Violence Content Classification Using Audio Features

Theodoros Giannakopoulos [1], Dimitrios Kosmopoulos [2],
Andreas Aristidou [1], and Sergios Theodoridis [1]

[1] Department of Informatics and Telecommunications, National and Kapodistrian
University of Athens, Panepistimiopolis, Ilissia Athens 15784
{tyiannak, stud01144, stheodor}@di.uoa.gr
[2] Institute of Informatics and Telecommunications, National Center for Scientific Research
"Demokritos", Agia Paraskevi, Athens 15310, Greece
dkosmo@iit.demokritos.gr

Abstract. This work studies the problem of violence detection in audio data, which can be used for automated content rating. We employ some popular frame-level audio features both from the time and frequency domain. Afterwards, several statistics of the calculated feature sequences are fed as input to a Support Vector Machine classifier, which decides about the segment content with respect to violence. The presented experimental results verify the validity of the approach and exhibit a better performance than the other known approaches.

1 Introduction

In the following years a huge increase of the available multimedia content is expected. Almost everyone will be able to provide content, accessible by large portions of the population with limited central control. However, the increasing use of the related technology by sensitive social groups creates the need for protection from harmful content. The goal of the work presented here is part of a multimodal approach, which aims to contribute to the automated characterization of multimedia content with respect to *violence*. This approach will make possible for content providers to rate automatically their content and for the end user to filter the violent scenes in client terminal devices.

The violence characterization is quite subjective and this creates difficulties in defining violent content unambiguously. As violence we may define any situation or action that may cause physical or mental harm to one or more persons. Violent scenes in video documents regard the content that includes such actions. Such scenes are usually manifested through characteristic audio signals (e.g., screams, gunshots etc). For the specific problem the related literature is very limited and in most cases it examines only visual features such as in [1], and [2]. Audio data for violent detection is used as an additional feature to visual data in [3], where abrupt changes in energy level of the audio signal are detected using the *energy entropy* criterion.

We deduce from existing literature that although the audio is a very useful source of information, much simpler to process than video and in most cases self-sufficient for violent scene characterization, it has been rather overlooked. The use of additional

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 502 – 507, 2006.

features as well as better classification methods is able to provide much better results and this is what we do in this work.

2 Audio Features

We extract six segment-level audio features, which will be used at a next stage by a Support Vector Machine classifier. For the feature calculation we assume that the signal $\boldsymbol{x}$ has already been segmented into semantically coherent *segments* (scenes). The segments are divided into *W time-windows* (*frames*) of predefined duration *S*, and for each one of them we calculate the frame-level features. Therefore, for each audio segment, six *feature sequences* of length *W* are calculated. In order to extract semantic content information it is necessary to follow how those sequences change from frame to frame. To quantify this variation, a number of statistics (e.g., mean value) have been calculated, for each feature sequence. In this paper, we use six popular frame-level features extracted both from the time and frequency domain ([4]). Afterwards, 8 statistics $f_1,\ldots,f_8$ have been calculated from the feature sequences, as described in the following. Those statistics are then used by the classifier as single-feature values of each audio segment.

2.1 Time-Domain Features

The *energy entropy* expresses abrupt changes in the energy level of the audio signal. In order to calculate this feature, the frames are further divided into *K* sub-windows of fixed duration. For each sub-window *i*, the normalized energy σ_i^2 is calculated, i.e., the sub-window's energy divided by the whole window's energy. Then, the energy entropy is computed for frame *j* using eq. (1). The value of the energy entropy is small for frames with large changes in energy level. Therefore, we can detect many violent actions like shots, which are characterized by sudden energy transitions in a short time period. This becomes obvious in figures 1 and 2, where the energy entropy sequences of two audio signals with violent content are presented. The features f_1 and f_2 that we use based on the energy entropy are the ratios of maximum to mean and maximum to median value of the energy entropy (eq. (2) and (3)).

Another feature stems from *the absolute value of the signal amplitude* ($A_{0\mathrm{i}}$ for each sample *i*). We use here as feature f_3 the ratio of the max to the mean absolute signal amplitude (4).

Short time energy N_j is another widely used feature in audio classification applications (*eq. (5)*). In the current project, we use the mean and variance of the short time energy (eq. (6) and (7)).

Zero crossing rate ([4]) is one of the most widely used time-domain audio features. It is calculated by the number of time-domain zero-crossings, divided by the number of samples in the frame, as presented in eq. (8) (*sgn* is the signum function). The statistic used for ZCR is the ratio of maximum to mean value (eq. (9)). In figure 3, the ZCR sequences for three different audio segments containing violence (gunshots), music and speech are presented. It is obvious that the plot justifies the selection of f_6.

Table 1. The features extracted from the employed criteria

Feature Name	Feature Equation	Statistics (Single Features)
Energy Entropy	$I_j = -\sum_{i=1..K} \sigma_i^2 \log_2 \sigma_i^2$ (1)	$f_1 = \max_{j=1..W}(I_j) \Big/ \left(\frac{1}{W}\sum_{j=1..W} I_j\right)$(2) $f_2 = \frac{\max_{j=1..W}(I_j)}{\mathit{median}_{j=1..W}(I_j)}$(3)
Signal Ampl.		$f_3 = \max_{i=1..L}(A_{0i}) \Big/ \left(\frac{1}{L}\sum_{i=1..L} A_{0i}\right)$(4) *(L: signal length)*
Short Time Energy	$N_j = \sum_{i=1..S} x_i^2$ (5)	$f_4 = \frac{1}{W}\sum_{j=1..W} N_j$(6) $f_5 = \frac{1}{W}\sum_{j=1..W}(N_j - f_4)^2$(7)
Zero Crossing Rate	$Z_j = \frac{1}{2S}\sum_{i=1..S} \lvert \mathrm{sgn}(x_i) - \mathrm{sgn}(x_{i-1}) \rvert$(8)	$f_6 = \max_{j=1..W} Z_j \Big/ \left(\frac{1}{W}\sum_{j=1..W} Z_j\right)$(9)
Spectral Flux	$F_j = \sum_{k=0..S-1}(N_{j,k} - N_{j-1,k})^2$(10)	$f_7 = \max_{j=1..W} F_j \Big/ \left(\frac{1}{W}\sum_{j=1..W} F_j\right)$(11)
Spectral Rolloff	$\sum_{k=0}^{m_c^R(j)} \lvert X_{jk} \rvert = \frac{c}{100}\sum_{k=0}^{S-1} \lvert X_{jk} \rvert$(12)	$f_8 = \max_{j=1..W} m_c^R(j) \Big/ \left(\frac{1}{W}\sum_{j=1..W} m_c^R(j)\right)$ (13)

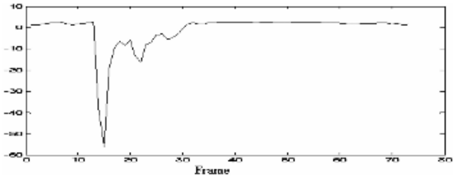

Fig. 1. Energy entropy of a shot

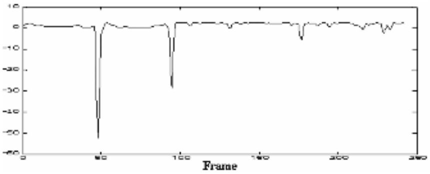

Fig. 2. Energy entropy of beatings

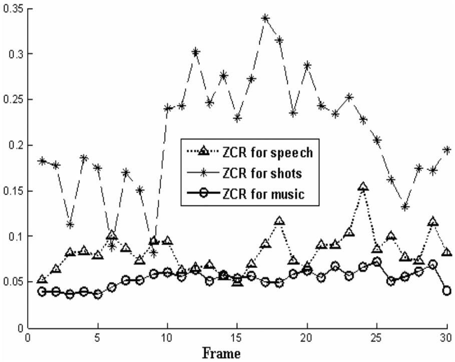

Fig. 3. ZCR for gunshots, music and speech

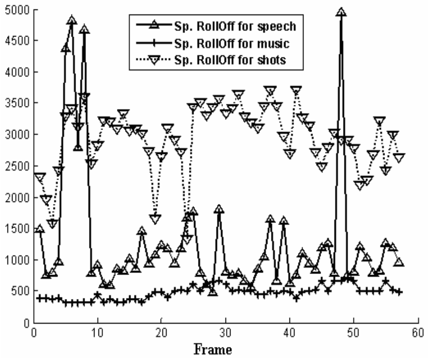

Fig. 4. Spectral rolloff for gunshots, music and speech

2.2 Frequency-Domain Features

Spectral flux is a frequency-domain measure of the local spectral change between successive frames, and it is defined as in equation (10). The related feature is the ratio of maximum to mean of the spectral flux, as presented in (11). $N_{j,k}$ is the spectral energy of the j-th frame for the k-th sample.

This frequency domain feature is called *spectral rolloff*, and it is defined as the frequency bin $m_c^R(j)$ below which the c percent (e.g., c=90) of the magnitude distribution of the Discrete Fourier Transform (DFT) X_k coefficients is concentrated for frame j (eq. (12)). It is a measure of skewness of the spectral shape, with brighter sounds resulting in higher values. The related statistic is the ratio of max to mean of the spectral rolloff (eq. 13). In fig. 4 an example of spectral rolloff sequences of three different audio segments (shots, music and speech) is presented. These differ in their mean values, but also in the way their peaks are distributed over time.

3 Classification

For the classification task we used the Support Vector Machine Classifier (SVM), which is known for its computational efficiency and effectiveness even for high dimensional spaces for classes that are not linearly separable ([5]). The classification task is twofold: (a) training, and (b) testing. During training we provided the normalized features extracted from audio segments to the classifier, along with labels Y indicating if the features correspond to violent (+1) or non-violent content (-1). The SVM gave as output a number of support vectors, which were used for the classification task. We have applied the linear, the Gaussian radial basis, the polynomial and the sigmoid hyperbolic tangent kernel functions for classification purpose with varying the C parameter, which expresses the trade-off between training error and margin.

4 Experimental Results

A database of audio segments, extracted from several movie genres, was created for training/testing purposes. The total duration of the samples was 20 minutes: 50% of that data was used for training, and the remaining 50% for testing. The sampling rate was 16 KHz and the sample resolution 16 bits (1 channel was used). The violent audio segments were extracted from scenes such as shots, explosions, fights and screams, while non-violent audio segments of music and speech were also extracted, along with non-violent sounds like fireworks, which have similar characteristics to violent sounds. The signal was divided into frames using Hamming windows of 400 msec, with a step of 200 msec (50% overlapping). All audio features described in section 2 were used. It has been experimentally found that the testing set was classified with minimum error rate when a polynomial kernel function had been used in the SVM classifier.

In table 2, we present the classification error rates of each individual classifier (feature), along with the error rates when all features are used (8-D space). Apart from the average error, the false negative and false positive errors are presented. It is obvious than no individual feature has overall performance better than 80%. Though, the error rate itself is not a criterion for deciding which of the features are better when

Table 2. Classification error rate for each feature

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	**8-D**
% False Negative (FN)	9	12.5	13.5	12	9	11	10	12.5	**4.75**
% False Positive (FP)	13.5	13	15	17	12.5	12.5	10.5	12	**9.75**
Overall % Error	22.5	25.5	28.5	29	21.5	23.5	20.5	24.5	**14.5**

used in combination with the others. For this purpose, feature selection techniques may be used. On the other hand, the SVM has a classification error rate of 14.5% when all features are used.

For the 8-D case, the following measures were extracted (see Table 2): (a) *Recall (R):* The proportion of the violent segments, which have been correctly identified as violent, (b) *Precision (P):* The proportion of the data classified as violent, whose true class label was indeed violent and (c) *Accuracy (A):* The overall proportion of data classified correctly. The measures R, P, A are given by: $R=TP/(TP+FN)$, $P=TP/(TP+FP)$ $A= (TP+TN)/(TP+TN+FP+FN)$, where TP is the true positive rate and TN is the true negative rate. In our case, the two classes shared the same probability, i.e., the number of violent and non-violent test segments are the same. Therefore: $TP+FN=TN+FP=0.5$. Combining the results of Table 2, the definitions for R, P, A and the last equation, we found that recall was equal to 90.5%, precision 82.4% and accuracy 85.5%.

5 Conclusions and Future Work

In this work, we have used some popular audio features for detecting violence in audio segments with an SVM classifier. The whole system was tested using audio segments extracted from real movies and the results were promising. In average, 85.5% of the audio data was classified correctly, while the percentage of the violent segments that where correctly identified (recall) was 90.5%. This is a significantly better rate than that provided by any individual features, some of which have been used separately in the past for the same task.

In the future, other audio features (e.g., Mel-frequency cepstral coefficients - MFCCs) could be added to those presented in this paper. Furthermore, techniques for feature selection could be used, to find an optimal sub-set of features for the specific classification problem. Also, other classification algorithms can be employed like Hidden Markov Models (HMMs). Another direction of research could be the usage of multi-class recognition algorithms, for characterizing the audio segments as shots, screams, etc, instead of facing the problem as binary. On the other hand, it is needed to implement an audio *segmentation* algorithm, for dividing large audio streams into shorter homogenous segments. The computed segments will be fed as input to the feature calculation stage. Finally, is obvious, that the results of any audio violence detection system can be combined with synchronized visual cues.

References

1. Vasconcelos, N.; Lippman, A., *Towards semantically meaningful feature spaces for the characterization of video content*, Proc. International Conference on Image Processing, 1997., Volume 1, Oct 1997, Pages: 25 - 28 vol.1
2. A. Datta, M. Shah, N. V. Lobo, "Person-on-Person Violence Detection in Video Data", *IEEE International Conference on Pattern Recognition*, Canada, 2002.
3. J. Nam, A.H. Tewfik, “Event-driven video abstraction and visualisation”, Multimedia Tools and Applications, 16(1-2), 55-77, 2002
4. Sergios Theodoridis, Konstantinos Koutroumbas, *Pattern Recognition*. Academic Press, 2005, 3rd Edtition.
5. N. Cristianini and J. Shawe-Taylor, "Support Vector Machines and other kernel-based learning methods", Cambridge University Press, ISBN 0-521-78019-5, 2000.