Violent Scenes Detection with Large, Brute-forced Acoustic and Visual Feature Sets

Florian Eyben, Felix Weninger, Nicolas Lehment, Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
Theresienstrasse 90, 80333 München, Germany
eyben@tum.de

Björn Schuller
JOANNEUM RESEARCH*
DIGITAL - Institute for Information and Communication Technologies
Graz, Austria
bjoern.schuller@joanneum.at

ABSTRACT

This paper describes the TUM approaches for violent scenes detection in movies, submitted for the MediaEval 2012 Affect Challenge. Score fusion is used to fuse Support-Vector Machine (SVM) confidence scores assigned to short fixed length windows within each movie shot. SVM predictors for acoustic and visual channels are trained. For the acoustic channel, a large set of acoustic features based on the set from the INTERSPEECH 2012 Speaker Trait Challenge is employed. A comprehensive set of common video low-level descriptors such as optical flow, gradients, and hue and saturation histograms is used for the visual channel.

1. INTRODUCTION

The objective of the violent scenes detection task is to detect violence in movies on a shot level using multi-modal features. For details on the task, we refer to the paper describing the affect task [1]. In the following we describe our approach.

2. METHOD

Our method for detection of violent scenes uses Support Vector Machine (SVM) classifiers which are trained on features extracted from the development data. Independent classifiers are trained on acoustic and visual features. To obtain a single decision and confidence score for each shot in the test data, the predictions made by the acoustic and visual SVMs are fused by simple score averaging.

Our feature extraction method follows our standard static brute-force approach from the domain of affect recognition and paralinguistic audio analysis. Thereby low-level descriptors (LLDs) are summarised over segments of variable or fixed length by applying statistical functionals, such as mean, standard deviation, quartiles, regression coefficients, etc to the LLDs. This way, LLD series of variable length can be mapped onto a single feature vector. However, the feature vectors are affected by variations in segment length. Longer segments will contain more information, possibly violence mixed with non-violence or simply different violent or non-violent content. As the shots provided for the affect task vary considerably in length, we decided to use fixed length sub-segments of the shots. In pre-evaluation runs, we found that 2 seconds long sub-segments gave good results. We investigated both, overlapping sub-segments sampled at a rate of .5 seconds, and non-overlapping sub-segments. The label for each of the sub-segments is inferred from the violent segment ground truth annotation as follows: If a shot sub-segment overlaps with a violent segment in some way, the shot sub-segment is labelled as violent; it is labelled as non-violent otherwise. We would like to note here that a single shot can contain violent and non-violent sub-segments because the boundaries of the violent segments are not aligned to the shot boundaries.

Table 1: Acoustic and visual low-level descriptors.

4 acoustic energy LLDs
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
Logarithmic energy, and zero-crossing rate
33 acoustic spectral LLDs
MFCC 1–16
Spectral energy 40–150, 250–650 Hz, 1 k–4 kHz, 5 k–15 kHz
Spectral roll-off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewness, kurtosis, slope, psychoacoustic sharpness, harmonicity, centroid
95 visual LLDs
Normalised HSV histograms (20, 20, 10 bins)
Normalised dense Optical Flow histograms (20 bins)
Normalised Laplacian edge histograms (20 bins)
Mean Optical Flow
Optical Flow standard deviation
Strongest edge in lower 98 % of Laplacian edges

Extraction of acoustic features is done with our open-source feature extraction toolkit openSMILE [2]. The 37 acoustic LLDs, given in Table 1 are extracted from overlapping frames with a length of 25 ms at a rate of 10 ms. For F_0 based features (fundamental frequency, probability of voicing, jitter/shimmer) the frame size is 60 ms. The frame sampling rate of 10 ms is unchanged. 51 functionals (cf. Table 2) are applied to the acoustic LLDs and their first order delta coefficients over windows with a 2 second maximum length. A total of $37 \cdot 2 \cdot 51 = 3\,774$ acoustic features is obtained.

The low level video features are computed for each frame and consist of Hue-Saturation-Value (HSV) histograms, an optical flow analysis and a Laplacian edge detection. Three, dimensionally independent, normalised HSV histograms (20, 20 and 10 bins) are computed. A dense Farneback optical flow analysis compares consecutive frames for pixel-wise displacements. The magnitudes of the resulting 2D displacement vectors are computed, thresholded to a maximum displacement of 15 % of the normalised frame size and

*This author is further affiliated with Technische Universität München, Germany

Table 3: Results on test set, mean average precision (MAP) at top 100 and Acoustic Event Detection (AED) MediaEval (cf. [1]) cost. Results on development set with 3-fold cross validation (CV), MAP at top 100 and top 20, and unweighted/weighted average recall (UAR/WAR).

	Test data		Development data (3-fold CV)			
Configuration	MAP100	AED MediaEvalCost	MAP100	MAP20	UAR	WAR
TUM-1	0.484	7.83	0.397	0.525	0.584	0.848
TUM-2	0.376	6.85	0.445	0.515	0.648	0.830
TUM-3	0.360	6.83	0.428	0.518	0.648	0.826
TUM-4	0.392	7.27	0.442	0.503	0.634	0.829
TUM-5	0.320	6.67	0.224	0.213	0.537	0.832

Table 2: 51 functionals applied to acoustic and visual low-level descriptors and delta coefficients.

quartiles 1–3 and 3 inter-quartile ranges
1 % percentile (≈ min), 99 % percentile (≈ max)
percentile range 1 %–99 %
position of min / max, range: max-min
arithmetic mean, root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above 90 % / below 25 % of range
rel. duration LLD is rising / falling
gain of linear prediction (LP) and LP Coefficients 1–5
range of peaks (absolute and rel. to arith. mean)
mean value of peaks (absolute and rel. to arith. mean)
mean value of peaks – arithmetic mean
mean value of minima rel. to arith. mean
max, min, mean, std. dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
linear regression slope, offset, and quadratic error
quadratic regression coefficient 1, and quadratic error

sorted into 20 bins. The resulting histogram is then normalised. Next, the mean optical flow and its standard deviation are determined. These frame-to-frame motions are expected to yield information concerning the overall pacing of the current scene. Furthermore, high standard deviations on optical flow would signify non-uniform scene flow while high mean flows could indicate a fast-paced scene. Finally, Laplacian edge detection is used for a simple detection of motion blur. An edge image is computed per frame, the 2 % strongest edges are discarded as noise and the remaining strongest edge is used as a feature. Additionally, a normalised magnitude histogram of the edge image is calculated, ignoring values close to zero (histogram range: 16–255, 8-bit edge image). All 95 visual descriptors are given in table 1. 51 functionals (cf. Table 2) are applied to the frame-wise visual LLDs and their first order delta coefficients with openSMILE in order to summarise the low-level descriptor features over windows with a 2 second maximum length. In this way a total of $95 \cdot 2 \cdot 51 = 9\,690$ video features are obtained.

All SVMs in our experiments use a linear kernel function and a complexity $C = 0.01$. Training has been done with the WEKA [3] implementation of the Sequential Minimal Optimisation Algorithm (SMO). In order to obtain a shot-level result, we linearly fuse the SVM scores of all sub-segments within a shot. In the following the configurations for the 5 run submissions are summarised:

- TUM-1: audio+video, train (audio): overlapping sub-segments (rate: .5 seconds); train (video) and test (both): non-overlapping sub-segments.
- TUM-2: audio, train: overlapping sub-segments (rate: .5 seconds); test: non-overlapping sub-segments.
- TUM-3: audio, train+test: overlapping sub-segments (rate: .5 seconds).
- TUM-4: audio, train+test: non-overlapping sub-segments.
- TUM-5: video, train+test: non-overlapping sub-segments.

The differences are mainly in the rate at which the 2 second sub-segments of the shots are sampled from the training and testing data. Run TUM-1 is the only run where acoustic and visual features are fused, and TUM-5 is a run with the visual features alone.

3. RESULTS

Table 3 shows the results of our 5 approaches on the MediaEval affect test set and the development set. We use 3-fold cross validation on the development set. The fold division by movie title is available upon request. We report the official challenge metric mean average precision (MAP) for the top 100 scored shots. The best performing configuration on the test data set is TUM-1. It is also the best on the development data, concerning MAP at 20. However, MAP at 100 is worse than the other audio only configurations.

4. CONCLUSION

The obtained results demonstrate the feasibility of our approach. Audiovisual fusion gives the best results on the test set. Yet overall, we found the results, esp. MAP, to be quite sensitive to variations in the way the data are segmented and the labels are assigned (i. e., if the labels are assigned per segment or per shot, and if the segments are aligned to the shot boundaries, or to the ground truth of the violent segments). In future work the issue of proper segmentation and a deeper analysis of the discriminative power of single features should be carried out, to improve the precision of the systems.

5. REFERENCES

[1] C. H. 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.

[2] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *Proceedings of ACM Multimedia 2010*, Florence, Italy, October 2010, pp. 1459–1462.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, “The WEKA Data Mining Software: An Update,” in *SIGKDD Explorations*, vol. 11. 2009.