DAI Lab at MediaEval 2012 Affect Task: The Detection of Violent Scenes using Affective Features

Esra Acar [1], Sahin Albayrak [2]
DAI Laboratory, Technische Universität Berlin
Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
[1] esra.acar@dai-labor.de [2] sahin.albayrak@dai-labor.de

ABSTRACT

We propose an approach to detect violence in movies at video shot level using low-level and mid-level features. We use audio energy, pitch and Mel-Frequency Cepstral Coefficients (MFCC) features to represent the affective audio content of movies. For the affective visual content, we extract average motion information. To learn a model for violence detection, we choose a discriminative classification approach and use a two-class support vector machine (SVM). Within this task, we investigate whether affect-related features provide good representation of violence and also whether making abstractions from low-level features are better than directly using low-level data for the task.

1. MOTIVATION AND RELATED WORK

The MediaEval 2012 Affect Task aims at detecting violent segments in movies. Detailed description of the task, the dataset, the ground truth and evaluation criteria are given in the paper by Demarty et al. [3].

The affective content of a video is defined as the intensity (i.e. arousal) and type (i.e. valence) of emotion (both are referred to as affect) that are expected to arise in the user while watching that video [4]. Recent research efforts on affective content analysis of movies ([5], [6], [7], [8]) propose methods to map low-level features (e.g. low-level audio-visual features of videos and users' physiological signals) to high-level emotions. A recent work on horror scene recognition [1] has demonstrated that describing violence by affect-related features is a promising approach. Film-makers intend to elicit some particular emotions which is usually referred as expected emotions in the audience. When we refer to violence as an expected emotion in videos, affect-related features are applicable to represent violence in movies.

2. PROPOSED APPROACH

We propose an approach that uses affect-related audio and visual features to represent violence in the video shots of movies. For the representation of video content, low-level audio and visual features are extracted. For the MFCC features of the video shots, an abstraction is applied and mid-level representations are constructed. The details of this process are explained in the following subsection. The audio and visual features are then fused at feature-level and a two-class SVM is trained to learn a violence model.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2.1 Audio Representation

Audio signals of video shots of movies are represented by audio energy, pitch and MFCC that are commonly used to represent the affective content of videos. We use audio energy and MFCC features to describe the arousal aspect of emotion, where pitch features are used for the valence aspect. The durations of the annotated video shots vary and as a result, each video shot has different numbers of audio energy, pitch and MFCC feature vectors. To obtain audio representations having the same dimension, we compute mean and standard deviation for audio energy, pitch and each dimension of the MFCC features.

For MFCC features, we also apply an abstraction to generate mid-level audio representations for video shots. We use an MFCC-based Bag of Audio Words (BoAW) approach. In [2], constructing a dictionary by only using normal event videos is proposed and reconstruction cost of videos is computed to detect abnormal events. Inspired by this work, we think that violence can also be seen as abnormality. The frequency of violent words is expected to be higher and the frequency of non-violent words is expected to be lower for video shots containing violence. Therefore, we construct two different audio vocabularies, one containing violence words and one containing non-violence words. For the construction of violence words, we use the MFCC of movie segments that are annotated as violent, whereas non-violence words are constructed by using non-violent movie segments. Each vocabulary is constructed by clustering MFCC feature vectors with the k-means clustering algorithm. Each resulting cluster is treated as an audio word. Once a vocabulary of size k ($k = 222$ for violence dictionary and $k = 959$ for non-violence dictionary in this work) is built, each MFCC feature is assigned to the closest audio word (Euclidean distance is used), a BoAW histogram is computed that represents the violence and non-violence word occurrences within a video shot.

2.2 Visual Representation

For the visual representation of video shots, average motion information which is also commonly used to represent arousal aspect of affective content of videos is preferred. Motion vectors are computed using block-based motion estimation and then average motion is found as the average magnitude of all motion vectors.

2.3 Results and Discussion

The aim of this work was to assess the performance of low-level and mid-level affect-related features for violence detec-

tion. We evaluated our approach on 3 Hollywood movies from the MediaEval 2012 dataset [3]. We submitted five runs in total for the MediaEval 2012 Affect Task: *r1-low-level*, *r2-mid-level-100k*, *r3-mid-level-300k*, *r4-mid-level-300k-default* and *r5-mid-level-500k*. In *r1-low-level*, we used low-level audio and visual features for representation, while in the remaining submissions, we used mid-level MFCC representations instead of low-level MFCC features. The differences between these remaining runs are the number of instances that are used to construct the BoAW dictionaries and also the SVM parameter values. For *r2-mid-level-100k*, *r3-mid-level-300k*, *r4-mid-level-300k-default* and *r5-mid-level-500k*, in the construction of dictionaries, we used 100k, 300k, 300k and 500k samples (where k represents the number of words in the related dictionary), respectively. The difference between *r3-mid-level-300k* and *r4-mid-level-300k-default* which use the same number of samples to construct dictionaries is the values of SVM parameters. In the training phase, we applied a two-class SVM with RBF kernel. Since the number of non-violent samples is much higher than the number of violent ones, we perform undersampling by choosing random non-violent samples to balance the training data. Parameter optimization was performed by 5-fold cross validation and the values that provide the best accuracy are chosen as SVM parameters. LibSvm[1] was used as the SVM implementation. We employed the MIR Toolbox v1.4[2] to extract the audio energy, pitch and 13-dimensional MFCC, where for average motion extraction, we used a public motion estimation toolbox[3]. Table 1 reports AED [3] precision, AED recall and AED F-measure values, where Table 2 shows the evaluation results for the submitted runs.

Table 1: AED precision, recall and F-measures at video shot level

Run	AED-P	AED-R	AED-F
r1-low-level	0.141	0.597	0.2287
r2-mid-level-100k	0.140	0.629	0.2285
r3-mid-level-300k	0.144	0.625	0.2337
r4-mid-level-300k-default	0.190	0.627	0.2971
r5-mid-level-500k	0.154	0.603	0.2457

Table 2: Mean Average Precision (MAP) values at 20 and 100

Run	MAP at 20	MAP at 100
r1-low-level	0.2132	0.18502
r2-mid-level-100k	0.20369	0.14492
r3-mid-level-300k	0.35925	0.18538
r4-mid-level-300k-default	0.15469	0.15083
r5-mid-level-500k	0.15	0.11527

As shown in Table 2, *r3-mid-level-300k* which uses mid-level representation achieves the best MAP values, especially for MAP at 20. For MAP at 100 though, *r1-low-level* achieves comparable results with *r3-mid-level-300k*. From these results, we observe that with mid-level representations, it is possible to achieve a slightly better performance compared to low-level representations. Another notable result is that *r4-mid-level-300k-default* achieved a better performance than *r3-mid-level-300k* in terms of AED precision, AED recall and AED F-measure values, where we used the default SVM parameter values. This also suggests that we should consider revising our SVM parameter optimization process. Finally, from the overall results, we observe that using affect-related features to describe violence needs some improvements. For instance, as mid-level representations we consider to include facial features of persons appearing in videos. We think that this will improve the performance of the system in terms of MAP values.

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
[2] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox`
[3] `http://www.mathworks.com/matlabcentral/fileexchange/8761`

3. CONCLUSIONS AND FUTURE WORKS

The aim of this work is to investigate whether affect-related features are well-suited to describe violence. To detect violence in movies, we merge affect-related audio and visual features in a supervised manner using a two-class SVM. Our main finding is that more sophisticated affect-related features are necessary to describe the content of videos. Especially, the visual representation part needs further development. Our next step in this work is therefore to use mid-level features such as human facial features to represent video contents.

4. REFERENCES

[1] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. Horror video scene recognition via multiple-instance learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[3] 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.

[4] A. Hanjalic and L. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, pages 143–154, 2005.

[5] G. Irie, K. Hidaka, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Latent topic driving model for movie affective scene classification. In *ACM international conference on Multimedia*, 2009.

[6] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *IEEE International Symposium on Multimedia*, 2008.

[7] M. Xu, J. S. Jin, S. Luo, and L. Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *ACM International Conference on Multimedia*, 2008.

[8] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu. A three-level framework for affective content analysis and its case studies. *Multimedia Tools and Applications*, 2012.