MV-MAP: Multiresolution Video Visualization and Summarization on MAPs

Yingge Wang * Qiang Cheng Jie Cheng Thomas S. Huang

Abstract

This paper considers visualizing and summarizing image sequences using manifold learning and multiresolution techniques. The images in a video are found usually lying on a significantly low-dimensional manifold, which provides intrinsic information on the video content and formation. The parametrization of the manifold is discovered using a nonlinear subspace method preserving underlying geometry, especially local topology, in the original space. Two modes of video roadmaps have been constructed using VMAPS [1]. The first discovers the landmark points signaling dramatic changes in video content in the temporal order. The second reveals the global content coherence, without the temporal ordering. To facilitate the browsing of long sequences with complicated contents and structures, we build multiresolution visualization and summarization tools on VMAPS. Experimental results validate the proposed method. It may find applications to video monitoring and surveillance for interactive exploitation of video contents, intrusion detection, etc.

1. Introduction and VMAP

Recent years have seen increasing research and business interest in visualization and summarization of image sequences, including video, with applications to image and video indexing, retrieval, content adaptation, and so on. In the literature, video browsing and summarization usually detect shot transitions, and select representative frames from each shot [2] - [5]. Some techniques cluster the key frames to provide a hierarchical representation of video segments [6] - [8]. By making use of the principal component analysis (PCA), motion and color features are condensed before a supervised classification method, hidden Markov model (HMM), is applied [9]. These methods provide meaningful browsing or classification of video frames; however, important, underlying nonlinear articulation structures are not taken into account by them. In recording or making videos, different image frames are coherently and intrinsically articulated into a semantically meaningful entity, e.g., movie, by the underlying articulation structures, which are naturally created by camera panning, zooming, tilting, and the motion or evolution of objects. In general, due to montage or the time-varying nature of the scenes, these underlying articulation structures are globally highly nonlinear, even for short clips, as shown experimentally [1]. We have proposed to exploit the underlying manifold information of image frames for visualization and summarization [1]. The manifold articulation primitives (MAP) are learned, and further processed to obtain better spatial layout for efficient browsing and summarization. One- and two-dimensional MAPs are experimentally found to capture the manifold information of video clips well, with very small residual variance [1].

In this paper, based on the embedded MAPs using a particular subspace learning technique, Isomap, we build a multiresolution video visualization and summarization system, called MV-MAP. Because the temporal sequences of the MAPs are nonstationary and have long durations, it is difficult to visualize all the frames in one limited space. We exploit wavelet analysis to flexibly and adaptably describe the sequence in the joint time-frequency domain, at multiple levels. Compared to the PCA, Isomap preserves the local topology when mapping the neighborhoods in the original feature space to those in the lower-dimensional one. The local topology is more important and reliable than global one in revealing the content coherence [10] [1], particularly for visualization and summarization purposes. We adopt Isomap for its robustness to noise and outliers. Features are extracted from image sequences for efficient representations. Color provides robust information about the contents [11] and will serve as our main features. Usually, the set of visual features may include color histogram, the first two moments in RGB space, and Tamura’s features: coarseness, contrast, directionality, linelikeness, regularity, or roughness [12]. To better capture the visual information or to obtain fast representation, more visual features or a subset can be utilized. For videos with audio component, the mel-frequency cepstral coefficients can be included, which

* The first and third authors are independent. The second is with ECE Dept., Wayne State Univ., Detroit, MI 48202. The fourth is with ECE Dept. and Beckman Inst., Univ. of Illinois, Urbana, IL 61801. Email: yinggewang@hotmail.com, qcheng@ece.eng.wayne.edu, chengjie_2000@yahoo.com, huang@ifp.uiuc.edu.
are often used in speech recognition and watermarking [13]. Those features are useful as many as several hundreds or even thousands. Isomap extracts the intrinsic manifolds and reduces the dimensionality. Based on the MAPs, two modes of video visualization and summarization are constructed [1]. The first discovers the landmark points signaling dramatic changes in video content in the temporal order. The second takes into account the global content coherence, without the temporal ordering. In the following, to facilitate the browsing of long sequences with complicated contents and structures, we build multiresolution visualization and summarization tools.

2. Multiresolution Roadmaps

For long image sequences or movies that have complex stories or structures, the roadmaps constructed using VMAP [1] may appear very complicated. Many details may become unidentifiable, because a large number of frames are squeezed within one limited displaying window. A natural approach to solving this problem is to construct roadmaps at multiple levels of details. At a high level, a viewer goes through the most important contents, with image sequence organized coarsely. When more details are to be found, the viewer can go down to lower levels to explore specific scenes. This way, the constructed roadmaps can satisfy different requirements for details or watching times. This strategy is similar in spirit to the roll-up and drill-down operations in the on-line analytic processing (OLAP) in a database [14]. To this end, we utilize multiscale techniques to build our MV-MAP based on VMAPs [1]. Specifically, we exploit wavelet analysis (WA) techniques. WA techniques have been utilized in many fields ranging from biomedicine to geographical information systems. They offer a flexible and adaptable means of describing nonstationary signals in the joint time-frequency domain. The basic idea includes using filter banks to select the information with different time-frequency resolutions. The filtered information at different levels contains hierarchical details. Both high- and low-frequency components are exploited by the MV-MAP.

The first mode of VMAP is obtained by first extracting the most significant 1-D manifold, then considering the changing rates (derivatives) of the manifold. The time course of this VMAP is a long vector, mathematically. We apply WA techniques to this time series. By analyzing the detailed signals, i.e., high-frequency components, we aggregate the time series of data into different clusters. The resulting cluster boundaries find landmark frames, indicating the beginning or ending of scale-dependent scenes, and the center of each cluster is a representative frame used for summarization. For quick visualization and summarization, the ratio of desired watching time against the duration of the whole sequence can lead to an estimated scale where we should access the roadmap. At this scale, the representative images from all clusters collectively provide a summarization of the video content, and the cluster boundaries give structural information to facilitate the visualization or further exploration of the contents. The second mode of VMAP is also exploited, in contrast to the first, we utilize the approximate signals to capture the content coherence. Both construction methods are given in the following.

3. First Mode of MV-MAP

The goal is to get a number of well-structured roadmaps with multiple scales. WA provides an efficient way to obtain such structures with hierarchical time-scale resolutions.

3.1. Multiresolution WA of VMAP for scalability

The first mode of VMAP of video trails is built by utilizing the difference signal of the most significant 1-D MAP. It provides structural information on significant changes in time. Because the difference (or derivative for continuous signal) can be regarded as the detail signal from the simple Harr transform, we extend this one-level construction to multiple scales by explicitly exploiting WA. This turns out to be a useful means for the visual description of the inherent structures of VMAP, especially, when there are rich details in a long image sequence with complicated contents. We adopt WA for multiresolution analysis with flexible time-scale localization capability. The salient landmark frames are revealed by significant spikes at consecutive levels.

We make use of the Harr wavelet for multiple-level transforms. Harr wavelet is well localized in time (able to provide arbitrarily sharp resolution in time). A discontinuity such as a significant landmark point can be localized with arbitrary precision, however, the frequency localization is not as good, since the Fourier transform of the Harr wavelet decays only as 1/ω, with ω → ∞. For better frequency resolution (thus poorer time resolution by the Heisenberg uncertainty principle), other wavelets such as Choi-William, Daubechies, Cohen families, may be utilized. We adopt the Harr wavelet from the broad wavelet families for two reasons: First, it helps us pinpoint the landmark points sharply; and second, it serves as a natural extension to the VMAP. By decomposing into multi-level detailed signals, the significant spikes are observed through consecutive levels, and are indicators to the salient landmark frames. The positions of the landmarks at level l need to be expanded by 2^l to restore their locations at the original scale. The multiresolution roadmap is helpful in accessing and browsing the video content at different scales.
3.2. Windowing long or fast-varying sequences

A long image sequence has rich details and vibrant scene changes. Consider, for example, a TV show has short commercial advertisements inserted, or a music video has fast changes of lighting conditions or frequent switches between foreground and background. The underlying manifolds may be complicated, giving rise to very spiky roadmaps. To better describe the underlying manifold structure, we low-pass filter the contents of the images with a short-time window. The processing can be sped up. More importantly, the roadmaps can be smoothened by filtering out short-lived changes or noise. The length of the window need be determined according to the content of the video sequence. For example, for teleconferencing or tele-education sequences, long windows may be used, since there are many static scenes with flat VMAPs. In contrast, if an image sequence has many fast-varying contents, we need adopt a relatively short window. In this paper, we choose to place those images within one second into a group, and use their average feature vector to represent them. We extract the underlying manifold utilizing the average vectors, based on which, the multiresolution roadmaps are built.

3.3. Highlighting salient landmarks

It is observed that the difference sequences of the embedded manifold have many variations at high levels. To highlight those important spots, we take a data-driven approach to eliminate insignificant variations. In a similar spirit to wavelet denoising, at each level, we regard those coefficients that are less than a threshold insignificant. The threshold is both scale and data dependent. We consider the choice of threshold in wavelet denoising [15], particularly, we employ a data-driven adaptive threshold selection method using the principle of Stein’s Unbiased Risk Estimate (SURE) [15]. Those coefficients whose values are below the threshold are set to zeros, while those above retained. Since the signs of detailed signals are not very relevant (the magnitudes of the changes are informative), we only utilize the magnitudes in our final MV-MAPs. The number of levels need be chosen so that desired resolution can be achieved. Especially, the decomposition can be terminated when all the images in the sequence are taken as a single cluster, i.e., when the detailed signal becomes flat. After constructing the multi-scale roadmap with the above procedure, summarization is performed, at each level, by selecting the representative images from those clusters delineated by consecutive significant landmarks.

For a Sports Motorcycle sequence, e.g., the detailed signals after an 8-level decomposition are obtained using the above procedure and plotted in Figure 1. They are the first mode of the MV-MAP for this sequence. At the coarsest level, the whole sequence can be regarded as one cluster; at the 7th, there are two clusters partitioned by one landmark; at the 6th, we can have three clusters, and so on. The first level corresponds to the most detailed roadmap for the image sequence. Visual inspection reveals that the correspondence between the landmark frames indicated by the MV-MAP and the shot/scene changes in the original sequence are well matched. The extracted frames for summarization using MV-MAP are semantically meaningful.

4. Second mode of MV-MAP

For long image sequences, the second mode of MV-MAP can be built based on the second mode of the VMAP. The procedures of feature extraction, windowing, manifold embedding are similar to those of the first mode described above. In exploiting multiresolution WA, however, the second mode makes use of approximate signals instead of detailed ones. This is determined by the purpose of the second mode which aims at global content coherence instead of content changes that are obtained from the first mode. The second mode of MV-MAP roadmaps reveals the similarity of image content by data aggregation. Relevant images aggregate into the same blob. The distance measure is important as pointed out in [1]. Making use of multiresolu-
lution WA, we decompose each dimension of the VMAP separately, that is, along horizontal and vertical directions independently. The number of decomposition levels is determined similar to the method in the first mode: The decomposition can be terminated when the detailed signals along both directions become flat. We summarize the original sequence, especially, the blobs of the MV-MAP well in the underlying the video articulation manifold. Two modes of multiresolution roadmaps are then built, revealing the dramatic changes or content coherence. Experiments corroborate the proposed method. Besides summarization and visualization, it may also find applications to video monitoring and surveillance. Integrating extracted video, audio, and textual information for interactive exploration on the roadmaps is a line of future research.

5. Conclusion

In this paper, we construct multiresolution roadmaps guiding the visualization and summarization of image sequences by exploiting the intrinsic geometric structure underlying the video articulation primitives. Two modes of multiresolution roadmaps are then built, revealing the dramatic changes or content coherence. Experiments corroborate the proposed method. Besides summarization and visualization, it may also find applications to video monitoring and surveillance. Integrating extracted video, audio, and textual information for interactive exploration on the roadmaps is a line of future research.

References