Audio Classification for Scene Change Detection in Video Sequences*

C. Saraceno & R. Leonardi
Signals & Communications Lab., Dept. of Electronics for Automation,
University of Brescia, Brescia I-25123, Italy
E-mail: {saraceno,leon}@ing.unibs.it

ABSTRACT

The organization of video databases according to the semantic content of data, is a key point in multimedia technologies. In fact, this would allow algorithms such as indexing and retrieval to work more efficiently.

The segmentation of a video sequence into scenes and the characterization of each scene has been suggested as a technique for organizing video information. Typically, this is performed by analyzing the video signal only. Human beings, on the other hand, use both their visual and auditory systems to perceive the semantics of a film. The associated audio signal can thus be useful to extract information which cannot be simply derived from the sole analysis of the video signal or at least to make the video processing more robust.

In this work a technique which uses audio together with video information is proposed in order to improve the performance of existing scene change detection algorithms. Tests are performed on material coming from advertisement and movie sequences.

1 INTRODUCTION

In order to generate indices that can describe a video sequence in terms of objects and their interaction, attempts to characterize a sequence according to the video content of data have been proposed in the literature [6]. It is well known in the image processing society how difficult is the identification and the characterization of objects in a video sequence as well as their tracking over time. This is mainly due to the fact that no robust algorithm has been found for this purpose. Therefore, a simpler way to operate has been to determine the editing features such as cuts, fades and dissolves that have been used in the composition process of the video material. Attempts to extract these characteristics have been performed in the past on compressed [5] or uncompressed [4] material and using only video information. Efforts have been devoted in segmenting the video in shots[1] and for each shot assigning a representative frame, etc.

Video shots, however, are usually inefficient to convey semantic information. Depending on the type of video material, shots can last 3-4 seconds or less. For advertisements, usually, a shot lasts roughly 2 seconds; in the case of movies, it depends on the director, but rarely it lasts longer than a minute, and on the average about 10 sec. Further, a shot does not always coincide with a finite action of the movie; usually consecutive shots are semantically correlated to each other. Based on these considerations, it is clear that shots cannot be used alone as a successful base for extracting semantic features from a video sequence. For this reason, efforts have been recently devoted in trying to segment video through a more accurate analysis and characterization of the image sequence. In Yeung and Yeo [7], for example, three major models of temporal events were defined: dialogues, action and story units. A *time-constrained clustering* was performed for automatic labeling of shots. Based on this clustering the video sequence is analyzed for the extraction of video segments which fit one of the three models.

Only recently, the idea of jointly combine audio and visual information for indexing and retrieval purposes has become subject of research. In other words, audio is used to retrieve information that cannot be obtained by the video processing, such as presence of music, number of speakers, ... Audio can also be used to confirm the results of a classification obtained by a previous video analysis. For example, Patel and Sethi [8] have examined the potential of speaker identification techniques for characterizing video clips in terms of actors present in the scene. In other words, assuming to know the major actors present in the video, the goal is to label all video clips with descriptors indicating for each one the presence of the different actors.

In our work, a possible general scheme for

*This work was partially funded by the Italian Ministry of the University, and of the Scientific and Technological Research.

[1] A shot is define as the interval between the beginning and the end of a camera record.

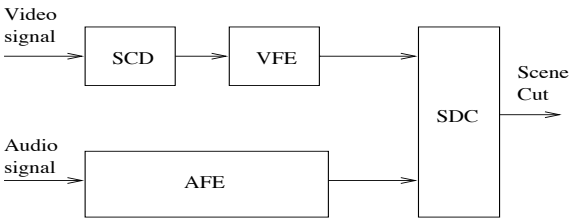

Figure 1: General scheme for audio-visual scene change detection.

AFE : Audio Feature Extractor
SCD : Shot Cut Detector
VFE : Video Feature Extractor
SDC : Scene Detection and Characterization.

jointly using audio and video information for scene change detection and characterization is proposed. An audio classification is suggested and scene change detection is performed using the audio classification together with a shot cut detection module. The number of audio classes has been kept small for simplicity and applicability. Simulation results are very promising especially considering the wide range of multimedia material that has been tested.

In the next section, the general scheme to combine audio and video information is explained. This section introduces also the simple audio model being considered. Section 2 starts by explaining the proposed audio classification procedure; it ends demonstrating its performance on various type of simulation material. In section 3, the algorithm to identify coherent audio-visual scenes using jointly video and audio processing units is described. Finally, the performance of the proposed "semantic" characterization of audio-visual material are shown in section 4.

2 COMBINING AUDIO-VISUAL INFORMATION

The goal is to segment video sequences according to the semantic content of data. The segmentation can be performed in two stages by identifying first the shot cuts and then trying to merge successive correlated shots.

For this purpose, we define a "scene" as a set of one or more consecutive shots which are "semantically" correlated. According to this definition of a "scene", the goal is to show that better scene change detection can be obtained by jointly using audio and video information.

A possible general scheme for jointly using audio and video information for scene change detection and characterization is proposed in Fig. 1 On one side the video signal is first analyzed to detect shot cut (SCD), then different characteristic features of the shot (such as number of moving objects present in the shot, representative frames of the shot etc.,) are extracted (VFE).

On the other side, the audio file is processed in order to extract acoustic or meaningful features for interpretation purposes such as presence of speech or music, number of speaker etc. (AFE). Both audio and video processing unit outputs are passed to a unique module (SDC) which tries to combine the corresponding information in order to merge shots having similar characteristics. For example, a decision rule could be: when objects identified in one shot are also present in the consecutive shot and the two shots exhibit the same type of music, these will most likely belong to the same scene; thus they can be grouped together.

In what follows, the focus is placed in the characterization of the audio component. The literature has been placing a lot of attention on the segmentation of video material into shots but little research has been carried out so far in separated audio signals into different classes.

An easy way to extract relatively abstract features from an audio file is to simply group first consecutive audio frames of 2^p samples (typically p ranges from 9 to 11) in segments that are of a given class. That is consecutive frames are merged into one segment as long as they are all recognized as belonging to the same class, e.g. speech. Then each segment is analyzed in order to extract other features which are specific to that particular audio class. For example, when speech segments have been identified, a subsequent process can try to recognize a specific speaker.

For the first type of audio processing that allows to separate an audio file into a series of consecutive segments of one class, we assume the audio signal be represented as a linear combination of 4 types of fundamental signals: **Silence** ($s[n]$), **Speech** ($v[n]$), **Music** ($m[n]$) and **Noise** ($q[n]$), i.e.

$$a[n] = \alpha_s s[n] + \alpha_v v[n] + \alpha_m m[n] + \alpha_q q[n] \quad (1)$$

Further, the assumptions for each signal category are the following:

- **Silence** segments define those audio frames which only contain a quasi-stationary background noise, with a low energy content with respect to surrounding signals belonging to other classes.
- **Speech** segments contain voiced, unvoiced and plosive sounds [1].
- **Music** segments contain a combination of sounds with peculiar characteristics of periodicity.
- **Noise** segments correspond to all other categories, i.e. everything which was not classified as one of the previous classes.

For simplicity, a segment will be classified in only one of the previous categories, regardless if more

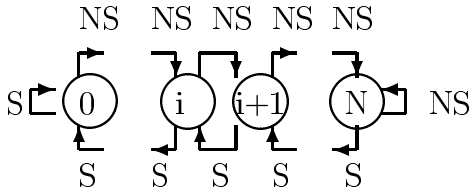

Figure 2: Finite State Machine

than one class is simultaneously present in the original signal. This choice is not as restrictive as it appears as we are interested on extracting meaningful semantic information for audio-visual scene change detection (as it will be explained later). This allows to place a certain prioritization in the classification process. This may seem restrictive, but in many cases it may turn out sufficient for a high level analysis of audio-visual data. For example, we believe that it is more relevant to recognize talking people rather than if they are talking on a silent or noisy environment. We thus simply establish a priority in the classification process, so that more relevant attributes are considered first. For example, we do not want to classify as "Noise" segments containing also speech. Accordingly, Voice will have the highest priority, followed by Music, then Noise and finally Silence.

3 AUDIO CLASSIFICATION PROCEDURE

Briefly, the audio classification is performed as follows: first, the algorithm processes the audio file in order to detect silence segments. Frames which are not classified as silence are further processed in order to discriminate voice from music. Those frames which correspond neither to silence nor to voice or music are finally labeled as noise.

3.1 Silence detection

Silence detection is performed with an algorithm based on energy information [2] which uses a Finite State Machine (FSM) (see Fig. 2). The algorithm does not require any a priori information on noise characteristics. An initial training must occur in order to evaluate the statistics of the background noise (local energy mean and its standard deviation). This is initially performed on the first few seconds of the audio signal (typically 0.4 sec.) assuming that the audio signal starts with silence and with audio frames of 512 samples (44.1kHz sampling rate, 16 bit uniform quantization). After the initialization procedure, the FSM is used to discriminate between silence and non silence. The FSM has N+1 states (typically N is equal to 8). State 0 corresponds to silence, state N corresponds to non silence while states between 1 to $N-1$ are intermediary states. The initial state is by hypothesis set to 0 (i.e., the audio signal starts in silence). Every time the energy value of a frame falls below $m + K * \sigma$, where m is the background noise energy mean, σ its standard deviation and K a constant (set to 0.4), there is a transition from state i to state $i-1$ till state 0 is reached. If the energy value is above the aforementioned threshold, there is a transition from state i to state $i+1$, till state N is reached. Frames which belong to intermediate states are not classified until one of the two states (0 or N) has been reached. These are then classified according to the reached state. In other words, the FSM uses past and future information, to reduce the possibility of wrong classifications. In fact, due to the stochastic nature of noise, there may be silence segments with high energy values (In this case, noise is not the surrounding non speech/music audio ($q[n]$ in equation (1)), rather it corresponds to the electronic or quantization noise level). On the other hand, voice or music segments may have low energy value. All this would result in a misclassification if the energy information was used without the FSM, and its statistics were not locally estimated. The more the states of the FSM, the less sensitive the classification results to transient energy fluctuation.

3.2 Speech and music characteristics

Everything which has not been classified as silence is further analyzed to identify speech and music segments.

Speech segments present the characteristics of being usually limited in frequency to about 8 to 10kHz and they usually alternate tonal ($\sim$periodic) to noise-like segments and high to low energies. On the other hand music segments present the characteristics of covering a broader range of frequencies with respect to speech: they can go up to 20kHz or more. However, their energy is usually concentrated at low frequencies. Further, they also present periodicity characteristics over short intervals of time, but their period can be larger than the period present on voiced segments [3]. Besides, music segments do not present as much tonal noise-like alternation as speech segments do.

3.3 Speech and music detectors

Based on these considerations, our classification procedure will use energy information, the zero crossing rate (ZCR), a short-time autocorrelation measure and a contextual majority based decision key. The music/speech discrimination algorithm is summarized in the block diagram of Fig.3. The autocorrelation is used to detect the periodicity of the signal over a short interval of time [1], the short-time average ZCR is used to measure the frequency content of the signal. The contextual majority based decision is used to reduce the probability of misclassification.

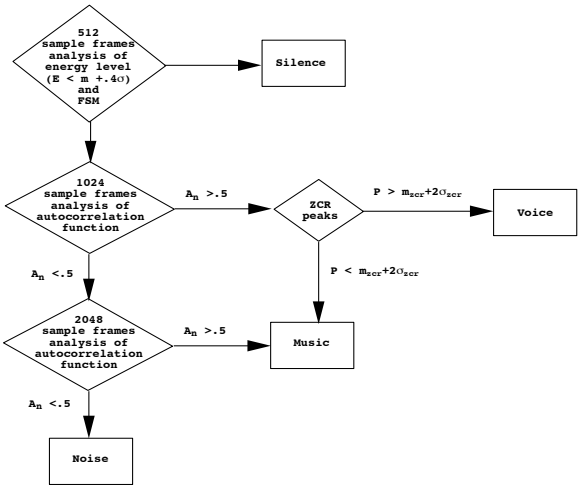

Figure 3: Block diagram of music/speech discriminator

Audio samples which do not belong to silence segments are grouped, when consecutive, into frames of M samples each (typically $M = 1024$, i.e. ~ 23 msec) and for each frame a short-time autocorrelation function is evaluated as follows:

$$R_n[k] = \sum_{i=-\infty}^{\infty} a[i]w[i-n]a[i+k]w[i-n+k] \tag{2}$$

where $w[n]$ is a rectangular window of length M

$$w[n] = \begin{cases} 1 & 0 \leq n \leq M-1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The short-time autocorrelation has been used to evaluate what we define as the periodicity content of an audio frame. For each function the second major peak $R_n[k]$ with

$$k = \text{argmax}_{i=i_0,i_0+1,..,M-1} \; R_n[i] \tag{4}$$

is detected (i_0 is typically set to the position of the first local minimum of $R_n[i]$). The ratio $A_n = R_n[k]/R_n[0]$ is then calculated. A_n provides a measure of the periodicity content of the processed frame. The higher the ratio A_n the higher the periodicity content of the frame. Frames having $A_n > 0.5$ are considered periodic. Speech frames having periodicity characteristics usually lead to high A_n values. Unfortunately also music frames can have high A_n values. To discriminate between music and speech segments, a short-time average ZCR is used at this point. The short-time average ZCR function is defined as:

$$ZCR[n] = \frac{1}{2M} \sum_{i=-\infty}^{\infty} |sgn(a[i]) - sgn(a[i-1])|.w[i-n] \tag{5}$$

where sgn(.) is the regular sign function.

In fact, ZCR values of music segments exhibits a smaller variation when compared to speech segments [3]. The ZCR is therefore evaluated on consecutive periodic frames and on frames which are not considered periodic ($A_n \leq 0.5$) but which are surrounded by periodic ones (in order to allow for a context majority based decision as indicated previously). The mean and the standard deviation of the ZCR is estimated over a period of 50 such consecutive frames or less (if the number of consecutive periodic frames is less than 50). If at least one of those frames has a ZCR value greater than its mean plus L times the standard deviation (with L typically equal to 2), the entire group of frames is labeled as a speech segment. It is labeled as music otherwise. Finally, all frames which have not been labeled so far are resized so as to contain 2048 samples each. The corresponding A_n's are then calculated and those with $A_n > 0.5$ are considered periodic and labeled as music again. Frames which are not periodic but which are surrounded by periodic frames are also labeled as music frames, to allow again for the context majority based decision. This new computation of A_n has been performed for those music frames which have a larger period than previously (a pecular characteristics of music, as indicated in the previous paragraph), and which could not be detected when using a rectangular window of 1024 samples for computing the short-time autocorrelation function defined in equation (2).

All frames which have not been labeled so far are finally classified as noise.

3.4 Performance of the audio classification

Simulations were carried out on:

A1 15 min. of audio containing silence and speech recorded in a silence environment;

A2 15 min. of audio containing classical music;

A3 10 min. of audio extracted from the "Pulp fiction" movie. All of them had a CD audio quality (44.1kHz sampling rate and 16 bits).

Results are shown in Tables 1 to 3. Performance of the proposed classification process are quite satisfactory. It should be noted that misclassifications in A1 and A2 mainly occurred at the boundary of different audio segments. Further, the singing segments of the A3 audio files were mostly recognized as speech. Further studies could lead to a more specific decision by explicitly trying to identify such peculiar mixture of music and voice.

4 INTEGRATION OF VIDEO AND AUDIO

On one hand, the audio signal is segmented using the procedure explained in the previous sec-

	Silence	Music	Speech	Noise
32% silence	95.5%	0%	4.5%	0%
0% music	0%	0%	0%	0%
68% speech	6%	0%	94%	0%
0% noise	0%	0%	0%	0%

Table 1: A1 audio classification results

	Silence	Music	Speech	Noise
10% silence	98%	2%	0%	0%
90% music	10%	85%	0%	5%
0% speech	0%	0%	0%	0%
0% noise	0%	0%	0%	0%

Table 2: A2 audio classification results

tion, on the other hand, the video signal can be segmented in shots using one of the techniques proposed in the literature [4, 5, 6]. In our work, shot cuts were identified by hand. This choice was dictated by the idea of having a 100% reliable shot cut detector. Possible errors of the shot cut detector might have affected the overall evaluation of our scene characterization procedure performance. The idea was to estimate the effective improvement that audio analysis and classification was bringing for scene identification.

Now, let us see how the audio classification together with the shot cut detection can be used to help detecting scene changes. We have noticed that the strategy and performance largely depends on the type video material, because of the specific composition strategies that are used for each type of audio-visual material.

In the case of video advertisements, it may be reasonable to consider each advertisement as a single scene. A scene change corresponds therefore simply to a change of advertisement. It was statistically noted that this one occurred in presence of a joint occurrence of a shot cut (video) and of a silence segment (audio). Therefore, a joint occurrence of silence and shot cut was used to recognize different scenes (i.e. different advertisement) with a high probability of being correct. In Fig. 4, a scene (advertisement) change occurs while in Fig. 5, there is only a shot cut: both audio signal and iconized video frames are shown to demonstrate the consideration made previously.

	Silence	Music	Speech	Noise
10% silence	96%	1%	3%	0%
25% music	6%	87%	2%	5%
37% speech	3%	0%	95.5%	1.5%
8% noise	0%	0%	0%	100%
20% M&S	0%	11%	82%	7%

Table 3: A3 classification - M&S: music & singing.

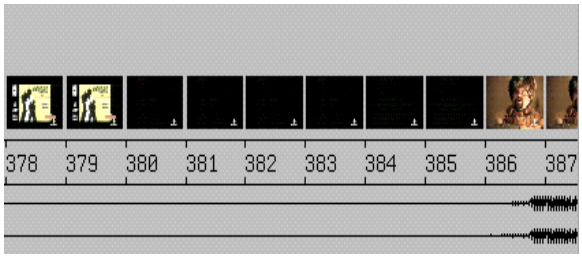

Figure 4: Advertisement change.

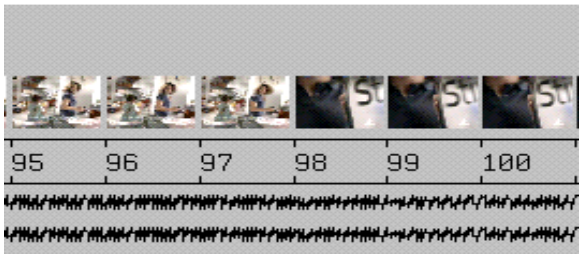

Figure 5: Cut inside the same advertisement

In the case of movies, TV broadcast news, TV shows or documentaries, the joint occurrence of shot cut and silence rarely indicates a change of topic, allowing poor discrimination of scenes. Reasonable performance to characterize movies can be reached instead by merging successive shots exhibiting the same type of audio signal, on both sides of the shot cut. To be more specific, assuming that the sequence of images (video) has been decomposed in a sequence of shots S_i, identifying different camera records, the following strategy was used:

- when S_i and S_{i+1} exhibit the same type of audio signal, they are merged to form a single scene; however, it was noticed that this correspondence must occur for some period of time. When two adjacent shots contain different classes of audio, they are not merged if the the same audio signal class occurs across the shot boundary between the two shots only for a few seconds (up to 5 seconds), and then changes.

- when S_i and S_{i+1} exhibit the same class of audio signal except at the boundary across the two shots, it is expected that they belong to different scenes;

- when S_i and S_{i+1} exhibit different types of audio classes across the shot boundary, it was expected that clearly define different scenes.

5 SIMULATION RESULTS

One hour of advertisement material was analyzed using the SDC by looking for joint occurrences of shot cut and silence, as described in the

previous section. 100% of the scenes (i.e. advertisements) were properly detected, with a 5% ratio of false alarm (1 out of 20 scene changes did not correspond to real advertisement changes). In these simulations, the shot to scene ratio equal to 11 : 1.

20 minutes of the "First Knight" movie were analyzed and shots were merged whenever they exhibited the same class of audio across video shot boundaries according to the procedure described in the previous section. 98% of adjacent shots exhibiting different classes of audio signals across the shot boundary belong to different scenes. 93% of adjacent shots having different classes of audio signals across the shot boundary were detected, 97% of which corresponded to effective scene changes while 3% were wrongly split. 76% of adjacent shots showing the same class of audio signal across the shot boundary belong to different scenes. 95% of adjacent shots having this type of audio characteristic were identified. 78% of them corresponded to effective scene changes while 22% of them were wrongly separated. 67% of adjacent shots having the same class of signal, even across the shot boundary, belong to the same scene. 89% of adjacent shots having this type of audio characteristic were detected. 60% of them corresponded effectively to the same scene while 40% of them corresponded to different scenes, thus resulting in a miss of scene change detection. This strategy worked remarkably well in the case of music.

In the case of speech, a further analysis to identify features like change of speaker, would have improved the overall system performance. The identification of dialogue situations both in terms of visual and phonetic correspondences should improve the identification of semantically meaningful scenes.

One hour of a national geographic documentary was tested. 82% of adjacent shots exhibiting different types of audio signals across shot boundaries belong to different scenes. 90% of such occurrences were detected, 93% of which corresponded to effective scene changes while the remaining 7% did correspond to the same scene, thus determining false alarms. 50% of adjacent shots having the same type of audio signal but across the shot boundary belong to the same scene. 98% of adjacent shots having this type of audio characteristic was detected, 50% of which corresponded to effective scene change while the other half should not have been separated. 78% of adjacent shots having the same type of audio signal, even across the boundary, belong to the same scene, thus the other 22% resulted in false alarms. 92% of them were detected, 90% of which were correctly merged, while the other 10% were improperly grouped. The scene to shot ratio was 1 : 6, when considering only regions that had been merged.

6 CONCLUSION

We have shown in this paper how scene identification and characterization can be performed using jointly audio and video information. We have used a simple audio model and classification strategy together with a shot cut detector to achieve a good segmentation. Improvements may be expected if a further analysis on the audio and video signals is performed and if a deeper relationship between audio and video is exploited. The advantages of the proposed strategy lie in its simplicity though obtaining relatively accurate results. Efforts need to be further devoted to improve the audio classification strategy and extensively carry out simulations to determine its performance over large samples of audio material.

References

[1] L. Rabiner & B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1994.

[2] Peter De Souza, "A Statistical Approach to the Design of an Adaptive Self-Normalizing Silence Detector", *IEEE trans. Acoust., Speech, Signal Processing*, Vol. ASSP-31(3): 678-684, Jun. 1983.

[3] J. Saunders, "Real Time Discriminator of Broadcast Speech/Music", *Proceedings of ICASSP96*, pp. 993-996, 1996.

[4] I. K. Sethi & N. Patel, "A Statistical Approach to Scene Change Detection", *Storage and Retrieval for Image and Video Databases III*, SPIE Vol. 2420: 329-338, Feb. 1995.

[5] H. Zhang, C. Y. Low and S. W. Smoliar, "Video Parsing and Browsing Using Compressed Data", *Multimedia Tools and Application*, Kluwer Academic Publishers, Boston, Vol. 1: 89-111, 1995.

[6] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full motion search for object appearances", *Proceedings IFIP TC2/WG2.6 Second Working Conf. on Database Sys.*, Sep. 30 - Oct. 3 1991, pp. 980-989, 1991.

[7] M. M. Yeung and B. L. Yeo, "Video content characterization and compaction for digital library application", *Storage and Retrieval for Image and Video Databases V*, Vol. SPIE-3022: 45-58, Feb. 1997.

[8] N. Patel and I. K. Sethi, "Video Classification Using Speaker Identification", *Storage and Retrieval For Image and Video Databases V*, Vol. SPIE-3022: 218-225, Feb. 1997.