Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 892124, 16 pages
doi:10.1155/2010/892124

Research Article

Hierarchical Keyframe-based Video Summarization Using QR-Decomposition and Modified k-Means Clustering

Ali Amiri and Mahmood Fathy

Computer Engineering Department, Iran University of Science and Technology, 1684613114 Narmak, Tehran, Iran

Correspondence should be addressed to Ali Amiri, a_amiri@iust.ac.ir

Received 20 May 2010; Revised 8 September 2010; Accepted 28 October 2010

Academic Editor: Moon Kang

We propose a novel hierarchical keyframe-based video summarization system using QR-decomposition. Specially, we attend to the challenges of defining some measures to detect the dynamicity of a 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 that must be selected from each shot. Also, we derive a theorem that illustrates an important property of QR-decomposition. We utilize it in order to summarize video shots with low redundancy. The proposed algorithm is implemented and evaluated on TRECVID 2006 benchmark platform. Compared with other algorithms, our results are among the best.

1. Introduction

The latest rapid growths in multimedia applications, expansion of the digital multimedia in the Internet, and large databases of video have initiated an 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 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 illustrate users the essential parts of video for efficient browsing. During story board video summarization, users can guess the general content of the video more quickly from keyframes, while 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. Keyframes 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 keyframe-extraction algorithms assume that the video file has been segmented into shots and then extracts within each shot a small number of keyframes. We expect that the keyframe-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.

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 many of these approaches, 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 spatiotemporal 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 keyframes, reducing redundancy, and linking the semantic gap between low-level descriptors used by computer systems and the high-level concepts perceived by human users, are very important and until now have 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 some well-known benchmark data. 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. An overview of some modern works regarding story board video summarization which have been done over the past few years is considered in Section 2. In Section 3, the problems and demanding issues concerning video summarization are presented and this section demonstrates the reason for which we have proposed this solution. In Section 4, a brief description of QR-decomposition is discussed. Section 5 explains our QR-decomposition-based video summarization approach. Section 6 includes the experimental evaluations of our approach on video summarization tasks in the some well-known test beds. Section 7 clarifies our conclusions and considers some ideas that could enhance the performance of our present solution.

2. Related Works

Video summarization is a method that summarizes video content by choosing a set of significant frames called keyframes and demonstrates the video in a compressed style to give users a well-organized method to browse or search video content. A video summary can be illustrated without the concern of timing problems. In addition, extracted keyframes could be utilized for content-based video indexing and retrieval.

Recent works in the video summarization area can be classified into three categories, based on the method of keyframe extraction: sampling-based, shot based, and segment based. In the sampling based approaches, keyframes 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 keyframes, 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 keyframes 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 keyframes 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 keyframes. More complicated systems based on color clustering, global motion, or gesture analysis could be found in [5–7]. It is necessary to notice that, the regular keyframes cannot reflect well the essential video dynamics. Therefore, many authors have worked on exploring a different method to display the shot content using a combined panoramic image called the mosaic. In [8], Irani and Anandan propose the static background mosaics to represent the underlying shot dynamics. They transform the video from the frame-based representation to an explicit and compact scene-based representation. They explain some of techniques based on geometric and dynamic information for indexing video using the scene-based representation. The authors of [9] have developed synopsis mosaics. They establish global motion representation model that is parametric in both space and time and can be fitted to the entire sequence at once. It leads to temporal consistency and considerably better video summarization. In [10], Taniguchi et al. design a video browsing system that contains two mode of operation for each video: panoramic and keyframe. The panoramic mode estimates the camera operations such as pan and tilt and creates the panorama image from video segment. The keyframe mode is used to supplement the panoramic mode. In some other works, mathematical tools have been applied to video summarization. In [11], the feature trajectory is extracted to model the video content in a high-dimensional feature space. In [12], a keyframe-based video summarization is explained using visual attention clue. They utilized visual attention index (VAI) descriptor to bridge the semantic gap between low-level features used by computer systems and the high-level concepts perceived by human users and to select the keyframes for video summaries. The keyframes are considered as frames of video that correspond to curvature points. In spite of recent improvements, the main drawback of the shot-based keyframe extraction methods is that they do not scale up efficiently for long video. In [13], Ćalić et al. utilized the common and perceptive rules of story grammar of comics to illustrate great amounts of information in the form of a keyframe summary. In [14], the authors proposed two automatic approaches to estimate the very suitable number of keyframes by supervised and unsupervised discovering of content of video. They also presented a keyframe extraction algorithm utilizing three Iso-content principles. In [15], a keyframe-based video summarization framework has been proposed. It analysis the visual cues of the video and extracts four perspective models to make a fusion model. This new model illustrates the view variations of users while looking at the video data and has been used to summarize video sequences.

In segment-based approaches, the video segments are first extracted from the clustering of frames, and then the

keyframes are chosen as those frames of video that are closest to the centroids of each calculated clusters. In [16], Uchihashi et al. present a pictorial video summarization method that resembles comic books. They have applied audio and images analysis to detect the significant scenes and the relevant keyframes of video. The selected keyframes are arranged by significance and then powerfully packed into a pictorial summary. In [17], the authors have used the hierarchical clustering to determine keyframes. In this method, the desired number of keyframes has been controlled by restricting the number of clusters. Also, some temporal limitations has been utilized to ignore inappropriate clusters and to choose a representative frame for each cluster. Yeung and Yeo [18] proposed a scene-level video summarization system. They presented techniques to differentiate the dominance of the content in subdivisions of the segment based on analysis results, select a graphic layout pattern according to the relative dominances, and create a set of video posters, each of which is a compact, visually pleasant, and intuitive representation of the story content. The collection of video posters arranged in temporal order then forms a pictorial summary of the sequence to tell the underlying story. The authors of [19] studied a temporal summarization of video to solve the challenge of extracting a constant number of keyframes to summarize a certain video. They mapped the video frames in a high-dimensional space corresponding to a low-level feature such as color, motion, shape, and texture. Then, they extracted representative images (R-images) as keyframes of the video. All extracted R-images of a story unit were resized and organized into a single regular-sized image following a predefined visual layout called the poster. In [20], Otsuka et al. proposed an automatic sports highlights detection system using only audio features and Gaussian mixture models. Also, in [21], the authors proposed a probabilistic framework to detect sport events using webcast texts. In [22], Ciocca and Schettini suggested a postprocessing algorithm for video summarization methods to generate visual video summaries that are exhaustive but not redundant; their algorithm employs both supervised and unsupervised classification approaches to achieve these goals. In [23], the HMM and Bayesian network are utilized to model audio effects and background sounds and realize the high-level semantics of and auditory context, respectively. This framework utilized to detect significant scenes in the sport digital video. In [24], the authors presented an approach to summarize the instructional video of chalk board presentations. They first extend an algorithm to extract content text and figures from video as middle-level features. These features are used to discover a set of keyframes that contain most of visual content. Finally, the summary video is constructed based on the Kth Hausdorff distance and connected component decomposition.

Recently, researches on segment-based video summarization have been focused on mapping video into a hierarchical structure to build the video summary. In [25], Ratakonada et al. suggest a hierarchical keyframe-extraction technique for video summarization which attends to solve the challenge of video summarization in the absence of any information about its content. They studied a pairwise k-means clustering approach with temporal consecutiveness constrain to build the hierarchical summaries from video. Also, they extend their work for summary from MPEG-2 video without fully decoding the bit stream. Also, another hierarchical video summarization algorithm has been addressed by Zhu et al. in [26]. They first segment the video into shots and apply keyframe extraction algorithm to detect the discrete keyframes. Then, they utilize an affinity matrix to merge visually similar shots into clusters. Temporal information is used to merge temporary adjacent shots from the cluster level into video group level. Consequently, a hierarchical video-content structure with growing granularity is created. Finally, they designed a hierarchical video summarization system based on the hierarchical video structure to build the video summary. In [27], Ngo et al. analyze the video structure and detect video highlights to build summaries from digital video. They represent the video using a complete undirected graph and utilize the normalized cut algorithm to partition efficiently the video into clusters to make a directed temporal graph. They design some algorithms on this graph to detect the scenes, clusters, shots, and subshots, and construct the hierarchical video summary. In [28] the authors introduce a framework that uses Intermedia as an intermediate video description to illustrate the video content hierarchically. The authors of [29] presented a relational graph-based video summarization system that uses a structural video-content browsing system. This structural representation utilizes four classes of entities: who, what, where, and when. The maximum entropy criterion is used to incorporate visual, text features, and speech transcripts to create high-level concept entities and construct relational graph for the video. This graph is used to construct meaningful shots and summaries from video. In [30], a MINMAX optimization model is designed to summarize digital video smoothly and hierarchically. In [31], the authors design a hierarchical video summarization system that offers a two-level redundancy elimination routine to summarize the video contents. In [32], the Laplacian Eigenmap has been utilized to design a hierarchical video summarization tool. The authors show that the system accomplished reasonable results in keyframe extraction and shot level video summarization.

We suggest the hierarchical QR-decomposition-based video summarization method and exhibit its effectiveness through a theoretical and practical analysis. As opposed to the aforementioned approaches, our solution is capable of summarizing the video in shot and scene levels, and it changes the size of video summaries by user interests. Finally, we test our algorithm on the some well-known datasets.

3. Problems and Motivations

In this section, we discuss the problem of keyframe-based video summarization and concentrate to a number of exacting issues. Next, we argue the motives and philosophy of our approach for resolving these challenges.

In the recent context of video summarization, the keyframe-based techniques have been modeled as a hierarchical system to extract the video summaries in various levels such as subshot, shot, and scenes level. These systems

include four major modules: video shot boundary detection, representation of shot content, keyframe extraction, and hierarchical keyframe grouping. 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). Shot boundary detection, also known as temporal video segmentation, is the process of detecting the transitions between the adjoining shots [33]. After shot boundary detection, the visual aspect of video shots is also studied. In this perspective, video can be viewed as a three-dimensional signal, in which two dimensions disclose the visual content in the horizontal and vertical frame direction, and the third dimension discloses variations of the visual content over the time axes. This formulation extracts certain kinds of visual features from each frame, obtains a compact content representation, and then extracts some keyframes from each shot. Finally, the researchers proposed various approaches to construct the video summaries for different levels.

These researches on video summarization make it a much more challenging task, especially compared to those of a conventional video summarization task. Some of these challenges are as follows.

(1) Defining criteria for measuring dynamicity in shot, scene, and movie levels plays a very important role in video summarization and can be very helpful in determining the number of keyframes required for representing each shot to create video summarization. However, no criteria have ever been defined, and this remains to be a very important challenge.

(2) Although a great deal of works have been done on this issue so far, defining criteria for identifying redundancies and duplicative frames in shot level is still a critical challenge in video summarization. In other words, there is not any spatiotemporal analysis to detect the temporal and semantically dependency between the frames and to eliminate the redundancy and repetitive frames with similar spatial concepts at shot levels.

(3) Eliminating redundancies in shot level is not equivalent to removing redundancies in scene level. Clustering is usually an important technique in eliminating redundancies in scene level. Despite these improvements, designing a clustering algorithm with variable clusters is still a challenging issue.

To attack these challenges in a unified way, we advise using QR-decomposition and modified k-mean algorithm-based approach, which can significantly increase the effectiveness of video summarization tasks while at the same time reducing the computational cost. The main ideas in our solution are summarized below.

(1) Some criteria which are based on the basic characteristics of QR-decomposition have been suggested for determining the dynamicity in shot, scene and video levels.

(2) A theorem revealing a new characteristic of QR-decomposition has been presented and proven. This theorem presents a clear and definite criterion for identifying and removing redundancies. This feature also makes it possible for the users to set the duration of the video summarization. The most independent and important frames are used in each summarization.

(3) A modified version of the k-mean clustering algorithm has been proposed for summarization in shot level. This algorithm starts clustering with two clusters and then if necessary increases the number of clusters.

4. 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 accurate 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 interframes dependency detection and intrashot dynamicity detection are offered in the following subsections.

4.1. Review of QR-decomposition. It is evident that the study of singular values of a matrix represents valuable and useful information. The singular value decomposition (SVD) of a matrix is a factorization of the matrix into a product of three matrices. Given an $m \times n$ matrix A, where $m \geq n$, the SVD of A is defined as [34]

$$A = U\Sigma V^T, \tag{1}$$

where $U = [u_{i,j}]$ is an $m \times n$ column-orthogonal matrix whose columns are referred to as left singular vectors; $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix whose diagonal elements are nonnegative singular values arranged in descending order; and $V = [v_{i,j}]$ is an $n \times n$ orthogonal matrix whose columns are referred to as right singular vectors. If $\text{rank}(A) = r$, then Σ satisfies

$$\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0. \tag{2}$$

The SVD has been utilized by a great number of researchers for rule base reduction [35, 36]. The main reason for using SVD in complexity reduction is that SVD decomposes a given system into different parts and specifies the level of the importance of each decomposed part. We can reduce the rank of the matrix by selecting its most important columns, known as the subset selection problem [9]. In order to do this, we can simply truncate the vectors that have the least level of importance in accordance with SVD.

The QR-decomposition of a matrix A of order $m \times n$, where $m \geq n$ is given as [34]

$$A \cdot \Pi = Q \cdot R, \tag{3}$$

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

If there is a well-defined gap in the singular values of A, $\sigma_{r+1} \ll \sigma_r$, then the subset selection will tend to yield a subset that contains the most significant columns (rules) of A. Nevertheless, the singular values usually reduce smoothly without any clear gap. In such situations, the truncation index r is specified through calculating the number of zero or close to zero singular values in the SVD of A. Because it has been professed that the smaller the singular values, the less significant the related rules will be. As for the singular values, the R-values also help to specify which number to choose [35].

QR-decomposition has many applications in image and video processing. Recently, in [37], the QR-decomposition has been used for background estimation and object detection and in [38] for video segmentation and shot boundary detection. In [35], the QR-decomposition is utilized for rule base reduction. The authors of [39–41] utilized QR-decomposition for data dimension reduction and discriminant analysis.

Also, other decompositions such as SVD and or eigenvalue decomposition (EVD) have many applications in various domains of computer science. For example, in [33, 42, 43], the SVD has been utilized for shot boundary detection and video retrieval and summarization. Although they are successful, as mentioned above, the singular values usually reduce smoothly without any clear gap, and as a result calculating the truncation index is not efficient. But, in QR-decomposition, the R-values has a clear gap, and the calculation of truncation index is very efficient. Also, in the eigenstructure-based decompositions such as singular SVD or EVD, the computational complexity is very high, and it turns out to be a major obstacle to real-time implementation. However, the computational complexity of QR factorization is lower than that of SVD and EVD. In addition, the QR factorization can be recursively implemented as an updating factorization [34, 44] or as a rank-revealing ULV decomposition [45].

4.2. Intrashot Redundancy Detection Based on QR-decomposition. 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 keyframes 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 theorem.

Example 1. Let $A = [A_1, A_2, A_3]$, be a 6×3 matrix as follow:

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

Then, the QR-decomposition of A is given as $A \cdot \Pi = Q \cdot R$, where Π and R are as follows:

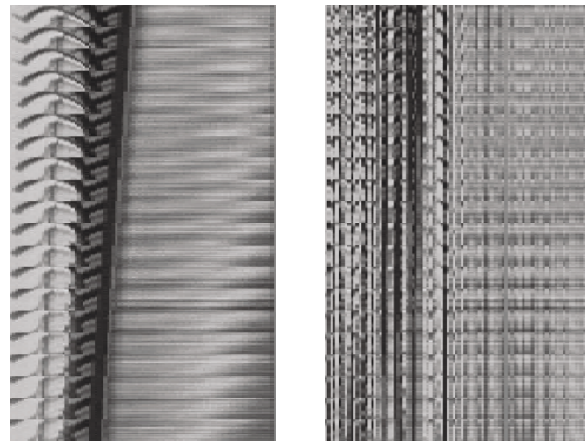
$$\Pi = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} -14.9666 & -9.2873 & -13.4299 \\ 0 & 10.1856 & 1.4994 \\ 0 & 0 & 5.1371 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{5}$$

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

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

Then, the QR-decomposition of $\widetilde{A}$ is given as $\widetilde{A} \cdot \widetilde{\Pi} = \widetilde{Q} \cdot \widetilde{R}$, where and are as follows:

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



(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.

$$\widetilde{R} = \begin{bmatrix} -14.9666 & -9.2873 & -13.4299 & -13.4299 & -13.4299 \\ 0 & 10.1856 & 1.4994 & 1.4994 & 1.4994 \\ 0 & 0 & 5.1371 & 5.1371 & 5.1371 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \underbrace{0 \quad 0 \quad 0}_{R} & & & 0 & 0 \end{bmatrix}. \tag{7}$$

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 theorem.

Theorem 4.1. *Let the QR-decomposition of A be given by* (3), $A = [A_1, \ldots, A_i, \ldots, A_n]$, $R = [R_1, \ldots, R_n]$. *One supposes that Rank$(A) = n$; in other words, one considers that 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* ($A_i^{(1)} = \cdots = A_i^{(k)} = A_i$), *then the corresponding R-matrix obtained from its QR-decomposition will be as* $\widetilde{R} = [R_1, \ldots, R_n, \widetilde{R}_1, \ldots, \widetilde{R}_k]$ *and* $\widetilde{R}_i(n+i)$, *for* $i = 1, \ldots, k$ *(the k latest R-values of $\widetilde{R}$) will be zero.*

Proof. See the appendix. □

According to this theorem, we can extract a suitable feature matrix from the input video and reduce the visual redundancy by eliminating duplicative frames. Hence, we provide a mathematical analysis to extract some independent keyframes with little redundancies. The details of usage of this theorem will be studied in Section 5.

4.3. Shot Dynamicity Measurement Using QR-decomposition. 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 ($b = 1, \ldots, M$), over time as a time series of intensity values X, by time t

$$X = \left\{X_{i,1}^b, X_{i,2}^b, \ldots, X_{i,t}^b\right\}. \tag{8}$$

We construct matrix A^b for bth block as

$$A^b = \begin{bmatrix} X_{1,1}^b & \cdots & X_{1,t}^b \\ \vdots & \ddots & \vdots \\ X_{N,1}^b & \cdots & X_{N,t}^b \end{bmatrix}, \tag{9}$$

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:

$$P\left(D \mid Y_{i,j}^b\right) = \begin{cases} 0 & \text{if } j > (1-\beta) * t, \\ 1 & \text{otherwise,} \end{cases} \tag{10}$$

$$j = 1, \ldots, t; \quad i = 1, \ldots, N,$$

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 depends on the type of the video. For example, in news videos, relaying on our experiments 1/3 is a proper value for β.

Figure 1 demonstrates the sorting result based on R-values on 100 frames of a 20×20 test block. Figure 1(a) shows matrix A, and Figure 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.

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

$$\text{Block-Dynamicity}(b,j) = \frac{\text{number of pixels marked as dynamic}}{\text{Total number of pixels } (N)}. \tag{11}$$

The Block-Dynamicity 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 jth frame as follows:

$$\text{Frame-Dynamicity}(j) = \frac{1}{M}\sum_{i=1}^{M}\text{Block-Dynamicity}(i,j). \tag{12}$$

For an intricate frame with many moving objects, the Frame-Dynamicity 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:

$$\text{Shot-Dynamicity} = \frac{1}{K}\sum_{j=1}^{K}\text{Frame-Dynamicity}(j). \tag{13}$$

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

$$\text{Video-Dynamic} = \sum_{i=1}^{n}\text{Shot-Dynamicity}(i). \tag{14}$$

Now, let the input video has been segmented into n different shots such as $S_1, S_2, \ldots, S_n$ with shot dynamicity Shot-Dynamicity(1), ..., Shot-Dynamicity(n), respectively. We utilized the shot boundary-detection algorithm proposed in [38], which is 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

$$\text{NOK}_i = \left\lfloor \frac{\text{Len}}{\text{Video-Dynamic}} \times \text{Shot-Dynamicity}(i) \right\rfloor + 1, \quad \text{for } i = 1,\ldots,n. \tag{15}$$

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

5. Hierarchical Video Structure Summarization

In this paper, we present the video structure analysis in three steps. First, video shot boundaries are detected using QR-decomposition-based algorithm has been presented in [38]. In the second step, the shot level keyframes are detected by eliminating near duplicated frames. Third, shot-level keyframes are further clustered to find common scenes. With this scheme, a three-layer video structure is summarized. The following gives details of the technical implementation using QR-decomposition.

5.1. Shot-Level Video Summarization. In this paper, to detect the video shot boundaries, we apply a QR-decomposition based algorithm presented in [38]. 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 $i = 1, 2, \ldots, n$, we created an m-dimensional feature vector for each frame j such as $X_j^{(i)} = [X_{1,j}^{(i)}, X_{2,j}^{(i)}, \ldots, X_{m,j}^{(i)}]^T$. using $X_j^{(i)}$ as column vector j, we obtained feature matrix $X^{(i)}$ for shot i as follows:

$$X^{(i)} = \left[X_1^{(i)}, X_2^{(i)}, \ldots, X_t^{(i)}\right] = \begin{bmatrix} X_{1,1}^{(i)} & X_{1,2}^{(i)} & \ldots & X_{1,n_i}^{(i)} \\ X_{2,1}^{(i)} & X_{2,2}^{(i)} & \ldots & X_{2,n_i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m,1}^{(i)} & X_{m,2}^{(i)} & \ldots & X_{m,n_i}^{(i)} \end{bmatrix}. \tag{16}$$

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 [46]. Specially, we created a 1728-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 jth column, we generated the feature matrix $X^{(i)}$ for shot i in the video sequence.

Also, suppose that the user of video summarization system is interested to take out a video summary of length Len. We compute NOK_i, $i = 1, 2, \ldots, n$, the number of keyframes must be selected from each shot according to

previous section. In addition, for each shot i, the QR-decomposition of its feature matrix is given as

$$X^{(i)}\Pi_i = Q_i R_i, \quad \text{for } i = 1, \ldots, n. \tag{17}$$

Now, according to Theorem 4.1, the corresponding frames to the first NOK_i columns of R_i are selected as the keyframes of each shot i, where $i = 1, \ldots, n$.

5.2. Scene-Level Summarization. After shot-level keyframes are extracted, these keyframes may still share common scenes. To eliminate this redundancy, further clustering and selection are required.

Among various clustering algorithms, k-means clustering is a practical and easy method for this kind of problem. k-means is a clustering algorithm to partition a set of n data items into K clusters (where $K < n$), which is very similar to the expectation-maximization (EM) algorithm for mixtures of Gaussians in that they both attempt to find the centers of clusters in the data. To attain the target of clustering, k-means algorithms are based on minimization of the following objective function:

$$\Gamma = \sum_{j=1}^{K} \sum_{x_{i \in C_j}} \left\| x_i - c_j \right\|, \tag{18}$$

where there are K clusters S_j, $i = 1, 2, \ldots, K$, and c_j is the centroid or mean point of all the points belong to S_j, and $\| \cdot \|$ is any norm denoting the distance between any data item and the cluster center. Here, we utilize the Euclidian norm to cluster shot level keyframes.

In order to find a suitable solution to the above equation, the number of clusters should be determined before the k-means begin. We apply a modified version of k-means to calculate the proper number of clusters iteratively. The algorithm starts initially with two clusters. We use a predefined threshold $T_{\max}$ to terminate the iterative algorithm. In each iteration, the algorithm compute distance between the cluster centers. If their mean distances are greater than the predefined threshold, then the number of clusters is increased. Otherwise, the procedure is stopped, and the current number of clusters is considered as a proper solution. Suppose that $\{c_i\}^k$ is the centers of clusters at kth iteration, then this procedure can be summarized as follows.

(1) Let there are initially $k = 2$ clusters in the data set. We apply the standard k-means algorithm to obtain two cluster centers with corresponding coordinates.

(2) Compute the Euclidean distances between centers of clusters. If their mean distances are greater than the predefined threshold $T_{\max}$, then the optimal cluster number k is increased to $k + 1$. Otherwise, the procedure is completed and k is selected as suitable cluster number. The final cluster is what we are looking for.

Consequently, when the clustering procedure is finished, the corresponding frames of cluster centers are considered as scene level keyframes.

The value of $T_{\max}$ controls the number of scene level keyframes and the amount of similarity between clusters. If $T_{\max}$ is small, then the number of keyframes and the similarity between clusters will be high, and if $T_{\max}$ is large, then clusters will be dissimilar and the number of scene level clusters will be low. In our experiments, we conclude that $T_{\max} = 0.2$ is proper, and the extracted scene level summaries are desirable.

Here, we utilize some feature vectors that extracted from keyframes of the shot-level summaries as inputs for clustering algorithm. In particular, at first, the color histogram matrix of the keyframes is computed according to Section 5.1. Then, the QR-decomposition of the feature matrix is computed. The columns of Q-matrix are used as inputs for clustering algorithm. The proposed clustering algorithm is run and clusters are computed.

6. 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 large-scale test set provided by the TRECVID 2006 [47], which has assessments for keyframe 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.

6.1. Evaluation Criteria. Evaluation of the video summaries obtained by the various keyframe extraction techniques is one of the important issues in the field of video analysis and summarization. The mainly universal assessment of a summary is based on the personal opinion of a group of clients. For example, in [48], Fayzullin et al. define three properties that must be taken into account when creating video summary: continuity, priority, and repetition; in [49, 50], the authors considered a global subjective assessment to measure the goodness of the summary; in [51], Liu et al. asked from users to mark the summary as good, acceptable, or bad rely on their approval.

The problem of these measures is that their assessment is mostly subjective and cannot be used to analyze video sequences automatically. In this paper, we have utilized a more objective, general purpose to summary assessment; one that does not consider the type of video being proceed and can be automatically applied to all video sequences without requiring the services of video experts.

A video summary is considered good if the set of keyframes effectively represents the pictorial content of the video sequence. From a broad variety of objective measures for assessment of goodness of video summary, we have chosen two well-known measures: fidelity measure which is defined in [52] and Shot Reconstruction Degree (SRD) which is suggested in [53]. Fidelity applies a global strategy, while the SRD utilizes a local assessment of the keyframes.

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

$$\text{dist}(F_i, \text{Keys}) = \min_{1 \le j \le \text{len}} \left\{ \left\| \text{Feature}(F_i) - \text{Feature}\left(F_{k_j}\right) \right\| \right\}, \tag{19}$$

where $\|\cdot\|$ is a proper distance function and Feature$(\cdot)$ 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 the set of keyframes is defined as

$$\text{Dist}(V, \text{keys}) = \max_{1 \le i \le n} \{\text{dist}(F_i, \text{Keys})\}. \tag{20}$$

And consequently, the fidelity criterion is defined as follows:

$$\text{Fidelity}(V, \text{keys}) = \max_{\text{Dist}} - \text{Dist}(V, \text{keys}), \tag{21}$$

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

6.1.2. SRD Measure. This measure evaluates the goodness of the keyframe 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 keyframes will summarize the visual content of the video appropriately.

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

$$\widetilde{F}_i = \text{FIA}(\text{Keys}, i), \tag{22}$$

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

$$\text{SRD}(V, \text{Keys}) = \sum_{n=i}^{\gamma} \text{Sim}\left(F_i, \widetilde{F}_i\right). \tag{23}$$

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

6.2. Results. In order to evaluate the effectiveness of the proposed system, we carry out the three various experiments. First, video shot boundaries are detected using QR-decomposition-based algorithm has been presented in [38]. In the second step, the shot level keyframes are detected by eliminating near duplicated frames. Third, shot-level keyframes are further clustered to find common scenes. With this scheme, a three-layer video structure is summarized. In addition, there are two parameters in the proposed algorithm that must be determined manually: β in (10) and $T_{\max}$ in Section 5.2. The other parameters such as the length of video summary and the number of keyframes should be determined by external user. In our experiments, we consider $T_{\max} = 0.2$ and $\beta = 0.33$. The following gives details of the technical implementation using QR-decomposition.

In Figure 2, the shot dynamicity measure of (12) 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, semidynamic, and dynamic. A 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.

In Figure 3, some of the frames of a static shot are presented. This shot has been summarized through the proposed system using 3 and 7 keyframes. As we can see, the visual contents obtained from the three summarizations are the same. The obtained dynamicity criterion of (12) for this shot is 0.17.

A semidynamic 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. In Figure 4, some of the frames of a semidynamic shot are displayed. This shot has been summarized through the proposed system using 3 and 10 keyframes. As you can see, the visual content of the summarization with 3 keyframes differs from that of the 10 keyframes summarization. But the visual contents of the summarizations with 10 keyframes and the 21 frames of the shot are almost the same. The obtained dynamicity criterion of (13) for this shot is 0.63.

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. In Figure 5, some of the frames of a dynamic shot are shown. This shot has been summarized through the proposed system using 5 and 15 keyframes. As we can see, the visual contents of the summarizations are completely different. The obtained dynamicity criterion of (13) for this shot is 0.89. As is seen, many keyframes have been used to summarize this shot precisely.

Therefore, as mentioned in a previous section, to control the number of keyframes will be extracted from each shot, we can use shot-dynamicity measure.

To compare the effectiveness of the proposed system with other systems, together using our QR-decomposition-based algorithm, we experienced a simple Time-Sampled (TS) method [54], the k-Means with SVD (KMSVD) approach [55] and the Information Theory (IT-) based algorithm [56]. In our implementation, all methods are required to produce the same number of keyframes which is determined by user. Also, to judge the performance of our algorithm with respect to the other algorithms, it is useful to express the results

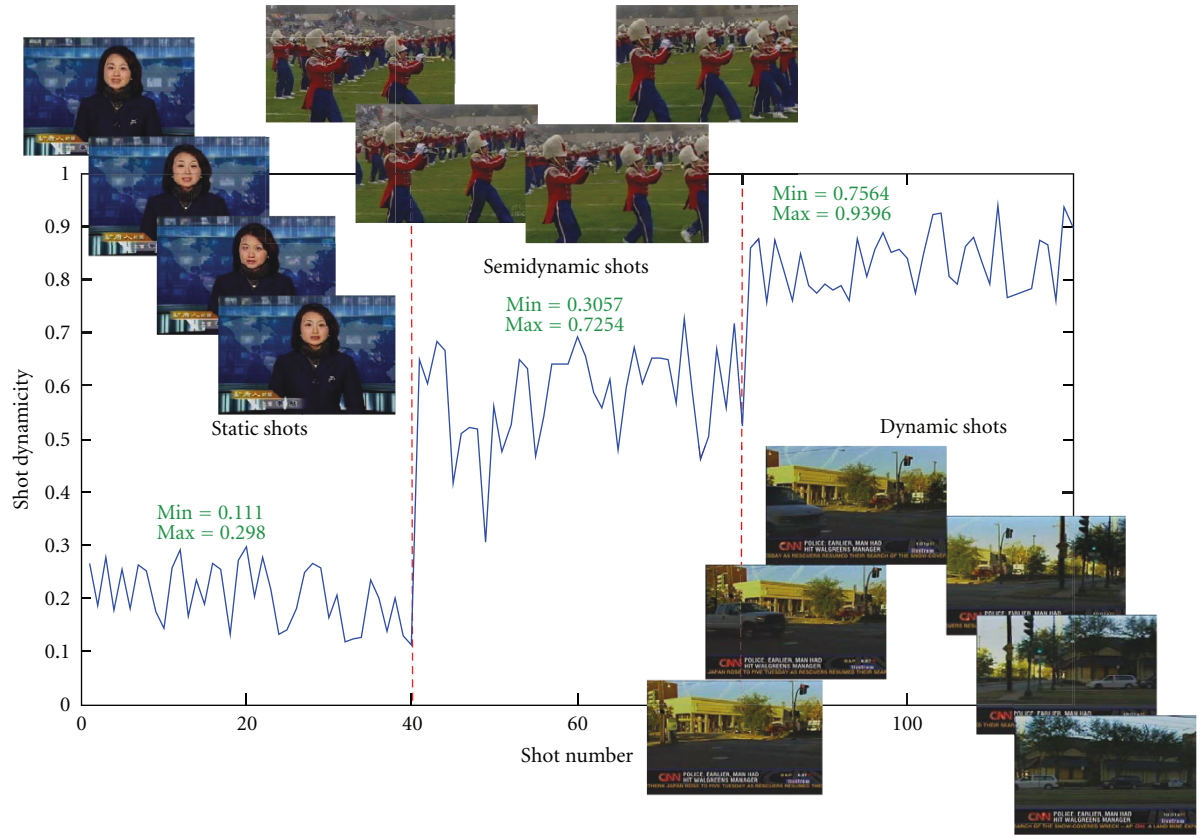


FIGURE 2: Three categories of the video shots based on shot dynamicity measure: static, semidynamic, and dynamic.

(a) 18 Frames of a static shot

(b) Summarization with NOK = 3

(c) Summarization with NOK = 7

FIGURE 3: Some of the frames of a static shot which has been summarized with 3 and 7 keyframes.

as a measure of relative improvement using the following formula:

$$\Delta(\mathrm{QR}_{\mathrm{Method}}, X_{\mathrm{Method}}) = \frac{\mathrm{Measure}(\mathrm{QR}_{\mathrm{Method}}) - \mathrm{Measure}(X_{\mathrm{Method}})}{\mathrm{Measure}(X_{\mathrm{Method}})}, \quad (24)$$

where measure corresponds to the fidelity or the SRD measure, and we substitute X_{Method} with the algorithm we used to compare with our proposed $\mathrm{QR}_{\mathrm{Method}}$ algorithm.

Figures 6 and 7 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

Table 1: Details of all the video of TRECVID 2006 and the average (AV) and standard derivation (STD) of their shot dynamicity.

File Name	Hard cut	Dissolve	Fade in/out	Other	Shot dynamicity	
					AV	STD
20051101_142800_LBC_NAHAR_ARB.mpg	45	156	12	30	0.212	0.051
20051227_125800_CNN_LIVEFROM_ENG.mpg	44	17	1	9	0.851	0.037
20051227_105800_MSNBC_NEWSLIVE_ENG.mpg	339	96	6	79	0.273	0.077
20051213_185800_PHOENIX_GOODMORNCN_CHN.mpg	81	164	1	25	0.569	0.095
20051209_125800_CNN_LIVEFROM_ENG.mpg	57	25	1	11	0.495	0.041
20051208_182800_NBC_NIGHTLYNEWS_ENG.mpg	424	142	9	63	0.347	0.053
20051208_145800_CCTV_DAILY_CHN.mpg	139	279	5	18	0.832	0.069
20051208_125800_CNN_LIVEFROM_ENG.mpg	244	74	8	41	0.174	0.038
20051231_182800_NBC_NIGHTLYNEWS_ENG.mpg	120	103	4	39	0.880	0.041
20051114_091300_NTDTV_FOCUSINT_CHN.mpg	121	59	0	18	0.183	0.033
20051115_192800_NTDTV_ECONFRNT_CHN.mpg	74	91	0	3	0.312	0.081
20051129_102900_HURRA_NEWS_ARB.mpg	24	107	1	16	0.159	0.043
20051205_185800_PHOENIX_GOODMORNCN_CHN.mpg	132	196	3	29	0.432	0.055

(a) 21 frames of a semidynamic shot

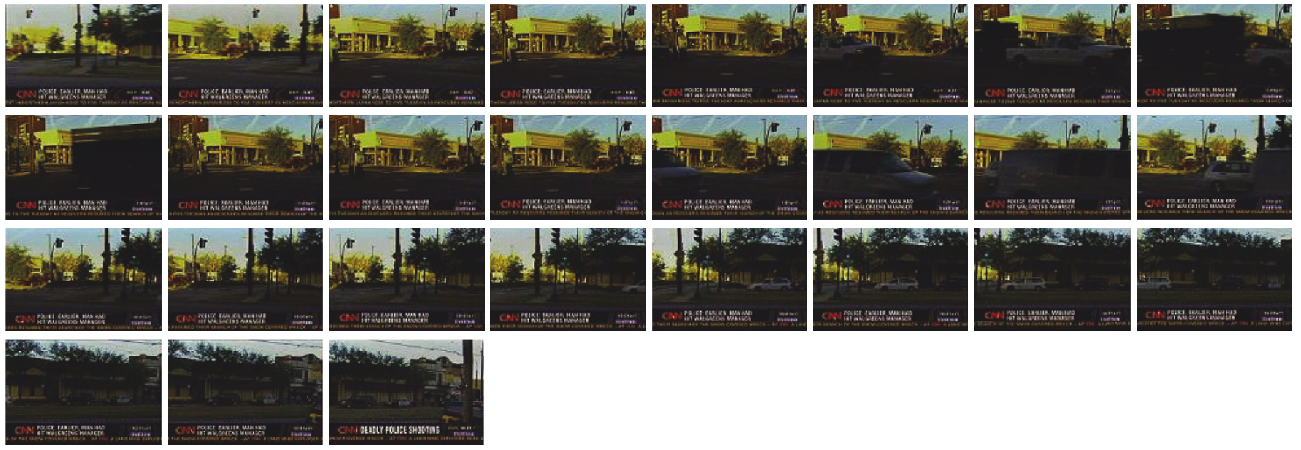
(b) Summarization with NOK = 3

(c) Summarization with NOK = 10

Figure 4: Some of the frames of a semidynamic shot which has been summarized with 3 and 10 keyframes.

of keyframes per shot is small the differences between the proposed and other algorithms are slight and as the number of keyframes per shot increases, the gap is more marked.

The reference video test set TRECVID 2006 is a large set of video data that can be used for comparing keyframe-based video summarization algorithms. This set consists of 13 video files with size 4.46 GB and contains 3785 shots of different types. The names of the video files together with the types and number of their shots, and also the average and standard derivation of shot dynamicity measure for each



(a) 27 frames of a dynamic shot

(b) Summarization with NOK = 5

(c) Summarization with NOK = 15

Figure 5: Some of the frames of a dynamic shot which has been summarized with 5 and 15 keyframes.

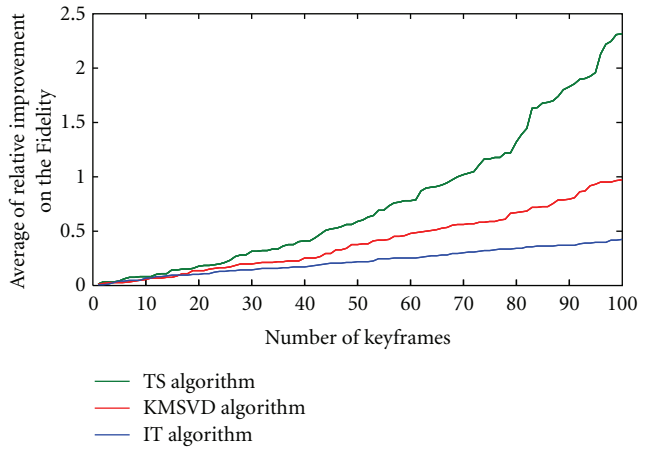


Figure 6: the average of relative improvement of fidelity (the results are in percentages).

Table 2: Average (AV) and standard derivation (STD) of the relative improvement (Δ) measured on the fidelity and SRD results of the proposed algorithm with respect to each of the other algorithms using all the videos of TRECVID 2006.

Video name	Fidelity						SRD					
	IT		KMSVD		TS		IT		KMSVD		TS	
	AV	STD	AV	STD	AV	STD	AV	STD	AV	STD	AV	STD
20051101_142800_LBC_NAHAR_ARB.mpg	4.3	1.2	16.3	3.7	34.8	5.7	15.7	2.9	28.6	4.9	104.7	8.9
20051227_125800_CNN_LIVEFROM_ENG.mpg	3.9	1.9	14.7	2.8	35.1	4.1	17.1	3.4	37.2	2.4	98.2	10.3
20051227_105800_MSNBC_NEWSLIVE_ENG.mpg	3.2	0.6	8.5	1.9	20.1	3.7	12.6	2.3	31.3	5.6	89.4	6.1
20051213_185800_PHOENIX_GOODMORNCN_CHN.mpg	3.5	1.1	10.6	2.4	2.7	1.9	11.1	3.8	42.7	4.8	101.3	8.7
20051209_125800_CNN_LIVEFROM_ENG.mpg	3.7	1.3	12.1	1.4	23.3	3.1	13.9	4.9	45.2	3.5	109.7	5.1
20051208_182800_NBC_NIGHTLYNEWS_ENG.mpg	3.8	1.7	14.0	2.6	26.5	2.9	11.3	3.2	34.8	3.9	85.3	6.2
20051208_145800_CCTV_DAILY_CHN.mpg	3.1	0.8	7.6	0.9	15.3	1.6	10.9	3.6	36.6	3.2	91.9	6.9
20051208_125800_CNN_LIVEFROM_ENG.mpg	4.1	0.5	11.8	2.3	23.5	3.5	9.8	4.3	34.1	2.7	87.6	4.6
20051231_182800_NBC_NIGHTLYNEWS_ENG.mpg	3.2	1.1	6.8	0.8	15.6	2.5	12.1	1.8	43.4	1.8	102.3	7.1
20051114_091300_NTDTV_FOCUSINT_CHN.mpg	4.2	1.6	14.5	1.5	27.2	3.4	10.8	2.1	29.8	4.6	79.4	4.8
20051115_192800_NTDTV_ECONFRNT_CHN.mpg	3.6	0.7	8.9	1.1	19.1	3.9	16.4	2.0	40.4	3.1	112.1	3.9
20051129_102900_HURRA_NEWS_ARB.mpg	3.9	1.0	12.3	1.6	20.2	2.7	13.0	3.3	37.1	2.8	105.4	8.3
20051205_185800_PHOENIX_GOODMORNCN_CHN.mpg	3.3	1.8	7.2	1.9	15.3	3.3	9.1	1.9	32.8	3.6	93.6	5.4
Weighted average (based on number of shots in each video)	3.6	1.1	10.7	1.9	22.3	3.1	11.8	2.9	35.4	3.8	93.7	6.4

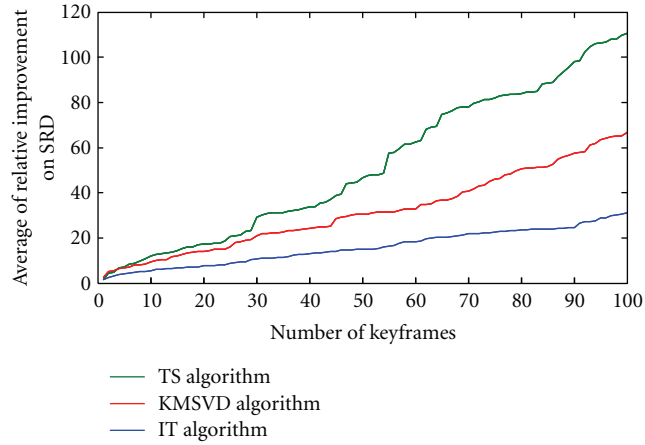


Figure 7: The average of relative improvement of SRD (the results are in percentages).

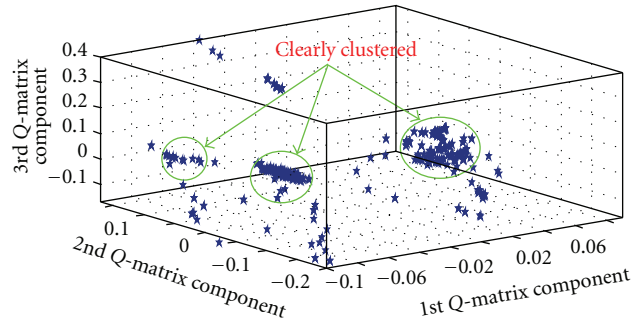


Figure 8: Projection of feature vectors of keyframes frames of four shot-level summaries into the Q-matrix subspace.

video file have been demonstrated in Table 1. In addition, the average and standard derivation of the relative improvement ($\Delta(\mathrm{QR}_{\mathrm{Method}}, X_{\mathrm{Method}})$) of fidelity and SRD for each video file have been illustrated in Table 2.

Summarization in shot level may not be able to decrease the video redundancies in the scene level. Therefore, in addition to eliminating shot-level redundancies, it is necessary to remove scene-level redundancies as well. In other word, many times it is possible that in some different shots of a scene the visual contents are nearly the same, and therefore they contain scene-level redundancies. In order to remove these redundancies, we use the proposed method

in Section 5.2. Figure 8 is the projected result of feature vectors of 460 keyframes of four shot-level summaries into the Q-matrix subspace. As shown, we can clearly see frames are clustered together. This confirms that we can detect redundancy between keyframes using Q-space and modified k-means clustering algorithm.

7. 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 QR-decomposition. 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.

Appendix

Proof of Theorem 4.1. Let $A = [A_1, \ldots, A_i, \ldots, A_n]$ be an $m \times n$ matrix with linearly independent columns. The Gram-Schmidt [34] orthonormalization method is applied to the columns of A. The result is an orthonormal basis $= [q_1, \ldots, q_i, \ldots, q_n]$, where

$$q_1 = \frac{A_1}{v_1}, \qquad q_j = \frac{A_j - \sum_{h=1}^{j-1} \langle q_h, A_j \rangle q_h}{v_j} \quad \text{for } j = 2, 3, \ldots, n, \tag{A.1}$$

where $v_1 = \|A_1\|$, $v_j = \|A_j - \sum_{h=1}^{j-1} \langle q_h, A_j \rangle q_h\|$ for $j > 1$, and $\langle q_h, A_j \rangle$ is the inner product of vectors q_h and A_j. The above relations can be written as

$$A_1 = v_1 q_1,$$
$$A_j = \langle q_1, A_j \rangle q_1 + \cdots + \langle q_{j-1}, A_j \rangle q_{j-1} + v_j q_j \quad \text{for } j > 1, \tag{A.2}$$

which in turn can be expressed in matrix form by writing

$$A = [A_1, \ldots, A_i, \ldots, A_n] = \underbrace{[q_1, \ldots, q_i, \ldots, q_n]}_{Q} \times \underbrace{\begin{bmatrix} v_1 & \langle q_1, A_1 \rangle & \langle q_1, A_2 \rangle & \cdots & \langle q_1, A_n \rangle \\ 0 & v_2 & \langle q_2, A_2 \rangle & \cdots & \langle q_2, A_n \rangle \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \ddots & v_i & & \langle q_i, A_n \rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & v_n \end{bmatrix}}_{R}. \tag{A.3}$$

This procedure demonstrates an algorithm based on Gram-Schmidt approach for computing matrices Q and R. Now, 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 $(A_i^{(1)} = \cdots = A_i^{(k)} = A_i)$, then according to the above procedure, the QR-decomposition for or columns $A_i^{(1)}, \ldots, A_i^{(k)}$ of $\widetilde{A}$ can be computed as follows:

$$\begin{aligned} A_i^{(1)} &= \langle q_1, A_i \rangle q_1 + \cdots + \langle q_{i-1}, A_i \rangle q_{i-1} + v_i q_i, \\ A_i^{(2)} &= \langle q_1, A_i \rangle q_1 + \cdots + \langle q_{i-1}, A_i \rangle q_{i-1} + \langle q_i, A_i \rangle q_i + v_{i+1} q_{i+1}, \\ &\vdots \\ A_i^{(k)} &= \langle q_1, A_i \rangle q_1 + \cdots + \langle q_{i-1}, A_i \rangle q_{i-1} + \langle q_i, A_i \rangle q_i \\ &\quad + \ldots + v_{i+k-1} q_{i+k-1}. \end{aligned} \tag{A.4}$$

The coefficient v_j for $j = i, i+1, \ldots, i+k-1$, in these relation can be computed as follows:

$$\begin{aligned} v_j &= \left\| A_j - \sum_{h=1}^{j-1} \langle q_h, A_j \rangle q_h \right\| = \left\| A_i - \sum_{h=1}^{j-1} \langle q_h, A_i \rangle q_h \right\| \\ &= \|A_i - A_i\| = 0. \end{aligned} \tag{A.5}$$

In other word, in the corresponding R-values of columns $A_i^{(1)}, \ldots, A_i^{(k)}$ in the QR-decomposition of matrix $\widetilde{A}$ will be zero. Now, by applying the Rank-Revealing algorithm [57], the R-values are sorted in decreasing order, and the proof is completed. □

References

[1] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," *Multimedia Tools and Applications*, vol. 11, no. 3, pp. 347–358, 2000.

[2] M. Mills, J. Cohen, and Y. Y. Wong, "Magnifier tool for video data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 93–98, May 1992.

[3] Y. Taniguchi, A. Akutsu, Y. Tonomura, and H. Hamada, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing," in *Proceedings of the 3rd*

International Multimedia Conference and Exhibition (Multimedia '95), pp. 25–33, November 1995.

[4] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 90–105, 2002.

[5] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings of the International Conference on Image Processing (ICIP '98)*, pp. 866–870, October 1998.

[6] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber, "Summarization of videotaped presentations: automatic analysis of motion and gesture," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 686–696, 1998.

[7] C. Toklu and S. Liou, "Automatic key-frame selection for content-based video indexing and access," in *Proceedings of the Storage and Retrieval for Media Databases*, vol. 3972 of *Proceeding of SPIE*, pp. 554–563, January 2000.

[8] M. Irani and P. Anandan, "Video indexing based on mosaic representations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 86, no. 5, pp. 905–921, 1998.

[9] N. Vasconcelos and A. Lippman, "A spatiotemporal motion model for video summarization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 361–366, June 1998.

[10] Y. Taniguchi, A. Akutsu, and Y. Tonomura, "PanoramaExcerpts: extracting and packing panoramas for video browsing," in *Proceedings of the 5th ACM International Multimedia Conference (Multimedia '97)*, pp. 427–436, November 1997.

[11] A. D. Doulamis, N. Doulamis, and S. Kollias, "Non-sequential video content representation using temporal variation of feature vectors," *IEEE Transactions on Consumer Electronics*, vol. 46, no. 3, pp. 758–768, 2000.

[12] J. Peng and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," *IEEE Multimedia*, vol. 17, no. 2, pp. 64–73, 2010.

[13] J. Ćalić, D. P. Gibson, and N. W. Campbell, "Efficient layout of comic-like video summaries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 7, pp. 931–936, 2007.

[14] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent key frames selection based on iso-content principles," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 447–451, 2009.

[15] J. You, G. Liu, L. Sun, and H. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 273–285, 2007.

[16] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video manga: generating semantically meaningful video summaries," in *Proceedings of the 7th International Multimedia Conference and Exhibition (Multimedia '99)*, pp. 383–392, November 1999.

[17] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," in *Proceedings of the 6th IEEE International Conference on Multimedia Computing and Systems (ICMCS '99)*, pp. 756–761, June 1999.

[18] M. M. Yeung and B.-L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 5, pp. 771–785, 1997.

[19] X. Sun and M. S. Kankanhalli, "Video summarization using R-sequences," *Real-Time Imaging*, vol. 6, no. 6, pp. 449–459, 2000.

[20] I. Otsuka, K. Nakane, A. Divakaran, K. Hatanaka, and M. Ogawa, "A highlight scene detection and video summarization system using audio feature for a personal video recorder," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 1, pp. 112–116, 2005.

[21] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1342–1355, 2008.

[22] G. Ciocca and R. Schettini, "Supervised and unsupervised classification post-processing for visual video summaries," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 2, pp. 630–638, 2006.

[23] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1026–1038, 2006.

[24] C. Choudary and T. Liu, "Summarization of visual content in instructional videos," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1443–1455, 2007.

[25] K. Ratakonda, M. I. Sezan, and R. Crinon, "Hierarchical video summarization," in *Visual Communications and Image Processing '99*, vol. 3653 of *Proceedings of SPIE*, pp. 1531–1541, January 1999.

[26] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref, "Exploring video content structure for hierarchical summarization," *Multimedia Systems*, vol. 10, no. 2, pp. 98–115, 2004.

[27] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296–305, 2005.

[28] Q. Shen, Y. Guo, H. Li, and F. Wu, "Intermediate description for multiple video adaptation," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 919–926, 2009.

[29] B.-W. Chen, J.-C. Wang, and J.-F. Wang, "A novel video summarization based on mining the story-structure and semantic relations among concept entities," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 295–312, 2009.

[30] Z. Li, G. M. Schuster, and A. K. Katsaggelos, "MINMAX optimal video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1245–1256, 2005.

[31] Y. Gao, W.-B. Wang, and J.-H. Yong, "A video summarization tool using two-level redundancy detection for personal video recorders," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 521–526, 2008.

[32] R. M. Jiang, A. H. Sadka, and D. Crookes, "Hierarchical video summarization in reference subspace," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1551–1557, 2009.

[33] Y. Gong and X. Liu, "Video summarization and retrieval using singular value decomposition," *Multimedia Systems*, vol. 9, no. 2, pp. 157–168, 2003.

[34] G. Golub and C. Loan, *Matrix Computations*, Johns-Hopkins, Baltimore, Md, USA, 2nd edition, 1989.

[35] M. Seines and R. Babuška, "Rule base reduction: some comments on the use of orthogonal transforms," *IEEE Transactions on Systems, Man and Cybernetics C*, vol. 31, no. 2, pp. 199–206, 2001.

[36] O. Kaynak, K. Jezernik, and A. Szeghegyi, "Complexity reduction of rule based models: a survey," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 1216–1221, May 2002.

[37] M. Amintoosi, F. Farbiz, and M. Fathy, "A QR decomposition based mixture model algorithm for background modeling," in *Proceedings of the 6th IEEE International Conference on Information, Communications and Signal Processing (ICICS '07)*, pp. 1–5, December 2007.

[38] A. Amiri and M. Fathy, "Video shot boundary detection using QR-decomposition and gaussian transition detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 509438, 12 pages, 2009.

[39] J. Ye, J. Ravi, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1312–1322, 2006.

[40] M. Cheng, B. Fang, Y.-Y. Tang, and J. Wen, "An efficient regularized neighborhood discriminant analysis through QR decomposition," in *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR '08)*, pp. 304–309, August 2008.

[41] J. Ye and Q. Li, "A two-stage linear discriminant analysis via QR-decomposition," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 27, no. 6, pp. 929–941, 2005.

[42] X. Liu, "Video shot segmentation and classification," in *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 1, pp. 860–863.

[43] Z. Černeková, C. Kotropoulos, and I. Pitas, "Video shot-boundary detection using singular-value decomposition and statistical tests," *Journal of Electronic Imaging*, vol. 16, no. 4, pp. 51–59, 2007.

[44] C. H. Bischof and G. M. Shroff, "On updating signal subspaces," *IEEE Transactions on Signal Processing*, vol. 40, no. 1, pp. 96–105, 1992.

[45] G. W. Stewart, "Updating a rank-revealing ULV decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 14, pp. 494–499, 1993.

[46] W. Cheng, D. Xu, Y. Jiang, and C. Lang, "Information theoretic metrics in shot boundary detection," in *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES '05)*, vol. 3683 of *Lecture Notes in Computer Science*, pp. 388–394, 2005.

[47] NIST, "Homepage of Trecvid Evaluation," http://www-nlpir.nist.gov/projects/trecvid/.

[48] M. Fayzullin, V. S. Subrahmanian, A. Picarello, and M. L. Sapino, "The CPR model for summarizing video," in *Proceedings of the 1st ACM International Workshop on Multimedia Databases*, pp. 2–9, New Orleans, La, USA, 2002.

[49] R. Narasimha, A. Savakis, R. M. Rao, and R. De Queiroz, "A neural network approach to key frame extraction," in *Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307 of *Proceedings of SPIE*, pp. 439–447, 2004.

[50] J. Calic and E. Izquierdo, "Efficient key-frame extraction and video analysis," in *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC '02)*, pp. 28–33, 2002.

[51] T. Liu, H.-J. Zhang, and F. Qi, "A novel video keyframe-extraction algorithm based on perceived motion energy model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 10, pp. 1006–1013, 2003.

[52] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1269–1279, 1999.

[53] T. Liu, X. Zhang, J. Feng, and K.-T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1451–1457, 2004.

[54] Y. Rui and T. S. Huang, "Exploring video structure beyond the shots," in *Proceeding of IEEE International Conference on Multimedia Computing and Systems (ICMCS '98)*, pp. 237–240, Austin, Tex, USA, 1998.

[55] S. Lee and M. H. Hayes, "Properties of the singular value decomposition for efficient data clustering," *IEEE Signal Processing Letters*, vol. 11, no. 11, pp. 862–866, 2004.

[56] Z. Cernekova and I. Pitas, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, 2006.

[57] Y. P. Hong and C.-T. Pan, "Rank-revealing QR factorizations and the singular value decomposition," *Mathematics of Computation*, vol. 58, no. 197, pp. 213–232, 1992.