A Tool for News Video Scene Indexing

Nicola Adami and Riccardo Leonardi

University of Brescia
Via Branze 38 - 25123 Brescia - IT
{adami,leon}@ing.unibs.it

Abstract. Indexing of audiovisual sequences is a difficult task, and it is often necessary to build ad-hoc techniques according to the considered audiovisual class. In this work we present an interactive approach to news programs scene indexing and classification. The entire process is fully automated but the user is allowed to manually correct intermediate and final results. Basically the news video is first segmented in shots. Then, each shot is associated to a codebook obtained by a vector quantization process. Shots in which the anchorperson appears can be easily separated as their associated codebook exhibits peculiar characteristics. Sequences of consecutive shots can be identified forming scenes classified as "Outdoor" or "Indoor" following a certain semantics. This process is achieved by establishing correlation among shots and using the statistical information regarding the temporal distribution of anchorperson shots with respect to all the other shots of the news programme.

Introduction

The explosion of digitally represented audio-visual material imposes a need to define suitable frameworks for retrieving and browsing relevant information according to user specific requirements. In October 1996, the International Standard Organization (ISO) started a standardization process for the description of the content of multimedia documents, namely MPEG7 [3]. This standardization effort defines a set of standard Descriptors (D) and Description Schemes (DS) expressed according to a Description Definition Language (DDL). A DS can be used to structure a description of a multimedia document, by organizing descriptors (D) characterizing features such as shape, colour, texture, motion (for objects), or audio class type.

For several reasons news programs contain very important multimedia information, and they are thus receiving more increasing attention with respect to other classes of audio-visual programmes. Indeed a generic user can use a news video description for fast browsing and content understanding, on the other hand, professionals may retrieve a specific story unit from old material in a digital archive to construct a reportage. A rich description of news story can also be used to track the number of times a particular politician has been on the air and for how long each time. Due to the strong request of this kind of description, several automatic and semiautomatic tools are being proposed in the literature. Most of the techniques are based on textual and linguistic properties [6], while other approaches use face recognition in order to identify relevant events in the video [8]. More sophisticated system use both video cues and text extracted from closed captions and speech [5].

In this work we propose an alternative method to news video indexing. The main intention is to show how, using video features, a news program can be segmented in video scenes which can be further classified. This represent a basic tool that can be integrated with a more sophisticated one to reach a fully automatic process, and obtain a rich description of the audiovisual content. The main idea is to use a codebook obtained by a vector quantization algorithm to describe the content of a video shot. As it has been shown in a previous work [2], this descriptor expresses very well the block image color structure and distribution. It can be thus used to easily correlate shots among a video sequence. In addition a news video has a typical organization of the story unit where at least the anchorperson appears at the begin of each unit. Based on this consideration, and taking into account the high correlation between anchorperson shots, it is possible to detect all of them by looking at the evolution of some parametric representation of the anchor shot with respect to the same representation of all other shots in the news program. The tool described hereafter can serve as a support for a user in order to help in generate a description of a news content. While all the indexing and classification process are fully automated, the user is allowed to manually correct intermediate and final results so as to speed up the content description creation and to obtain a final result of the same quality level of a fully manually annotated description.

This paper is organized as follows. In Section 1 a general description of the tool is proposed. In Section 2 the codebook descriptor is introduced together with the extraction process and the similarity measure

used to establish shot correlation. In Sections 3 and 4 all the steps required to achieve a final scene segmentation and classification are presented. Finally, an example of application of the presented tool is shown in Section 5, demonstrating the high accuracy that can be achieved.

1 Scene segmentation classification tool: a global view

In this section we present the structure of the scene identification and classification algorithm. As it is shown in Fig. 1, the video sequence is, first of all, temporarily separated into shots. To achieve this goal the algorithm described in [1] has been used. The aforementioned technique provides the shot boundary and also the type of editing effect used to link two consecutive camera records. The editing effects information can be useful for scene identification considering that usually the "editing style" is kept constant for some categories of scenes in a news program. For example, in news programs used to test the here proposed method, the first anchor-shot of a news scene is always separated from the subsequent shot, belonging to that same scene, by means of a dissolve.

For each shot a codebook descriptor is generated, as described later in Section 2. In the following stage the shots containing the anchorperson are identified. As explained in detail in Section 3 this is accomplished using the peculiarity of the distribution of the similarity error between the "Anchor" shot and all the other shots in the program. At this level the output result is presented to the user that can refine it, by excluding all shots which could have been labeled improperly as "Anchor". All video segments separated by anchorperson shots are labeled as possible scenes of interest as described in Section 4, on the basis of the average time duration separating two consecutive "Anchor" shots. Depending on the pattern and number of "Anchor" shots belonging to the remaining video segments these are divided further into possible "Outdoor" or "Indoor" subsegment.

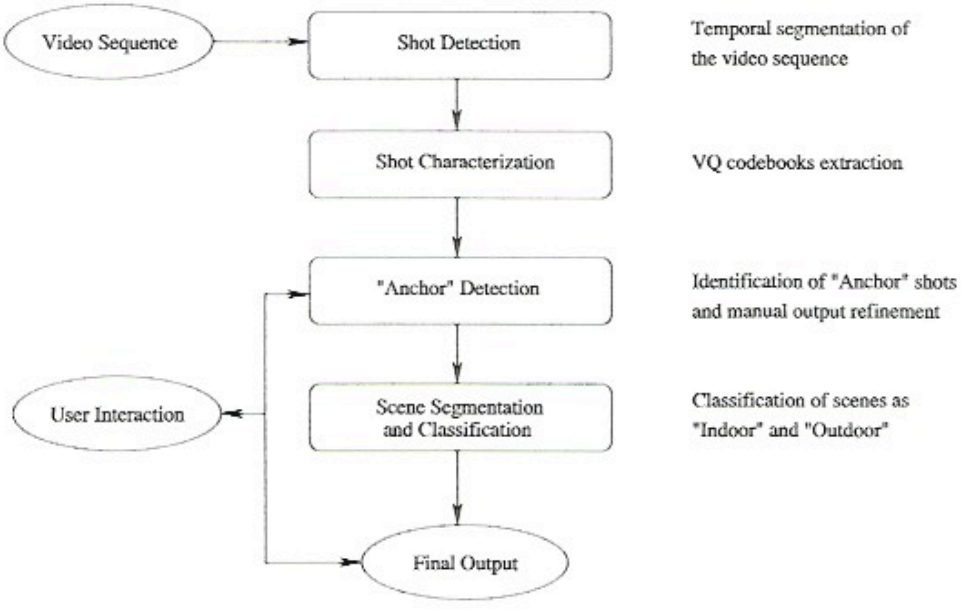

Fig. 1. Block diagram of the scene identification and classification process.

2 Shot content modelling

As described in [2], the VQ codebook descriptor of a given shot can be efficiently used to compactly represent the color pattern distribution of the image blocks forming a video shot. A similar approach is used in [7] in order to group sequences of shots into clusters to identify scenes with a particular semantics. The VQ codebook descriptor includes the set of codewords, their relative frequency with respect to the training data, and the variances of the training data with respect to the each code word. The training

data are obtained from blocks of frames forming a shot. To add the relative frequency and variance gives a more complete information about the data distribution instead to use only the codewords as proposed in [7].

2.1 Codebook estimation

The generation of the codebook of a given shot is obtained using the *LBG* vector quantization algorithm. The training set is composed of the vectors obtained by decomposing each frame into blocks of 4×4 pixels. More specifically, we have limited the frame selection to equally sub-sampled frames. In addition, each frame is spatially sub-sampled by a factor of two, in order to reduce further the amount of data to process. All blocks are stored to form vectors where the original *RGB* color space is converted to a representation in the *Luv* space. The criterion imposed to stop the LBG iteration and the splitting is either a maximum number of codewords, set to 100, or a threshold on the decrease of the global distortion value. For each codeword, the relative number of vectors in the Dirichlet region associated with the codeword is computed so as to obtain a relative frequency parameter. We evaluate also as an additional indicator the standard deviation of the vectors in each training region with respect to its centroid (i.e. the codeword).

2.2 The Dissimilarity Measure

The comparison of two different codebooks, composed by words of the same size, is not a trivial task even when they have the same number of words. By possibly using the maximum distortion as the convergence criterion for designing the codebook, the codebook size may depend on the complexity of the shot. The dissimilarity is estimated by calculating the average of the minimum distance between each pair of words belonging to different shots. The metric proposed in [4] well fits the need to deal with descriptors with variable dimension components. Given two mixtures $A(x)$ and $B(x)$ where:

$$A(x) = \sum_{i=1}^{N} \alpha_i a_i(x) \qquad \text{and} \qquad B(x) = \sum_{j=1}^{M} \beta_j b_j(x)$$

with $\sum_{j=1}^{M} \beta_j = 1$ and $\sum_{i=1}^{N} \alpha_i = 1$, the distance between A and B is measured by [4]:

$$D(A,B) = \min_{\mathbf{W}=[w_{i,j}]} \sum_{i=1,j=1}^{NM} w_{i,j} d(a_i, b_j)$$

with the constraints

$$\forall \quad i,j \quad w_{i,j} \geq 0, \qquad \sum_{j=1}^{M} w_{i,j} = \alpha_i, \qquad \sum_{i=1}^{N} w_{i,j} = \beta_j.$$

As explained in [4], the weights $w_{i,j}$ represent the solution to a minimization problem. In case of measuring the distance between two codebooks with a large number of codewords, finding the weights could represent a significant computational burden. By using the simplex method, for example, the problem can be reduced to finding the solution of a linear system of dimension $(N + M, M * N)$. To reduce further the complexity we propose the following measure:

$$\begin{aligned} D(A,B) = & \sum_{i=1}^{N} \min_{\mathbf{W}=[w_{i,j}]} \sum_{j=1}^{M} w_{i,j} d(a_i, b_j) \\ & + \sum_{j=1}^{M} \min_{\bar{\mathbf{W}}=[\bar{w}_{i,j}]} \sum_{i=1}^{N} \bar{w}_{i,j} d(a_i, b_j) \end{aligned} \tag{1}$$

with the constraints

$$\forall \quad i \quad \sum_{j=1}^{M} w_{i,j} = \alpha_i, \qquad \text{and} \qquad \forall \quad j \quad 0 \leq w_{i,j} \leq \beta_j$$

$$\forall \quad j \quad \sum_{i=1}^{N} \bar{w}_{i,j} = \beta_j, \qquad \text{and} \qquad \forall \quad i \quad 0 \leq \bar{w}_{i,j} \leq \alpha_i.$$

With the above constraints, the computation of the weights $w_{i,j}$ and $\bar{w}_{i,j}$ can be directly obtained because the minimization problem can be applied separately on each row and column. This new measure, compared with the original one, has lead to same behaviors and comparable performance, but it allows a faster computation of the distance between codebooks with large number of codewords. The component distance $d(a_i, b_j)$ used in (1) between two codewords of dimension K has been set to:

$$d(a_i, b_j) = \sqrt{(1-\alpha)\sum_{k=1}^{K}(\mu_{a_{i,k}} - \mu_{b_{j,k}})^2 + \alpha\sum_{k=1}^{K}(\sigma_{a_{i,k}} - \sigma_{b_{j,k}})^2}$$

where $\alpha \in [0,1]$ establishes the relative weight between mean value and standard deviation for the distance computation.

3 Anchor man detection

The "Anchor" shot detection is a fully automatic process based on the analysis of the statistic distribution of the dissimilarity between a candidate shot to be labeled as "Anchor", and all other shots of the news program.

The probability to find such a shot inside a temporal windows placed at the beginning of the video sequence is very high. It is of 100% if the window extent is sufficiently large. For this reason, each shot inside this initial window is compared with all shots forming the news program by calculating the dissimilarity measure (1). By using this information we estimate the "Anchor" shot positions in the initial window and all the similar ones in the news program. Let us make some consideration on the dissimilarity error distribution. As it is shown in Fig. 2a, such dissimilarity measure values, in case of an "Anchor" shot, are concentrated in two well separated intervals, where the values around the lowest one are those of the shot with a high probability to be classified as "Anchor". In the case of a "Generic", shot the error distribution is in general more uniform (see Fig 2b). Hence a criteria to detect the "Anchor" shot can be to identify an isolated peak in the dissimilarity distribution in correspondence to the lowest error values, and select as similar all the shots with dissimilarity error below the upper bound of the interval covered by the peak. In order to avoid the use of any threshold this operation is made by using a GMM to model the dissimilarity distribution. For each shot in the initial windows the corresponding dissimilarity values are used to fit a GMM with $N = 3$ gaussian functions. With this value it is possible to well fit the two types of considered distributions. In particular this is the minimum number of gaussians to model the lowest peak of an "Anchor" shot with a single gaussian function. The analysis is then performed on the GMMs parameters. The GMMs associated to all the shots in the initial window are ordered on the basis of the lowest gaussian mean value. For each of the first M=10 GMMs in this ordered list, all the shots of the news program, which exhibit a dissimilarity measure lower than the gaussian with lowest mean value plus its associated standard deviation are declared "similar". However this operation is not sufficient to select the "Anchor" shots because in some cases the GMM associated to a "Generic" shot can incorporate gaussian with a lowest mean error comparable with the one of an "Anchor" GMM. For this reason a final test is performed by looking at the temporal distribution of the distance between consecutive shots, similar to each of the 10 candidate one. As shown in Fig. 3a the distribution of the time intervals, in terms of number of shots between two consecutive "Anchor", shows several lobes while in case of "Generic" shots there is a single lobe, (see Fig. 3b). All shots that satisfy this condition are labeled as "Anchor" together with those declared similar to it. The results of this automatic identification can then be presented to the user so that remaining improperly classified results can be discarded.

4 Scene detection and classification

Once the position of the shots in which the anchorperson appears have been identified it is possible to build and classify the scenes forming the news program.

This operation is performed in two steps. In the first, the mean interval T_{av}, in terms of number of shots, between two consecutive "Anchor" is computed. In Fig. 4 a state diagram describes how the first stage of the segmentation is implemented. Any time two "Anchor" shots are found, at a distance greater than T_{av}, a scene is detected and labeled as "Outdoor" (see Fig. 5). This label indicates a story unit referring to something happened outside the TV studio, from where the news program is being recorded. All the sequences of shots between "Outdoor" scenes are labeled as "Possible Dialog" and further processed to detect dialog pattern and/or other "Outdoor" scenes.

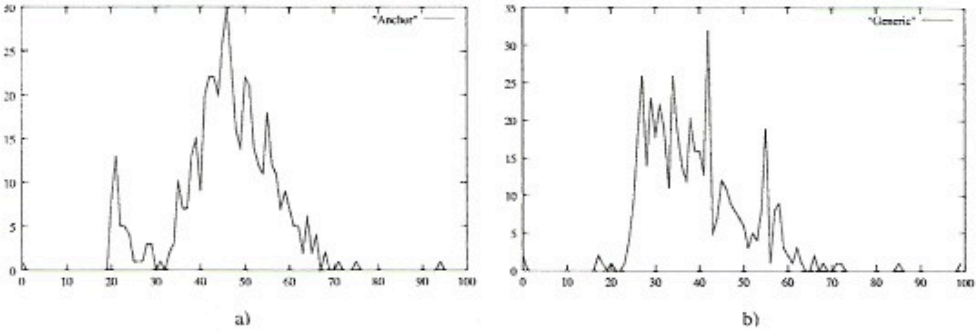

Fig. 2. Similarity error distribution a) for a "Anchor" shot and b) for a "Generic" shot.

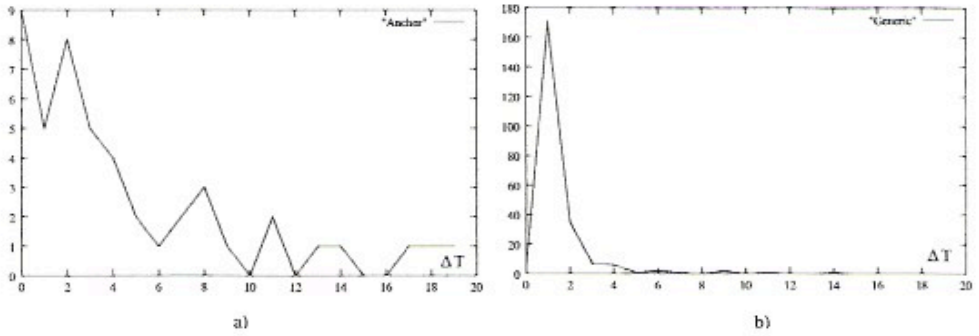

Fig. 3. Shot interval distribution a) for a "Anchor" shot set and b) for a "Generic" shot set.

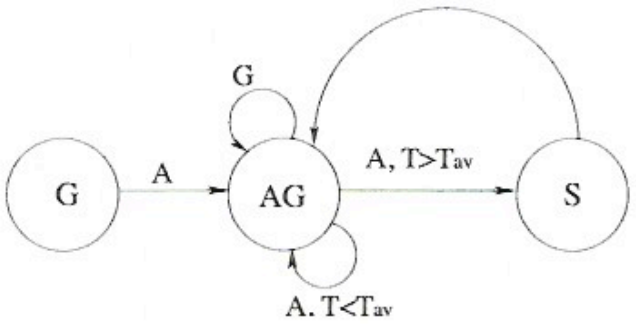

Fig. 4. State diagram for the scene identification.

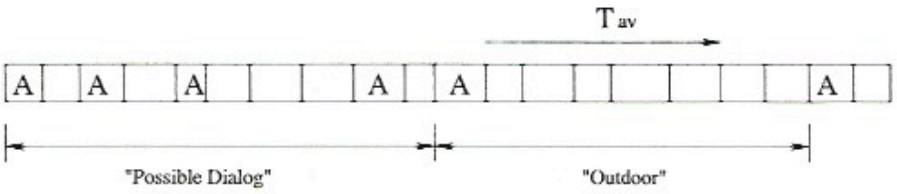

Fig. 5. The two alternative state sequence patterns.

In the final step the scenes previously labeled as "Possible Dialog" are analyzed and classified. This is accomplished by looking at the shots correlation pattern. All the shots in a scene, except those labeled as "Anchor", are clustered by using the dissimilarity measure introduced previously. A label is then assigned to each cluster obtaining patterns of labels as shown in Fig. 6 ("A" represent the anchorperson, while a same letter is attached to shots exhibiting similar content). This label list is then analyzed and a sequence of shots is marked as "Indoor" whenever it contains at least two "Anchor" and two shots labelled with the same letter. Otherwise the subsegment is marked as "Outdoor".

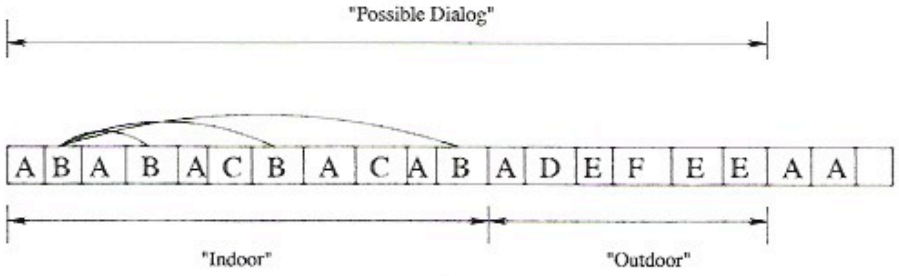

Fig. 6. Dialog identification.

5 Example

In this section we present a scene segmentation and classification results obtained using the method described previously with a news program extracted from the MPEG-7 content set. Once the program has been segmented into shots, a codebook is built as described in Section 2 for each of the shots. The "Anchor" shots are then identified as described in Section 3 and using a window of "W=50" for the initial correlation estimation. In Fig. 7 the k-frame of shots candidate to be "Anchor" are shown. They satisfy both the dissimilarity measure distribution and temporal distribution test.

Fig. 7. Shot identified as "Anchor" after the anchorperson first identification step.

In Fig. 8 we can see all the shots labeled as "Anchor" which are presented to the user for validation. To facilitate the validation these are ordered according to their probability to be "Anchor", starting from the minimum to the maximum. In other words at the beginning of the presentation shots more distant from the initial candidate are displayed, as their number is expected to be relatively small.

In the final two stages, the scenes are identified and classified and the results are presented in Fig. 9. An "Outdoor" scene is described by the "Anchor" shot and the first shot of the story unit while an "Indoor" scene is described again by the "Anchor" plus the k-frame of the most frequent shot occurring in the dialog.

6 Concluding Remarks

In this work an interactive tool for scenes detection and classification in news programs has been presented. The proposed techniques starting from a video sequence identify the temporal boundary of each scene present in the program and label this group of shots as "Indoor" if the scene represents a dialog between the Anchor man and guests and as "Outdoor" otherwise.

Fig. 8. Shot identified as "Anchor" after the anchorperson final identification step.

Fig. 9. Example of the final output results.

A real example has been presented where a news program has been segmented into scenes and each of those has been automatically classified in one of the two considered class "In door" and "Out door". The proposed technique outperforms most of the other approaches in terms of segmentation and classification while enabling a manual control of the intermediate and final results. This technique can be useful for semi automatic news program indexing while reducing drastically the amount of time required with respect to a fully manual scene segmentation. It can be expected that the results can be improved by using also audio information. For example for each shot a GMM model could be trained on selected audio features of the associated audio signal. Then all GMMs could be used in a similar way as this has been performed for the codebook descriptor so as to establish correlations among shots.

References

1. N. Adami and R. Leonardi. Identification of editing effect in image sequences by statistical modelling. Portland, Oregon, April 21-23 1999. PCS99 Picture Coding Symposium.
2. N. Adami and R. Leonardi. Evaluation of different descriptors for identifying similar video schot. Tokyo, Japan, August 22-25 2001. Submitted to ICME 2001.
3. MPEG7 Requirement Group. Introduction to mpeg7 iso/iec jtc1/sc29/wg11 n3751. La Baule, France, October 2000.
4. Z. Liu and Q. Huang. A new distance measure for probability distribution function of mixture type. Istanbul, Turkey, June 5-9 2000. ICASSP-2000.
5. P. Pala M. Bertini, A. Del Bimbo. Content based annotation and retrieval of news videos. New York City, NY USA, 30 July - 2 August 2000. ICME 2000.
6. Y. Nakamura and T. Kanade. Semantic analisys for video contents extraction - sspotting by association in news video. Seattle, WA USA, November 1997. ACM Multimedia Conference.
7. C. Saraceno and R. Leonardi. Identification of story units in audio-visual sequences by joint audio and video processing. Chicago-IL, USA, October 1998. Proc. International Conference on Image Processing.
8. N. Tsapasoulis Y. Avrithis and S. Kollias. Broadcast news parsing using visual cues: A robust face detection approach. New York City, NY USA, 30 July - 2 August 2000. ICME 2000.