RECOD at MediaEval 2015: Affective Impact of Movies Task

Daniel Moreira[1], Sandra Avila[2], Mauricio Perez[1], Daniel Moraes[1], Vanessa Testoni[3], Eduardo Valle[2], Siome Goldenstein[1], Anderson Rocha[1]*

[1]Institute of Computing, University of Campinas, SP, Brazil
[2]School of Electrical and Computing Engineering, University of Campinas, SP, Brazil
[3]Samsung Research Institute Brazil, SP, Brazil

ABSTRACT

This paper presents the approach used by the RECOD team to address the challenges provided in the MediaEval 2015 Affective Impact of Movies Task. We designed various video classifiers, which relied on bags of visual features, and on bags of auditory features. We combined these classifiers using different approaches, ranging from majority voting to machine-learned techniques on the training dataset. We only participated in the Violence Detection subtask.

1. INTRODUCTION

The MediaEval 2015 Affective Impact of Movies Task challenged its participants to automatically classify video content, regarding three high-level concepts: valence, arousal and violence [5].

The activities of classifying video valence and of classifying video arousal were grouped under the same subtask: the Induced Affect Detection. The classification of violence, in turn, was related to the Violence Detection subtask, where participants were supposed to label a video as violent or not. For both subtasks, the same video dataset was annotated and provided. It consisted of short clips, extracted from 199 Creative Commons-licensed movies of various genres. A detailed overview of the two subtasks, metrics, dataset content, license, and annotation process can be found in [5].

In the following sections, we detail the classifiers we designed to solve the task. Thereafter, we explain the setup of the submitted runs, and report the results, with the proper discussion.

2. SYSTEM DESCRIPTION

We designed video classifiers based on bags of visual features, and on bags of auditory features. Following the typical bags-of-features-based approach, these classifiers implement a pipeline that is composed by three stages: (*i*) low-level video/audio description, (*ii*) mid-level feature extraction, and (*iii*) supervised classification. These classifiers are then combined either in a majority-voting fashion, or in a machine-learned scheme.[1]

*Corresp. author: A. Rocha, anderson.rocha@ic.unicamp.br

[1]As we are patenting the developed approach, a few technical aspects are not reported on this manuscript.

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

2.1 Bags of Visual Features

First of all, similarly to Akata et al. [1], as a preprocessing step, and for the sake of saving low-level description time, we reduce the resolution of all videos, keeping the original aspect ratio.

We developed two classifiers based on bags of visual features. These classifiers differ from each other mainly with respect to the employed low-level local video descriptors. We have a solution based on a static frame descriptor (Speeded-UP Robust Features, SURF [2]), and another solution based on a space-temporal video descriptor.

In the particular case of the SURF-based classifier, SURF descriptions are extracted on a dense spatial grid, at multiple scales. In the case of the space-temporal-based one, we apply a sparse description of the video space-time (i.e., we describe only the detected space-temporal interest points).

Prior to the mid-level feature extraction, for the sake of saving extraction time, we also reduce the dimensionality of the low-level descriptions.

In the mid-level feature extraction, for each descriptor type, we use a bag-of-visual-words-based representation [4].

In the high-level video classification, we employ a linear Support Vector Machine (SVM) to label the mid-level features, as suggested in [4].

2.2 Bags of Auditory Features

We developed three classifiers based on bags of auditory features. Analogously to the visual ones, these classifiers differ from each other with respect to the employed low-level audio descriptors. We thus use the OpenSmile library [3] to extract audio features.

Prior to the mid-level feature extraction, for the sake of saving extraction time, we also reduce the dimensionality of the low-level descriptions.

To deal with the semantic gap between the low-level audio descriptions, and the high-level concept of violence, we adapt a bag-of-features-based representation [4] to quantize the auditory features.

Finally, concerning the high-level video classification, we employ a linear SVM.

2.3 Combination Schemes

To combine various classifiers, we adopt two late fusion schemes.

In the first one, we combine the scores returned by the various classifiers in a voting fashion. After counting the votes, we designate the video class as being equal to the most voted one. To attribute a final score, we pick the score of

the classifier that presents the strongest certainty regarding the video class.

In the second combination scheme, we concatenate the positive scores of the classifiers in a predefined order, and feed them to an additional classifier.

2.4 External Data and Data Augmentation

In the dataset of this year, 6,144 short video clips were provided in the development (i.e., training) group [5]. From this total, only 272 video clips were from the positive class, a small number for an effective train of our techniques. Therefore, in order to augment such content and obtain a more balanced training set, we incorporated the 86 YouTube web videos that were provided in the competition of last year [6], as an external data source.

Given that these web videos were, in average, longer than the videos of this year, we decided to segment the positive annotated chunks in parts of $10 - 12$ seconds. That leaded to a total of 252 additional positive segments to augment our positive training dataset.

3. SUBMITTED RUNS

This year, participants were allowed to submit up to five runs for the violence detection subtask, with at least one requiring the use of no external training data [5]. The official evaluation metric is mean average precision (MAP), which is calculated using the NIST *trec_eval*[2] tool.

Table 1 summarizes the runs that were submitted this year to the competition. In total, we generated five different runs. In two, we did not use external data, while on the remaining other three, we employed external data, as explained in Section 2.4.

Run	External Data	Visual Features	Auditory Features	Combined	MAP
1	No	All	All	Majority Voting	0.1143
2	No	All	All	Classifier	0.0690
3	Yes	All	All	Majority Voting	0.1126
4	Yes	No	Tone	Majority Voting	0.0924
5	Yes	Space-temporal	No	Majority Voting	0.0960

Table 1: Official results obtained for the Violence Detection subtask.

4. RESULTS AND DISCUSSION

The best result (related to run *1*) was achieved by the classifier that used a majority-voting late combination of visual and auditory features, trained with no external data (i.e., $MAP = 0.1143$). It performed better than the exact same solution (at run *3*, with $MAP = 0.1126$), whose only difference was the use of external data in the training phase (as explained in Section 2.4).

Therefore, we failed to augment the training data. Reason for that may be related to the use of different types of video sources, given that — in this year — Hollywood-like movie segments were provided [5], contrasting to the predominantly amateur web videos of last year [6].

Notwithstanding, the majority-voting late combination of visual and auditory features indeed improved the classification performance. Although trained with the same videos (with external data), runs *4* (auditory only, with $MAP = 0.0924$) and *5* (visual only, with $MAP = 0.096$) achieved results that were below the combined solution (related to run *3*, with $MAP = 0.1126$).

Regarding our results, in general terms, we did not have enough positive samples to learn a better classifier, a mandatory requirement of the machine learning techniques that we employed.

5. CONCLUSIONS

This paper presented the video classifiers used by the RECOD team to participate in the violence detection subtask of the MediaEval 2015 Affective Impact of Movies Task. The reported results show that a late combination of visual- and auditory-feature-based classifiers lead to a better final classification system, in the case of violence detection. Finally, given the machine leaning nature of our solutions, the challenging dataset of this year did not contain enough positive video samples to learn a better classifier, what strongly impacted on our results.

Acknowledgments

Part of the results presented in this paper were obtained through the project "Sensitive Media Analysis", sponsored by Samsung Eletrônica da Amazônia Ltda., in the framework of law No. 8,248/91. We also thank the financial support from CNPq, FAPESP and CAPES.

6. REFERENCES

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(3):507–520, 2014.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[3] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462. ACM, 2010.

[4] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[5] 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 *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015.*, 2015.

[6] M. Sjöberg, B. Ionescu, Y. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty. The mediaeval 2014 affect task: Violent scenes detection. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, October 16-17, 2014.*, 2014.

[2] `http://trec.nist.gov/trec_eval/`