High-resolution Video Mosaicing for Documents and Photos by Estimating Camera Motion

Tomokazu Sato[ab], Sei Ikeda[a], Masayuki Kanbara[ab],
Akihiko Iketani[b], Noboru Nakajima[b], Naokazu Yokoya[ab], Keiji Yamada[b]

[a] Nara Institute of Science and Technology, Ikoma, Nara, Japan
[b] NEC, Ikoma, Nara, Japan

ABSTRACT

Recently, document and photograph digitization from a paper is very important for digital archiving and personal data transmission through the internet. To realize easy and high quality digitization of documents and photographs, we propose a novel digitization method that uses a movie captured by a hand-held camera. In our method, first, 6-DOF(Degree Of Freedom) position and posture parameters of the mobile camera are estimated in each frame by tracking image features automatically. Next, re-appearing feature points in the image sequence are detected and stitched for minimizing accumulated estimation errors. Finally, all the images are merged as a high-resolution mosaic image using the optimized parameters. Experiments have successfully demonstrated the feasibility of the proposed method. Our prototype system can acquire initial estimates of extrinsic camera parameters in real-time with capturing images.

Keywords: Video mosaicing, Extrinsic camera parameter recovery, Image feature tracking

1. INTRODUCTION

In recent years, document and photograph digitization from a printed or drafted paper is very important for digital archiving and personal data transmission through the internet. Though many people wish to digitize documents on a paper easily, now heavy and large image scanners are required to obtain high quality digitization. To realize easy and high quality digitization of documents and photographs, a hand-held camera that can be freely moved is competent. A high-resolution image of a document is expected to be synthesized by applying a video mosaicing technique to a movie taken by the mobile camera.

In the literature of video mosaicing, a number of methods have been explored and proposed. Szeliski[1] developed an image based video mosaicing method using 8-DOF projective image transformation parameters between pairs of input images. His method can be used when a target is a plane (planar image mosaicing) or optical centers of images are approximately fixed throughout the video capturing (panorama image mosaicing). After his work, several methods are proposed for extending this mosaicing technique.[2–7] One of the major extensions is the use of image features instead of all the pixels in images in order to reduce calculation cost.[2–4] Although the calculation cost is drastically decreased by such an extension, a synthesized image from a long image sequence is usually distorted because they only stitch features between successive frames and ignore the re-appearing features. There exists an error accumulation problem in estimating projective image transformation parameters between images.

In order to synthesize a high-resolution mosaic image with minimum distortion from a plane object tracing movie, we propose a novel video mosaicing method that estimates 6-DOF extrinsic camera parameters instead of using general 8-DOF projective image transformation parameters. First, our method estimates 6-DOF extrinsic camera parameters for each frame tracking image features in the captured images. After the initial estimate, re-appearing features are stitched throughout the input images and accumulated errors are minimized. Our method is based on the assumption that the documents are given on a plane paper and the image plane of

Further author information: (Send correspondence to T.S)
T.S, S.I, M.K, N.Y: E-mail: {tomoka-s,sei-i,kanbara,yokoya}@is.aist-nara.ac.jp
A.I, N.N, K.Y: E-mail: {iketani,noboru,yamada}@ccm.cl.nec.co.jp

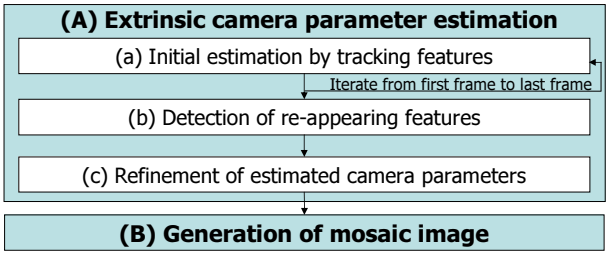


Figure 1. Flow diagram of image mosaicing.

camera in the first frame is approximately parallel to the plane object. Additionally, for reducing the degree of freedom of camera parameters, intrinsic camera parameters must be estimated in advance and they must be fixed throughout the image capturing.

This paper is structured as follows. Section 2 describes a method for generating a high-resolution mosaic image. In Section 3, experimental results with documents and photos show the feasibility and the accuracy of the proposed method. Finally, Section 4 describes conclusion and future work.

2. VIDEO MOSAICING BY ESTIMATING EXTRINSIC CAMERA PARAMETERS

This section describes a method for generating a mosaic image with minimum distortion by estimating extrinsic camera parameters in capturing images. As shown in Figure 1, the proposed method first estimates extrinsic camera parameters of a freely moving mobile camera (A), and a mosaic image is then generated by using the estimated parameters (B).

In the following sections, first, extrinsic camera parameters and error functions are defined. The stages (A) and (B) are then described in some detail.

2.1. Definition of extrinsic camera parameters

In this section, extrinsic camera parameters and an error function for estimating them are defined. In general mosaicing methods for planar mosaicing, the 8-DOF projective image transformation parameter $\mathbf{H}_f$ is used for image registration and is defined for each successive image pair as follows:

$$(au_{(f+1)p}, av_{(f+1)p}, a)^T = \mathbf{H}_f (u_{fp}, v_{fp}, 1)^T, \tag{1}$$

$$\mathbf{H}_f = \begin{pmatrix} m_{1f} & m_{2f} & m_{3f} \\ m_{4f} & m_{5f} & m_{6f} \\ m_{7f} & m_{8f} & 1 \end{pmatrix}, \tag{2}$$

where a is a parameter, and (u_{fp}, v_{fp}) denotes the 2-D position of the feature p in the f-th frame image. By using this equation, images can be registered even if intrinsic parameters of the camera are unknown. However, this expression easily accumulates estimation errors because transformation parameters are defined only between two successive frames.

To reduce accumulative errors, we define the 6-DOF transformation matrix $\mathbf{M}_f$ instead of $\mathbf{H}_f$ in Eq. (2) using known intrinsic parameters. In this paper, as shown in Figure 2, the transformation matrix $\mathbf{M}_f$ of the f-th frame is defined between the mosaic image plane and the f-th frame image plane using camera position (t_{1f}, t_{2f}, t_{3f}) and camera posture (r_{1f}, r_{2f}, r_{3f}) parameters of the f-th frame.

$$(a\hat{u_{fp}}, a\hat{v_{fp}}, a)^T = \mathbf{M}_f (x_p, y_p, 1)^T, \tag{3}$$

$$\mathbf{M}_f = \begin{pmatrix} c_1c_3 + s_1s_2s_3 & s_1c_2 & t_{1f} \\ -s_1c_3 + c_1s_2s_3 & c_1c_2 & t_{2f} \\ c_2s_3 & -s_2 & t_{3f} \end{pmatrix}, \tag{4}$$

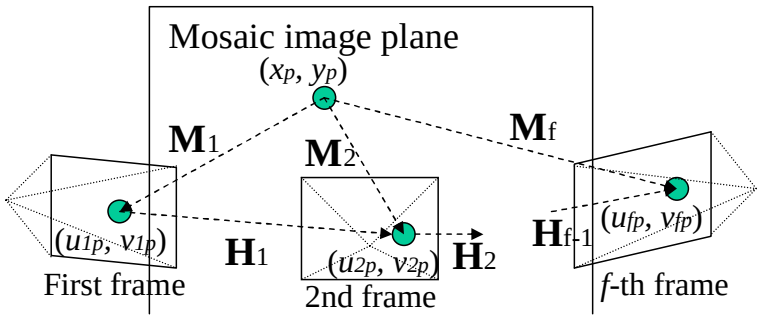


Figure 2. Mosaic image plane and camera.

$$s_i = \sin(r_{if}),\ c_i = \cos(r_{if}) \qquad (i = 1, 2, 3), \tag{5}$$

where (x_p, y_p) is the position of the feature p in the mosaic image plane, $(\hat{u}_{fp}, \hat{v}_{fp})$ is the projected position of (x_p, y_p) to the f-th frame image with the ideal camera model. The position $(\hat{u}_{fp}, \hat{v}_{fp})$ in the ideal camera image is transferred to the position (u_{fp}, v_{fp}) in the real camera image by known intrinsic camera parameters including focus, aspect, optical center and distortion parameters. The transformation matrix $\mathbf{M}_f$ defined in Eq. (4) is essentially the same as a usual extrinsic camera matrix except the omission of z-axis parameters, because a target object is always on the z=0 plane.

On the other hand, the position $(u_{fp},\ v_{fp})$ computed by Eq. (3) and known intrinsic camera parameters is not always consistent with an actually detected position (u'_{fp}, v'_{fp}) in the real image due to quantization and detecting errors. In this paper, the squared errors E_{fp} is defined as an error function for the feature p in the f-th frame as follows.

$$E_{fp} = \left\{(u_{fp} - u'_{fp})^2 + (v_{fp} - v'_{fp})^2\right\}. \tag{6}$$

The sum of E_{fp} is employed for estimating $\mathbf{M}_f$ and (x_p, y_p) in the following section.

2.2. Extrinsic camera parameter estimation

As shown in Figure 1, the proposed extrinsic camera parameter estimation method is constructed of the three processes. First, extrinsic camera parameters $\mathbf{M}_f$ from the first frame to the last frame are estimated by tracking features with the iterating process for each frame (a). Re-appearing features in the whole acquired images are then detected for bounding features even if they do not appear in successive frames (b). Finally, estimated extrinsic camera parameters are refined globally in the whole input image sequence (c). The following briefly describes each process.

2.2.1. Initial estimation by tracking features

In this process for each frame, an initial estimate of extrinsic camera parameter $\mathbf{M}_f$ is computed by tracking image features. This process is basically an extension of our previous work.[8] In the first frame, $\mathbf{M}_f$ is set as an identity matrix in accordance with the assumption that the image plane in the first frame is approximately parallel to the target object. The positions (x_p, y_p) in the mosaic image plane of image features in the first frame are also computed based on this assumption. Note that even if the target object and the image plane of the first frame are not accurately parallel to each other, they are corrected at the refinement process.

In the succeeding frames ($f > 1$), $\mathbf{M}_f$ is determined by iterating the following steps until the last frame.

Tracking of image features: All of image features are automatically tracked from the previous frame to the current frame by using a standard template matching with Harris corner detector.[9] The RANSAC approach[10] is also employed for eliminating outliers.

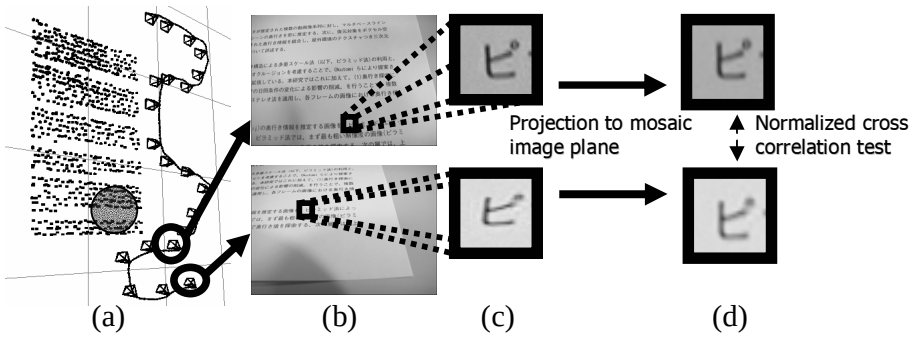


Figure 3. Detection of re-appearing features. (a) camera path, posture and feature positions on mosaic image plane, (b) sampled frames of input images, (c) templates of a feature on different images, (d) templates projected to a mosaic image plane.

Extrinsic camera parameter estimation: The tracked position (u_{fp}, v_{fp}) in the f-th frame and the position in the mosaic image plane (x_p, y_p) are used for estimating the extrinsic camera parameter $\mathbf{M}_f$. In this step, the error function $\sum_p E_{fp}$ is minimized by the non-linear squared minimization.

Estimation of feature position on mosaic plane: The position (x_p, y_p) in the mosaic image plane of each feature p is computed and refined in every frame by minimizing the error function $\sum_f E_{fp}$.

Addition and deletion of features: In order to obtain accurate estimates of camera parameters, good features should be selected. The set of natural features is automatically updated by checking conditions of features using multiple criteria.[8]

By iterating the above steps, initial estimates of extrinsic parameters $\mathbf{M}_f$ and feature positions (x_p, y_p) in the mosaic image plane are automatically computed from the first frame to the last frame.

2.2.2. Detection of re-appearing features

In an input movie, some of image features usually come in to and come out from the sight of the camera in several times due to the camera motion. In this research, these re-appearing features are detected for bounding features to reduce accumulative estimation errors.

As shown in Figure 3, the same image feature in a non-successive image pair often exhibits different looks as an effect of camera motion. To remove this effect, first, templates of all the features are projected to the mosaic image plane. Next, feature pairs whose distance is less than a given threshold are selected and tested with the normalized cross correlation function using multi-resolution templates. By using this approach, re-appearing features can be detected even if the templates of the features are rotated and distorted by camera motion.

2.2.3. Refinement of estimated camera parameters

In this process, the accumulation of estimation errors is minimized over the whole input. The accumulated estimation error E is given by the sum of re-projection errors as in Eq. (7) and is minimized with respect to the camera parameters $\mathbf{M}_f$ and the feature positions (x_p, y_p) in the mosaic image plane over the whole input.

$$E = \sum_f \sum_p E_{fp}. \tag{7}$$

By minimizing this error function, extrinsic camera parameters are refined and accumulated errors of the initial estimates are minimized.

2.3. Generation of mosaic image

At the final stage, a mosaic image is generated by projecting pixels in the mosaic image plane to all the frame images using Eq. (3) with refined 6-DOF extrinsic camera matrix $\mathbf{M}_f$. In this paper, a blending method is employed to generate a smooth mosaic image.

A color of pixel on the mosaic plane $C(x, y)$ is computed by colors of input images $I_f(u, v)$ and blending function $W(u, v)$ as follows.

$$C(x, y) = \frac{\sum_f W(u_f, v_f) I_f(u_f, v_f)}{\sum_f W(u_f, v_f)}, \tag{8}$$

$$W(u, v) = \begin{cases} (u^2 + v^2)^{-1}; & (u, v) \subseteq \text{image region} \\ 0; & \text{otherwise} \end{cases}, \tag{9}$$

where (u_f, v_f) is a projected position of (x, y) to the f-th frame image plane obtained by applying Eq. (3) and known intrinsic camera parameters.

3. EXPERIMENTS

To show the feasibility of the proposed method, a prototype image mosaicing system is constructed with a desktop PC (Pentium-4 3.2GHz, Memory 2GB) and a calibrated IEEE1394 CCD camera (Aplux C104T). The camera is hand-held. In experiments, two kinds of plane papers are used for the target objects. One is a printed A4 size document (sequence 1). The other is a printed photograph on an A4 size paper (sequence 2). In both papers, plus marks (+) are printed on 40mm grid points for quantitative evaluation.

3.1. Mosaicing for a document

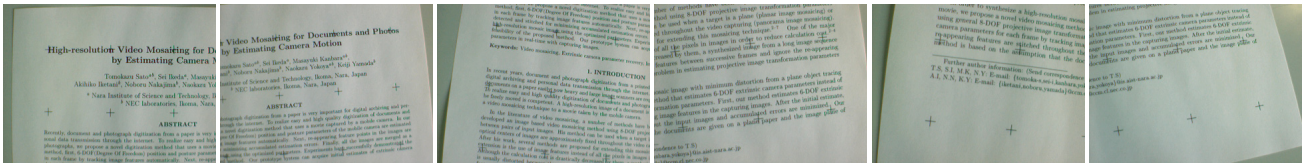
As shown in Figure 4, a target paper of document is captured as 640×480 images of 150 frames at 15 fps. First, initial estimates of extrinsic camera parameters are acquired by automatically tracking image features as shown in Figure 5. In this experiment, 95 points on average of the image features are tracked per frame. After the initial estimate, re-appearing features are detected and accumulated errors are minimized. The number of detected re-appearing features is 129 points in this experiment, and the average re-projection error of the features after refinement is 0.75 pixel. Figure 6 illustrates the acquired extrinsic camera parameters and the feature positions on the mosaic image plane after the refinement process. The curved line shows the estimated camera path and pyramids show the camera postures in every 10 frames.

Finally, all the input images are unified into a mosaic image by using the estimated camera parameters as shown in Figure 7. The generated mosaic image size is 1600×1916. We can confirm that the distortion of synthesized images is very little. However, some part of the mosaic image is unclear due to the shortage of the input image resolution. Super-resolution techniques should be attempted to solve it. The performance of our system for this sequence is as follows: 15 fps for image acquisition and initial parameter estimation, 1 second for detecting re-appearing features in 150 frames, 25 seconds for camera parameter refinement, and 16 seconds for mosaic image generation.

3.2. Mosaicing for a photograph

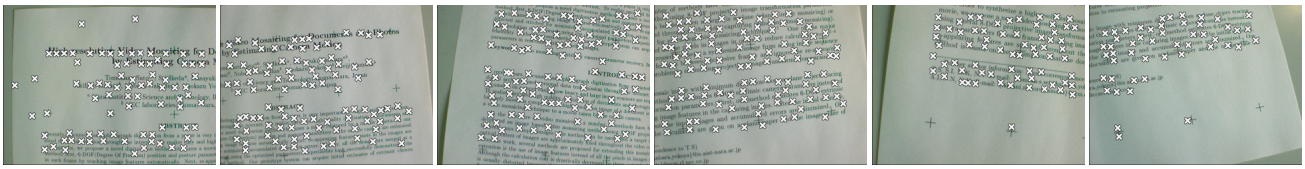
In this experiment, the target is a printed photograph. The target photograph is captured as 640×480 images of 150 frames at 15 fps as shown in Figure 8. Simultaneously, image features are automatically tracked for acquiring initial estimates as shown in Figure 9. The average number of tracked features is 90 points per frame. After the detection of 210 re-appearing features, the extrinsic parameters are refined. The average re-projection error of the features is 0.98 pixel. Figure 10 shows the extrinsic camera parameters and the feature positions on the mosaic image plane after refinement. It can be observed that the recovered camera path is smooth and there is no discontinuity.

Finally, the mosaic image is generated as a 1600×1997 size image as shown in Figure 11. The performance of our system for this sequence is as follows: 15 fps for image acquisition and initial parameter estimation, 1 second for detecting re-appearing features in 150 frames, 42 seconds for camera parameter refinement, and 21 seconds for mosaic image generation. Although we can confirm that the distortion of synthesized images is very little, the re-projection error of the features is bigger and the computational time for the refinement is longer than the results for the document because there are more similar patterns in the photograph than in the document.



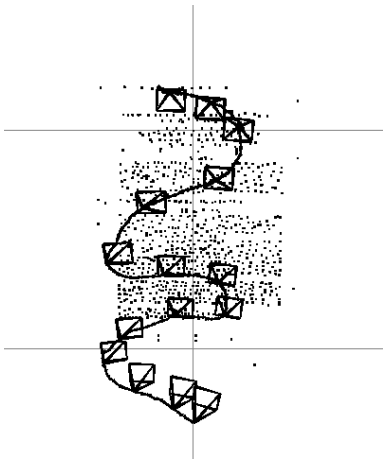
first frame 30-th frame 60-th frame 90-th frame 120-th frame 150-th frame

Figure 4. Sampled frames of an input movie (sequence 1).

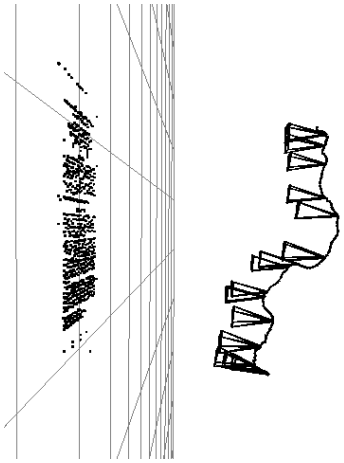


first frame 30-th frame 60-th frame 90-th frame 120-th frame 150-th frame

Figure 5. Result of feature tracking (sequence 1).



(a) top view



(b) side view

Figure 6. Estimated extrinsic camera parameters and feature positions (sequence 1).

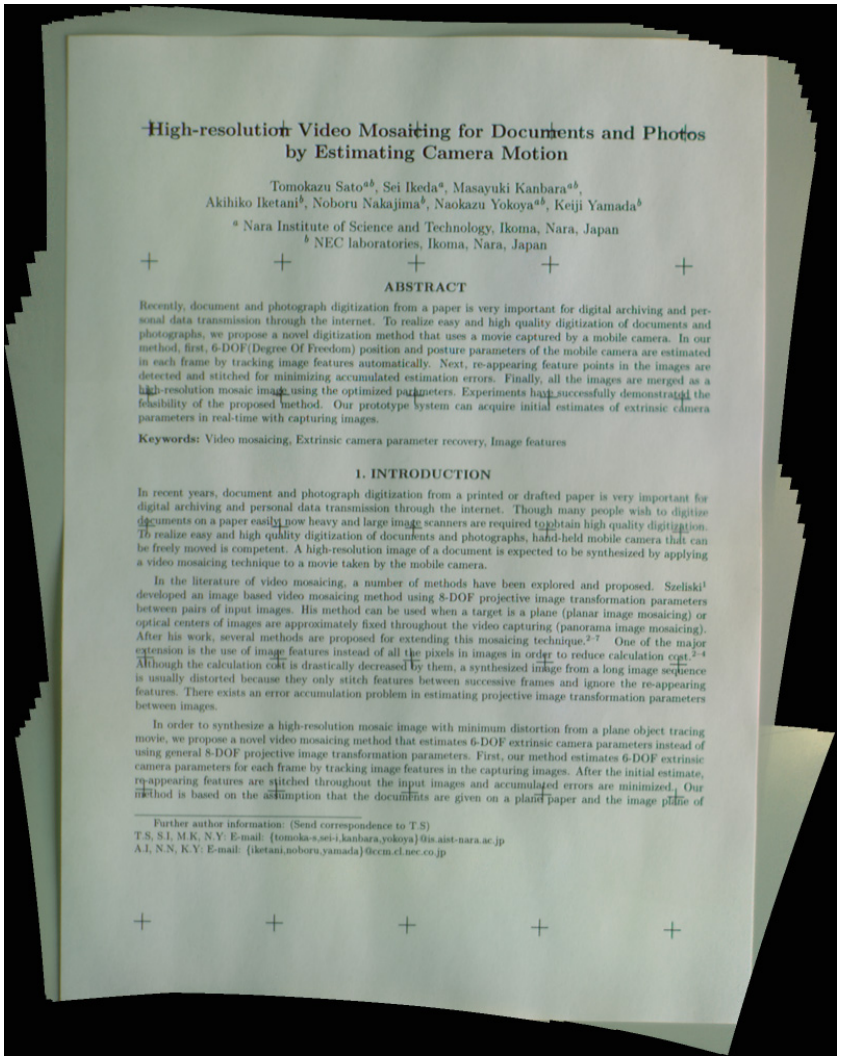


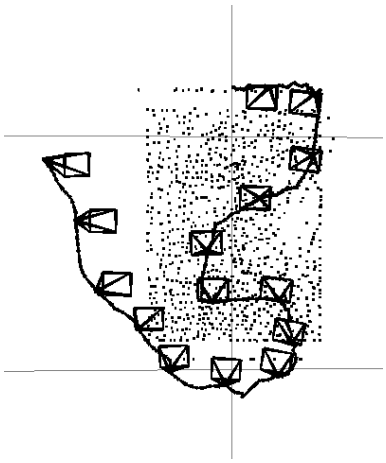
Figure 7. Generated mosaic image (sequence 1).

first frame 30-th frame 60-th frame 90-th frame 120-th frame 150-th frame

Figure 8. Sampled frames of an input movie (sequence 2).

first frame 30-th frame 60-th frame 90-th frame 120-th frame 150-th frame

Figure 9. Result of feature tracking (sequence 2).



(a) top view

(b) side view

Figure 10. Estimated extrinsic camera parameters and feature positions (sequence 2).

Figure 11. Generated mosaic image (sequence 2).

3.3. Quantitative evaluation

In this section, the distortions of the generated mosaic images are quantitatively evaluated by measuring the distances between adjacent grid positions. In the above two experiments, plus marks (+) have been printed on the target paper at every 40mm grid positions for this evaluation. First, the positions of the plus marks are acquired manually in the generated mosaic image. The distances between adjacent plus marks are then computed in the unit of pixel. The average, maximum, minimum and standard deviation of the distances are shown in Table 1. The percentage of each value from the average distance is also shown in parenthesis. In this table, the standard deviation is considered as the average distortion of the generated image. Although the average distortion of the photograph is a little worse than that of the document, both the average distortions are sufficiently little for the purpose of digital archiving and personal data transmission.

Table 1. Distances of adjacent grid points in generated mosaic images [pixels(percentage from average)]

target	average	maximum	minimum	standard deviation
document	255.5(100.0)	258.0(101.0)	253.0(99.0)	1.20(0.47)
photograph	245.6(100.0)	249.0(101.4)	243.0(98.9)	1.27(0.52)

4. CONCLUSION

This paper has proposed a method for generating a mosaic image from a freely moving camera with minimum distortion. In experiments, two kinds of mosaic images are successfully generated in short time. In future work, super-resolution techniques should be attempted using estimated camera parameters for acquiring higher resolution images. The mosaicing method for non-plane targets will also be explored to realize a more useful and practical mosaicing system.

REFERENCES

1. R. Szeliski: "Image Mosaicing for Tele-Reality Applications," Proc. IEEE Workshop on Applications of Computer Vision, pp. 230–236, 1994.
2. N. Chiba, H. Kano, M. Higashihara, M. Yasuda and M. Osumi: "Feature-based Image Mosaicing," Proc. IAPR Workshop on Machine Vision Applications, pp. 5–10, 1998.
3. S. Takeuchi, D. Shibuichi, N. Terashima and H. Tominage: "Adaptive Resolution Image Acquisition Using Image Mosaicing Technique from Video Sequence," Proc. IEEE Int. Conf. on Image Processing, vol. I, pp. 220–223, 2000.
4. C.T. Hsu, T.H. Cheng, R.A. Beuker and J.K. Hong: "Feature-based Video Mosaicing," Proc. IEEE Int. Conf. on Image Processing, vol. II, pp. 887–890, 2000.
5. M. Lhuillier, L. Quan, H. Shum and H. T. Tsui: "Relief Mosaicing by Joint View Triangulation," Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, vol. I, pp. 785–790, 2001.
6. W. Du and H. Li: "Construction of Image Mosaics with Video Texture," Proc. Asian Conf. on Computer Vision, vol. II, pp. 871–876, 2002.
7. U. Bhosle, S. Chaudhuri and S.D. Roy: "A Fast Method for Image Mosaicing Using Geometric Hashing," IETE Journal of Research, Special Issue on Visual Media Processing, vol. 48, no. 3-4, pp. 317–324, 2002.
8. T. Sato, M. Kanbara, N. Yokoya and H. Takemura: "Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-baseline Stereo Using a Hand-held Video Camera," Int. Jour. of Computer Vision, vol. 47, no. 1-3, pp. 119–129, 2002.
9. C. Harris and M. Stephens: "A Combined Corner and Edge Detector," Proc. Alvey Vision Conf., pp. 147–151, 1988.
10. M.A. Fischler and R.C. Bolles: "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981.