Violence Detection in Hollywood Movies by the Fusion of Visual and Mid-level Audio Cues

Esra Acar, Frank Hopfgartner, Sahin Albayrak
DAI Laboratory, Technische Universität Berlin
Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
{name.surname}@tu-berlin.de

ABSTRACT

Detecting violent scenes in movies is an important video content understanding functionality e.g., for providing automated youth protection services. One key issue in designing algorithms for violence detection is the choice of discriminative features. In this paper, we employ mid-level audio features and compare their discriminative power against low-level audio and visual features. We fuse these mid-level audio cues with low-level visual ones at the decision level in order to further improve the performance of violence detection. We use Mel-Frequency Cepstral Coefficients (MFCC) as audio and average motion as visual features. In order to learn a violence model, we choose two-class support vector machines (SVMs). Our experimental results on detecting violent video shots in Hollywood movies show that mid-level audio features are more discriminative and provide more precise results than low-level ones. The detection performance is further enhanced by fusing the mid-level audio cues with low-level visual ones using an SVM-based decision fusion.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms, Performance, Experimentation

Keywords

Bag-of-Audio-Words, Mel-Frequency Cepstral Coefficients, Motion, Decision Fusion, Support Vector Machine

1. INTRODUCTION

The advances in digital media management techniques have facilitated delivering digital videos to consumers. Therefore, accessing online movies through services such as Video-On-Demand has become extremely easy. As a result, parents are not always able to precisely monitor what their children watch. Children are, consequently, exposed to movies, documentaries, or reality shows which have not necessarily been checked by parents, and which might contain inappropriate content. One of these inappropriate contents is violence. Psychological studies have shown that violent content in movies has harmful effects, especially on children [5]. As a con-sequence, there is a need for automatically detecting violent scenes in videos, where the legal age ratings are not available. Although defining scenes as "violent" is subjective (i.e., person-dependent), in our work, we aim at sticking to the common definition of vio-lence: "physical violence or accident resulting in human injury or pain" [7].

An important step in the task of movie violent content detection is the representation of movie segments. Many of the existing works (e.g., [6, 9]) proposed for violence detection represent videos using low-level representations, especially for the representation of audio signals. Making abstractions is better than directly using low-level features in order to bridge the gap between the features and high-level human perception of violence. However, high-level semantics are difficult to detect and state-of-the-art detectors are far from perfect. Therefore, using mid-level representations may help modeling video segments one step closer to human perception.

This paper aims at investigating the discriminative power of low-level and mid-level audio features to model violence in Hollywood movies. We also investigate how MFCC-based mid-level features perform when fused with average motion features for the detection of violent content and show that promising results are obtained by fusing these cues at the decision level.

The paper is organized as follows. Section 2 explores the recent developments and reviews methods which have been proposed to detect violence in movies. In Section 3, we introduce our method. We provide and discuss evaluation results on Hollywood movies in Section 4. Concluding remarks and future directions to expand our current approach are presented in Section 5.

2. RELATED WORK

Although video content analysis has been studied extensively in the literature, violence analysis of movies is restricted to a few studies. Due to paper length limitations, we only discuss some of the most representative ones which use both audio and visual cues.

Wang et al. [6] apply Multiple Instance Learning (MIL) (MI-SVM [4]) using color, textual and MFCC features. Video scenes are divided into video shots, where each scene is formulated as a bag and each shot as an instance inside the bag for MIL.

Giannakopoulos et al. [9] propose to use a multi-modal two-stage approach. In the first step, they perform audio and visual analysis of segments of one second duration. The classifications obtained in this first step are then used to train a *k*-NN classifier.

In [10], a three-stage method is proposed. In the first stage, they apply a semi-supervised cross-feature learning algorithm [14] on the extracted audio-visual features for the selection of candidate violent video shots. In the second stage, high-level audio events (e.g., screaming, gun shots, explosions) are detected via SVM training for each audio event. In the third stage, the outputs of the classifiers generated in the previous two stages are linearly weighted for final decision.

Lin et al. [12] train separate classifiers for audio and visual analysis and combine these classifiers by co-training. Probabilistic latent semantic analysis is applied in the audio classification part. In the visual classification part, the degree of violence of a video shot is determined by using motion intensity, the (non-)existence of flame, explosion and blood appearing in the video shot.

In order to provide a better description of audio signal, Rabiner et al. [13] propose to use discrete HMM through the use of Vector Quantization (VQ) in speech processing. In this work, we apply the VQ coding scheme on MFCCs for violence detection. Our approach differs from the aforementioned works in the following aspects: (1) we stick to a broad definition of "violence" [7], (2) we evaluate our approach on a diverse benchmarking dataset [2] (i.e., not a restricted dataset which contains only action movies), (3) we construct mid-level audio representations by a Bag-of-Audio-Words (BoAW) approach with the VQ coding scheme and show that these representations are more discriminative than low-level audio and visual ones, and (4) we further improve the performance by fusing mid-level audio representations with visual ones at the decision level by an SVM-based fusion and manage to be in the top 25% among submissions in the MediaEval Violent Scenes Detection (VSD) task [2] in terms of *average precision at 20*.

3. THE VIOLENCE DETECTION METHOD

In this section, we discuss the representation of video shots and the learning of a violence model which are the two main components of our method.

3.1 Video Representation

Sound effects and background music in movies are essential for stimulating people's perception. Therefore, the audio signals are important for the representation of videos. We represent the audio content at two different levels: low-level and mid-level. Both representations are based on MFCC features extracted from the audio signals of video shots as illustrated in Figure 1(a).

3.1.1 Low-level Audio Representation

Due to the variability in duration of the video shots annotated as violent or non-violent, each shot comprises a different number of MFCC feature vectors. Aiming at constructing audio representations having the same dimension, mean and standard deviation for each dimension of the MFCC feature vectors are computed. The resulting mean and standard deviation compose the low-level audio representations of shots.

3.1.2 Mid-level Audio Representation

In order to generate mid-level audio representations for video shots, we apply an abstraction process which uses MFCC-based BoAW. The first step in the BoAW scheme is to construct a dictionary of audio words. We follow an unsupervised way of constructing the audio dictionary. First, we cluster MFCC feature vectors extracted from shots with a k-means clustering, in which the centroid of each of the k clusters is treated as an audio word. For the dictionary construction, $400 \times k$ MFCC feature vectors are sampled from the training data (this figure has experimentally given satisfactory results). Once an audio vocabulary of size k ($k = 1000$ in this work) is built, each MFCC feature is assigned to the closest audio word in terms of Euclidean distance. Subsequently, a histogram is computed for each shot extracted from movies in the training dataset and the related shot is represented by a BoAW histogram representing the audio word occurrences.

3.1.3 Visual Representation

For the visual representation of video shots, the average motion which film-makers usually make use of in order to elicit some particular perception in the audience is adopted. Motion vectors are computed by block-based estimation and then average motion is deduced as the average magnitude of all motion vectors. This process is performed only around the keyframe of video shots (i.e., average motion values are computed between the keyframe and its preceding and succeeding frames, respectively).

3.2 Violence Detection Model

We train three two-class SVMs in order to learn violence models. One SVM model is constructed using low-level audio features, the second one using low-level visual features, and the third using mid-level audio features. As the last step, we fuse the predictions of the latter two SVM models using another two-class SVM for the final prediction as shown in Figure 1(b). In the learning step, the main issue is the problem of imbalanced data. In the training dataset, the number of non-violent video shots is much higher than the number of violent ones. This results in the learned boundary being too close to the violent instances. Consequently, the SVM tends to classify every sample as non-violent. Different strategies to "push" this decision boundary towards the non-violent samples exist. Although more sophisticated methods dealing with the imbalanced data issue have been proposed in the literature (see [11] for a comprehensive survey), we choose, in the current framework, to perform random undersampling to balance the number of violent and non-violent samples. This method proposed by Akbani et al. [3] appears to be particularly adapted to the application context of our work. In [3], different under- and oversampling strategies are compared. According to the results, SVM with the undersampling strategy provides the most significant performance gain over standard two-class SVMs. In addition, the efficiency of the training process is improved as a result of the reduced training data and, hence, is scalable to large datasets similar to the ones used in the context of our work.

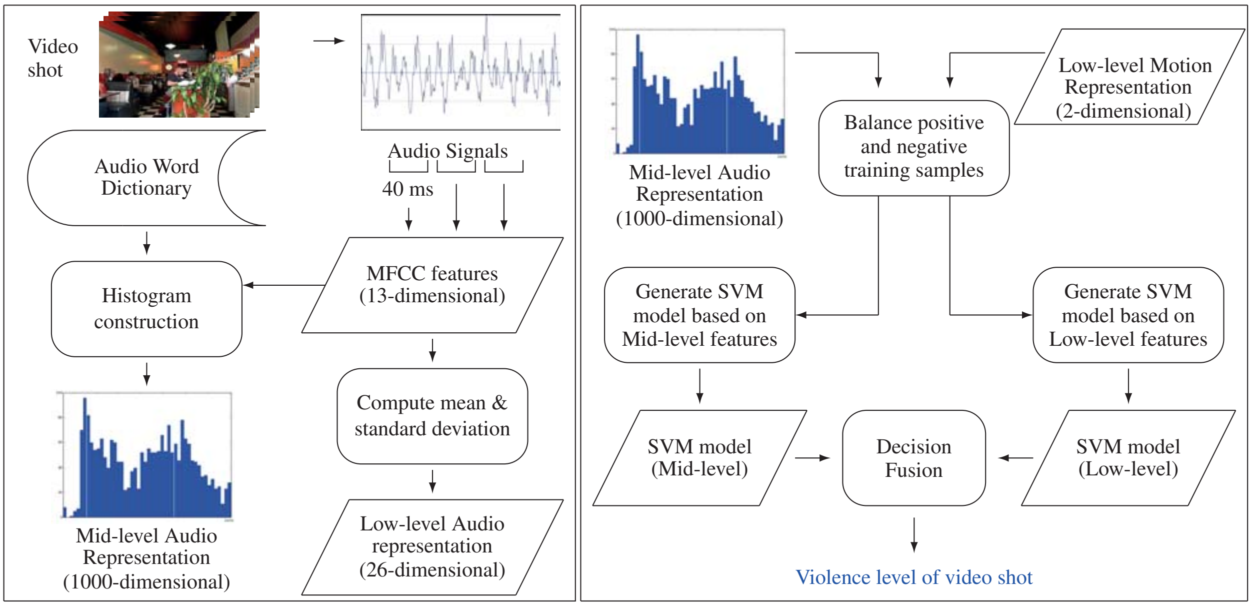
4. PERFORMANCE EVALUATION

The experiments presented in this section aim at comparing the discriminative power of mid-level against low-level audio features. We also evaluate the performance of multi-modal (i.e., mid-level audio and low-level visual) SVM-based decision fusion. A direct comparison of our results with other works discussed in Section 2 is not straightforward due to the differences in the definition of "violence" in published works. However, we compare our method with the methods in the MediaEval VSD task which also stick to the same "violence" definition.

4.1 Dataset and Ground Truth

The MediaEval VSD dataset[1] consists of 32.708 video shots from 18 Hollywood movies of different genres (ranging from extremely violent movies to movies without violence), where each shot is labeled as violent or non-violent. The data set is divided into a training set consisting of 26.138 shots from 15 movies and a test set consisting of 6.570 shots from the remaining 3 movies. The movies of the training and test set were selected in such a manner that both

[1] https://research.technicolor.com/rennes/vsd/


Video shot
Audio Signals
40 ms
Audio Word Dictionary
MFCC features (13-dimensional)
Histogram construction
Compute mean & standard deviation
Mid-level Audio Representation (1000-dimensional)
Low-level Audio representation (26-dimensional)
Mid-level Audio Representation (1000-dimensional)
Balance positive and negative training samples
Low-level Motion Representation (2-dimensional)
Generate SVM model based on Mid-level features
Generate SVM model based on Low-level features
SVM model (Mid-level)
Decision Fusion
SVM model (Low-level)
Violence level of video shot

(a)

(b)

Figure 1: (a) The generation process of audio representations for video shots of movies. (b) The learning phase of the method.

training and test data contain movies of variable violence levels (extreme to none). On average, around 11.5% of shots are annotated as violent in both datasets. Ground truth was generated by 9 human assessors, partly by developers, partly by possible users. Violent movie segments are annotated at the frame level. Automatically generated shot boundaries with their corresponding key frames are also provided for each movie. A detailed description of the dataset and the ground truth generation are given in [7] and [8], respectively.

4.2 Experimental Setup

We employed the MIR Toolbox v1.4[2] to extract the MFCC features (13-dimensional). Frame sizes of 40 ms without overlap are used to align with the 25-fps frames. Features are extracted as explained in Section 3. We trained the two-class SVMs with an RBF kernel using libsvm[3] as the SVM implementation. Training was performed using audio and visual features extracted at the video shot level. We trained one SVM using low-level audio, a second SVM using average motion and a third SVM using mid-level audio features. SVM parameters were optimized by 5-fold cross-validation on the training data.

4.3 Evaluation

Providing a ranked list of violent video shots to the user is more important for our use case. Thus, the evaluation metrics we used are *average precision at 20* and *100* which are also official metrics used in the MediaEval VSD task and *R-precision* which can be seen as an alternative to the *precision at k* in information retrieval.

4.4 Results and Discussions

Table 1 reports the *average precision at 100* values for a baseline method (i.e., random classification) provided by the organizers and for our methods based on mid-level audio and multi-modal (i.e., mid-level audio and low-level visual) SVM-based fusion. The results show that significant improvement is achieved with our approach compared to the baseline method in terms of *average precision at 100*.

Table 1: Average Precision at 100 for the Baseline and Our Methods

Movie	Baseline	Mid-level Audio	Multi-modal
DeadPoets Society	2.17%	15.6%	22.90%
Fight Club	13.27%	29.2%	28.34%
Independence Day	13.98%	72.2%	74.64%

Table 2 provides a comparison of our approach with the best run of participating teams (in terms of *average precision at 20*) in the MediaEval VSD task. The method where we only exploit the audio and disregard the visual modality of videos, manages to be in the top 35% of the submissions in the MediaEval VSD task. The method where we fuse audio cues with motion cues at the decision level, manages to be in the top 25% of the submissions. In addition, the *Mid-level Audio* method, among approaches making use of only one modality (unimodal, i.e., either audio or visual), ranks third among 16 other unimodal submissions in the MediaEval VSD task in terms of *average precision at 100*. The main differences between our approach and the best performing methods in the MediaEval are the fact that we achieve promising results even though we perform no post-processing on violence predictions such as temporal smoothing and the fact that our approach is independent from the concept annotations provided by the organizers, which reduces the need for training data.

Table 3 shows *average precision* (at 20 and 100) as well as *R-precision* for the low- and mid-level audio, low-level visual (i.e., average motion) and the SVM-based decision fusion methods. We observe that the mid-level audio representation provides more precise detections when compared to low-level audio or visual representation. We also note that the performance is further improved by fusing these mid-level audio cues with motion cues using SVM-

[2] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Table 2: Average Precision (AP) at 20 for the Best Run of Teams in the MediaEval VSD Task and Our Methods (VQ: vector quantization, SIFT: Scale Invariant Features Transform, STIP: Spatial-Temporal Interest Points, VSD: Violent Scenes Detection) [1]

Team	Features	Modality	Method	APat20
Shanghai-Hongkong	Trajectory-based features, SIFT, STIP, MFCCs	audio-visual	SVM with chi-squared kernel + temporal smoothing	0.736
ARF	Color, texture, audio and concepts	audio-visual	Multi-layer perceptron	0.701
TEC	Color, motion, acoustic	audio-visual	Bayesian network with temporal integration post-processing	0.669
Multi-modal (ours)	Bag-of-Audio-Words (BoAW) with VQ, average motion	audio-visual	SVM with RBF kernel	**0.545**
TUM	Acoustic energy and spectral, color, texture, optical flow	audio-visual	SVM with linear kernel	0.504
Mid-level Audio (ours)	BoAW with VQ	audio	SVM with RBF kernel	**0.489**
NII	Visual concepts learned from color and texture	visual	SVM with RBF kernel (with chi-square distance)	0.401
LIG-MRIM	Color, texture, bag of SIFT and MFCCs	audio-visual	Fusion of SVMs and *k*-NNs with conceptual feedback	0.286
DYNI-LSIS	Multi-scale local binary pattern	visual	SVM with linear kernel	0.026

based decision fusion.

Table 3: Average Precision (AP) at k (k = 20 and 100) and R-precision (RP) on the Test Dataset

Method	APat20	RPat20	APat100	RPat100
Low-level Visual	0.317	0.253	0.244	0.188
Low-level Audio	0.403	0.323	0.353	0.318
Mid-level Audio	0.489	0.445	0.387	0.355
Multi-Modal	0.545	0.418	0.420	0.357

The evaluation results demonstrate that our method is able to suitably detect violent content such as disasters with explosions (e.g., a plane crash or a man hitting his head). On the other hand, the method wrongly classifies a video shot as violent when the shot contains disasters such as explosions (such explosions were, indeed, not annotated as "violent", since no one is injured) or exciting moments such as strong applauses. Shots which contain no excitement or action, e.g., containing normal speech or music in the background are also easily classified as non-violent. The most challenging violent shots are the ones which are "violent" according to the general definition of violence, but actually only contain actions such as self-injuries, or other moderate actions such as an actor pushing or hitting slightly another actor.

One significant point which can be inferred from the overall results is that the average precision variation of the proposed method is high for movies of varying violence levels (Table 1). Additionally, the method performs better when the violence level of a movie is higher (*Independence Day* is the one having the most violent shots in the test dataset - around 14.5% of all shots). The difference between the results obtained on *Fight Club* and *Independence Day* is most probably due to the nature of violent content present in these movies. The violent actions present in *Fight Club* are under-represented in the training dataset and, consequently, no related audio word(s) could be extracted for these actions (i.e., *Fight Club* has no proper representation in terms of audio words).

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an approach for the detection of violent content in movies at video shot level. We showed that mid-level audio features (BoAW) provide a better performance than low-level audio and visual features. We also fused these mid-level audio cues with motion cues at the decision level for further improvement and achieved promising results in terms of average precision. Incited by the promising results obtained for this work, we currently investigate the construction of more sophisticated mid-level feature representations. Within this context, an interesting research question is whether augmenting the feature set by including mid-level motion features helps further improving classification. In addition, we plan to extend our approach to user-generated videos.

6. ACKNOWLEDGMENTS

We would like to thank *Technicolor* (`http://www.technicolor.com/`) for providing the ground truth, video shot boundaries and the corresponding keyframes which have been used in this work.

7. REFERENCES

[1] MediaEval VSD task proceedings, 2012. `http://ceur-ws.org/Vol-927/`.

[2] VSD affect task, 2012. `http://www.multimediaeval.org/mediaeval2012/`.

[3] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. *ECML*, pages 39–50, 2004.

[4] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15:561–568, 2002.

[5] B. Bushman and L. Huesmann. Short-term and long-term effects of violent media on aggression in children and adults. *Archives of Pediatrics & Adolescent Medicine*, 160(4):348, 2006.

[6] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. Horror video scene recognition via multiple-instance learning. In *ICASSP*, 2011.

[7] C. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2012 Affect Task: Violent Scenes Detection in Hollywood Movies. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[8] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. A benchmarking campaign for the multimodal detection of violent scenes in movies. In *ECCV*, 2012.

[9] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. Audio-visual fusion for detecting violent scenes in videos. *Artificial Intelligence: Theories, Models and Applications*, pages 91–100, 2010.

[10] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao. Detecting violent scenes in movies by auditory and visual cues. *Advances in Multimedia Information Processing 2008*, pages 317–326, 2008.

[11] H. He and E. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[12] J. Lin and W. Wang. Weakly-supervised violence detection in movies with audio and video based co-training. *Advances in Multimedia Information Processing*, pages 930–935, 2009.

[13] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. *Signal Processing Series*, 1993.

[14] R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *CVPR*, 2005.