VIDEO CONTENT EXTRACTION AND REPRESENTATION USING A JOINT AUDIO AND VIDEO PROCESSING

Caterina Saraceno

PRIP Institute for Automation
Vienna University of Technology, Vienna, Austria A-1040
saraceno@prip.tuwien.ac.at

ABSTRACT

Computer technology allows for large collections of digital archived material. At the same time, the increasing availability of potentially interesting data makes difficult the retrieval of desired information. Currently, access to such information is limited to textual queries or characteristics such as color or texture. The demand for new solutions allowing common users to easily access, store and retrieve relevant audio-visual information is becoming urgent. One possible solution to this problem is to hierarchically organize the audio-visual data so as to create a nested indexing structure which provides efficient access to relevant information at each level of the hierarchy. This work presents an automatic methodology to extract and hierarchically represent the semantic of the contents, based on a joint audio and visual analysis. Descriptions on each media (audio, video) will be used to recognize higher level of meaningful structures, such as specific types of scenes, or, at the highest level, correlations beyond the temporal organization of information, allowing to reflect classes of visual or audio or audio-visual types. Once a hierarchy is extracted from the data analysis, a nested indexing structure can be created to access relevant information at a specific level of detail, according to the user requirements.

1. INTRODUCTION

In order to generate indices that can be used to access a video database, a description of each video sequence is necessary. A first number of attempts described in the literature have focused on determining the different editing stages that took place to compose the video material (abrupt cuts, dissolves, fades,..). The segmentation of the video sequence into its individual shots and the characterization of each shot have thus been suggested as a technique for organizing, at low level, video information. [1] [2]. Although shots can be used as a base for the characterization of video material, they often lead to a far too fine segmentation of the audio-visual sequence with respect to the semantic meaning of data.

On the other hand, information such as the number of speakers, the presence of music etc. cannot be extracted from the analysis of video information though they provide important features for the characterization of the audio-visual material.

In this work, a possible structure of video information based on features extracted from both media (audio,video) and on their correlation is presented. In the next section, four different layers of abstraction are identified: 1. video frames and audio samples; 2. shots and audio segments; 3. scenes; 4. video idioms. An automatic technique which detects certain types of video idioms is presented in section 3. Simulation results which were obtained using such a technique are shown in section 4. Conclusions are drawn in the last section.

2. DATA REPRESENTATION

A hierarchical organization of audio-visual material can, at a low level, characterize on one side the visual information, and on the other side the audio signal. Afterwards, the analysis of correlations existing on the visual signal and/or on the audio signal can lead to higher level descriptions (see Fig.1).

The analysis of visual content should provide informations such as object identification, presence of motion, etc. At the end of the visual content analysis, the visual sequence can thus be described as suggested in [4] by:

• **Frames** with associated descriptors such as color, texture, shape, edge features,

• **Video micro segments**, i.e. sets of consecutive frames characterized by the same type of camera movement or significant object movement. To characterize each video micro segment, a representative frame, also called K-frame, can be chosen from the set of frames forming the shot. Additional informations such as time interval, camera motion (zoom, pan, etc.), object shape, etc. can also serve as visual de-

This work was supported by the Austrian Science Foundation under grant S 7000-MAT.

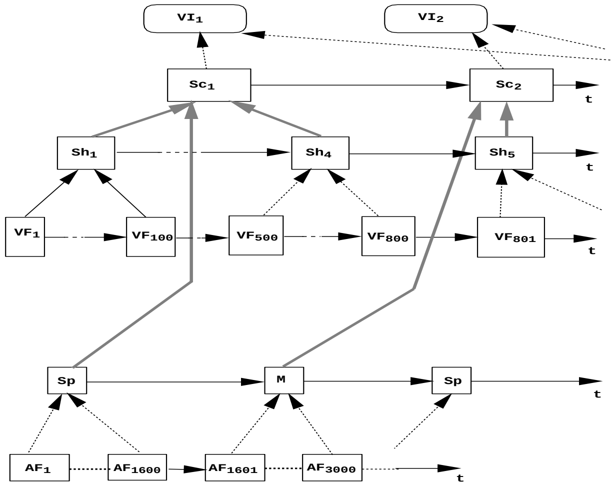

Figure 1: Hierarchical data representation of audio-visual information
VI: Video Idiom; Sc: Scene; Sh: Shot; VF: Video Frame; AF: Audio Frame; Sp: Speech; M: Music.

scriptors at this level;

• **Shots** which are sets of consecutive frames obtained from a continuous camera record. To characterize each shot, descriptors can include the list of video micro segments contained in the shot, the shot duration, the salient still (i.e. K-frame expressing the content of the entire shot), the type of transition at boundaries, etc.

The separation of the audio signal into segments with the same characteristics could be advantageously utilized for the content characterization of the audio-visual material. Audio descriptors can be obtained by extracting information such as number of speakers, type of music, etc. At the end of an audio content analysis, the audio stream may be decomposed as follows:

• **Audio micro segments** are sets of consecutive audio samples characterized by being produced by the same speaker, the same instrument etc. To characterize each audio micro segment, information such as time interval, speaker name, music type, can be used.

• **Audio segments** are sets of consecutive audio micro segments characterized by the same type of audio (music, speech, etc.). Once an audio segment is identified, it might be decomposed in the list of audio micro segments that form it. Descriptors such as duration, number of speakers, etc. may be used to further characterize the audio segments.

Once a segmentation of the audio and visual signals has been performed, consecutive shots may be grouped where some correlation exists between the underlying audio and/or visual information. These grouped shots form the so called *scenes*. Scenes can be sequences of contiguous shots or scenes with the same environment (such as “outside/day”), the same set of characters, common audio/visual object(s), etc. Each scene can be described using information such as time interval, associated shots, associated audio segments, number of speakers, name of speaker, types of sounds, type of scene (e.g. dialogue, etc.). Furthermore, it may occur that scenes, shots or audio segments, which are not consecutive in time, can be semantically correlated as well. If such a correlation is exploited, a different video sequence could be generated by following a different logic of presentation with respect to a simple temporal concatenation of events. We call this representation into classes of video scenes *video idioms*. Video idioms can be sequences of non contiguous shots or scenes with the same environment, a common set of video and/or audio objects, etc. To characterize a video idiom information such as common audio/video object(s), the position of the associated scenes/shots can be considered as suggested in [5]. In the next section, a technique for the identification of certain types of video idioms is presented.

3. VIDEO IDIOM IDENTIFICATION

The identification of video idioms requires, first, an analysis of the visual and the audio information. In our work, the visual information is temporally segmented in shots (as presented in [6]). In order to temporally segment an audio stream, we use a classification based on the types of sounds forming the signal as proposed in [3], where a classification in four classes (silence, music, speech and noise) is presented and the audio signal is temporally segmented according to the given classification. Once the segmentation of the audio and video signals is performed, 4 different types of scenes can be identified (as presented in [6]):

Dialogues: The audio signal is mostly speech and the change of the associated visual information occurs in an alternated fashion, that is, the associated visual labels (which should ideally reflect a change of speaker) follow a pattern of the type ABABAB... [7];

Stories: The audio signal is mostly speech while the associated visual information exhibits the repetition of a given visual content, to create a shot pattern of the type **A**BC**A**DE FG**A**H...;

Actions: The audio signal belongs mostly to one class (which is not speech) and the visual information exhibits a progressive pattern of shots with contrasting visual contents of the type ABCDEF...;

Generic scenes: Consecutive shots which do not belong to one of the aforementioned scenes but their associated audio is of a consistent type.

In order to verify the repetition of similar visual cues among non consecutive shots, a similarity measure was defined in [6]. It is calculated by first detecting for each shot a VQ codebook which reconstructs the shot with a certain distortion with respect to the original visual information. Once a codebook has been associated to a shot, a similarity measure between two shots is defined based on the code vectors representing the shots, as follows:

$$S_{vq}(S_i, S_j) = \|D_j(S_i) - D_i(S_i)\| + \|D_i(S_j) - D_j(S_j)\| \quad (1)$$

where $D_i(S_i)$ is the average distortion obtained when shot S_i is quantized using its associated codebook.

Once scenes have been identified and characterized as belonging to one of the four aforementioned classes, *video idioms* are created by grouping together scenes which have common characteristics according to some correlation measures. Common characteristics among scenes can be exploited by analyzing either the audio signal only, the visual signal only, or both. Depending on the type of analyzed information, video idioms can group scenes based upon different levels of abstraction. For example, if a speaker recognition module were available, a possible video idiom could group together scenes where a particular speaker is present.

In our work, video idioms are identified based on using jointly audio and video information. Four different types of video idioms have been defined. Several more could be considered. Here, video idioms are created by merging together scenes of the same type which exhibit joint audio and visual correlations. In particular, two scenes are grouped together in one video idiom if one of the following four cases is detected, that is *dialogue, action, story* and *generic video idioms*.

Dialogue Video Idiom. If Sc_h and Sc_k are both dialogue scenes, then, according to the dialogue definition, each of them has two recurrent visual patterns (ABABAB). Let $\mathcal{A}_i = \{A_{i1}, A_{i2}, ..., A_{iM}\}$ be the ordered set of shots of Sc_i (with $i = h, k$) belonging to one visual pattern, and let $\mathcal{B}_i = \{B_{i1}, B_{i2}, ..., B_{iM}\}$ be the ordered set of shots belonging to the other visual pattern of Sc_i. The two sets $\mathcal{A}_i$ and $\mathcal{B}_i$ are ordered, therefore, the generic shot $A_{ij} \in \mathcal{A}_i$ temporally precedes shot $A_{i(j+1)} \in \mathcal{A}_i$.

At this point, the rule to group scene dialogues can be formulated. Two dialogue scenes Sc_h and Sc_k are grouped together if $\exists A_{hi} \in \mathcal{A}_h$ and $B_{hj} \in \mathcal{B}_h$, and $\exists A_{kl} \in \mathcal{A}_k$ and $B_{km} \in \mathcal{B}_k$ such that $S_{vq}(A_{hi}, A_{kl}) < \varepsilon$ and $S_{vq}(B_{hj}, B_{km}) < \varepsilon$ where S_{vq} is defined as in Eq. (1) and A_{hi}, A_{kl}, B_{hj} and B_{km} are not noisy shots. In fact in the scene identification procedure, in order to manage possible errors occurring during the visual correlation analysis, a small number of "noisy" shots, i.e. shots which do not follow the alternation of pattern, are also allowed (as explained in [6]).

Story Video Idiom. If Sc_h and Sc_k are both story scenes, then, according to the definition of story scenes, each of them has only one recurring visual pattern (ABACAD). Let $\mathcal{A}_i = \{A_{i1}, A_{i2}, ..., A_{iM}\}$ be the ordered set of shots of a story scene Sc_i, which belongs to the recurrent visual pattern. Furthermore, let $Ssize(Sc)$ be the size of scene Sc, i.e. the number of shots which constitutes the scene Sc. If $Ssize(Sc_h) \leq Ssize(Sc_k)$, the two scenes Sc_h and Sc_k are grouped together if there exist two shots $A_{kl}, A_{km} \in \mathcal{A}_k$, such that A_{h1} is similar to A_{kl} and A_{hM} is similar to A_{km}, where A_{h1} is the first shot of $\mathcal{A}_h$, and A_{hM} is the last one. The idea of considering two shots per scene belonging to the same recurrent visual pattern of the smaller scene is dictated by two different reasons. First, when considering two shots of the same scene, the probability that two scenes which are not correlated are wrongly grouped together is lower than when using techniques which consider only one shot per scene. Tests have shown that more than two shots per scene do not increase the performance. In a scene, shots belonging to the same pattern can be considered similar, i.e the associated codebooks can reconstruct with low distortion all other shots belonging to the same visual pattern. In practice, this is not always the case. Scenes are detected by first analyzing correlations among close shots. Therefore, in cases where there is a slow change in the visual content during the scene evolution, this is not properly taken into account during the scene identification process. On the other hand, the slow change of visual content can create problems in the identification of video idioms, if this effect is not taken into account. That is the second reason why most distant shots of the smaller scene are considered for correlation analysis.

Action Video Idiom If Sc_i and Sc_j are both action scenes, each of them has only one associated audio class and no recurring visual patterns. Let N_{hk} be the number of similar shots between two scenes Sc_h and Sc_k. Two action scenes Sc_h and Sc_k are grouped together if they have the same audio classification and

$$\frac{N_{hk}}{\min(Ssize(Sc_h), Ssize(Sc_k))} > T_{avi} \quad (2)$$

where T_{avi} is typically equal to 0.9, that is, at least 90% of shots belonging to the scene with smaller size should have a similar shot in the other scene.

The number N_{hk} is determined by evaluating the similarity between all shots of the two scenes. A 2D matrix is built for this purpose, where each entry specifies the distance between any two shots of the two scenes, and where Eq. (1) is used to measure the distance between shots. The smallest distance in the matrix identifies a pair of matching shots. This pair is added to a list while the corresponding row and column are removed from the matrix. Subsequently, the next pair is searched by scanning the matrix to identify the next smallest distance. The process is iterated until there are no more cells in the matrix. N_{hk} is the number of all shot pairs which can be considered similar, i.e. for which $S_{vq} < \varepsilon$.

Generic Video Idiom. If Sc_h and Sc_k are both general scenes, each of them has only one associated audio class. Two general scenes Sc_h and Sc_k are grouped together if they have the same audio classification and the number of similar shots between the two scenes satisfies Eq.(2) as well.

	Time (min)	Dialogue	Story	Action	generic scene
movie	30	15	20	19	38
Talk Show	10	20	6	0	0
News	10	17	4	3	2

Table 1: Type and number of detected scenes

Once scenes have been detected, the algorithm for video idiom identification starts by processing the first scene of the video. For each scene, future scenes of the same type are searched for. In case a correspondence is detected between two or more scenes, all similar scenes are merged to form a video idiom.

4. SIMULATION RESULTS

Simulation results were carried on a 30 min. movie, 10 min. news and 10 min. Talk Show. The number and type of scenes detected by the scene detection and characterization algorithm (proposed in [6]) is shown on Table1. For each type of video material the number of correctly grouped, missed and erroneously grouped scenes are shown on Table 2. The decision if two scenes have been grouped correctly or not was made has subjectively, according to information on the visual and audio content. In other words, once the audio-visual material has been automatically segmented in scenes, the scenes are grouped using subjective consideration. Afterwards, the subjective video idioms identification is compared with results obtained by the algorithm. It must be noted that most of the uncertainty and errors are due to general scenes. This is mainly due to the soft definition of generic scenes, which is too general. Besides, a generic scene does not contain any type of specific structure in its visual content, thus making hard the process of its identification, even from a subjective point of view.

5. CONCLUSION

This work deals with the generation of hierarchical representation of audio-visual material for automatic indexing and fast retrieval. The approach we have proposed is based on a joint analysis of video and associated audio signals, presenting a technique to create video idioms exploiting correlations among non consecutive scenes of the same type. A higher video description can be achieved if further analysis on the audio and video signals is performed and if deeper relationships between audio and video are exploited. The advantages of the proposed strategy lie in the combination of its simplicity and its relatively accurate results.

movie	Detect	Miss	False Ident.
Dialogue	11	1	3
Story	14	5	1
Action	13	1	5
Generic Scene	20	4	14
Talk Show	Detect	Miss	False Alarm
Dialogue	7	1	2
Story	5	0	1
Action	0	0	0
Generic Scene	0	0	0
News	Detect	Miss	False Ident.
Dialogue	15	1	1
Story	4	0	0
Action	3	0	0
Generic Scene	0	1	1

Table 2: Video Idiom Results

6. REFERENCES

[1] F. Arman, A. Hsu and M.Y. Chiu, “Feature management for large video databases,” *Proc. of the SPIE Conf. on Storage and Retrieval for Image and Video Databases*, SPIE-1908:2-12, 1993.

[2] A. Nagasaka and Y Tanaka, “Automatic video indexing and full motion search for object appearances,” *Proc. IFIP TC2/WG2.6 Second Working Conf. on Visual Database Sys.*, pp. 980-989, 1991.

[3] C. Saraceno & R. Leonardi, “Video Indexing Using Joint Audio-Visual Semantically Correlated Information” to appear in *International Journal of Imaging Systems and Technology*.

[4] MPEG Requirements Group, “Third Draft of MPEG-7 Requirements,” *document ISO/MPEG N1921*, Fribourg MPEG Meeting, Oct. 1997.

[5] C. Saraceno & R. Leonardi, “Indexing Audio-Visual Sequences by Joint Audio and Video Processing,” to appear in *Proc. of the Workshop on Ontologies for MPEG 7*, Gifu, Japan, 19 Nov. 1998.

[6] C. Saraceno and R. Leonardi, “Identification of Story Units in Audio-Visual Sequences by Joint Audio and Video Processing,” to appear in *Proc. of ICIP’98,* Oct. 4-7, 1998.

[7] M. Yeung & B.L. Yeo, “Video Content Characterization and Compaction for Digital Library Applications”, *Storage and Retrieval for Image and Video Databases V* SPIE 3022: 45-58, San Jose, Feb. 1997