Video Content Representation using Optimal Extraction of Frames and Scenes

Nikolaos D. Doulamis, Anastasios D. Doulamis, Yannis S. Avrithis and Stefanos D. Kollias
National Technical University of Athens
Department of Electrical and Computer Engineering
9, Heroon Polytechniou str., 157 73 Zografou, Athens, Greece
e-mail: ndoulam@image.ntua.gr

Abstract

In this paper, an efficient video content representation is proposed using optimal extraction of characteristic frames and scenes. This representation, apart from providing browsing capabilities to digital video databases, also allows more efficient content-based queries and indexing. For performing the frame/scene extraction, a feature vector formulation of the images is proposed based on color and motion segmentation. Then, the scene selection is accomplished by clustering similar scenes based on a distortion criterion. Frame selection is performed using an optimization method for locating a set of minimally correlated feature vectors.

1. Introduction

The rapid development of video and multimedia applications has enabled users to handle large amounts of visual information. However, tools and algorithms for effective organization and management of video databases and for content-based search and retrieval are still limited. For this reason, a new standardization phase is currently in progress by the MPEG group in order to develop algorithms for audiovisual coding (MPEG-4) [7] and content-based video storage, retrieval and indexing in multimedia applications, based on object extraction from scenes (MPEG-7) [8][10]. In the context of this paper, we present an efficient video content representation using optimal extraction of characteristic frames and scenes of video sequences. This representation, apart from providing browsing capabilities to digital video databases, also allows content-based queries and indexing to be performed more efficiently.

Several approaches for indexing and retrieval from video sequences have been proposed in the recent literature. In [5] a framework which enables content-based retrieval of video sequences using motion and texture cues has been proposed. Another approach dealing with indexing and retrieval using relevance feedback of mars was presented in [9]. In [11] it is developed a method for building an image representation using library basis elements that are facilitated by a joint adaptive space and frequency graph. Automatic indexing of TV news recordings has been analyzed in [6] where shots containing persons are identified and news items are recovered. An approach for automatic video segmentation and content-based retrieval based on a temporally windowed principal component analysis of a subsampled version of the video sequence is presented in [4].

The above techniques, either exploit color, motion or texture information in order to provide content-based query capabilities or use eigenvalue decomposition to reduce the image dimension. Our approach is oriented to extracting a small amount of information which is sufficient to provide a meaningful representation of a video sequence. This approach not only provides a more efficient way for video indexing, but also results in reducing the storage requirements and thus permits easy management of multimedia databases.

In one of our earlier works [1], an integrated framework for automatic extraction of characteristic frames has been proposed. The extraction mechanism was based on time variations of the frame feature vectors, generated using color and motion segmentation. However, since similar frames may be characterized by different segments, the overall procedure was rather sensitive and heavily dependent on the adopted segmentation algorithm. In this paper, the frame selection mechanism is enhanced by introducing an optimization method for locating a set of minimally correlated feature vectors. Furthermore, a scene selection mechanism is proposed, based on minimization of a distortion criterion for clustering the scene feature vectors.

2. Feature extraction

The feature vector extraction procedure is performed in a way similar to [1] and is briefly discussed in the sequel. The scene and frame selection mechanisms are then described, and experimental results are presented.

2.1. Scene cut detection

The first stage of the feature extraction procedure includes a scene cut detection technique, in order to locate the main shots of a video stream. Since visual content is

typically stored in MPEG compressed format, it is preferable to perform the feature extraction directly in the compressed domain. As a result in our approach scene cut detection is achieved by computing the sum of the block motion estimation error over each frame and detect frames for which this sum exceeds a certain threshold [1].

2.2. Color and motion segmentation

Color and motion segmentation provide a powerful representation of each video frame, more oriented to the human perception. In general, the number, size and location of objects as well as their color, motion, or texture characteristics give more meaningful information for an image than raw pixels. Thus, a color and motion segmentation technique is applied to each video frame. Block resolution has been adopted both for reducing the required computational time and exploiting information which already exists in the MPEG coding standard. To avoid oversegmentation problems, we have proposed a hierarchical block-based segmentation algorithm described in [1]. Apart from information provided by color or motion segmentation other features are also included in the feature vector, such as information of color and motion histograms or appropriate ac coefficients of the DCT transform.

2.3. Feature vector formulation

A multidimensional feature vector is generated for each frame by transforming the image domain to another domain (the feature one), more efficient for video content description. Color/motion segment properties cannot be directly used as feature vector elements since their size is different for each frame. To overcome this problem and to achieve better feature representation fuzzy classification of the extracted properties is performed as described in [1]. Finally, based on the feature vectors of all frames within a scene, a multidimensional scene feature vector is constructed, describing the average frame properties of the scene.

3. Scene selection mechanism

Based on scene feature vectors, an optimal extraction of the most characteristic scenes is performed. This is accomplished by clustering similar scene feature vectors and selecting a limited number of cluster representatives. Let $\mathbf{s}_i \in \Re^M$, $i = 1,2,...,N_S$ be the scene feature vector for the *i*-th scene, where N_S is the total number of scenes. Then $S = \{\mathbf{s}_i, i = 1,2,\ldots,N_S\}$ is the set of all scene feature vectors. Let also K_S be the number of scenes to be selected and $\mathbf{c}_i$, $i = 1,2,...,K_S$ the feature vectors which best represent those scenes. For each $\mathbf{c}_i$, an influence set is formed which contains all scene feature vectors $\mathbf{s} \in S$ which are closer to $\mathbf{c}_i$:

$$Z_i = \{\mathbf{s} \in S : d(\mathbf{s}, \mathbf{c}_i) < d(\mathbf{s}, \mathbf{c}_j)\ \forall\ j \neq i\} \quad (1)$$

where $d(\cdot)$ denotes the distance between two vectors. A common choice for $d(\cdot)$ is the Euclidean norm. In effect, the set of all Z_i defines a partition of *S* into clusters of similar scenes which are represented by the feature vectors $\mathbf{c}_i$. Then the average distortion, defined as

$$D(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{K_S}) = \sum_{i=1}^{K_S} \sum_{\mathbf{s} \in Z_i} d(\mathbf{s}, \mathbf{c}_i) \quad (2)$$

is a performance measure of the representation of scene feature vectors by the cluster centers $\mathbf{c}_i$. The optimal vectors $\hat{\mathbf{c}}_i$ are thus calculated by minimizing *D*:

$$(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, ..., \hat{\mathbf{c}}_{K_S}) = \underset{c_1, c_2, ..., c_{K_S} \in \Re^M}{\arg\min}\ D(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{K_S}) \quad (3)$$

Direct minimization of the previous equation is a tedious task since the unknown parameters are involved both in distances $d(\cdot)$ and influence zones. For this reason, minimization is performed in an iterative way using the generalized Lloyd or *K-means* algorithm [2]. Starting from arbitrary initial values $\mathbf{c}_i(0)$, $i = 1,2,...,K_S$, the new centers are calculated through the following equations for $n \geq 0$:

$$Z_i(n) = \{\mathbf{s} \in S : d(\mathbf{s}, \mathbf{c}_i(n)) < d(\mathbf{s}, \mathbf{c}_j(n))\ \forall\ j \neq i\} \quad (4)$$

$$\mathbf{c}_i(n+1) = cent(Z_i(n)) \quad (5)$$

where $\mathbf{c}_i(n)$ denotes the *i*-th center at the *n*-th iteration, and $Z_i(n)$ its influence set. The center of $Z_i(n)$ is estimated by the function

$$cent(Z_i(n)) = \frac{1}{|Z_i(n)|} \sum_{\mathbf{s}_i \in Z_i(n)} \mathbf{s}_i \quad (6)$$

where $|Z_i(n)|$ denote the cardinality of $Z_i(n)$. The algorithm converges to the solution $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, ..., \hat{\mathbf{c}}_{K_S})$ after a small number of iterations.

Finally, the K_S most representative scenes are extracted as the ones whose feature vectors are closest to $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, ..., \hat{\mathbf{c}}_{K_S})$:

$$\hat{\mathbf{s}}_i = \underset{\mathbf{s} \in S}{\arg\min}\ d(\mathbf{s}, \hat{c}_i), \quad i = 1,2,\ldots,K_S \quad (7)$$

4. Frame selection mechanism

After extracting the most representative scenes, the next step is to select the most characteristic frames within each one of the selected scenes. This is achieved by

minimizing a correlation criterion, so that the selected frames are not similar to each other. In particular, the most characteristic frames are selected as the ones with the minimum correlation among them. The selection could also be performed using the previous optimization technique. However, that approach does not exploit the temporal relation of feature vectors, which is significant for the frame selection procedure, as it is described in the sequel.

Let us denote by $\mathbf{f}_i \in \Re^M$, $i \in V = \{1,\ldots,N_F\}$ the feature vector of the *i*-th frame, where $N_F = 2^L$ is the total number of frames of a scene, and suppose that the K_F most characteristic ones should be selected. The correlation coefficient of the feature vectors $\mathbf{f}_i, \mathbf{f}_j$ is defined as

$$\rho_{ij} = C_{ij} / (\sigma_i \sigma_j) \tag{8}$$

where $C_{ij} = (\mathbf{f}_i - \mathbf{m})^T (\mathbf{f}_j - \mathbf{m})$ is the covariance of the two vectors, $\mathbf{m} = \sum_{i=1}^{N_F} \mathbf{f}_i / N_F$ is the average feature vector of the scene and $\sigma_i^2 = C_{ii}$ is the variance of $\mathbf{f}_i$. In order to define a measure of correlation between K_F feature vectors, we first define the *index* vector $\mathbf{x} = (x_1,\ldots,x_{K_F}) \in W \subset V^{K_F}$ where

$$W = \{(x_1,\ldots,x_{K_F}) \in V^{K_F} : x_1 < \cdots < x_{K_F}\} \tag{9}$$

is the subset of V^{K_F} which contains all sorted index vectors $\mathbf{x}$. Thus, each index vector $\mathbf{x} = (x_1,\ldots,x_{K_F})$ corresponds to a set of frame numbers. The correlation measure of the feature vectors $\mathbf{f}_i$, $i = x_1,\ldots,x_{K_F}$ is then defined as

$$R(\mathbf{x}) = R(x_1,\ldots,x_{K_F}) = \left(\sum_{i=1}^{K_F-1} \sum_{j=i+1}^{K_F} (\rho_{x_i,x_j})^2 \right)^{1/2} \tag{10}$$

Based on the above definitions, it is clear that searching for a set of K_F minimally correlated feature vectors is equivalent to searching for an index vector $\mathbf{x}$ that minimizes $R(\mathbf{x})$. Searching is limited in the subset *W*, since index vectors are used to construct sets of feature vectors. Therefore any permutations of the elements of $\mathbf{x}$ will result in the same sets. The set of the K_F least correlated feature vectors, corresponding to the K_F most characteristic frames, is thus represented by

$$\hat{\mathbf{x}} = (\hat{x}_1,\ldots,\hat{x}_{N_F}) = \arg\min_{\mathbf{x} \in W} R(\mathbf{x}) \tag{11}$$

Unfortunately, the complexity of an exhaustive search for the minimum value of $R(\mathbf{x})$ is such that a direct implementation would be practically unfeasible, since the multidimensional space *W* includes all possible sets (combinations) of frames. A dramatic reduction in complexity is achieved, however, through *logarithmic search*, which is performed in a way similar to the search for block motion estimation in video sequences [3]. The algorithm is described as follows:

By letting $\mu = 2^{L-1} - 1$, we define the *initial index* $\mathbf{x}_0 \in W$ as the element of *W* which is closest to the *middle point* $\tilde{\mathbf{x}}_0 = (\mu,\ldots,\mu)$. It can be shown that

$$\mathbf{x}_0 = (\mu - \lfloor K_F/2 \rfloor,\ldots,\mu-1,\mu+1,\ldots,\mu+\lfloor K_F/2 \rfloor) \tag{12}$$

if K_F is even, and

$$\mathbf{x}_0 = (\mu - \lfloor K_F/2 \rfloor,\ldots,\mu-1,\mu,\mu+1,\ldots,\mu+\lfloor K_F/2 \rfloor) \tag{13}$$

if K_F is odd, where $\lfloor \cdot \rfloor$ denotes integer part. The *neighborhood* of $\mathbf{x}_0$ is defined as

$$N(\mathbf{x}_0, S_0) = \{\mathbf{x} \in W : \mathbf{x} = \mathbf{x}_0 + S_0 \mathbf{p},\ \mathbf{p} \in G^{K_F}\} \tag{14}$$

where $S_0 = 2^{L-2} = N_F/4$ is the initial *step size* and $G = \{-1,0,1\}$. Based on these definitions, we calculate the next index vector $\mathbf{x}_1 = \arg\min_{\mathbf{x} \in N(\mathbf{x}_0, S_0)} R(\mathbf{x})$. By letting $S_1 = S_0/2$, we repeat the same steps:

$$\mathbf{x}_n = \arg\min_{\mathbf{x} \in N(\mathbf{x}_{n-1}, S_{n-1})} R(\mathbf{x}), \quad S_n = S_{n-1}/2 \tag{15}$$

for $n = 1,\ldots,L-2$ (until $S_n = 1$) and get the final result $\hat{\mathbf{x}} = \mathbf{x}_{L-2}$.

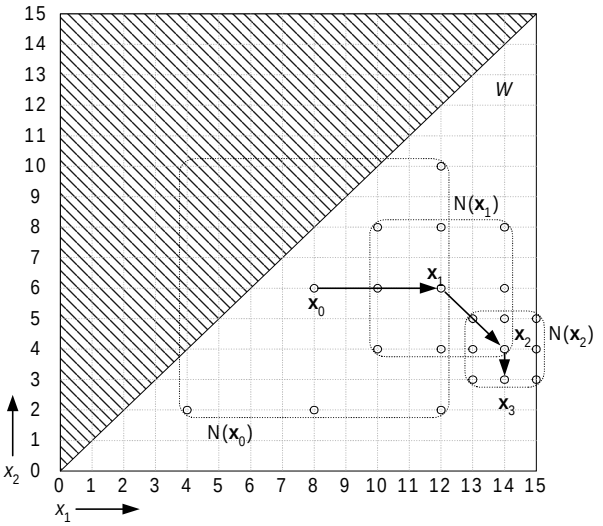

Figure 1. Illustration of the logarithmic search procedure for the simple 2-dimensional case of $K_F = 2$.

The algorithm is based on the assumption that frames which are close to each other (in time) should have similar properties, and therefore indices which are close to each other (in *W*) should have similar correlation measures. However, the technique performs equally well even in the case of random feature vectors, as shown by experiments.

The overall procedure for the very simple case of $K_F = 2$ is illustrated in Figure 1.

Figure 2. The 20 scenes of the test video sequence.

5. Experimental results.

The proposed algorithms were integrated into a system that was tested using several video sequences from video databases. The results obtained from a TV news reporting sequence of total duration 2.5 minutes (3750 frames) are presented in the following figures. The sequence was first partitioned in 20 scenes and then the frame and scene feature vectors were extracted using the aforementioned methodology. Figure 2 illustrates for each scene the frame whose feature vector is closest to the respective scene feature vector. We have chosen to keep four representative scenes ($K_S = 4$), and thus four scene clusters are generated. Each cluster contains the scenes whose feature vectors were closest to respective cluster center. The results of the scene selection mechanism are shown in Figure 3, where the representative scene of each cluster is shown with black border. It is clear that the four selected scenes give a meaningful representation of the content of the whole video sequence. Furthermore, it can be seen that each cluster contains scenes with similar properties, such as number and complexity of objects.

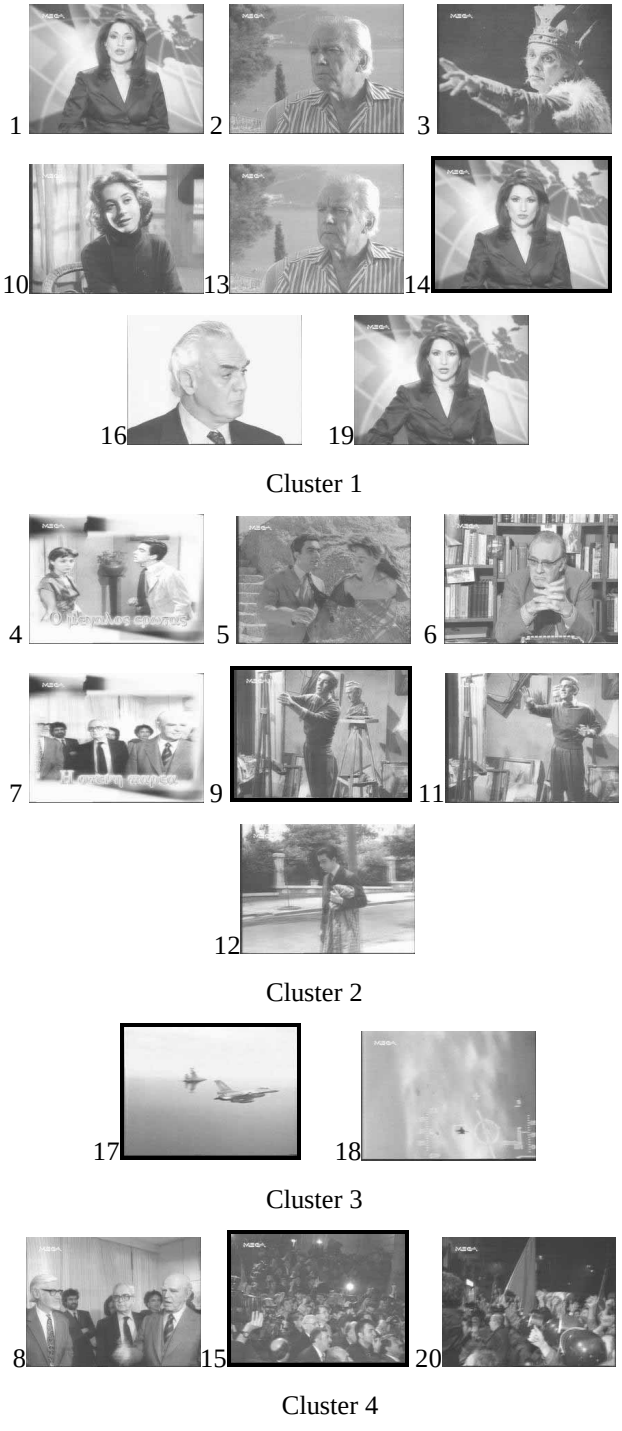

Figure 3. The four scene clusters generated by the scene selection mechanism. The respective selected (representative) scenes are shown with black border.

The frame selection mechanism was tested with the last scene of cluster 2 (scene 12). Four frames were extracted out of a total of 255 frames, using logarithmic search with $K_F = 4$, and the representative frames are shown in Figure 4. Although a very small percentage of frames is retained, one can perceive the content of scene by just examining the four selected frames. The correlation measure $R(\mathbf{x})$ was also tested using a large number of random index vectors and its probability density function (histogram) is depicted in Figure 5. Although our search algorithm requires about 1% of the computational time of the random search, the located minimum value of $R(\mathbf{x})$ was indeed very close to the actual minimum, as shown by the vertical dashed line of Figure 5.

Figure 4. The four selected frames of scene 12.

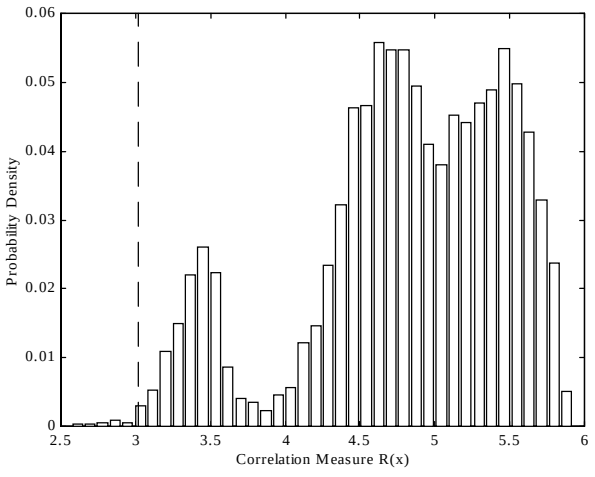

Figure 5. The probability density function of the correlation measure $R(\mathbf{x})$. The vertical dashed line shows the minimum value located by the logarithmic search algorithm

6. Conclusions

In this paper, a mechanism for automatic extraction of the most representative scenes and frames in video databases was proposed. The scheme includes on the one hand, a minimization of a distortion criterion and one the other, an optimization technique for indicating the indices of the most characteristic frames within each selected scene. To accomplish the optimal extraction, we first applied a color or motion segmentation technique to video frames in order to obtain an image representation more suitable for classification. Furthermore, to make the proposed architecture more robust, a fuzzy representation of the feature vectors was introduced. Experimental results indicating the good performance of the proposed scheme were provided by examining real TV programs.

7. References

[1] A. Doulamis, Y. Avrithis, N. Doulamis and S. Kollias, "Indexing and Retrieval of the Most Characteristic Frames/Scenes," *Workshop on Image Analysis for Multimedia Interactive Systems, pp. 105-110,* Louvain-la-Neuve, Belgium, 1997.

[2] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1993.

[3] H. Gharaviand and M. Mills, "Block-Matching Motion Estimation Algorithms: New Results," *IEEE Trans. on Circ. & Systs*. vol. 37, pp. 649-651, 1990.

[4] K. J. Han and A. H. Tewfik, "Eigen-Image Video Segmentation and Indexing," *Proc. of IEEE ICIP*, pp. 538-541, Santa Barbara, USA, Oct. 1997.

[5] G. Iyerngar and A.B. Lippman, "Videobook: An Experiment in Characterization of Video," *Proc. of IEEE ICIP*, pp. 855-858, Lausanne Switzerland Sept. 1996.

[6] B. Merialdo, "Automatic Indexing of TV News," *Workshop on Image Analysis for Multimedia Interactive Systems*, pp. 99-104, Louvain-la-Neuve, Belgium, 1997.

[7] MPEG Video Group, "MPEG-4 Requirements," *ISO/IEC GTC1/SC29/WG11 N1679*, Bristol MPEG Meeting, April 1997.

[8] MPEG Video Group, "MPEG-7: Context and Objectives (v.3)," *ISO/IEC GTC1/SC29/WG11 N1678*, Bristol MPEG Meeting, April 1997.

[9] Y. Rui, T.S. Huang and S. Mehrotra, "Content-based Image Retrieval with Relevance Feedback in Mars," *Proc. of IEEE Inter. Conf. on Image Processing (ICIP)*, vol. 2, pp. 825-818, Santa Barbara USA, October 1997.

[10] Y. Rui, T. S. Huang and S.-F. Chang, "Digital Image/Video Library and MPEG-7: Standardization and Research Issues," *Proc. of IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol.6, pp. 3785-3788, Seattle USA, May 1998.

[11] J. R. Smith and S.-F. Chang, "Joint Adaptive Space and Frequent Basis Selection," *Proc. of IEEE Inter. Conf. on Image Processing (ICIP)*, vol. 3, pp. 702-705, Santa Barbara USA, October 1997.