RECOD at MediaEval 2014: Violent Scenes Detection Task

Sandra Avila[‡], Daniel Moreira[†], Mauricio Perez[†], Daniel Moraes[§], Isabela Cota[†],
Vanessa Testoni[§], Eduardo Valle[‡], Siome Goldenstein[†], Anderson Rocha[†]

[†]Institute of Computing, University of Campinas (Unicamp), SP, Brazil
[‡]School of Electrical and Computing Engineering, University of Campinas (Unicamp), SP, Brazil
[§]Samsung Research Institute Brazil, SP, Brazil

sandra@dca.fee.unicamp.br, daniel.moreira@ic.unicamp.br, {mauricio.perez, daniel.moraes, isabela.cota}@students.ic.unicamp.br, vanessa.t@samsung.com dovalle@dca.fee.unicamp.br, {siome, anderson.rocha}@ic.unicamp.br

ABSTRACT

This paper presents the RECOD approaches used in the MediaEval 2014 Violent Scenes Detection task. Our system is based on the combination of visual, audio, and text features. We also evaluate the performance of a convolutional network as a feature extractor. We combined those features using a fusion scheme. We participated in the main and the generalization tasks.

1. INTRODUCTION

The objective of the MediaEval 2014 Violent Scenes Detection task is to automatically detect violent scenes in movies and web videos. The targeted violent scenes are those "*one would not let an 8 years old child see in a movie because they contain physical violence*".

In this year, two different datasets were proposed: (i) a set of 31 Hollywood movies, for the main task, and (ii) a set of 86 short YouTube web videos, for the generalization task. The training data is the same for both subtasks. A detailed overview of the datasets and the subtasks can be found in [6].

In the following, we briefly introduce our system and discuss our results[1].

2. SYSTEM DESCRIPTION

2.1 Visual Features

In low-level visual feature extraction, we extract SURF descriptors [4]. For that, we first apply the FFmpeg software [1] to extract and resize the video frames. Low-level visual descriptors are extracted on a dense spatial grid at multiple scales. Next, they are reduced using a PCA algorithm.

Besides that, in order to incorporate temporal information, we compute dense trajectories and motion boundary descriptors, according to [7]. Again, for the sake of processing time, we decide to resize the video. Also, we reduce the dimensionality of the video descriptors.

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

Furthermore, we use a visual feature extractor based on Convolutional Networks, which were trained on the ImageNet 2012 training set [5]. It has been chosen due to its very competitive results on detection and classifications tasks. Additionally, as far as we know, deep learning methods have not yet been employed in the MediaEval Violent Scenes Detection task.

2.2 Audio Features

Using the OpenSmile library [3], we extract several types of audio features. A bag of visual words-based representation is employed to quantize the audio features and a PCA algorithm is also used to reduce the dimensionality of the features.

2.3 Text Features

To represent the movie subtitles, we apply the bag of words approach: the most common, simple and successful document representation used so far. The bag of words vector is normalized using a term's document frequency.

Also, before creating the bag of words representation, we remove the stop words and we apply a stemming algorithm to reduce a word to its stem.

2.4 Classification

Classification is performed with Support Vector Machines (SVM) classifiers, using the LIBSVM library [2]. Moreover, classification is done separately for each descriptor. The outputs of those individual classifiers are then combined at the level of normalized scores. Our fusion strategy is done by the combination of classification outcomes optimized on the training set.

3. RUNS SUBMITTED

In total, we generated 10 different runs: 5 runs for each subtask. For main task (**m**), we have:

- **m1:** 3 types of audio features + 3 types of visual features (including a visual feature extractor based on Convolutional Networks) + text features;
- **m2:** 1 type of audio features + 3 types of visual features (including a visual feature extractor based on

[1] There are some technical aspects which we cannot put directly in the manuscript given we are patenting the developed approach.

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

Convolutional Networks) + text features;
- **m3:** 1 type of audio features + 3 types of visual features (including a visual feature extractor based on Convolutional Networks);
- **m4:** 1 type of audio features + 2 types of visual features + text features;
- **m5:** 1 type of audio features.

For generalization task (**g**), we have:

- **g1:** 3 types of audio features + 3 types of visual features (including a visual feature extractor based on Convolutional Networks);
- **g2:** 1 type of audio features + 3 types of visual features (including a visual feature extractor based on Convolutional Networks);
- **g3:** 1 type of audio features + 2 types of visual features;
- **g4:** 1 type of audio features;
- **g5:** 1 type of visual features.

4. RESULTS AND DISCUSSION

Tables 1 and 2 show the performance of our system for main and generalization task, respectively. We can notice that, despite the diversity of fusion strategies, the differences among most runs (**m1**, **m2**, **m3** and **g1**, **g2**, **g3**) are quite small. We are currently investigating such results. Also, we observe that, for run **m4**, we selected a wrong threshold[2] by mistake.

	8mil	Brav	Desp	Ghos	Juma	Term	Vven	MAP
m1	0.204	0.477	0.337	0.567	0.188	0.479	0.378	0.376
m2	0.239	0.459	0.308	0.348	0.362	0.465	0.431	0.373
m3	0.189	0.545	0.277	0.465	0.212	0.418	0.489	0.371
m4	0.115	0.319	0.209	0.270	0.159	0.502	0.167	0.249
m5	0.373	0.301	0.307	0.423	0.175	0.308	0.317	0.315

Table 1: Official results obtained for the main task in terms of MAP2014.

	g1	**g2**	**g3**	**g4**	**g5**
MAP	0.618	0.615	0.604	0.545	0.515

Table 2: Official results obtained for the generalization task in terms of MAP2014.

For the main task, our results are considerably below our expectations (based on our training results). By analyzing the results, we pointed out a crucial difference between training and test videos. In the Violent Scenes Detection task, the participants are instructed in how to extract the DVD data and convert it to MPEG format. For the sake of saving disk space, we opted to convert the MPEG video files to MP4 or to M4V. However, that choice introduced a set of problems.

First, with respect to the training data, we converted the MPEG video files to MP4 or to M4V, depending on which video container we were able to successfully synchronize the extracted frames, regarding the numbers given by the groundtruth. Despite both containers store the video stream in H.264 format, we did not notice that the M4V conversion resulted in a different video aspect ratio (718×432 pixels). Similarly, the audio encoding was also divergent: MP3 audio for MP4, while AAC audio for M4V. Next, due to frame synchronization issue, we kept the test data in its original format (MPEG-2, 720×576 pixels, with AC3 audio). Therefore, we faced the problem of dealing with different aspect ratios in training and testing data, as well as distinct audio formats.

For the generalization task, the problem is alleviated because the test data is provided in MP4.

Tables 3 reports the unofficial (**u**) results for main task that we evaluated ourselves. Here, the results are obtained by using the data (training and test sets) in MPEG format. The first column indicates which input features were used: **u1** for 1 type of audio features and **u2** for text features. Unfortunately, due to time constraints, we were not able to prepare more runs.

It should be mentioned first that, the results for run **u2**, are independent of the video format, since we directly extracted the movie subtitles from DVD. For run **u1**, we can notice a considerable improvement of classification performance, from 0.315 (run **m5**) to 0.493 (run **u1**), confirming the negative impact of using distinct audio formats. We are currently investigating the impact on visual features.

	8mil	Brav	Desp	Ghos	Juma	Term	Vven	MAP
u1	0.351	0.601	0.636	0.530	0.521	0.352	0.463	0.493
u2	0.402	0.237	0.407	0.345	0.232	0.277	0.188	0.298

Table 3: Unofficial results obtained for the main task in terms of MAP2014.

5. ACKNOWLEDGMENTS

This research was partially supported by FAPESP, CAPES, CNPq and Project “Capacitação em Tecnologia de Informação” financed by Samsung Eletrônica da Amazônia Ltda., using resources provided by the Informatics Law no. 8.248/91.

6. REFERENCES

[1] FFmpeg. http://www.ffmpeg.org/.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):1–27, 2011.

[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, 2010.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575, 2014.

[6] M. Sjöberg, B. Ionescu, Y. Jiang, V. Quang, M. Schedl, and C. Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 16–17 2014.

[7] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103:60–79, 2013.

[2] Scenes are classified as violent or non-violent based on a certain threshold.