RFA at MediaEval 2015 Affective Impact of Movies Task: A Multimodal Approach

Ionuţ Mironică
University Politehnica of Bucharest, Romania
imironica@imag.pub.ro

Bogdan Ionescu
University Politehnica of Bucharest, Romania
bionescu@imag.pub.ro

Mats Sjöberg
Helsinki Institute for, Information Technology HIIT
University of Helsinki, Finland
mats.sjoberg@helsinki.fi

Markus Schedl
Johannes Kepler University, Linz, Austria
markus.schedl@jku.at

Marcin Skowron
Austrian Research Institute for Artificial Intelligence, Vienna, Austria
marcin.skowron@ofai.at

ABSTRACT

The MediaEval 2015 Affective Impact of Movies Task challenged participants to automatically find violent scenes in a set of videos and, also, to predict the affective impact that video content will have on viewers. We propose the use of several multimodal descriptors, such as visual, motion and auditory features, then we fuse their predictions to detect the violent or affective content. Our best-performing run with regard to the official metric received a MAP of 0.1419 in the violence detection task, and an accuracy of 45.038% for the arousal estimation and 36.123% for the valence estimation.

1. INTRODUCTION

The MediaEval 2015 Affective Impact of Movies Task [6] challenged participants to develop algorithms for finding violent scenes in movies. Also, in contrast to previous years, the organizers introduced a completely new subtask for detecting the emotional impact of movies. The task provided a dataset of 10,900 short video clips extracted from 199 Creative Commons-licensed movies. Detailed description of the task, the dataset, the ground truth and evaluation criteria are given in the paper by Sjöberg et al. [6].

Our system this year is largely based on several multimodal systems that already obtained good results on similar problems [3, 4, 5].

2. METHOD

Our system builds on a set of visual, motion and auditory features, combined with a Support Vector Machine (SVM) classifier to obtain a violence or an affect score for each video document. First, we perform the feature extraction at the frame level. The resulting features are aggregated in one video descriptor using different strategies: the average of features, Fisher kernel(FK) [4] or Vector of Locally Aggregated Descriptors(VLAD) [3]. Finally, the global video descriptors are fed into a SVM multi-classifier framework. These steps are detailed in the following.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2.1 Feature set

Visual: We extracted ColorSIFT features [8] using the opponent colour space and spatial pyramids with two different sampling strategies: the Harris-Laplace salient point detector and dense sampling. We employed the Bag-of-Visual-Words (BoVW) approach where each spatial pyramid partition is represented by a 1,000 dimensional histogram over its ColorSIFT features. We also computed the CENsus TRansform hISTogram (CENTRIST) descriptor proposed in [9]. In addition, we used a total of four Convolutional Neural Networks (CNN) features, using the protocol laid out in [1]. The used CNNs were trained on either ImageNet 2010 or 2012 training datasets, following as closely as possible the network structure parameters of Krizhevsky *et al* [2]. Furthermore, the input images were resized to 256×256 pixels either by distortion or center cropping, thus giving in total four different CNNs from which we extract four different sets of feature vectors. We use the activations of the first fully-connected layer of each network as our features, which results in 4096-dimensional feature vectors. Ten regions were extracted from the test images as suggested in [2] (four corners, center patch plus flipping) and then a component-wise maximum is taken of the region-wise features.

Auditory: As for audio features, we used descriptors provided within the block-level framework [5]. They have been proven to be useful for retrieval, classification, and similarity tasks in the audio and music domain. More precisely, we computed for the audio channel of each video its spectral pattern (considers the cent-scaled spectrum on a 10-frame-basis to characterize frequency and timbre), delta spectral pattern (computes the difference between the original spectrum and a copy of the spectrum delayed by 3 frames), variance delta spectral pattern (considers the variance between the delta spectral blocks), logarithmic fluctuation pattern (applies several psychoacoustic models and characterizes the amplitude modulations), correlation pattern (computes Pearson's correlation between all pairs of 52 cent-scaled frequency bands), and spectral contrast pattern (computes the difference between spectral peaks and valleys in 20 cent-scaled frequency bands). We eventually end up with each clip being characterized by a 9,448-dimensional feature vector that models its audio content.

Table 1: Results for the submitted runs.

		Violence task	Affect task	
	Description	MAP	Accuracy valence	Accuracy arousal
run_1	Average on audio descriptors & nonlinear SVM	0.0485	33.032%	45.038%
run_2	Average on visual features & nonlinear SVM	0.0452	36.123%	34.104%
run_3	Modified VLAD with motion features & linear SVM	0.0768	29.731%	39.865%
run_4	Fisher kernel with CNN visual features [2] & linear SVM	0.1419	30.320%	44.365%
run_5	Late fusion between all the previous runs	0.0824	29.752%	37.595%

Motion: We computed the Histogram of Oriented Gradients (3D-HoG) and Histograms of Optical Flows (3D-HoF) cuboids motion features [7]. First of all, we computed each feature in 3D blocks with a dense sampling strategy: first the gradient magnitude responses in horizontal and vertical directions are computed. Then, for each response the magnitude was quantized in k orientations, where $k = 8$. Finally, these responses were aggregated over blocks of pixels in both spatial and temporal directions and concatenated.

2.2 Frame aggregation

Results from the literature showed that adopting Fisher kernel [4] and VLAD [3] representations in many video classification tasks allow for achieving higher accuracy than the use of traditional Bag-of-Words histogram representations. This is because these representations capture temporal variation over the frames within a video. We used two classical methods to encode the temporal variation over frame-based features, the Fisher Kernel [4] and a modified version of Vector of Locally Aggregated Descriptors [3]. Then, we aggregated the frame features already presented in Section 2.1.

2.3 Classifier

The final component of the system consists of the data classifier which is fed with the multimodal descriptors computed on previous steps. Among the broad choice of existing classification approaches, we selected a SVM classifier. We tested several type of kernels, i.e., a fast linear kernel and two nonlinear kernels: RBF and Chi-Square. While linear SVMs are very fast in both training and testing, SVMs with nonlinear kernels are more accurate in many classification tasks due to better adaptation to the shape of the clusters in the feature space.

Finally, in the case of multimodal features, we combine the SVMs output confidence values using max late-fusion combination:

$$CombMean(d, q) = \max_{i=1}^{N} cv_i \quad (1)$$

where cv_i is the confidence value of classifier i for class q ($q \in \{1, ..., C\}$), C represents the number of classes, d is the current video, and N is the number of classifiers to be aggregated.

3. EXPERIMENTAL RESULTS

3.1 Submitted runs

We submitted five runs for both tasks: the violence detection task and the induced affect detection task. For the first run we combined the audio features with a nonlinear SVM classifier. For the second run, we combined several visual features (BoVW-ColorSIFT, CENTRIST histograms and CNN features) with nonlinear SVM classifier. The next run uses a combination of modified VLAD with motion 3D-HoG/3D-HoF motion features with nonlinear SVM classifiers. In the fourth run, we propose the aggregation of the CNN frame features with the Fisher kernel representation. Then, we used a linear SVM classifier. Finally, for the fifth task we performed a late fusion strategy of the first four runs.

3.2 Results and discussion

Table 1 details the results for all our runs. The third column presents the MAP results obtained on the violence task, while the next two columns provide the final accuracy on the second task: the valence and arousal predictions.

Audio features and standard visual features performed poorly in the violence task. On the other side, the combination of VLAD with motion features obtained better results. The best results are obtained using Fisher kernel with CNN visual features. Fusing all the features together did not improve the results above the FK-CNN only result. In contrast, in the induced affect detection task all combinations perform similarly, except for audio features which have a clearly better result.

4. CONCLUSIONS

In this paper, we presented several multimodal approaches for the detection of violent content in movies. We obtained the best results on the violence task by using motion and visual features. On the other side, we obtained the best results on the affect task using the audio features only. The visual / motion features obtained lower results for both valence and arousal predictions. One reason for this is that the visual features do not fit on the purpose of the affect task. It also indicates that the affect task is more challenging than the violence task.

Acknowledgements

We received support by the Austrian Science Fund (FWF): P25655 and the InnoRESEARCH POSDRU /159/1.5/S/ 132395 program.

5. REFERENCES

[1] M. Koskela and J. Laaksonen. Convolutional network features for scene recognition. In *Proceedings of the 22nd International Conference on Multimedia*, Orlando, Florida, November 2014.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2012.

[3] I. Mironică, I. Duţă, B. Ionescu, and N. Sebe. A Modified Vector of Locally Aggregated Descriptors

Approach for Fast Video Classification. *Multimedia Tools and Applications (MTAP)*, 2015.

[4] I. Mironică, J. Uijlings, N. Rostamzadeh, B. Ionescu, and N. Sebe. Time Matters! Capturing Variation in Time in Video using Fisher Kernels. In *ACM Multimedia*, Barcelona, Spain, 21-25 October 2013.

[5] K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees. Automatic Music Tag Classification based on Block-Level Features. In *Proceedings of the 7th Sound and Music Computing Conference (SMC 2010)*, Barcelona, Spain, July 2010.

[6] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The MediaEval 2015 Affective Impact of Movies Task. In *MediaEval 2015 Workshop*, Wurzen, Germany, September 14-15 2015.

[7] J. Uijlings, I. Duta, E. Sangineto, and N. Sebe. Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. *International Journal of Multimedia Information Retrieval*, pages 1–12, 2014.

[8] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1582–1596, 2010.

[9] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1489–1501, 2011.