Audio-Visual Fusion for Detecting Violent Scenes in Videos

Theodoros Giannakopoulos[1], Alexandros Makris[1], Dimitrios Kosmopoulos[1], Stavros Perantonis[1], and Sergios Theodoridis[2]

[1] Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center of Scientific Research Demokritos, Greece
tyiannak@gmail.com,{amakris,dkosmo,sper}@iit.demokritos.gr
[2] Department of Informatics and Telecommunications,
University of Athens, Greece
stheodor@di.uoa.gr

Abstract. In this paper we present our research towards the detection of violent scenes in movies, employing fusion methodologies, based on learning. Towards this goal, a multi-step approach is followed: initially, automated auditory and visual processing and analysis is performed in order to estimate probabilistic measures regarding particular audio and visual related classes. At a second stage, a meta-classification architecture is adopted, which combines the audio and visual information, in order to classify mid-term video segments as "violent" or "non-violent". The proposed scheme has been evaluated on a real dataset from 10 films.

Keywords: Violence detection, multi-modal video classification.

1 Introduction

During the last decade, a huge increase of video data has occurred, mainly due to the existence of several file-sharing web communities and new facilities regarding digital television. Therefore, the provided multimedia content is becoming easily accessible by large portions of the population, while limited control over the content exists. Due to this vast amount of multimedia content, the manually performed annotation is obviously a hard task. It is therefore obvious that the need of protection of sensitive social groups (e.g. children), using automatic content-based classification techniques, is imperative. In this paper, we present a method for automatic *violence* detection in films, based on audio-visual information.

There are not many works in the literature which attempt to detect violence scenes using visual features. Most of the methods concern surveillance cameras and use background substraction techniques to detect the people in the scene [1], [2]. These approaches however are not suitable for movies where the camera moves abruptly and there are many shot changes. In [3], a generic approach to determine the presence of violence is presented. Two features are used, which measure the average activity and the average shot-length. Experiments with

S. Konstantopoulos et al. (Eds.): SETN 2010, LNAI 6040, pp. 91–100, 2010.

movie trailers show that the features are able to discriminate violent from non-violent movies. However, no attempt to characterize the specific segments of the movie which contain the violence is carried out. In [4], three visual features are used, measuring the level of activity, the presence of gunfires/explosions and the presence of blood.

In video data, most violent scenes are characterized by specific audio events (e.g. explosions). The literature related to the detection of violent content is limited and usually it examines only visual features ([5], [6]). In [7] a simple audio feature, in particular, the energy entropy, is used as additional information to visual data. In [8], a film classification method is proposed that is mainly based in visual cues. The only audio feature adopted in this paper is the signal's energy. A more detailed examination of the audio features for discriminating between violent and non-violent sounds was presented in [9]. In particular, seven audio features, both from the time and frequency domain, have been used, while the binary classification task (violent and non violent) was accomplished via the usage of Support Vector Machines. In [10], a multi-class classification algorithm for audio segments from movies has been proposed. Bayesian networks along with the one vs all architecture has been used, while the definition of the classes has been violence-oriented (three violent classes have been adopted).

In this work we have used a variant of the classifier proposed in [10], on a segment basis, in order to generate a sequence of audio class labels. As far as the visual part is concerned, in this work we have employed motion and person activity related features, in order to derive three visual-related classes. The two individual modules are combined in a meta-classification stage, which is responsible for classifying video segments in two classes, i.e., "Violence" and "Non-Violence".

2 Audio Classifier

2.1 Audio Class Definition

In order to create an audio-based characterization scheme, we have defined seven audio classes, three from which are violent and four non-violent. The class definitions have been motivated by the nature of the audio signals met in most movies. The non-violent classes are: *Music*, *Speech*, *Others1*, and *Others2*. Classes Others1 and Others 2 are environmental sounds met in movies. Others1 contains environmental sounds of low energy and almost stable signal level (e.g. silence, background noise, etc). Others2 is populated by environmental sounds with abrupt signal changes, e.g. a door closing, thunders, etc. The *violent*-related classes are: *Shots*, *Fights* (beatings) and *Screams*.

2.2 Audio Feature Extraction

12 audio features are extracted for each segment on a short-term basis. In particular, each segment is broken into a sequence of non-overlapping short-term

windows (frames). For each frame 12 feature values are calculated. This process leads to 12 feature sequences, for the whole segment. In the sequel, a *statistic* (e.g. standard deviation, or average value) is calculated for each sequence, leading to a 12-D feature vector for each audio segment. The features, the statistics and the window lengths adopted are presented in Table 1. For more detailed descriptions of those features, please refer to [10].

Table 1. Window sizes and statistics for each of the adopted features

	Feature	**Statistic**	**Window (msecs)**
1	Spectrogram	σ^2	20
2	Chroma 1	μ	100
3	Chroma 2	*median*	20 (mid term:200)
4	Energy Entropy	*max*	20
5	MFCC 2	σ^2	20
6	MFCC 1	*max*	20
7	ZCR	μ	20
8	Sp. RollOff	*median*	20
9	Zero Pitch Ratio	–	20
10	MFCC 1	max/μ	20
11	Spectrogram	*max*	20
12	MFCC 3	*median*	20

The selection of the particular audio features, and of the respective statistics, was a result of extensive experimentation. Though, most of the adopted features have a physical meaning for the task of classifying an audio sample in the particular seven classes. For example, in Figure 1 an example of an Energy Entropy sequence is presented for an audio stream that contains: classical music, gunshots, speech and punk-rock music. Also, the maximum value and the $\frac{\sigma^2}{\mu}$ ratio statistics are presented. It can be seen that the maximum value of the energy entropy sequence is higher for gunshots and speech. This is something expected, since the energy entropy feature ([10]) has higher values for audio signals with abrupt energy changes (such as gunshots).

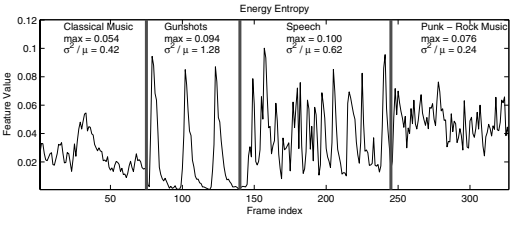

Fig. 1. Example of Energy Entropy sequence

2.3 Class Probability Estimation

In order to achieve multi-class classification, the "One-vs-All" (OVA) classification scheme has been adopted. This method is based on decomposing the K-class classification problem into K binary sub-problems ([11]). In particular, K binary classifiers are used, each one trained to distinguish the samples of a single class from the samples in the *remaining* classes. In the current work, we have chosen to use Bayesian Networks (BNs) for building these binary classifiers. At a first step, the 12 feature values described in section 2.2, are grouped into three 4D separate feature vectors (feature sub-spaces). In the sequel, for each one of the 7 binary sub-problems, three k-Nearest Neighbor classifiers are trained on the respective feature sub-space. This process leads to three binary decisions for each binary classification problem. Thus, a 7x3 kNN decision matrix R is computed. $R_{i,j}$ is 1 if the input sample is classified in class i, given the j-th feature sub-vector, and it is equal to 0 if the sample is classified in class "not i".

In order to decide to which class the input sample will be classified, according to R, BNs have been adopted: each binary subproblem has been modelled via a BN which combines the individual kNN decisions to produce the final decision. In order to classify the input sample to a specific class, the kNN binary decisions of each subproblem (i.e. the rows of matrix R) are fed as input to a separate BN $i, i = 1, \ldots, 7$, which produces the following probabilistic measure for each class, for each input each sample k: $P_i(k) = P(Y_i(k) = 1 | R_{i,1}^{(k)}, R_{i,2}^{(k)}, R_{i,3}^{(k)})$. This is the probability that the input sample's true class label is i, given the results of the individual kNN classifiers. After the probabilities $P_i(k)$, $i = 1, \ldots, 7$ are calculated for all binary subproblems, the input sample k is classified to the class with the largest probability, i.e.: $WinnerClass(k) = \arg\max_i P_i(k)$. This combination scheme can used as a classifier, though, in this work we use it as a probability estimator for each one of the seven classes. For more details on this classification scheme, please refer to [10].

3 Visual Classifier

The problem of violence detection in videos is challenging because of the big variability of the violent scenes, and the unconstrained camera motion. It is impossible to accurately model the scene and the objects within. Instead, we define classes that represent the amount of human activity in the scene and use features that can discriminate the video segments between these classes. The amount of activity is strongly correlated with the existence of violence as can be seen from Figure 2.

3.1 Video Class Definition

For the video based characterization we define three activity classes. These classes are defined by the amount of human activity in the scene as no-normal-high activity. The first class contains scenes that do not show humans or that

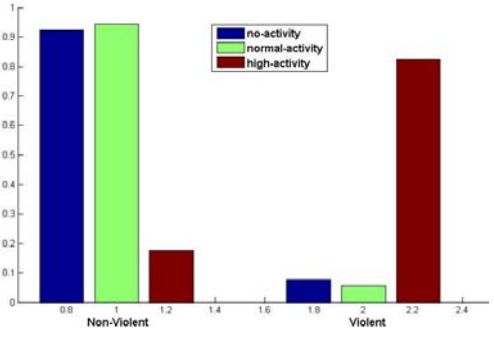

Fig. 2. Correlation between Violence and Activity classes: The first set of bars represents the percentage of segments with no, normal, and high activity which are labeled as non-violent whereas the second set contains the violent. As can be seen most of the segments that contain high activity are violent and vice-versa. The plots are derived from a randomly selected dataset of hand-labeled movie segments.

show humans that are completely inactive. The second class contains scenes in which one or more persons perform an activity that does not include erratic motion (e.g. walking, joking, chatting). The third class which is strongly correlated with violence contains scenes with people with erratic motion (e.g. fighting, falling).

3.2 Video Features

The used features can be split in two categories: the first contains simple motion related features, the second contains higher level features which originate from detecting semantic visual objects in the scene. The motivation for the motion features is that most scenes with high activity contain fast erratic motion. Additionally, the detectors are used to determine the presence and an estimation of the trajectories of people in the scene. The visual signal is split into 1-seconds mid-term segments. Each segment is comprised of several video frames, and the visual features are calculated on every frame. We average over the values of each frame to derive the value of the feature which characterizes the whole segment.

Motion Features

– *Average Motion (AM)*: This is the average motion of the frame. The frame is split in blocks of which we calculate the motion vectors using the previous frame as reference. The feature is derived by averaging the motion vector's lengths. The motion vectors lengths and the block size are determined as fractions of the frame size so that the feature will be invariant to frame scale changes. The feature is defined by: $AM = \frac{1}{N_b} \sum_{i=1}^{N_b} v_i$, where N_b is the number of blocks and v_i is the length of the motion vector of the i-th block.

– *Motion Orientation Variance (MOV)*: The feature measures the variance of the motion vectors orientations. The mean orientation is derived by calculating the mean motion vector. Then the variance is calculated by the following: $MOV = \frac{1}{N_b}\sum_{i=1}^{N_b} d_a^2(a_i, a_\mu)$, where a_i is the orientation of the i-th motion vector, a_μ is the orientation of the mean motion vector and $d_a()$ denotes the difference between the two orientations in $(-\pi, \pi)$.

Detection Features. Using an object recognition method we detect the presence of people or faces in the scene. The detected visual objects are tracked to establish the correspondences between them in consecutive frames. The tracked objects are used to derive a feature that measures their motion in the scene.

We used the object detection algorithm of [12] which exploits shape information to detect people in the scene. The detector uses a set of haar-like features and a boosted classifier. We recognize frontal faces, profile faces, full body and, upper body. The algorithm results in bounding boxes containing the detected objects in each frame.

The detector runs in standard intervals and each time the produced bounding boxes are used to initialize the trackers. When the next interval starts we measure the degree of overlap between a tracked object and each detected object of the same type (e.g. face). If there is a significant overlap then the detected object is linked to the tracked object. This way the final output of the system is a set of object trajectories from the frame where first detected up to the frame were it has been lost by the tracker or the position of the tracker does not match with any new detection. The tracking algorithm used is the one described in [13]. The algorithm uses the spot and blob tracking models.

The tracked objects are used to derive a metric for their motion. The metric is calculated as the average degree of overlap over all the tracked objects between two consecutive frames. The degree of overlap for a tracked object is defined as:

$$OTD = \frac{1}{2}\left[\frac{I^t(b) - I^t_{int}(b)}{I^t(b)} + \frac{I^{t-1}(b) - I^t_{int}(b)}{I^{t-1}(b)}\right] \quad (1)$$

where: $I^t(b)$ is the number of pixels of the b-th tracked object's bounding box at time t, and $I^t_{int}(b)$ is the number of pixels of the intersection between the bounding boxes of the b-th tracked object at times t and $t-1$.

3.3 Video Class Probability Estimation

The classification of the mid-term segments in the activity classes is performed using a weighted kNN classifier using the three features described in Section 3.2. The classifier is trained using a dataset of hand-labelled scenes containing no, normal, or high human activity. As described in Section 4.1, the individual audio and visual decisions are fused, using a meta-classifier. Therefore, it was necessary to use the weighted kNN algorithm as a *class probability estimator*. In particular, the probabilities of the "normal activity" and the "high activity" classes are estimated, using the weighted kNN algorithm, for each segment.

4 Fusion Approaches towards Violence Detection

4.1 Multi-modal Fusion

The seven audio class probabilities described in Section 2, along with the two visual-based class probabilities, described in Section 3, are combined in order to extract a final binary decision. This process is executed on a mid-term basis, i.e. in a sequence of successive segments from the original stream. In particular, *for each mid-term window of* 1 *sec length, a* 10*-D feature vector is created* with elements the seven audio probabilities, described in Section 2.3, the label of the winner audio class and the two visual-based classification decisions.

The combined 10-D feature vector is used by a k-Nearest Neighbor classifier, which extracts the final binary (violence vs non-violence) decision, for the respective mid-term segment. The same process is repeated for all 1 sec segments of the whole video stream. In Figure 3 a scheme of this process is presented.

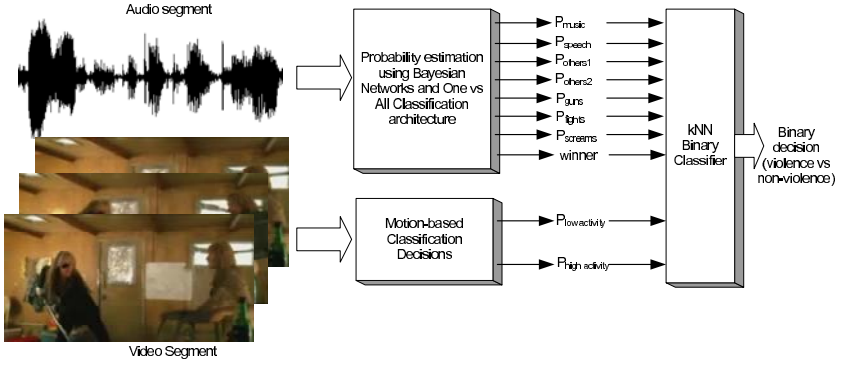

Fig. 3. Multi-modal fusion process

For comparison purposes, apart from the fused classifier, two individual kNN classifiers, an audio-based and a visual-based, have been trained, in order to distinguish between violence and non-violence, on a mid-term basis. In other words, these two individual classifiers have been trained on the 8D feature sub-space (audio-related) and on the 2D feature sub-space (visual-related) respectively. In Figure 4, an example of the violence detection algorithm is presented, when using a) only audio features b) only visual features and c) the fused feature vector. The gray line corresponds to the true class labels over time. Also, for each case, the precision and recall rates are presented. It is obvious that the fused approach performs significantly better than both audio and visual based approaches.

5 Experimental Evaluation

5.1 Scenario and Setup

For training and evaluation purposes, 50 videos have been ripped from 10 different films. The overall duration of the data is 2.5 hours. The video streams

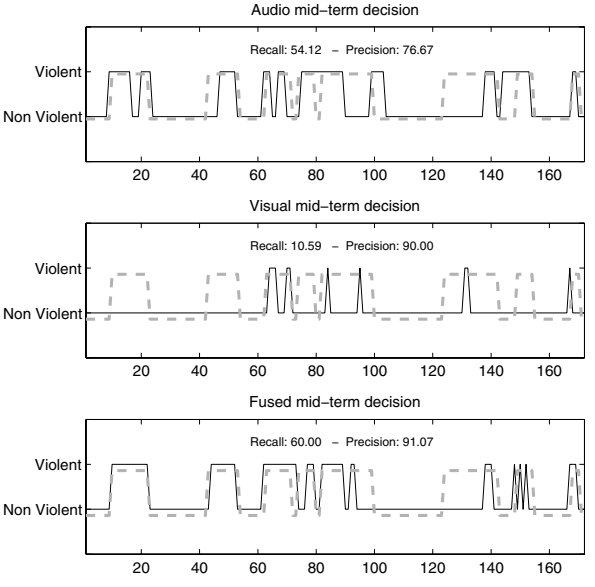

Fig. 4. Violence detection example for a movie audio stream

have been manually annotated by three humans. In particular, the humans annotated the parts of the videos that contained violent content and the results of this annotation have been used as ground truth for training and evaluating the proposed methods. According to the manual annotations 19.4% of the data was of violent content. Almost 9000 mid-term segments were, in total, available for training and testing. Though, the evaluation (as described in the following Section) was carried out on a video stream basis.

5.2 Classification and Detection Results

In this Section the results of the proposed binary classification method are presented. The performance of the fused classification process is compared to the individual performances, if only the audio and the visual features were used. In all three methods, the "Leave One Out" evaluation method has been used, on a video file basis, i.e., in each cross-validation loop, the mid-term segments of a single video file have been used for evaluation, while the rest of the data has been used for training purposes. The following types of performance measures have been computed:

1. Classification Precision (P): This is the proportion of mid-term segments that have been classified as violence and were indeed violence.
2. Classification Recall (R): This is the proportion of mid-term violent segments that were finally classified as violent.

3. Classification F_1 measure.
4. Detection Precision (Pd): This is the number of detected violent segments, that were indeed violence, divided by the total number of detected violent segments.
5. Detection Recall (Rd): This is the number of correctly detected violent segments divided by the total number of **true** violent segments.
6. Detection Fd_1 measure.

Performance measures P, R and F_1 are associated to the **classification** performance of the algorithm on a mid-term (1-second) basis, while the measures Pd, Rd and Fd_1 are related to the **event detection** performance of the algorithm. Note that a violent segment is correctly detected if it overlaps with a true violent segment. In addition, for comparison purposes, we have computed the same performance measures for the random mid-term classifier. The performance results are displayed in Table 2.

Table 2. Classification and Detection Performance Measures

	Classification Performance Measures		
	Recall	Precision	F_1
Audio-based classification	63.2%	45.2%	52.7%
Visual-based classification	65.1%	40.7%	50.1%
Random classification	19 %	50%	28%
Fused classification	60.1%	47%	52.8%

	Detection Performance Measures		
	Recall	Precision	F_1
Audio-based detection	82.9%	38.9%	53%
Visual-based detection	75.6%	34%	46.9%
Fused detection	83%	45.2%	58.5%

6 Conclusions

We have presented a method for detecting violence in video streams from movies. Both audio and visual based classes have been defined, and respective soft-output classifiers have been trained. Then, a simple meta-classifier has been adopted, in order to solve the binary classification task: Violence Vs Non Violence. Experimentation has been carried out on a real film dataset. Experiments indicated that audio classification and detection was respectively 1.6% and 6.1% better than the visual-based method. Furthermore, the fused meta - classification achieved a **boosting** at the overall performance compared to the best individual method (i.e., the audio-based method). Finally, the overall **event detection** performance indicated that only 17% of the violent events are not detected, while almost 1 out of 2 detected events are indeed violent ones.

Acknowledgment

This work has been supported by the Greek Secreteriat for Research and Technology, in the framework of the PENED program, grant number TP698.

References

1. Datta, A., Shah, M., da Vitoria Lobo, N.: Person-on-person violence detection in video data. In: ICPR, vol. 1, pp. 433–438 (2002)
2. Zajdel, W., Krijnders, J., Andringa, T., Gavrila, D.: Cassandra: audio-video sensor fusion for aggression detection, pp. 200–205 (2007)
3. Vasconcelos, N., Lippman, A.: Towards semantically meaningful feature spaces for the characterization of video content. In: ICIP 1997: Proceedings of the 1997 International Conference on Image Processing (ICIP 1997), Washington, DC, USA, 3-Volume Set-Volume 1, p. 25. IEEE Computer Society, Los Alamitos (1997)
4. Nam, J., Alghoniemy, M., Tewfik, A.H.: Audio-visual content-based violent scene characterization. In: ICIP(1), pp. 353–357 (1998)
5. Vasconcelos, N., Lippman, A.: Towards semantically meaningful feature spaces for the characterization of video content. In: International Conference on Image Processing, pp. 25–28 (1997)
6. Datta, A., Shah, M., Lobo, N.V.: Person-on-person violence detection in video data. In: IEEE International Conference on Pattern Recognition, Canada (2002)
7. Nam, J., Tewfik, A.H.: Event-driven video abstraction and visualization. Multimedia Tools Appl. 16(1-2), 55–77 (2002)
8. Rasheed, Z., Shah, M.: Movie genre classification by exploiting audio-visual features of previews. In: Proceedings 16th International Conference on Pattern Recognition, pp. 1086–1089 (2002)
9. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006)
10. Giannakopoulos, T., Pikrakis, A., Theodoridis, S.: A multi-class audio classification method with respect to violent content in movies, using bayesian networks. In: IEEE International Workshop on Multimedia Signal Processing, MMSP 2007 (2007)
11. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141 (2004)
12. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: ICIP(1), pp. 900–903 (2002)
13. Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Hierarchical feature fusion for visual tracking. In: IEEE International Conference on Image Processing 2007, September 16 -October 19, vol. 6, pp. VI –289–VI –292 (2007)