Audio-Visual Processing for Scene Change Detection

Caterina Saraceno and Riccardo Leonardi

Signals and Communications Lab.,
Dept. of Electronics for Automation, School of Engineering,
University of Brescia,
via Branze, 38 I-25123, Italy
E-mail: {*saraceno,leon*}*@bsing.ing.unibs.it*

Abstract. The organization of video data-bases according to 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.
As an attempt to extract semantic information, efforts have been devoted in segmenting the video in shots[1] and for each shot trying to extract informations such as representative video frame, etc. As a video sequence is constructed from a 2-D projection of a 3-D scene, processing video information only has shown its limitations especially in solving problems such as object identification or object tracking. Further not all information is contained in the video signal and more can be achieved by analyzing the audio signal as well. Information can be obtained from the audio signal either to confirm the results obtained by a video processing unit or to acquire information that cannot be extracted from video (such as presence of music).
This paper presents a technique which combines video and audio information for classification and indexing purposes.

1 Introduction

The segmentation of a video sequence in shots and the characterization of each shot has been suggested as a technique for organizing video information. Traditionally, algorithms to perform these tasks have been carried out both on compressed [6-7] or uncompressed material [4-5], using only video information.

Depending on the type of video, shots can be as long as 3-4 seconds or less. For advertisements, usually, a shot lasts roughly 2 seconds; in case of movies, rarely it lasts longer than a minute, and around 10 sec. on the average. Further, a shot does not always coincide with a finite action of a movie; also consecutive shots are often semantically correlated to each other.

For this reason, efforts have been devoted to group together shots belonging to the same scene [11], where a scene can be defined as a set of one or more consecutive shots which are "semantically" correlated.

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

Due to the high correlation between the video signal and the associated audio, a scene change detection can be performed using jointly audio and video information. In fact, a joint approach may lead to better performance for the analysis and characterization of audio-visual multimedia information. Since the audio signal is usually composed by several types of audio, such as silence, speech etc, a classification of the audio signal can be performed as a first step for the extraction of audio "semantic" information.

In the next section a classification of the audio signal and a technique to segment the audio signal on the basis of the suggested classification will be proposed. Results obtained from the audio analysis will be shown. Section 3 will then propose a general scheme for scene change detection using jointly audio and video information. Section 4 will show some results obtained by a joint analysis of audio and video. Finally conclusion and future research issues will be discussed in the last section.

2 Audio Processing

In the field of audio processing, big efforts have been devoted to speech recognition and speaker's identification and verification. In case of speech recognition, audio frames composed of speech, and background noise frames are considered [8]. In case of speaker's identification and verification, a speaker is recognized among a set persons [9]. Usually the process is divided into two steps. First, a learning procedure is applied to extract characteristic features for each person belonging to the group to determine his/her class, then the talking person is identified by a simple classification procedure. In practice, the audio signal is simply composed by speech and background noise. Here complex forms of audio are considered. They may thus involve more complex form of preprocessing, to separate the different components of the audio. In the same framework, a recent attempt to separate speech from music on broadcast FM radio has been proposed by Saunders [10].

In a real situation, all audio components are usually mixed together and no a priori information is available, such as number of speakers, type of background noise, etc. The first step is to try to separate the components. Once an audio classification is performed a further analysis can take place on each single type of signal in order to extract informations such as number of speakers, type of music, etc.

In what follows, a classification of the audio file and a technique to perform such a classification is proposed. No other analysis on each class of signal is performed, while the result of the audio classification for scene change detection and characterization is directly demonstrated.

2.1 Audio Segmentation and Classification

Let us suppose that the audio signal is composed of a linear combination of 4 types of signals: **Silence**, **Speech**, **Music** and **Noise**.

• **Silence** segments are those audio frames which only contain a quasi-stationary background noise, with a low energy level with respect to signals belonging to other classes.
• **Speech** segments contain voiced, unvoiced and plosive signals [1].
• **Music** segments contain composition of sound with peculiar characteristics of periodicity.
• **Noise** segments are all other categories, i.e. everything which does not belong to the previous classes. In particular, this class contains non stationary background noise.
The audio signal is split in segments[2] which are consistent with the above classification. For simplicity, a frame will be classified in only one of the previous categories, regardless if more than one of them is simultaneously present in the original signal. This choice is not as restrictive as it appears as we are interested on extracting semantic information: e.g., it is sufficient to identify talking people rather than if they are talking on a silent or noisy environment. On the other hand we do not want to classify as "Noise", frames containing also speech. Therefore, a priority has been created in the classification process: 1.Voice, 2.Music, 3.Noise, 4.Silence.

The audio classification is performed using non-overlapped frames of data. First, the algorithm processes the audio frames in order to detect silence segments, this is performed with an algorithm based on energy information [2] which will be described later. Frames which are not classified as silence are further processed in order to detect the voice portions [1; 3] by the evaluation of an autocorrelation and Zero Crossing Rate (ZCR) measures. Those frames which are neither silence nor voice are analyzed to detect the presence of music, using again an autocorrelation measure.

2.2 Silence detection

Silence segments are detected based upon an analysis of the signal energy using an estimate of its local mean and standard deviation. The basic idea is that the energy present in a silence frame is almost always lower than the energy present in a neighboring non silence frame.

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. The statistics are then dynamically updated. This requires the following assumptions:
• in the background noise there are no abrupt changes in statistics;
• the audio signal starts with a silence segment so as to provide for an initial estimation.
If the background noise can be considered wide stationary at least during short time intervals, the above hypotheses allow to update dynamically its statistics. Obviously using only energy information is not sufficient to discriminate between silence and non silence frames. Due to the stochastic nature of noise, there

[2] An audio segment is a set of one or more consecutive audio frames.

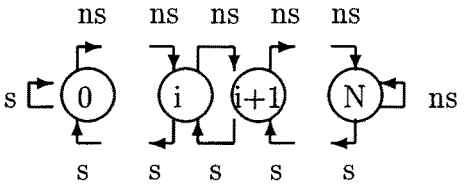

Fig. 1. FSM scheme

may be silence frames with high energy value. On the other hand, parts of voice frames present very low energy values. To reduce the probability of wrong classifications, past and future information is used. This can be obtained with a Finite State Machine (FSM). As shown in Fig. 2 every time the energy value of a frame falls below $m + K * \sigma$, where m is the background noise mean energy, σ its standard deviation and K a constant (typically 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. State 0 represents the silence state while state N represents the non silence state. Frames which belong to inter 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, using the FSM, a low energy frame surrounded by high energy frames is classified as non silence and vice versa. The more the states of the FSM the less sensitive the system to energy fluctuations. Tests have shown that for audio signals sampled at 44.1KHz, 8 is a good number of states for frames of 512 samples (11.6 ms).

2.3 Speech and Music detection

Speech detection has been carried out by evaluating an autocorrelation and Zero Crossing Rate (ZCR) measures. If only continuous speech is present (apart from the background noise), the identification of formants gives a certainty of voice presence. Because just energy information is used to discriminate between silence and non silence frames, unvoiced frames which are at the boundaries of a voice segment could be classified as silence. To avoid this wrong classification a check on the ZCR is sufficient [1]. However, formants are not present on all speech frames and sometimes even voiced segments lead to anomalous sounds. To reduce the probability of misclassification a context based classification has been used: a non-speech frame surrounded by voiced frames is classified as speech.

On the other hand, problems can occur when speech is combined with music. In fact, music has periodicities which sometimes fall in the same range of speech. Music-Speech detection/discrimination can be performed noting that usually music frames present periodicity with a fundamental period which is somehow longer when compared to the voice counterpart and that the ZCR function is smoother. The algorithm tries to detect the speech segments evaluating an autocorrelation measure on frames of 1024 samples. If a periodicity is detected the ZCR is evaluated in order to confirm that such a frame is a speech

frame rather than music. All frames which are not recognized as speech are rearranged in groups of two and for each group again the short-time autocorrelation function is evaluated. If a periodicity is detected, the group of frames is labelled as music, otherwise as noise.

2.4 Audio classification results

Simulations were carried out on:
- 4 min. of audio containing silence and speech ($A1$);
- 4 min. of audio containing classical music ($A2$);
- 4 min. containing audio extracted from the movie "Pulp Fiction" with noise, people shouting, singing and speaking ($A3$);
- 3 min containing audio extracted from the movie "Pulp Fiction" with applause, music, speech and songs ($A4$).

$A1$ was recorded in a very silent environment 93% of silence frames and 96% of voice frames were detected correctly. 2% of silence frames were labelled as voice while 5% of voice frames were labelled as silence. The mismatch occurred only at the boundaries of the different audio frames. 80% of $A2$ frames were recognized as music while 15% were classified as voice and 5% as noise. $A3$ contained noise such as crashing dishes, slamming doors, a woman and a man talking, shouting and singing. The noise frames were all recognized (100%), the silence frames were recognized 94% of the time, voice frames 95% of time, while music was identified 50% of the time.

3 Scene detection

So far, we have proposed a technique to classify audio signals. Let us see, now, how this classification can be used for scene detection and characterization.

According to the previous definition of "scene", scene change detection can be performed using the shot cut detection together with other modules which classify the audio signal and then exploit audio/video semantic correlation among shots. A possible scheme for jointly using audio and video information for scene change detection and characterization is proposed in Fig. 3. The split-and-merge procedure described hereafter on both video and audio signals is used to identify each scene.

3.1 Split procedure

On one hand, the audio file is split in segments according to the classification described above. On the other hand the video signal is split in shots. The shot cut detection can be performed using techniques such as [4-7]. In our algorithm, the Hampapur [5] procedure has been implemented.

Note that during the shot cut detection procedure, it is important not to have misses, while false shot cuts are less critical as scene change detection will be

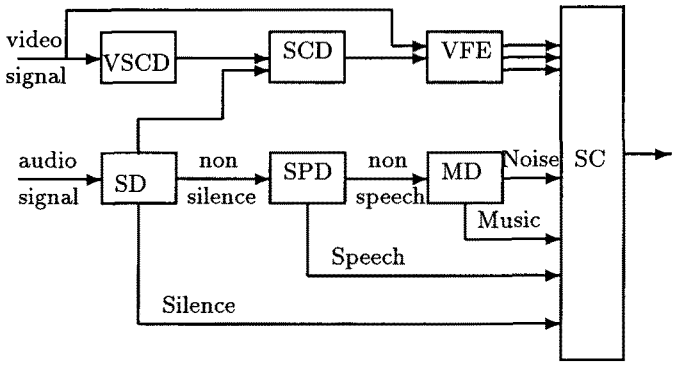

Fig. 2. Scene characterization block diagram

VSCD	: Video Shot Cut Detector	SCD	: Shot Cut Detector
VFE	: Video Features Extractor	SC	: Scene detector & Characterizator
SD	: Silence Detector	SPD	: Speech Detector
MD	: Music Detector		

performed by trying to merge subsequently the detected shots. If a miss happens, the scene will be affected by this error, while for a false cut it is likely that the two shots will be grouped together during the merging stage described in the next subsection.

3.2 Merge procedure

Once the shot change detection has been performed (see Fig. 3), the "VFE" module tries to extract features from each shot by identifying the largest object, using for example a joint segmentation and tracking strategy.

The merging procedure is then performed by the "SC" module so as to provide for scene changes and so as to characterize the resulting scenes. The SC module takes into account information coming both from video and audio classifiers and tries to figure out if a correlation between adjacent shots exists. As an example, a correlation between adjacent shots exists if the two shots have the same audio classification. Besides, an higher correlation exists if the number of speakers in the two shots is the same. If so the corresponding shots are grouped together under one single scene. After all scenes have been identified, the SC module may further characterize them by identifying the most representative frames, the number of objects they contain, the number of speakers, the type of music, ...

We have noted that, depending on the video, a scene change may occur jointly with an audio silence frame.

In the case of advertisement, which can be said to form a single scene, information on silence audio segments can be used together with shot cut detection to

make the scene change detection more robust. In case of movies, an audio silence does not always take place at the boundary of a scene change. Very often, audio anticipates video, i.e. the audio related to the next scene starts a few seconds before the scene changes.

In case of advertisement consecutive shots are grouped together in a single scene if no silence has been detected at the boundaries of the shots.

In all other cases information obtained by the audio classificator and by the shot cut detector are combined in order to detect scene changes. It appears that a good strategy is to group adjacent shots in a single scene when they have the same audio classification.

4 Simulation results

Simulations using shot cut an silence information were carried out on:
- 10 min. of video corresponding to 14 different advertisements with graphics, special effects, fades, dissolves and abrupt changes;
- 10 min. of a dubbed movie with dissolves and abrupt changes.

In case of video advertisements, scene changes were confirmed by the presence of silence at shot cut locations. In 100% of the processed advertisements there was a scene change when a shot cut occurred jointly with a silence segment. The proposed algorithm allowed to detect all occurrences of silence and shot cuts. Only 5% of them did not correspond to scene changes.
In case of movies, only 2% of scene changes occurred jointly with silence frames.

For movies and documentaries, the audio classification was thus used jointly with the shot cut detection module to discriminate different scenes.

One hour of a national geographic documentary was analyzed. 82% of adjacent shots having different type of audio signals 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. 78% of adjacent shots having the same type of audio signal 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 the merged regions.

20 min. of the "First Knight" movie was also analyzed. 98% of adjacent shots having different type of audio signals belong to different scenes. 93% of such occurences were detected, 97% of which corresponded to effective scene changes while 3% were wrongly split. 67% of adjacent shots having the same type of signal 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. A more complete analysis on the video and/or audio signals would have resulted in an improvement of the detection performance. The audio analysis, for example, would have required a speaker recognition unit. The scene to shot ratio was 1 : 5, when considering only the merged regions.

We can summarize that for advertisement, silence detection allows to improve the scene change detection whereas for movies the results are not remarkable. On the other hand, in case of movies, the classification of the audio file can improve the performance of scene change detection.

5 Conclusion

We have shown that audio and video combined together can outperform any separate analysis of each source of information for extracting semantics. Only preliminary solutions to implement the processing units of the proposed scheme (see Fig. 1) have been suggested. In particular more efforts must be devoted especially to improve the SC block (such as the creation of a speaker discriminator, a correlation detector on video and audio frames, etc.) Further to this effort, a systematic evaluation of the simulations must be carried out, to adequately estimate the improvement made possible by a joint audio/video analysis. To further improve the analysis and indexing of multimedia information, it would be also desirable to extract all classes of audio information when occurring simultaneously, rather than simply detecting the one with the highest priority.

References

1. L. Rabiner & B. H. Juang, *Fundamentals of Speech Recognition*, ed. Prentice Hall, 1994
2. P. De Souza, "A Statistical Approach to the Design of an Adaptive Self-Normalizing Silence Detector", *IEEE trans. Acoust., Speech, Signal Processing*, ASSP-31(3):678-684, Jun. 1983.
3. H. Kobataker, "Optimization of Voiced/Unvoiced Decision in Nonstationary Noise Environments", *IEEE Transaction on Acoustic, Speech & Signal Proc.*, ASSP-35(1):9-18, Jan. 1987.
4. I. K. Sethi & N. Patel, "A Statistical Approach to Scene Change Detection", *Storage and Retrieval for Image and Video Databases III*, SPIE-2420:329-338, Feb. 1995.
5. A. Hampapur, R. Jain and T Weymouth, "Digital Video Segmentation", *Proc. of Multimedia 94 Conf.*, San Francisco, pp. 357-363, 1994.
6. H. Zhang, C. Y. Low and S. W. Smoliar, "Video Parsing and Browsing Using Compressed Data", *Multimedia Tools and Applications*, Kluwer Academic Publishers, Boston, Vol. 1, pp. 89-111, 1995.
7. J. Meng, Y. Juan & Shih-Fu Chang, "Scene Change Detection in a MPEG Compressed Video Sequence", SPIE-2419:14-25, 1995.
8. J.W. Pitton, K. Wang and B.H. Juang, "Time-Frequency Analysis and Auditory Modeling for Automatic Recognition of Speech", *Proceedings of the IEEE*, 84(9):1199-1215, Sep. 1996.
9. G.R. Doddington, "Speaker Recognition Identifying People by their Voices", *Proceedings of the IEEE*, 73(11):1651-1664, Nov. 1985.
10. J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music" *Proc. of the 1996 ICASSP Conf.*, 993-996, 1996.
11. 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*,SPIE-3022, pp. 45-58, Feb. 1997.