ViVoLab and CVLab - MediaEval 2014: Violent Scenes Detection Affect Task

Diego Castán, Mario Rodríguez, Alfonso Ortega, Carlos Orrite, Eduardo Lleida
Universidad de Zaragoza - ViVoLab and CVLab
Zaragoza, Spain
[dcastan, mrodrigo, ortega, corrite, lleida]@unizar.es

ABSTRACT

This paper describes the ViVoLab/CVLab system to provide segments of violent scenes from Hollywood movies and "wild-user" videos from the internet. We propose a system based on a fusion of acoustic features, audio concepts and video features. Our joint audio-visual approach achieves MAP2014 values of 17.81% and 43.03%, for the main task and the generalization task, respectively.

1. INTRODUCTION

Due to the popularity of video-sharing websites, there is a significant interest in multimedia analysis. The MediaEval evaluations aim at analyzing multimedia content for several purposes. We evaluate the detection of violent scenes in movies and in "wild-user" videos from the internet in the "Violent Scenes Detection" task. A complete description of the task, the metric and the databases can be found in [2].

We propose a system based on Gaussian Mixture Models with Hidden Markov Models (GMM/HMM) and Support Vector Machine (SVM) that is able to segment video streams into violent or non-violent. The GMM/HMM approaches have been widely used in audio to provide a clear segmentation for different tasks and the SVM provides good results in classification in low-level video features approaches. The classes are modeled with mid-level and low-level features extracted from the audio track and low-level features extracted from the video frames.

2. SYSTEM

We have developed a multimodal system that makes use of different audio concepts. The concepts are modeled with GMMs and the normalized log-likelihood of each concept is used as mid-level features. The mid-level features are merged with video features and with low-level acoustic features. Figure 1 shows a diagram of the different parts of the system. The development set has been divided in two sets. The first set is composed of five movies ("Saving Private Ryan", "Pirates of the Caribbean", "The Sixth Sense", "The Bourne Identity" and "Fight Club") and it has been used to model the audio concepts and the violence/non-violence detector with the video features. The second set is composed of three movies ("Fargo", "Pulp Fiction" and "The Pianist") and it has been used to train the final classification system that employs a fusion of audio and video features. We restricted ourselves to a smaller subset for training due to resource constraints.

MediaEval 2014 Workshop October 16-17, 2014, Barcelona, Spain

2.1 Audio Approach

2.1.1 Acoustic Features

The audio from each video has been extracted and resampled to 16KHz. First, we extract 16 MFCC (including C0) computed in time windows of 25ms with 10ms steps. We concatenate these MFCC with six features that have been shown to perform well to describe the audio concepts related with violence [1]. These features are the entropy energy, the short-time energy, the spectral centroid, the spectral entropy, the spectral roll-off and the spectral flux. The delta and delta-delta are computed for all the features (16-dim.+6-dim.) and are normalized to get zero mean and unit variance. The delta and delta-delta are concatenated with the original feature vector yielding a 66-dim. vector. After that, the mean and standard deviation are computed over 1 second windows with an overlap of 0.25 seconds. Therefore, the system uses 132 features (mean(66-dim.)+std(66-dim.)) every 0.25 seconds to model the audio concepts and the violence/non-violence decisions.

2.1.2 Audio Concept Models

We have modeled three sets of audio concepts: gunshots, explosions and screams. A complete description about the specific audio concepts can be found in [2]. Each concept is modeled with a GMM of 1024 Gaussians trained with an Expectation-Maximization algorithm. The log-likelihoods of the concepts are concatenated in a 13-dim. vector known as mid-level features in the literature.

2.2 Video Approach

The video approach has been performed using the Improved Dense Trajectories (IDT) [3], which are the state of the art in unconstrained video action recognition. We first have extracted 256,000 randomly selected IDT from the 5 pre-selected movies. Each of these low level features is extracted from a 15 frames window obtaining a vector of 426-dim. (30 from the trajectory, 96 from the Histograms of Oriented Gradients, 108 from the Histograms of Optical Flow, and 96 for each of the x and y Motion Boundary Histograms). These vectors are reduced to half using a PCA dimensionality reduction process. Then, the obtained features are used to create a GMM of 256 Gaussians with which we obtain Fisher Vectors (FV) from the videos. We depict the recognition system in Figure 1. From each of the 5 movies in the pre-selected set we obtain 100 video segments of several lengths (30, 90, 150, 210, 270, 330 frames), with the position and the length of each segment randomly selected. From those videos segments, 40 belong to violent

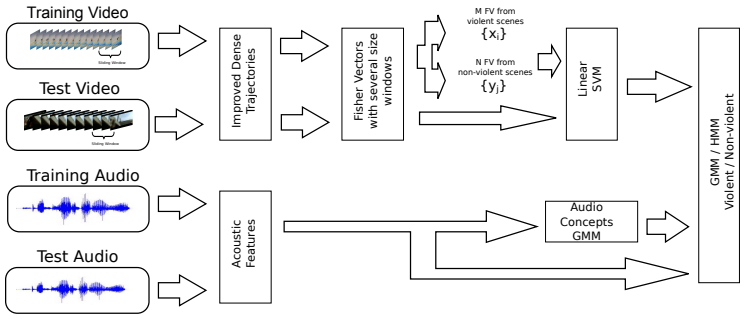

Figure 1: General diagram of the ViVoLab/CVLab for Violent Scenes Detection in MediaEval 2014

scenes, and 60 to non-violent scenes in each movie, resulting in a total of 200 violent scenes $\{x_i\}$ and 300 non-violent scenes $\{y_j\}$ for training. In each of the segments we extract the IDTs which after dimensionality reduction are used to obtain a FV of 109,056-dim. The FVs are then normalized with power and L2 normalizations. These final 500 FVs of violent and non-violent scenes are used to train a SVM. Due to the high dimensionality of the FV we train a Linear-SVM. Finally, scene recognition is performed in segments selected with six sliding windows. All of the windows move every 30 frames (1.2s) but each one has a different size (30, 90, 150, 210, 270, 330 frames). The score is then obtained every 30 frames, for each of the windows we obtain the final score $f_t^w = \frac{1}{N}\sum_{j=0}^{N-1} s_{t-jw}^w$ where s_t^w is the score obtained from the window of size w starting at frame t, and N is the number of times a 30 frames segment fits in the window. In the boundaries of the movies there are less than N scores and therefore, the final score is the average of the available ones. Finally, we have selected two scores from the visual approach, f^{150} and $f = (f^{30} + f^{90} + f^{150})/3$, after evaluating all the window sizes with the MAP2014 metric into the second pre-selected set of 3 movies.

2.3 Fusion

The acoustic features (low-level features), the audio concept log-likelihoods (mid-level features) and the violence/non-violence scores from video are concatenated in a into 147-dim. vectors. These vectors are used as features to model a GMM/HMM-based violence classification system. Each class has the same number of Gaussians and states with a left-to-right topology to smooth the transitions between classes.

3. RESULTS

Five systems were evaluated for both tasks without changing any of their parameters. Runs #1 and #2 are based only on the audio approach with 64-Gauss/6-states and 128-Gauss/7-states respectively. Runs #3, #4 and #5 are based on the video/audio fusion described in the last section with 128-Gauss/6-states, 512-Gauss/7-states, and 1024-Gauss/8-states respectively.

Figure 2 shows the results with the metric proposed in MediaEval 2014. It can be seen how the runs based on audio/video (#3, #4 and #5) perform better than the runs with only audio (#1 and #2). The best run for the main task is #4 with 17.81% MAP2014 and the best system for the generalization task is #3 with 43.03% MAP2014.

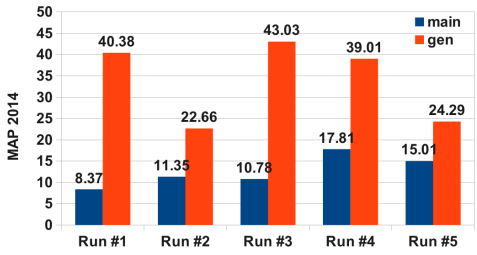

Figure 2: Results for the main task and the generalization task with the MAP2014 metric

4. CONCLUSIONS

We have proposed 5 different runs based on audio and audio/video. The joint audio/video systems perform better than the audio-only systems for both tasks thanks to the complementary information. The system identifies better the violence in the generalization task maybe because the violence is more evident in the internet videos. The runs with less states identifies the short violent segments (like in the generalization task) better than the runs with more states. Also, shorter windows in the video approach perform better probably because they do not mix violent and non-violent scenes. On the other hand, the runs with more states achieve better results in the main task since the violent segments are longer.

Acknowledgements

This work has been funded by the Aragon Government, Spanish Government and the European Union (FEDER) under the projects TIN2011-28169-C05-02, TIN2013-45312-R and TAMA project. Mario Rodriguez has got a FPI grant.

5. REFERENCES

[1] T. Perperis and T. Giannakopoulos. Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. *Expert Systems with Applications*, 38(11):14102–14116, 2011.

[2] M. Sjöberg, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty. The MediaEval 2014 affect task: Violent scenes detection. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, October 2014.

[3] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV 2013 - IEEE International Conference on Computer Vision*, pages 3551–3558, December 2013.