Dynamic point selection in image mosaicking

Huibao Lin
Jennie Si
Arizona State University
Department of Electrical Engineering
Tempe, Arizona 85287

Glen P. Abousleman
General Dynamics C4 Systems
8201 East McDowell Road
Scottsdale, Arizona 85257

Abstract. Image mosaicking is a procedure of integrating information from a series of images to create a comprehensive view of the scene. It is typically carried out by selecting a subset of pixels from each of the individual images, matching these selected pixels from different images, and then mapping all the images onto a common image grid. The number of selected pixels is a critical parameter that affects both computational complexity and mosaicking accuracy. An image mosaicking algorithm is developed by using a novel dynamic point selection concept. The algorithm automatically determines the number of pixels to select according to the similarity of the images. Simulations show that the proposed algorithm generates mosaic accurately and efficiently.
[DOI: 10.1117/1.2180794]

Subject terms: images; mosaics; matching; correspondence; registration; dynamic.

Paper 050848LR received Nov. 2, 2005; revised manuscript received Dec. 26, 2005; accepted for publication Jan. 3, 2006; published online Mar. 13, 2006.

An image mosaic combines a series of images or video frames to form a comprehensive view of the scene. It has wide applications in a variety of areas such as video compression,[1] global environment understanding,[2] video editing and indexing,[3] and panoramic image generation.[4]

In order to combine a series of images, correspondence among images has to be established. This can be achieved by making use of common areas in the images, or the relationship between two images, which is usually represented by a mathematical transform. Due to the overwhelmingly large number of image pixels, it is prohibiting to find correspondence between images by doing a pixel by pixel search. Instead, only those pixels that convey critical information about the images are chosen, and they are referred to as interest points. The geometric and optical properties of these interest points are then evaluated to form the so-called local descriptors. The name comes from the fact that local descriptors are calculated from a neighborhood of individual interest points. Interest points from different images can then be matched by comparing their local descriptors. By using the matching points, correspondence between images can then be established. Finally, the images are mapped onto a common grid to form a comprehensive view.

0091-3286/2006/$22.00 © 2006 SPIE

The selection of interest points is critical for mosaicking. A significant amount of research has gone into developing criteria for choosing interest points. For example, some criteria are based upon target corners or junction of edges, to name a few.[5] The interest points are chosen from those pixels with the largest distinguishing value according to their respective criteria. However, little has been done to determine how many interest points should be chosen for an image series. This is not considered an issue if one takes into account the fact that the more interest points are selected, the better the matching. However, the increase in computational complexity may prohibit the image mosaicking system from realistic applications. To illustrate, let the number of interest points be n. Then there are n^2 pairs of points to be matched given two images. Thus the computation time is quadratic to the number of interest points.

In this paper, we propose a dynamic point selection procedure to automatically choose the number of points according to similarities among images. Specifically, fewer interest points are chosen if the images are similar to each other; while more are chosen if the images are more disparate. Simulations show that by this mechanism, computation time is significantly reduced without compromising mosaicking accuracy.

Figure 1 shows a novel automatic, dynamic point selection procedure for image mosaicking.

First of all, the saliency map for each image frame is calculated. The saliency information of a pixel corresponds to its likelihood of being a corner point, and is calculated by the algorithm developed by Harris and Stephens.[6] Once the salience for every pixel is calculated, those pixels with locally maximal saliency values are detected and sorted. From the sorted list, a number (N_1) of the most salient points are chosen as interest points.

Local descriptors are calculated from a neighborhood of the interest points. Among a variety of local descriptors, scale-invariant feature transform (SIFT)[7] has been shown to be robust[8] and thus used in this paper. Refer to Ref. 7 for implementation details. The descriptors are then normalized to eliminate the influence of luminance change. The

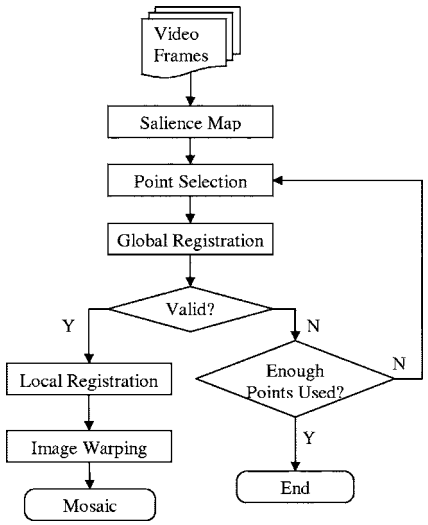

Fig. 1 Procedure for image mosaicking with dynamic point selection.

(a) (b) (c)

Fig. 2 Testing images 1, 2, and 3. The interest points are superimposed on the images.

similarity of two interest points is determined by the inner product of their descriptors: the larger the inner product, the stronger the similarity.

The interest points are used to establish the correspondence between two images. According to the fundamental principle of camera geometry, if camera lens distortion and target occlusion are not considered, two images, $\mathbf{I}_1$ and $\mathbf{I}_2$, generated for the same target are related by the following projective transform:

$$x_t = \frac{a_{11}x_1 + a_{12}y_1 + a_{13}}{a_{31}x_1 + a_{32}y_1 + 1},$$

$$y_t = \frac{a_{21}x_1 + a_{22}y_1 + a_{23}}{a_{31}x_1 + a_{32}y_1 + 1}, \quad (1)$$

where (x_1, y_1) and (x_t, y_t) are the coordinates of the same target point in $\mathbf{I}_1$ and $\mathbf{I}_2$, respectively.

Note that due to quantization error in digital images, Eq. (1) cannot be satisfied strictly for two points corresponding to the same target point. Hence, correspondence between two interest points is defined as follows: if there exists an interest point at (x_2, y_2) in $\mathbf{I}_2$ such that the distance between (x_t, y_t) and (x_2, y_2) is no greater than $\sqrt{2}$ pixels, then (x_1, y_1) and (x_2, y_2) correspond to each other. The choice of $\sqrt{2}$ pixels as the distance threshold is because $\sqrt{2}$ is the maximum distance two connecting pixels can have, if 8-connectness is considered.

Once the similarity between every interest point in $\mathbf{I}_1$ and that in $\mathbf{I}_2$ is calculated, a number of N_2 ($N_2 \leqslant N_1$) pairs of the most similar interest points are found. Based on these N_2 pairs of interest points, the projective transform coefficients $[a_{ij}]$ are obtained.

To register a pair of images, a certain number of corresponding interest points have to be found. For example, at least 4 pairs of corresponding interest points are necessary if projective transform as shown in Eq. (1) is used. To ensure a sufficient number of corresponding points generated, more interest points have to be found when the images are more disparate. On the other hand, fewer interest points are preferred for computation efficiency.

The proposed dynamic point selection procedure automatically increases the number of interest point, N_1, when the selected interest points do not have enough corresponding ones. This is carried out by applying the estimated projective model on all the N_1 interest points. If $N_3 (N_2 \leqslant N_3 \leqslant N_1)$ pairs of corresponding points are found, and the descriptors for these interest points are similar, then the projective model $[a_{ij}]$ is validated. Otherwise, N_1 is increased and the aforementioned procedure is reiterated.

On the contrary, if the selected interest points consist of more than enough corresponding ones, fewer interest points are chosen for the next pair of images. To be exact, if the previous N_4 or more pairs of images have been successfully registered without increasing N_1, then N_1 is reduced.

Typical settings for N_2, N_3, and N_4 are 6, 12, and 4, respectively. In the experiments, N_1 is initialized to 50, and N_1 is increased by 10 each time for a failure of point registration, or reduced by 10 each time N_4 or more successive pairs of images have been registered without increasing N_1. The choice of using 10 as the step size for adjusting N_1 is

Table 1 Number of interest point necessary for 4,8,12,16,20,24,28 pairs of corresponding points. The amount of computation saved is shown in the last row.

# of corresponding point		4	8	12	16	20	24	28
# of interest point	1 to 2	4	14	18	23	27	39	44
	1 to 3	14	23	29	41	53	56	62
	1 to 4	20	33	53	66	78	87	97
	1 to 5	18	33	50	56	70	87	97
	1 to 6	155	191	247	273	334	429	474
Computation saved (%)		79.22	78.41	77.88	77.40	77.40	77.85	77.81

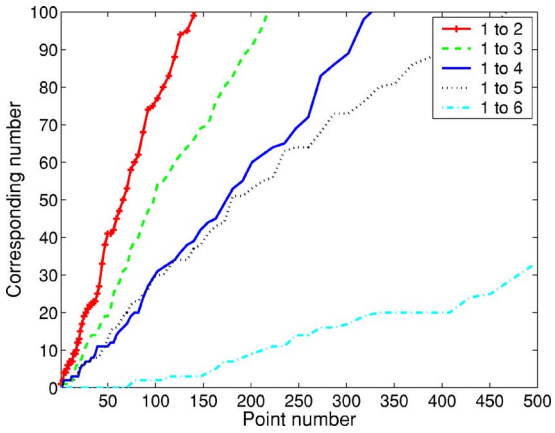

Fig. 3 Correspondence for the interest points.

not critical: any step size, as long as it is not too large, can reduce computation.

Upon validation of the matching points, a local registration is carried out to accommodate changes in target appearance. Moving targets are removed and luminance variance is corrected. Finally, the images are warped to a common image grid to form a comprehensive view.

The proposed dynamic point selection is evaluated on the dataset in Ref. 8. This dataset includes the transformation parameters in addition to the images. Figure 2 shows three of the images and 100 interest points detected on each of them, respectively.

As images become more disparate, e.g., when the common area of the image is less, the corresponding points are fewer. Refer to Figs. 2 and 3 for examples. However, to ensure correctly identifying correspondence between images, the number of corresponding points should be greater than a certain number. For example, if the projective model is used to represent the transformation between images, at least 4 pairs of corresponding points have to be identified. Therefore, more interest points should be used for a pair of more disparate images. As shown in Table 1, only 4 interest points are needed from images 1 and 2, to result in 4 pairs of corresponding points between them; while 14 interest points are needed for images 1 and 3, for the same number of corresponding points.

To register a series of images without using dynamic point selection, the number of interest points has to be the maximum for matching any pair of images. As an example, refer to Table 1. To ensure 4 pairs of corresponding interest points between image 1 and image 2, 3, 4, 5, or 6, at least

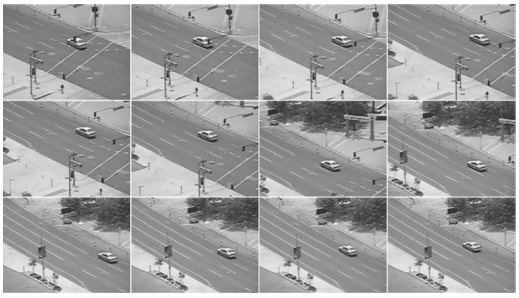

Fig. 4 Testing video sequence.

Fig. 5 Mosaic generated from the image sequence shown in Fig. 4.

4, 14, 20, 18, or 155 interest points have to be found, respectively. Without using dynamic point selection, at least 155 interest points have to be found from each of the images. However, by using dynamic point selection, only enough interest points are required. Because the computation complexity of point matching is quadratic to the number of points, the computation time can be reduced as much as 79.22%. Experiments are carried out to find 4, 8, 12, 16, 20, 24, or 28 pairs of corresponding interest points. As shown in Fig. 3 and Table 1, computation time can be reduced around 78%.

The proposed mosaicking algorithm has been well tested on real-world, monocular video sequences. It is shown to be accurate and robust, and runs in real time on a P4 CPU of 2.4 GHz. One of the sequences is shown in Fig. 4 and the mosaic is shown in Fig. 5.

In conclusion, an image mosaicking algorithm is developed by using a novel dynamic point selection procedure. It automatically selects a sufficient number of interest points. Simulations show that the proposed algorithm generates mosaic accurately and efficiently. Simulations also show that the dynamic point selection procedure can reduce computation time for point matching by about 78%.

Acknowledgment

This work was supported by General Dynamics C4 Systems, and in part by the National Science Foundation under grant ECS-0002098.

References

1. M. Irani, S. Hsu, and P. Anandan, "Video compression using mosaic representations," *Signal Process. Image Commun.* **7**, 529–552 (1995).
2. F. H. Moffitt and E. M. Mikhail, *Photogrammetry*, 3rd ed., Harper & Row, New York (1980).
3. H. S. Sawhney and S. Ayer, "Compact representation of videos through dominant multiple motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.* **18**(8), 814–830 (1996).
4. R. Szeliski, "Video mosaics for virtual environments," *IEEE Comput. Graphics Appl.* **16**, 22–30 (1996).
5. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.* **65**(1-2), 43–72 (2005).
6. C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, Manchester, NH, pp. 147–151 (1988).
7. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* **60**(2), 91–110 (2004).
8. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in Intl. Conf. Comput. Vision Patt. Recog. 2, 257–263 (2003).