Ghost Error Elimination and Superimposition of Moving Objects in Video Mosaicing

Tomio Echigo†, Richard Radke‡, Peter Ramadge‡,
Hisashi Miyamori*, and Shun-ichi Iisaku*

†*IBM Research Tokyo Research Laboratory* echigo@jp.ibm.com,

‡*Princeton University Electric Engineering* {rjradke, ramadge}@ee.princeton.edu,

**Communications Res. Lab. Ministry of Posts & Tel.* {miya, iisaku}@crl.go.jp

Abstract

This paper presents a new approach for region-based video mosaicing, treating moving objects separately from the background, and with improved ghost-like noise elimination. The mosaic images show the moving objects superimposed over a stationary background. Conventional technologies can reduce the ghost-like noise that occurs from moving objects by using temporal median filtering, but its efficiency depends on the ratio between the speeds of the camera and the moving object. Our technology eliminates these noises more efficiently by using segmented images of a spatio-temporal video sequence. Segmentation is performed using a novel technique that uses different configurations of quad-trees for the initial separation in the split-and-merge process. The segmented images are also used to display tracked moving objects on the panoramic image.

1. Introduction

Video mosaicing is useful in a variety of tasks for applications envolving video, such as synthesizing, summarization, and compression. In conventional video mosaicing technologies[3][4][5][7][8], all frames from a video sequence are projected on an adaptive surface, and a panoramic image is created by determining pixel values from successive frames. In the process, pixels reminiscent of moving objects are also blended, resulting in the presence of a ghost-like noise in the panoramic image. In order to solve this problem, temporal median filters have been used and claimed to be efficient. However, they actually depend on the speed difference between the camera and the moving object. Therefore, slowly moving objects can not be completely eliminated.

In our method, firstly we segment regions that consist of moving objects and a stationary background, tracking the moving objects in a spatio-temporal buffer over multiple frames. Secondly, motion parameters of the camera are estimated from motion vectors of feature points in the background region. These feature points are obtained from correspondence between frames. Background regions of the video sequence are then projected on the most suitable reference frame, but as the regions of moving objects are cut out from the images, values of pixels in the regions occluded by moving objects can not be defined at the moment. They are obtained from the pixels in the background regions that are disclosed after the objects move to other locations. Since the pixels of moving objects are not taken in account, the false pixel values that yield the ghost-like noise in the panoramic image are eliminated. Finally, our system augments visualization by showing not only the panoramic image with a realistic wide-angle view, but also the images of tracked objects (moving objects selected by the user) superimposed in a stroboscopic fashion.

2. Failure Case of a Temporal Median Filter

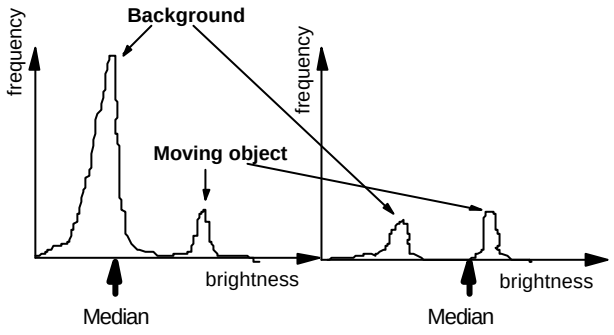


Figure 1: Histgram of temporary projection on each pixel

In the creation of a panoramic image, the stationary background is supposed to be obtained from pixels in the video sequence corresponding to a same location in the panoramic image, sampled over a period of time long enough for moving objects to pass. In this case, as shown in Fig. 1 (a), a temporal median filter is effective in eliminating fast moving objects. However, when an object moves slowly, as shown in Fig. 1 (b), the camera during few frames captures the region of the background occluded by the object. The period of capturing background is not much longer than one of capturing an object on a particular pixel. In this case the temporal median filter fails to output the correct value for the background pixel, so that a ghost of the moving object appears in the final panoramic image. In order to eliminate this ghost, each pixel should be manipulated as being connected to adjacent pixels. It requires region segmentation, as explained next.

3. Spatio-Temporal Region Segmentation

In order to solve the problem of segmenting and tracking an object, many conventional approaches use snakes and active tubes techniques of dynamic contour detection. They require a correct initialization, so that objects' initial contours include the objects themselves. However, it is not convenient to tie the initial contours to a particular frame of the video sequence.

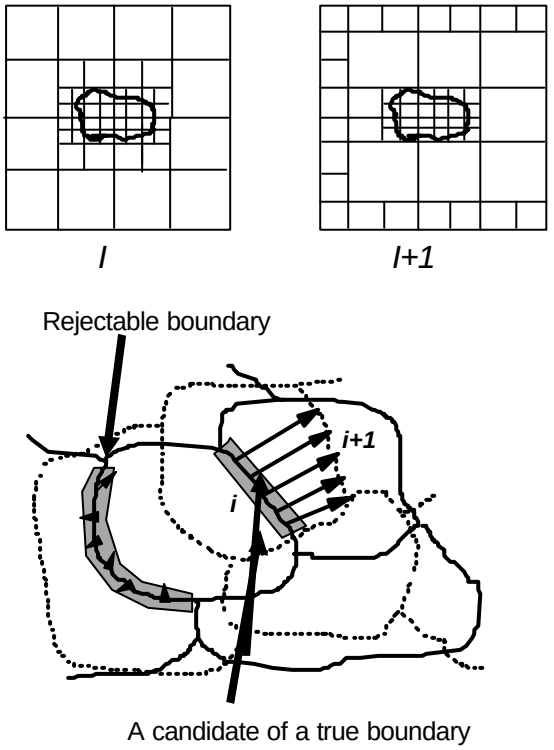


Figure 2. Different configurations of quad-trees between successive frames

The proposed technique segments regions and tracks them as predefined objects in a spatio-temporal buffer, where multiple frames are stored without initial contours.

First, all images in the buffer are spatially segmented by the split-and-merge method, which separates and merges regions on the basis of colors in the structure of a quad-tree. The shapes of the initial regions obtained by the split-and-merge method depend on the configurations of the quad-tree. Our approach uses different configurations of quad-trees for each frame, that is, splitting home positions are different in successive frames. These images have different approaches of region growing, so that initial shapes of regions overlapped between successive frames are different from each other, as shown in Fig. 2. When the camera does not move and all objects are stationary, all pixel value differences between successive frames are nearly equal zero. In this case, the results of segmentation should depend on the spatial features. As spatial features, we use not only colors, but also the parameters of Gaussian Markov Random Field (GMRF) model[2][5], which can express texture features and is effective for obtaining the exact boundaries of texture regions. We use the four neighbors model of GMRF and the intensity plane of the image for texture segmentation. Assuming Gaussian distribution of potential of image *Xij*, GMRF is expressed as follows:

$$P(X_{ij} / \boldsymbol{e}_{mn}) = \frac{K}{\sqrt{|\Sigma|}} \exp(-U(i, j)),$$

where K is constant value for normalization and Σ is squared mean value of the region. The spatial correlation at the element (i, j) expressed as follows:

$$u(i, j) = \{I(i, j) - \bar{m}\} - \sum_{(m,n)\in e} \boldsymbol{a}(m, n)\{I(i+m, j+n) - \bar{m}\}.$$

The parameters of GMRF are determined from the coding method by Besag[1] and the maximum likelihood estimation (MLE). The initial regions are separated according to their colors and the parameters of GMRF in the entire region; however, it is computationally expensive to segment and track entire objects from their spatial features, because the parameters of the GMRF model are obtained from MLE. When a region is not perfectly overlapped between successive frames and the gradient along the boundary is too small, the part of the boundary that separates regions can be reduced, and thus the regions can be merged at a lower computational cost than by using the GMRF parameters. Additionally, regions merged by only

spatial features are not robust with respect to exact tracking over successive frames, because the gradient given by texture features often splits initial regions in overly small sizes, and motion is blurred by erosion of the shapes of the objects. Therefore, we use motion displacement along the boundaries calculated from spatio-temporal intensity gradients. When motion displacements are the same along a portion of the boundary that separates regions, that portion of the boundary can be taken as a part of the true contour of an object. Otherwise- that is, when motion displacements consist of disparate values along a candidate boundary- the candidate should be considered a false contour and be rejected. A dominant motion of the camera movement is determined from corresponding features in the whole image between successive frames, as described in the next section. It can be judged from differentiation between successive frames shifted by a dominant motion whether a candidate boundary is rejected or not. All boundaries should be verified in the buffer of multiple frames not only between successive frames, but also between every third and every sixth frame. Regions inside the true contour can then be tracked as parts of the same object. To take account of occlusion, motion displacement of a region inside the contour is determined from its maximum value along the true contour.

In order to segment a player's region, adjacent regions that have close motion displacements should be merged into a single object, even if their colors and texture features are different, since the region may be a part of a face, an arm, a shirt, a pair of shorts, or a leg. We employ typical color map of players' regions of both teams that are cut out from a frame. Hence, the proposed algorithm is effective for segmenting and tracking predefined objects.

4. Recovery of Camera Motion

We employ a pin-hole camera model that is generally used to model the camera perspective projection. The perspective model gives the exact representation to account for all the possible camera motions, compared with the other approximated models; the parallel transformation and the affine one (including para-perspective and weak perspective transformation), although its parameters are mathematically and computationally too hard to estimate. In our application, the camera is fixed on a tripod, giving images which have small translational displacements because the rotation axes does not coincide with the optical center of the camera (considering that the camera performs only rotational movements). However, the depth of the field of view is at least 40 meters far from the camera, so that the translation is much smaller than the scene depth. Therefore, our approach assumes a zero translation model between successive images of the video sequence.

Selecting image features in an image whose corresponding locations in order images can be precisely measured is an important problem for estimating the exact parameters. In our approach, we employ Tan's method[9] that selects block features that have rich enough intensity textures and consistent inter-frame motions and finds correspondences between images. With this method, a quantitative measure can be obtained to select good motion features from images in the sense of the maximum likelihood estimation for estimating motion parameters about the multiple motion models of the rotation and scaling factors. In order to the estimation stable when a few data points are wrong, a robust estimation is also needed. We employ the M-estimation for robustness, which is applied with the Geman-McLure function. Since M-estimation requires an initial estimate, firstly we use the initial estimate determined from the least squares method. Although the least squares method for estimating parameters has non-linear equation, we can assume the rotational transformation should be small between successive frames, so that the initial estimate can be determined from linear equations. By using the initial estimate, the typical value of the standard residuals is defined as the median of the absolute residuals . The revised parameters are calculated by using the adjusted effective weight iteratively. When the residual becomes smaller than the threshold, the estimation process is finished.

5. Representation of Regions

In our approach, all pixels in the video sequence are classified as regions. We use the video of the soccer game as the target contents. The camera was set on the stand at the distance of 80 m from the center of the field, whose field of the view was covered over both sidelines mostly. The focal length of the camera was not made change suddenly. Almost players moved from the right to the left or from the left to the right. The lawn area on the field and the stand were defined as the background. In the segmentation process, the only object in the field was extracted from the background. The objects except for the background were defined as a ball and the players. The stationary lines, that is, side lines, goal lines, and penalty lines, and the goal post were merged in the background.

The region is described as the minimal bordered rectangle that has a binary bitmap presented the inside region as "1" and the outside one as "0". All regions have the bitmaps. We applied the maximum region that is defined as the background for generating the planar panoramic image.

6. Rendering Results

In our examinations, the dominant motions of image

features were generated mainly from panning around the Y axis in the camera coordinate systems. The secondary important parameter was tilting around the X axis, but it was much smaller than the panning. There was no rolling around the Z axis.

We defined the projection surface as planar. The most suitable projective plane is defined by calculating the span of the camera panning movement and choosing the frame nearest the central position. After the projective transform, there are cases when a pixel does not have a correspondent in the original frame, thus bilinear interpolation is used to set its value. A more vivid image could be obtained by processing the pixels as they are read from the original sequence of frames, so that the most recently inputted pixels are given priority over the ones already read. It is done by taking the mean value between the average of the pixels already read and the newly inputted one.

The interlaced video yields a large displacement between the odd and the even fields. In our approach, while video mosaicing of the stationary background is generated from both fields of all images, moving objects are drawn from newly inputted images of odd fields only, ignoring objects in the even fields. The image of a moving object copied from the odd field is drawn on the even field. The video has D1 quality, with 704 by 480 pixels of resolution, 30 fps. interlaced, and 4:2:2 YUV colors. Fig. 3 shows the background panoramic image created from eight frames sampled from a sequence of 31 frames. The image plane of the 15th field was selected as the most suitable projective image. In this sequence, the camera moved with rapid panning, so that the temporal median filter left a moving object partially like as ghost-like noise, as shown in Fig. 3 (b). The detail results of the effectiveness of the temporal median filter on particular pixels are shown in Fig. 5. In Fig. 5 (b), the lower peak was yielded by a moving object and the value remained after temporal median filtering.

(a) A result by using segmented regions

(b) A result by the temporal median filter

Figure 3. Comparison of results by using segmented region and temporal median filter

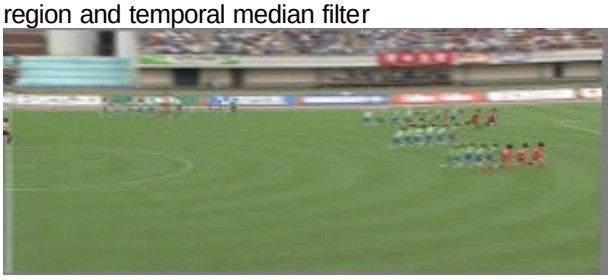
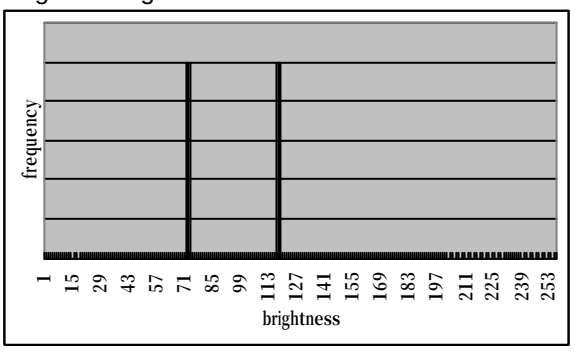


Figure 4. Mosaiced background and superimposed objects

(a) Frequency of captured images on a pixel of the background region



(b) Frequency of captured images on a pixel of the region including a moving object

Figure 5. Examples of the effective temporal median

filter and occurrence of ghost-like noise

Figure 6. Selected objects superimposed on mosaiced background

Figure 7. Mosaiced background and superimposed moving objects

Our approach output the result of mosaiced background by using segmented regions and superimposed moving objects on it, as shown in Fig. 3 (a) and Fig. 4, respectively.

Since all objects are independent, we can select the objects to be tracked or watch the animation of those selected moving objects. Fig. 6 shows the stroboscopic painting of the selected three players by three frames on the background panoramic image.

Fig. 7 presents the realistic efficiency of the panoramic image in the wide angle from a long sequence of 141 frames.

7. Conclusions

We developed a new technique to generate a planar projective panoramic image from segmented regions. The background of the panoramic image was created from the largest regions segmented by spatio-temporal constraint, giving a realistic visualization from a wide-angle panoramic view. The moving objects were drawn independently of other regions, using newly inputted images. The advantage of our method is the elimination of false pixels that occur from moving objects in the planar projective panoramic image, because the panoramic image is made solely from the stationary background. In addition, another advantage is the ability to display tracked objects of or an animation of those moving object selected by the user, because all previously segmented objects can be handled independently.

References

[1] J. E. Besag, "Spacial Interaction and the Statistical Analysis of Lattice System," J. Roy. Statis. Soc., B36, pp. 192-236, 1974.

[2] T. Echigo and S. Iisaku, "Unsupervised Segmentation of Colored Texture Images by Using Multiple GMRF Models and Hypothesis of Merging Primitives," Trans. IEICE D-II, vol. J81-D-II, no. 4, pp. 660-670, 1998.

[3] M. Irani, P. Anandan, and S. Hsu, ``Mosaic Based Representations of Video Sequences," ICCV, pp. 605-611, 1995.

[4] S. Mann and R. W. Picard, ``Video Orbits of the Projective Group: A New Perspective on Image Mosaicing," IEEE Trans. Image Processing, vol. 7, 1997.

[5] D. K. Panjwani and G. Healey, "Markov Random Field Models for Unsupervised Segmentation of textured Color Images," IEEE Trans. PAMI, vol. 17, no. 10, pp. 939-954, 1995.

[6] H. S. Sawhney and S. Ayer, ``Compact Representations of Videos Through Dominant and Multiple Motion Estimation," IEEE PAMI, vol. 18, no. 8, pp. 814-830, 1996.

[7] R. Szeliski, ``Image Mosaicing for Tele-Reality Application," Technical Report RL 94/2, Digital Equipment Corp., 1994.

[8] R. Szeliski and H. Y. Shum, ``Creating Full View Panoramic Image Mosaics and Environment Maps," SIGGARAPH, 1997.

[9] Y. P. Tan, ``Digital Video Analysis and Manipulation," PhD. Thesis, Princeton University, 1997.