Hierarchical Video Summarization in Reference Subspace

Richard M. Jiang, Abdul H. Sadka, Danny Crookes

Abstract *--- In this paper, a hierarchical video structure summarization approach using Laplacian Eigenmap is proposed, where a small set of reference frames is selected from the video sequence to form a reference subspace to measure the dissimilarity between two arbitrary frames. In the proposed summarization scheme, the shot-level key frames are first detected from the continuity of inter-frame dissimilarity, and the sub-shot level and scene level representative frames are then summarized by using K-mean clustering. The experiment is carried on both test videos and movies, and the results show that in comparison with a similar approach using latent semantic analysis, the proposed approach using Laplacian Eigenmap can achieve a better recall rate in keyframe detection, and gives an efficient hierarchical summarization at sub shot, shot and scene levels subsequently.*

Index Term **---** Hiearchical Video Summarization, Latent Semantic Analysis, Laplacian Eigenmap, Representative Frame.

I. INTRODUCTION

With the explosion of multimedia databases due to the growth in internet and wireless multimedia technology, the management of vast video content demands automatic summarization to abstract the most relevant content or useful information from a massive visual data set [1]. Recent advances [1-4] in this area have successfully generated a number of practical systems, such as VideoCollage and VideoSue [5].

Video summarization refers to creating an excerpt of a digital video, which must contain high priority entities and events from the video and exhibit reasonable degrees of continuity with little repetition. The challenge in video summarization is how to effectively extract certain content of the video while preserving the essential message of the original video [1-4]. There are basically two types of video abstraction: static video summarization and dynamic video skimming. Video summarization is a process that selects a set of salient images called key frames to represent the video content, while video skimming represents the original video in the form of a short video clip. Both ways are actually similar to each other; they share temporal video structure analysis for implementing the finding of content or frames of specific interest to a user for video browsing. Fig.1 shows an outline of such video summarization systems.

Richard M. Jiang is with Computer Science, Loughborough University, UK. Abdul H. Sadka is with CMCR, Brunel University. Danny Crookes is with ECIT, Queen's University Belfast, UK. Correspondence Email: M.Jiang@lboro.ac.uk.

Contributed Paper
Manuscript received July 6, 2009

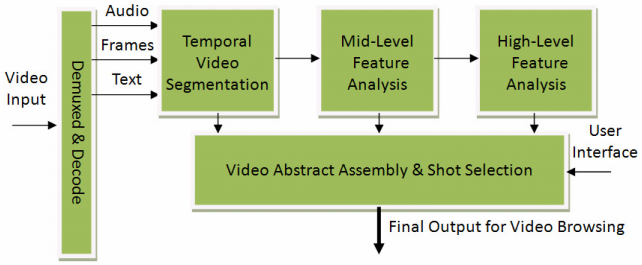


Fig. 1. A Three Layer Video Abstraction System

Based on the way a key frame is extracted for video summarization, existing work in this area can be categorized into three classes: sampling based, shot based, and segment based. Most of the earlier summarization work belongs to the first class, where key frames were either uniformly sampled or randomly chosen from the original video. The MiniVideo [6] and the magnifier [7] systems are such two examples. This approach is the simplest way to extract key frames, but such an arrangement may fail to capture the real video content, especially when it is highly dynamic.

More sophisticated work has since been developed to extract key frames by adapting to dynamic video content. Since a shot is defined as a video segment taken from a continuous period, a natural and straightforward way is to extract one or more key frames from each shot using low-level features such as color and motion. A typical approach in [8] extracts key frames in a sequential fashion via thresholding. Other schemes may include the use of color clustering, global motion, or gesture analysis [9-11].

Various clustering-based extraction schemes at the higher representative scene-level have been also proposed. In these schemes, segments are first generated from frame clustering and then the frames that are closest to the centroid of each qualified segment are chosen as key frames [12~13]. Yeung and Yeo [14] reported their work on video summarization at the scene level based on a detected shot structure, where they classified all shots into a group of clusters and then extracted meaningful scenes, namely representative images (R-images) to represent its component shot clusters.

Though there may have various video summarization schemes, they all share a basic infrastructure in two aspects: using specific features as the continuity metric for temporal video segmentation, and setting up decision methods (include fixed thresholds, adaptive thresholds, and statistical detection methods) for dissimilarity-based classification. Feature selection for video summarization involves a number of choices, such as HSV/YUV color feature [15~16], color or color-spatial histogram [17], edge information [18], motion features [19~20], and DFT/DCT/DWT coefficients [21~22].

Recently, singular value decomposition (SVD) and latent semantic analysis (LSA) [23~25] emerges as an attractive computational model for video summarization because

Eigenfeatures are usually the most representative and discriminant features for frame comparison. It can also put all frames into a balanced comparison, and thus the overall video structure can be hierarchically organized with adaptive thresholds. However, this approach is computationally intensive since it operates directly on video frames.

LSA is usually considered as a linear scale-space approach, which has disadvantages such as low dimensional reduction rate and low accuracy. Recently, nonlinear scale-space approaches, such as kernel principle component analysis (KPCA) [26] and Laplacian Eigenmaps [27~29], have been considered as more efficient ways for discriminant information extraction. In this paper, we propose a reference frame subspace approach using Laplacian Eigenmap, which selects a limited number of reference frames to form a Laplacian Eigen subspace to measure the inter-frame dissimilarity. The proposed video summarization scheme is tested with standard test videos and movies, which shows the proposed approach can efficiently perform video summarization.

In the rest of the paper, section II describes the proposed approach, section III gives the experimental results, and section IV concludes the paper.

II. Video Analysis in Laplacian Subspace

A. Laplacian Eigenmap

Conventional latent semantic analysis (LSA) [23~25] projects the frames into its Eigen subspace, where LSA is applied to extract a subspace in which the variance is maximized. Its objective function is as follows:

$$\max_W \sum_{i=1}^{n} (y_i - \bar{y})^2, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

Recent non-linear scale-space approaches [26-30] have recently been intensively researched. Among these approaches, Laplacian Eigenmap [27-30] has been considered among the best ways that can outperform the traditional linear SVD or LSA approaches. The Laplacian approach seeks to preserve the intrinsic geometry of the data and local structure. The objective function of Laplacian Eigenmap is as follows:

$$\max_W \sum_{ij} (y_i - y_j)^2 S_{ij}, \quad \text{where, } S_{ij} = \begin{cases} \exp\left(-\|x_i - x_j\|^2 / \alpha\right), \|x_i - x_j\|^2 < \varepsilon \\ 0 \qquad \text{otherwise.} \end{cases} \tag{2}$$

where ||.|| is the Frobenius form, S is a similarity matrix, and ε defines the radius of the local neighbourhood that is sufficiently small, and greater than zero. The objective function with the choice of symmetric weights S_{ij} incurs a heavy penalty if neighbouring data points x_i and x_j are mapped far apart, i.e., if $(y_i - \bar{y})^2$ is large. Therefore, minimizing it is an attempt to ensure that, if x_i and x_j are "close," y_i and y_j then should be close as well. Following some simple algebraic steps, we have,

$$\begin{aligned} &\frac{1}{2}\sum_{ij} (y_i - y_j)^2 S_{ij} \\ &= \frac{1}{2}\sum_{ij} (W^T x_i - W^T x_j)^2 S_{ij} \\ &= \sum_{ij} W^T x_i S_{ij} x_i^T W - \sum_{ij} W^T x_i S_{ij} x_j^T W \\ &= W^T X L X^T W \end{aligned} \tag{3}$$

where

$$L = D - S, \text{ with } Dij = \begin{cases} \sum_j S_{ij}, i = j \\ 0, i \neq j \end{cases} \tag{4}$$

L is called the Laplacian matrix. The problem then becomes:

$$\arg\min_W W^T X L X^T W \tag{5}$$

With its solution W_{Lap}, we have Laplacian projection,

$$x \Rightarrow y = W_{Lap}{}^T x \tag{6}$$

In this projection, the Laplacian graph model is embedded to map the nonlinear data separation problem into a linear separation problem. Details about Laplacian Eigenmap can be found in Ref.[27-30].

B. Dissimilarity in Reference Laplacian Subspace

Similar to LSA, directly applying Laplacian Eigenmap to all frames will encounter the problem of computing efficiency. For instance, a video with 10000 frames will require the computation of a 10000×10000 matrix to obtain its projection matrix W_{Lap}.

In order to resolve this bottleneck, instead of using all frames to extract the projection matrix W_{Lap}, we use a subset of frames selected uniformly from the video sequence as reference frames. Considering K frames selected from the video stream with a regular interval Δ ($K\Delta=N$, N is the total number of frames), we have a set of K frames:

$$\Lambda = [\Lambda_1, \Lambda_2, \ldots, \Lambda_K] \tag{7}$$

From eq.(5), we can obtain the Laplacian projection matrix $W_{Lap}{}^T$ from Λ. As stated in eq.(6), a given image x_i can then be projected into this reference subspace as y_i. The distance of the image x_i from a reference image Λ_k can be calculated as,

$$d_k = \|W_{Lap}{}^T x_i - W_{Lap}{}^T \Lambda_k\| = \|y_i - y_k\| \tag{8}$$

where $\|\cdot\|$ is Hilbert-Schmidt norm, and d_k is the dissimilarity between Mahalanobis distances of two images x_i and x_k.

In this paper, the dissimilarity between a given image and a set of the selected frames $\{\Lambda_k\}$ forms a dissimilarity vector D^i,

$$D^i = \{d_1, d_2, \ldots, d_K\} \tag{9}$$

Thus, any frame can be simply featured by its similarity projection D^i in the dissimilarity subspace $\mathbb{R}^K$ provided by reference frames.

With this Laplacian subspace projection, the dissimilarity between any two frames x_i and x_j can be computed by their distance in this reference subspace,

$$ds^{\{i,j\}} = \|D^i - D^j\| \tag{10}$$

In temporal video structure analysis, the most useful information is the dissimilarity between neighbouring frames i and (i+1), namely $ds^{\{i, i+1\}}$.

C. Hierarchical Video Structure Summarization

With the overall dissimilarity measure in Eq.(10), it is not difficult to estimate the hierarchical temporal video structure. In this paper, we perform the video structure analysis in three steps. First, scene changes are evaluated from the frame-level dissimilarity continuity in video sequence. In the second step, the sub-shot key frames are detected by comparing intra-shot dissimilarity of all frames in a shot. Third, shot-level key frames 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 the Laplacian Eigen features.

1) Shot-level Temporal Video Segmentation

To detect a video shot boundary, in this paper, rather than using any predefined threshold, we apply an adaptive threshold obtained from the dissimilarity distribution *ds*,

$$d_{TH} \rightarrow \arg\min_{d_{TH}}\left(Q\left(ds, d_{TH}\right)\right) \quad (11)$$

where Q is the target function to estimate the best threshold values. A frame with its neighbourhood dissimilarity *ds* greater than d_{TH} can be defined as a segment boundary, and the middle frame between two shot boundary frames is defined as the representative frame of this segment.

2) Scene-level Summarization

After shot-level key frames are extracted, these key frames may still share common scenes. To eliminate this redundancy, further clustering and selection is required.

Among various clustering algorithms, k-means clustering (KMC) is a practical and easy method for this kind of problem. KMC 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 centres of clusters in the data.

To attain the target of clustering, KMC algorithms are based on minimization of the following objective function:

$$Z = \sum_{j=1}^{K}\sum_{x_i \in S_j}^{N}\left\|x_i - c_j\right\|^2 \quad (12)$$

where there are K clusters S_j, $i = 1, 2, ..., K$, and c_j is the centroid or mean point of all the points x_i. ||*|| is any norm denoting the distance between any data item and the cluster centre.

In order to find an appropriate solution to the above equation, a cluster number must be provided before the KMC iteration starts. Assuming $\{c_i\}^{(k)}$ is the set of $K^{(k)}$ initial cluster centres for KCM clustering, to obtain an optimal result of cluster centres $\{c_i\}^{(k+1)}$ that matches the optimization target in the clustering model Ω, maximum likelihood estimation can be obtained through an Expectation-Maximization iteration. The iterative procedure can be summarized as follows:

- Assume there are initially k=2 clusters in the data set. The KMC algorithm is applied to $\{x_i\}$, resulting in two cluster centres with corresponding coordinates.
- Euclidean distances between the cluster centres $\{c_i\}^{(m)}$ are computed. If their mean distances are greater than the predefined thresholds $\{T^{(k)}\}$, the optimal cluster number K is increased to K+1. Otherwise, the process is terminated, and we take K=m as our final result.
- Using similar iterations to this, one is able to determine a cluster number that brings us a Euclidean distance less than the threshold. This cluster number is what we are looking for.

At the end, the key frame at the cluster centre is considered as the most representative scene frame of this cluster.

3) Sub-shot Level Representative Frame Selection

After the shot-level video segment structure is determined, its hierarchical sub-shot structure can be extracted subsequently. The basic scheme is similar to the above KM procedure for scene summarization. The difference is that we use all intra-shot frames in one shot to generate the dissimilarity matrix for this shot. After we obtain the intra-shot dissimilarity matrix, the KM approach is applied to this matrix to find out most representative sub-shot frames, which cannot be represented well by shot-level key frames.

With the above three-level summarization scheme, we can have sub-shot level, shot level, and scene level results for browsing and skimming. However, mostly we may use two of these three levels. Some videos may need sub-shot summarization, but no common scenes between shots. Others may have frequent scene changes with no apparent sub-shot structure, while these shots may share the same scenes. In the following experiment, the experiment is conducted with both cases.

III. Experimental Results & Discussion

In an experiment, the proposed scheme and the state-of-the-art LSA-based scheme are coded using MATLAB, and tested with the same set of test videos for comparison. The test videos are encoded in MPEG-4 format, which can be read through MATLAB multimedia interface. The experiment is carried out on a 2GH AMD 64 Turion PC with 512MB RAM.

In all experiments, the video frames are resized to 50% to save computation time. In the test, the video stream is first input through the MATLAB multimedia codec interface, and reference frames are selected from the sequence at a regular interval, e.g. one reference frame per 200 frames. For a 30-minute video sequence, there are about 250 reference frames. Then the Laplacian and LSA approaches are applied to these reference frames, respectively, to obtain the projection matrix W_{Lap} and W_{LSA}. With the projection matrices, all frames can be projected into the Laplacian and LSA subspaces, and the dissimilarity between them can be defined as their Euclidean distance in the projection subspace. With the dissimilarity measure from LSA-based or Laplacian-based subspace, shot-level key frames, sub-shot key frames and above-shot scene frames are then extracted successively.

In the following sections, we report two kinds of test. One uses a standard test video database [30], where every video may have long shots that require sub-shot summarization as well. Another test uses movies that have short shots with common scenes that need further scene-level summarization.

A. Experiment with Test Videos

In this test, five videos from RUSHES video database [30] are used, as shown in Fig.2. These videos have long shots that may need further sub-shot level temporal segmentation, which means the video summarization at both shot and sub-shot level. Here, the reference frames are selected, one per 200 frames. With this subset of reference frames, the matrix size for SVD computation is reduced dramatically from $N \times N$ to $N/200 \times N/200$, where N is the number of total frames a video has.

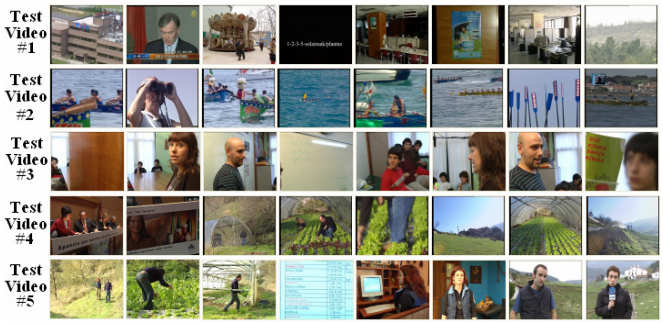
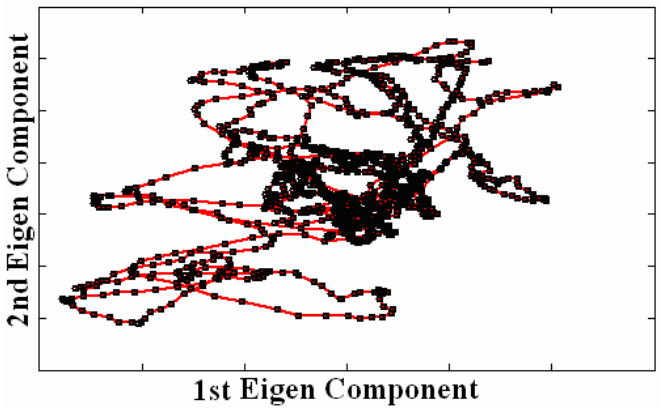
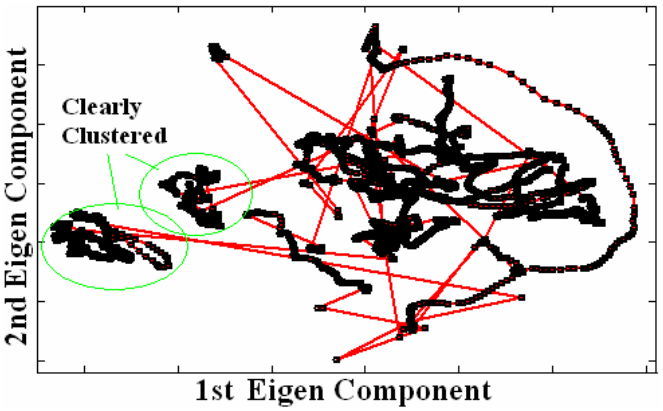


Fig. 2. Frames in five test videos



a) Projection of the 1st 1000 Frames in LSA space



b) Projection of the 1st 1000 Frames in Laplacian space

Fig. 3. Projections in LSA and Laplacian Eigenmap subspaces.

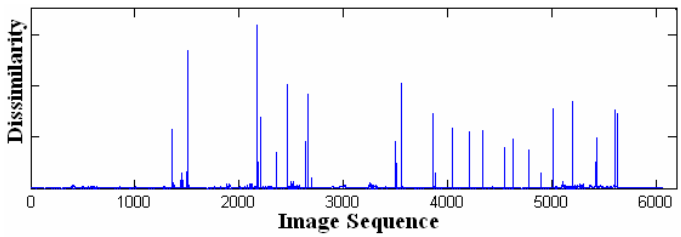


Fig. 4. Dissimilarity distribution *ds* between neighbouring frames in reference-based Laplacian subspace.

a) Shot-level keyframes detected by Laplacian Eigenmap

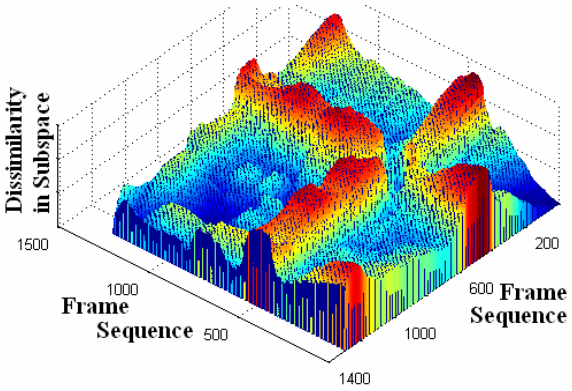
b) Shot-level keyframes detected by LSA-based approach

c) Lost keyframes in LSA-based summarization

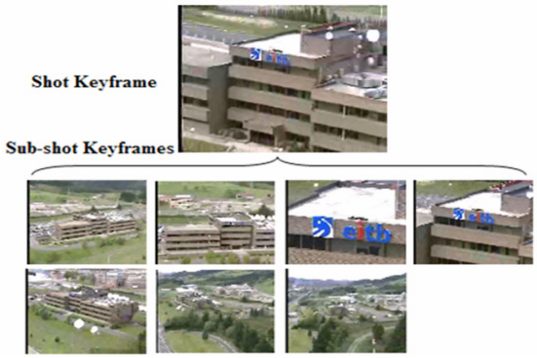
Fig. 5. Detected shot-level key frames in test video #1

TABLE I
COMPARISON OF RECALL AND PRECISION IN SHOT-LEVEL KEYFRAME DETECTION.

Test Videos	LSA		Laplacian Eigenmap	
	Recall	Precision	Recall	Precision
#1	0.8	0.91	1	0.89
#2	0.76	0.95	0.87	0.92
#3	0.71	0.76	0.8	0.85
#4	0.68	0.82	0.75	0.81
#5	0.75	0.87	0.81	0.88
Total	0.74	0.86	0.85	0.87



a) Dissimilarity matrix of frames in the same shot



b) Sub-shot representative frames

Fig. 6. Sub-shot level key frame detection

Table II
AVERAGE COMPUTE TIME FOR LSA AND LAPLACIAN EIGENMAP (UNIT: SECONDS, 1000 FRAMES)

Method	Reference Eigenspace	Frame Projection	Total Time
LSA	1.2	0.325/frame	326.5
Laplacian	1.1	0.256/frame	257.3

Fig.3 is the projected result of 1000 frames of the test video #1 into the Laplacian and the LSA subspaces, respectively. As shown in Fig.3-b, we can clearly see frames are clustered together in the Laplacian subspace, which implies a better temporal segmentation than LSA. This gives a very comprehensible explanation as to why Laplacian Eigenmap can give better results in the dissimilarity measurement for video summarization.

Fig.4 shows the corresponding dissimilarity between neighbouring frames measured in Laplacian Eigenmap subspace, which is used to extract shot-level keyframes. With the computed dissimilarity, the temporal video structure can be easily obtained. Fig.5 gives the detected representative frames of one test video. Fig.5-a shows the results of the proposed Laplacian approach, and Fig.5-b shows the results of the conventional LSA approach. In comparison, the LSA-based approach failed in the detection of several shot-level key frames, as shown in Fig.5-c.

The performance of both approaches can usually be evaluated in terms of the recall rate and precision, which are defined as:

$$\text{Recall Rate} = \frac{K_P}{K_A}, \quad \text{Precision} = \frac{K_P}{K_P + K_N}$$

where K_P is the number of correctly detected key frames, K_A is the number of all real key frames, and K_N is the number of wrongly detected frames. Table I gives the overall test results of both approaches over all five videos in RUSHES database [31~32]. From the benchmark results, we can see that although both approaches have similar precision, the proposed Laplacian approach can have much better recall rate in keyframe detection.

Because these test videos contain long shots, it is also easy to find out the hierarchical sub-shot video structure after the shot-level structure is detected. As introduced in section II-c, the first step is to establish the dissimilarity matrix of all intra-shot frames. Fig.6-a shows this intra-shot dissimilarity matrix of the first shot structure in the test video #1 measured in the Laplacian Eigenmap subspace. This shot has 1356 frames and seven sub-shots are detected, as shown in Fig.6-b.

Table II also gives the computing time of both approaches. We can see that Laplacian-based approach is slightly better in saving time. This is due to the dramatic dimensional reduction in Laplacian Eigenmap. While LSA subspace may have several hundred dimensions, Laplacian subspace can only have several ten.

B. Summarizing Real Movies

Since video abstarction appears more natural and attractive to viewers, most recent work on movie abstraction focuses on the generation of a short synopsis of a long feature film. In this section, the proposed Laplacian-based approach is applied to two well-known test movies, *Friends* and *Titanic*.

In the experiment, we first select reference frames by one per 200 frames. For a 30 minute story, this gives about 225 reference frames. As described above, the Laplacian Eigenmap approach is applied to extract the projection matrix W_{Lap}. After W_{Lap} is obtained, all frames can be projected into its Eigen space.

Fig.7 shows the frames in the movie *Friends*. We can see there are frequent scene transitions that may make up the video story. Fig.8 shows the projection result of about 8000 frames. With this projection result, the distance between any two frames can be easily measured. A useful dissimilarity measure is the dissimilarity between two neighbouring frames, which may represent the scene transition process. Fig.9 gives the measured neighbourhood dissimilarity distribution along the frame sequence. With this dissimilarity measure, shot transitions and key frames can be detected, as shown in Fig.10. In total, about 80 keyframes are found from the first 5 minutes of the movie.

Fig. 7. Frames in Movie --- *Friends* Episode

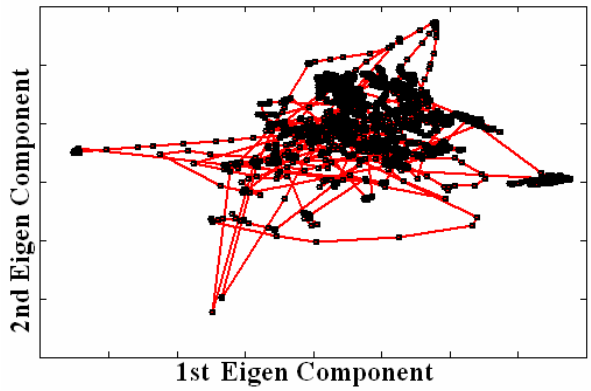


Fig. 8. Projection of frames in Laplacian Eigen Space (8000 frames shown in total).

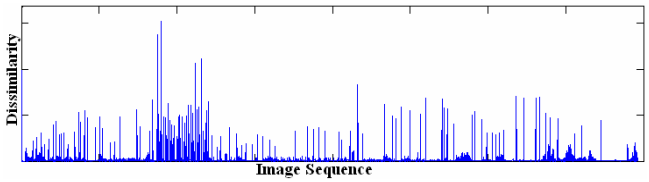


Fig. 9. Dissimilarity between the neighboured frames in Eigen space

Fig. 10. Extracted shot-level representative frames in *Friends*

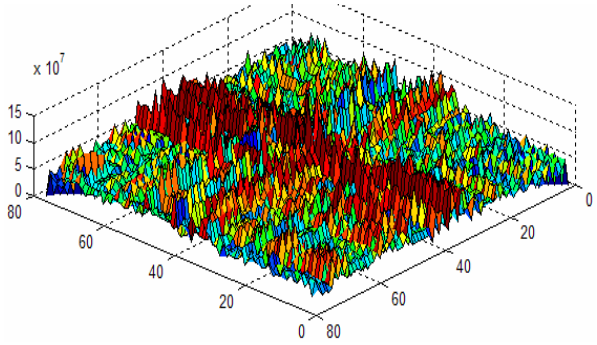


Fig. 11. Dissimilarity matrix between keyframes

Fig. 12. Summarized representative scenes in *Friends*

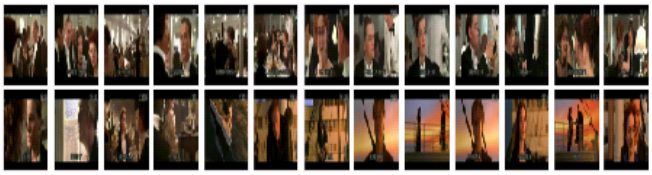


Fig. 13. Frames in Movie --- *Titanic*

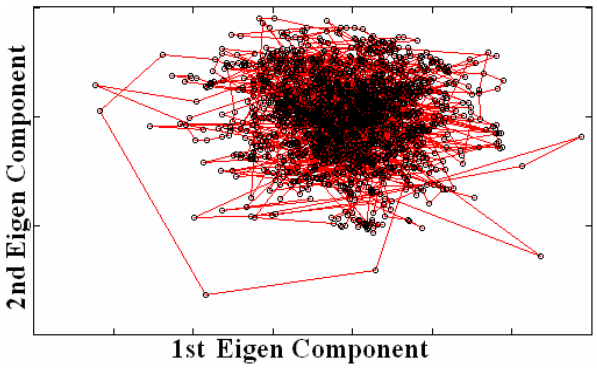


Fig. 14. Projection of frames in Laplacian Eigen Space (about 27000 frames for the 30 minutes video).

Fig. 15. Extracted shot-level representative frames in *Titanic*

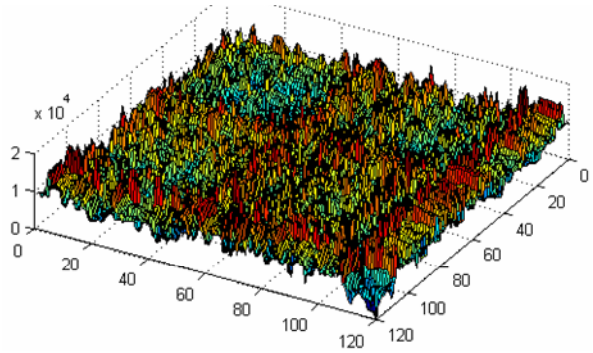


Fig. 16. Dissimilarity matrix between keyframes

Fig. 17. Summarized representative scenes in *Titanic*

As it can be seen from the detected key frames in Fig.10, shot transitions happen frequently in such movies, while many shots are taken from the same scene. Thus, scene-level movie summarization becomes necessary to find the most representative scene frames.

In order to achieve this purpose, the dissimilarity of all detected keyframes can be computed by measuring the distance in their Eigen space projection. Fig.11 gives the calculated dissimilarity matrix of all shot-level key frames. Similarly, as described in section II-C, K-means clustering is applied to the dissimilarity matrix to find out which are the most representative scene frames. Fig.12 is the result from the EM procedure, where about 20 scene-level representative frames are selected from about 80 key frames. It can be seen that these scene frames have obviously less redundancy or scene similarity than shot-level keyframes.

Fig.13 shows the frames in another classical movie, *Titanic*. Fig.14 shows the projection result of about 27000 frames of the 30 minutes video. After the frames are projected into the subspace, the distance between any two frames can be measured as their dissimilarity. With the disismilarity of neighbouring frames, the shot transitions are detected and their corresponding key frames are listed, as shown in Fig.15, where we can see that many similar keyframes are actually from the same scene.

In order to provide further scene level suimmarization, the dissimilarity matrix between key frames is computed, as shown in Fig.16. KM clustering is then applied to the dissimilarity matrix, and scene-level representative frames are obtained, as shown in Fig.17.

From the experiment on two typical movies, it is shown that the proposed approach can work well for movie summarization. With the proliferation of digital videos, as a useful video abstraction tool for fast content browsing, skimming, transmission, and retrieval of massive video

databases, video summarization and skimming has become an indispensable tool of any practical video content management system, which illustrates the potential of the proposed approach for commercial applications in web multimedia, mobile multimedia, interactive TV, and emerging 3D TV.

IV. CONCLUSIONS

In conclusion, a novel video summarization approach using Laplacian Eigenmap for hierarchical video summarization is presented, and the experiments on test videos show that the proposed approach can adaptively give the hierarchical video summary with higher recall accuracy for finding representative content in comparison with the similar LSA-based approach. A further examination was performed on movie summarization, which shows the proposed hierarchical scheme can also effectively eliminate the redundancy in key frames and find the most representative scene frames from shot-level key frames.

REFERENCES

[1] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV", IEEE Signal Proc. Mag., Vol.23, No.2, 2006, pp.90.

[2] W. Zhang, Q. Ye, L. Xing, Q. Huang, W., Gao, "Usupervised sports video scene clustering and its application to story units detection", Proc. SPIE – VCIP, 2005.

[3] J. Nam, A. Tewfik, "Dynamic Video Summarization and Visualization", ACM Multimedia, 1999.

[4] M. Jiang, A. Sadka, D. Crookes, "Advances in Video Summarization and Skimming", in "Recent Advances in Multimedia Signal Processing and Communications", Springer-Verlag, 2009.

[5] Y. Li, S. Lee, C. Yeh, C. Kuo, "Techniques for Movie Content Analysis and Skimming", IEEE Signal Processing Magazine, Mar. 2006, pp.76.

[6] Y. Taniguchi, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing", Proc. ACM Multimedia'95, Nov. 1995, pp. 25.

[7] M. Mills, "A magnifier tool for video data", Proc. ACM Human Computer Interface, May 1992, pp. 93.

[8] H. J. Zhang, J. Wu, D. Zhong, S. Smoliar, "An integrated system for content-based video retrieval and browsing", Pattern Recognition, Vol.30, No.4, Apr. 1997, pp.643.

[9] Y. Zhuang, Y. Rui, T. Huang, S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering", Proc. ICIP'98, Oct. 1998, pp.866.

[10] S. Ju, M. J. Black, S. Minneman, D. Kimber, "Summarization of video-taped presentations: Automatic analysis of motion and gestures", IEEE Trans. Circuits Syst. Video Technol., Vol. 8, No.5, Sept. 1998, pp. 686.

[11] C. Toklu, S. P. Liou, "Automatic keyframe selection for content-based video indexing and access", Proc. SPIE, Vol.3972, Jan. 2000, pp.554.

[12] S. Uchihashi, J. Foote, A. Girgensohn, J. Boreczky, "Video manga: Generating semantically meaningful video summaries", Proc. ACM Multimedia'99, Oct. 1999, pp.383.

[13] A. Girgensohn, J. Boreczky, "Time-constrained keyframe selection technique", Proc. ICMCS'99, June 1999, pp.756.

[14] M. M. Yeung, B. L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content", IEEE Trans. Circuits Syst. Video Technol., Vol.7, No.5, Oct. 1997, pp.771.

[15] D. Lelescu, "Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream", IEEE Trans. Multimedia, Vol. 5, No.1, Mar. 2003, pp.106.

[16] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?", IEEE Trans. Circuits Syst. Video Technol., Vol.12, No.2, pp. 90–105, Feb. 2002.

[17] H. Zhang, S. Kankanhalli, "Automatic partitioning of full-motion video", ACM Multimedia Syst., Vol.1, No.1, Jan. 1993, pp.10.

[18] Zabih, R., Miller, J., Mai, K.: "A feature-based algorithm for detecting and classification production effects", ACM Multimedia Syst., Vol.7, No.1, Jan. 1999, pp.119.

[19] M. Jiang, D. Crookes, "Approach to Automatic Video Object Motion Segmentation", Electronics Letters, 2007, Vol.43, No.18, pp.968.

[20] M. Jiang, D. Crookes, S. Davidson, R. Turner, "A Low-Power Systolic Array Processor Architecture for FSBM Video Motion Estimation", Electronics Letters, 2006, Vol.42, No.20, pp.1146.

[21] M. Jiang, D. Crookes, "FPGA-based minutia matching for biometric fingerprint image database retrieval", Journal of Real-Time Image Processing, Springer, Vol.3, No.2, September 2008, pp.177.

[22] M. Jiang, D. Crookes, "A High-Speed Area-Efficient 3D Discrete Wavelet Transform Architecture", Electronics Letters, 2007, Vol.43, No.9, pp.502.

[23] Y. H. Gong, X. Liu, "Video Summarization Using Singular Value Decomposition", Int. Conf. Computer Vision & Pattrn Recognition, 2000.

[24] M. Slaney, D. Ponceleon, "Hierarchical segmentation using latent semantic indexing in scale space", 2001 IEEE International Conf. Acoustics, Speech & Signal Processing, Vol.3, May 2001, pp.1437.

[25] F. Souvannavong, B. Merialdo, B. Huet, "Latent semantic indexing for semantic content detection of video shots", 2004 IEEE International Conf. on Multimedia & Expo, Vol.3, June 2004, pp.1783.

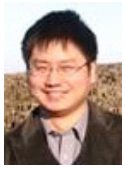
[26] K. Kim, S. Park, H. Kim, "Kernel principal component analysis for texture classification", IEEE Signal Processing Letters, Vol.8, Issue 2, Feb. 2001, pp.39.

[27] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science, 290 (5500), 2000, pp.2323.

[28] Mikhail Belkin, Partha Niyogi,"Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering", Advances in Neural Information Processing Systems 2001, pp.14.

[29] He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: "Face Recognition Using Laplacianfaces", IEEE Trans. Patten Analysis and Machine Intelligence, Vol.27, No.3, Mar. 2005.

[30] Schreer, O., Ardeo, L., Sotiriou, D., Sadka, A., Izquierdo, E.: "User Requirements for Multimedia Indexing and Retrieval of Unedited Audio-Visual Footage – RUSHES", Ninth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '08), 7-9 May 2008, pp.76.



Richard M. Jiang obtained his PhD in computer science from Queen's University Belfast. He is currently with Computer Science, Loughborough University. His research interest includes content-based video retrieval, 3D computer vision, and VLSI vision processing.

Abdul H. Sadka obtained his PhD in Electrical and Electronic Engineering from the University of Surrey in 1997. He was appointed a Professor and the head of Electronic and Computer Engineering at Brunel University in 2004. His research interests include video coding and transcoding, video transmissions over networks, computer vision and content-based multimedia retrieval.

Danny Crookes graduated with BSc in Mathematics and Computer Science in1977, and PhD in Computer Science in 1980, both from Queen's University Belfast. He was appointed Professor of Computer Engineering at Queen's University Belfast in 1993. His research interests include software tools for high performance computing, real-time image and vision processing, and medical image analysis.