Key-Frame Based Video Summarization Using QR-Decomposition

Ali Amiri, Mahmood Fathy
Computer Engineering Department
Iran University of Science and Technology
Tehran, Iran
A_amiri@iust.ac.ir, Mahfathy@iust.ac.ir

ABSTRACT: *In this paper, we propose a novel keyframe based video summarization system using QR-Decomposition. Specially, we attend to the challenges of defining some measures to detect the dynamicity of shot and video and extracting appropriate keyframes that assure the purity of video summary. We derive some efficient measures to compute the dynamicity of video shots using QR-Decomposition and we utilize it in detecting the number of keyframes must be selected from each shot. Also, we derive a corollary that illustrates a new property of QR-Decomposition. We utilize this property in order to summarize video shots with low redundancy. The proposed algorithm is implemented and evaluated on TRECVID 2006 benchmark platform. Compared with results reported by others, our results are among the best. These results confirm the high performance of the proposed algorithm.*

Keywords: Video summarization, Key frame extraction, QR Decomposition, Shot dynamicity

Received: 19 April 2010, Revised 18 May 2010, Accepted 28 May 2010

1. Introduction

The explosive proliferation of multimedia content made available on broadcast and Internet initiated and increasing demand for efficient tools and methods for fast browsing and accessing the information pursued. Supplying information for pervasive access and use of multimedia is the most important objective in multimedia researches. To attain this, it is necessary to extend technologies to extract the interesting points from the media large pieces.

Video summarization is a novel technology of content-based video compression, which efficiently finds significant information from video and eliminates redundant data. It provides a small summary for a long video data, and includes two categories: story board and video skimming. The former, in addition identified as a static keyframes, is a set of motionless frames that have been extracted or created from shots or scenes, while the latter is a set of unstill frames which illustrates users the essential parts of video for efficient browsing. While story board video summarization, users can guess the general content of the video more quickly from keyframes, dynamic video skimming includes some important pictorial, audio, and motion information. Both techniques should demonstrate a summary of the essential events existed in the video documents. Here, we focus on the construction of a summary from video using keyframes. Key frames can be defined as a subset of a video sequence that can represent the video visual content as close as possible, with a limiting number of frame information [1]. Usually, the key frame extraction algorithms assume that the video file has been segmented into shots and then extract within each shot a small number of key frames. We expect the key frame extraction algorithm preserves the most important content of the video while eliminating all redundancy. Also, it should be automatic and content-based. Theoretically, all primary components of video such as relevant objects, actions and events that reflect the semantic primitives must be used.

Recent works in the story based video summarization area can be classified into three categories, based on the method of key frame extraction: sampling based, shot based, and segment based. In the sampling based approaches, key frames were extracted by randomly choosing of uniformly sampling from the original video. It has been used in earlier systems such as magnifier [2] and the MiniVideo [3]. It is the straightforward and easy way to extract key frames, yet such an arrangement may fail to capture the real video content, especially when it is highly dynamic.

In the shot based approaches, the video is segmented into separated shots and one or more key frames extracted from each shot. A sequence of frames captured by one camera in a single continuous action in time and space is referred to as a video shot [4]. Normally, it is a group of frames that have constant visual attributes, (such as color, texture, and motion). It is a more significant and straightforward method to extract key frames by adapting to dynamic video content. In shot based approaches a typical and easy manner is utilizing low-level features such as color and motion to extract key frames. More complicated systems based on color clustering, global motion, or gesture analysis could be found in [5]–[7]. In segment based approaches, the video segments are first extracted from the clustering of frames and then the key frames are chosen as those frames of video that are closest to the centroids of each calculated clusters. Some systems based on this approach can be found in [8]-[12].

Video summarization is not novel to researchers in the content-based multimedia mining community. Since 1990, many algorithms have been widely presented in multimedia area. Also, over the past few years research on video summarization rapidly has been increased. In spite of recent improvements, video summarization is still a very difficult task, with many unsolved problems. Among them, two problems are very important and yet unsolved. First, in of many of these approaches is that they utilize spatial low-level visual features (e.g., color histogram) of the video for keyframes extraction and video summarization, which generally reflect only spatial changes between frames. In other words, in these algorithms, there is not any spatio-temporal analysis to detect the temporal and semantically dependency between the frames and to eliminate the redundancy and repetitive frames with similar spatial concepts. The second drawback is that in these approaches, users are not able to search and browse the video using high-level concepts intuitively. In these methods there is a semantic gap between low-level features and semantic interpretation of the video. In other words, maintaining of content balance in key-frames and reducing redundancy and linking the semantic gap between low-level descriptors used by computer systems and the high-level concepts perceived by human users, is very important and until now has not been solved efficiently. In order to achieve these goals, we put forward a hierarchical QR-decomposition-based approach that will be used to summarize video documents efficiently. We have carried out our solution and assessed it according to the TRECVID 2006 data set. Our method produced very hopeful results, in comparison with the best results reported in the literature.

The rest of this paper is organized as follows. A brief description of QR-decomposition is discussed in Section 2. Section 3 explains our QR-Decomposition based video summarization approach. Section 7 includes the experimental evaluations of our approach on video summarization tasks in the some well-known test beds. Section 8 clarifies our conclusions and considers some ideas that could enhance the performance of our present solution.

2. Video Analysis Using Qr Decomposition

In order to design an efficient video summarization system, two presumptions are required, the first one of which is that a mathematical criterion to measure the video dynamicity for detecting the number of keyframes in each shot needed to produce a summary with a predefined length. The second presumption is an accurately method that detects the independent keyframes within shots.

In this section we will present some QR-Decomposition based algorithms to afford these presumptions in the proposed video summarization system. A review of QR-Decomposition and the details of the QR-Decomposition based inter frames dependency detection and intra shot dynamicity detection are offered in the following subsections.

2.1 Review of QR-Decomposition

The QR-Decomposition of a matrix A of order m × n where m ≥ n is given [13]

$$A. \Pi = Q. R, \tag{1}$$

where $\Pi = [\rho_{i,j}]$ is a permutation matrix; $Q[q_{i,j}]$ is an m × n column-orthogonal matrix; and $R[r_{i,j}]$ is an n × n upper triangular matrix whose diagonal elements, the R-values, are arranged in decreasing order and incline to track the singular values of A.

2.2 QR-Decomposition based Frame Dependency Detection

It is obvious that the major grounds of visual redundancy due to the repetition of each frame with little alterations during its temporary adjacent frames in the video. Subsequently, we need to design an algorithm that able to detect dependencies between frames, eliminate the repetitive frames with small alterations and extract key frames with maximum visually information.

QR-Decomposition is a powerful mathematical tool that provides these requirements. In the subsequent, first we will give an example to clarify the most important unique property of QR-Decomposition that grants these necessities, and next we will evidence this property in a corollary.

Example. Let A = [A_1 ,A_2 ,A_3] , be a 6 × 3 matrix as follow:

$$A = \begin{bmatrix} 7 & 8 & 7 \\ 3 & 8 & 8 \\ 10 & 2 & 3 \\ 0 & 5 & 7 \\ 4 & 4 & 7 \\ 4 & 6 & 2 \end{bmatrix},$$

Then, the QR-Decomposition of A is given as A. Π = Q. R, where Π and R are as follows:

$$\Pi = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

$$R = \begin{bmatrix} -14.9666 & -9.2873 & -13.4299 \\ 0 & 10.1856 & 1.4994 \\ 0 & 0 & 5.1371 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

Now, let $\widetilde{A} = [A_1, \underbrace{A_2, A_2, A_2}_{\text{redundancy}}, A_3]$ be a 6 × 5 matrix as follows:

$$\widetilde{A} = \begin{bmatrix} 7 & 8 & 8 & 8 & 7 \\ 3 & 8 & 8 & 8 & 8 \\ 10 & 2 & 2 & 2 & 3 \\ 0 & 5 & 5 & 5 & 7 \\ 4 & 4 & 4 & 4 & 7 \\ 4 & 6 & 6 & 6 & 2 \end{bmatrix},$$

Then, the QR-Decomposition of Ã is given as Ã. $\widetilde{\Pi}$ = $\widetilde{Q}$. $\widetilde{R}$,where $\widetilde{\Pi}$ and $\widetilde{R}$ are as follows:

$$\widetilde{\Pi} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\widetilde{R} = \left[\underbrace{\begin{matrix} -14.9666 & -9.2873 & -13.4299 \\ 0 & 10.1856 & 1.4994 \\ 0 & 0 & 5.1371 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}}_{R} \begin{matrix} -13.4299 & -13.4299 \\ 1.4994 & 1.4994 \\ 5.1371 & 5.1371 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{matrix} \right]$$

According to this numerical example, we find that QR-Decomposition sort the columns of matrix in decreasing columns independency order and if some R-values of R-matrix in the QR-Decomposition of a given matrix be zero (or close to zero), then the matrix has redundancy in its columns, and we can delete the corresponding columns to reduce these redundancies. Consequently, we give the following corollary.

Corollary 1. Let the QR-Decomposition. of A be given by (1), A = [A_1,...,A_i,...,A_n], R = [R_1,...,R_n]. If Rank(A) = n, then, we all columns of A are linearly independent. Let $\widetilde{A} = [A_1, \ldots, \underbrace{A_i^{(1)}, \ldots, A_i^{(k)}}_{k}, \ldots, A_n]$ be the matrix obtained by k time duplicating of column vector A_i in ($A_i^{(1)} = \cdots = A_i^{(k)} = A_i$), and $\widetilde{R}$ = [R_1,...,R_n, $\widetilde{R}_1$,..., $\widetilde{R}_k$] be the corresponding R-matrix obtained from its QR-Decomposition then $\widetilde{R}_i$(n + i), for i = 1,…,k (the k latest R-values of $\widetilde{R}$) will be zero.

According to this corollary, we can extract a suitable feature matrix from the input video and reduce the visual redundancy it by eliminating duplicative frames. Hence, we provide a mathematical analysis to extract some independent key frames with little redundancies.

2.3 QR-Decomposition based Shot Dynamicity Measurement

In order to define a measure to estimate the amount of dynamicity in an arbitrary shot, we utilize QR-Decomposition. At first, we split each gray scaled input frame into M small blocks and apply QR-Decomposition on each block to identify the static part. We consider the values of a particular pixel (x_i,y_j) at bth block (b = 1,… ,M), over time as a time series of intensity values X, by time t:

$$X = \{X^b_{i,1}, X^b_{i,2}, X^b_{i,t}\}$$

We construct matrix A^b for bth block as:

$$A^b = \begin{bmatrix} X^b_{1,1} & \cdots & X^b_{1,t} \\ \vdots & \ddots & \vdots \\ X^b_{N,1} & \cdots & X^b_{N,t} \end{bmatrix}$$

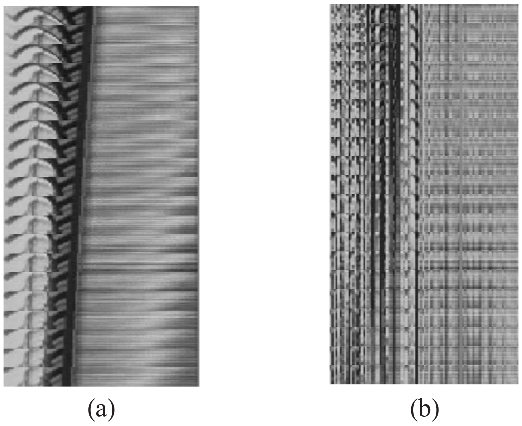
Where N is the number of pixels in each block. Because the proposed method is the same for all blocks, in the rest of this section we name A^b as A. Each R-value taken from QR decomposition of matrix A, is related to one of the columns of A. Since those columns of A containing only static data are almost similar to each other, the R-values corresponding to these columns will be smaller than those containing dynamic and motion objects. If we define the series Y as a sorted list of X according to the R-values, for the ith pixel's intensity values at bth block, we can estimate the dynamicity probability as follows:

$$P(D|Y^b_{i,j}) = \begin{cases} 0 & \text{if } j > (1-\beta)*t \\ 1 & \text{otherwise} \end{cases} ; \qquad (2)$$

$$j = 1,...,t \; ; \; i = 1,...,N$$

Where β demonstrates the percentage of the blocks containing static data only. Since the block sizes are small, each block shows nothing but static data in many image frames. It is depend on type of the video. For example in news video, relay on our experiments, 1⁄3 is a proper value for β .

Fig. 1 demonstrates the sorting result based on R-values on 100 frames of a 20×20 test block. Fig. 1.(a) shows matrix A and Fig. 1.(b) shows the same matrix where its columns are sorted based on QR-Decomposition's R-values (Y series). As can be seen, those columns containing only the background are shifted to the end.



(a) (b)

Figure 1. The sorting result based on R-values on 100 frames of a 20×20 test block. (a): Matrix A. (b): Matrix where its columns are sorted based on QR-Decomposition's R-values

We define the fraction of pixels that belong to the dynamic pixels as dynamicity of that block So, the dynamicity of bth block at jth frame can be written as:

$$BD(b, j) = \frac{\text{number of pixels marked as dynamic}}{\text{Total number of pixels (N)}} \qquad (3)$$

The BD measure is a number between 0 and 1. If a block be a part of moving object then this measure will be near 1. Also, we can define the dynamicity of j^th^ frame as follows:

$$\mathrm{FD(j)} = \frac{1}{\mathrm{M}} \sum_{\mathrm{i=1}}^{\mathrm{M}} \mathrm{BD(i, j)} \tag{4}$$

For an intricate frame with many moving objects, the FD measure will limit to 1, and for a simple frame with small moving parts, this measure will limit to 0. Similarly, for a shot with k frames, the dynamicity can be defined as follows:

$$\mathrm{SD} = \frac{1}{\mathrm{K}} \sum_{\mathrm{j=1}}^{\mathrm{K}} \mathrm{FD(j)}, \tag{5}$$

This measure can be utilized to control the number of required key frames for each shot. If a shot has many self-motivated frames with different visual concepts, then the SD measure will be limit to 1, and we will need to extract numerous key frames to summarize the shot. Also, if a shot include some static frames, then the SD measure will be limit to 0, and we extract few key frames to summarize the shot. Also, for a video with n shots, the total dynamicity can be defined as follows:

$$\mathrm{VD} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{SD(i)} \tag{6}$$

Now, let the input video has been segmented into n different shots such as $S_1, S_2, \ldots, S_n$ with shot dynamicity SD(1),…,SD(n) respectively. We utilized the shot boundary detection algorithm proposed in [14], which based on QR-Decomposition, to partition the video into shots. Also, suppose the user of video summarization system interested to extract a video summary of length Len, then the number of keyframes must be selected from shot i is:

$$\mathrm{NOK_i} = \left\lfloor \frac{\mathrm{Len}}{\mathrm{VD}} \times \mathrm{SD(i)} \right\rfloor + 1 \tag{7}$$

$$\text{for i = 1,...,n}$$

Consequently, by using this approach we introduced a mathematical criterion to measure the dynamicity of a shot, and to control the number of key frames will be extracted from each shot.

2.4 QR-Based Video Summarization

In this paper to detect the video shot boundaries, we apply a QR-Decomposition based algorithm presented in [14]. Let the input video has been segmented into n various shots $S_1, S_2, \ldots, S_n$. For each shot i with n_i frames and i = 1,2,...,n, we created an m-dimentional feature vector for each frame j such as $X_j^{(i)} = [X_{1,j}^{(i)}, X_{2,j}^{(i)}, X_{m,j}^{(i)}]^T$. using $X_j^{(i)}$ as column vector j, we obtained feature matrix $X^{(i)}$ for shot i as follows:

$$X^{(i)} = [X_1^{(i)}, \ldots, X_t^{(i)}] = \begin{bmatrix} X_{1,1}^{(i)} & \cdots & X_{1,n_i}^{(i)} \\ X_{2,1}^{(i)} & \cdots & X_{2,n_i}^{(i)} \\ \vdots & \ddots & \vdots \\ X_{m,1}^{(i)} & \cdots & X_{m,n_i}^{(i)} \end{bmatrix} \tag{8}$$

In order to extract spatial features of each frame, from a broad range of image features, we used color histograms which are essential features for signifying the overall spatial features of each frame [15]. Specially, we created a 1331-dimensional feature vector $X_j^{(i)}$. To compute the feature vector in our system implementation, we made three-dimensional histograms in RGB color space with twelve bins for R, G and B, respectively, leading to a total of 1728 bins. These produced a 1728-dimensional feature vector for the frame. Finally, utilizing the feature vector of frame j as the j^th^ column, we generated the feature matrix $X^{(i)}$ for shot i in the video sequence.

Also, suppose the user of video summarization system interested to take out a video summary of length Len. We compute NOK_i,i = 1,2,…,n, the number of keyframes must be selected from each shot according to (7). In addition, for each shot i, the QR-Decomposition of its feature matrix is given as:

$$X^{(i)}\Pi_i = Q_i R_i, \quad \text{for } i = 1, \ldots, n \tag{9}$$

Now, according to the corollary (1), the corresponding frames to the first NOK_i columns of R_i are selected as the key frames of each shot i, where i = 1,…,n.

3. Experimental Results

In this section, a set of experiments will be presented to confirm the efficiency of the proposed video summarization system. In order to evaluate the system with standard data sets, we have demonstrated the outcomes of the tests using a largescale test set provided by the TRECVID 2006 [16], which has assessments for key frame detection and summary extraction. Also, we utilized two well-known objective criteria to compare the effectiveness of our system with some other recently presented systems.

In the following subsections, the details of test set, evaluation criteria, and results of experiments to assess the results have been presented.

3.1 Evaluation Criteria

Evaluation of the video summaries obtained by the various key frame extraction techniques is one of the important issues in the field of video analysis and summarization. From a broad variety of objective measures for assessment of goodness of video summary, we have chosen two well-known measures: Fidelity measure which defined in [17], and Shot Reconstruction Degree (SRD) which suggested in [18]. Fidelity applies a global strategy, while the SRD utilizes a local assessment of the key frames.

3.1.1 Fidelity Measure

Let $V = \{F_1,\ldots,F_\gamma\}$ be the frames of the input video, and $Keys = \{F_{k_1},\ldots,F_{k_{len}}\}$ be the set of k_{len} key frames extracted from the video sequences. The distance between an arbitrary frame F_i, $i = 1,2,\ldots,\gamma$ and set of key frames Keys is defined as follows:

$$\mathrm{dist}(F_i,Keys) = \min_{1\le j\le len}\{\|\mathrm{Feature}(F_i) - \mathrm{Feature}(F_{k_j})\|\}$$

Where $\|.\|$ is a proper distance function and Feature(.) is a feature extraction method that used to describe each frames in the video. Here we used color histograms. Also, the distance between the input video and set of key frames is defined as:

$$\mathrm{Dist}(V,keys) = \max_{1\le i\le n}\{\mathrm{dist}(F_i, Keys)\}$$

And consequently the Fidelity criterion is defined as follows:

$$\mathrm{Fidelity}(V,keys) = \max_{Dist} - \mathrm{Dist}(V,keys) \qquad (10)$$

Where $\max_{Dist}$ is the largest possible value that the $\|.\|$ distance function can suppose. According to this measure, high Fidelity values demonstrate that the key frames extracted from the input video give a high-quality global description of the visual content of the video.

3.1.2 SRD Measure

This measure evaluates the goodness of the key frame extraction algorithm in reconstructing of the entire input video from the set of keyframes by utilizing an appropriate frame interpolation algorithm. If the reconstructed video approximates the original video accurately, then the key frames will summarize the visual content of the video appropriately.

Let $V = \{F_1,\ldots,F_\gamma\}$ be the frames of the input video, and $Keys = \{F_{k_1},\ldots,F_{k_{len}}\}$ be the set of k_{len} key frames extracted from the video sequences. Given FIA be an interpolation algorithm that reconstructs an arbitrary frame i, $i = 1,\ldots,\gamma$ from extracted key frames:

$$\tilde{F}_i = \mathrm{FIA}(Keys, i)$$

Where $\tilde{F}_i$ is the approximation of frame i using FIA. Also, let $\mathrm{Sim}(F_i,F_j)$ be a similarity measure between frames i and j, then the SRD measure will be computed as follows:

$$\mathrm{SRD}(V,keys) = \sum_{n=i}^{\gamma} \mathrm{Sim}(F_i,\tilde{F}_i) \qquad (11)$$

This measure analyzes local details in the video and high SRD values demonstrate that the extracted key frames will give more details of visual content of the input video.

3.2 Results

In Fig. 2 the shot dynamicity measure of (5) has been plotted for some different shots. According to this result and from the dynamicity point of view, we classify the video shots into three categories: static, semi-dynamic, dynamic. A

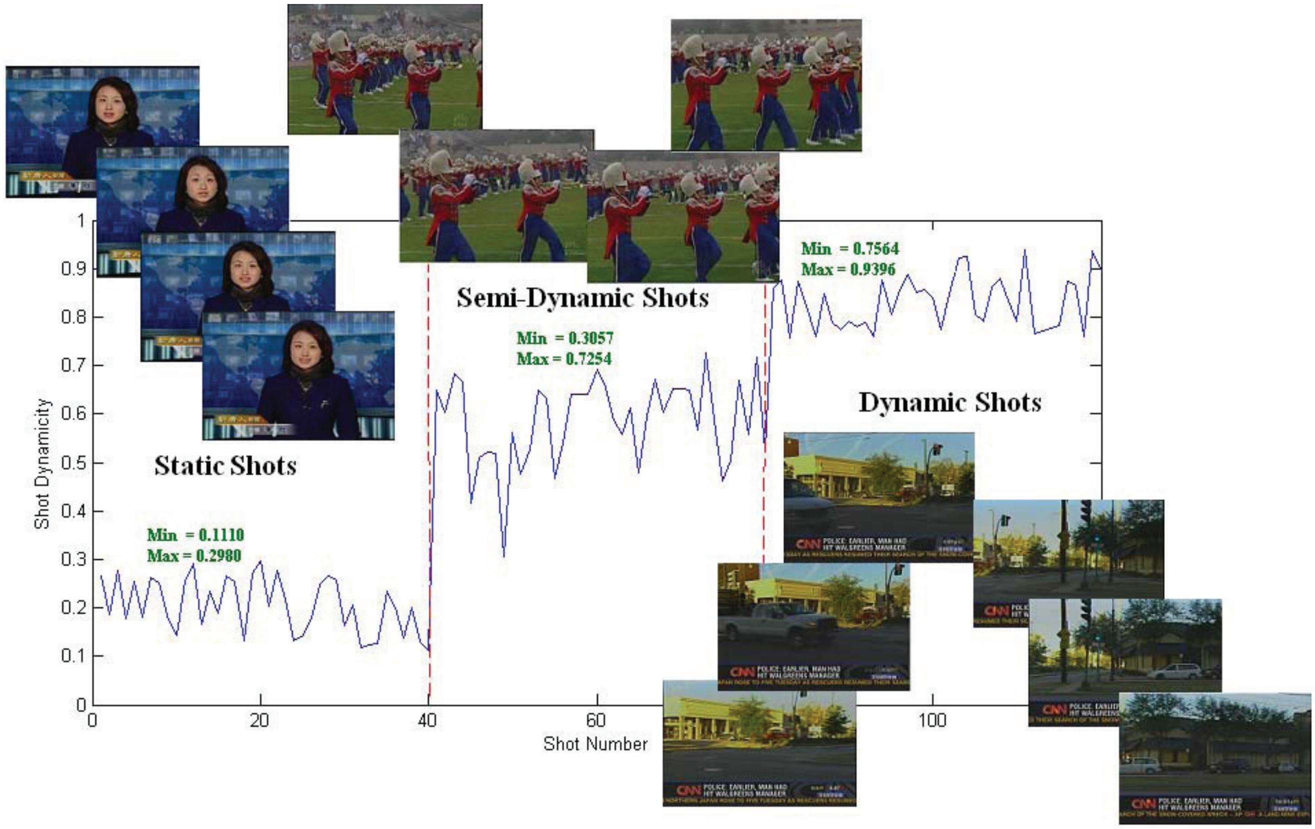


Figure 2. Three categories of the video shots based on Shot Dynamicity Measure: static, semi-dynamic, dynamic

static shot such as anchor shots include a sequence of frames with very small object motions. In these shots, a small number of keyframes could summarize all visual content of the shot. A semi-dynamic shot as a shot of tennis game contains a sequence of frames with medium object or camera motions. In these shots, a medium number of keyframes could summarize all visual content of the shot. A dynamic shot as a shot of freeway includes a set of frames with high object or camera motions and we need many numbers of keyframes to summarize the shot. Therefore, as mentioned in previous section, to control the number of key frames will be extracted from each shot, we can use shot dynamicity measure.

Together using our QR-Decomposition based algorithm, we experienced a simple Time Sampled (TS) method [19], the KMeans with SVD (KMSVD) approach [20] and the Information Theory-based (IT) algorithm [21]. In our implementation, all methods are required to produce same number of key-frame which determined by user. Also, to judge the performance of our algorithm with respect to the other algorithms, it is useful to express the results as a measure of relative improvement using the following formula:

$$\Delta\,(QR_{Method}\,, X_{Method}) = \frac{Measure(QR_{Method}) - Measure(X_{Method})}{Measure(X_{Method})} \tag{12}$$

where Measure corresponds to the Fidelity and the SRD measure, and we substitute X_{Method} with the algorithm we used to compare with our proposed QR_{Method} algorithm.

Fig.3 and Fig.4 illustrates the average of relative improvement of Fidelity and SRD for different number of keyframes for 300 various shots. It can be seen that when the number of key frames per shot is small the differences between the proposed and other algorithms are slight and as the number of key frames per shot increases, the gap is more marked.

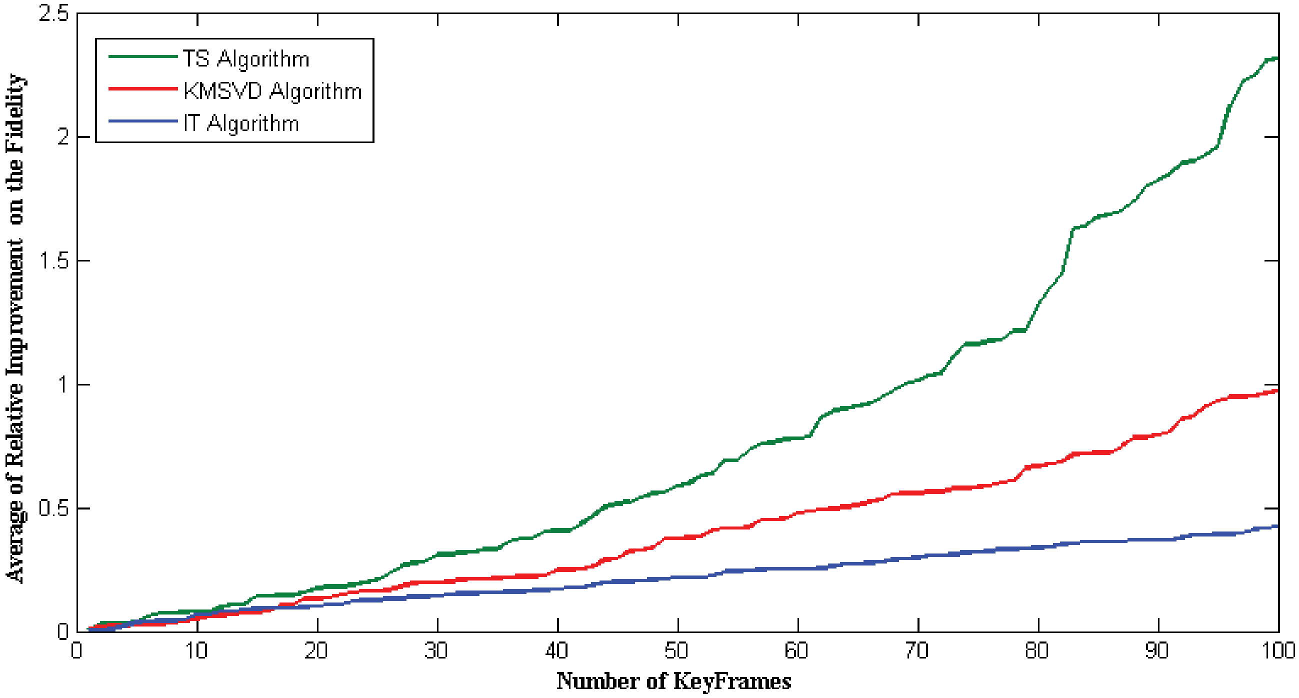


Figure 3. The average of relative improvement of Fidelity (The results are in percentages)

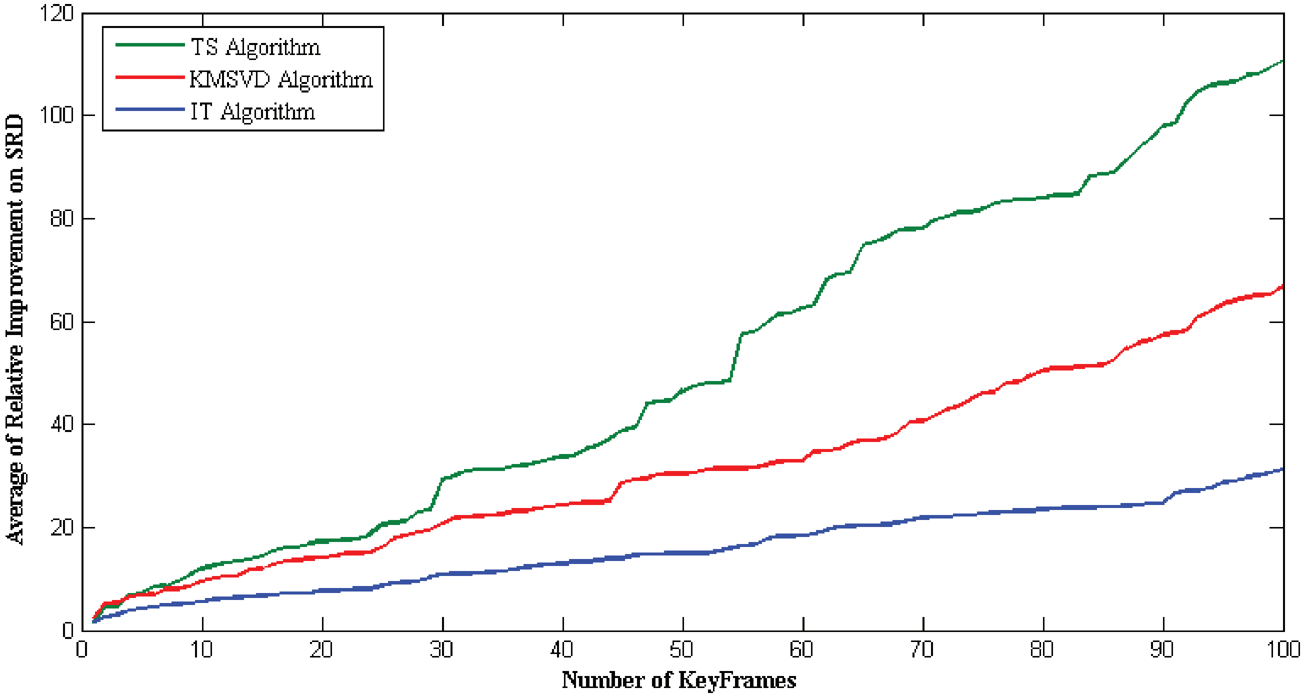


Figure 4. The average of relative improvement of SRD (The results are in percentages)

Table 1 illustrates the average of the relative improvement ($\Delta(QR_{Method}, X_{Method})$ measured on the Fidelity and SRD results of the proposed algorithm with respect to each of the other algorithms tested using all TRECVID 2006 video set.

Algorithm	**Relative Improvement Measured on**	
	Fidelity	***SRD***
IT	3.6	11.8
KMSVD	10.7	35.4
TS	22.3	93.7

Table 1. Average of the relative improvement (Δ) measured on the fidelity and srd results of the proposed algorithm with respect to each of the other algorithms

4. Conclusion

In this paper, a novel video summarization algorithm is developed based on QR-Decomposition. We derive some efficient measures to compute the dynamicity of video shots using QR-Decomposition and we utilize it in detecting the number of keyframes must be selected from each shot. Also, we derive a corollary that illustrates a new property of QRDecomposition. We utilize this property in order to summarize video shots with low redundancy. The proposed algorithm is implemented and evaluated on TRECVID benchmark platform. Compared with results reported by others, our results are among the best.

References

[1] Girgensohn, A., Boreczky, J. S. (2000). Time-constrained keyframe selection technique, *Multimedia Tools Applications* 11 (3) 347–358.

[2] Mills, M. (1992). A magnifier tool for video data, *In*: Proc. ACM Human Computer Interface, p. 93–98.

[3] Taniguchi, Y. (1995). An intuitive and efficient access interface to real-time incoming video based on automatic indexing, *In*: Proceeding of ACM Multimedia'95, p. 25–33.

[4] Hanjalic, A. (2002). Shot-boundary detection: unraveled and resolved?" IEEE Transactions on Circuits and Systems for Video Technology, 12 (2) 90-105.

[5] Zhuang, Y.T., Rui, Y. Huang, T.S., Mehrotra, S (1998). Adaptive key frame extraction using unsupervised clustering, *In*: Proceeding of ICIP'98, p. 866–870.

[6] Ju, S.X., Black, M.J., Minneman, S., Kimber, D. (1998). Summarization of videotaped presentations: Automatic analysis of motion and gestures, *IEEE Transaction on Circuits Syst. Video Technol*., 8 (5) 686–696.

[7] Toklu, C., Liou, S.P. (2000). Automatic keyframe selection for contentbased video indexing and access, *In*: Proceeding of SPIE, V. 3972, p. 554–563.

[8] Cernekova, Z., Pitas, I. (2006). Information theory-based shot cut/fade detection and video summarization*, IEEE Transactions on Circuits and Systems for Video Technology* 16 (1) 82-91.

[9] Turaga, P.,Veeraraghavan, A (2009). Unsupervised View and Rate Invariant clustering of Video Sequences," *Computer Vision and Image Understanding*, 113 (3) 353-371.

[10] Liu, Lijie., Fan, Guoliang (2005). Combined Key-Frame Extraction and Object- Based Video Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology* 15 (7) 869-884.

[11] Parkhurst, D.J., Niebur, E (2003). Scene Content Selected by Active Vision, Spatial Vision, 16 (2) 125-154.

[12] Castelhano, M. S., Henderson, J. M (2007). Initial scene representations facilitate eye movement guidance in visual search, *Journal of Experimental Psychology: Human Perception and Performance*, 33 (8) 753-763, 2007.

[13] Setnes, M., Babuska, R (2001). Rule base reduction: Some Comments on the use of Orthogonal Transforms, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 31 (2) 199-206.

[14] Amiri, Ali., Fathy, Mahmood (2009). Video Shot Boundary Detection Using QR- Decomposition and Gaussian Transition Detection, *EURASIP Journal on Advances in Signal Processing*, 12.

[15] Cheng, W., Xu, D., Jiang, Y., Lang , C (2005). Information Theoric Metrics in Shot Boundary Detection, LNAI, V.3683, p.388-394.

[16] NIST, Homepage of Trecvid Evaluation. Online Available: http://wwwnlpir. nist.gov/projects/trecvid/.

[17] Chang, H. S., Sull, S., Lee, S. U (1999). Efficient Video Indexing Scheme for Content-Based Retrieval, *IEEE Transactions on Circuits Systems and Video Technology,* 9 (8) 1269–1279. [18] Tieyan, L., Zhang, X., Feng, J., Lo, K. T (2004). Shot reconstruction degree: a novel criterion for key frame selection, *Pattern Recognition Letter*, 25, p. 1451–1457.

[19] Rui, Y., Huang, T. S. (1998). Exploring Video Structure Beyond the Shots, *In:* Proceeding of IEEE International Conference on Multimedia Computing and Systems (ICMCS), Texas USA, p. 237-240.

[20] Lee, Sangkeun, Hayes, M. H (2004). Properties of the Singular Value Decomposition for Efficient Data Clustering, *IEEE Signal Processing Letters*, 11 (11) 862-866.

[21] Cernekova, Z., Pitas, I. (2006). Information theory-based shot cut/fade detection and video summarization, *IEEE Transactions on Circuits and Systems for Video Technology*, 16 (1) 82-91.