AUDIO AS A SUPPORT TO SCENE CHANGE DETECTION AND CHARACTERIZATION OF VIDEO SEQUENCES

Caterina Saraceno and Riccardo Leonardi

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

ABSTRACT

A challenging problem to construct video databases is the organization of video information. The development of algorithms able to organize video information according to semantic content of the data is getting more and more important. This will allow algorithms such as indexing and retrieval to work more efficiently.

Until now, an attempt to extract semantic information has been performed using only video information. As a video sequence is constructed from a 2-D projection of a 3-D scene, video processing has shown its limitations especially in solving problems such as object identification or object tracking, reducing the ability to extract semantic characteristics. A possibility to overcome the problem is to use additional information. The associated audio signal is then the most natural way to obtain this information.

This paper will present a technique which combines video and audio information together for classification and indexing purposes. The classification will be performed on the audio signal; a general framework that uses the results of such classification will then be proposed for organizing video information.

1. INTRODUCTION

The segmentation of a video sequence into several clips and the characterization of each clip has been suggested as a technique for organizing video information. Algorithms which try to locate boundaries between consecutive camera shots in a video sequence have been developped as methods for determining scene cuts. Until now, only video information has been used in order to locate cut points [6-10]. Even if good results have already been achieved for abrupt scene change detection, problems appear when fades[1] and dissolves[2] are present. Storically, "shot" and "scene" have been used interchangeably, always refering to the group of frames between two consecutive camera shots. In this work, we want to give a different definition to the term "scene", taking into account the fact that consecutive shots could be related to each other. We define a "scene" as a set of one or more consecutive shots which are "semantically" correlated. Let us consider the following clarifying example. Suppose that in a shot there is a group of persons talking in a room while in the following shot the same group is talking in the same room from a different viewpoint. The video analysis will identify the shot cuts, regardless of the fact that the two shots are semantically correlated to each other. This problem could be overcome through an analysis of the video frames, identifying the inherently present objects and finding a correspondence between each object contained in the two shots. This task is particularly difficult if the correlation between the two boundary frames is low. It is at this point that the audio analysis can be helpful in trying for example to establish a correspondence between the audio segments corresponding to the two shots. In other words, because of the respective significance of audio and video signals, a joint approach may lead to better performance for the analysis and characterization of audio-visual multimedia information.

Further, for indexing or browsing purposes, a simple scene change detection is not sufficient to allow for a complete characterization of the video sequence. Often a more structural organization of the information which does not follow only the temporal evolution of the events, is desirable.

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

2. SCENE DETECTION

Typically, a system for shot cut detection and characterization can be summarized in two steps. First, the whole video is processed in order to detect the cuts. Then, a further processing is performed on each shot, in order to extract semantic features. Algorithms to perform these tasks have been carried out both on compressed [9-10] or uncompressed material [6-8].

According to the previous definition of "scene", scene change detection can be performed using the shot cut detection together with other modules which exploit audio/video semantic correlation among shots. A possible scheme for jointly using audio and video information for scene change

[1] A fade causes a picture to come gradually in or out of view on a screen. The frames gradually brightens in case of fade in and gradually darkens in case of fade out.

[2] A dissolve is a gradual change from one picture into another.

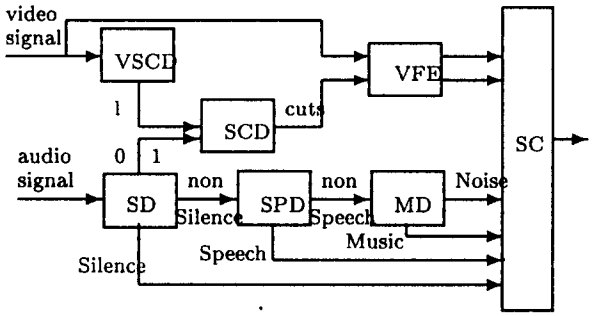

Figure 1: Scene characterization block diagram

detection and characterization is proposed in Fig.1.[3] A split-and-merge procedure on both video and audio signals is performed so as to identify each scene.

2.1. Split procedure

2.1.1. Segmentation and classification of audio

The audio signals is split in segments which are consistent from a classification point of view. We propose to divide the audio signal into 4 classes: **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 phonemes (vowels, diphthongs, semivowels and consonants: the storical 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.

An audio signal can also be obtained as a linear combination of signals belonging to the different classes above. For simplicity, a segment will be classified in only one of the previous categories, regardless if more than one is simultaneously present in the original signal the dominant class will preveil. This choice is not as restrictive as it appears as we are interested on extracting semantic information, it is suficient 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" segments containing also speech. Therefore, a priority has been created in the classification process: 1.Voice, 2.Music, 3.Noise, 4.Silence, where silence is always present on audio signals.

The audio classification is performed as follows: first, the algorithm processes the audio file in order to detect silence segments (SD). This is performed with an algorithm based on energy information [2-3] which will be described later. Segments which are not classified as silence are further processed in order to detect the voice parts [1; 4-5] by the evaluation of an autocorrelation measure (SPD). Those segments which are neither silence nor voice are analyzed to detect the presence of music (MD), using again an autocorrelation measure.

2.1.2. Segmentation of video

The video signal is split in shots. The shot cut detection can be performed using techniques such as [6-8; 10]. In our algorithm, a simple procedure has been used in order to detect shot cuts. For abrupt change detection, the difference energy between the luminance of two successive frames is evaluated. Whenever a local peak is detected the corresponding frame is identified as a cut point. For fade and dissolves, the difference energy between frames at a distance N from each other (tipically $N = 16$ frames at a sampling rate of 25Hz) is misureded. A local peak will correspond to a cut and the frame at the median position (on the interval N) is chosen as a cut point.

2.1.3. Cut detection by joint audio-video analysis

It has been noted that, depending on the video, a scene change may occur jointly with an audio silence segment. For video news and advertisements, this is very likely, while for movies it depends on the director's "style" and type of movie. However, we noticed that this happens rarely. In fact, very often audio anticipates video, i.e. the audio related to the next scene starts a few seconds before the scene changes. Therefore, information on silence audio segments can be used to make the shot cuts detection more robust, as shown in Fig.1, especially for news and advertisement.

In this case, the "VSCD" block of Fig.1 will associate to each frame two probability measures, one corresponding to the probability that the given frame has an abrupt change, the other indicating the probability of fade/dissolve presence on the same frame. For assigning each probability values, we use the two difference energy measures described previously in relationship with the abrupt changes and the fades/dissolves. Whenever the function for abrupt change detection has a local peak, a high probability value is associated to the corresponding frame. The higher the peak with respect to its neighbors the higher the associated probability value. All other frames will be assigned a zero probability measure. For fades/dissolves detection, whenever the corresponding function has a local peak a high probability value is associated with the median frame. The higher the local peak the higher the associated probability value. All the others will be assigned a zero probability. The resulting probabilities will serve as in input to the "SCD" block which will receive a 1 or 0 auxiliary measure from the SD block depending whether the associated audio segments has been classified as silence (0) or non silence (1). Now, four adaptive local thresholds are set, two of them for detecting abrupt changes and two for fades and dissolves[11]. For abrupt change detection (for fades/dissolves a similar procedure can be designed) two thresholds λ_1 and λ_2 are considered. Every frame having a P_a above λ_1 is labelled

[3]

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

as a cut. Whenever this value falls below λ_1 but above λ_2 the audio signal status is checked. If the audio segment corresponding to the video frame has been labelled as silence, the frame will be classified as a cut. All video frames with an associated probability value below λ_2 are never labelled as cut frames, independently of the audio characteristics. The use of audio information allows a more robust shot detection. It is important not to have misses, while false shot cuts are less critical as scene change detection will be reach by trying to merge subsequentlythe 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.

2.2. Merge procedure

Once the shot change detection has been performed, the "VFE" module tries to extract features from each shot by identifying the most important object, using 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. If so the corresponding shots are grouped together under one single scene. After all scenes have been identified, the SC module will further characterize them by finding the most representative frames, the number of objects present, the number of speaker, the type of music, etc. This aspect is beyond the scope of this paper, and remains under investigation.

3. AUDIO PROCESSING

3.1. 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 segment is almost always lower than the energy present in a non silence segment.

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 that the following assumptions are valid:

- 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 hypothesis allow to update dynamically its statistics. Obviously using only energy information is not sufficient to discriminate between silence and non silence segments. Due to the stochastic nature of noise, there may be silence segments with high energy value. On the other hand, parts of voice segments present very low energy values. To reduce the probability of wrong classifications, past and future information will be used. This can be obtained with a Finite State Machine

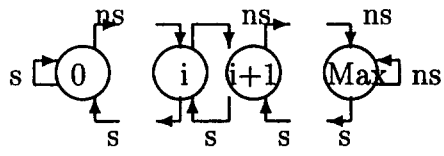

Figure 2: FSM scheme

(FSM). As shown in Fig. 2 every time the energy value of a segment 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 Max is reached. State 0 represents the silence state while state Max represents the non silence state. Segments which belong to inter states are not classified until one of the two state (0 or Max) has been reached. These are then classified according to the reached state.

3.2. Speech and Music detection

Speech detection, on the other hand, has been carried out by evaluating an autocorrelation measure. If only continous speech is present (apart from the background noise), the identification of formants gives the certainty of voice presence. Because just energy information is used to discriminate between silence and non silence segments, unvoiced parts which are at the boundaries of a voice segment could be classified as silence. To avoid this wrong classification a check on the zero crossing rate is sufficient[1].

Music detection has been performed noting that usually music segments present periodicity with a fundamental period which is somehow longer when compared to the voice counterpart. When music or noise is added to speech, it becomes difficult to determine the end points of a speech segment. It is especially hard to discriminate between voice and music (because both of them present periodicities that sometimes falls in the same range) or between voice and noise (due to the unvoiced components of the speech signal and to the fact that sometimes even voiced leads to anomalous sounds, which could be considered noise).

Because formants are not present on all speech segments, the same logic used to determine silence segments has been used: short segments of unvoiced signal are classified as voice if they are surrounded by voiced segments.

3.3. Audio classification results

Simulations were carried out on:

- 4 min. of audio containing silence and speech (let us call it $A1$);
- 4 min. of audio containig classical music ($A2$);
- 4 min. containing audio extracted from the movie pulp fiction with noise, people screaming, 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 segments and 96% of voice segments were detected correctly. 2% of silence segments were labelled as voice while 5% of voice segments were labelled as silence. The mismatch occurred only at the boundaries of the different segments.
$A2$ was taken from a compact disc record. 80% of segments were recognized as music while 15% were classified as voice and 5% as noise.
$A3$ was taken from a compact disc record as well. It contained noise such as crashing dishes, slamming doors, a woman and a man talking, screaming and singing. The noise segments were all recognized (100%), the silence segments were recognized 94 of time, voice segments 95% of time while music was identified 50% of the time.

4. SIMULATION RESULTS

Simulations were carried out on:
• 5 min. of video news containing short reportages with abrupt changes;
• 5 min. of video corresponding to 7 different advertisements with graphic special effects, fades, dissolves and abrupt changes;
• 10 min. of a dubbed movie with dissolves and arbupt changes.
The processed news had two journalists speaking, various reportages were shown. 80% of shot cuts corresponded also to silence audio segments (with a lower percentage in the case of reportage). In case of video advertisements, only 10% of shot cuts corrsponded also silence audio segments. If we consider as a single scene each advertisement, the only way to detect a scene change would have been to use silence-information: 100% of the processed advertisements defined a scene change when a shot cut occurred jointly with a silence segment. The processed movie was dubbed, i.e. for example that people start to talk before the original audio does. This is due to the necessity of the dubbed process: sometimes less words are needed to express a concept in the original language with respect to the language of translation. In this case 2% of scene changes occurred jointly with silence segments. On the other hand, a consistent audio could be detected among shots belonging to the same scene. 30% of adjacent shots had the same music (in terms of average amplitude) 40% of adjacent shots included the same speaker (this was performed in a supervised fashion).

We can summarize that for news and 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, an analysis of the audio file (such as type of music, number of speakers etc.) 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 segments, 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. It would be also desirable to be able to recover all classes of audio information when occurring simultaneously.

References

[1] L. Rabiner & B. H. Juang, *Fundamentals of Speech Recognition*, ed. Prentice Hall, 1994
[2] B.S. Atal & L.S. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition",*IEEE trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201-212, June 1976.
[3] 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, No.3, pp. 678-684, June 1983.
[4] H. Kobataker, "Optimization of Voiced/Unvoiced Decision in Nonstationary Noise Environments", *IEEE Transaction on Acoustic, Speech & Signal Proc.*, Vol. ASSP-35, No.1, pp. 9-18, Jan. 1987.
[5] J.R. Deller Jr., J.G. Proakis and J.H.L. Hansen, *Discrete-Time Processing of Speech Signal*, ed. Macmillan, 1993.
[6] I. K. Sethi & N. Patel, "A Statistical Approach to Scene Change Detection", *Storage and Retrieval for Image and Video Databeses III*, SPIE Vol. 2420, pp. 329-338, Feb. 1995.
[7] G. W. Donohoe, D. R. Hush and N. Ahmed, "Change detection for Target detection and Classification in Video Sequences", *Proceedings of ICASSP 1988*, pp. 1084-1087, 1988.
[8] Y. Nakajima, "A Video Browsing Using Fast Scene Cut Detection for an Efficient Networked Video Database Access", *IEIC Trans. Inf. & Syst.*, Vol. E77-D. No. 12, pp. 1355-1364, Dec. 1994.
[9] 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, pp. 89-111, 1995.
[10] J. Meng, Y. Juan & Shih-Fu Chang, "Scene Change Detection in a MPEG Compressed Video Sequence", *SPIE*, Vol 2419, pp. 14-25, 1995.
[11] A. Hampapur, R. Jain and T Weymouth, "Digital Video Segmentation", *Proceedings of Multimedia 94* 10/94 S. Francisco, pp.357-363, 1994.