Indexing Audio-Visual Sequences by Joint Audio and Video Processing[ϑ]

C. Saraceno♣♦ & R. Leonardi♦
♦DEA Univ. of Brescia, Via Branze 38, 25123, Brescia, Italy
♣PRIP Vienna Univ. of Technology, Treitlstr. 3, 1040, Vienna, Austria

Abstract. The focus of this work is oriented to the creation of a content-based hierarchical organisation of audio-visual data (a description scheme) and to the creation of meta-data (descriptors) to associate with audio and/or visual signals. The generation of efficient indices to access audio-visual databases is strictly connected to the generation of content descriptors and to the hierarchical representation of audio-visual material. Once a hierarchy can be extracted from the data analysis, a nested indexing structure can be created to access relevant information at a specific level of detail. Accordingly, a query can be made very specific in relationship to the level of detail that is required by the user. In order to construct the hierarchy, we describe how to extract information content from audio-visual sequences so as to have different hierarchical indicators (or descriptors), which can be associated to each media (audio, video). At this stage, video and audio signals can be separated into temporally consistent elements. At the lowest level, information is organised in frames (groups of pixels for visual information, groups of consecutive samples for audio information). At a higher level, low-level consistent temporal entities are identified: in case of digital image sequences, these consist of shots (or continuous camera records) which can be obtained by detecting cuts or special effects such as dissolves, fade in and fade out; in case of audio information, these represent consistent audio segments belonging to one specific audio type (such as speech, music, silence, ...). One more level up, patterns of video shots or audio segments can be recognised so as to reflect more meaningful structures such as dialogues, actions, ... At the highest level, information is organised so as to establish correlation beyond the temporal organisation of information, allowing to reflect classes of visual or audio types: we call these classes idioms. The paper ends with a description of possible solutions to allow a cross-modal analysis of audio and video information, which may validate or invalidate the proposed hierarchy, and in some cases enable more sophisticated levels of representation of information content.

1. Introduction

Today, computer technology allows for large collections of archived digital material. To name a striking example, the Internet itself can be seen as a large unorganised multimedia database, where information of all kinds are stored. Until recently, access to such information has been limited to textual queries. The demand for new solutions allowing common users to easily access, store and retrieve relevant audio-visual information is becoming urgent. Disregarding, in this work, networking and communication issues, an easy access to multimedia information can be achieved by organising data based on its content. If meta-information is associated to each type of data so as to describe its content in sufficient detail, intelligent agents residing on the end user's system could help the user in filtering and retrieving information based on his/her interest.

In this work, meta-information is generated through the analysis of the audio and visual signals, and a hierarchical structure of audio-visual material is created based on the extracted meta-information.

In the next section, a brief overview of the ongoing efforts is presented. In section 2, a possible structure for video information is proposed. Four different layers of abstraction are identified: 1. video frames and audio samples, 2. shots and audio segments, 3. Scenes, 4. video idioms.

[ϑ] This work was partially supported by the Italian Ministry of the University (MURST), by the Austrian Science Foundation under grant S7000-MAT, and by the European Commission ESPRIT contract 28798-AVIR.

2. Previous work

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 identifying the objects contained in each frame, then on trying to track their motion and interaction in space and time by combining the processing results of several consecutive frames [1]. This approach, even if quite attractive from a content description point of view, is relevant only as long as the video sequence is formed from a single camera record (a single shot.). Moreover, it is well known to the image processing community that this task appears quite difficult. When dealing with arbitrary source material, given the current state of the art performance of segmentation techniques, there is a high number of instances that may not be easy to solve. A simpler way to operate has been to focus on low level operations that try to recover the temporal organisation of any given video sequence. Since the early 90's, efforts have been devoted to determine the different editing stages that took place to compose the video material (abrupt cuts, dissolves, fade in, fade out,..). Such segmentation of the video sequence into its individual shots and the characterisation of each shot have thus been suggested as a technique for organising, at low level, video information. Traditionally, algorithms to perform these tasks have been carried out both on compressed [2] or uncompressed material [3].

A key aspect is that they operate only by analysing video information. Although shots can be considered as a base to begin the characterisation of the video material, they do not lead to a complete and efficient description of the temporal organisation of the video sequence. Shot separation often leads to a far too fine segmentation of the audio-visual sequence with respect to the semantic meaning of data. Thus it is necessary to provide other means to combine shots into meaningful scenes often referred to as story units [4]. 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 these provide important features for the characterisation of the audio-visual material. In addition to its own characterisation capability, processing audio information could be efficiently used jointly with video processing to extract relevant content description. Patel and Sethi [5] have examined the potential of speaker identification techniques for characterising video clips in terms of actors present in the scene. Assuming to know the major actors in a movie, the goal is to label each video clip with descriptors indicating the presence of a certain actor.

Another example of using audio and video information for scene classification was proposed by Wang et al. [6]. They presented a technique to classify scenes based on motion information and audio classification. They tried to classify four types of TV programs: news, advertisement, weather forecast and football game. On one hand motion was used based on the idea that usually news are characterised by little overall motion, football games have spread motion (players are moving in different direction), and baseball games have a lot of camera panning effects. On the other hand, the audio classification was performed by using the volume signal distribution, the silence interval distribution, the spectrogram, the central frequency and the audio signal bandwidth. They showed that some of those features, especially the mean and standard deviation of silence intervals can discriminate between news, commercials and football games. Nam and Tewfik [7] detected shot cut using a 3D wavelet decomposition. The audio signal was divided into five unequal-widths subbands. Scenes were detected if a shot cut occurred in proximity of a sharp temporal variation in the power of subband audio signals.

Yeung and Yeo [1] have proposed a technique to characterise the video in temporal events by analysing only visual information. They first segment the video in shots and then they merge consecutive shots that present specific homogeneous characteristics. They define three models of temporal events: *dialogues*, *actions* and *story units*. The term dialogue is referred to an event where a conversation-like montage is present. If a label is associated to each shot, representing the content of the shot, ABABABAB is a possible label sequence of a dialogue. An action is characterised by a progressive presentation of shots with contrasting visual contents, often used to express fast movements or emotional impact. For example, the label sequence ABCDEFBGHI characterises an action event. A story unit is a collection of contiguous shots such that the associated label sequence is the minimal label sequence that contains all the identical (or recurring) labels. As an example, ABACDEFCGF is composed by two story units, the first story unit is ABA and the second is CDEFCGF. In other words, a story unit is considered as a higher level of abstraction, compared to dialogues and action. A story unit is a dialogue and/or an action, but it can also combine them. For example, the story unit ABABCDEFABA is composed by a dialogue (ABAB), an action (CDEF) and again the same dialogue (ABA).

In our work, we do not want to restrict the levels of abstraction to only two levels. Besides, we do not want to restrict the higher level (i.e. the story unit) to contain only identical labels. As an example, according to [1] the label sequence ABABCDCD is composed of two dialogues ABAB, and CDCD, which are also two different story units. Our idea is to identify the two dialogues at a lower level, and to have the possibility to group together consecutive dialogues at a higher level, even though they do not contain the same speakers, but they have a common action: the ``dialogue''. Indeed, a mere analysis of the video signal cannot guarantee that a label sequence ABABAB is a dialogue. It can be defined as a dialogue only if the associated audio signal contains mostly speech. Therefore, we decided to define four different types of scenes based on both audio and video information, as explained in the following section, where a hierarchical representation of audio-visual material is presented.

3. Hierarchical structure

A hierarchical organisation of audio-visual material can, at a low level, characterise on one side the visual information, and on the other side the audio signal. Afterwards, the analysis of correlation existing on the visual signal and/or on the audio signal can lead to higher level descriptions.

In the following paragraph, a possible hierarchical organisation of visual information is presented, followed by the organisation of audio information. Then, a joint audio-visual description is presented.

3.1 Visual content description

The analysis of visual content should provide information 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 [8] by:

- **Frames** with associated descriptors such as colour, texture, shape, edge features.
- **Video micro-segments**, i.e. sets of consecutive frames characterised by having a same camera movement or a same significant object movement. To characterise each video micro segment, a representative frame, also called K-frame, can be chosen from the set of frames forming the shot. Additional information such as time interval, camera motion (zoom, pan, etc.), object shape, etc. can also serve as visual descriptors at this level.
- **Shots** being sets of consecutive frames obtained from a continuous camera record. To characterise 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.

3.2 Audio content description

The separation of the audio signal into segments having same characteristics, could be advantageously utilised for the content characterisation 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 characterised by being produced by a same speaker, a same instrument etc. To characterise 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 characterised by having 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 characterise the audio segments.

3.3 Joint description

Once a segmentation of the audio and video signals has been performed, consecutive shots may be then grouped when some correlation exists on the basis of the underlying audio and/or visual information. These grouped shots form what we shall refer to as *scenes*. Scenes can be sequence of contiguous shots or scenes with same environment (such as ``outside/day''), same set of characters, common video object(s), common audio object(s), 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.).

This approach allows for organising the audio-visual data for any subsequent analysis, classification or indexing task. Furthermore it may occur that scenes, which are or 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 same environment, common set of video and/or audio objects, etc. To characterise a video idiom information such as common audio/video object(s), position of the associated scenes/shots can be considered.

Video idioms may also be convenient for compression purposes, both in terms of audio and video. The content of all scenes, that combine into a single video idiom, may simply be reconstructed using a tuned (or a set of tuned) dictionary of words that match well the information content of the underlying source(s), from an information theoretic point of view. By a proper training, it may be possible (for example) to match a given speaker (identified by its audio) to a certain face (being present in the video sequence). This information can be efficiently used to merge consecutive shots into a single scene. In a more complex framework, an alternation of speakers may identify a dialogue situation. This alternation could be observed both in terms of the visual and audio signals.

The key point is to establish a hierarchical data structure (a relational or a transition graph) which allows to handle different levels of abstraction of the audio-visual data. At the lowest level, we have individual audio and video frames. These are grouped separately at the next level into shots for video frames, and consecutive audio frames of one class for audio information. The next level up identifies each scene, with a descriptor of its content: e.g.

1. the length of the scene, a series of key video/audio frames;
2. a dictionary of visual words and another dictionary of audio words which may be used to generate all possible audio/video frames using a VQ based decoding algorithm;
3. present correlation between audio and video components, classification information (such as type of audio, identification of speaker, type of video material...).

From this level downward, all nodes of this relational graph are connected to each other according to the temporal evolution of the sequence (see Fig. 1). Finally a top level may be used to break the temporal relationship among scenes (defining the video idioms). Links are created between different nodes at this level, when correlation exists, so as to group the different scenes. Additional descriptors may be incorporated to further enable a characterisation of the information.

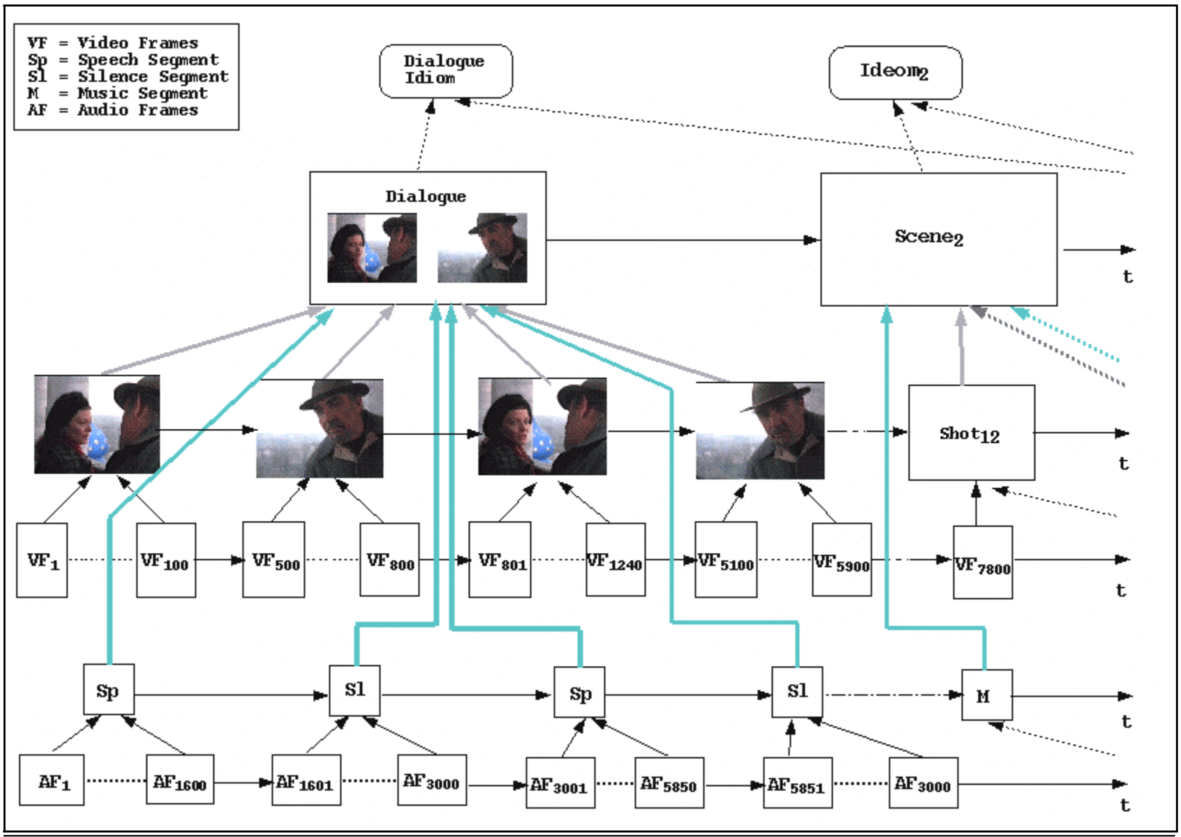

Figure 1: Hierarchical data representation of audio-visual information

The advantage of such a representation is, of course, to have a faster access to specific information, allowing for an efficient navigation through the audio-visual data. Operations such as coding (as mentioned previously), editing, retrieval from specialised queries may all benefit from this type of representation.

A more attractive solution may be to enable easy composition of novel audio-visual material. In fact, by reversing the order of scenes, shots or consistent type of audio segments, new audio-visual sequences may be generated with little diversity with respect to the semantic content of the original sequence. If different, it may help each user to create information that better expresses his/her ideas.

In the next section, a definition of each component of the graph of Fig. 1 will be provided.

4. Data representation

At the lowest level, video frames are segmented in shots. Each shot has an associated *meta-information*. A *meta-information* Sh describes the time interval in which the shot resides, the predominant audio classification C_i, the percentage x_i of samples belonging to such a classification (evaluated over the entire shot), a reference to the codebook Q_i associated with the shot, a reference to the K-frame KF_i representative of the shot, a label of the visual content in the shot, plus a variable number of parameters to be defined by the user.

For the *i*-th shot, the associated meta-information is

$$Sh_i\left(\left\lfloor t_l^i, t_m^i \right\rfloor C_i, x_i, Q_i, KF_i, ...\right)$$

On the other hand, the audio signal is split in segments according to some criteria. Each audio segment has an associated *meta-information AF*. This information indicates the time interval, the classification C_i, plus a variable number of information, which can be defined by the user (such as the name of the speaker, the type of music, etc.). For the *i*-th audio segment, the associated meta-information is

$$AF_i\left(\left\lfloor t_l^i, t_m^i \right\rfloor C_i, ...\right)$$

Consecutive shots can be merged together to create a scene. A scene can also be created by merging several scenes. The only constraint is that a scene represents a set of contiguous frames. As an example, let us assume to have two contiguous scenes, both labelled as dialogues, but with different speakers. At a first level of abstraction, an accurate algorithm would identify them as two different dialogues, due to the different speakers involved, but at a higher level of abstraction may be, only the concept “dialogue” is of interest. In such a case, a scene which connects the two “dialogue” scenes is needed. In general a scene contains also an associated *meta-information*. Each meta-information contains the first and last shot of the scene, the predominant audio classification C_i among the shots, a label VL_i of the visual content in the scene, a series of K-frames $KF_{il},\ldots,KF_{iN}$, plus a variable number of parameters to be defined by the user, such as audio K-frames, a *salient still frame*, i.e. a frame representative of the entire sequence of shots (for example a panoramic image of the shot), etc. For the *i*-th scene, the associated meta-information is

$$Sc_i\left(\left\lfloor Sh_l^i, Sh_m^i \right\rfloor C_i, VL_i, KF_{i1}, ..., KF_{iN}, ...\right)$$

Video idioms are at highest level of the hierarchy. They are formed by merging scenes or other video idioms (depending on the level of abstraction). The difference between scenes and video idioms is that the former represents a set of contiguous frames, while the latter does not. Video idioms are created to exploit correlation among non contiguous scenes. A video idiom is represented by the list of scenes composing the video idioms, a label VL_i of the video content, a label to the audio content AL_i, a series of K-frames $KF_{i1},...,KF_{iN}$, plus a variable number of parameters to be defined by the user. Therefore the *i*-th video idiom can be referred to

$$VI_i\left(Sc_j, Sc_k, ...Sc_l, VL_i, AL_i, KF_{i1}, ..., KF_{iN}, ...\right)$$

An example of a technique to hierarchically organise audio-visual information and, in particular, to identify video idioms is presented in the following section.

5. 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 [9]). In order to temporally segment an audio stream, we use a classification based on the types of sounds forming the signal as proposed in

[10], 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 have been performed, different types of scenes are identified (as presented in [9]). In particular, *Dialogue* scenes are identified when the audio signal is mostly speech and the the associated visual information exhibits an alternation of visual content to create a shot pattern of the type ABABAB (which should ideally reflect a change of speaker).

In order to verify the repetition of similar visual cues among non consecutive shots, a similarity measure has been defined in [9], and it can be used also to evaluate the visual correlation among scenes. 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:

$$S_{vq}\left(S_i S_j\right) = \left\| D_j\left(S_i\right) - D_i\left(S_i\right) \right\| + \left\| D_i\left(S_j\right) - D_j\left(S_j\right) \right\|$$

where $D_i\left(S_i\right)$ is the average distortion obtained when shot S_i is quantized using its associated codebook. Once *dialogue scenes* are identified, *dialogue video idioms* can be created by grouping together *dialogue scenes* which have common characteristics, according to some correlation measures (e.g. same speakers).

In our work, dialogue video idioms are identified by grouping together dialogue scenes which have same recurrent visual patterns (ABABAB). Several other techniques can be defined, which, for example, group together dialogue scenes having a common speaker, etc. Experiments were carried out on 10 min. Talk Show having 20 automatically detected dialogue scenes. 17 dialogue scenes were grouped correctly, 1 dialogue scene was missed (i.e. the algorithm did not recognise the visual correlation between two dialogue scenes) and 2 scenes were group in wrong video idioms (i.e. they were group with scenes which did not have same visual correlation).

6. Conclusion

This work deals with the generation of hierarchical representation of audio-visual material for automatic indexing and fast retrieval. The proposed approach is based on a joint analysis of video and associated audio signals. A technique to create video idioms exploiting correlation among non consecutive dialogue scenes is presented. Several more techniques could be considered. A higher level semantic video description will be achieved if the relationship between audio and video is further exploited.

References

[1] M.M Yeung and B.L. Yeo, Video Content Characterization and Compaction for Digital Library Application, *Proc. of the SPIE Conf. on Storage and Retrieval for Image and Video Databases V*, SPIE-3022:45-58, Feb. 1997.

[2] 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.

[3] 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.

[4] M.M. Yeung and B.L. Yeo, Time-constrained clustering for segmentation of video into story units, *Proc. of ICPR'96*, III:375-380, Aug. 1996.

[5] N. Patel and I.K. Sethi, Video Classification Using Speaker Identification, *Proc. of EI'97: Storage and Retrieval for Image and Video Databases V*, SPIE-3022:218-225, Feb. 1997.

[6] Y. Wang, J. Huang, Z. Liu and T. Chen, Multimedia Content Classification using Motion and Audio Information, *Proc. of IEEE ISCAS'97*, 2:1488-1491, 1997.

[7] J. Nam and A. H. Tewfik, ``Combined Audio and Visual Streams Analysis for Video Sequence Segmentation, *Proc. of ICASSP'97*, 3:2665-2668, 1997.

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

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

[10] 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.