FAR at MediaEval 2014 Violent Scenes Detection: A Concept-based Fusion Approach

Mats Sjöberg
University of Helsinki,
Finland
mats.sjoberg@helsinki.fi

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

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

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

ABSTRACT

The MediaEval 2014 Violent Scenes Detection task challenged participants to automatically find violent scenes in a set of videos. We propose to first predict a set of mid-level concepts from low-level visual and auditory features, then fuse the concept predictions and features to detect violent content. With the objective of obtaining a higly generic approach, we deliberately restrict ourselves to use simple general-purpose descriptors with limited temporal context and a common neural network classifier. The system used this year is largely based on the one successfully employed by our group in 2012 and 2013, with some improvements and updated features. Our best-performing run with regard to the official metric received a MAP2014 of 45.06% in the main task and 66.38% in the generalization task.

1. INTRODUCTION

The MediaEval 2014 Violent Scenes Detection task [4] challenged participants to develop algorithms for finding violent scenes in two settings: popular Hollywood-style movies (main task), and YouTube web videos (generalization task). The organizers provided a training set of 24 movies with frame-wise annotations of segments containing physical violence as well as several violence-related concepts (e.g. blood or fire) for part of the data. The test set consisted of 7 movies for the main task, and 86 short web videos for the generalization task.

Our system this year is largely based on the one successfully employed by us in 2012 [3] and 2013 [5]. We tackle the task as a machine learning problem, employing general-purpose features and a neural network classifier. The main novel contribution is an updated set of low-level features.

2. METHOD

Our system builds on a set of visual and auditory features, employing the same type of neural network classifier at different stages to obtain a violence score for each frame of an input video. First, we perform feature extraction at the frame level. The resulting data is then fed into a multi-classifier framework that operates in two steps. The first step consists of training the system using ground truth data. Training is performed at two levels. At mid-level, a bank of classifiers is trained using ground truth related to concepts that are usually present in the violent scenes, e.g., presence of "fire", presence of "gunshots", or "gory" scenes. Then, high-level violence detection is ensured by a final classifier that is fed either with the previous concept predictions and/or the low-level content descriptors. The violence classifier is also trained on the provided ground truth for the violent segments. The final step consists of classifying the new unlabeled data (e.g. the test set) which is achieved by employing the previously trained multi-classifier framework. These steps are detailed in the following.

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

2.1 Feature set

Visual (225 dimensions): For each video frame, we extract several standard color and texture-based descriptors, such as: Color Naming Histogram, Color Moments, Local Binary Patterns, Color Structure Descriptor, and Gray Level Run Length Matrix. Also, we compute the Histogram of Oriented Gradients, that exploits the local object appearance and shape within a frame by using the distribution of edge orientations. For a more detailed description of the visual features, see [1].

Auditory (29 dimensions): In addition, we extract a set of low-level auditory features: amplitude envelop, root-mean-square energy, zero-crossing rate, band energy ration, spectral centroid, spectral flux, bandwidth, and Mel-frequency cepstral coefficients. We compute the features on frames of 40 ms without overlap to make alignment with the 25-fps video frames trivial.

2.2 Classifier

For classification, we use multi-layer perceptrons with a single hidden layer of 512 units and one or multiple output units. All units use the logistic sigmoid transfer function. The input data is normalized by subtracting the mean and dividing by the standard deviation of each input dimension. Training is performed by backpropagating cross-entropy error, using random dropouts to improve generalization. We follow the dropout scheme of [2, Sec. A.1] with some minor modifications to the parameters.

For the concept training set of 18 movies, each video frame was annotated with the 10 different concepts as detailed in [4]. We divide the concepts into visual, auditory and au-

Table 1: Results for different features (%)

	feat.	prec.	recall	F-score	MAP2014
main_1	a	28.04	71.26	40.24	**45.06**
main_2	v	17.88	93.62	30.03	32.64
main_3	c	28.65	44.94	34.99	25.02
main_4	av	19.34	77.18	30.92	31.96
main_5	ac	29.16	63.08	39.88	40.77
gen_1	a	46.04	85.81	59.93	57.81
gen_2	v	43.42	86.05	57.72	59.63
gen_3	c	49.68	85.80	62.92	**66.38**
gen_4	av	44.76	83.38	58.25	58.07
gen_5	ac	46.86	83.94	60.14	60.92

Table 2: Movie specific results, MAP2014 (%)

movie (main task)	a	v	c	av	ac
Ghost in the Shell	82.67	20.38	25.26	23.72	67.30
Braveheart	29.01	36.26	17.22	22.65	24.79
Jumanji	29.27	16.13	2.71	14.07	23.70
Desperado	37.78	42.58	18.25	34.85	27.65
V for Vendetta	48.48	24.98	36.80	45.07	49.10
Terminator 2	56.17	27.27	48.82	43.25	55.26
8 Mile	32.03	60.84	26.09	40.08	37.61

diovisual categories, depending on which low-level feature domains we think are relevant for each. Next, we train and evaluate a neural network for each of the concepts, employing leave-one-movie-out cross-validation.

2.3 Fusion scheme

The final violence predictor is trained using both low-level features and all mid-level concept predictions as inputs. For comparison, we also train classifiers to predict violence just from the features or just from the concepts.

Training the violence detector requires inputs that are similar to those that will be used in the testing phase, thus using the concept ground-truth for training will not work. Instead we use the concept prediction cross-validation outputs on the training set (see previous section) as a more realistic input source – in this way the system can learn which concept predictors to rely on.

The final violence prediction score is generated by applying a sliding median filter as temporal smoothing. We used a filter length of 5 seconds (125 frames), this was selected from experimenting in the training set. The final detection as violent or non-violent is generated by thresholding the prediction score. The thresholds were determined by maximizing the MAP2014 performance measure in the training set using cross-validation.

3. EXPERIMENTAL RESULTS

We submitted five runs for both the main task and the generalization task. Table 1 details the results for all our runs. The first five lines show our runs submitted to the main task, the next five lines are those for the generalization task. The second column indicates which input features were used, 'a' for auditory, 'v' for visual, and 'c' for concept predictions. Multiple feature modalities indicate that they were integrated using early fusion. For the main task, the auditory features achieved the highest MAP2014 result. Concept detectors and visual features performed poorly in the main task, and fusing them with the auditory features did not improve the results above the audio-only result. In contrast, in the generalization task all combinations perform similarly, except for the concepts which have a clearly better result.

Another observation is that all results have a strong imbalance between precision and recall. Our analysis indicates that this is not due to a poor selection of the violence judgment threshold (in fact our thresholds are relatively close to optimal), but instead due to the new MAP2014 measure favoring high recall.

Table 2 shows the movie specific results for each of our main task runs. Interestingly auditory features perform particularly well on the anime movie "Ghost in the Shell", while the visual features perform strongly on "8 Mile", a drama movie with more realistic violence such as fist fights etc. "Jumanji" and "Braveheart" are the two movies with the poorest results. This can perhaps be explained by the fact that they differ from the training set more than the other movies. In particular "Braveheart" depicts brutal medieval fights, which are not represented in the training set.

4. CONCLUSIONS

Our results show that violence detection can be done well using general-purpose features and generic neural network classifiers, without engineering domain-specific features. The selection of feature modalities is highly dependent on the type of material, for Hollywood-style movies auditory features performed best, while concepts are useful for the more mixed style found in YouTube videos. Based on the results, we can also conclude that our violence detection framework generalises well: even though it was trained on only feature length movies it performs accurate violence detection on YouTube videos as well.

Acknowledgements

We received support by the Academy of Finland, grants no. 255745 and 251170, ESF POSDRU/159/1.5/S/132395 InnoRESEARCH programme, the EU-FP7 project no. 601166 and the Austrian Science Fund (FWF): P22856 and P25655.

5. REFERENCES

[1] B. Boteanu, I. Mironică, and B. Ionescu. A relevance feedback perspective to image search result diversification. In *Proc. ICCP*, 2014.

[2] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv, 2012.

[3] B. Ionescu, J. Schlüter, I. Mironică, and M. Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *Proc. ICMR*, pages 215–222, New York, NY, USA, 2013. ACM.

[4] 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.

[5] M. Sjöberg, J. Schlüter, B. Ionescu, and M. Schedl. FAR at MediaEval 2013 violent scenes detection: Concept-based violent scenes detection in movies. In *Proc. MediaEval Workshop*, Barcelona, Spain, 2013.