ARF @ MediaEval 2012: An Uninformed Approach to Violence Detection in Hollywood Movies

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

Bogdan Ionescu, Ionuţ Mironică
LAPI, University Politehnica of Bucharest, Romania
bionescu@imag.pub.ro
imironica@imag.pub.ro

Markus Schedl
Department of Computational Perception, Johannes Kepler University, Linz, Austria
markus.schedl@jku.at

ABSTRACT

The MediaEval 2012 Affect Task challenged participants to automatically find violent scenes in a set of Hollywood movies. We propose to first predict a set of mid-level concept annotations from low-level visual and auditory features, then fuse the concept predictions and features to detect violent content. Instead of engineering features suitable for the task, we deliberately restrict ourselves to simple general-purpose features with limited temporal context and a generic neural network classifier, setting a baseline for more sophisticated approaches. On 3 test movies, our system detects 49% of violent frames at a precision of 28%, outperforming all other submissions.

1. INTRODUCTION

The MediaEval 2012 Affect Task [1] challenged participants to develop algorithms for finding the most violent scenes in a Hollywood movie from DVD content such as video, audio and subtitles. The organizers provided a training set of 15 movies with frame-accurate annotations of segments containing physical violence as well as several violence-related concepts (such as screams or fire), and a test set of 3 unannotated movies.

We chose to tackle the task as a machine learning problem, employing only a minimum amount of human intelligence in order to set a baseline for more informed approaches. In Section 2, we describe the set of features and classifier we used, and explain how we incorporated the concept annotations into the training process of our violence detector. Section 3 shows our results, and Section 4 gives a conclusion and an outlook on future work.

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.

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 [5] 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 [4]: 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 aligment 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: Weights are initialized to all zeroes, mini-batches are 900 samples, the learning rate starts at 1.0, 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 preliminary experiments on 5 movies.

2.3 Fusion scheme

Given the high variability in appearance of violent scenes in movies and the low amount of training data, training a classifier to predict violent frames directly from the low-level visual and auditory features seems impossible. Instead of designing more advanced feature extractors (such as face and blood detectors), we try to use the concept annotations as a stepping stone: Predicting mid-level concepts from low-level features should be more feasible than directly predicting all forms of physical violence, and predicting violence from mid-level concepts should be easier than from low-level features.

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.

3. EXPERIMENTAL RESULTS

We will first evaluate the performance of the concept predictors, then evaluate the violence predictor and report the official results of our submission on the test set.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

Table 1: Evaluation of concept predictions

concept	vis.	aud.	dim.	prec.	rec.	F-sc.
blood	✓		5	0.07	1.00	0.12
coldarms	✓		1	0.11	1.00	0.19
firearms	✓		1	0.17	0.45	0.24
gore	✓		1	0.05	0.33	0.09
gunshots		✓	4	0.10	0.14	0.12
screams		✓	5	0.08	0.19	0.12
carchase	✓	✓	1	0.01	0.08	0.01
explosions	✓	✓	1	0.08	0.17	0.11
fights	✓	✓	5	0.14	0.29	0.19
fire	✓	✓	1	0.24	0.30	0.26

3.1 Concept prediction

For the training set of 15 movies, each video frame was annotated with 10 different concepts detailed in [1, Sec. 4]. We divide the concepts into visual, auditory and audiovisual categories, then train and evaluate a neural network for each of the concepts in leave-one-movie-out cross-validation.

Table 1 shows our results. For each concept, we list the input features (visual, auditory or both), the number of output dimensions,[1] and precision, recall and F-score at the binarization threshold giving the best F-score. We see that *fire* detection performs best, presumably because it is always accompanied by prominent yellow tones captured well by the visual features. The purely visual concepts (first four rows) obtain high F-scores only because they are so rare that setting a low threshold gives a high recall without hurting precision. Manually inspecting some concept predictions shows that *fire* and *explosions* are accurately detected, *screams* and *gunshots* are mostly correct (although singing is mistaken for screaming, and accentuated fist hits in fights are mistaken for gunshots), but the *blood* predictor does not find any of the numerous blood scenes in "Kill Bill".

3.2 Violence prediction

Equipped with a set of concept predictors of different quality, we proceed to train a frame-wise violence predictor. In the final system, it will get the visual and auditory features as well as all concept predictions as inputs, so we need to provide similar inputs during training. Using the concept ground truth as a substitute for concept predictions will not work – the system would learn to associate blood with violence, then provide inaccurate violence predictions on the test set where we only have highly inaccurate blood predictions. Instead, we train it on the real-valued concept predictor outputs obtained during the cross-validation described in Section 3.1. This allows the system to learn which predictions to trust and which to ignore.

In a final cross-validation, we achieve a frame-wise violence detection precision of 0.23, recall of 0.41 and F-score of 0.30 at a threshold of 0.09. As predictions are noisy, we employ a sliding median filter for temporal smoothing. Trying a selection of filter lengths, we end up smoothing over 150 frames (6 seconds), improving results to a precision of 0.27, recall of 0.46 and F-score of 0.34 at a threshold of 0.07.

[1]Some concepts consist of multiple tags that may or may not be mutually exclusive, e.g., the *gunshots* concept includes guns and cannons. The table gives results for the best tag per concept (space does not allow us to report results for all tags, and averaging over tags gives inconclusive results).

Table 2: Official shot-level results on test set

movie	prec.	rec.	F-sc.	MAP@100
(all)	0.31	0.66	0.42	0.65
Dead Poets Society	0.13	0.38	0.19	0.50
Fight Club	0.30	0.54	0.39	0.57
Independence Day	0.34	0.79	0.48	0.89

3.3 Official results

We submitted two runs on the task's test set. The *segment-level* run forms segments of consecutive frames our predictor tagged as violent or non-violent, the *shot-level* run uses the shot boundaries provided by the task organizers. For both runs, each segment (whether obtained from our predictor or given by the organizers) 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 0.07, the threshold we found above.

We achieve a frame-wise precision of 0.28, recall of 0.49 and F-score of 0.36 in the segment-level run. Table 2 details the results for the shot-level run used for the official ranking. Results vary dramatically with the movie considered: Our system works well on "Independence Day", an action movie featuring fire and explosions, but gives many false positives on "Dead Poets Society", a comparatively peaceful movie.

4. CONCLUSION AND OUTLOOK

Our results show that a naive attempt at violence detection – with features too simple to allow any higher-level understanding of movie segments – can do fairly well, possibly due to cinematic techniques commonly used in Hollywood action scenes. This sets a high baseline to be challenged by more sophisticated approaches. For further insights, we will check if and how the concept predictions helped detecting violence, and compare to other participants' methods.

5. ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF) under project no. Z159, and by the research grant EXCEL POSDRU/89/1.5/S/62557.

6. REFERENCES

[1] C. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2012 Affect Task: Violent Scenes Detection in Hollywood Movies. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[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] 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.

[5] 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.