UMons at MediaEval 2015 Affective Impact of Movies Task including Violent Scenes Detection

Omar Seddati[1], Emre Kulah[2], Gueorgui Pironkov[1], Stéphane Dupont[1], Saïd Mahmoudi[1], Thierry Dutoit[1]

[1] University of Mons, Belgium
{omar.seddati, gueorgui.pironkov, stephane.dupont, said.mahmoudi, thierry.dutoit}@umons.ac.be

[2] Middle East Technical University, Ankara, Turkey
emre.kulah@ceng.metu.edu.tr

ABSTRACT

In this paper, we present the work done at UMons regarding the MediaEval 2015 Affective Impact of Movies Task (including Violent Scenes Detection). This task can be divided into two subtasks. On the one hand, Violent Scene Detection, which means automatically finding scenes that are violent in a set if videos. And on the other hand, evaluate the affective impact of the video, through an estimation of the valence and arousal. In order to offer a solution for both detection and classification subtasks, we investigate different visual and auditory feature extraction methods. An i-vector approach is applied for the audio, and optical flow maps processed through a deep convolutional neural network are tested for extracting features from the video. Classifiers based on probabilistic linear discriminant analysis and fully connected feed-forward neural networks are then used.

1. INTRODUCTION

With the increasing amount of video content available, the aim of MediaEval 2015 "Affective Impact of Movies Task" is to show users (depending on their age, preferences or mood) the content they are looking for. More precisely, this year the task focuses on two different aspects.

The first subtask is Violent Scene Detection (VSD), the goal being to alert parents about the potentially violent content of a video. Thus, the criterion for VSD used for annotation is: "*videos one would not let an 8 years old child see because of their physical violence*". Another possible application could be facilitating video surveillance alerts, as monitoring several screens simultaneously is a complicated task, even for humans.

Additionally to VSD, and for the first time at this year's MediaEval workshop, a second subtask is examined: Induced Affect Detection. This subtask focuses on the impact emotions can have for video or movie suggestions. Each video scene is categorized depending of its valence class (positive - neutral - negative) and its arousal class (active - neutral - passive). The purpose here is to predict the feelings that a particular video will cause to an user in order to recommend him similar or completely different content.

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

Both subtasks are examined on the same dataset. Around 10,000 video clips from professional and amateur movies are used, all under Creative Commons license. More information about these subtasks can be found in [6].

2. APPROACH

We use the same techniques for the VSD and affect detection subtasks. In our approach audio and video information are analyzed separately. Thus, two different feature extraction methods are applied depending of the features.

2.1 Audio approach

For the audio processing we use the same method as [2], where i-vectors and Probabilistic Linear Discriminant Analysis (pLDA) are used to classify environments (wedding ceremony, birthday party, parade, etc.). The i-vector approach consists of extracting a low-dimensional feature vector from high-dimensional data without losing most of the relevant acoustic information. This method was introduced by the speaker recognition community and has also proven its efficiency in language detection or in speaker adaptation for speech recognition.

In order to extract the i-vectors and classify them through pLDA, we have used the Matlab MSR Identity Toolbox [5]. For each audio track of the video shots, we extract 20 Mel-frequency cepstral coefficients, and the associated first and second derivatives. Thus, we use as input 60-dimensional features with a fixed length of 800 frames for each shot. For each shot a 100-dimensional i-vector is extracted. All the i-vectors are then processed through three independent classifiers. The first one is trained to classify violent and non-violent scenes. The second one differentiate positive, neutral and negative valence. The third one is trained on the three different levels of arousal.

2.2 Video approach

Convolutional neural networks (ConvNets) are a state-of-the-art technique in the field of object recognition within images. ConvNets applied to 2D images are adapted to capture spatial configurations. Using them to capture temporal information related to changes between video frames requires using several frames as input. A drawback is that it significantly increases the dimensionality of the input. Thus, an alternative approach consists of using optical flow maps as

Table 1: ConvNet architecture

Ind	Type	Filter size	Filter num	Stride
1	Conv	7x7	32	2
2	ReLU	-	-	-
3	Maxpool	3x3	-	2
4	Normalization			
5	Conv	5x5	96	1
6	ReLU	-	-	-
7	Maxpool	3x3	-	2
8	Normalization			
9	Conv	3x3	96	1
10	ReLU	-	-	-
11	Maxpool	3x3	-	2
12	Conv	3x3	96	1
13	ReLU	-	-	-
14	Maxpool	3x3	-	2
15	Conv	3x3	96	1
16	ReLU	-	-	-
17	FC	-	1024	-
18	Dropout	-	-	-
19	ReLU	-	-	-
20	FC	-	512	-
21	Dropout	-	-	-
22	ReLU	-	-	-
23	FC	-	2 or 3	-
24	LogSoftMax	-	-	-

input. Each map represents the motion of each pixel between two successive frames.

We used TV-L1 [4] algorithm from the OpenCV toolbox for optical flow extraction. We use 10 stacked optical flow frames as input. Note that 10 stacked optical flows equals 20 maps given that both horizontal and vertical components have to be provided. In order to reduce overfitting we use dropout, as well as data augmentation by cropping and flipping randomly the maps of the input sequence. We also estimate the motion of the camera by calculating the mean across the maps of the same component (horizontal and vertical), then we subtract the corresponding mean. Our system is tested on the publicly available Torch toolbox [1] which offers a powerful and varied set of tools, especially for building and training ConvNets. The details for the used architecture are listed in Table 1.

Using dense optical flow maps, means that the size of the neural network increases rapidly with the length of the sequence used as input. This implies that short sub-sequences of video frames (or rather optical flow maps) have to be used as input to the ConvNet. This increases the risk that those sub-sequences fall on parts of the video where there is no useful information for the identification of the category. To tackle this problem, we use a sliding window approach at test time, estimating the probability for each category in several sub-sequences of the video. The class with the highest probability after averaging over all the different subsequence probabilities is selected as the most likely class.

We also train a ConvNet with the same architecture on the HMDB-51 dataset [3] (action recognition benchmark), in order to build a more robust motion feature extractor leveraging this additional external data. Then, we extract features from the MediaEval annotated data and train a two layers fully connected neural network for each of the three subtasks.

Table 2: Mean Average Precision (MAP) on Violence detection

Run	MAP (%)
i-vector - pLDA	9,56
optical flow maps - ConvNets	9,67
optical flow maps - ConvNets - HMDB-51	6,56

Table 3: Global accuracy for Affect detection. (OFM stands for optical flow maps)

Run	Valence (%)	Arousal (%)
i-vector - pLDA	37.03	31.71
OFM - ConvNets	35.28	44.39
OFM - ConvNets - HMDB-51	37.28	52.44

3. RESULTS AND DISCUSSION

We have submitted three runs for both subtasks. The results for VSD tasks are presented in Table 2. The Mean Average Precision (MAP) is computed for each run. We can see that using external data from HMDB in order to train the feature extractor is less efficient than training the feature extractor on the MediaEval dataset. The i-vector & pLDA technique present similar results as the optical flow maps & ConvNets association.

The global accuracy for the affect detection task is shown if Table 3. For the valence, all methods give similar results. A difference appears for the arousal task. The audio features perform poorly in comparison to the other runs. Using external data proves here to be more interesting as the last run significantly outperforms the second run. Motion seems to be an important discriminative factor for arousal estimation.

3.1 Discussion

We have also investigated merging the audio and visual features together. The features from the ConvNets extractor and the i-vectors were used as input to another neural network. But the results were poorer than using the features separately. Further work will investigate audio-visual fusion more in depth.

4. CONCLUSION

In this paper we presented two approaches for both affect and violent scene detection. Visual and audio features are processed separately. Both features are giving similar results for violence detection and valence. For arousal, video features are far more interesting, especially when the ConvNets feature extractor is trained on external data. Our future work will focus on the merging the audio and video features.

5. ACKNOWLEDGEMENTS

This work has been partly funded by the Walloon Region of Belgium through the Chist-Era IMOTION project (Intelligent Multi-Modal Augmented Video Motion Retrieval System) and by the European Regional Development Fund (ERDF) through the DigiSTORM project.

6. REFERENCES

[1] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[2] B. Elizalde, H. Lei, and G. Friedland. An i-vector representation of acoustic environments for audio-based video event detection on user generated content. In *Multimedia (ISM), 2013 IEEE International Symposium on*, pages 114–117. IEEE, 2013.

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[4] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo. TV-L1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013.

[5] S. O. Sadjadi, M. Slaney, and L. Heck. MSR identity toolbox v1. 0: A MATLAB toolbox for speaker recognition research. *Speech and Language Processing Technical Committee Newsletter*, 2013.

[6] M. Sjöberg, B. Ionescu, H. Wang, Y. Baveye, E. Dellandréa, L. Chen, V. L. Quang, M. Schedl, and C.-H. Demarty. The MediaEval 2015 affective impact of movies task. In *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.