A Naïve Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies

Bogdan Ionescu
LAPI, University Politehnica of Bucharest
061071 Bucharest, Romania.
bionescu@imag.pub.ro

Jan Schlüter
Austrian Research Institute for Artificial Intelligence
A-1010 Vienna, Austria.
jan.schlueter@ofai.at

Ionuţ Mironică
LAPI, University Politehnica of Bucharest
061071 Bucharest, Romania.
imironica@imag.pub.ro

Markus Schedl
Department of Computational Perception, JKU
A-4040 Linz, Austria.
markus.schedl@jku.at

ABSTRACT

In this paper we approach the issue of violence detection in typical Hollywood productions. Given the high variability in appearance of violent scenes in movies, training a classifier to predict violent frames directly from visual or/and auditory features seems rather difficult. Instead, we propose a different perspective that relies on fusing mid-level concept predictions that are inferred from low-level features. This is achieved by employing a bank of multi-layer perceptron classifiers featuring a dropout training scheme. Experimental validation conducted in the context of the Violent Scenes Detection task of the MediaEval 2012 Multimedia Benchmark Evaluation show the potential of this approach that ranked first among 34 other submissions, in terms of precision and F_1-score.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding; I.5.3 [**Pattern Recognition**]: Classification—*violence detection*.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

violence detection, multimodal video description, multi-layer perceptron, Hollywood movies.

1. INTRODUCTION

Accessing multimedia information or "content" is now part of our daily routine. The Internet, social media and -networks have skyrocketed and video content is available abundantly, basically on every kind of visual terminal. Video broadcasting footage (e.g., YouTube, Dailymotion, Blip.tv) is now the largest broadband traffic category on the Internet, comprising more than a quarter of total traffic.[1] In this context, one of the emerging research areas is the automatic filtering of video contents. The objective is to select appropriate content for different user profiles or audiences. A particular case is the filtering of affect content related to violence, for instance for banning children from accessing it or for automatic video content rating.

Defining the term "violence" is not an easy task, as this notion remains subjective and thus dependent on people [1]. Definitions range from literal ones such as "actions or words which are intended to hurt people"[2] or "physical violence or accident resulting in human injury or pain" [2] to more technical film-making related where this notion is defined by specific visual-auditory indicators, e.g., high-speed movements or fast-paced music [3].

In this paper we address the problem of violence detection in the context of typical Hollywood movies. Our approach relies on fusing mid-level concept predictions made using multi-layer perceptron classifiers. The final goal is to automatically localize the occurrence of violence within a video.

The remainder of the article is organized as follows: Section 2 presents a detailed overview of the current state-of-the-art of the research in violent scene detection. Section 3 introduces the proposed approach, while Section 4 details the classification scheme involved. Experimental validation is presented in Section 5. Section 6 concludes the paper.

2. PREVIOUS WORK

Due to the complexity of the research problem, starting with the formulation of the task (i.e., defining violence) to the inference of highly semantic concepts out of low-level information, the problem of violence detection in videos has been marginally studied in the literature. Some of the most representative approaches are reviewed in the sequel.

A related domain is the *detection of affective content* in

[1]Source: CISCO systems, `http://www.cisco.com`.
[2]Source: Cambridge dictionary, `http://dictionary.cambridge.org`.

WOODSTOCK '97 El Paso, Texas USA

videos, which refers to the intensity (i.e. arousal) and type (i.e. valence) of emotion that are expected to arise in the user while watching a certain video [4]. Existing methods attempt to map low-level features (e.g., low-level audio-visual features, users' physiological signals) to high-level emotions [5, 6]. If we refer to violence as an expected emotion in videos, affect-related features may be applicable to represent the violence concept [7].

Another related domain is *action recognition*, which focuses on detecting *human violence in real-world scenarios*. An example is the method in [8] that proposes an in-depth hierarchical approach for detecting distinct violent events involving two people, e.g., fist fighting, hitting with objects, kicking. The information used consists of computing the motion trajectory of image structures (acceleration measure vector and its jerk). The framework is preliminarily validated on 15 short-time sequences including around 40 violent scenes. Another example is the approach in [9] that aims to detect instances of aggressive human behavior in public environments. The authors use a Dynamic Bayesian Network (DBN) as a fusion mechanism to aggregate aggression scene indicators, e.g., "scream", "passing train" or "articulation energy". Evaluation is carried out using 13 clips featuring various scenarios, such as "aggression toward a vends machine" or "supporters harassing a passenger". The method reports an accuracy score close to 80%.

The use of Bag-of-Visual-Words (BoVW) statistical models has also been exploited. For instance, [10] addresses fight detection using BoVW along with Space-Time Interest Points (STIP) and Motion Scale-Invariant Feature Transform (MoSIFT) features. In addition, for the purpose of evaluation and to foster research on violence detection, the authors attempt to introduce a standard testing data set consisting of 1,000 clips of action scenes from hockey games. Ground truth is provided at frame level (as "fight" or "non-fight" labeling). Highest reported detection accuracy is near 90%. A similar experiment is the one in [11] that uses BoVW with local spatio-temporal features. Experimental tests show that for this scenario motion patterns tend to provide better performance than spatio-visual descriptors. Tests are conducted on sports and surveillance videos.

A broader category of approaches focus on a more general framework, such as detecting *video shots/segments with violent content* that may be considered disturbing for different categories of viewers. These methods are typically addressing video TV broadcasting materials, such as Hollywood entertainment movies. One of the early approaches in this direction is the one in [12], where the violent events are located using multiple audio-visual signatures, e.g., description of motion activity, blood and flame detection, violence/non-violence classification of the soundtrack and characterization of sound effects. Only qualitative validation is reported. Other examples include the following: [3] exploits shot length, motion activity, loudness, speech, light, and music. Features are combined using a modified semi-supervised learning scheme that uses Semi-Supervised Cross Feature Learning (SCFL). The method is preliminarily evaluated using 4 Hollywood movies, yielding a top F_1-score of 85%; [13] combines audio-visual features (e.g., shot length, speech, music ratios, motion intensity) to select representative shots in typical action movies with the objective of producing automatic video trailers. Content classification is performed with Support Vector Machines (SVMs); [14] uses various audio features (e.g., spectrogram, chroma, energy entropy, Mel-Frequency Cepstral Coefficients (MFCC)) and visual descriptors (e.g., average motion, motion orientation variance, measure of the motion of people or faces in the scene). Modalities are combined by employing a meta-classification architecture that classifies mid-term video segments as "violent" or "non-violent". Experimental validation is performed on 10 movies and highest F_1-score is up to 58%; [15] proposes a violent shot detector that uses a modified probabilistic Latent Semantic Analysis (pLSA) to detect violence from the auditory content while the visual information is exploited via motion, flame, explosion and blood analysis. Final integration is achieved using a co-training scheme (typically used when dealing with small amounts of training data and large amounts of unlabeled data). Experimental validation is conducted on 5 movies showing an average F_1-score of 88% (however there is no information on the ground truth used). More recently, approaches also consider the benefits of temporal integration of features and late fusion integration schemes, e.g., [16].

Although most of the approaches are multimodal, there are some attempts to exploit the benefits of single modalities, e.g., [18] uses Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) for modeling audio events over time series. For experimentation, authors model the presence of gunplay and car racing scenes with audio events such as "gunshot", "explosion", "engine", "helicopter flying", "car braking", and "cheers". Validation is performed on a very restrained data set (excerpts of 5 minutes extracted from 5 movies) leading to average F_1-scores of up to 90%; [19] uses face, blood, and motion information to determine whether an action scene has violent content or not. The specificity of the approach is in addressing more semantics-bearing scene structure of video rather than simple shots.

In general, most of the existing approaches focus more or less on engineering content descriptors that may be able to highlight the specificity of violent contents or on the detection of concepts associated with it. Unfortunately, there is a lack of a unified evaluation framework. Almost all of the existing approaches are tested either on very limited data sets (just a few excerpts), on "closed" data or on specific domains (e.g., only sports). Another problem lies in the violence related ground truth that reflects different understandings of the concept. It tends to vary dramatically from method to method and to be adapted to each of the data set (proof is the high disparity of reported results and also the very high accuracy in some cases). This hinders reproducibility of the results and renders impossible performance comparison.

In the context of movie violence - which is the subject of this work - there is a sustained effort made by the community of the Violent Scenes Detection task of the MediaEval Multimedia Benchmark Evaluation [17] to constitute a reference evaluation framework for validating violence detection methods. It proposes a standardized data set together with a detailed annotation ground truth of several audio-visual concepts related to violence [2] (see Section 5).

In this paper we propose a different perspective that exploits the use of mid-level concepts in a multiple neural network fusing scheme. The proposed approach goes beyond the current state-of-the-art along these dimensions:

- by addressing a highly complex scenario where violence is considered to be any scene involving human injury or pain;

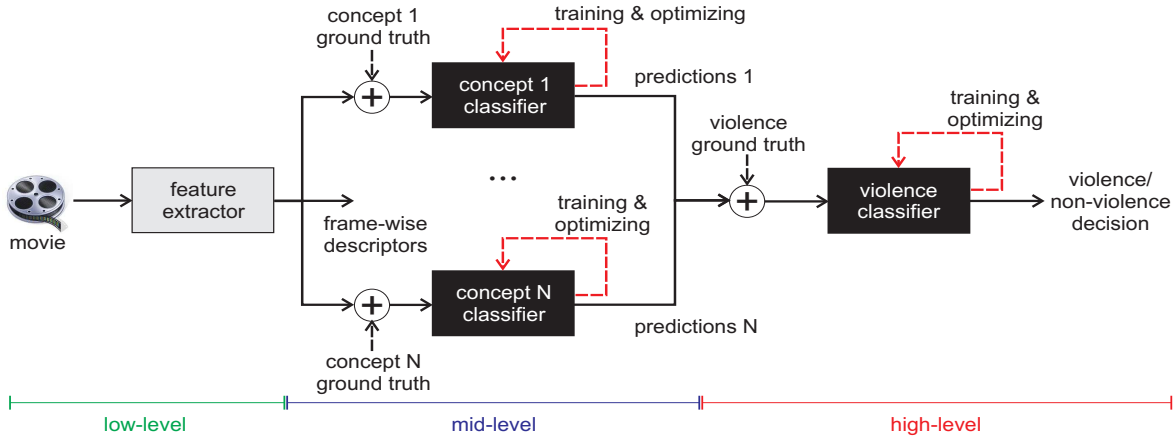


Figure 1: Method diagram (semantic level increases from left to right).

- thanks to the fusion of mid-level concept predictions, the method is feature-independent in the sense that it does not require the design of adapted features;
- violence is predicted at frame level which facilitates detection of arbitrary length segments, not only fixed length (e.g., shots);
- evaluation is carried out on a standard data set [2] making the results both relevant and reproducible.

3. PROPOSED APPROACH

Instead of focusing on engineering a best content description approach suitable for this task, as most of the existing approaches do, we propose a novel perspective. Given the high variability in appearance of violent scenes in movies and the low amount of training data that is usually available, training a classifier to predict violent frames directly from visual and auditory features seems rather ineffective.

We propose instead to use high-level concept ground-truth obtained from manual annotation to infer mid-level concepts as a stepping stone towards the final goal. Predicting mid-level concepts from low-level features should be more feasible than directly predicting all forms of violence (highly semantic). Also, predicting violence from mid-level concepts should be easier than using directly the low-level content features.

A diagram of the proposed method is shown in Figure 1. First, we perform feature extraction. Features are extracted at frame level (see Section 5.2). 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. Once we captured data characteristics we may *classify* unseen video frames into one of the two categories: "violence" and "non-violence". Consecutively, violence frames are aggregated to segments. Each of the two steps is presented in the sequel.

3.1 Concept and violence training

To train the system we use ground truth data at two levels: ground truth related to concepts that are usually present in the violence scenes, such as presence of "fire", presence of "gunshots", or "gory" scenes (more information is presented in Section 5) and ground truth related to the actual violence segments. We used the data set provided in [2].

The mid-level concept detection consists of a bank of classifiers that are trained to respond to each of the target violence-related concepts. At this level, the response of the classifier is optimized for best performance. Tests are repeated for different parameter setups until the classifier yields the highest accuracy. Each classifier state is then saved. With this step, initial features are therefore transformed into concept predictions (real valued between [0;1]).

The high-level concept detection is ensured by a final classifier that is fed with the previous concept predictions and acts as a final fusion scheme. The output of the classifier is thresholded to achieve the labeling of each frame as "violent" or "non-violent" (yes/no decision). As in the previous case, we use the violence ground truth to tune the classifier to its optimal results (e.g., setting the best threshold). The classifier state is again conserved.

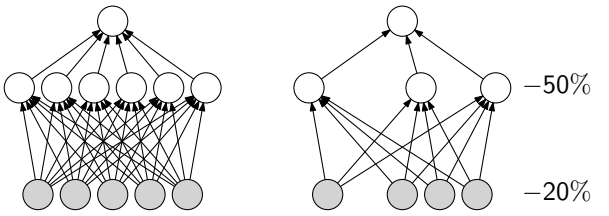
3.2 Violence classification

Equipped with a violence frame predictor, we may proceed to label new unseen video sequences. Based on the previous classifier states, new frames are now labeled as "violent" or "non-violent". Depending on the final usage, aggregation into segments can be performed at two levels: arbitrary length segments and video shot segments. The *segment-level* tagging forms segments of consecutive frames our predictor tagged as violent or non-violent and the *shot-level* tagging uses preliminary shot boundary detection [20].

For both approaches, each segment (whether obtained at the level of frames or shots) 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 the optimal threshold found previously in the training of the violence classifier.

4. NEURAL NETWORK CLASSIFIER

To choose the right classification scheme for this particular fusion task, we conducted several preliminary experimental tests using a broad variety of classifiers, from functional-based (e.g., Support Vector Machines), decision trees to neural networks. Most of the classifiers failed in providing rele-



(a) a feed-forward neural network

(b) the same network with randomly dropped units

Figure 2: Illustrating random dropouts of network units: 2a shows the full classifier, 2b is one of the possible versions trained on a single example.

vant results when coping with high amount of input data, i.e. labeling of individual frames rather than video segments. The inherent parallel architecture of neural networks fitted well these requirements, in particular the use of multi-layer perceptrons. Therefore, for the concept and violence classifiers (see Figure 1) we employ a multi-layer perceptron with a single hidden layer of 512 logistic sigmoid units and as many output units as required for the respective concept. Some of the concepts from [2] consist of independent tags, for instance "fights" encompasses the five tags "1vs1", "default", "distant attack", "large" and "small" (in which case we use five independent output units).

Networks are trained by gradient descent on the cross-entropy error with backpropagation [26], using a recent idea by Hinton et al. [25] to improve generalization: For each presented training case, a fraction of input and hidden units is omitted from the network and the remaining weights are scaled up to compensate. Figure 2 visualizes this for a small network, with 20% of input units and 50% of hidden units "dropped out". The set of dropped units is chosen at random for each presentation of a training case, such that many different combinations of units will be trained during an epoch.

This helps generalization in the following way: By randomly omitting units from the network, a higher-level unit cannot rely on all lower-level units being present and thus cannot adapt to very specific combinations of a few parent units only. Instead, it is driven to find activation patterns of a larger group of correlated units, such that dropping a fraction of them does not hinder recognizing the pattern. For example, when recognizing written digits, a network trained without dropouts may find that three particular pixels are enough to tell apart ones and sevens on the training data. A network trained with input dropouts is forced to take into account several correlated pixels per hidden unit and will learn more robust features resembling strokes.

Hinton et al. [25] showed that features learned with dropouts generalize better, improving test set performance on three very different machine learning tasks. This encouraged us to try their idea for our data as well, and indeed we observed an improvement of up to 5%-points F_1-score in all our experiments. As an additional benefit, a network trained with dropouts does not severely overfit to the training data, eliminating the need for early stopping on a validation set to regularize training.

5. EXPERIMENTAL RESULTS

The experimental validation of our approach was carried out in the context of the 2012 MediaEval Benchmarking Initiative for Multimedia Evaluation, Affect task: Violent Scenes Detection [17]. It proposes a corpus of 18 Hollywood movies of different genres, from extremely violent movies to movies without violence. Movies are divided into a development set, consisting of 15 movies: "Armageddon", "Billy Elliot", "Eragon", "Harry Potter 5", "I am Legend", "Leon", "Midnight Express", "Pirates of the Caribbean 1", "Reservoir Dogs", "Saving Private Ryan", "The Sixth Sense", "The Wicker Man", "Kill Bill 1", "The Bourne Identity", and "The Wizard of Oz" (total duration of 27h 58min, 26,108 video shots and violence duration ratio 9.39%); and a test set consisting of 3 movies - "Dead Poets Society", "Fight Club", and "Independence Day" (total duration 6h 44min, 6,570 video shots and violence duration ratio 4.92%). Overall the entire data set contains 1,819 violence segments [2].

Ground truth is provided at two levels. Frames are annotated according to 10 violence related high-level concepts, namely: "presence of blood", "fights", "presence of fire", "presence of guns", "presence of cold weapons", "car chases" and "gory scenes" (for the video modality); "presence of screams", "gunshots" and "explosions" (for the audio modality) [1]; and frame segments are labeled as "violent" or "non-violent". Ground truth was created by 9 human assessors.

For evaluation, we use classic precision and recall:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN} \tag{1}$$

where TP stands for true positives (good detections), FP are the false positives (false detections) and FN the false negatives (the misdetections). To have a global measure of performance, we also report F_1-scores:

$$F_1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

Values are averaged over all experiments.

5.1 Parameter tuning and preproceesing

The proposed approach involves the choice of several parameters and preprocessing steps.

In what concerns the multi-layer perceptron, we follow the dropout scheme in [25, 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. We use a single hidden layer of 512 units. These settings worked well in preliminary experiments on 5 movies.

To avoid multiple scale values for the various content features, the input data of the concept predictors is normalized by subtracting the mean and dividing by the standard deviation of each input dimension. Also, as the concept predictors are highly likely to yield noisy outputs, we employ a sliding median filter for temporal smoothing of the predictions. Trying a selection of filter lengths, we end up smoothing over 125 frames, i.e. 5 seconds.

5.2 Video descriptors

We experimented with several video description approaches that proved to perform well in various video and audio classification scenarios [17, 22, 23, 27]. Given the specificity of the task, we derive *audio*, *color*, *feature description* and *temporal structure* information. Descriptors are extracted at frame level as follows:

- **audio** (196 dimensions): we use a general-purpose set of audio descriptors: *Linear Predictive Coefficients* (LPCs), *Line Spectral Pairs* (LSPs), *MFCCs*, *Zero-Crossing Rate* (ZCR), and *spectral centroid*, *flux*, *rolloff*, and *kurtosis*, augmented with the variance of each feature over a window of 0.8 s centered at the current frame [24, 27];

- **color** (11 dimensions): to describe global color contents, we use the Color Naming Histogram proposed in [22]. It maps colors to 11 universal color names: "black", "blue", "brown", "grey", "green", "orange", "pink", "purple", "red", "white", and "yellow";

- **features** (81 values): we use a 81-dimensional Histogram of Oriented Gradients (HoG) [23].

- **temporal structure** (single dimension): to account for temporal information we use a measure of visual activity. We use the cut detector in [21] that measures visual discontinuity by means of difference between color histograms of consecutive frames. To account for a broader range of significant visual changes, but still rejecting small variations, we lower the threshold used for cut detection. Then, for each frame we determine the number of detections in a certain time window centered at the current frame (e.g., for violence detection good performance is obtained with 2 s windows). High values of this measure will account for important visual changes that are typically related to action.

5.3 Cross-validation training results

In this experiment we aim to train and evaluate the performance of the neural network classifiers according to concept and violence ground truth. We used the development set of 15 movies. Training and evaluation are performed using a leave-one-movie-out cross-validation approach.

5.3.1 Concept prediction

First, we train 10 multi-layer perceptrons to predict each of the violence-related high-level concepts. The results of the cross-validation are presented in Table 1. For each concept, we list the input features (visual, auditory, or both) and average precision, recall and F_1-score at the binarization threshold giving the best F_1-score (real valued outputs of the perceptrons are thresholded to achieve yes/no decisions).

Results show that the highest precision and recall are up to 24% and 100%, respectively, while the highest F_1-score is of 26%. Detection of fire 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_1-score 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 frequently mistaken for screaming, and accentuated fist hits in fights are often mistaken for gunshots).

5.3.2 Violence prediction

Given the previous set of concept predictors of different qualities, we proceed to train the frame-wise violence predictor. Using the concept ground truth as a substitute for concept predictions will likely yield poor results - the system would learn, for example, 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 on the real-valued concept predictor outputs obtained during the cross-validation described in Section 5.3.1. This allows the system to learn which predictions to trust and which to ignore.

Table 1: Evaluation of concept predictions.

concept	visual	audio	precision	recall	F_1-score
blood	✓		7%	100%	12%
coldarms	✓		11%	100%	19%
firearms	✓		17%	45%	24%
gore	✓		5%	33%	9%
gunshots		✓	10%	14%	12%
screams		✓	8%	19%	12%
carchase	✓	✓	1%	8%	1%
explosions	✓	✓	8%	17%	11%
fights	✓	✓	14%	29%	19%
fire	✓	✓	24%	30%	26%

The violence predictor achieves precision and recall values of 28.29% and 37.64%, respectively and an F_1-score of 32.3% (results are obtained for the optimal binarization threshold). The results are very promising considering the difficulty of the task and the diversity of movies. The fact that we obtain better performance compared to the detection of individual concepts may be due to the fact that violent scenes often involve the occurrence of several concepts, not only one, which may compensate for some low concept detection performance. A comparison with other techniques is presented in the following section.

5.4 MediaEval 2012 results

In the final experiment we present a comparison of the performance of our violence predictor in the context of the 2012 MediaEval Benchmarking Initiative for Multimedia Evaluation, Affect task: Violent Scenes Detection [2, 17].

In this task, participants were provided with the development data set (15 movies) for training their approaches while the official evaluation was carried out on 3 test movies: "Dead Poets Society" (34 violent scenes), "Fight Club" (310 violent scenes) and "Independence Day" (371 violence scenes) - a total of 715 violence scenes (ground truth for the test set was released after the competition). A total number of 8 teams participated providing 36 runs. Evaluation was conducted both at video shot and segment level (arbitrary length). The results are discussed in the sequel.

5.4.1 Shot-level results

In this experiment, video shots (shot segmentation was provided by organizers [2, 1]) are tagged as being "violent" or "non-violent". Frame-to-shot aggregation is carried out as presented in Section 3.2. Performance assessment is conducted on a per-shot basis. To highlight the contribution of the concepts, our approach is assessed with different feature combinations (see Table 3). A summary of the best team runs is presented in Table 2 (results are presented in decreasing order of F_1-score values).

The use of mid-level concept predictions and multi-layer perceptron (see ARF-(c)) ranked first and achieved the high-

Table 2: Violence shot-level detection results at MediaEval 2012 [2][17].

team	descriptors	modality	method	precision	recall	F_1-score
ARF-(c)	concepts	audio-visual	**proposed**	46.14%	54.40%	49.94%
ARF-(a)	audio	audio	**proposed**	46.97%	45.59%	46.27%
ARF-(av)	audio, color, HoG, temporal	audio-visual	**proposed**	32.81%	67.69%	44.58%
ShanghaiHongkong [30]	trajectory, SIFT, STIP, MFCC	audio-visual	temp. smoothing + SVM with χ^2 kernel	41.43%	46.29%	43.73%
ARF-(avc) [34]	audio, color, HoG, temporal & concepts	audio-visual	**proposed**	31.24%	66.15%	42.44%
TEC [33]	TF-IDF B-o-AW [16], audio, color	audio-visual	fusion SVM with HIK and χ^2 kernel & Bayes Net. & Naive Bayes	31.46%	55.52%	40.16%
TUM [29]	energy & spectral audio	audio	SVM linear kernel	40.39%	32.00%	35.73%
ARF-(v)	color, HoG, temporal	visual	**proposed**	25.04%	61.95%	35.67%
LIG [31]	color, texture, SIFT, B-o-AW of MFCC	audio-visual	hierarch. fusion of SVMs & k-NNs with conceptual feedback	26.31%	42.09%	32.38%
TUB [7]	audio, B-o-AW MFCC, motion	audio-visual	SVM with RBF kernel	19.00%	62.65%	29.71%
DYNI [32]	MS-LBP texture [35]	visual	SVM with linear kernel	15.55%	63.07%	24.95%
NII [28]	concept learned from texture & color	visual	SVM with RBF kernel	11.40%	89.93%	20.24%

Notations: SIFT - Scale Invariant Features Transform, STIP - Spatial-Temporal Interest Points, MFCC - Mel-Frequency Cepstral Coefficients, SVM - Support Vector Machines, TF-IDF - Term Frequency-Inverse Document Frequency, B-o-AW - Bag-of-Audio-Words, HIK - Histogram Intersection Kernel, k-NN - k Nearest Neighbors, RBF - Radial Basis Function, MS-LBP - Multi-Scale Local Binary Pattern.

Table 3: Feature combinations.

run	description
ARF-(c)	use of only mid-level concept predictions;
ARF-(a)	use of only audio descriptors (the violence classifier is trained directly on the audio features);
ARF-(v)	use of only visual features;
ARF-(ac)	use of only audio-visual features;
ARF-(avc)	use of all concept and audio-visual features (the violence classifier is trained using the fusion of concept predictions and features).

est F_1-score of 49.94%, that is an improvement of more than 6 percentage points over the other teams' best runs, i.e. team ShanghaiHongkong [30], F_1-score of 43.73%. For our approach, the lowest discriminative power is provided by using only the visual descriptors (see ARF-(v)), where the F_1-score is only 35.65%. Compared to visual features, audio features seem to show better descriptive power, providing the second best F_1-score of 46.27%. The combination of descriptors (early fusion) tends to reduce their efficiency and yields lower performance than the use of concepts alone, e.g., audio-visual (see ARF-(av)) yields an F_1-score of 44.58%, while audio-visual-concepts (see ARF-(avc)) 42.44%.

Another observation is that, despite the use of general purpose descriptors (see Section 5.2), the representation of feature information via mid-level concepts allows better performance than other, more elaborate content description approaches or classification, such as the use of SIFTs, B-o-AW of MFCC or motion information.

Figure 3 details the precision-recall curves for our approach. One may observe that the use of concepts alone (red line) provides significantly higher recall than the sole use of audio-visual features or the combination off all for a

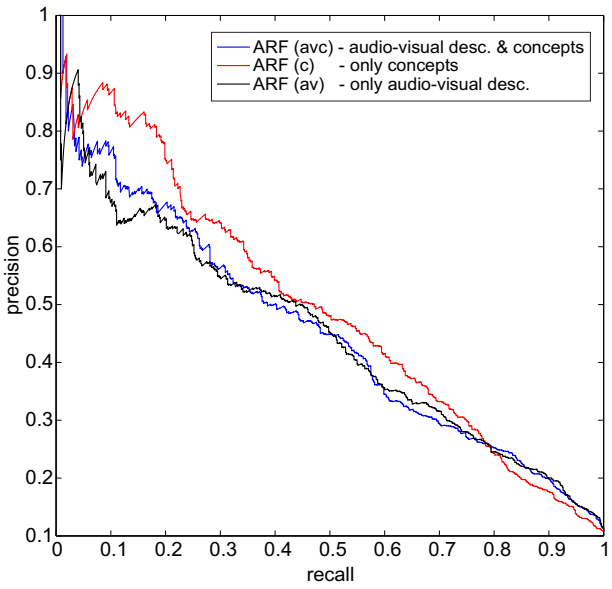


Figure 3: Precision-recall curves for the proposed approach.

precision of 25% and above.

5.4.2 *Arbitrary segment-level results*

The final experiment is conducted at segment level. Video segments of arbitrary length are tagged as "violent" or "non-violent". Frame-to-segment integration is carried out as presented in Section 3.2. The performance assessment in this case is conducted on a per-unit-of-time basis.

Using the mid-level concepts, we achieve average precision

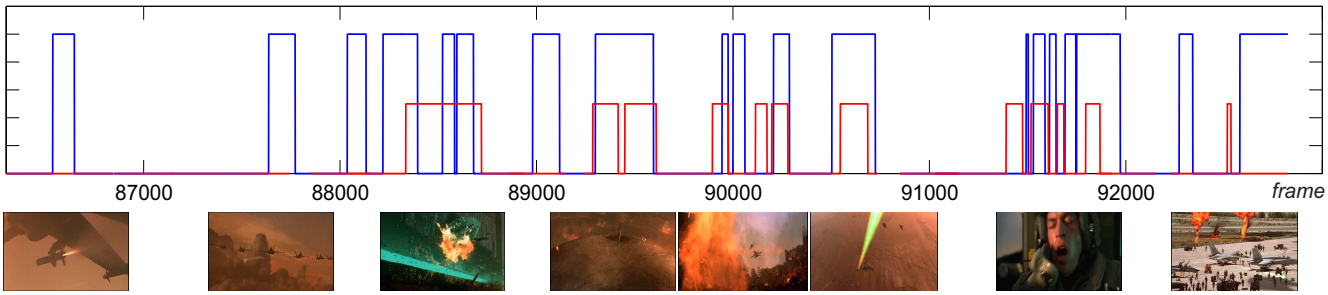


Figure 4: Examples of violent segment detection in movie "Independence Day" (the oX axis is the time axis, the values on oY axis are arbitrary, ground truth is depicted in red while the detected segments in blue).

and recall values of 42.21% and 40.38%, respectively, while the F_1-score amounts to 41.27%. This yields a miss rate (at time level) of 50.69% and a very low false alarm rate of only 6%. These results are also very promising considering the difficulty of detecting precisely the exact time interval of violent scenes, but also the subjectivity of the human assessment (reflected in the ground truth). Comparison with other approaches was not possible in this case as all other teams provided only shot-level detection.

5.4.3 *Violence detection examples*

Figure 4 illustrates an example of violent segments detected by our approach in the movie "Independence Day". For visualization purposes, some of the segments are depicted with a small vignette of a representative frame.

In general, the method performed very well on the movie segments related to action (e.g., involving explosions, firearms, fire, screams) and tends to be less efficient for segments where violence is encoded in the meaning of human actions (e.g., fist fights or car chases). Examples of false detections are due to visual effects that share similar audio-visual signatures with the violence-related concepts. Common examples include accentuated fist hits, loud sounds or the presence of fire not related to violence (e.g., see the rocket launch or the fighter flight formation in Figure 4, first two images). Misdetection is in general caused by limited accuracy of the concept predictors (see last image in Figure 4, where some local explosions filmed from a distance have been missed).

6. CONCLUSIONS

We presented a naive approach to the issue of violence detection in Hollywood movies. Instead of using concept descriptors to learn directly how to predict violence, as most of the existing approaches do, the proposed approach relies on an intermediate step consisting of predicting mid-level violence concepts. Predicting mid-level concepts from low-level features seems to be more feasible than directly predicting all forms of violence. Predicting violence from mid-level concepts proves to be much easier than using directly the low-level content features. Content classification is performed with a multi-layer perceptron whose parallel architecture fits well the target of labeling individual video frames. The approach is naive in the sense of its simplicity. Nevertheless, its efficiency in predicting arbitrary length violence segments is remarkable. The proposed approach ranked first in the context of the 2012 Affect Task: Violent Scenes Detection at MediaEval Multimedia Benchmark (out of 36 total submissions). However, the main limitation of the method is its dependence on a detailed annotation of violent concepts, inheriting at some level its human subjectivity. Future improvements will include exploring the use of other information sources, such as text (e.g., subtitles that are usually provided with movie DVDs).

7. ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF): Z159 and P22856-N23, and by the research grant EXCEL POSDRU/89/1.5/S/62557. We acknowledge the 2012 Affect Task: Violent Scenes Detection of the MediaEval Multimedia Benchmark `http://www.multimediaeval.org` for providing the test data set that has been supported, in part, by the Quaero Program `http://www.quaero.org`. We also acknowledge the violence annotations, shot detections and key frames made available by Technicolor [1].

8. REFERENCES

[1] Technicolor, `http://www.technicolor.com`, last accessed 2012.

[2] C.-H. Demarty, C. Penet, G. Gravier, M. Soleymani, "The MediaEval 2012 Affect Task: Violent Scenes Detection in Hollywood Movies", Working Notes Proc. of the MediaEval 2012 Workshop [17], `http://ceur-ws.org/Vol-927/mediaeval2012_submission_3.pdf`.

[3] Y. Gong, W. Wang, S. Jiang, Q. Huang, W. Gao, "Detecting Violent Scenes in Movies by Auditory and Visual Cues", 9th Pacific Rim Conf. on Multimedia: Advances in Multimedia Information Processing, pp. 317-326. Springer-Verlag, 2008.

[4] A. Hanjalic, L. Xu, "Affective Video Content Representation and Modeling", IEEE Trans. on Multimedia, pp. 143-154, 2005.

[5] M. Soleymani, G. Chanel, J. J. Kierkels, T. Pun. "Affective Characterization of Movie Scenes based on Multimedia Content Analysis and User's Physiological Emotional Responses", IEEE Int. Symp. on Multimedia, Berkeley, California, USA, 2008.

[6] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, H. Lu, "A Three-level Framework for Affective Content Analysis and its Case Studies", Multimedia Tools and Applications, 2012.

[7] E. Açar, S. Albayrak, "DAI Lab at MediaEval 2012 Affect Task: The Detection of Violent Scenes using

Affective Features", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_33.pdf.

[8] A. Datta, M. Shah, N. Da Vitoria Lobo, "Person-on-Person Violence Detection in Video Data", Int. Conf. on Pattern Recognition, 1, p. 10433, Washington, DC, USA, 2002.

[9] W. Zajdel, J. Krijnders, T. Andringa, D. Gavrila, "CASSANDRA: Audio-Video Sensor Fusion for Aggression Detection", IEEE Conf. on Advanced Video and Signal Based Surveillance, pp. 200-205, London, UK, 2007.

[10] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence Detection in Video using Computer Vision Techniques", Int. Conf. on Computer Analysis of Images and Patterns, LNCS 6855, pp. 332-339, 2011.

[11] Fillipe D. M. de Souza, Guillermo C. Chávezy, Eduardo A. do Valle Jr., Arnaldo de A. Araújo, "Violence Detection in Video Using Spatio-Temporal Features", 23rd SIBGRAPI Conf. on Graphics, Patterns and Images, pp. 224-230, 2010.

[12] J. Nam, M. Alghoniemy, A.H. Tewfik, "Audio-Visual Content-Based Violent Scene Characterization", IEEE Int. Conf. on Image Processing, 1, pp. 353 - 357, 1998.

[13] A.F. Smeaton, B. Lehane, N.E. O'Connor, C. Brady, G. Craig, "Automatically Selecting Shots for Action Movie Trailers", 8th ACM Int. Workshop on Multimedia Information Retrieval, pp. 231Ű-238, 2006.

[14] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, S. Theodoridis, "Audio-Visual Fusion for Detecting Violent Scenes in Videos", Artificial Intelligence: Theories, Models and Applications, LNCS 6040, pp 91-100, 2010.

[15] J. Lin, W. Wang, "Weakly-Supervised Violence Detection in Movies with Audio and Video based Co-training", 10th Pacific Rim Conf. on Multimedia: Advances in Multimedia Information Processing, pp. 930-935, Springer-Verlag, 2009.

[16] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Kyoto, 2012.

[17] MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, http://www.multimediaeval.org/, last accessed 2012.

[18] W.-H. Cheng, W.-T. Chu, J.-L. Wu, "Semantic Context Detection based on Hierarchical Audio Models", ACM Int. Workshop on Multimedia Information Retrieval, pp. 109 - 115, 2003.

[19] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, C.-W. Su, "Violence Detection in Movies", 8th Int. Conf. Computer Graphics, Imaging and Visualization, pp.119-124, 2011.

[20] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved?", IEEE Trans. on Circuits and Systems for Video Technology, 12(2), pp. 90 - 105, 2002.

[21] B. Ionescu, V. Buzuloiu, P. Lambert, D. Coquin, "Improved Cut Detection for the Segmentation of Animation Movies", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, France, 2006.

[22] J. Van de Weijer, C. Schmid, J. Verbeek, D. Larlus, "Learning color names for real-world applications", IEEE Trans. on Image Processing, 18(7), pp. 1512-1523, 2009.

[23] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", IEEE Int. Conf. On Intelligent Transportation Systems, 1, pp. 432-437, St. Louis, 2009.

[24] Yaafe core features, http://yaafe.sourceforge.net/, last accessed 2012.

[25] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors", arXiv.org, http://arxiv.org/abs/1207.0580, 2012.

[26] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-Propagating Errors", Nature, 323, pp. 533–536, 1986.

[27] C. Liu, L. Xie, H. Meng, "Classification of Music and Speech in Mandarin News Broadcasts", Int. Conf. on Man-Machine Speech Communication, China, 2007.

[28] V. Lam, D.-D. Le, S.-P. Le, Shin'ichi Satoh, D.A. Duong, "NII Japan at MediaEval 2012 Violent Scenes Detection Affect Task", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_21.pdf.

[29] F. Eyben, F. Weninger, N. Lehment, G. Rigoll, B. Schuller, "Violent Scenes Detection with Large, Brute-forced Acoustic and Visual Feature Sets", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_25.pdf.

[30] Y.-G. Jiang, Q. Dai, C.C. Tan, X. Xue, C.-W. Ngo, "The Shanghai-Hongkong Team at MediaEval2012: Violent Scene Detection Using Trajectory-based Features", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_28.pdf.

[31] N. Derbas, F. Thollard, B. Safadi, G. Quénot, "LIG at MediaEval 2012 Affect Task: use of a Generic Method", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_39.pdf.

[32] V. Martin, H. Glotin, S. Paris, X. Halkias, J.-M. Prevot, "Violence Detection in Video by Large Scale Multi-Scale Local Binary Pattern Dynamics", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_43.pdf.

[33] C. Penet, C.-H. Demarty, M. Soleymani, G. Gravier, P. Gros , "Technicolor/INRIA/Imperial College London at the MediaEval 2012 Violent Scene Detection Task", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_26.pdf.

[34] J. Schlüter, B. Ionescu, I. Mironică, M. Schedl , "ARF @ MediaEval 2012: An Uninformed Approach to Violence Detection in Hollywood Movies", Working Notes Proc. of the MediaEval 2012 Workshop [17], http://ceur-ws.org/Vol-927/mediaeval2012_submission_36.pdf.

[35] S. Paris, H. Glotin, "Pyramidal Multi-level Features for the Robot Vision @ICPR 2010 Challenge", 20th Int. Conf. on Pattern Recognition, pp. 2949 - 2952, Marseille, France, 2010.