TUB-IRML at MediaEval 2014 Violent Scenes Detection Task: Violence Modeling through Feature Space Partitioning

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

ABSTRACT

This paper describes the participation of the TUB-IRML group to the MediaEval 2014 Violent Scenes Detection (VSD) affect task. We employ low- and mid-level audio-visual features fused at the decision level. We perform feature space partitioning of training samples through k-means clustering and train a different model for each cluster. These models are then used to predict the violence level of videos by employing two-class support vector machines (SVMs) and a classifier selection approach. The experimental results obtained on Hollywood movies and short Web videos show the superiority of mid-level audio features over visual features in terms of discriminative power, and a further enhanced performance resulting from the fusion of audio-visual cues at the decision-level. Finally, the results also demonstrate a performance gain obtained by partitioning the feature space and training multiple models, compared to a unique violence detection model.

1. INTRODUCTION

The MediaEval 2014 VSD task aims at detecting violent segments in movies and short Web videos. Detailed description of the task, the dataset, the ground truth and evaluation criteria are given in the paper by Sjöberg et al. [6].

2. THE PROPOSED METHOD

2.1 Video Representation

We represent the audio content using mid-level representations, whereas the visual content is represented at two different levels: low-level and mid-level.

Mid-level audio representations are based on Mel-Frequency Cepstral Coefficient (MFCC) features extracted from the audio signals of video segments of 0.6 second length. We experimentally verified that the 0.6 second time window was short enough to be computationally efficient and long enough to retain sufficient relevant information. In order to generate the mid-level representations, we apply an abstraction process which uses an MFCC-based Bag-of-Audio Words approach with a sparse coding scheme. We employ the dictionary learning technique presented in [5]. In order to learn the dictionary of size k ($k = 1024$ in this work) for sparse coding, $400 \times k$ MFCC feature vectors are sampled from the training data (this $400 \times k$ figure was determined experimentally). In the coding phase, we construct the sparse representation of audio signals by using the LARS algorithm [1]. In order to generate the final sparse representation of video segments, which is a set of MFCC feature vectors, we apply the *max-pooling* technique.

We use **motion-related descriptors** for the visual representation of video segments. One of the motion descriptors is Violent Flow (ViF) which is proposed for real-time detection of violent crowd behaviors. We compute a ViF descriptor for each video segment to represent statistics of flow-vector magnitude changes over time. For a detailed explanation of the computation of this descriptor, the reader is referred to [3].

We also use **static content representations**. More specifically, we employ affect-related static visual descriptors. We compute mean and standard deviation of saturation, brightness and hue in the HSL color space. We also compute the colorfulness of the keyframe of video segments using the method in [2], where the keyframe is deemed to be the frame in the middle of a video segment.

Mid-level visual representations are based on histogram of oriented gradient (HoG) and histogram of oriented optical flow (HoF) features extracted from the visual content of video segments of 0.6 second length. HoG and HoF descriptors are densely sampled and computed for subvolumes of video segments (HoG descriptors are subsampled every 6 frames and HoF descriptors are subsampled every 2 frames as recommended in [7]). The Horn-Schunk method [4] is applied to compute optical flow vectors which are used for the extraction of HoF descriptors. The resulting HoG and HoF descriptors are subsequently used to generate mid-level HoG and HoF representations separately.

2.2 Violence Detection Model

"Violence" is a concept which can audio-visually be expressed in diverse manners. Therefore, learning multiple models for the "violence" concept instead of a unique model constitutes a more judicious choice. This justifies that we first perform feature space partitioning by clustering video segments in our training dataset and learn a different model for each violence sub-concept (i.e., cluster). We use two-class SVMs in order to learn violence models (i.e., one SVM for each sub-concept).

In the learning step, the main issue is the problem of imbalanced data. This is caused by the fact that, in the training dataset, the number of non-violent video shots is

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

Table 1: The MAP2014 and MAP@100 of our method with different representations (i.e., *Run1* where we use MFCC-based mid-level audio representations, *Run2* where HoG- and HoF-based mid-level features and ViF descriptors are used, *Run3* where we use affect-related color features, *Run4* where we use audio and visual features (except color), and *Run5* where all audio-visual representations are linearly fused at the decision level) on the Hollywood movie and the Web video dataset, respectively.

Method	MAP2014 − Movies	MAP@100 − Movies	MAP2014 − Web videos	MAP@100 − Web videos
Run1	0.169	0.368	0.517	0.582
Run2	0.139	0.284	0.371	0.478
Run3	0.080	0.208	0.477	0.495
Run4	0.172	0.409	0.489	0.586
Run5	0.170	0.406	0.479	0.567
SVM-based unique model	0.093	0.302	-	-

much higher than the number of violent ones. We choose, in the current framework, to perform random undersampling to balance the number of violent and non-violent samples (with a balance ratio of 1:2).

In the test phase, the main challenge is to combine the classification results of the violence models. We perform a classifier selection to solve this. More precisely, we first determine the nearest cluster to a video segment of the test set using Euclidean distance measures. Once the classifier for the video sample is determined, the output of the chosen model is used as the final prediction for that video sample.

3. RESULTS AND DISCUSSION

Table 1 reports the mean Average Precision (MAP2014) and MAP@100 metrics on the Hollywood movie dataset and the Web video dataset. We observe that the mid-level audio representation based on MFCC and sparse coding (*Run1*) provides promising performance and outperforms all other representations (*Run2* and *Run3*) that we use in this work. We also note that the performance is further improved by fusing these mid-level audio cues with low- and mid-level visual cues at the decision level by linear fusion (*Run4*). However, the fusion of affect-related color features with the other audio-visual features does not help improving the performance (*Run5*).

The results on the Web video dataset in terms of MAP metrics even demonstrate superior results compared to the ones obtained on the Hollywood movie dataset. Therefore, we can conclude that our violence detection method generalizes particularly well to other contents, including other types of video content not used for training the models. Another interesting observation is that affect-related color features seem to provide better results in terms of MAP metrics on the Web video dataset in comparison to the Hollywood movie dataset (*Run3*).

Table 1 also provides a comparison of our methods in terms of MAP2014 and MAP@100 metrics with an SVM-based unique violence detection model (i.e., a model where no feature space partitioning is performed). We can conclude that our method outperforms the SVM-based detection method where the feature space is not partitioned and all violent and non-violent samples are used to build a unique model.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an approach based on feature space partitioning for the detection of violent content in movies and short Web videos at the video segment level. We employed low- and mid-level audio-visual features to represent videos. We showed that the mid-level audio representation based on MFCC and sparse coding provides promising performance in terms of MAP2014 and MAP@100 metrics and also outperforms our visual representations. We also fused these mid-level audio cues with low- and mid-level visual cues at the decision level using linear fusion for further improvement and achieved better results than unimodal video representations in terms of the MAP metrics. We observed from the overall evaluation results that our method performs better when violent content is better expressed in terms of audio features (a typical example would be a gun shot scene). Hence, as a future work, we need to extend/improve our visual representation set with more discriminative representations. Another possibility for future work is to further investigate the feature space partitioning concept and optimize the distribution or number of subconcepts in order to enhance the classification performance of our method.

Acknowledgments

The research leading to these results has received funding from the European Community FP7 under grant agreement number 261743 (NoE VideoSense).

5. REFERENCES

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[2] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *Electronic Imaging 2003*, pages 87–95. Int. Society for Optics and Photonics, 2003.

[3] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–6. IEEE, 2012.

[4] B. K. Horn and B. G. Schunck. Determining optical flow. In *1981 Technical Symposium East*, pages 319–331. Int. Society for Optics and Photonics, 1981.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[6] 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 *MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[7] J. R. Uijlings, I. Duta, N. Rostamzadeh, and N. Sebe. Realtime video classification using dense hof/hog. In *ICMR*, page 145. ACM, 2014.