The 5th Asian Conference on Computer Vision, 23--25 January 2002, Melbourne, Australia

Construction of Image Mosaics with Video Texture*

Wei Du, Hua Li
Lab of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, P.R.China, 100080
duwei@ict.ac.cn, lihua@ict.ac.cn

Abstract

Image mosaics are useful for a variety of tasks in applications of vision and computer graphics. One obvious disadvantage of image mosaics is that it cannot represent dynamic scene. Recently, a new type of medium named Video Texture is proposed to replace static photography with video clip. In this paper, we present a novel system to combine image mosaics with video texture. The system first composites a sequence of images into a final mosaic; then, some periodic/random moving objects in scene are captured into video clips, the system converts these video clips into video texture; at last, video texture is registered with image mosaics and both are combined into a compact representation. This system infuses image mosaics with dynamic qualities, while keeping the advantages of providing full view of scene (360 degree).

1. Introduction

Image mosaics are useful for a variety of tasks in applications of vision and computer graphics, such as virtual environment, video compression, video index, and scene stabilization. To construct a typical panoramic image, a user takes a hand-held camcorder, pans around a scene capturing the region of interest, and registers video frames to form a complete description of the surrounding environment [1][2]. However, image mosaics has an obvious drawback, it cannot represent dynamic scene—that is, every object in image mosaics is static, which gives users a false impression. In practice, two issues must be considered when constructing “dynamic” image mosaics.

First, image registration can be biased by moving objects. For example, correlation-based registration algorithms [3][4] that minimize the pixel intensity differences fail because moving regions contribute significant false residual to the minimization. Feature-based registration algorithms [5] also produce unreliable results because features may arise on the boundary of moving objects. Davis [6] extended Mellin transformation to register images, which remains unbiased by movement. However, his method can’t deal with large movement. New methods should be proposed to solve this problem.

Second, an effective representation of dynamic scenes is required. Video is an obvious alternative to static photography, but video is a very specific embodiment of a very specific period of time, we are forced to use a video clip periodically. Recently, Schodl [7] proposed a new type of medium named Video Texture. This new medium can provide a continuous infinitely varying stream of images. While its individual frames may be repeated from time to time, the video sequence as whole is never repeated exactly. Although video texture can be used to infuse a static image with dynamic qualities and explicit action, its field of view is as narrow as common image.

In this paper, we present a novel system to construct “dynamic” image mosaics. The system first composites a sequence of images into a final mosaic, a dominant motion estimation algorithm is used to register images, and outlier mask is detected to avoid bias; then, some periodic/random moving objects in scene are captured into video clips, the system converts these video clips into video texture; at last, video texture is registered with image mosaics and both are combined into a compact representation. The system infuses image mosaics with dynamic qualities, while keeping the advantages of providing full view of scene. For example, imagine a scene with a flag. Using our system, you can not only construct a panoramic image mosaic to look around, but also find the flag in the mosaic is flapping in breeze.

The remainder of this paper describes, in more detail, the method of constructing image mosaics in the presence of moving objects, in section 2, the method of creating video texture from common video clips, in section 3. The combination of these two mediums is proposed in section 4.

2. Image mosaicing with moving objects

Mosaics creation requires that images be registered with respect to one another. The problem of image registration can be regarded as the estimation of motion parameters between images. Sawhney [8][9] proposed a

* Supported by National 973 Project of China (No. G1998030608) and Youth Foundation of Institute of Computing Technology.

dominant motion analysis technology to automatically separate moving objects and significant components of the scene. Our registration method is similar to his.

2.1. Dominant motion estimation

Given two images, their motion transformation is recovered by minimizing the pixel intensity error between them,

$$\min E(M) = \sum_i [I_1(p_i - u(p_i;M)) - I_0(p_i)]^2 ,$$

where p_i is the vector of image coordinates, and $u(p_i;M)$ is the displacement vector at p_i described using a parametric vector M. Assuming the images taken from the same viewpoint but in different directions, the relationship between two overlapping images can be described by a planar perspective motion model

$$p' \sim M \cdot p = \begin{bmatrix} m_0 & m_1 & m_2 \\ m_3 & m_4 & m_5 \\ m_6 & m_7 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix},$$

where p' and p are homogeneous coordinates. To recover the parameter M, we iteratively update the transformation matrix using an incremental form

$$M \leftarrow (I + \delta M) \cdot M .$$

Then, the optimization problem can be solved by standard M-estimation, as described in [8][10]. In order to handle large initial displacements, the images are represented at multiple scales using Gaussian and Laplacian pyramids, and optimization is done at each scale [11].

With this method, a planar panorama can be created by mosaicing together numerous frames. Figure 1 illustrates a planar mosaic constructed by 3 consecutive images. Although a walking man is recorded in the video sequence, the registration result is still perfect. Building full view panorama is the similar process. For example, to construct a cylindrical panorama, we first warp each perspective image into cylindrical coordinates, and then recover only a 2-parameter translation model between neighboring images, shown in figure 2.

2.2. Composition

After registration, the images are warped and composited into a complete mosaic. To reduce discontinuities in intensity and color, typical systems for building mosaics blend the overlapping sections by a weighting function, which produces blurred results in the case of moving objects, shown in figure 3. To solve this problem, we employ Davis's idea [6] that segments the overlapping section into small regions and samples each pixel in these regions from a single source image.

Consider the overlapping section in a pair of registered images. A method for dividing this section into two regions so that no discontinuities occur along the boundary is desirable. The relative difference image, $abs(I_0 - I_1) / \max(I_0 - I_1)$, provides a measure of similarity between the source pair on each pixel. A dividing boundary falling along a path of low intensity in the difference image will produce minimum discontinuity between regions in the final mosaic. Davis [6] uses Dijkstra's algorithm to detect the best path dividing the overlapping section.

Although this method avoids producing blurred images of moving objects, it suffers large time consuming especially when image resolution is high. We improve Davis's method by decomposing the original difference image into a dynamic quad-tree, each node of which is the mean difference in corresponding area. Furthermore, paths through low intensity nodes with larger area are preferred because such nodes suggest these areas are well registered. Thus, we update the cost function of each node by

$$\cos t(x, y) = w_0 * diff(x, y) + \frac{w_1}{area(x, y)},$$

where $diff(x, y)$ is the mean difference of node (x,y) and $area(x, y)$ is the pixel number in node (x,y). Applying Dijkstra's algorithm, our method can detect the best path of lowest cost much faster. Shown in figure 4, the best path detected only pass through 6 nodes. The path should be refined in each node. Figure 5 gives the final composited mosaic.

3. Video texture

Dynamic scene cannot be adequately represented by a static panorama. The best alternative is video. But video has its own drawbacks. If video is played in loop mode, there is an obvious discontinuity between the last frame and the first frame. Since video captures the time-varying behavior and shows strong structure, it is possible to generate a similar looking video of arbitrary length. Video texture is such type of medium, which has the qualities somewhere between those of a photograph and a video. In this section, we introduce the method of generating video texture from original video clips.

3.1. Synthesis video texture

The general approach of synthesizing video texture is to find places in the original video where a transition can be made to some other place in the video clip without introducing noticeable discontinuities.

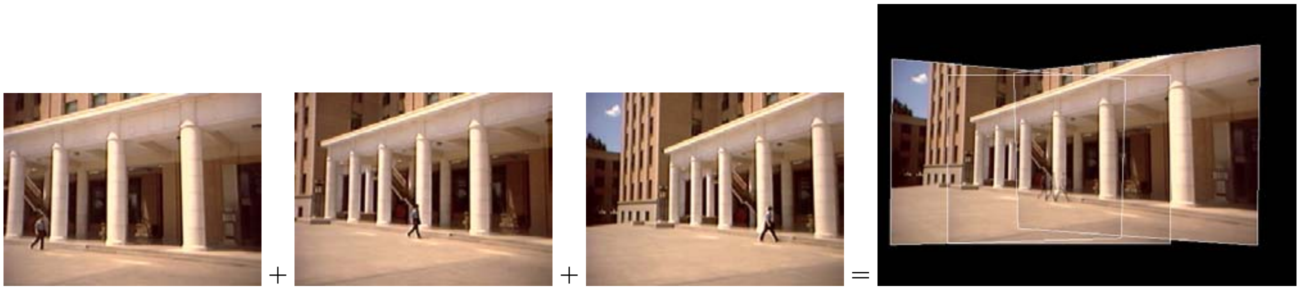


Figure 1. Construction of a planar mosaic from three images. Note that although a walking man is captured in the sequence, the moving region of the walking man doesn't bias the result.

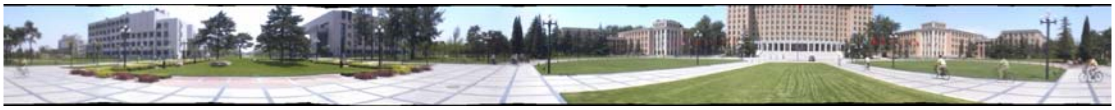


Figure 2. Construction of a cylindrical panorama.

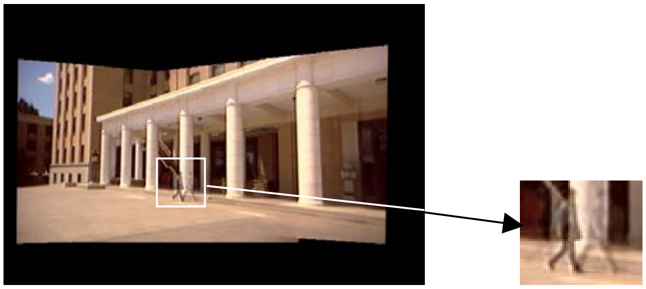


Figure 3. The mosaic composited with a feathering algorithm. Overlapping sections are blended with a weighting function, producing blurred result in regions of the walking man (in white rectangle).

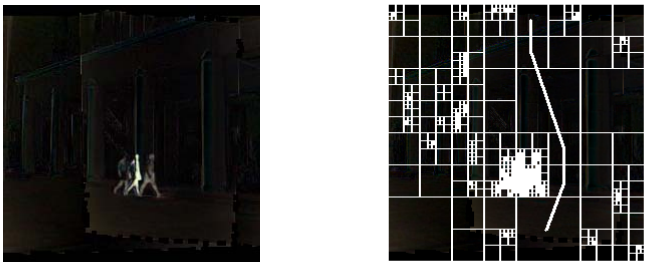


Figure 4. Best path detection in quad-tree difference image. Left image is the relative difference image of a pair of registered images, only regions of moving objects have large intensity. Right image is the result of quad-tree decomposition, white thick line segments shows the best path detected.

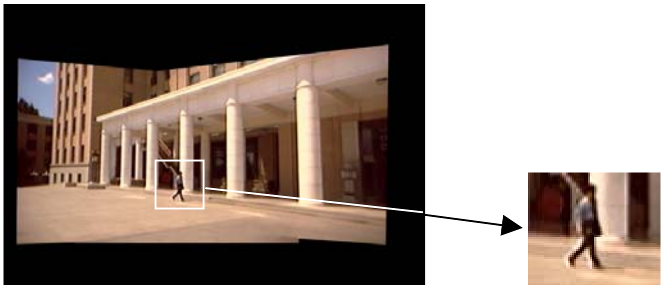


Figure 5. The final composited mosaic composited with our method.

The first step is to compute the frame-to-frame distances, and store them in matrix $D_{ij} = \left\| I_i - I_j \right\|_2$, where D_{ij} is the L2 distance between frame i and j. The transition can be made from i to j if the successor of i is similar to j—that is, whenever $D_{i+1,j}$ is small.

Such distance function can only preserve the similarity across frames; it can't keep the dynamics of motion. The problem can be solved by matching subsequences within a weighted window instead of individual frames. Define dynamic distance by filtering the difference matrix with binomial weights

$$D'_{ij} = \sum_{k=-m}^{m-1} w_k D_{i+1+k,j+k} .$$

To avoid dead end—the transition might lead to some portion of the video from which there is no exit, we try to predict the anticipated future cost of choosing a given transition. The anticipated future cost reflects the expected average cost of future transitions

$$D''_{ij} = (D'_{ij})^p + \alpha \cdot \min_k D''_{jk} .$$

Thus, the probabilities that transition can be made from i to j are defined as

$$P_{ij} \propto = \exp(-D''_{ij} / \sigma^2) ,$$

where σ is a small multiple of the average D''_{ij} and P_{ij} is normalized so that $\sum_j P_{ij} = 1$. Figure 6 gives two demos of creating video texture from common video clips. The first clock video is obtained from CD of Siggraph' 2000. The second flag video is captured by us with a camcorder. The four gray images followed are the distance image, the dynamic distance image, the future cost image and the probabilities image respectively.

Video texture can be played in two modes: random play and video loop. The first mode uses a Monte-Carlo technique to decide which frame should be played after a given frame according to the probabilities of the transitions. The second mode selects a small number of transitions with low average cost. This mode is necessary especially when video texture is played by a conventional video player.

3.2. Blending and morphing

Although this method makes transitions that introduce only small discontinuities in the motion, there are cases where no unnoticeable transitions are available in the sequence. To reduce such discontinuities, Schodl uses a cross-fading algorithm to blend the frames before and after the transition, instead of simply jumping from one frame to another. The algorithm computes a weighted average of all frames participating in a multi-way fade,

$$B(x,y) = \sum_i \alpha_i I_i(x,y) .$$

In my experience, there are still some cases that can't get satisfactory result with cross-fading algorithm. We prefer Beier's feature based metamorphing technology [12] to reduce discontinuities and align features, shown in figure 7.

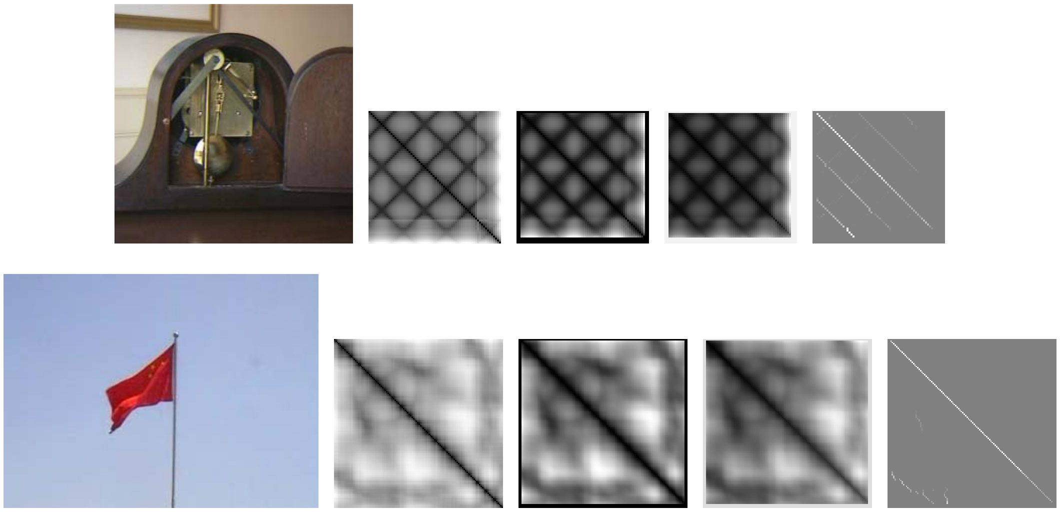
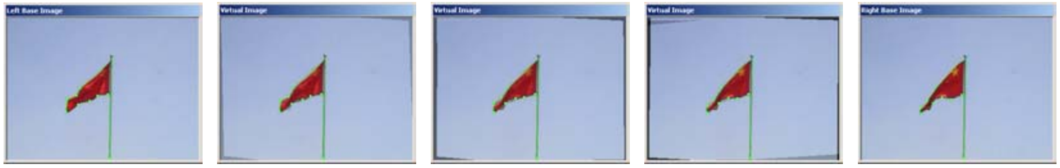


Figure 6. Two demos of creating video texture from common video clips. The four gray images followed are the distance image, the dynamic distance image, the future cost image and the probabilities image respectively. Note that clock video shows strong periodicity and flag video shows only randomness.



frame i 25% 50% 75% frame j

Figure 7. The morphing result between frame i and j.

4. Image mosaics with video texture

After image mosaic and video texture are created, we can combine them together to construct "dynamic" image panorama. First, video texture is registered with image mosaics by applying our motion estimation method proposed in section 3.1(the system requires video texture and image mosaic taken from the same viewpoint). The registration is very simple except that users capture mosaics and video texture with different focal length. In this case, we employ an fft-based registration method to detect the scale first, which is described in [13].

Then, the video texture is projected to mosaic according to the parameter computed. Since video texture shows strong similarity between frames, a background model can be detected with a time median filter. The difference between each frame and the background image is relative small, and can be compressed by entropy compression algorithm. In these different images, only pixels of moving objects have large values. With a suitable threshold, motion areas can be masked easily.

To eliminate the color discontinuities between video texture and image mosaic, static areas in video texture should be blended with a weighting function. In this way, video texture and image mosaic are stitched seamlessly into a dynamic panorama.

Rendering such a dynamic panorama is quite simple. Besides the parameter of view direction, time parameter is added to tell which frame of video texture should be played. If there is no motion area in view window, the image is rendered as common static panorama; otherwise, video texture is played with either random play mode or video loop mode. Figure 8 and 9 show the results of combining image mosaic and video texture.

5. Conclusions and future works

In this paper, we have presented our system for constructing dynamic panorama. The system automatically creates image mosaics in the presence of moving objects. Video clips of periodic/random moving objects are converted into video texture. The system stitches these two mediums together and infuses image mosaics with dynamic qualities, while keeping the advantages of providing full view of scene. Compared with previous works, two main contributions have been proposed. First, the combination of video texture and image mosaics is a very important idea. Second, a new composition method is proposed to blend mosaics of moving regions.

The system can be enhanced in the following way. First, A better registration algorithm is needed to construct perfect panorama. Dominant motion estimation method can only register images with small motion. For the case of images with large motion, the result of this method is biased. New method should be proposed to solve this problem. Second, the key to synthesis video texture is to compute the frame-to-frame distance, and the simplest L2 distance is adopted now. We hope to find a better distance function, improving the quality of video texture. Motion blur is also disgusting in registration and composition of image mosaics. An automatic deblurring method is required.

6. References

[1] Chen, S.E, Quicktime VR——an image-based approach to virtual environment navigation, In Computer Graphics (SIGGRAPH'95), pp.29~38

[2] Laura Teodosio, Michael Mills, Panoramic overviews for navigating real-world scenes, Proceedings of the First ACM International Conference on Multimedia '93, Anaheim, CA, pp.359-364

[3] L.G. Brown, A survey of image registration techniques, ACM Computing Surveys, 24:4, 1992

[4] R. Szeliski, and H.-Y. Shum, Creating full view panoramic image mosaics and texture-mapped models, in Computer Graphics (SIGGRAPH'97), pp.251-258

[5] D. Cpel, and A. Zisserman, Automated mosaicing with super-resolution zoom, CVPR'98, pp.885-891

[6] J. Davis, Mosaics of scenes with moving objects, ICCV'98, pp.354-360

[7] A. Schodl, and R. Szeliski, Video Textures, in Computer Graphics (SIGGRAPH'2000), pp.489-498

[8] H.S. Sawhney, and S. Ayer, Compact representations of videos through dominant and multiple motion estimation, IEEE trans. PAMI, vol.18, No.8, pp.814-830

[9] H.S. Sawhney, S. Hsu, and R. Kumar, Robust video mosaicing through topology inference and local to global alignment, ECCV'98, Freiburg, Germany, vol.2, pp.103-119, 1998

[10] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, Numerical recipes in C, second version, Cambridge university press, 1992

[11] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani, Hierarchical model-based motion estimation, ECCV'92, Santa Margherita Liguere, Italy, 1992, pp.237-252

[12] T. Beier, Feature-Based Image Metamorphosis, In Computer Graphics (SIGGRAPH'92), pp.35-42

[13] B.S. Reddy, and B.N. Chatterji, A FFT-based technique for translation, rotation, and scale-invariant image registration, IEEE trans. Image Processing, 5:8, 1996

[14] C. Morimoto, R. Chellappa, Fast 3D stabilization and mosaic construction, CVPR'97

[15] Yin Li, Weixin Kong, Hanqing LU and Songde MA, Mosaic representation of video shots based on edges of contrast by watershed, ACCV'2000

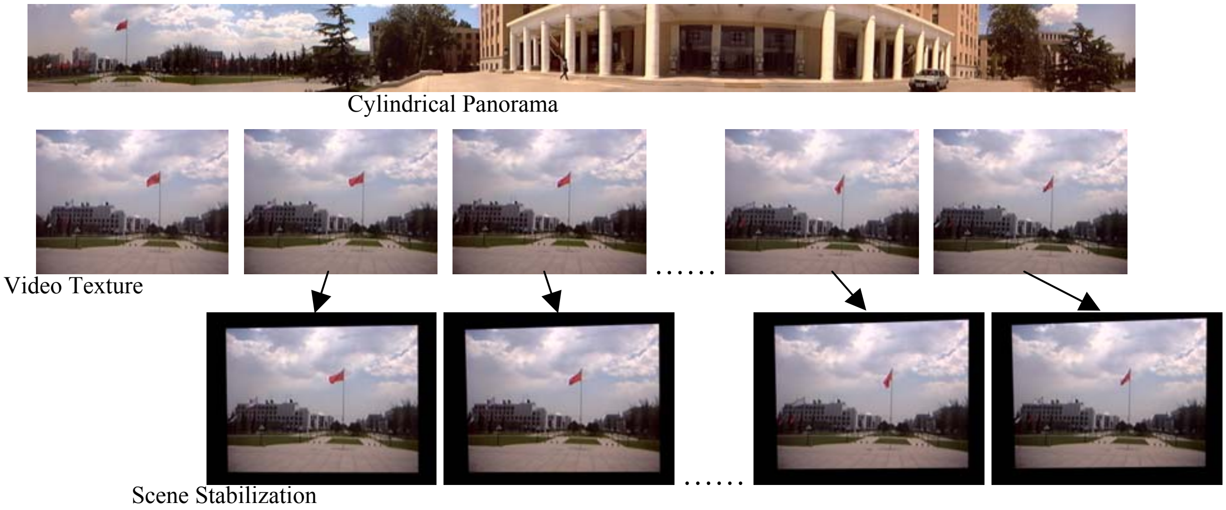


Figure 8. The first result of combining image mosaic and video texture. Since the original video clip is captured with a camcorder held in hand, there is small dithering between image frames. We register each frame of the video clip with our motion estimation method, which is also called scene stabilization [14].

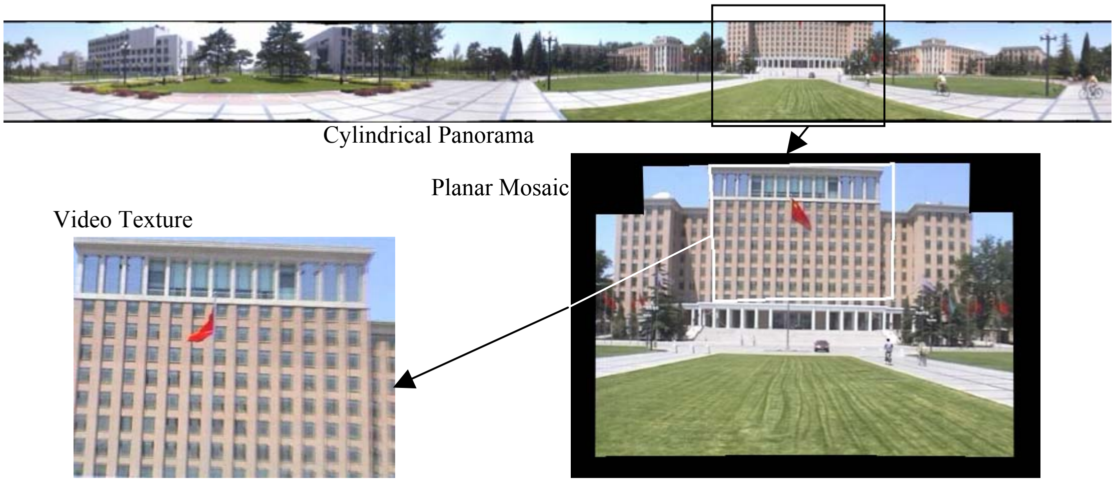


Figure 9. The second result of combining image mosaic and video texture. Note that we capture mosaic and video texture with different focal length (video texture is zoomed in), and an fft-based registration method is adopted to detect the scale first. The white quadrangle in the planar mosaic is registered with the video texture.