Automatic Violence Scenes Detection: A Multi-Modal Approach

Gabin Gninkoun
Computer Science Department
University of Geneva
Switzerland
gabin.gninkoun@gmail.com

Mohammad Soleymani
Computer Science Department
University of Geneva
Switzerland
mohammad.soleymani@unige.ch

ABSTRACT

In this working note, we propose a set of features and a classification scheme for detecting automatically violent scenes in movies. The features are extracted from audio, video, and subtitles modalities of the movies. In violent scenes classification, we found the following features relevant: the short time audio energy, motion component, and shot words rate. We classified the shots into violent and non-violent using naïve Bayesian, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) targeting to maximize the precision of the detection in the first two minutes of retrieved content.

Keywords

Violence, audio feature extraction, visual feature extraction, text-based features, subtitles, violence scenes detection, classification

1. INTRODUCTION

Visual media is nowadays full of violent scenes. Therefore, multimedia content is rated to protect minors or warn the viewers for graphic or inappropriate images.
Manual rating of all existing content is not feasible for the fast growing digital media. An automatic method that can detect violence in movies, including verbal, semantic and visual violence can help video on demand services as well as online multimedia repositories to rate their content.
In this task, we have used audio, visual and text modalities to detect violent scenes in movies at shot level. Despite its importance, this problem has not been extensively addressed in literature. Giannakopoulos et al. combined visual and audio features to design a multi-modal fusion process [3]. Two individual kNN classifiers (audio based and visual based) were trained in order to distinguish violence and non-violence at segment level. de Souza et al. developed a violence segment detector based on the concept of visual codebook with usage of a Linear Support Vector Machines (LSVM). The visual codebook was defined using a k-means clustering algorithm. The input video data was segmented into shots, which were converted into bags of visual words [1]. The current study's task and its dataset are provided by Technicolor for the MediaEval benchmarking initiative 2011. The details about the task, the dataset and annotations are given in the task overview paper [2].

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

2. FEATURES AND METHODS

2.1 Proposed Content-Based Features

2.1.1 Audio-Visual features

The extracted audio features are: *energy entropy, signal amplitude, short time energy, zero crossing rate, spectral flux, spectral rolloff.* A more detailed description can be found in [3]. In visual modality, we extracted the *shot length*, the *shot motion component*, the *skewness of the motion vectors* and the *shot motion content.* The description of the technique used to compute the shot motion component is given in [6].

2.1.2 Text-based features

The subtitles available for all DVDs carry a semantic information of the movie content. We have parsed the file content in a set of *CaptionsElement* where *CaptionsElement* is an object having the four attributes (*num*, *startTime*, *stopTime*, *stemmedWords*). The attribute *num* corresponds to the dialogue position in the subtitles file content.
Each dialogue text is first tokenized. The English stops words were first removed. Then, we used WordNet [5] to remove names from the remaining words. Afterwards, we applied the Porter stemming algorithm [8] on to get the stemmed words. Two features have been derived from the text modality: *shot words rate* (SWR) and *shot swearing words rate* (SSWR). We defined SWR as the estimated amount of words in a shot. Similar to SWR, SSWR corresponds to the estimated amount of swearing words in a shot. To compute SSWR, a list of the 341 most currently used swearing words was obtained from a swearing words dictionary (http://www.noswearing.com/dictionary) and used in our swearing words detector.
All the proposed content-based features have been extracted using shot boundaries provided by MediaEval [2]. In total, we extracted 15 features/statistics from three modalities.

2.2 Discriminant Analysis and Post-Processing

Three different classifiers were applied to detect violent shots, namely, QDA, LDA and naïve Bayesian classifier. A post-processing on the results of QDA was also done to consider the temporal correlation between consecutive shots. The post-processing consists on smoothing the confidence scores for the violent class from QDA using weights found from the transition probabilities on the training set.

3. EXPERIMENTS AND RESULTS

According to the requirements, we have generated five different runs trying different classifiers with prior probabilities. These characteristics are listed in Table 1.

Table 1: The classifiers and prior probabilities for five submitted runs. p_n is the prior probability for the non-violent and p_p is for violent.

Run	Classifier	p_n	p_p
1	LDA	0.5	0.5
2	LDA	0.3	0.7
3	Naïve Bayesian	0.5	0.5
4	QDA	0.5	0.5
5	QDA + Post-processing	-	-

3.1 Evaluation criteria and Classifier selection

The goal of the violence detection in the proposed use case scenario is to provide the user with the most violent shots in the movie. We defined an evaluation criteria based on this use case scenario as follows. The detected violent shots were first ranked based on their confidence scores, the first 2 minutes on the top of the list were set aside as the retrieved content. The precision, recall and F1 score were then computed for the top ranked two minutes shots. We used a K-fold cross validation with $K = 11$ and different prior probabilities for each class. The best performance was achieved using LDA and QDA methods with equal prior probabilities for both classes.

3.2 Post-processing

The results of the last run correspond to the post-processing of the fourth run. A weighted average of the confidence scores was used to smooth the violent shots' decisions. The weights are given in Table 2 where the first row represents the probabilities of transition to a violent shot while the four neighbouring shots are non-violent. The second row of the table represents the transition probabilities for transition to a violent shot while the neighbouring shots are violent. These values were obtained from the training set. The post-processing reduced the false positives significantly.

Table 2: The transition probabilities were computed on a five consecutive shots window.

	1	**2**	**3**	**4**	**5**
non-violent	0.04	0.03	0	0.03	0.04
violent	0.67	0.77	1	0.77	0.67

We ultimately obtained the best result with the with minimum MediaEval cost $C \approx 2.02$ and recall $r = 0.87$ (Table 3) using LDA with prior probabilities 0.3 and 0.7 respectively for non-violent class and violent. However, if we look at both F1 score and MediaEval cost the fourth run which was with QDA and equal prior probabilities performed better. These results matched our expectations from the cross validation results on the training set.

4. CONCLUSIONS

We have proposed a set of features to automatically detect violent material at shot level for commercial movies. The performance of the proposed system have been evaluated based on a detection cost function weighting by false alarms and missed detections rate. The short time energy, the motion component and the shot words rate are proposed and used as relevant features to classify a movie's shots as violent or non-violent. The proposed methods were unable to detect all the violent scenes without sacrificing the false positive rate. This is due to the fact that the proposed features are not enough to capture all violent actions or events. Automatic detection of more high level concepts such as scream, explosion, or blood are needed to improve the detections.

Table 3: Violence detection system evaluation results for the 5 submitted runs at shot level.

Run	Precision	Recall	F-measure	MediaEval cost
1	0.174	0.377	0.238	6.522
2	0.164	0.870	0.276	2.024
3	0.183	0.426	0.256	6.049
4	0.178	0.774	0.289	2.838
5	0.252	0.077	0.119	9.252

5. ACKNOWLEDGEMENTS

This work is supported by the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement Petamedia No. 216444.

6. REFERENCES

[1] F. D. M. de Souza, G. C. Chavez, E. A. do Valle Jr., and A. de A. Araujo. Violence detection in video using spatio-temporal features. *Graphics, Patterns and Images, SIBGRAPI Conference on*, 0:224–230, 2010.

[2] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2011 Affect Task: Violent Scenes Detection in Hollywood Movies. In *Working notes Proceeding of Medieval workshop*, Pisa, Italy, September 2011.

[3] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In S. Konstantopoulos et al., editor, *Artificial Intelligence: Theories, Models and Applications*, volume 6040 of *Lecture Notes in Computer Science*, pages 91–100. Springer Berlin / Heidelberg, 2010.

[4] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao. Detecting violent scenes in movies by auditory and visual cues. In Y.-M. Huang et al., editor, *Advances in Multimedia Information Processing - PCM 2008*, volume 5353 of *Lecture Notes in Computer Science*, pages 317–326. Springer Berlin / Heidelberg, 2008.

[5] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[6] Z. Rasheed and S. Mubarak. *Video categorization using semantics and semiotics*. PhD thesis, Orlando, FL 32816, USA, 2003. AAI3110078.

[7] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Trans. Circuits Syst. Video Technol.*, 15(1):52–64, 2005.

[8] P. Willett. The Porter Stemming Algorithm: Then and Now. *Program: Electronic Library and Information Systems*, 40(3):219–223, 2006.