Multi-modal fusion for violent scene detection in Hollywood movies

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[a,b], Global Customer Service[b,]*

[a]1600 John F Kennedy Boulevard, Philadelphia
[b]360 Park Avenue South, New York

Abstract

Violent scene detection (VSD) is a challenging problem because of the heterogeneous content, large variations in video quality, high semantic meaning of the concepts. In the last few years, combining multiple features from multimodalities is proven as an effective strategy for a wide range of video classification tasks and almost existing works have focused on general multimedia event detection (MED), the specific event detection like VSD has been comparatively less studied. In this work, we try to study the value of multi features and combination of these features for VSD in Hollywood movies. We rigorously analyze and combine a set of low-level features and deep learning feature that capture appearance, color, texture, motion and audio in videos. We also evaluate the utility of mid-level visual information obtained from detecting related violent concepts. Experiments are performed on MediaEval VSD 2014 dataset, made publicly available. Results show the performance of visual and motion features are better than audio features. The performance of mid-level features was nearly as good as that of low-level visual features. Experiments with a number of fusion methods show that all single features are complementary and help to

[☆]Fully documented templates are available in the elsarticle package on CTAN.

[*]Corresponding author
Email address: support@elsevier.com (Global Customer Service)
URL: www.elsevier.com (Elsevier Inc)

[1]Since 1880.

improve overall performance. This study also provides an empirical foundation for selecting feature sets that are capable of dealing with heterogeneous content data as violent scenes in movies.
Keywords: violent scene detection,video retrieval,multi-modal fusion,multiple features

1. Introduction

Nowadays, the movie industry generates thousands of movies each year. However, not all movies are suitable for young people (especially children, teenagers, e.g.) to watch, because they might have violent contents. The research from University of Pittsburgh [2] reports that watching violence in movies or TV programs tends to make children more aggressive and leads to unhealthy attitudes. It is crucial to have violent scene detection (VSD) system which enables the parents to choose movies that are suitable for their children. More general, for content provider, the violent scene detection technique can be used to assist in movie rating; for general end users, it can block the violent content in client terminal devices.

[2] http://www.ocd.pitt.edu/Files/PDF/Parenting/TvAndMovieViolence.pdf

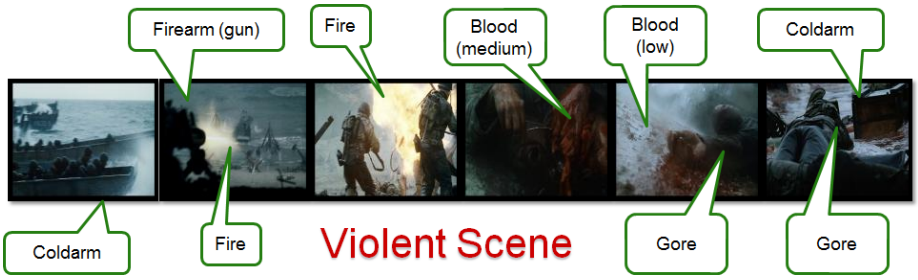

Figure 1: Following the violent definition in [1], this is an example of a violent scene in Saving Private Ryan Movie because it contains physical violence or accident resulting (fighting scene: boats approaching, guns shooting, fire guns shooting,..) in human injury or pain (blood and gory, dead soldiers)

Detecting violent scene is a challenging task mainly due to the large variation of violence concept and the semantic gap between the concept and low-level features provided in the media. To capture violence insights, multiple features (e.g. global or local static visual features, motion-based featuers, and audio features) are extracted for learning and prediction. Besides, to narrow the semantic gap, a set of various mid-level concepts and their spatial temporal relationships can be used to characterize violence e.g. **object-based concepts**: blood, gun, cold arm, firearm; **action-based concepts**: fighting, running, shooting, etc.; **activity-based concepts**: car chase, boat chase; **scene-based concepts**: gory scene, overcast, scene; **sound-based concepts** such as gun shot, scream, explosion. Figure 1 demonstrate an example of violent scene in a Hollywood movie. However, how to precisely detect mid-level concepts and efficiently utilize their relationships is yet another challenging and unsolved problem.

Given a movie as input, the expected output of a VSD system is all shots and segments having violent information. A general framework contains 4 main processing steps: video segmentation, feature extraction, learning models, and prediction. In the first step, the input movie is divided into fundamental processing units i.e. video shots. Then, features are extracted from the shots in second step. The goal of this step is to extract meaningful features which will be used for learning and predicting concepts in shots in later steps. Related studies

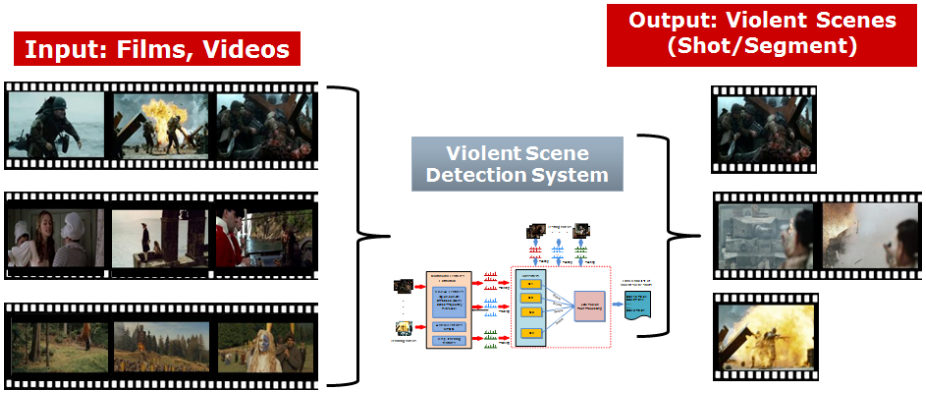

Figure 2: Overview of violent scene detection system.

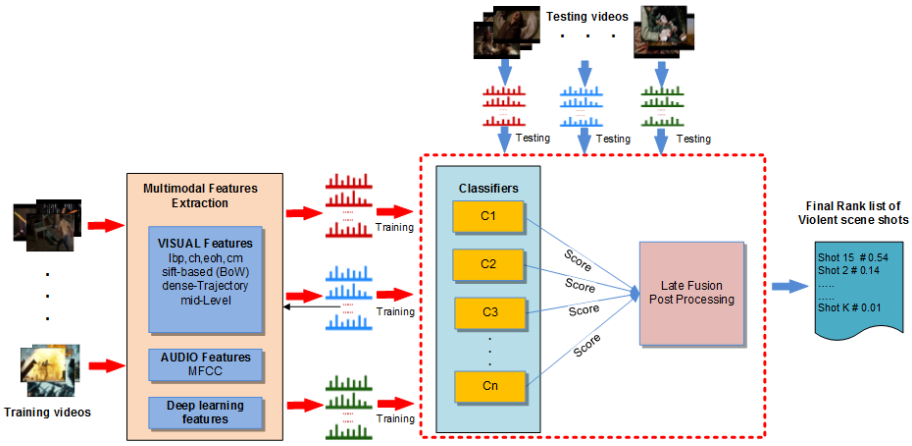

Figure 3: Overview of violent scene detection framework.

for similar system (e.g. multimedia event detection, action recognition, scene classification) have shown that selecting relevant features and their combination is essential for accurate detection.

In this paper, we evaluate various types of features and their combinations for VSD.

- **Visual features.** We select following features for evaluation: color moments, color histogram, edge orientation histogram, and local binary patterns, SIFT, Color-SIFT, Opponent-SIFT. Visual features are expected to be able to convey visual characteristics of the original violence concept, the object-based concepts, and the scene-based concepts.

- **Audio feature.** We use the standard MFCC as audio feature. It is used for capturing specific sound signals in violent scenes to discriminate them from non-violent scenes (e.g. sounds of gun shots, a scream, or explosions).

- **Motion features.** We employ Dense Trajectory feature with Motion Boundary Histogram (MBH), Histograms of Oriented Gradients (HoG), and Histograms of Optical Flow (HoF) as features in presentation of action-based and activity-based concepts. This is to take the advantages of videos compared to static images.

- **Mid-level based feature.** To narrow the semantic gap, we use mid-level features to describe the original violence concept. Scores from classifiers of mid-level concepts are composed to form a unified feature for each shot. We then use these features for shot classification.

The main focus is to provide a comparative analysis on features and their contributions to the final performance of a VSD system. The observation drawn from our work may be helpful to future works in the field.

We perform our experiments on a standard benchmark dataset, MediaEval VSD 2014[1], which are made publicly available [3] . Our system is trained and tested on Hollywood movies within MediaEval VSD Dataset. MediaEval VSD official metric (MAP-mean average precision) is used for evaluation. By using standard dataset and evaluation protocol, the performance of our system can be fairly compared to those from other MediaEval VSD participants (without using external data). Results show the performance of visual features are the best. Low-level visual features and motion feature play very important role when fusing all modalities. Audio and mid-level features are not good as low-level visual features but they are complementary to improve overall performance. Experiments with several fusion methods show that all single features are complementary and help to improve overall performance.

The rest of this paper is organized as follows. We review some related work in section 2. Section 3 introduces our framework, our approaches for feature extraction, and detail of feature configurations, fusion methods. The experimental results and their analysis will be in section 4, as well as the conclusion and future works are described in the section 5 of this paper.

[3] http://www.technicolor.com/en/innovation/research-innovation/scientific-data-sharing/violent-scenes-dataset

2. Related work

Violent scene detection is a kind of multimedia event detection. Although combining multiple features from multi-modalities has been proven as an effective strategy for multimedia event detection, relatively few works have been proposed to apply these approaches for violent scene detection in video. The main reason is that the definition of violence is ambiguous. It is difficult to describe this high-level concept using mathematical formulation precisely. In general, the violence concept is not well-defined and recent approaches addressed the problem by their own definitions.

Some of the previous works applied different kind of visual features to detect flame, blood, explosion as the informative cues for violence. As a first approach in this field, Jeho et al.[2] propose an approach to recognize violent scenes in videos by detecting flame and blood, capturing the degree of motion activity, the soundtrack and characterization of sound effects of violent events. Meanwhile, Chen et al.[3] decompose violent scene detection into action scene detection and bloody frame detection. Clarin et al.[4] present a system which uses a Kohonen self-organizing map, which is for detecting skin and blood pixels in each frame, and motion intensity analysis to detect violent actions involving blood. More recently, Gong Yu et al.[5] introduce a violence detector using low-level visual and auditory features and high-level audio effects to identify potential violent content in movies. Jian [6] describe a weakly-supervised audio violence classifier combined using co-training with a motion, explosion and blood video classifier to detect violent scenes in movies. Penet et al.[7] compared two modality fusion methods, namely Early Fusion and Late Fusion. Early Fusion concatenates features from both modalities before machine learning, while Late Fusion fuses probabilities of both modalities already calculated. They reported Late Fusion was superior to Early Fusion.

Beside low-level based approaches, violent scene detection is also a kind of high level recognition task. The major challenge is to deal with the gap of semantic meaning. In recent years, there are some approaches using attributes

to narrow the semantic gap in high-level recognition task, such as object recognition using attributes, scene classification using attributes, action recognition using actions bank[8]. Li et al.[9] propose Object Bank (OB), a new representation of natural images based on objects (using objects as attributes to describe scenes). Liu et al.[10] use high-level semantic concepts, also called attributes, to represent human actions from videos and argue that attributes enable the construction of more descriptive models for human action recognition. Sasanand et al.[8] present Action Bank, a new high-level representation (kind of attributes) of video. For VSD, Bodgan et al.[11] rely on fusing mid-level concept predictions made using multi-layer perception classifiers to automatically localize the occurrence of violence within a video. They proposed a frame-level violence prediction, applying a multi-layer perceptron in order to utilize these concepts. They put the first layer for the concept prediction, and the second layer for the violence prediction. In addition to those provided concepts, Tan et al.[12] have utilized extra 42 violence concepts such as bomb and war from ConceptNet[13]. ConceptNet is composed of nodes representing concepts in the form of words or short phrases with their relationships. On their system those extra concepts are trained using YouTube videos which are crawled additionally.

In the last few years, MediaEval Violent Scene Detection affect Task is become popular, many state-of-the art systems have been developed and reported recently[1]. Most systems extract a large number of multimodal features from visual, audio signals and text. The features include low-level features such as color histogram, edge of histogram, local binary pattern, SIFT based, Space-Time Interest Points (STIP), Histograms of Oriented Gradients (HoG), Histograms of Optical Flow (HoF), Motion Boundary Histograms (MBH), visual activity, MFCCs, trajectory-based features (cite each team). Different state-of-the-art features encoding are used, such as Bag-of-Visual-Words representation, Fisher vector representations. In addition, some approaches use of pre-defined concept detectors like part-level attributes FUDAN Team[14], mid-level concept description, with provided concepts from ConceptNet (e.g., punishment, victim, rape, etc) VIREO team[12]. The video shots are classified by using dif-

ferent machine learning techniques (Support Vector Machines (SVMs), k nearest neighbors (kNN), Bayesian Network). Multimodal integration is achieved via early fusion[15] and late fusion[15, 16, 17, 14] where score fusion is often used to combine scores independently computed from different subsets of features.

In this paper, we present a novel violent scene detection framework which is designed to incorporate many of the above-mentioned success principles in its architecture. In particular, our work incorporates novel developments into the system, which can be summarized into major contributions. First, our work developed and incorporated multi modal features at diverse granularities to evaluate the value of different types of features for VSD. Second, our work explores the use of mid-level concept features, which are detected based on low-level features, aiming to provide more semantic understanding capability into our system. Third, we present several approaches to learn fusion functions which combine violent scores (late fusion) to improve overall performance.

3. VSD system

3.1. Framework Overview

We use a unified framework to evaluate the performance of each feature as well as the performance of feature combination methods. The framework should be flexible so that we can easily test different kind of features or fusion strategies. It also should be designed in a component-based manner so that we can evaluate each component separately while keeping other components intact. In this spirit, our framework can be decomposed into following components: (1) Pre-processing, (2) Feature Extraction, (3) Feature Encoding, (4) Feature Classification and (5) Feature Fusion. The detail of each component is described in the following sections. We especially focus on the feature encoding and feature fusion component. The former is used to evaluate features while the later is used to evaluate fusion strategies for VSD. An overview of our framework is shown in Figure 3.

3.2. Pre-processing

The pre-processing step prepares data for further processing at later components. At first, videos are resized to the width of 320 pixels, while the resized height is scaled relatively so that the aspect ratio is kept. Motion features are extracted from the resized videos. This simple processing technique significantly reduce the processing time and still does not affect the detection performance, as shown in [18] for multimedia event detection.

For image features, keyframes are sampled at five frame per second. We found this sampling technique a good trade off between time and accuracy, as also suggested in [19]. After that, blank keyframes, which are often filled with single color, are removed because they do not contain informative feature.

For audio features, we only extract the audio channel from the original video. Since the scale of the video does not affect the audio information, we use the original video for audio extraction and save as audio files with standard WAV format. Audio features will be extracted from these audio files directly.

3.3. Feature Extraction

The feature extraction component aims to make a discriminative vector representation for each shot that is extracted from the pre-processing step. The extraction method depends on what type of feature will be used. To conduct a comprehensive evaluation of features for VSD, we support a large variety of features including global and local visual features. Global features capture the global statistics of each extracted shot. These statistics can be calculated directly from sub regions of a sampled frame and then concatenated to form the vector representation for that frame, before being aggregated to make the final representation for each shot. It is more complicated to calculate the feature vector representation for local features. The number of local features vary from frame to frame, therefore it requires a special encoding technique which will be described in 3.4.

Besides global and local features, other features are also supported in our evaluation framework. Audio feature can be extracted from pre-defined tempo-

ral windows. Feature of each temporal window provide a local audio characteristic at that temporal location. Therefore, audio feature can be considered as a local feature and can be well-integrated into our feature extraction framework. Another feature that is also supported is mid-level feature, which is represented from violent concept detectors. We consider both the in-domain and out-domain concept detectors. In case the concepts are used from off-the-shelf datasets, we employ the state-of-the-art deep learning features which are extracted from a pre-trained model. The detail description of each feature is presented in Section 4.

3.4. Feature Encoding

As for local features, we employed the popular bag-of-words (BOW) model to generate a fixed-length feature representation from its local descriptors. The model was first applied to represent text document as mentioned in [20]. It was adopted to represent image by Csurka et al. [21]. Its extension to motion and audio features is also straightforward as described in [22] and [23].

We follow the experiment setup in [24] to implement our bag-of-words models. We fix the codebook size to 1000 because we observe that in [24] the performance does not significantly improve when the larger codebooks were used. On the other hand, smaller codebook can significantly reduce the computational time for feature encoding as well as feature learning. In order to train the codebook, we randomly select 1M local descriptors and cluster using the K-means algorithm. The local descriptors are assigned to each codeword in a soft-weighting manner [25] to improve the discriminative of the encoded feature.

The main drawback of the bag-of-words model is that it does not incorporate spatialinformation. The simplest way to overcome this problem is to partition the image into sub-regions and encode local features in each region independently. After that feature from all regions are concatenated into a single feature vector. There are many ways to partition an image into sub-region. To this end, we follow [24] and[26] to use 2×2 and 1×3 spatial configurations. We

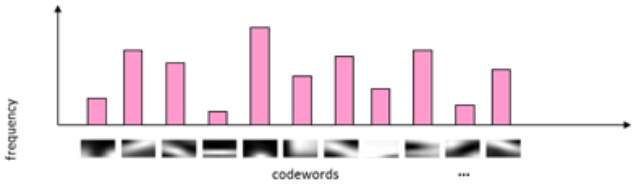

Figure 4: Bag of visual words.

found that these spatial configurations a good trade-off between performance and computational cost due to the high dimensional feature vector.

3.5. Feature Classification

LibSVM [27] is used for training and testing at shot level. To generate training data, shots which fall into positive segments more than 80% will be considered as positive shots. The remaining shots are considered as negative. Extracted features are scaled to [0, 1] using the svm-scale tool of LibSVM. We use chi-square kernel to calculate the distance matrix. The optimal (C;g) parameters for learning SVM classifiers are found by conducting a grid search with 5-fold cross validation on the original dataset.

3.6. Feature Fusion

3.6.1. Late Fusion

Fusing information coming from different media seems a natural way to handle multimedia content. Fusing of multi-modal information has been widely used for tasks including multimedia event detection, video search, etc. Naturally different semantic events, types of multimedia data have their own characteristics so their fusion strategies could be different also. For violent scene detection, we investigated several simple but effective fusion schemes to construct a better violence prediction system that exploits the advantages of each type of modalities.

4. VSD features

4.1. Visual global features

We evaluate both global features and local features. We selected the best configuration of each types of features from our previous work[28]. The global features include color moments, color histogram, edge orientation histogram, and local binary patterns with different configuration:

- Granularity: Since global features do not capture spatial information, to overcome this problem, a grid nm usually used to divide the input image into non overlapping sub-regions. The features extracted from these regions are concatenated to form the feature vector for the image.
- Color space: Local binary patterns (LBP) and edge orientation histogram (EOH) are extracted from gray scale image. For color moments and color histogram, color spaces including HSV, RGB, Luv, and YCrCb are used.
- Quantization: For color histogram, we only use 8-bin histogram for each channel. For edge orientation histogram, we quantize orientations into histograms of 36 (edge) + 1(non-edge) bins. For local binary patterns, we quantize binary patterns into histograms of 30, 59 bins.

4.2. Visual local features

4.2.1. Still image feature

For local feature, we use popular SIFT with both Hessian Laplace interest points and dense sampling at multiple scales. For dense sampling, besides normal SIFT descriptor, we also use Opponent-SIFT and C-SIFT. For Harlap interest point detector, we only use normal SIFT descriptor.

4.2.2. Motion feature

As shown by Wang et al. [29], Dense Trajectory feature is one of the best approaches for action classification. Dense Trajectory has an efficient solution to remove camera motion. In the Hollywood movies, there are a lot of action with

different movie effects in the violent scenes. We try to apply the dense trajectory to capture these information. Trajectories are obtained by tracking the densely sampled points in the optical flow fields. As it is suggested by Wang [29], we use Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) to describe each trajectory. HOG is used to capture appearance characteristics of the moving object while HOF captures the speed of the moving object. The last descriptor, MBH, can capture boundary of the motion and is good for handling camera motion.

4.2.3. Audio feature

We use the popular Mel-frequency Cepstral Coeffcients (MFCC) [30] for extracting audio feature. We choose a window size of 25 ms for audio segment and a step size of 10 ms. The 13-dimensional MFCC vectors along with each first and second derivatives are used for representing each audio segment. Raw MFCC features are also encoded using BoW. These configurations are also used by AXES/LEAR team in TRECVID Multimedia Event Detection 2013 [18] and THUMOS Challenge 2014 [31] in which they won the best performance.

4.3. Mid-level feature

4.3.1. In domain: VSD concepts

Violent scene detection is also a kind of highlevel recognition task. Violence concept has high semantic meaning and high variability in appearance. Besides that, due to lack of training data for this concept, training violence classifier directly from low-level features is not effective. Instead of focusing on selecting the low-level features for VSD, we try to investigate how to use related violent information as mid-level feature to detect violent scene in the movie. We manually select the violent information which are annotated by human assessors as related attributes of violence concept.Then we propose to use such concepts as related semantic attributes of the original violence concept. Figure 1 shows how to describe the violent scene by their semantic attributes. These attributes which can be more well-presented by low-level features (e.g. the attribute about

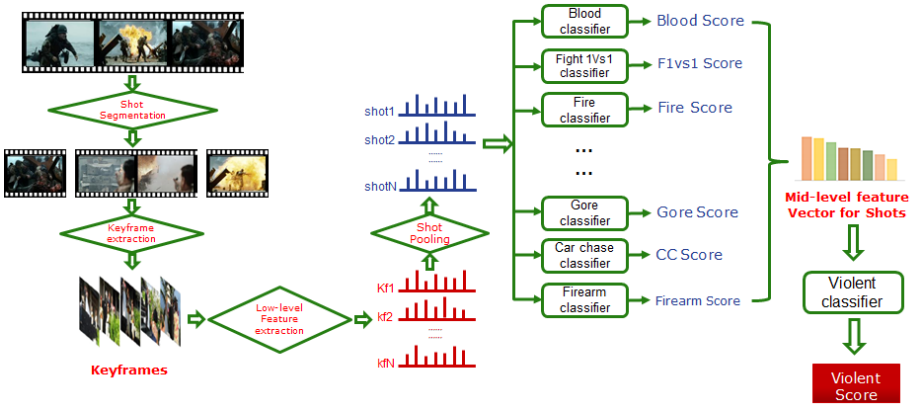

Figure 5: Overview of Mid-level framework.

blood can be sufficiently presented by color) are used to create the mid-level features. In other words, the attributes have smaller gap to the low-level features, compared to the original violence concept. For example, in Figure 1, because violence occurred during the whole scene, it is very hard to use low-level features to represent the violence concept. But, we can use the low-level visual features to represent some attributes such as fire, blood. By doing this, we narrow the semantic gap between the original violence concept and low-level features extracted from video shots. Attributes related to violence concept are manually defined. We used seven related concepts provided in MediaEval Dataset as violent attributes. Our mid-level framework is shown in Figure 5. Firstly, each corresponding attribute classifier is trained by using low-level features. Secondly, the mid-level features of a training (or test) video shots are formulated by concatenating scores returned by attribute classifiers. At the moment, we use the same weights for concatenating different scores. Then thirdly, we use this feature to train the mid-level feature-based violence classifier. Finally, we apply this violence classifier on test set to get the violence score for each shot (these shots are also represented by mid-level features).

4.3.2. Out domain: Deep learning feature

5. Experiments

5.1. Dataset

In this paper, we used the dataset from MediaEval Affect Task 2014 [1], this is a set of 31 Hollywood movies that must be purchased the original DVD due to copyright issues. The movies are of different genres (from extremely violence movies to movies without violence). In this dataset, we focus on the violent concept with subjective definition, which is defined as those which one would not let an 8 years old child see because they contain physical violence. Follow the proposed method, we divide this dataset into 2 parts:

- DEVEL (Table 1): this is training dataset; used to train violence classifiers; has 24 movies, total 34779 shots, 48.29 hours.
- TEST (Table 2): this is testing dataset; used to test and evaluate the system; has 7 movies, total 10006 shots and 13.89 hours.

Total duration are about 62.18 hours, with 44.785 shots. To reduce the computation cost, when we extract keyframes, we resize the keyframes to 500x400 pixels.

5.2. Groundtruth

By using subjective definition in MediaEval VSD 2014[1], the ground truth is created by human assessors and provided by the MediaEval organizers. In addition to segments containing physical violence, annotations also include the following high-level concepts: presence of blood, fights, presence of fire, presence of guns, presence of cold arms, car chases and gory scenes, for the visual modality; gunshot, explosion and scream for the audio modality. The ground truth data are provides in segment. To generate training data, we consider the positive shots are the ones which have 80% overlapping with ground truth segments.

No.	Video Name	Length (in Seconds)	#keyframes	#shot
1	Armageddon	8,681.05	217,026	1,737
2	BillyElliot	6,349.36	158,734	1,270
3	Eragon	5,985.57	149,639	1,198
4	Harry Potter 5	7,954.72	198,868	1,591
5	I Am Legend	5,780.58	144,514	1,157
6	Leon	6,344.49	158,612	1,269
7	Midnight Express	6,960.96	174,024	1,393
8	Pirates Of The Caribbean 1	8,241.01	206,025	1,649
9	Reservoir Dogs	5,712.98	142,825	1,143
10	Saving Private Ryan	9,750.89	243,772	1,951
11	The Sixth Sense	6,178.01	154,450	1,236
12	The Wicker Man	5,870.89	146,772	1,175
13	The Bourne Identity	6,816.29	170,407	1,364
14	The Wizard of Oz	5,859.29	146,482	1,172
15	Dead Poets Society	7,415.17	185,379	1,484
16	Fight Club	8,006.34	200,158	1,602
17	Independence Day	8,834.96	220,874	1,767
18	The God Father	10,194.96	254,874	2,039
19	Pulp Fiction	8,887.97	222,199	1,778
20	Forrest Gump	8,176.97	204,424	1,636
21	Fargo	5,646.34	141,158	1,130
22	The Pianist	8,567.10	214,177	1,714
23	Fantastic Four 1	6,094.41	152,360	1,219
24	Legally Blond	5,523.49	138,087	1,105
	otal	**173,833.8**	**4,345,840**	**34,779**

Table 1: DEVEL set includes 24 Hollywood movies.

No.	Video Name	Length (in Seconds)	#keyframes	#shot
1	V for Vendetta	7,626.49	190,662	1,526
2	Terminator 2	8,831.37	220,784	1,767
3	Jumanji Collectors Edition	5,993.98	149,849	1,199
4	Ghost in the Shell	4,966.00	124,150	994
5	Desperado	6,012.89	150,322	1,203
6	Brave Heart	10,224.49	255,612	2,045
7	8 Mile	6,355.53	158,888	1,272
	Total	**50,010.75**	**1,250,267**	**10,006**

Table 2: TEST set includes 7 Hollywood movies.

5.3. Evaluation Metrics

We use mean average precision (mAP), an evaluation metric that is widely used for classification and retrieval systems. The mAP is computed based on the ranked list of shots returned by the detection system and the ground truth provided by the task organizers. Mean average precision (mAP) is calculated as below:

$$MAP = \frac{\sum_{v=1}^{V} AP(v)}{V}$$

, where V is number of test videos and AP is average precision for each video.

6. Conclusion

We evaluate the performance of multi features for the violent scene detection. The performance of global features and local features that are widely used in state of the art classification systems are compared. Our study can be served as a baseline for comparison of advanced algorithms or systems in the violent scene detection. Still image based feature (SIFT) has better performance than motion based feature (Dens-trajectory + MBH). Fusion of global features is effective, while fusion of local features is not (only SIFT is enough). Fusion of local features, global features, motion features and audio feature achieves the best performance. Post-processing method is very effective to improve overall performance. Mid-level representation shows promising results compared to using raw feature only, however the performance is still limited. But adding mid-level feature in fusion schema is not effective. Experimental results on MediaEval 2013 VSD benchmark dataset show the validity of the approach and its comparable performance to state-of-the-art methods.

References

[1] C.-H. Demarty, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, C. Penet, Benchmarking violent scenes detection in movies, in: Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on, IEEE, 2014, pp. 1–6.

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

[3] C. Liang-Hua, H.-W. Hsu, L.-Y. Wang, , C.-W. Su, Violence detection in movies, Computer Graphics, Imaging and Visualization (CGIV) (2011) 119–124.

[4] C. C., J. Dionisio, M. Echavez, P. C. Naval., Dove: Detection of movie violence using motion intensity analysis on skin and blood, Workshops and Demonstrations - ECCV (2005) 150–156.

[5] G. Yu, W. Wang, S. Jiang, Q. Huang, W. Gao, Detecting violent scenes in movies by auditory and visual cues, Advances in Multimedia Information Processing-PCM (2008) 317–326.

[6] L. Jian, W. Wang, Weakly-supervised violence detection in movies with audio and video based co-training, Advances in Multimedia Information Processing-PCM (2009) 930–935.

[7] P. Cdric, C.-H. Demarty, G. Gravier, P. Gros, Technicolor and inria/irisa at mediaeval 2011: Learning temporal modality integration with bayesian networks, MediaEval Multimedia Benchmark Workshop.

[8] S. Sadanand, J. J. Corso, Action bank: A high-level representation of activity in video, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1234–1241.

[9] L. Li-Jia, H. Su, L. Fei-Fei, E. P. Xing, Object bank: A high-level image representation for scene classification & semantic feature sparsification, Advances in Neural Information Processing Systems (2010) 1378–1386.

[10] L. Jingen, B. Kuipers, S. Savarese, Recognizing human actions by attributes, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011) 3337–3344.

[11] I. Bogdan, J. Schluter, I. Mironica, M. Schedl, A naive mid-level concept-based fusion approach to violence detection in hollywood movies, ACM Conference on International Conference on Multimedia Retrieval (2013) 215–222.

[12] C. C. Tan, C.-W. Ngo, The vireo team at mediaeval 2013: Violent scenes detection by mid-level concepts learnt from youtube., in: MediaEval, 2013.

[13] H. Liu, P. Singh, Conceptneta practical commonsense reasoning tool-kit, BT technology journal 22 (4) (2004) 211–226.

[14] Q. Dai, J. Tu, Z. Shi, Y.-G. Jiang, X. Xue, Fudan at mediaeval 2013: Violent scenes detection using motion features and part-level attributes., in: MediaEval, 2013.

[15] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, et al., Technicolor/inria team at the mediaeval 2013 violent scenes detection task, MediaEval 2013 Working Notes.

[16] M. Sjöberg, J. Schlüter, B. Ionescu, M. Schedl, Far at mediaeval 2013 violent scenes detection: Concept-based violent scenes detection in movies., in: MediaEval, 2013.

[17] N. Derbas, B. Safadi, G. Quénot, et al., Lig at mediaeval 2013 affect task: Use of a generic method and joint audio-visual words., in: MediaEval, Citeseer, 2013.

[18] R. Aly, R. Arandjelovic, K. Chatfield, M. Douze, B. Fernando, Z. Harchaoui, K. McGuinness, N. E. O'Connor, D. Oneata, O. M. Parkhi, et al., The axes submissions at trecvid 2013.

[19] M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, Semantic model vectors for complex video event recognition, Multimedia, IEEE Transactions on 14 (1) (2012) 88–101.

[20] Z. S. Harris, Distributional structure., Word.

[21] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on statistical learning in computer vision, ECCV, Vol. 1, 2004, pp. 1–2.

[22] J. Sivic, A. Zisserman, Efficient visual search of videos cast as text retrieval, Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (4) (2009) 591–606.

[23] Y.-G. Jiang, X. Zeng, G. Ye, D. Ellis, S.-F. Chang, S. Bhattacharya, M. Shah, Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching., in: TRECVID, 2010.

[24] Y.-G. Jiang, J. Yang, C.-W. Ngo, A. G. Hauptmann, Representations of keypoint-based semantic concept detection: A comprehensive study, Multimedia, IEEE Transactions on 12 (1) (2010) 42–53.

[25] Y.-G. Jiang, C.-W. Ngo, J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of the 6th ACM international conference on Image and video retrieval, ACM, 2007, pp. 494–501.

[26] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, IEEE, 2006, pp. 2169–2178.

[27] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[28] V. Lam, D.-D. Le, S.-P. Le, S. Satoh, D. A. Duong, Nii, japan at mediaeval 2012 violent scenes detection affect task., in: MediaEval, Citeseer, 2012.

[29] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action Recognition by Dense Trajectories, in: IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, United States, 2011, pp. 3169–3176.
URL http://hal.inria.fr/inria-00583818/en

[30] L. R. Rabiner, R. W. Schafer, Introduction to digital speech processing, Foundations and trends in signal processing 1 (1) (2007) 1–194.

[31] D. Oneata, J. Verbeek, C. Schmid, The lear submission at thumos 2014.