing*: *Algorithms and Systems VII*, San Jose, CA USA (2009), 72451F, pp. 10.
8. H.-Y. Huang, W.-S. Shih and W.-H. Hsu, A film classifier based on low-level visual features, *J. Multimedia* (2008) 36–43.

9. H. Cheng and R. Wang, Semantic modeling of natural scenes based on contextual Bayesian networks, *Pattern Recogn.* (2010) 4042–4054.
10. J. R. R. Uijlings, A. W. M. Smeulders and R. J. H. Scha, Real-time visual concept classification, *IEEE Trans. Multimedia* (2010) 665–681.
11. C. Huang, T. Fu and H. Chen, Text-based video content classification for online video-sharing sites, *J. Am. Soc. Inform. Sci. Technol.* (2010) 891–906.
12. F. Perronnin, Universal and adapted vocabularies for generic visual categorization, *IEEE Trans. Pattern Anal. Mach. Intell.* (2008) 1243–1256.
13. P. Muneesawang, L. Guan and T. Amin, A new learning algorithm for the fusion of adaptive audio–visual features for the retrieval and classification of movie clips, *J. Signal Process. Syst. Signal Image Video Technol.* (2010) 177–188.
14. R. Cai, L. Lu and A. Hanjalic, Co-clustering for auditory scene categorization, *IEEE Trans. Multimedia* (2008) 596–606.
15. J.-J. Aucouturier, The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music, *J. Acoust. Soc. Am.* (2007) 881–891.
16. S. Y. Liu, D. Xu and S. H. Feng, Region contextual visual words for scene categorization, *Exp. Syst. Appl.* (2011) 11591–11597.
17. T. Li, T. Mei, I.-S. Kweon *et al.*, Contextual bag-of-words for visual categorization, *IEEE Trans. Circuits Syst. Video Technol.* (2011) 381–392.
18. J. Qin and N. H. C. Yung, Scene categorization via contextual visual words, *Pattern Recogn.* (2010) 1874–1888.
19. A. Koike and T. Takagi, Classifying biomedical figures using combination of bag of keypoints and bag of words, *Int. Conf. Complex*, *Intelligent and Software Intensive Systems*, Fukuoka, Japan, 16–19, March 2009, pp. 848–853.
20. L. Weizman and J. Goldberger, Detection of urban zones in satellite images using visual words, *Geoscience and Remote Sensing Symp.*, Vol. 5 (2008), pp. 160–163.
21. M. Montagnuolo and A. Messina, Parallel neural networks for multimodal video genre classification, *Multimedia Tools Appl.* (2009) 125–159.
22. W. Zajdel, J. D. Krijnders, T. Andringa *et al.*, CASSANDRA: Audio-video sensor fusion for aggression detection, *IEEE Conf. Advanced Video and Signal Based Surveillance*, London, England, 5–7 September 2007, pp. 200–205.
23. T. Giannakopoulos, A. Pikrakis and S. Theodoridis, A multi-class audio classification method with respect to violent content in movies using Bayesian networks, *Multimedia Signal Processing*, *2007. MMSP 2007. IEEE 9th Workshop* (*2007*), pp. 90–93.
24. A. Mecocci and F. Micheli, Real-time automatic detection of violent-acts by low-level color visual cues, *2007 IEEE Int. Conf. Image Processing, ICIP 2007*, San Antonio, TX USA (2007), pp. 345–348.
25. Y. Gong, W. Q. Wang, S. Q. Jiang *et al.*, Detecting violent scenes in movies by auditory and visual cues, *9th Pacific Rim Conf. Multimedia*, National Cheng Kung University Tainan, Taiwan, 9–13 December 2008, pp. 317–326.
26. F. D. M. De Souza, G. C. Cha vez, E. A. do Valle Jr. *et al.*, Violence detection in video using spatio-temporal features, *2010 23rd SIBGRAPI Conference on Graphics*, *Patterns and Images* (*SIBGRAPI 2010*), 2010, pp. 224–230|x+388.
27. T. Giannakopoulos, A. Pikrakis and S. Theodoridis, A multimodal approach to violence detection in video sharing sites, *2010 20th Int. Conf. Pattern Recognition* (*ICPR 2010*), Istanbul, Turkey (2010), pp. 3244–3247.
28. Z. Zeng, S. Zhang, H. Li, W. Liang and H. Zheng, A novel approach to musical genre classification using probabilistic latent semantic analysis model, *IEEE Int. Conf. Multimedia and Expo 2009*, New York, USA, June 28–July 3 2009, pp. 486–489.

29. B. S. Manjunath, P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface* (Wiley Press, 2002).
30. J. M. Martinez, R. Koenen and F. Pereira, MPEG-7: The generic multimedia content description standard, part 1, *IEEE Trans. Multimedia* **9**(2) (2002) 78–87.
31. G. Forman, An extensive empirical study of feature selection metrics for text classification, *J. Mach. Learn. Res.* **3** (2003) 1289–1305.
32. C. Ding and X. He, K-means Clustering via principal component analysis, *Int. Conf. Machine Learning*, Banff, Alberta, Canada, July 2004, pp. 225–232.
33. Y.-G. Jiang and C.-W. Ngo, Towards optimal bag-of-features for object categorization and semantic video retrieval, *Conference On Image And Video Retrieval*, Amsterdam, The Netherlands, July 2007, pp. 494–501.
34. Y. Gong and W. Wang, Detecting violent scenes in movies by auditory and visual cues, *Pacific Rim Conf. Multimedia*, Tainan, Taiwan, 12–16 December 2008, pp. 317–326.
35. A. F. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and TRECVid, in *Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval*, Santa Barbara, California, USA, 26–27 October 2006, pp. 321–330.
36. J. Yang and Y.-G. Jiang, Evaluating bag-of-visual-words representations in scene classification, *Multimedia Information Retrieval*, Augsburg, Bavaria, Germany, 28–29 September 2007, pp. 197–206.