Mining Home Video for Photos

Qian Lin, Tong Zhang, Mei Chen, Yining Deng, Pere Obrador
Imaging Systems Laboratory
HP Laboratories Palo Alto

HPL-2004-80
April 29, 2004*

home video analysis, video to photo, video printing, camera motion analysis, intelligent keyframe extraction, video panorama, super-resolution, user intent, motion mining

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.

* Internal Accession Date Only Approved for External Publication

Mining Home Video For Photos

Qian Lin, Tong Zhang, Mei Chen, Yining Deng, Pere Obrador

1. Introduction

To preserve precious memories of life, people record a vast amount of video using movie cameras and camcorders. Recently, many digital cameras implemented video capture functionality as well, capturing video up to VGA resolution. In a few years, compact digital cameras will be able to capture hours of high definition digital video.

People capture home video for a variety of purposes. One of the main purposes is to capture action and sound. Many trophy shots representing a fleeing moment have been captured on home video. Another one is to capture the environment. Still another is to capture an object that has caught the attention of the video photographer. In many of these cases a photo, whether capturing a moment or a scene, provides a complementary representation to the video. It can bring a better total experience to the user.

Our approach is to analyze the motion in home video, and derive the camera movement, such as panning and zooming. From the camera movement, we infer the intent of the user while capturing the video, and generate the appropriate photos that can represent the video.

For example, if the camera is in steady panning motion, it indicates 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. If the camera is zooming in, it indicates that there is an object the user is interested in, and wants to magnify it. We can enhance the frame to generate a photo with the object at higher resolution. If there is significant object motion, it indicates that the user is trying to capture the motion. We can extract representative frames from the video and produce action prints.

While there is a lot of research in this area described in the literature, most of the work only addresses some aspects of the problem and provides piece-wise solutions. Instead, we are proposing a new framework of mining video motion to determine user intent and derive the appropriate representations. And we have integrated individual technologies to provide a full solution to the problem.

The rest of the report is organized as follows. In section 2, the video to photo framework is presented. Methods for motion mining, keyframe extraction, video frame stitching, and video super-resolution are described in section 3. Implementation issues and experimental results are discussed in sections 4 and 5, respectively. Finally, concluding remarks are given in section 6.

2. The Video to Photo Framework

Figure 1 shows the overall framework of video to photo. We first perform motion analysis to determine the image motion between frames. We then cluster frames with similar motion together to identify segments of video with similar motion types. These motion types are further classified to infer the user's intent for particular segments of the video. Depending on the inferred intent of the user, we repurpose the video segments to automatically extract key frames, generate panorama, or generate high-resolution photos.

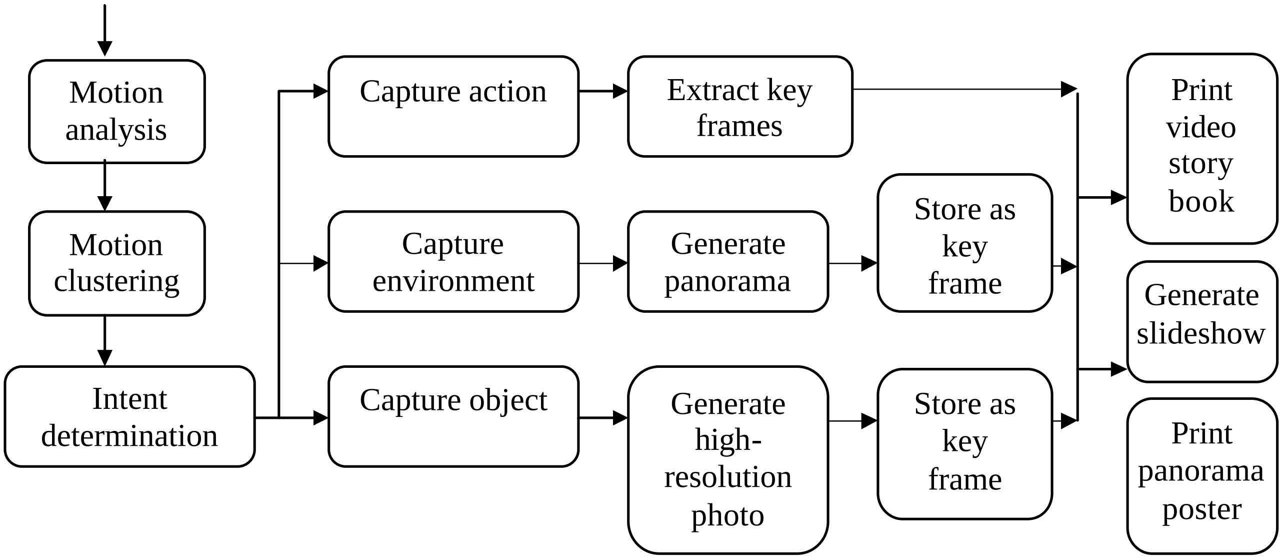


Figure 1: Overall framework for video to photo.

The resulted photos from video may be used in a number of applications, including printing video story book, video post card, and panorama posters. They also can be used to produce slide shows of video.

3. Methods and Algorithms

In this section, technologies for motion mining, keyframe extraction, panorama stitching and video super-resolution will be described in more details.

3.1. Motion Mining for User Intent Deduction

The proposed motion mining algorithm consists of three steps: it first analyzes the image motion between each pair of video frames; secondly, it finds clusters of adjacent video frames with similar motion types; and lastly, it derives the user's intent based on the each specific motion type, and repurpose the corresponding video segment for various creative rendering.

Prior research on motion mining focused on gathering data about human motion in video sequences. To the best of our knowledge, there hasn't been solution provided for automatic repurposing of various segments in a video clip. In the case of video 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, in the reality of consumer videos, a segment that is suitable for panorama stitching is often embedded within a long video sequence of various contents. Under existing technology, the user would have to manually cut out the relevant part in order to have a panorama created, which requires the inconvenience of using a video editing application. The proposed algorithm automates the above process, as well as having the ability to discover and repurpose video excerpts that are suitable for other kinds of creative rendering, such as action prints.

3.1.1. Motion Analysis

We adopt affine model for camera (global) motion, and compute it from the generally noisy optical flow estimates. For optical flow estimation, we used an iterative image alignment based matching technique that has been proven to be robust for creating dense correspondence maps under a variety of image motion scenarios [9]. For affine model computation, we adopt the least squared error (LSE) regression method [10,11].

3.1.2. Motion Type Clustering

After motion analysis, the proposed algorithm clusters video frames into a set of segments, with each segment corresponding to video frames having the same motion type, such as panning or zooming-in. This is done in the two steps:

1) Camera motion is quantized to create a motion profile of the video sequence. For example, panning parameters a classified into insignificant, steady, or fast, and zoom parameters are classified as insignificant or significant.

2) A hierarchical technique is used to cluster video frames based on their affine model similarity. The current implementation considers the following motion types: Panning, zooming, object motion, and still. At the bottom of the hierarchy, video frames within short temporal distances are grouped together based on estimated motion parameters. This results in a number of groups of video frames within which the camera motion is considered to be consistent (within a preset threshold), and average camera motion is computed for each group. The estimated average camera motion for each group is then used to merge groups into larger clusters of video frames of similar camera motion. This process is iterated to generate a set of video segments within which the camera motion is consistent.

3.1.3. User Intent Determination

The intent determination module infers a likely intent of the user by detecting a motion type in the video that may indicate an intent, 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 likely intent of the user, specifically,

1) Within a panning motion type, the algorithm determines user intent by analyzing the velocity and acceleration of the panning motion. If the velocity of the camera panning is very fast, it will be inferred that the user intended to quickly move to an area or object of interest, and had little or no interest in the intervening areas. On the other hand, if the

velocity of the panning motion is relatively constant, the algorithm will interpret that as an attempt to capture a panorama, and will record indices of the video frames at the beginning and the end of the panning period. In our integrated system, the panorama generation module will be invoked with these indices to automatically render a panoramic image.

2) In the case of a camera zoom-in motion, if it is followed by a still motion type that lasts over a predetermined length of time, the algorithm will decide that the user intent was to zoom in to record an object of interest and will record indices of the video frames at the beginning and the end of the still motion period. In our integrated system, the super-resolution enhancement module will be called with these indices to automatically render an image with the object of interest at a higher spatial resolution.

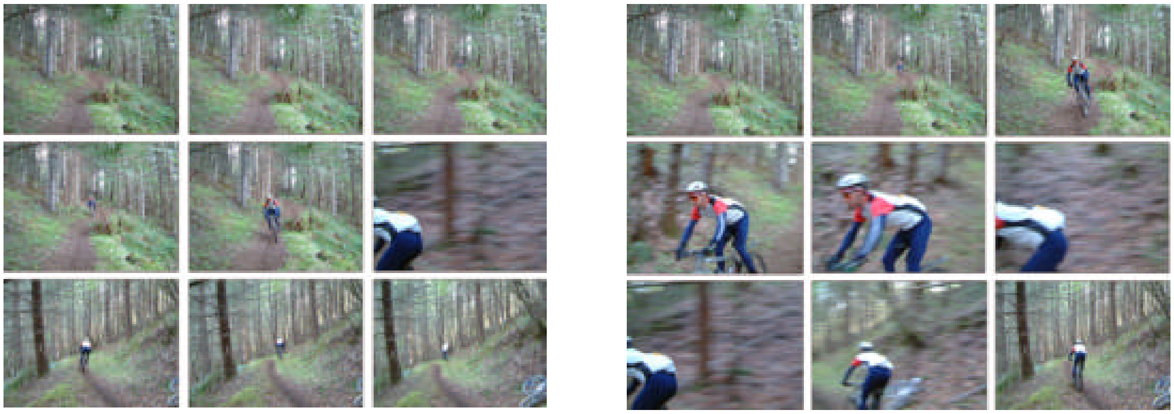
3) If the magnitudes and/or directions of local motion vectors within a video frame vary beyond a predetermined threshold, the algorithm will conclude that the user was trying to record an object or objects in motion, and will record the indices of the frames at the beginning and the end of that video segment. In our integrated system, the keyframe extraction module will be evoked with these indices to automatically select relevant key frames from the video segment.

3.2. Intelligent Keyframe Extraction

There are prior arts on extracting keyframes from video which, however, are all targeted at longer video recordings (movies, TV programs, tapes of home video, etc.) that contain complicated structures and multiple shots [1 - 3]. Distinguished from existing approaches, in this work, we focus on short video clips containing single shot (e.g. those taken with digital cameras).

In most current video printing systems, keyframes are obtained by evenly sampling the video clip over time. Such an approach, however, may not reflect highlights or regions of interest in the video. Keyframes derived in this way may also be improper for video printing in terms of either content or image quality. One example is shown in Figure 2(a), where nine keyframes were evenly sampled from a video clip, and none of them provides a good view of the major character (i.e. the person riding the bicycle) in the video.

In this work, we developed an intelligent keyframe extraction approach to derive an improved keyframe set by performing semantic analysis of the video content. This approach is targeted at both browsing and printing of video clips. The goal is to automatically generate a keyframe set from a video clip which can show different people and different views in the video; tell a complete story of the scene; or show highlights in an action video. This intelligent approach also reduces redundancies among keyframes; removes semantically meaningless frames from the keyframe set; as well as avoids bad frames (e.g. images which are too blurry, too dark or too bright). Figure 2(b) shows the keyframe set obtained from our developed method. It can be seen that more details of the video content are revealed in the improved keyframe set by automatically detecting highlight of the video.



(a) evenly-spaced keyframes

(b) keyframes from proposed approach

Figure 2. Extracting keyframes from an action video.

In the proposed scheme for intelligent keyframe extraction, for a video clip, a number of video and audio features are analyzed to first generate a candidate keyframe set. Then, the candidate key-frames are clustered and evaluated to obtain a final keyframe set.

3.2.1. Generating Candidate Keyframes

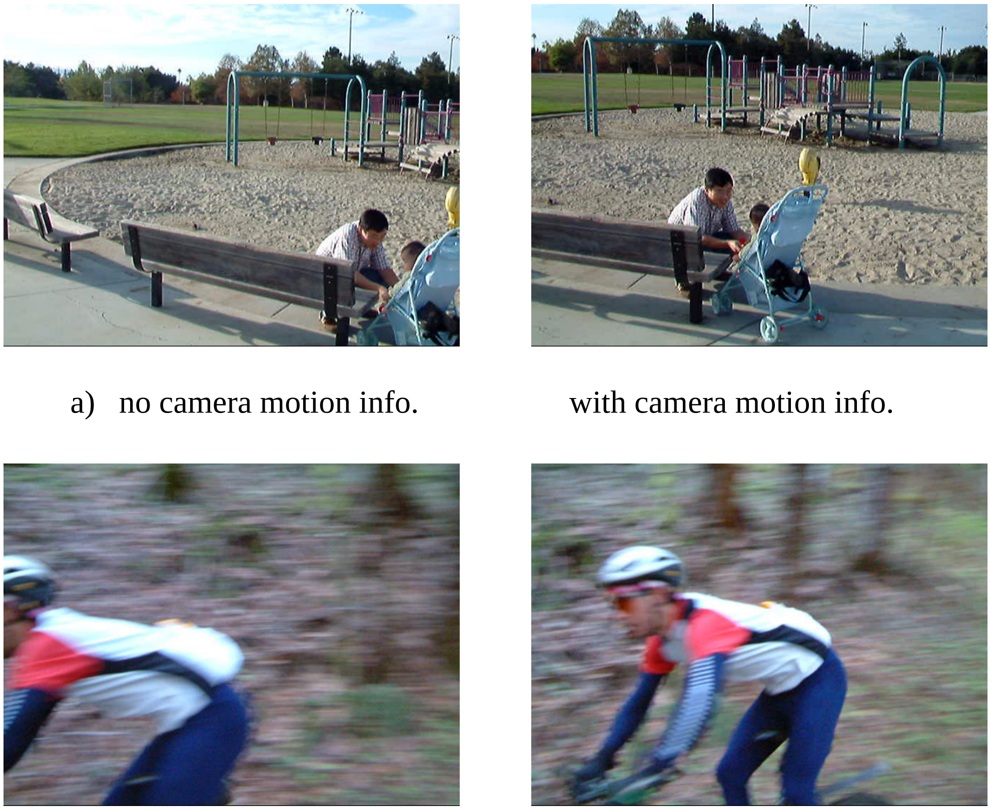
To show different views of the scene, we compute accumulative color histogram difference, as well as accumulative color layout difference in selecting candidate keyframes, so that the selected frames are quite different from each other in terms of either color histogram or color layout.

To detect highlights in the video content, camera motions are tracked in order to guess the intention of the video photographer. 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. In another type of case, if there is panning motion after a relatively long period of focus at the beginning of video, it might be an indication of coming highlight, and a candidate keyframe is chosen at the beginning of the motion. On the other hand, fast camera motions in a video suggest content of no interest, and no candidate keyframes are chosen from there. Skipping frames during fast motions also helps to avoid getting blurry keyframes. To track camera motions, we developed a method for fast camera motion estimation, as well as a procedure to extract semantic meanings from the camera motion information.

Another way to show highlights is to trace moving objects in the video, so that we can select a candidate keyframe with a moving object of the right size and position in the frame. While detecting moving objects without any a prior knowledge is a difficult problem and there is no existing approach for it, our proposed 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. Next, the moving object is tracked through following frames of the video using a fast searching algorithm.

In Figure 3, keyframe examples extracted with or without camera motion and object motion rules are shown, respectively. It can be seen that by following camera motion and tracking object motion, keyframes with much better content can be obtained.

Video highlights are also detected by searching for special audio events, such as screaming and the appearance of noise or speech after a period of silence. Audio processing tools we developed are able to detect such audio events by analyzing audio features in the time and frequency domains.



a) no camera motion info. with camera motion info.

(b) no moving object info. with moving object info.

Figure 3. Keyframes obtained with or without following
(a) camera motion and (b) object motion information.

To show people in the video, we include a face detection engine in our approach. Face detection is in general a time-consuming process. It may also generate an excessive number of false alarms. Special methods are being investigated to modify the face detection engine to overcome these problems. Promising results have been obtained in running face detection in real-time for video processing, with a relatively high accuracy.

3.2.2. 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 it is variable, then it will depend on the complexity of the video content, i.e. the more diversities in the scene, the more clusters. An importance score is computed for each candidate keyframe according to the camera motion rules; the number, size and position of human faces involved with the frame; the size and position of moving object in the frame; as well as audio events detected in the frame. Each candidate frame is also evaluated in terms of image quality, including sharpness and brightness, so as to avoid frames which are too blurry, too bright or too dark. Then, one representative frame is selected from each cluster, based on comparing importance scores, closeness to the center of the cluster, and sharpness of the frame.

3.3. Video Panorama

A panorama photo can be stitched from a sequence of video frames. One advantage of using video instead of still photos is ease-of-use. It is a lot more easier for a user to take a video shot that pans the scene than having to hold the camera, taking a sequence of images, and aligning them along the way to insure proper amount of overlaps between adjacent photos. Another advantage is that the accuracy of stitching video frames is higher than still photos. This is because at 30 frames/sec, video frames have much more overlapping space and less 3D perspective distortion in between them and therefore the alignment estimation between the frames is more accurate.

Our work on panorama stitching has been described in [4] and has several advantages over other approaches [5 - 8]. We briefly summarize this method here. The basic processing scheme 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 with significant motion shifts in between are selected. During the stitching phase, these select frames are stitched together to form a panorama photo.

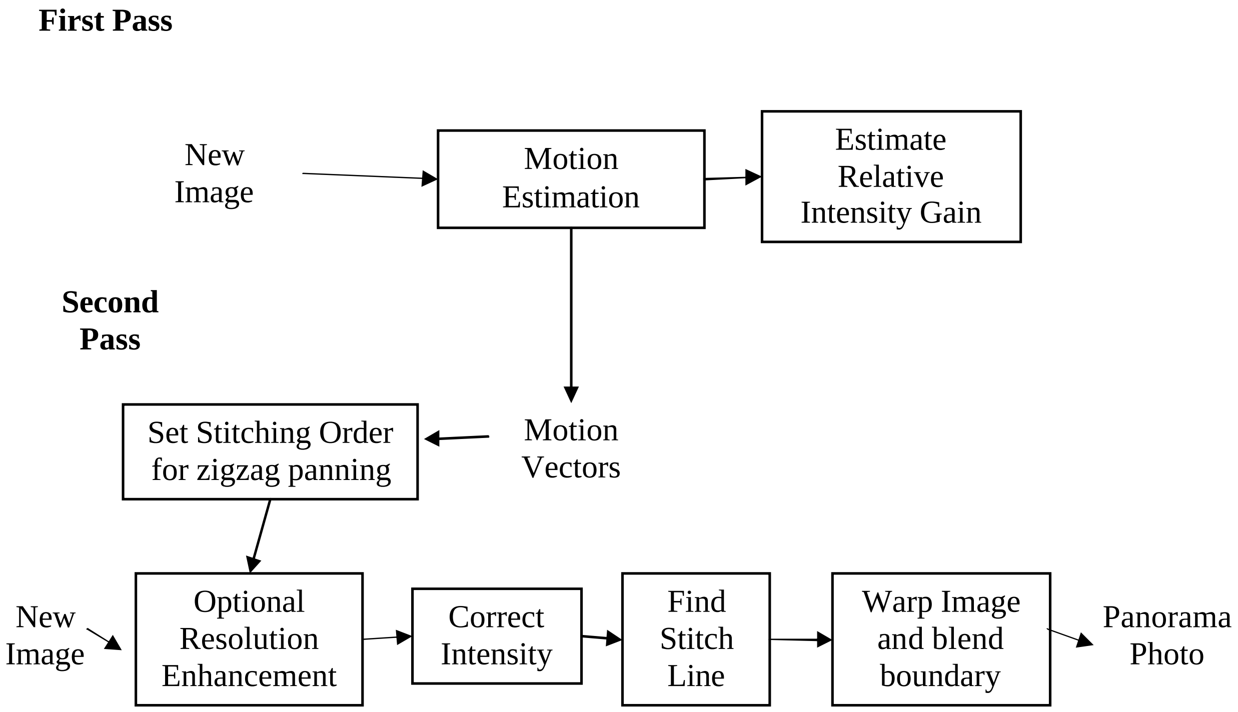

First Pass
New Image
Motion Estimation
Estimate Relative Intensity Gain
Second Pass
Set Stitching Order for zigzag panning
Motion Vectors
New Image
Optional Resolution Enhancement
Correct Intensity
Find Stitch Line
Warp Image and blend boundary
Panorama Photo

Figure 4. Two pass scheme to generate panorama

We developed a special stitching method that finds a line where two overlapping video frames have the best agreement between in other (see Figure5), then blend through this stitch line so that the stitching becomes seamless. When two images are aligned to each other, because there is perspective distortion and possible objects moving in the scene, alignment is not perfect at pixel level. Our method involves several steps:

1) It finds a strip in the overlapping region that gives the minimum error.

2) It finds a stitching curve within this strip that bypasses all the possible misalignments, including the ones resulting from moving objects, perspective distortion, or alignment errors.

3) Weighted blending is applied to pixels around the stitching line to mix the data from the overlapping images. This weighted blending is done together with Gaussian Pyramid filtering to blur the boundary and yet preserve the high frequency texture area in the image.

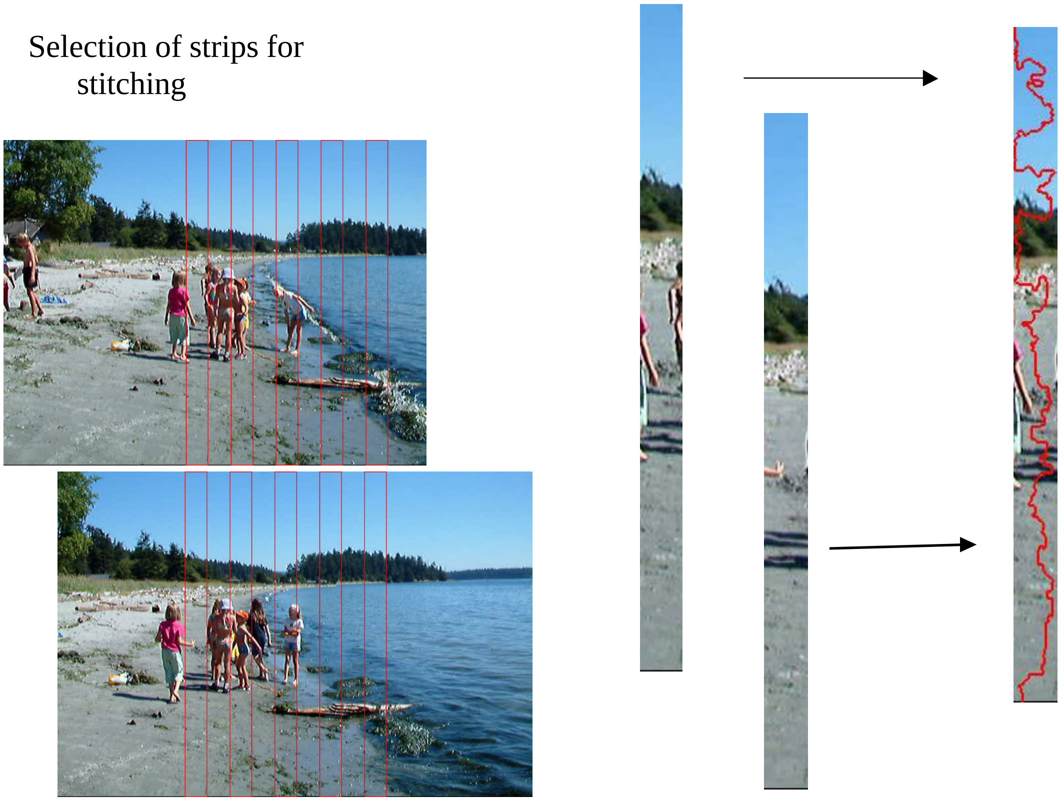


Figure 5(a). Stitching across the best agreement.

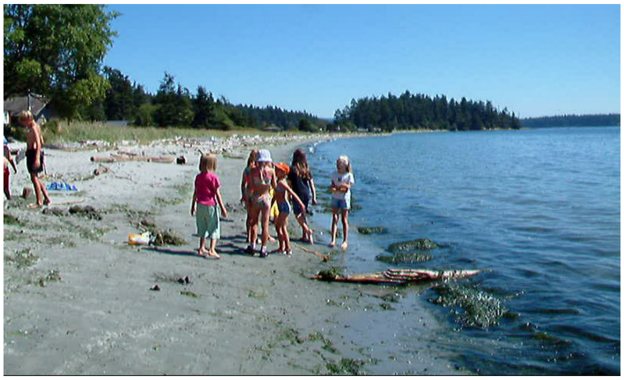


Figure 5(b). Stitching Result.

In additional to our unique stitching method, we compensate camera exposure variations by looking at the relative average intensity difference in the overlapping region. We also designed a way to group video frames for stitching to handle complicated camera motion such as the zigzag pattern. The result is a seamlessly stitched panorama photo generated from a video clip that is panning the scene.

3.4. Video Super-Resolution

As anyone who has ever tried to “blow up” a digital photograph knows, the resolution of the sensor is a fundamental limit. This limitation is even more pronounced for digital videos, for which a typical VGA video only has 0.3M pixels per frame. Interpolation can be used to display an image on an equipment with higher resolution, but this adds no additional information, and often results in visual degradation.

With still photographs interpolation is the best we can do, as information not captured by the sensor is permanently lost. However, this is not necessarily true for a video sequence. Imagine the scene as an array of N x N sub-pixels, with each of the sub-pixels much smaller than the M x M pixels captured by the camera sensor, and the camera is moving in a continuous manner relative to the scene. Each video frame will include a different M x M subset of the N x N sub-pixels. On the other hand, there will be macroscopic features of each sequential image that are related. For example, if the scene includes a car moving from the right to left, the car will be moving from right to left in many consecutive video frames, even though completely different M x M subsets of the N x N scene are being sampled. As a result, in real video sequences, each frame includes a mixture of new and previous information. Therefore, it should be possible to extract from each successive video frame the new information that could be used to enhance the resolution of a single frame.

The proposed system adopts a dynamic approach to enhance the spatial resolution of a video frame. It first estimates dense correspondence maps between the frame of interest and a number of neighboring auxiliary frames [9], it then classifies regions in the frame of interest according to the estimated motion magnitude. 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 motion vectors between the frame of interest and any of the 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 continuity. The higher resolution image is constructed by synthesizing complementing information from the auxiliary video frames. For regions of *high motion*, fewer auxiliary frames are used because only closely adjacent video frames share sufficient common feature; whereas for regions of *low motion*, a larger number of auxiliary images are employed as they will share more macroscopic feature with the frame of interest; and regions with intermediate motion are synthesized using a medium number of auxiliary frames. A final refinement process is done on the synthesized image to ensure consistency between regions (for details, refer to [14]).

This dynamic approach allows different regions of the scene to be treated differently, and is able to avoid artifacts that otherwise might result from treating all regions of the scene in the same way during the resolution enhancement process. In addition, it is also able to dynamically tailor the image resolution enhance process in an intelligent way, by deploying image processing resources to different regions of an image at varying computational intensity levels to achieve resolution enhancement in an efficient way.

3.5. High Resolution Video Panorama

One of the main disadvantages of a video panorama, with respect to a high resolution still image panorama (i.e., each high resolution still image is stitched with the adjacent one, and so forth) is the final resolution of the final panorama.

For a good print we need at least 150dpi source, therefore a VGA video panorama would have approximately 480/150=3.2” tall, if generated in landscape mode.

Some type of interpolation (e.g. resolution synthesis) can be applied to the panorama in order to obtain a better 8.5” tall panorama. This would allow to print on high quality 8.5” roll media or banner paper.

But the fact that the original image information is coming from a video signal allows us to do much better than that. So we built a system where both technologies introduced above (i.e., video panorama and video super-resolution) were combined together in order to generate a high resolution video panorama.

This system will first run the first pass of the video panorama algorithm on the original video data in order to identify the video frames with sufficient displacement that will be used for the final stitching. Each of those video frames is then resolution enhanced by making use of information from original neighboring video frames by using the video super-resolution algorithm.

The final result is a much higher quality video panorama, in which the compression/color artifacts do not show much, and there is a true increase in resolution. In our experiments, a super-resolution enhancement of 2x followed with an extra resolution synthesis enhancement of 2x generate very high quality 8.5” tall video panoramas.

4. Implementation Issues

For keyframe extraction, significant amount of efforts have been made in optimizing the algorithms so that the whole procedure could be run in real-time, and could be easily embedded into an appliance. A series of measures are taken at different parts of the procedure, including temporal and spatial down-sampling, fast searching of matching blocks in motion estimation, use of look-up tables, simplification of routine algorithms, repetitive use of intermediate results, etc.

For panorama generation, steps are taken to correct for camera exposure changes. When the camera is taking video, it usually adjusts the exposure automatically to capture the best picture quality for each video frame. However, the intensity differences between could cause errors in motion estimation and severe artifacts in stitching panorama image. The correction of intensity difference between frames is done in both places. During stitching, additional care is taken for over-saturated pixel (values near 255) to avoid the clipping effect.

5. Results and Discussions

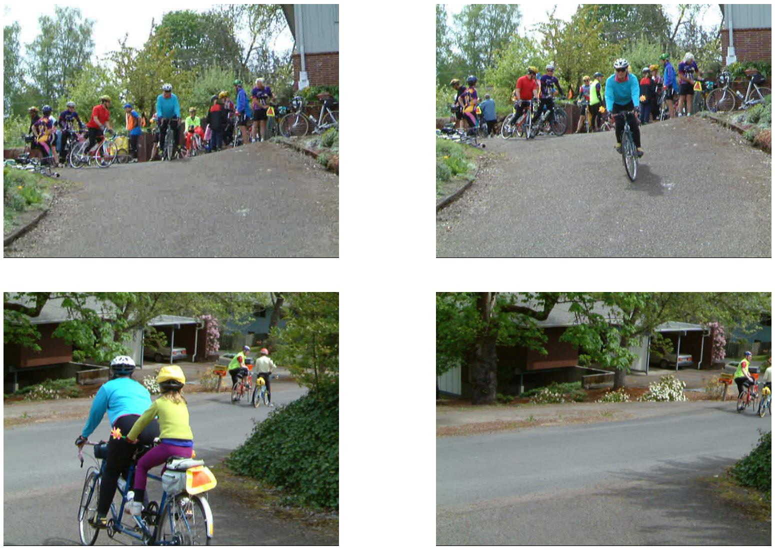
In this section, we present results from each module of our proposed system.

5.1. Motion Mining

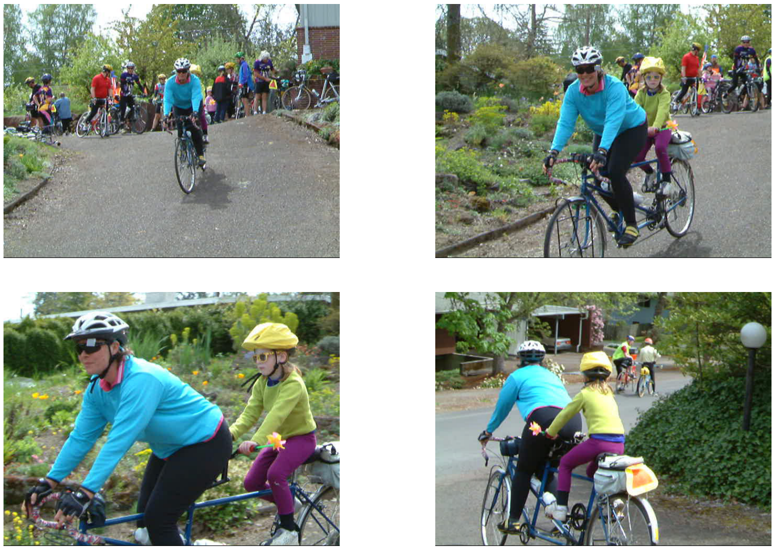
We processed 31 clips of home videos, out of which 23 contains panoramic segments. Among the rest 8, 2 have zoom-in sections. The current implementation failed on one panoramic clip, achieved partial-success on 4 clips by separating compound panoramas (panoramas with both rotation and panning, or panoramas with panning and tilting) into individual ones, and succeeded on the rest 18 clips. Among the 8 non-panoramic clips, there were 2 with camera zoom-in motion, the algorithm detected one of them, and rejected the other one because the duration of the *still* motion type following the zooming motion was below the preset threshold.

5.2. Keyframe Extraction

Results of applying the proposed intelligent keyframe extraction approach to one action video sample are shown in Figure 6. With content analysis as described above, we were able to grasp the highlight of this 10-second long short video clip. Compared to keyframes evenly sampled from the video file, the keyframe set generated from our algorithm shows much more details of the people in action including closer looks of the people, their faces and so on. Moreover, there is no semantically meaningless keyframe from our approach (like the fourth keyframe in the evenly sampled set). This improved set of keyframes are more proper for making print outs of the video.



(a) keyframes evenly sampled from video clip



(b) keyframes obtained from proposed approach

Figure 6. Keyframe extraction using proposed intelligent approach vs. evenly sampling video clip.

5.3. Video Panorama Generation

The following figures show the results of panorama after intent determination. They illustrate the advantages of using our approach. In particular, our stitching algorithm can handle moving objects very well. Our exposure correction and blending algorithm also smoothes out the differences between video frames and make the final stitching result seamless.

5.4. High Resolution Frame Generation

Examples of results from the proposed super-resolution enhancement are shown in Figure 9. Note that after the processing, both compression artifact and image noise are significantly reduced, previously illegible letters become clearly readable, and more details are revealed in the human face.

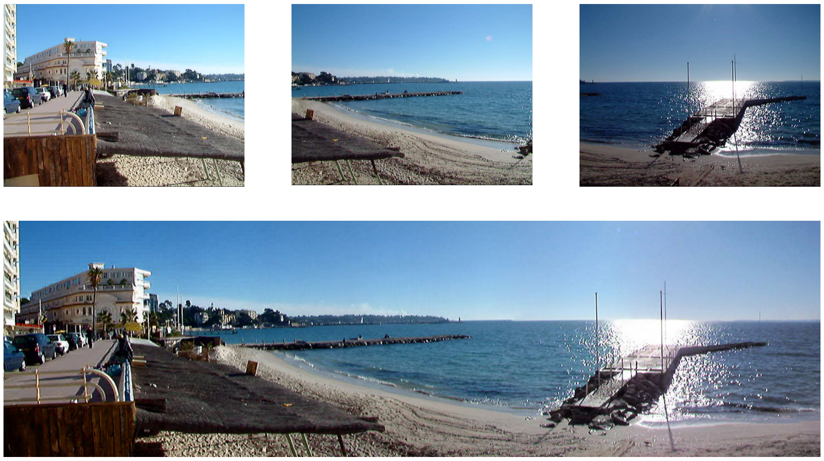


Figure 7. Top row shows samples of selected video frames used for actual stitching. Significant intensity differences can be seen between the frames due to camera exposure changes . The panorama result shows the corrected intensity and smooth blending.

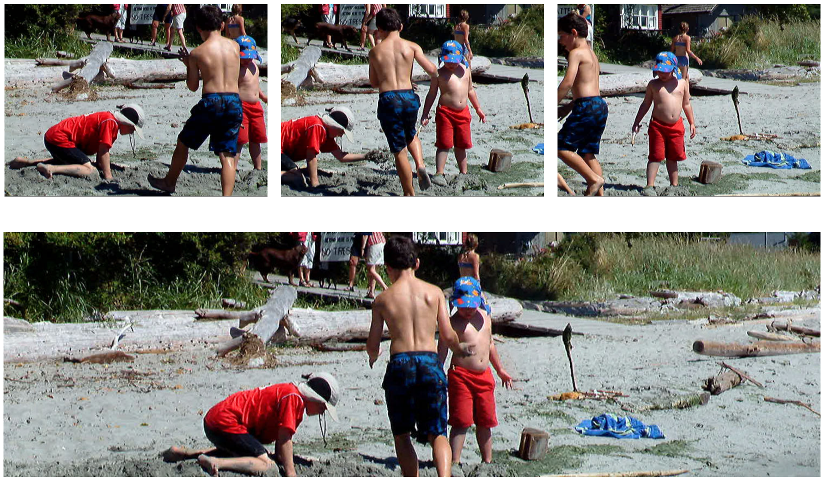
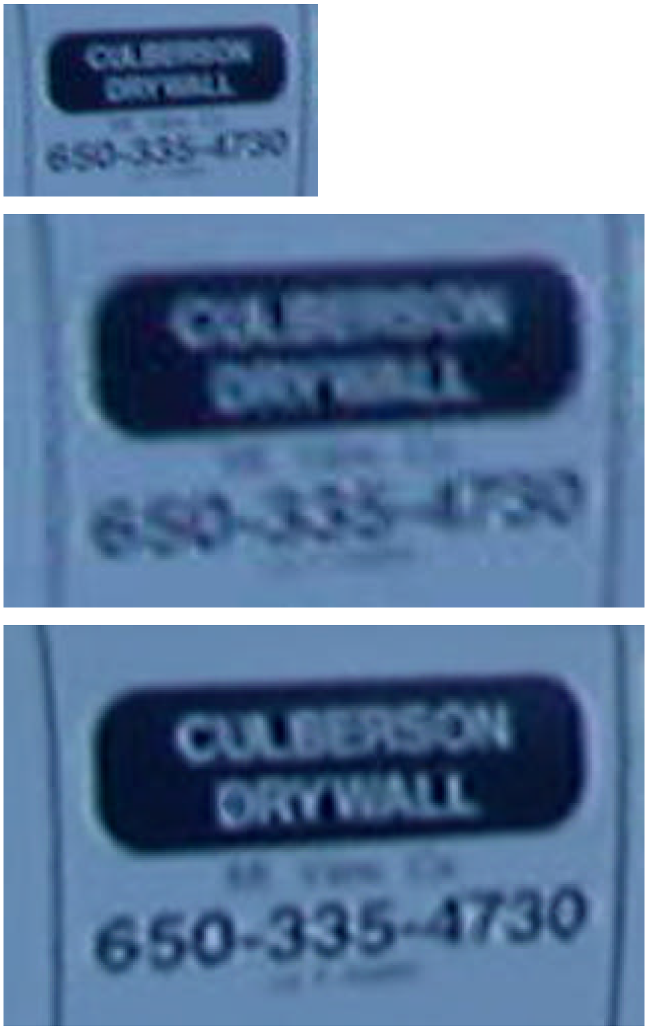


Figure 8. Top row shows samples of selected video frames used for actual stitching. Objects are moving across the scene. The panorama result avoids cutting through moving objects and stitches seamlessly.

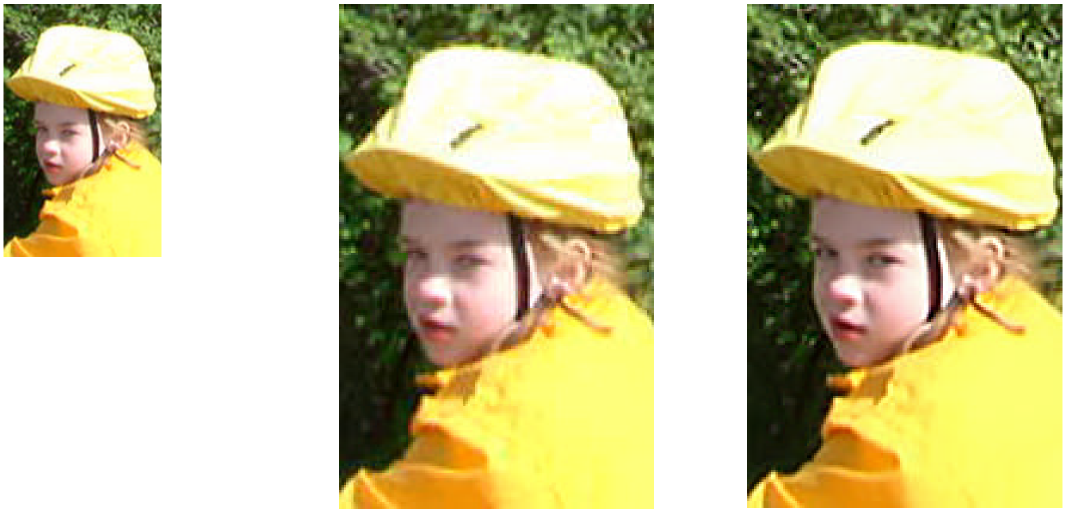


top: original video frame

middle: bi-cubic interpolation to 2x;

bottom: super-resolution enhancement to 2x

Figure 9. Comparison between original video frames
and results after super-resolution enhancement.



left: original video frame

middle: bi-cubic interpolation to 2x;

right: super-resolution enhancement to 2x

Figure 9. Comparison between original video frames and results after super-resolution enhancement (continued).

6. Conclusions

We have presented a framework and methods for mining video to photos to provide a complementary user experience. The proposed approach automatically analyzes the video through motion identification, clustering, and intent determination. Our current system processes three different intents: capturing of actions, of the environment, and of objects. Based on these intents, the system outputs the best image representations for these intents accordingly, including keyframes, panoramas, and high-resolution photos. In the future, we will extend the framework to include more output representations.

7. References

[1] N. Dimitrova, T. McGee and H. Elenbaas, "Video keyframe extraction and filtering: a keyframe is not a keyframe to everyone," Proceedings of the Sixth International Conference on Information and Knowledge Management. CIKM'97 : 113-20, 1997.

[2] F. Dirfaux, “Key frame selection to represent a video,” Proc. International Conference on Image Processing, (vol.2) 275-8, Vancouver, BC, Canada, 10-13 Sept. 2000.

[3] S. Uchihashi, J. Foote, A. Girgensohn, *et al.*, “Video manga: generating semantically meaningful video summaries,” Proc. ACM Multimedia 99 : 383-92, Orlando, Oct. 1999.

[4] Y. Deng and T. Zhang, “Generating Panorama Photos,” Proc. of SPIE Internet Multimedia Management Systems IV, vol. 5242, ITCOM, Orlando, Sept. 2003.

[5] H.-Y. Shum and R. Szeliski, “Construction and refinement of panoramic mosaics with global and local alignment,” ICCV, pp. 953-58, 1998.

[6] S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet, “Mosaicing on adaptive manifolds,” PAMI, pp. 1141-54, October, 2000.

[7] H. Sawyhney and S. Ayer, “Compact representation of video through dominant and multiple motion estimation,” PAMI, pp. 814-830, August, 1997.

[8] D. Capel and A. Zisserman, “Automated mosaicing with super-resolution zoom,” Proceedings of CVPR, page 885-891, June, 1998.

[9] J. Bergen, P. Anandan, and K. Hanna, “Hierarchical model-based motion estimation”, ECCV, 1992.J. Park, N. Yagi, K. Enami, K. Aizawa, and M. Hatori, “Estimatino of camera parameters from image sequence for model based video coding”, IEEE Trans. On Circuit and System for Video Technology. Vol. 4, No. 3, June 1994.

[10] R. L. Rardin, *Optimization in Operations Research*, Prentice Hall, 1998, ISBN: 0023984155.

[11] Y. Taniguchi, A. Akutsu, and Y. Tonomura, “PanoramaExcertpts: extracting and packing panoramas for video browsing”, Proc ACM Multimedia 97, pp. 427-436.

[12] H. Shekarforoush and R. Challapa, “adaptive super-resolution for Predator video sequences”, DARPA98, pp. 995-1002.

[13] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, “ Fast and robust super-resolution”, ICIP03, pp. 291-294, 2003.

[14] M. Chen, "Enhancing Image resolution", U.S. Patent Application, HP Docket No. 200310075.