A STOCHASTIC FRAMEWORK FOR OPTIMAL KEY FRAME EXTRACTION FROM MPEG VIDEO DATABASES

N. D. Doulamis, A. D. Doulamis, Y. Avrithis, S. D. Kollias
Dept.of Electrical and Computer Engineering
National Technical University of Athens
Athnens, Greece

Abstract - A framework for video content representation is proposed in this paper for extracting limited, but meaningful, information of video data directly from MPEG compressed domain. First, the traditional frame-based representation is transformed to a feature-based one. Then, all features are gathered together using a fuzzy formulation and extraction of several key frames is performed for each shot in a content-based rate sampling framework. In particular, our approach is based on minimization of a cross-correlation criterion among video frames of a given shot so as to be located a set of minimally correlated feature vectors. Experimental results indicating the good performance of the proposed scheme are also presented.

INTRODUCTION

Traditionally, video is represented by numerous consecutive frames, each of which corresponds to a constant time interval. However, such a representation is not adequate for the new emerging multimedia applications, such as content-based indexing, retrieval and video browsing. For this reason new methods for efficient video content representation should also be implemented. In particular, a "pre-indexing" stage should be introduced, extracting limited and meaningful information of the video content. The objective is to divide a video sequence into separate representative shots, i.e., scenes, and then extract the most characteristic frames (key frames) within the selected shots [2], [3], [7].

The first approaches [6], [10] in this research area are oriented to detecting shot changes; they can, therefore, be used as the first stage of video visualization algorithms. Exploitation of shot information by selecting one key frame for each shot has been presented in [1], [8]. However, a single key frame cannot provide sufficient information about the video content. Recently some other approaches dealing with construction of a compact image map or image mosaics have been described in [5], [9]. Although such a representation can be very good for specific applications, it cannot be effectively implemented in real world complex shots. A method for analyzing video and building a pictorial summary for visual representation has been proposed in [11]. This work is mainly concentrated on dividing a video into consecutive meaningful segments (story units), instead of extracting key frames.

0-7803-5610-1/99/$10.00 © 1999

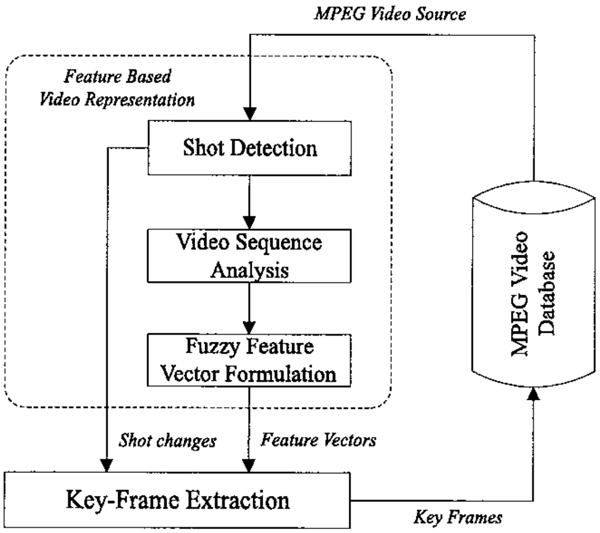


Figure 1: Block diagram of the proposed architecture.

In this paper, extraction of several key frames within a shot is proposed for efficiently describing the shot content. Firstly, video frame-based representation is transformed into a feature-based one, by applying several image processing and analysis techniques to each video frame. To reduce the required computations and simultaneously exploit information existing in MPEG video databases, such as block color average and motion vectors, our analysis is performed directly on the MPEG compressed domain. Based on the feature-based video representation, key frames are extracted by minimizing a cross-correlation criterion. A stochastic framework has been adopted for efficient estimation of those frames, which optimally represent, in the sense of cross-correlation, the content of a video shot.

FEATURE-BASED VIDEO REPRESENTATION

A block diagram of the proposed architecture is illustrated in Figure 1, consisting of mainly four modules; shot cut detection, video sequence analysis, fuzzy classification and key frame extraction. The first three modules, produce a feature vector representation of the video sequence by first segmenting it into distinct video shots and then applying image analysis techniques to the frames of each shot. Such a representation provides a more meaningful description of the video content.

Scene Cut Detection: A shot detection algorithm is first applied in order to temporally segment the sequence into shots. Several algorithms have been reported in the literature for shot change detection [6], [10]. In our approach the algorithm proposed in [10] has been adopted since it is based on the dc coefficients of the DCT transform of each frame.

Video Sequence Analysis: The next step of the analysis is segmentation of each shot into video objects and extraction of essential information describing those objects. In this paper, color and motion segmentation is applied to the image sequence representing each video shot. In particular, the number, size, location and average color (motion) components of all color (motion) segments are used for the construction of a color (motion) feature vector. To avoid the existence of small objects and also to accelerate the required time, a hierarchical color/ motion segmentation scheme have been proposed as it is described in [4].

Fuzzy Feature Vector Formulation: All features extracted by the video sequence analysis module can be used to describe the visual content of each video frame. However, they are not directly included in a vector to be used for this purpose, since their size differs between frames. To overcome this problem, we classify color as well as motion segments into pre-determined classes, forming a multidimensional histogram and then we assign a degree of membership to each class.

KEY-FRAME EXTRACTION

Key-frame extraction is achieved by minimizing a correlation criterion, so that the selected frames are not similar to each other.

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 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

$$\rho_{ij} = C_{ij}/(\sigma_i \sigma_j) \tag{1}$$

where $C_{ij} = (\mathbf{f}_i - \mu)^T(\mathbf{f}_i - \mu)$ is the covariance of the two vectors, $\mu = \sum_{i=1}^{N_F} \mathbf{f}_i / N_F$ is the average feature vector of the scene and $\sigma_i^2 = C_{ij}$ 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

$$W = \{(x_1, \ldots, x_{K_F}) \in V^{K_F} : x_1 < \cdots < x_{K_F}\} \tag{2}$$

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

$$R(\mathbf{x}) = R(x_1, \ldots, x_{K_F}) = \left(\sum_{i=1}^{K_F - 1} \sum_{j=i+1}^{K_F} \rho^2_{x_i, x_j} \right)^{\frac{1}{2}} \tag{3}$$

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 $R(\mathbf{x})$. Any permutations of the elements of $\mathbf{x}$ will result in the same sets. The set of the K_F least correlated vectors, is represented by

$$\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_{N_F}) = \arg\min_{\mathbf{x} \in w} R(\mathbf{x}) \tag{4}$$

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 W includes all possible combinations of frames. For that reason a logarithmic search has been proposed in [4]. The algorithm is restricted to the special case $N_F = 2^G$ and its implementation includes the definition of an initial step size $\delta(0) = s^{G-2} = N_F/4$ and an initial index $\mathbf{x}(0) \in W$ as the element of W which is closest to the middle point $\tilde{\mathbf{x}}_0 = (\mu, \ldots, \mu)$, where $\mu = 2^{G-1} - 1$. Successive index vector estimates are then obtained by the recursive equations

$$\mathbf{x}(n+1) = \arg \min_{\mathbf{x} \in N(x(n), \delta(n))} R(\mathbf{x}), \delta(n+1) = \delta(n)/2 \tag{5}$$

where the neighborhood $N(\mathbf{x}, \delta)$ of $\mathbf{x}$ is a small set of index vectors whose distance from $\mathbf{x}$ is δ. The final result $\hat{\mathbf{x}} = \mathbf{x}(G-1)$ is obtained by applying the above recursion for $n = 0, \ldots, G-2$ (until $\delta(n) = 1$). The algorithm, whose implementation details are fully described in [4], provides a very fast convergence to a sub-optimal solution. To make the algorithm more flexible, a stochastic approach is proposed in this paper, so that the optimal solution is obtained.

Stochastic Approach

The concept of this stochastic approach is to assign a probability to every neighbor point of the current examined point , i.e., every point belonging to the set and then to select the next index vector using the assigned probabilities; these probabilities are inversely proportional to the respective correlation measure. The search procedure is repeated several times, so that in effect multiple logarithmic search experiments take place in a random way. Due to the stochastic behavior of the algorithm, different neighbors are selected in every new experiment, resulting in the generation several random paths.

Let us denote by $\mathbf{x}_m(n)$ the index vector at the n-th iteration step for the m-th experiment, and by $\mathbf{y}_i$, $i = 1, \ldots, |N|$ its neighbors, i.e., the elements of set $N(\mathbf{x}_m(n), \delta(n))$, where $|N|$ is the set cardinality. Then, a probability value is assigned to each $\mathbf{y}_i$ according to the respective correlation measure as follows

$$p_i = 1 - R(\mathbf{y}_i) / \sum_{j=1}^{|N|} R(\mathbf{y}_j), i = 1, \ldots, |N| \tag{6}$$

A cumulative probability function is then constructed for all $\mathbf{y}_i$, i.e., $q_i = \sum_{j=1}^{i} p_j$, $i = 1, \ldots, |N|$, with $q_0 = 0$. Using a given random number r, uniformly distributed in the range [0,1], the next index vector $\mathbf{x}_m(n+1)$ is chosen among $\mathbf{y}_i$ as follows

$$\mathbf{x}_m(n+1) = \{\mathbf{y}_i \in N(\mathbf{x}_m(n), \delta(n)) : q_{i-1} < r \leq q_i\} \tag{7}$$

The iteration is repeated $n = 0, 1, \ldots, M-2$ times, as in the case of the logarithmic search algorithm, and the result of the m-th experiment is the index vector $\hat{\mathbf{x}}_m = \arg\min_{i=0,\ldots,M-1} R(\mathbf{x}_m(i))$ corresponding to the minimum correlation measure along the path of the experiment. The final result is the index vector corresponding to the minimum correlation measure of all vectors in all experiments. After J experiments, the optimal solution $\hat{\mathbf{x}} = \arg\min_{m=1,\ldots,J} R(\hat{\mathbf{x}}_m)$ is selected, containing the indices of the K_F key frames.

EXPERIMENTAL RESULTS

An MPEG video database consisting of real life video sequences is used in the following to test the performance of the proposed algorithm. The feature domains are partitioned in $Q = 3$ classes.

One shot of the database are used for demonstration of the performance of the proposed techniques. The shot, coming from a test drive sequence and consisting of $N_F = 203$ frames, is illustrated in Figure 2. One every 10 frames is depicted, resulting in 20 frame thumbnails.

The results of the proposed method for key-frame extraction are shown in Figure 3. The four selected key frames of the given shot are shown in the Figures 3(a,b) for the logarithmic and stochastic approach respectively. Although a very small percentage of frames is retained ($K_F = 4$), it is clear that, in all cases, one can visualize the content of the shot by just examining the four selected frames. Although a comparison of the two algorithms is rather subjective, it can still be argued that key frames selected by the the proposed method are more representative of the shot than that of the logarithmic one.

References

[1] F. Arman, R. Depommier, A. Hsu and M. Y. Chiu, "Content-Based Browsing of video Sequences," *ACM Multimedia*, pp. 77- 103, Aug. 1994.

[2] Y. Avrithis, N. Doulamis, A. Doulamis and S. Kollias, "Efficient Content Representation In MPEG Video Databases," *Proc. of IEEE workshop on Content-Based Access of Image and Video Libraries (CBAIVL)*, pp. 91- 94, Santa Barbara, USA, June 1998.

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

[4] N. Doulamis, A. Doulamis, Y. Avrithis and S. Kollias, "Video Content Representation using Optimal Extraction of Frames and Scenes," *Proc. of IEEE Conference on Image Processing (ICIP)*, Chicago USA, Oct. 1998.

[5] M. Irani and P. Anandan,"Video Indexing Based on Mosaic Representation," *Proc. of the IEEE*, vol. 86, no. 5., pp. 805- 921, May 1998.

[6] N. V. Patel and I. K. Sethi, "Video Shot Detection and Characterization for Video Databases," *Pattern Recognition*, vol. 30, no. 4, pp. 583- 592, April 1997.

[7] B. Shahraray, "Scene Change Detection and Content-Based Sampling of Video Sequences," *Proc. of SPIE 2419: Digital Video Compression: Algorithms and Technologies*, pp. 2- 13, Feb. 1995.

[8] S. W. Smoliar and H. J. Zhang, "Content-Based Video Indexing and Retrieval," *IEEE Multimedia*, pp. 62- 72, Summer 1994.

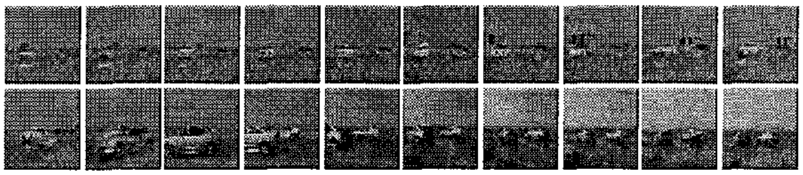
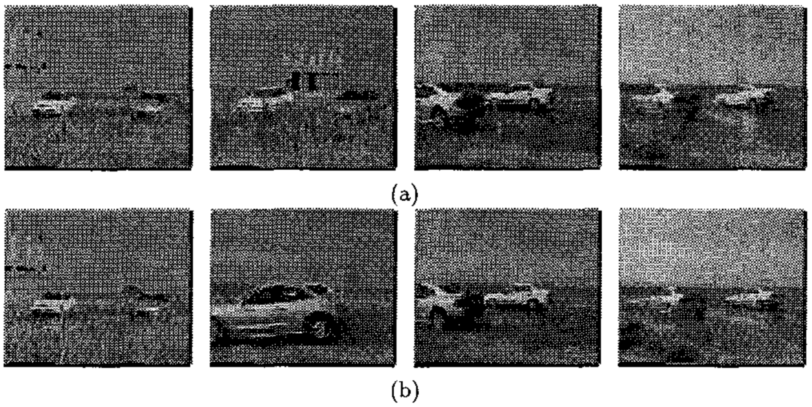


Figure 2: Test video sequence, frames #0 to #200.



(a)

(b)

Figure 3: Extracted key frames based (a) on logarithmic search, and (b) stochastic logarithmic

[9] N. Vasconcelos and A. Lippman, "A Spatiotemporal Motion Model for Video Summarization," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 361- 366, Santa Barbara, CA, June 1998.

[10] B. L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Videos," *IEEE Trans. Circ. and Syst. for Video Tech.*, vol. 5, pp. 533- 544, Dec. 1995.

[11] M. M. Yeung and B. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Trans. Circ. and Syst. for Video Tech.*, Vol. 7, No. 5, pp. 771- 785, Oct. 1997.