MediaEval 2011 Affect Task: Violent Scene Detection combining Audio and Visual Features with SVM

Esra Acar, Stephan Spiegel, Sahin Albayrak
DAI Labor, Berlin University of Technology, Berlin, Germany
{esra.acar, stephan.spiegel, sahin.albayrak}@dai-labor.de

ABSTRACT

We propose an approach for violence analysis of movies in a multi-modal (visual and audio) manner with one-class and two-class support vector machine (SVM). We use the scale-invariant feature transform (SIFT) features with the Bag-of-Words (BoW) approach for visual content description of movies, where audio content description is performed with the mel-frequency cepstral coefficients (MFCCs) features. We investigate the performance of combining visual and audio features in an early fusion manner to describe the violence in movies. The experimental results suggest that one-class SVM is a promising approach for the task.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Performance, Experimentation

Keywords

Violence detection, SVM, SIFT, Bag-of-Words, MFCCs

1. MOTIVATION AND RELATED WORK

Although video content analysis has been studied extensively in the literature, violence analysis of movies is restricted to a few studies [3], [4], [5]. Hence, the motivation of MediaEval 2011 Affect Task is the automatic violence content multi-modal analysis of movies to enable helping parents to review the most violent scenes in a movie to prevent their children from watching them. Detailed description of the task, the dataset, the ground truth and evaluation criteria are given in the paper by Demarty et al. [2].

Lin et al.[5] proposed a co-training based approach, where audio analysis by a modified pLSA algorithm, motion and high-level visual concept analysis was performed. Gong et al. applied a semi-supervised learning approach [4], where low-level visual and audio features were fused with high-level audio indicators. Giannakopoulos et al. proposed a multi-modal probabilistic late fusion approach [3]. For the MediaEval 2011 Affect Task, we apply multi-modal (audio and visual) analysis as in these studies [3], [4], [5] in an early fusion manner by one-class SVM [7] and two-class SVM [1]. We report and discuss our results on 3 Hollywood movies from the MediaEval 2011 dataset [2].

2. PROPOSED APPROACH

We propose an approach that merges visual and audio features in a supervised manner (with one-class and two-class SVM) for violence detection in movies. The main idea behind one-class SVM is to construct a hyper-sphere that contains most of the positive training examples. The hyper-sphere aims to separate the positive training examples from the rest of the world. The hyper-sphere is determined with two parameters which are v (an upper bound on the fraction of outliers) and σ (the kernel width). Two-class SVM on the other hand constructs a hyperplane in the feature space to achieve a good separation between positive and negative examples (i.e. maximum distance between the hyperplane and the nearest training examples of any two class).

For video content description, low-level visual and audio features of video shots of the movies are extracted. The low-level features are then combined in an early fusion manner to train SVMs. The multi-modal fusion scheme of our approach is given in Figure 1.

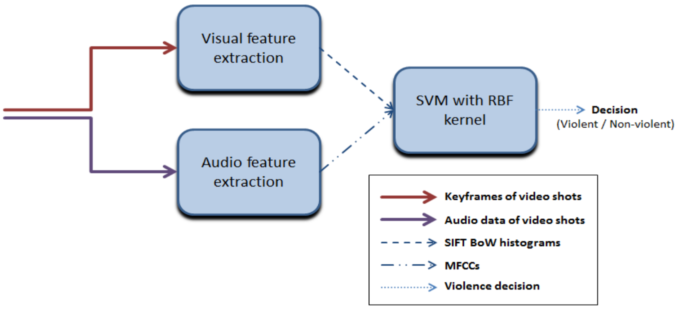

Figure 1: Multi-modal Fusion Scheme

2.1 Audio Features

To describe the audio content of the movies, we use MFCCs that are commonly used in audio recognition [6]. Due to the variability in duration of the annotated video shots, each video shot has different numbers of MFCCs. Since we want to describe each video shot by one significant feature vector, we compute the mean and standard deviation for each dimension of the MFCCs feature vectors to describe the audio signal.

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

2.2 Visual Features

SIFT based BoW approach is used for visual content description. As in BoW approaches known e.g. from [8], a visual vocabulary is constructed by clustering SIFT local feature vectors with the k-means clustering algorithm. Each resulting cluster is treated as a visual word. Once a visual vocabulary of size k ($k = 350$ in this work) is built, each SIFT feature is assigned to the closest visual word (Euclidean distance is used), a histogram is computed for the keyframe of a video shot and the related video shot is represented as BoW histogram that represents the visual word occurrences in its keyframe.

2.3 Results and Evaluation

The aim of this work is to assess the performance of one-class SVM and two-class SVM for violence detection. We evaluated our approach on 3 Hollywood movies from the MediaEval 2011 dataset [2]. We submitted three runs in total for the MediaEval 2011 Affect Task: *svm1(cf1:10)*, *svm1(cf1:1)* and *svm2(cf1:10)*. We applied one-class SVM with RBF kernel in *svm1(cf1:10)* and *svm1(cf1:1)* submissions, where two-class SVM with RBF kernel was applied in *svm2(cf1:10)* submission. The cost function mentioned in [2] was used during SVM parameter selection for *svm1(cf1:10)* and *svm2(cf1:10)*, where for *svm1(cf1:1)* submission cost function was adapted (i.e. C_{fa}=1 and C_{miss}=1). Parameter optimization process was performed by separating randomly the training data at hand into training, validation and test sets and the parameter values that gave the minimum cost according to the mentioned cost function were chosen. The optimized SVM parameters were v and σ for one-class SVM, where c and σ parameters were optimized for two-class SVM. LibSvm[1] was used as the SVM implementation. We employed the Auditory Toolbox[2] and David Lowe's SIFT demo software[3] to extract the 13-dimensional MFCCs and 128-dimensional SIFT features, respectively. Table 1 reports the number of false alarms (out of 3871) and miss detections (out of 629) and Table 2 gives the evaluation results. AED-P, AED-R and AED-F correspond to AED[2] precision, AED recall and AED F-measure, respectively.

Table 1: Misclassified video shots

Run	Miss	False alarm
svm1(cf1:10)	18 (%2,86)	3776 (%97,55)
svm1(cf1:1)	213 (%33,86)	2781 (%71,84)
svm2(cf1:10)	363 (%57,71)	1350 (%35,71)

Table 2: Evaluation results for the submitted runs

Run	AED-P	AED-R	AED-F	AED Cost
svm1(cf1:10)	0,1393	0,9714	0,2436	1,262
svm1(cf1:1)	0,1301	0,6614	0,2175	4,105
svm2(cf1:10)	0,1646	0,4229	0,237	6,12

The minimum miss rate is achieved with *svm1(cf1:10)*, where *svm2(cf1:10)* has the minimum false alarm rate. However, *svm2(cf1:10)* has the poorest cost value due to the miss rate. On the other hand, *svm1(cf1:10)* and *svm1(cf1:1)* have smaller miss rates, where their false alarm rate is higher compared to *svm2(cf1:10)*. The best cost is achieved with *svm1(cf1:10)*. However, the SVM classifier tends to classify almost every shot as violent in *svm1(cf1:10)*, because the cost of a miss is ten times higher than the cost of a false alarm. In *svm1(cf1:1)*, the number of video shots classified as violent gets lower, since the costs of a false alarm and a miss are equal. When all of the three runs are considered, two-class SVM achieves the poorest performance according to the cost measure.

We observed that one-class SVM tends to classify most of the video shots in the movies as violent even if equal costs are used for false alarm and miss. This may happen because of two reasons: (1) low-level audio and visual features are not selective enough to describe the violence, (2) sub-optimal parameters are being used for SVM model construction.

3. CONCLUSIONS

We applied one-class and two-class SVM approach for violence detection in movies. Our main finding is that one-class SVM seems a promising approach for the task. SVM parameters (v, σ) and the low-level audio and visual features used for the task need to be analyzed in more detail for better results in terms of false alarm rate. Future work will involve enhancing optimal SVM parameter selection process and more detailed analysis of the audio and visual features for content description to reduce the false alarm rate.

Acknowledgments We wish to thank Brijnesh Johannes Jain for his comments and suggestions that greatly contributed to the successful completion of this work.

4. REFERENCES

[1] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[2] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2011 Affect Task:Violent Scenes Detection in Hollywood Movies. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

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

[4] Y. Gong, W. Wang, S. Jiang, Q. Huang, and G. W. Detecting violent scenes in movies by auditory and visual cues. In *Advances in Multimedia Information Processing - PCM 2008, vol. 5353 of Lecture Notes in Computer Science*, pages 317–326, 2008.

[5] J. Lin and W. Wang. Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training. In *Advances in Multimedia Information Processing - PCM 2009, vol. 5879 of Lecture Notes in Computer Science*, pages 930–935, 2009.

[6] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Int. Symposium on Music Information Retrieval*, 2000.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[8] T. Zhang, C. Xu, G. Zhu, S. Liu, and L. H. A Generic Framework for Event Detection in Various Video Domains. In *ACM MM*, Firenze, Italy, October 25-29 2010.

[1] http://www.csie.ntu.edu.tw/˜cjlin/libsvm/

[2] http://cobweb.ecn.purdue.edu/˜malcolm/interval/1998-010

[3] http://www.cs.ubc.ca/˜lowe/keypoints/