VMAP: VIDEO VISUALIZATION AND SUMMARIZATION ON EMBEDDED MANIFOLD ARTICULATION PRIMITIVES

Yingge Wang, Qiang Cheng, Jie Cheng, and Thomas S. Huang

ABSTRACT

This paper considers visualizing and summarizing an image sequence using an unsupervised manifold learning method. The image sequence is articulated on a parameterized low-dimensional manifold and embedded in the high-dimensional image space. The manifold is discovered using a nonlinear subspace method preserving the underlying geometry and local neighborhoods. The visualization and summarization of video contents are performed on the manifold articulation primitives. We construct two modes of roadmaps serving like tourist maps guiding the sequence traversing. The first mode discovers and exhibits landmark points signaling the dramatic changes in video content in the temporal order. The second one displays the global content coherence by preserving the geodesic distances on the nonlinear manifold, while relaxing the temporal constraint. Experimental results demonstrate the effectiveness of the algorithm.

1. INTRODUCTION

Recent years have seen increasing interest in visualization and summarization of image sequences, including video, with applications to image and video indexing, retrieval, content adaptation, interactive service, and so on. Due to the ubiquitous use of imaging sensors, such as security surveillance cameras, satellite multispectral sensors, and computerized tomography (CT), image sequences are produced in huge quantity and in an anywhere, anytime, and anyhow manner. These sequences like surveillance videos are often highly redundant, and a substantial amount of time is spent in sifting important contents. To better utilize high-dimensional, large volumes of data, there is a strong need for efficient presentation of images preserving their essential information, and for automatic summarization facilitating the understanding, analysis, adaptation, and exploration of visual contents.

In the literature, video browsing and summarization usually detect shot transitions, and select representative frames from each shot [1] - [4]. Some techniques clustered the key frames to provide a hierarchical representation of video segments [5] - [7]. By making use of principal component analysis (PCA), motion and color features are condensed before a supervised classification method, hidden Markov model (HMM), is applied to the principal components [8]. Several works also use probabilistic unsupervised learning models for video, such as the Gaussian mixture model (GMM) [9] and nonparametric statistical models [10]. They mainly focus on the statistical properties of temporal-spatial data, not the geometrical ones. These methods provide meaningful and useful browsing, classification, or summarization of video data. The images usually lie on a significantly low-dimensional manifold (in the high-dimensional image space), which provides intrinsic information on the video content and formation. Camera panning, zooming, tilting, and the motion or evolution of objects can naturally create these manifolds, highly nonlinear globally. These inherent nonlinear structures, however, have not been fully exploited in the literature of video browsing and summarization. In this paper, we construct two modes of video roadmaps on top of the manifolds embedded with a subspace learning method ISOMAP for efficient visualization and summarization. Our algorithm can clearly show the landmark points of video trails and the compact groups of images, associated well with the contents semantically.

The rest of the paper is organized as follows. Section 2 introduces ISOMAP for subspace learning and manifold embedding. Section 3 constructs roadmaps of video on top of the embedded manifolds. Section 4 concludes the paper.

2. LEARNING INHERENT MANIFOLD OF VIDEO

Typical image sequences in a video consist of shots, that is, contiguous frames between fades, wipes, cuts, or large camera motions. Groups of shots may further constitute scenes that exhibit some consistency in the semantic context of the video. These scenes may have different shots or alternate between a few shots. In the formation of scenes, the object often is recorded by cameras with different motions such as panning, zooming, and tilting. The object can also change appearance in pose and articulation. The variation of these parameters gives rise to articulation manifolds in the high-dimensional image space [11] - [13]. The manifolds provide the underlying parametrization of video content and an arguably optimal representation of feature vectors from video frames. The importance of such a description is that a long segment of video can be treated as a whole, allowing one to build efficient and meaningful analysis, indexing, and classification applications. In general, the movement of objects is highly nonlinear rather than linear, and the video data are more likely to lie on a nonlinear manifold instead of a hyperplane. The PCA, or equivalently, singular-value decomposition (SVD), has long been adopted for dimensionality reduction and content description in pattern recognition, or shot change detection in image sequence analysis [2] [4] [8]. It can, however, provide only a linear subspace, a hyperplane, to approximate the underlying structure of the images. In general, the movement of objects is highly nonlinear rather than linear, and the video data are more likely to lie on a nonlinear manifold instead of a hyperplane. The PCA also provides a mapping by concentrating the energy in the eigenspace spanning all data, without preserving the local topology. Thus, some of the intrinsic structure may be revealed by the PCA projection; however, much relatively smooth trajectory will still be disguised. To discover the inherent mani-

The second author is with ECE Dept., Wayne State University, Detroit, MI 48202. The fourth author is with ECE Dept. and Coordinated Science Laboratory, University of Illinois at Urbana - Champaign. E-mails: yinggewang@hotmail.com, qcheng@ece.eng.wayne.edu, chengjie_2000_2000@yahoo.com, huang@ifp.uiuc.edu.

folds on a significantly low-dimensional space, a manifold learning technique ISOMSP is exploited in this paper.

The ISOMAP algorithm [12] used in this paper performs the following nonlinear manifold learning. Assume there are N points $\{x_i\}_{i=1}^N$ in the image feature space X. First, we employ a distance function to quantify perceptual similarity of images. As fully understanding of human perception is still immature, in this paper, we use a weighted L_1 and L_2 distance, $D(x_i,x_j) := 0.4L_1(x_i,x_j) + 0.6L_2(x_i,x_j)$. This measure works well in our experiments, and better perceptual distance may lead to improved performance. Secondly, a local neighborhood of M nearest neighbors is found for each point, and a graph G is defined by connecting each point with its neighbors. We have chosen $M = 8$ experimentally. Thirdly, all pair of distances $D_G = \{d(i,j)\}$ containing the shortest paths are built. This is done by initializing $d(i,j) = D(x_i,x_j)$ if i,j are neighbors or ∞ otherwise, and by replacing $d(i,j)$ by $\min\{d(i,j), d(i,k) + d(k,j)\}$ for each value of $k = 1,\cdots,N$. Lastly, define matrices S and H such that $S_{ij} = D_G^2(i,j)$, and $H_{ij} = \delta_{ij} - \frac{1}{N}$, then find the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N$ of $-\frac{1}{2}HSH$. The ith component of the corresponding eigenvectors v_t for λ_t is denoted as v_t^i. The ith component of embedded d-dimensional vector y_i is $\sqrt{\lambda_l}v_t^i$.

The geodesic distance between input data is used to capture information on the nonlinear manifold structure. Compared to PCA, ISOMAP preserves the local topology of input data, that is, the nearest neighbors of the input data in the original feature space are mapped to the nearest neighbors in the lower-dimensional space. The local topology is more intrinsic and reliable than the global one, particularly for the purposes of video visualization, because we look for compact groups preserving inherent topological structures in the original image space. Locally linear embedding (LLE) and Hessian Eigenmap [14] methods have also been proposed for subspace learning. We adopt ISOMAP for its relative robustness to noise and outliers. To demonstrate this, we add Gaussian noise to Swiss roll data [14]. Then, ISOMAP and several other methods are used to embed manifolds in the noisy data. It is found that ISOMAP has best preserved the local neighborhoods and the intrinsic geometry. Though ISOMAP only partially unfolds the manifold, the others have mixed up the local neighborhoods. These properties are important in considering an entire sequence that often reveals both long range and subtle short-time relationships.

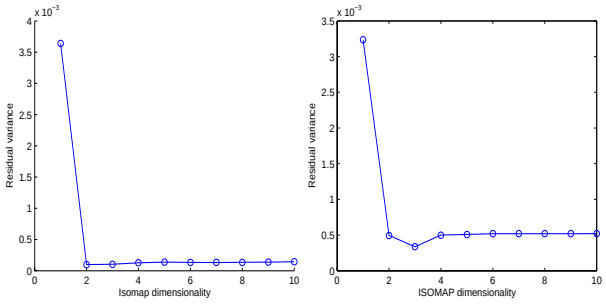
3. VIDEO ROADMAPS ON TOP OF EMBEDDED MANIFOLDS

Often times, there is an hour or more material in a long image sequence, and there is no roadmap to help viewers to find their way through the sequence. It would be much helpful if major contents or landmark frames can be filtered computationally to guide the viewers accessing and navigating the sequence.

We construct video roadmaps on top of the embedded manifolds. Two modes of roadmap are built: One is to follow the temporal trail of video for important landmark points; and the other is to display the coherent group on a two-dimensional plane, retaining the intrinsic similarity of images without the temporal order. By doing so, the "*sequential processing*" of images in a natural time series is transformed into the "*batch processing*" grouped by content similarities.

We take each single image or single frame in a video as our data unit. Since images have very high dimensions, we extract image features for efficient representations. The features need have high correlations if there is a perceived similarity. The pixels are not good features, because small camera motions usually lead to decorrelations of pixels in the same region for different frames, even though there are the same visual contents. Color provides robust information about the contents [15] and serve as our main features. For image sequence without audio component, the visual feature set may include color histogram, the first two moments in RGB space, and Tamura's features: coarseness, contrast, directionality, linelikeness, regularity, or roughness [16]. More visual features may be used to better capture the visual information, or, a subset can be utilized to have on-the-fly computation, e.g., in real-time interactive services. For video with audio component, the mel-frequency cepstral coefficients (MFCC) can be included, which were used in combined audio and video watermarking [17]. Those features usually are as many as several hundreds or even thousands. Each image is represented as a data point in the feature space, whose dimensionality is much lower than the original image space. However, for visualization purposes, the dimensionality of the feature space is still prohibitively high. Parameterized by camera or object movement, the underlying manifolds usually have significantly lower dimensionality, and convey key information on the articulation of video structures. We extract the intrinsic manifolds, and construct video roadmaps based on the manifolds.

It is observed in our experiments that, often times, the manifold can be well embedded into one- or two-dimensional space for video clips. For example, we learn manifolds for video clips Sports Car and Planet Formation, of 180 and 926 frames, respectively. These video clips include one scene change (e.g., Sports Car) and multiple dramatic changes in an evolutionary process (e.g., Planet Formation). The residual variances after the embedding with different dimensionalities are shown in Fig. 1. They are already very small with one-dimensional embedding, and close to zero with two-dimensional. This observation motivates us to construct two types of video roadmaps based on one- and two-dimensional manifolds, respectively.



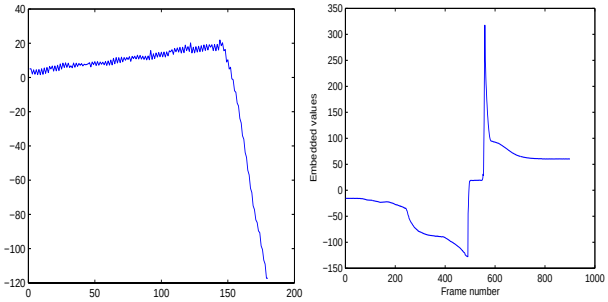
(a) (b)

Fig. 1. Residual variances after embedding into low-dimensional space with different dimensionalities for video clips (a) Sports Car, and (b) Planet Formation.

3.1. Roadmap along Temporal Sequence

Our first video roadmap is based on the embedded one-dimensional manifold as a temporal process. The resultant trail shows the underlying articulation of different images. Fig. 2 shows the primitive roadmap for Sports Car and Planet Formation. It is observed

that shot or scene changes correspond to landmark points of the trails. Usually, the manifolds consist of line segments or knobs. The landmark points are around the sharp turns of the manifolds. For the Sports Car clip, we demonstrate the correspondence in Fig. 3. For the Planet Formation, there are slow evolutionary as well as explosive changes, with the latter captured by the spiky landmark points.

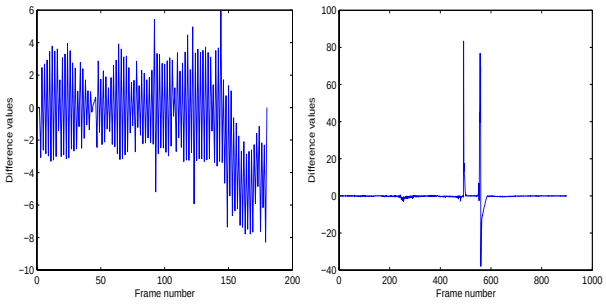


(a) (b)

Fig. 2. Trails for video clips (a) Sports Car and (b) Planet Formation.

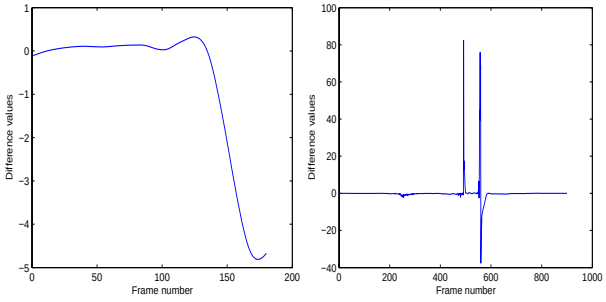
Fig. 3. The original images from a sports video clip. From the first to the 143th, a playing statue is displayed; and from the 143th to 180th frames, the scenes changes to car racing. The change of scene corresponds to the landmark point of the video trail.

Using ISOMAP, the intrinsic geometry and local neighborhoods of the data are preserved in the low-dimensional space. On the embedded manifold, a cluster or segment of data usually have small geodesic distances thus similar features; in this sense, each cluster corresponds roughly to a "story unit." Sharp changes of a manifold may reflect dramatic changes of the story or semantic contents. To quantify the rate of changes, we make use of the difference of adjacent trail points. For example, the difference signals for the embedded one-dim manifolds for Sports Car and Planet Formation are plotted in Fig. 4. The difference signals have many small oscillations, which may correspond to small changes in consecutive image, or noise introduced in image acquisition, or possibly by video coding errors. For a long image sequence, these small changes are regarded insignificant. We desire to retain significant changing points and obliterate the noisy or insignificant ones. Wavelet denoising algorithms [18] are employed for this purpose. Fox example, the denoised difference signals for Sports Car and Planet Formation using symlet are shown in Fig. 5. The landmark points can be easily captured for their association to the dramatic changes of the denoised difference signals. For visualization of image sequences, the denoised difference signals serve as our roadmap. The neighborhoods of landmark points of the roadmap are representatives of the video and need be extracted for summarization.



(a) (b)

Fig. 4. Difference signal for one-dim manifold of video clip (a) Sports Car and (b) Planet Formation.



(a) (b)

Fig. 5. Denoised difference signal for one-dim manifold of video clip (a) Sports Car and (b) Planet Formation.

3.2. Roadmap along the Video Trail

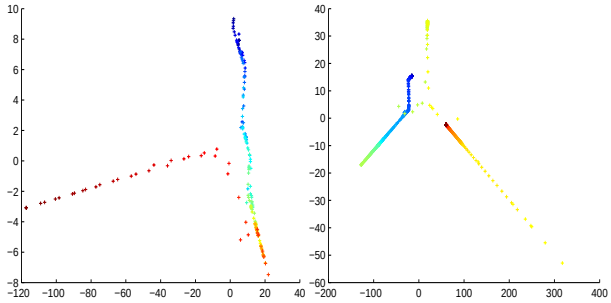
In the above we construct a mode of roadmap along the temporal sequence on top of embedded manifolds. Now we consider the second mode. In some videos, some scenes may have many recurrences. For example, in a sports video, a commercial advertisement may be inserted several times at different locations. Or, in surveillance video, there are many repeated scenes. In this case, viewers may have interest in browsing through *different* scenes. The first mode of roadmap is not proper for this purpose, since it gives landmark points along the temporal track, regardless of how many times a scene has occurred. We trade time for space so that viewers can have a global picture of the coherence of all frames simultaneously, instead of following their temporal ordering.

Our method is to reorganize those images by their content similarities on the articulation manifold. It plays a role similar to a tourist map, guiding a viewer to browse and navigate different images, indicating the direction in which the image sequence goes. Our algorithm can also be applied to the visualization and summarization of images in an image database, by discovering their coherent connections in terms of content similarities.

Based on the embedded manifold, we construct the second mode of roadmap with each image represented by a two-dim vector and displayed on the screen. For example, two-dim manifolds of Sports Car and Planet Formation are shown in Fig. 6. It can be seen compact groups are clearly formed and that each scene is located on a knob. The turning points of the knobs are images-

of-interest and can be taken as landmark points for this roadmap. For example, in the roadmap for Sport Car, there are mainly two knobs; and for Planet Formation, there are mainly three knobs. The *x*-axis and *y*-axis represent the first and the second dimensions of the embedded manifolds, respectively. They also provide directional information of the movements or poses of the objects. Visualization and summarization are performed on these knobs by choosing representative frames, or landmark points.

Fig. 6. Roadmap on top of two-dim manifold of video clip (a) Sports Car and (b) Planet Formation.



(a) (b)

4. CONCLUSION

In this paper, we construct roadmaps guiding the visualization and summarization of image sequences. The intrinsic parametric structure underlying the video is considered by embedding the articulation manifolds. Two modes of roadmaps for video trails are then built. The first mode discovers the landmark points signaling the dramatic changes in video content in the temporal order. The second one displays the global content coherence by preserving the geodesic distances on the nonlinear manifold. Multiresolution VMAP is being developed for scalable browsing and summarization. The method is being incorporated into video monitoring and surveillance systems for interactive explorations of video content. Integrating extracted video, audio, and textual information using graphical models for intelligent multimedia services on the constructed roadmaps, especially when high-level domain knowledge were available, is our future research topic.

5. REFERENCES

[1] G. Iyengar and A. Lippman, "Videobook: An experiment in characterization of video," *Proc. ICIP*, vol. 3, pp. 855-858, 1996.

[2] A. Tewfik and K. Han, "Eigen-image based video segmentation and indexing," *Proc. ICIP*, vol. 2, pp. 538-541, 1997.

[3] S. Srinivasan, D. Ponceleon, A. Amir, and D. Petkovic, "'What is in that video anyway?' In search of better browsing." *Proc. IEEE Int. Conf. on Multimedia Computing and Systems*, pp. 388-393, Florence, Italy, June, 1999.

[4] V. Kobla, D. Doermann, and C. Faloutsos, "Developing high-level representations of video clips using videotrails," *Proc. SPIE*, vol. 3312, pp. 81-92, 1998.

[5] M.M. Yeung and B.-L. Yeo, "Time-constrained clustering for segmentation of video into story units," *Proc. ICPR*, pp. 375-380, 1996.

[6] A. Hanjalic and H.J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1280-1289, 1999.

[7] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang, "On clustering and retrieval of video shots," *Proc. ACM Multimedia*, pp. 51-60, Ottawa, Canada, Sept. 2001.

[8] E. Sahouria and A. Zakhor, "Content analysis of video using principal components," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 8, 1290-1298, Dec. 1999.

[9] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, March 2004.

[10] D. DeMenthon, "Spatial-temporal segmentation of video by hierarchical mean shift analysis," *Proc. Stat. Methods in Video Processing Workshop*, June 2002.

[11] J.J. Koenderink. *Solid Shape*. MIT Press, 1990.

[12] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.

[13] S.T. Roweis and K. Lawrence, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000.

[14] D.L. Donoho and C. Grimes, "Hessian eigenmaps: new embedding techniques for high-dimensional data," *Proc. National Academy of Sciences*, 03-1596, 2003.

[15] Y. Rui, T.S. Huang, S. Mehrotra, and M. Ortega, "A relevance feedback architecture for content-based multimedia information retrieval systems," *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries*," pp. 82-89, 1997.

[16] H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-8, no. 6, pp. 460-473. 1978.

[17] Q. Cheng and T.S. Huang, "Combined audio and video watermarking using Mel-frequency cepstra," *Proc. Int. Conf. on Multimedia and Expo*, Tokyo, Japan, Aug., 2001.

[18] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptotia," *Jour. Roy. Stat. Soc.*, series B, vol. 57, no. 2, pp. 301-369, 1995.