# An Optimized Key-Frames Extraction Scheme Based on SVD and Correlation Minimization

*Klimis S. Ntalianis and Stefanos D. Kollias*
National Technical University of Athens, Electrical and Computer Engineering Department, 15773, Athens, Greece
*E-mail: kntal@image.ntua.gr*

## Abstract

*In this paper an optimized and efficient technique for key-frames extraction of video sequences is proposed, which leads to selection of a meaningful set of video frames for each given shot. Initially for each frame the Singular Value Decomposition method is applied and a diagonal matrix is produced, containing the singular values of the frame. Afterwards, a feature vector is created for each frame, by gathering the respective singular values. Next all feature vectors of the shot are collected to form the feature vectors basin of this shot. Finally a genetic algorithm approach is proposed and applied to the vectors basin, for locating frames of minimally correlated feature vectors, which are selected as key-frames. Experimental results indicate the promising performance of the proposed scheme on real life video shots.*

*Keywords: key-frames extraction, Singular Value Decomposition, correlation criterion, genetic algorithm.*

## 1. Introduction

Recent progress in the fields of video analysis and processing has led to a significant increase in the amount of visual information being stored, accessed and transmitted. This tremendous increase has stimulated the invention of new technologies for efficient searching, indexing and content-based retrieving. Towards this direction several methods for key-frames extraction, video summarization, content personalization and relevance feedback have been proposed. This paper focuses on the technology of key-frames extraction, which aims at providing compact representations of video shots / sequences. Key-frames extraction algorithms usually result in the selection of a small but meaningful number of characteristic frames from a shot, enabling the fast access to frames, shots, or events in video sequences. For example, in case of a 30-min video stream consisting of approximately 200 shots, a user can locate video segments of interest or compare video shots by examining only 1,000 out of 45,000 frames, if five key-frames per shot are extracted (on average).

Recently, some approaches have been proposed for non-linear content representation. Shot-cut detection algorithms [1] can be considered as early attempts. In [2] frames of a video sequence are chosen at regular time intervals, leading to a storyboard presentation. Selection of a single key-frame for each shot has been presented in [3], while in [4] three-dimensional iconic cubes are constructed, which contain the representative frame of a shot, together with camera breaks and relative duration (depth of the cube). However both approaches cannot provide sufficient information about the video content, especially for shots of long duration and high motion activity. In [5] construction of compact image maps is employed, while in [6] the video shot content is represented using image mosaics. However, although such approaches can be efficient for specific applications, such as sports programs, studio productions and other cases with specific motion characteristics, they cannot provide satisfactory results in real world complex shots where background / foreground changes or complex camera effects usually appear. Video abstraction using unsupervised cluster-validity analysis has been reported in [7]. In [8] a method for analyzing video and building a pictorial summary has been presented, while in [9] a fuzzy visual content representation scheme has been proposed with application to video summarization and content based indexing and retrieval. Another scheme for video summarization of three-dimensional video sequences has been reported in [10], where content description is accomplished using global and object-based characteristics.

However, common problems of the aforementioned techniques are that advanced schemes have high complexity, while simpler schemes do not provide enough visual information of each shot. To overcome these drawbacks, in this paper an SVD-based generalized framework for non-linear representation of video shots is proposed, regardless of the scene complexity. Towards this direction a content-based sampling algorithm is used, which extracts multiple representative frames (key-frames) for a given shot. This approach leads to summarization of visual information, in similarity to current document search engines. Thus, it is possible to automatically generate low-resolution video clip previews (trailers) or still image tabloids.

For this purpose and since video sequences usually contain large amounts of mostly redundant data, the Singular Value Decomposition method is incorporated [11]. The SVD method helps removing redundancy while retaining as much information from the data as possible, as it has optimal decorrelation and subrank approximation properties. In particular initially the SVD method is applied on each frame of a shot and the singular values are gathered to form a feature vector that compactly describes the content of the frame. The set of all feature vectors of each shot constructs a feature vectors' basin and in this paper key-frames are extracted from each basin (and consequently from each shot). Towards this direction a cross correlation criterion that measures correlation among sets of feature vectors is formulated and key-frames are extracted by minimizing this criterion. However as the computational complexity of a full search can be extremely large, a genetic algorithm approach is proposed to carry out the minimization task. Experimental results show that meaningful sets of key-frames are extracted, which provide a rough approximation of the visual content of each shot.

## 2. SVD and Feature Vector Formulation

The SVD is closely linked to the concepts of principal component analysis (PCA) and Karhunen–Loeve transform (KLT) and the relationships among them are discussed in detail in [12], [13]. In the context of key-frames extraction SVD can be very efficient in providing compact and meaningful representation of frame content.

Towards this direction let us denote by $s_j$ the $j$-th shot of a video sequence and by $F_{i,j}$ the $i$-th frame of the $j$-th shot. Then the

singular value decomposition (SVD) of a real-valued $M \times N$ frame $F_{i,j}$, with $M \leq N$ can be written as

$$F_{i,j} = U S_{i,j} V^t \quad (1)$$

where, $U$ is an orthogonal $M \times M$ matrix whose columns (called the "left singular vectors") are the eigenvectors of $F_{i,j}F_{i,j}^t$, $V$ is an $N \times N$ matrix whose columns (the "right singular vectors") are eigenvectors of $F_{i,j}^t F_{i,j}$, and $S_{i,j}$ is the $M \times N$ diagonal matrix whose diagonal elements (the "singular values") are the square roots of the corresponding eigenvalues of $F_{i,j}F_{i,j}^t$, which are ordered in descending order

$$\sigma_{i,j}^1 \geq \sigma_{i,j}^2 \geq \ldots \geq \sigma_{i,j}^N \quad (2)$$

Letting $\hat{F}_{i,j} = U^t F_{i,j} = S_{i,j} V^t$, the SVD may also be written as $F_{i,j} = U\hat{F}_{i,j}$.

In the proposed scheme the SVD method is applied on each frame and the resulting eigenvalues are used for construction of the feature vectors. In particular feature vector $\mathbf{f}_{i,j}$ of frame $F_{i,j}$ is given by:

$$\mathbf{f}_{i,j} = diag(U^T F_{i,j} V) = [\sigma_{i,j}^1 \sigma_{i,j}^2 \ldots \sigma_{i,j}^N]^T \quad (3)$$

where $diag(\cdot)$ denotes a vector whose elements are the main diagonal elements of the argument. As it can be observed by Equation (3), for given matrices $U$ and $V$ both having orthogonal columns, a feature vector can be produced for each frame $F_{i,j}$ of a specific shot $s_j$. Gathering all these vectors for shot $s_j$ a feature vectors basin $B_j$ is produced:

$$B_j = \bigcup_{i=1}^{L} \mathbf{f}_{i,j}, \quad \mathbf{f}_{i,j} \in s_j \quad (4)$$

Now the problem of key-frames extraction reduces to selection of key-feature-vectors from each basin (each shot) and the proposed solution is described next.

## 3. Key-Frames Extraction

In the proposed scheme key-frames are extracted by minimizing a cross correlation criterion, using a genetic algorithm. Cross correlation formulation and genetic minimization are discussed in the following subsections.

### *3.1 Cross Correlation and Problem Formulation*

In this paper extraction of key-frames within each given shot is achieved by minimizing a correlation criterion, so that the selected key-frames do not contain similar visual content. In particular, selected key-frames are those with the minimum correlation among the frames of a shot. Let us denote by $\mathbf{f}_{i,j} \in R^N$, $i \in D = \{1, \ldots, N_F\}$ the feature vector of the $i$-th frame of a given shot $s_j$, where $N_F$ is the total number of frames of this shot and $N$ is the length of the feature vector. Let us also suppose that the $K_F$ most characteristic ones should be selected. $K_F$ can be provided by the user interactively (according to the complexity of the shot), can be a priori set, or can be automatically set, expressing the minimum correlation between frames. In the following for simplicity purposes and without loss of generality we assume that shot $s_j$ is selected and for each feature vector $\mathbf{f}$ the shot index $j$ is omitted. Then the correlation coefficient of the feature vectors $\mathbf{f}_i$, $\mathbf{f}_j$ is defined as $\rho_{ij} = C_{ij}/(\sigma_i \sigma_j)$ where $C_{ij} = (\mathbf{f}_i - \mathbf{m})^T(\mathbf{f}_i - \mathbf{m})$ 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 shot 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 $\mathbf{a} = (a_1, \ldots, a_{K_F}) \in X \subset D^{K_F}$ where

$$X = \{(a_1, \ldots, a_{K_F}) \in D^{K_F} : a_1 < \cdots < a_{K_F}\} \quad (5)$$

$X$ is the subset of $D^{K_F}$ that contains all sorted index vectors $\mathbf{a}$. Thus, each index vector $\mathbf{a} = (a_1, \ldots, a_{K_F})$ corresponds to a set of frame numbers. The *correlation measure* of the feature vectors $\mathbf{f}_i$, $i = a_1, \ldots, a_{K_F}$ is then defined as

$$R_F(\mathbf{a}) = R_F(a_1, \ldots, a_{K_F}) = \\ = \frac{2}{K_F(K_F - 1)} \sum_{i=1}^{K_F - 1} \sum_{j=i+1}^{K_F} (\rho_{a_i, a_j})^2 \quad (6)$$

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 $\mathbf{a}$ that minimizes $R_F(\mathbf{a})$. Searching is limited in the subset $X$, since index vectors are used in order to construct sets of feature vectors, therefore any permutations of the elements of $\mathbf{a}$ will result in the same sets. The set of the $K_F$ least correlated feature vectors, corresponding to the $K_F$ key-frames is thus represented by

$$\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_{N_F}) = \arg\min_{\mathbf{a} \in X} R_F(\mathbf{a}) \quad (7)$$

### *3.2 The Genetic Algorithm Approach*

In order to find the solution of the previous problem (Eq. (7)), all different combinations of feature vectors should be examined. Unfortunately, the complexity of an exhaustive search for the minimum value of $R_F(\mathbf{a})$ is such that a direct implementation would be practically infeasible, since the multidimensional space $X$ includes all possible sets (combinations) of frames. A dramatic reduction in complexity can be achieved through *logarithmic search* [14], which provides very fast convergence to sub-optimal solutions. However, since the search procedure is by definition confined to a very small, pre-defined subset of the search space $U$, there is always a significant possibility of converging to a local minimum of $R_F(\mathbf{a})$, resulting in poor performance. For this reason, a *genetic algorithm* (GA) [15] approach is adopted in this paper. This approach seems to be very efficient for the particular optimization problem, given the size and dimensionality of the search space and the multimodal nature of the objective function. Possible solutions of the optimization problem, i.e., sets of frames, are represented by chromosomes whose genetic material consists of frame numbers (indices). Chromosomes are thus represented by index vectors $\mathbf{a} = (a_1, \ldots, a_{K_F}) \in D^{K_F}$ following an integer number encoding scheme, that is, using integer numbers for the representation of genes $a_i \in D$, $i = 1, \ldots, K_F$.

An *initial population* of $P$ chromosomes, $\mathbf{A}(0) = (\mathbf{a}_1, \ldots, \mathbf{a}_P)$ is produced and used for the creation of new generation populations $\mathbf{A}(n)$, $n>0$. The creation of $\mathbf{A}(n)$ at generation (or GA cycle) $n$ is performed by applying a set of operations on population $\mathbf{A}(n\text{-}1)$, described below. This procedure is repeated until $\mathbf{A}(n)$ converges to an optimal solution. Traditionally initial populations are randomly generated, but in this paper a temporal variation approach is used, where the temporal relation of feature vectors is exploited, increasing the possibility of locating sets of feature vectors with small correlation within the first few GA cycles. According to this

temporal variation approach, sets of frames are selected whose feature vectors reside in extreme locations of the feature vector trajectory. This selection is accomplished by locating points where the magnitude of the second-order derivative of feature vector trajectory is locally maximized.

The correlation measure $R_F$ (**a**) is used as an *objective function* to estimate the performance of all chromosomes $\mathbf{a}_i$, $i = 1, \ldots, P$ in a given population. However, a *fitness function* is used to map objective values to fitness values, following a *linear normalization scheme*. In particular, chromosomes $\mathbf{a}_i$ are ranked in ascending order of $R_F$ ($\mathbf{a}_i$), since the objective function is to be minimized. Let $r(\mathbf{a}_i) \in \{1, \ldots, P\}$ be the rank of chromosome $\mathbf{a}_i$, $i = 1, \ldots, P$. Defining an arbitrary fitness value $f_B$ for the best chromosome, the fitness of the $i$-th chromosome is given by the linear function

$$f(\mathbf{a}_i) = f_B - [r(\mathbf{a}_i - 1)]f_D, \quad i = 1, \ldots, P \tag{8}$$

where $f_D$ is a decrement rate. Thus, the average objective value of the population is mapped into the average fitness [16]. After fitness values, $f(\mathbf{a}_i)$, $i = 1, \ldots, P$, have been calculated for all members of the current population, *parent selection* is then applied so that a fitter chromosome gives a higher number of offspring and thus has a higher chance of survival in the next generation. A *proportionate scheme*, implemented by the *roulette wheel selection* procedure [17] is used for parent selection, ensuring that each chromosome has a growth rate proportional to its fitness value.

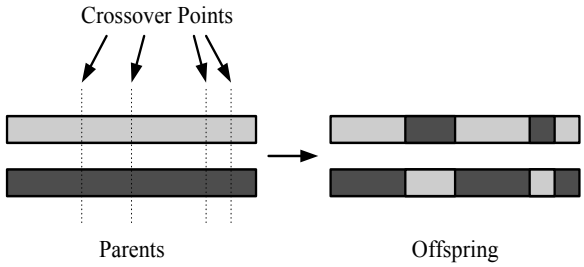


**Figure 1:** Example of the crossover operator with four crossover points.

A set of new chromosomes (offspring) is then produced by mating the selected parent chromosomes and applying a *crossover operator*. The genetic material of the parents is combined in a random way in order to produce the genetic material of the offspring. An example of the crossover operator with four crossover points used for exchanging genes is depicted in Figure 1. A generalized *uniform crossover* scheme is employed in the context of this paper, by considering each parent gene to be a potential crossover point. *Mutation* is then applied to the newly created chromosomes, introducing random gene variations that are useful for restoring lost genetic material, or for producing new material that corresponds to new search areas. In particular, each offspring gene $a_i$ is replaced by a randomly generated one $a_i \in D = \{1, \ldots, N_F\}$, if a probability test is passed. A small mutation probability ensures that only a small gene proportion is altered in each generation.

Once new chromosomes have been generated for a given population $\mathbf{A}(n)$, $n \geq 0$, the next generation population, $\mathbf{A}(n+1)$, is formed by inserting those new chromosomes into $\mathbf{A}(n)$ and deleting an appropriate number of older chromosomes, so that each population consists of $P$ members. The exact number of old chromosomes to be replaced by new ones defines the *replacement strategy* of the GA and greatly affects its convergence rate [15]. All steps of the above description refer to a simple GA cycle. Several cycles need to take place, that is, several generations $\mathbf{A}(n)$, $n > 0$ need to be produced until the population converges to an optimal solution. For this reason, the procedures of fitness evaluation, parent selection, crossover and mutation are repeated until a termination criterion is reached. Usually the GA terminates when the best chromosome fitness remains constant for a large number of generations, indicating that further optimization is unlikely.


**Figure 2:** One shot from a real-world sequence, consisting of 182 frames. One every 14 frames is depicted.

The above algorithm, as well as the logarithmic search algorithm, are 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 $X$) should have similar correlation measures. However, the proposed technique performs equally well even in the case of random feature vectors.

## 4. Experimental Results

In this section, the performance of the proposed key-frames extraction scheme is evaluated, using a real-world video sequence. As a result, very complicated content, with zooming, panning, complex camera effects and motion are encountered. Shot detection has been performed manually, while the SVD approach for feature extraction has been applied offline. After preprocessing, all information regarding shot change instances as well as feature vector representation of all frames is stored in a database and is readily available. Hence, key-frames extraction can be separately performed for each shot, using directly the feature vectors of all frames within the respective shot.

Afterwards one shot of the sequence is used for demonstration of the performance of the proposed technique. The shot depicts two persons walking through a crowd, consists of $N_F = 182$ frames and it is illustrated in Figure 2. For presentation purposes one every 14 frames is depicted, resulting in 14 frame thumbnails, so that an idea of the entire content is provided. Furthermore in order to determine $K_F$ (the number of key-frames to be extracted) for the cross-correlation minimization method, a temporal variation approach is adopted [10], where information of the trajectory formed by the vectors of all frames in the shot is exploited. Based on this approach, the number of key-frames is estimated to be $K_F = 4$ for the selected shot.

Next, the correlation minimization approach is activated. In particular the minimum value of the correlation measure (over the whole population) versus the cycle of the genetic algorithm is shown in Figure 3. As expected, $R_F$ (**a**) decreases as the GA cycle increases, until it reaches a minimum at generation 274. Since in the specific experiment half of the chromosomes are replaced by new ones at each generation, there are cases where all generated offspring have lower fitness than their parents. In these cases the value of the correlation measure remains at the same level, hence the "stepwise" appearance of the curve in Figure 3. Note that the step "width" generally increases with the GA cycle, since it is directly related to the probability of further optimization. The four extracted key-frames of the selected shot are shown in Figure 4. As it can be observed although a very small percentage of frames is retained (~2 %), it is clear that the selected four frames provide sufficient visualization of the total 182 frames, providing a meaningful representation of the content of the video shot.

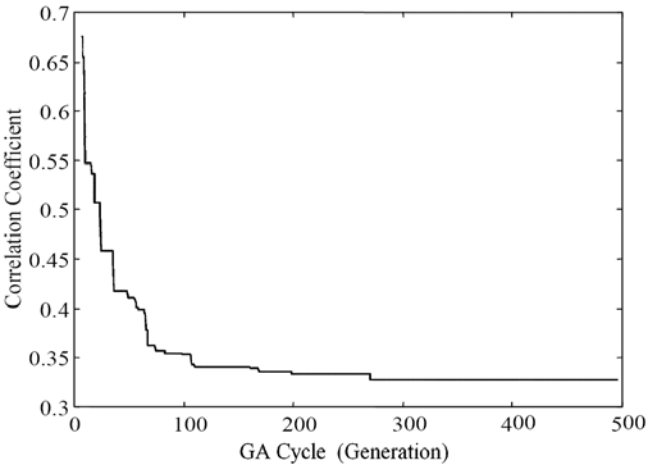


**Figure 3:** Minimum value of the correlation measure $R_F$(**a**) versus cycle of the genetic algorithm.

## 5. Conclusion

In this paper a novel content-based key-frames extraction scheme has been proposed. Since video sequences usually contain large amounts of mostly redundant data, the Singular Value Decomposition method is incorporated, which helps removing redundancy while retaining as much information of the data as possible. In particular initially the SVD method is applied on each frame of a shot and the singular values (eigenvalues) are gathered to form a feature vector that compactly describes frame content. Then, key-frames are optimally extracted for each shot, based on the minimization of an objective numerical criterion, i.e., the cross correlation function of frame feature vectors. Other criteria, which take into account human perception, can also be used by the proposed scheme. In this case extracted frames could be compared to those selected by several humans and examine which criterion is closer to human subjectivity. The formulated minimization problem is solved using a genetic algorithm approach.

In future works other ways of compactly and/or meaningfully representing visual information should also be examined. Furthermore other minimization methods could be incorporated to provide faster convergence to optimal solutions.


**Figure 4:** Extracted key-frames for the selected shot.

## 6. References


[1] N. V. Patel and I. K. Sethi, "Video Shot Detection and Characterization for Video Databases," *Pattern Recognition*, Vol. 30, pp. 583-592, 1997.

[2] M. Mills, J. Cohen, and Y. Y. Wong, "A magnifier tool for video data," in Proc. *ACM Computer Human Interface* (CHI), May 1992, pp. 93-98.

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

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

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

[6] N. Vasconcelos and A. Lippman, "A Spatiotemporal Motion Model for Video Summarization," *IEEE CVPR*, pp. 361-366, S. Barbara, 1998.

[7] Hanjalic and H. Zhang, "An integrated scheme for automated abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. on CSVT*, Vol. 9, No. 8, pp. 1280-1289, December 1999.

[8] B. L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Videos," *IEEE Trans. on CSVT*, Vol. 5, pp. 533- 544, 1995.

[9] A. Doulamis, N. Doulamis, and S. Kollias, "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Processing*, Vol. 80, pp. 1049-1067, June 2000.

[10] N. Doulamis, A. Doulamis, Y. Avrithis, K. Ntalianis, and S. Kollias, "Efficient Summarization of Stereoscopic Video Sequences," *IEEE Trans. on CSVT*, Vol. 10, No. 4, pp. 501-517, June 2000.

[11] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," Psychometrika, Vol. 1, pp. 211–218, 1936.

[12] J. J. Gerbrands, "On the relationships between SVD, KLT, and PCA," *Pattern Recognition*, Vol. 14, pp. 375–381, 1981.

[13] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, London, U.K.: Academic, 1979.

[14] M. Tekalp, *Digital Video Processing*, Eng. Cliffs. N. Jersey, Prentice Hall, 1995.

[15] E. Goldberg, *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison Wesley, 1989.

[16] K. S. Tang, K. F. Man, S. Kwong and Q. He, "Genetic Algorithms and Their Applications," *IEEE Signal Processing Magazine*, pp. 22-37, Nov. 1996.

[17] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.