FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies

Mats Sjöberg
Aalto University,
Espoo, Finland
mats.sjoberg@aalto.fi

Jan Schlüter
Austrian Research Institute
for Artificial Intelligence,
Vienna, Austria
jan.schlueter@ofai.at

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

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

ABSTRACT

The MediaEval 2013 Affect Task challenged participants to automatically find violent scenes in a set of popular movies. 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. We deliberately restrict ourselves to simple general-purpose features with limited temporal context and a generic neural network classifier. The system used this year is largely based on the one successfully employed by our group in the 2012 task, with some improvements based on our experience from last year. Our best-performing run with regard to the official metric received a MAP@100 of 49.6%.

Keywords

Violent Scenes Detection, Concept Detection, Supervised Learning, Neural Networks, MediaEval 2013

1. INTRODUCTION

The MediaEval 2013 Affect Task [1] challenged participants to develop algorithms for finding violent scenes in popular movies from DVD content based on video, audio and subtitles. The organizers provided a training set of 18 movies with frame-wise annotations of segments containing physical violence as well as several violence-related concepts (e.g. blood or fire), and a test set of 7 unannotated movies.

The system used by our group this year is largely based on the one successfully employed by us in the 2012 edition of the violent scenes detection task [4]. In this year we have tried new descriptor combinations, and tweaked the neural network training parameters based on experiments performed with the 2012 task setup.

2. METHOD

Our system builds on a set of visual and auditory features, employing the same type of classifier at different stages to obtain a violence score for each frame of an input video. The setup is largely the same as in 2012 [4].

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

2.1 Feature set

Visual (93 dimensions): For each video frame, we extract an 81-dimensional Histogram of Oriented Gradients (HoG), an 11-dimensional Color Naming Histogram [6] and a visual activity value. The latter is obtained by lowering the threshold of the cut detector in [3] such that it becomes overly sensitive, then counting the number of detections in a 2-second time window centered on the current frame.

Auditory (98 dimensions): In addition, we extract a set of low-level auditory features as used by [5]: Linear Predictive Coefficients (LPCs), Line Spectral Pairs (LSPs), Mel-Frequency Cepstral Coefficients (MFCCs), Zero-Crossing Rate (ZCR), and spectral centroid, flux, rolloff, and kurtosis, augmented with the variance of each feature over a half-second time window. We use frame sizes 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.

We normalize the input data 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 minor modifications: all weights are initialized to zero, mini-batches are 900 samples, the learning rate starts at 5.0, momentum is increased from 0.45 to 0.9 between epochs 10 and 20 and we train for 100 epochs only. These settings worked well in experiments with the 2012 training/testing split. In particular we increased the learning rate from what was used in 2012, because it improved performance.

2.3 Fusion scheme

As last year [4], we use the concept annotations as a stepping stone for predicting violence: We train a separate classifier for each of 10 different concepts on the visual, auditory or both feature sets, then train the final violence predictor using both feature sets and all concept predictions as inputs. For comparison, we also train classifiers to predict violence just from the features or just from the concepts.

3. EXPERIMENTAL RESULTS

3.1 Concept prediction

For the training set of 18 movies, each video frame was annotated with the 10 different concepts as detailed in [1]. We divide the concepts into visual, auditory and audiovisual 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. The evaluation results are very similar to our experiments in 2012 [4, Sec. 3.1], which is not surprising since the training set has only been supplemented with 20% new movies. For example, firearms and fire perform well, while carchase performs badly.

3.2 Violence prediction

Next, we train a frame-wise violence predictor, using visual and auditory low-level features, as well as the concept predictions, as input. Training 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.

3.3 Evaluation results

We submitted five runs for sub task 1, i.e., the objective violence definition. Due to time constraints we were not able to prepare any runs for sub task 2 which used the subjective violence definition. One of our runs was a *segment-level* run (run5), which forms segments of consecutive frames that our predictor tagged as violent or non-violent. The remaining four runs are *shot-level* (from run1 to run4), which use the shot boundaries provided by the task organizers. For each run, each partition (segment or shot) is assigned a violence score corresponding to the highest predictor output for any frame within the segment. The segments are then tagged as violent or non-violent depending on whether their violence score exceeds a certain threshold. We used the same thresholds as used by our system in 2012, which were determined by cross-validation in the training set of that year.

Table 1 details the results for all our runs. The first five lines show our runs submitted to the official evaluation. The first four are shot-level runs, the fifth our single segment-level run. The next three lines are additional unofficial runs that we evaluated ourselves. The second column indicates which input features were used, 'a' for auditory, 'v' for visual, and 'c' for concept predictions. The auditory features achieved the highest MAP@100 result, with no gains being provided by the other modalities.

For our submissions we reused the thresholds from [4]. Unfortunately, this gave a very imbalanced precision and recall for the concept-only submission (run 2), making it difficult to compare to our other runs. To better judge the relative performance of our submissions, Table 1 reports precision, recall and F-score for the threshold maximizing the F-score. Under this metric, the combination of auditory features and concept predictions gives the best result, but differences between most runs are quite small.

Table 2 shows the movie specific results for each of our submitted shot-level runs. Despite the bad threshold on run2 it performs very well on Pulp Fiction. The movie "Legally Blond" had very few violent scenes and these were hard to detect with any of our runs.

Table 1: Results for different features

	feat.	prec.	rec.	max F-sc.	MAP@100
run1	a	34%	48%	40.0%	49.6%
run2	c	35%	51%	41.5%	30.4%
run3	av	34%	52%	41.4%	39.6%
run4	avc	35%	48%	40.9%	40.4%
run5	avc	23%	28%	25.8%	35.0%
	v	20%	50%	29.0%	23.9%
	ac	37%	47%	41.6%	47.4%
	vc	22%	53%	31.0%	28.5%

Table 2: Movie specific results (MAP@100)

movie	run1	run2	run3	run4
Fantastic Four 1	73.1%	63.0%	60.5%	69.7%
Fargo	55.5%	0.0%	57.0%	60.6%
Forrest Gump	38.9%	19.3%	35.8%	37.0%
Legally Blond	0.0%	0.0%	4.3%	4.4%
Pulp Fiction	62.0%	90.9%	51.4%	52.1%
The God Father 1	84.7%	39.5%	49.3%	47.7%
The Pianist	32.9%	0.0%	19.3%	11.2%

4. CONCLUSIONS

Our results show that violence detection can be done fairly well using general-purpose features and generic neural network classifiers, without engineering domain-specific features. While auditory features give the best results, using mid-level concepts can give small overall gains, and more pronounced gains for particular movies.

5. REFERENCES

[1] C. Demarty, C. Penet, M. Schedl, B. Ionescu, V. Quang, and Y. Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[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, V. Buzuloiu, P. Lambert, and D. Coquin. Improved Cut Detection for the Segmentation of Animation Movies. In *IEEE ICASSP*, France, 2006.

[4] B. Ionescu, J. Schlüter, I. Mironica, and M. Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, ICMR '13, pages 215–222, New York, NY, USA, 2013. ACM.

[5] C. Liu, L. Xie, and H. Meng. Classification of music and speech in mandarin news broadcasts. In *Proc. of the 9th Nat. Conf. on Man-Machine Speech Communication (NCMMSC)*, Huangshan, Anhui, China, 2007.

[6] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Trans. on Image Processing*, 18(7):1512–1523, 2009.