Video Analysis for Browsing and Printing

Qian Lin, Tong Zhang, Mei Chen, Yining Deng, Brian Atkins

HP Laboratories
HPL-2008-215

Keyword(s):
video mining, video printing, user intent, video panorama, video keyframe, video superresolution

Abstract:

More and more home videos have been generated with the ever growing popularity of digital cameras and camcorders. In many cases of home video, a photo, whether capturing a moment or a scene within the video, provides a complementary representation to the video. In this paper, a complete solution of video to photo is presented. The intent of the user is first derived by analyzing video motions. Then, photos are produced accordingly from the video. They can be keyframes at video highlights, panorama of the scene, or high-resolution frames. Methods and results of camera motion mining, intelligent keyframe extraction, video frame stitching and super-resolution enhancement are described.

External Posting Date: December 18, 2008 [Fulltext] Approved for External Publication
Internal Posting Date: December 18, 2008 [Fulltext]

Submitted to 10th International Workshop on Image Analysis for Multimedia Interactive Services, London, May 6-8, 2009

VIDEO ANALYSIS FOR BROWSING AND PRINTING

Qian Lin, Tong Zhang, Mei Chen, Yining Deng, Brian Atkins

Hewlett-Packard Laboratories

ABSTRACT

More and more home videos have been generated with the ever growing popularity of digital cameras and camcorders. In many cases of home video, a photo, whether capturing a moment or a scene within the video, provides a complementary representation to the video. In this paper, a complete solution of video to photo is presented. The intent of the user is first derived by analyzing video motions. Then, photos are produced accordingly from the video. They can be keyframes at video highlights, panorama of the scene, or high-resolution frames. Methods and results of camera motion mining, intelligent keyframe extraction, video frame stitching and super-resolution enhancement are described.

1. INTRODUCTION

To preserve precious memories of life, people record vast amounts of video using cameras and camcorders. These days most digital cameras have video capture functionality, capturing video at VGA resolution or above. Some cameras are even able to capture hours of high definition video. People capture home video for a variety of purposes. One is to capture action and sound. Many trophy shots representing fleeting moments have been captured on home videos. Another one is to capture the environment. Still another is to capture an object that has caught the attention of the video-grapher. Currently, there is no good way of browsing home videos. For example, with a two-hour video on the hard disk, the prevalent way to know what's in the video is to open up the media player and either play, or move the scroll bar to a certain point to play from that point. Some DVD authoring software can generate chapters, but the chapters are typically represented by the first frame where there is a scene change. The goal of this research is to find better ways to represent videos by photos, so that video files can be easily browsed, and collages can be created for printing.

Our approach is to analyze the frame-to-frame motion profile in a home video to derive the camera movement, such as panning and zooming. Based on that, we can generate the appropriate photos that represent the video. For example, a panning motion may indicate that the user is trying to capture the environment. We can stitch the video frames together to generate a panoramic photo that represents the scene. A zooming-in motion may indicate that there is an object that the user is interested in, and wants to capture it at greater detail. We can enhance the frame of interest to generate a photo with the object at higher resolution. If there is considerable object motion, it indicates that the user is trying to capture the motion. We can extract representative frames and produce action prints. Using this process, we can derive a collection of photos in the form of panorama shots, detail shots, and action sequences. To browse a video file, one can either see a slide show of the photo collection, or a layout of the photos. In this paper, we describe the video analysis framework for browsing and printing, and also give an overview of technologies developed under the framework and their applications.

While there is plenty of published work in this area, most papers address only part of the problem. For example, one recent paper finds web video segments that are suitable as panorama sources[1]. In this work, we propose a framework of mining video motion to determine user intent, and derive the appropriate representations. We have integrated individual technologies to provide a complete solution to the problem.

The rest of the paper is organized as follows. In section 2, the proposed video to photo framework is presented. Algorithms and results for motion mining, keyframe extraction, video panorama, video super-resolution and video representation are described in section 3. Conclusions and future work are discussed in section 4.

2. VIDEO TO PHOTO FRAMEWORK

Figure 1 shows the proposed framework of video to photo. First, motion analysis is performed to determine motion types between neighboring frames, e.g. whether there is panning or zooming, and whether there is considerable object motion. Then frames with similar motions are clustered together. This will identify segments of video with similar motions. Depending on the capture type, we may extract keyframes, generate a panorama, or obtain a high-resolution photo. Technologies for motion mining, keyframe extraction, panorama generation, and video super-resolution are described in more details in the following sections. The resulted photos from video may be used in a number of applications, such as printing video story book, video post card, and panorama posters. They also can be used to produce slide shows of video.

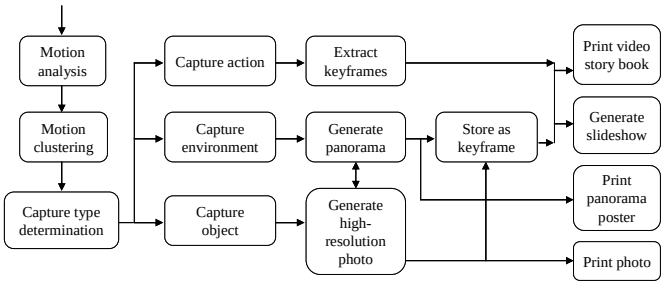

Fig. 1. Framework of the proposed approach for video to photo.

3. TECHNOLOGIES AND RESULTS

3.1. Motion Mining for User Intent Deduction

Prior research on motion mining was mostly focused on gathering data about human motion in video sequences. There has not been a solution provided for automatic repurposing of various segments in a video clip. In the case of panorama generation, although it has been researched extensively, prior solutions focus on rendering, under the premise that the whole video clip is known to have been taken with the intent for panorama creation. However, a segment

suitable for panorama stitching is often embedded within a long video sequence of various contents. Under existing technologies, the user would have to manually cut out the relevant part in order to have a panorama created, which requires the inconvenience of using video editing software. The proposed algorithm automates the above process, and it also has the ability to discover and repurpose video excerpts that are suitable for other kinds of creative rendering, such as action prints.

The proposed motion mining algorithm consists of three steps:

(1) **Motion Analysis.** We adopt the affine model for camera motion estimation, and compute it from the generally noisy optical flow estimates. For optical flow estimation, we use an iterative image alignment based matching technique that has been proven robust for creating dense correspondence maps under a variety of image motion scenarios[2].

(2) **Motion Type Clustering.** Next, camera motion information is quantized to generate a motion profile of the video sequence. For example, panning parameters are classified as insignificant, steady, or fast; and zooming parameters are classified as insignificant or significant. A hierarchical technique is used to cluster video frames based on their affine model similarity. Currently, we consider the following motion types: panning, zooming, object motion, and still. At the end, consecutive video frames are grouped based on estimated motion parameters. This results in a number of frame groups within which camera motion is considered to be consistent, and the average camera motion is computed for each group. It is then used to merge groups into larger clusters of similar camera motion. This process iterates to generate a set of video segments.

(3) **User Intent Determination.** This module infers user intent by detecting a video motion type, such as panning, zooming-in, still, and moving object(s). It enables the repurposing of a video into a variety of output forms that are adapted to the user's intent. More specifically: (a) Within a panning motion, the algorithm decides user intent by analyzing the velocity and acceleration of the motion. Fast camera panning suggests that the user intended to quickly move to an area or object of interest, and had little or no interest in the intermediate areas. Relatively constant camera panning suggests an attempt to capture a panorama. In this case, the panorama generation module may be invoked to automatically render a panoramic image. (b) A camera zoom-in motion followed by camera stillness lasting longer than a predetermined time interval indicates the user zoomed in to record an object of interest. As the result, the super-resolution enhancement module will be called automatically to render an image with the object of interest at a higher spatial resolution. (c) If the magnitudes and/or directions of local motion vectors vary beyond a predetermined threshold, it is determined that the user was trying to record an object(s) in motion. In this case the keyframe extraction module is invoked to select keyframes from the video segment.

We tested 31 home video clips, out of which 23 contain panoramic segments. Among the rest, 2 have zoom-in segments. Our approach failed on one panoramic clip, achieved partial-success on 4 clips by separating compound panoramas (panoramas with both rotation and panning, or both panning and tilting) into individual ones, and succeeded on the rest 18 clips. For the 2 camera zoom-in motions, the algorithm detected one, but rejected the other because the duration of camera stillness following the zoom-in motion was below the preset threshold.

3.2. Keyframe Extraction

Most prior work on extracting keyframes from video is targeted at longer video recordings that contain complicated structures and multiple shots. The common routine is to extract keyframes based on segmenting video into shots and/or scenes. For single-shot video clips (e.g. those taken with digital cameras), keyframes are normally obtained by sampling evenly over time. Such an approach, however, may yield keyframes that (a) do not reflect highlights or regions of interest and (b) are inappropriate for printing due to content or quality. To solve this problem, we developed an intelligent keyframe extraction approach to derive better keyframe sets by performing semantic analysis of the video content. One example is shown in Fig.2. On the left side are nine keyframes extracted by even sampling. It can be seen that none of these keyframes gives a good view of the major character (i.e. the person riding the bicycle). On the right side are nine keyframes extracted by our algorithm. Apparently the highlight of the video was detected, and by sampling more intensively at the highlight region an obviously better keyframe set was obtained.

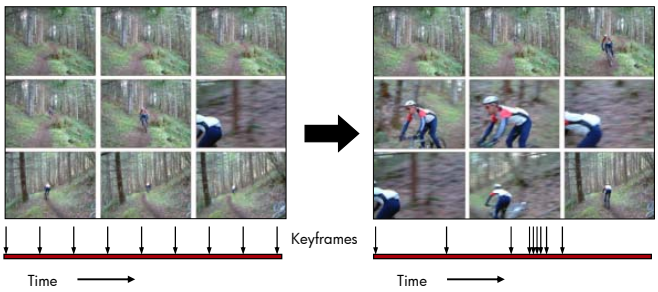

Fig. 2. Keyframe extraction from an action video.

Our solution consists of two main steps as follows.
(1) **Generating Candidate Keyframes.** To show different views of the scene, the accumulative color histogram difference and the accumulative color layout difference are computed, so that selected candidate keyframes are quite different from each other in terms of either color histogram or color layout. To detect highlights in video content, camera motions are tracked. For example, if there is a period of focus after a period of panning or zooming motion, it might indicate a scene or object of interest, and a candidate keyframe is selected. On the other hand, fast camera motions in a video suggest content of no interest, and no candidate keyframes are chosen. Another way to show highlights is to trace moving objects in the video, so that a candidate keyframe can be selected with a moving object of the right size and position in the frame. While detecting moving objects without *a priori* knowledge is a difficult problem, our method is based on the assumption that the foreground motion behavior is different from that of the background. Then, by applying a series of filtering to the residual image after camera motion estimation, the moving object is detected[3]. A face detection engine is included to get face number, positions, and sizes. Video highlights are also detected by searching for audio events such as screaming, music and speech.
(2) **Selecting Final Keyframe Set**. Candidate keyframes are clustered into a number of groups. The number of clusters can be either pre-determined by the user or variable. If variable, it is then determined by the complexity of the video content, i.e. the more diversity in the scene, the more clusters. An importance score is computed for each candidate keyframe based on camera motions; human faces in the frame; the size and position of moving object(s)

in the frame; as well as audio events detected. The image quality of each candidate frame is also evaluated. Then, one representative frame is selected from each cluster based on the importance score, closeness to the center of cluster, and sharpness of the frame.

This keyframe extraction approach was tested on 32 video clips from digital cameras with lengths ranging from 3 seconds to 2 minutes. Table 1 compares our results with keyframes evenly distributed over time. Keyframe sets were evaluated subjectively at a 3-level scale defined as: excellent – good for browsing and printing; fine – ok for browsing, but need tuning for printing; and poor – not complete, having semantically meaningless frames or with obvious redundancy among keyframes. In general, results from our approach tell more complete stories, have better image quality, and exhibit much less redundancy.

Table 1. Comparison of keyframe extraction results.

	Proposed scheme	Evenly spaced frames
Excellent	29	15
Fine	3	3
Poor	0	14

3.3. Video Panorama

Most software applications that generate panorama photos from videos do not handle complicated camera motions such as the zigzag pattern. Our goal is to produce panoramas in the presence of complicated camera motions as well as object motions. The basic processing scheme, as shown in Figure 3, has two phases: the alignment phase and the stitching phase. During the alignment phase, motion estimation is performed on each video frame, and a subset of frames that exhibit significant shifts are selected. During the stitching phase, the selected frames are stitched together to form a panorama photo.

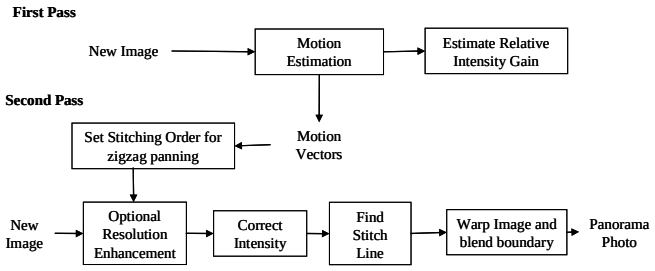

Fig. 3. Two pass scheme to generate panorama photos.

Fig. 4. Example of panorama generation: even though there are objects moving across the scene, the result avoids cutting through moving objects and stitches seamlessly.

We developed a special stitching method that finds a line where two overlapping video frames have the best agreement with each other, and then blend through this stitch line so that the stitching becomes seamless[4]. Because there is perspective distortion and object motion in the scene, alignment may not be perfect at the pixel level. We also consider variations in camera exposure and scene light level, and compensate by looking at the relative average intensity difference in the overlapped region. Figs.4 & 5 show two examples where at the top are samples of selected video frames for stitching, and at the bottom is the resulting panorama.

Fig. 5. Example of panorama generation: although there are significant intensity differences among the frames, the result shows corrected intensity and smooth blending.

3.4. Video Super-Resolution

The general approach of enhancing image quality by synthesizing information from multiple frames has been researched extensively. However, existing approaches attempt to combine information in frames without discriminating different content in the scene, and thus tend to be sub-optimal for dynamic scenes. For example, for a scene with a slow moving pedestrian and a fast moving car, there are more consecutive frames that share complementary information for the pedestrian than for the car. Therefore, if all images sharing information of the pedestrian are used to enhance the resolution of the frame, it's very likely that the result will be noisy around the car, due to higher possibility of errors in image correspondence. On the other hand, if only images sharing sufficient information of the car are used for resolution enhancement, the result for the pedestrian will be compromised since much available information is not used.

To address this issue, we developed a dynamic approach where information that contributes to the rendering of a distinct object is extracted from frames selected based on relevance to that specific object. Associations between image content and relevant frames are obtained using image cues such as motion information. Notice that this allows selective enhancement of the image scene, which is efficient when there is limited computing resources[5].

The proposed super-resolution system consists of three modules. The ***inter-frame motion estimation module*** computes dense correspondence maps between frames of interest (i.e. the reference frame and a number of neighboring auxiliary frames). (2) The ***motion-based scene segmentation module*** then classifies regions in the reference frame according to estimated motion vectors. Specifically, regions with small motion vectors between the frame of interest and all the auxiliary frames are classified as having low motion; regions with large locally coherent motion vectors between the frame of interest and any auxiliary frames are

classified as having high motion; whereas other regions are grouped as having intermediate motion. This motion segmentation process is done in a hierarchical manner to ensure region connectivity and continuity. (3) The ***dynamic content adaptive super-resolution synthesis module*** reconstructs a high resolution image integrating complementary information from neighboring auxiliary frames. In particular, for regions of high motion, fewer auxiliary frames are used as only closely adjacent video frames contain sufficient information that can be reliably registered and synthesized; whereas for regions of low motion, a larger number of auxiliary images are employed as more frames can be correctly corresponded and integrated; regions with intermediate motion are synthesized with a medium number of auxiliary frames. A final refinement process is conducted on the synthesized image to ensure consistency between regions.

This method was tested on a large number of videos with dynamic content, and achieved satisfactory results. Fig.5 demonstrates the enhancement effect on a parked van with text and on a girl riding a bicycle. The system was able to adaptively enhance the resolution without compromising accuracy by exploiting appropriate information. (Note that the previously illegible text on the van is now clearly readable, and the quality of the girl's face region is considerably improved without additional motion artifact.) We have also compared the performance of our system with that of a commercially available software package. While the performances are comparable for videos with moderate motion, our system outperforms considerably in cases of large image motion (i.e. the commercially available software tended to generate results with greater motion blur).

Fig. 5. Original video frame (left) and super-resolution enhanced to 2x (right). Upper: result at low motion region; lower: result at high motion region.

3.5. Video Representation

Once we obtain the keyframes that represent actions in video, panoramas that represent the environment, and high resolution photos that represent important shots, we can generate various output types. One example of representing the entire video collection is to generate a video thumbnail page using an automatic photo layout algorithm we developed[6]. One important advantage over a regular video thumbnail page is that we can provide a much better indication of the video content – what actions took place, the surroundings captured, and the important frames. Moreover, such an output can be printed out as a story book or a poster, representing highlights and details of the video in high quality.

Shown in Fig. 6 is an example of video thumbnail page generated with our approach. At the upper row, the image on the left is a super-resolution picture of a video frame which gives a clear view of the two riders in the video clip. In the middle and on the right side are two keyframe sets extracted from two action video clips, with the number of keyframes being four and nine, respectively. At the lower row, there are two panorama pictures produced from two video segments with panning motions. Overall, this video thumbnail page provides a complete and diversified view of video clips in the collection, including stories, actions, sceneries, and details.

Fig. 6. Video thumbnail page generated with proposed approach.

4. CONCLUSIONS AND FUTURE WORK

A complete solution of video to photo was presented in this paper. Under the proposed framework, we developed novel technologies to detect and extract semantics of camera and object motions, to produce panoramas from video, to generate high resolution photos from adjacent video frames, to generate keyframes for action shots, and to combine different types of outputs together with automatic layout generation. These technologies have been tested and proved to be superior to prior approaches; and they form a powerful system when integrated together. There are numerous applications of this system in video browsing and printing.

We are planning to continue working on the framework, focusing on an improved integration of these techniques to achieve better efficiency and effectiveness.

5. REFERENCES

[1] F. Liu, Y. Hu, M. Gleicher, "Discovering panoramas in web videos," *Proc. ACM Multimedia Conference*, p.329-38, Oct. 2008.

[2] H. Sawyhney and S. Ayer, "Compact representation of video through dominant and multiple motion estimation," *IEEE Trans. Pattern Analysis Machine Intelligence*, pp.814-830, August, 1997.

[3] Y. Wang, T. Zhang, D. Tretter, "Real time motion analysis toward semantic understanding of video content," *Int. Symposium Visual Communications and Image Processing*, Beijing, July 2005.

[4] Y. Deng, T. Zhang, "Generating Panorama Photos," *Proc. of SPIE Conf. on Internet Multimedia Management Systems IV*, vol. 5242, ITCOM, Orlando, Sept. 2003.

[5] M. Chen, "Dynamic content adaptive super-resolution," *Int. Conf. Image Analysis & Recognition*, Sep. 2004.

[6] C. B. Atkins, "Blocked Recursive Image Composition," *ACM Multimedia Conference*, Vancouver, BC, Canada, 2008.