NAIST-IS-DD0661201

Doctoral Dissertation

Video Mosaicing Based on Structure from Motion for Geometric Distortion-Free Document Digitization

Akihiko Iketani

September 24, 2008

Department of Information Systems
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Akihiko Iketani

Thesis Committee:

Professor Naokazu Yokoya	(Supervisor)
Professor Hirokazu Kato	(Co-supervisor)
Associate Professor Kazumasa Yamazawa	(Co-supervisor)
Assistant Professor Tomokazu Sato	(Co-supervisor)

Video Mosaicing Based on Structure from Motion for Geometric Distortion-Free Document Digitization*

Akihiko Iketani

Abstract

This thesis describes a novel video mosaicing method which is capable of generating a geometric distortion-free mosaic image. Video mosaicing, which stitches partial images of a target into a larger image called a mosaic image, has been used to obtain high resolution images of documents. Generally, mosaic images are prone to two types of geometric distortions: perspective distortion which appears when the target document is not fronto-parallel to the camera's image plane, and curvature distortion which is caused by projecting curved surface of the target document to the image plane of the camera.

This thesis first focuses on a flat document, and proposes a perspective distortion-free video mosaicing method for flat documents. In the proposed method, extrinsic camera parameters are estimated for each frame by applying structure from motion technique to the captured video. Using estimated extrinsic camera parameters, the method dewarps all the frame images and synthesizes them on a virtual fronto-parallel plane to generate a super-resolved mosaic image without perspective distortion.

Then, this method for flat documents is extended to deal with curved documents. This extended method generates a virtually flattened mosaic image of a curved surface. In this method, first, extrinsic camera parameters, along with 3-D feature positions are estimated by structure from motion. Then, by fitting

*Doctoral Dissertation, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0661201, September 24, 2008.

a parameterized surface to the 3-D features, the shape of the curved document is estimated. Using the estimated shape and extrinsic camera parameters, all the frame images are dewarped to remove curvature distortion and synthesized on a virtual plane to generate a mosaic image without curvature distortion.

Experiments on both flat and curved documents have been performed using a prototype system. In each experiment, the mosaic image has been proved to be distortion-free by quantitative analysis on distortion.

This work differs from previous works in document digitization in that it does not require any special hardware equipment besides a video camera. This makes the method suitable for mobile solution, and is one step forward to ubiquitous document digitization.

Keywords:

document digitization, geometric distortion, video mosaicing, extrinsic camera parameter estimation, super resolution

Acknowledgements

This work was completed under the supervision of Professor Naokazu Yokoya and Assistant Professor Tomokazu Sato of the Graduate School of Information Science at the Nara Institute of Science and Technology.

I would especially like to thank my advisor, Professor Naokazu Yokoya, who has been a great source of support and guidance throughout my Ph.D. study. Without the continuous encouragement and advice of Professor Yokoya, I could never have completed this thesis. I would also like to thank Assistant Professor Tomokazu Sato, who has also been a wonderful source of support and guidance throughout this study.

I am deeply grateful to Professor Hirokazu Kato and Associate Professor Kazumasa Yamazawa, who has given me insightful comments on this research as members of the thesis committee.

I have also benefited greatly through being able to meet and work with a number of people during my doctoral work at the Nara Institute of Science and Technology. Assistant Professor Masayuki Kanbara and Assistant Professor Sei Ikeda of the Graduate School of Information Science at the Nara Institute of Science and Technology provided helpful comments and invaluable discussions since the beginning of the joint project with NEC Corporation. I would also like to thank all the members of the Vision and Media Computing Laboratory of the Graduate School of Information Science at the Nara Institute of Science and Technology for their friendship and support.

I would also like to acknowledge the help and support from my supervisors and colleagues in NEC Corporation. Doctor Keiji Yamada of C&C Innovation Laboratories has given me an opportunity to start this work as the joint project of NEC Corporation and the Graduate School of Information Science

at the Nara Institute of Science and Technology, and has guided me in the early years of this study. Mr. Noboru Nakajima has given me numerous advices and guidance throughout this study. I would like to thank Mr. Atsushi Kashitani and Mr. Toshihiko Hiroaki of Common Platform Software Research Laboratories for their supervision and continuous understanding in this work. I would also like to thank all the members of Common Platform Software Research Laboratories who has supported and helped me to finish this work.

Finally, I especially want to thank my wife, Akiko, for understanding, supporting and encouraging me during my Ph.D. study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Document digitization technology has enabled us to store thousands of pages of printed and written documents in a small piece of storage device in a computer, and to share them on the Internet with people over the world. For many years, flat-bed scanners have been commonly used for this purpose for their image quality. Flat-bed scanners, however, are usually large and heavy, and are only available in limited situations such as in office environments. On the other hand, portable imaging devices such as digital cameras and cellular phones with a built-in camera have become so popular, and there is increasing interest in using these devices as alternatives to flat-bed scanners. Combined with wireless connection, these devices will turn into portable facsimile machines which allow us to scan and send documents anytime, anywhere. The concept of this ubiquitous document digitization is shown in Figure 1.1.

Camera-based document digitization, however, has several problems to be solved. One of the most critical problems is the resolution of the image acquired with these devices. For example, 10M pixel cameras enable full A4 pages to be sampled at about 320 dots per inch (dpi), whereas standard flat-bed scanners enable sampling at a few thousands dpi.

Another problem is the geometric distortion in the acquired image. There are two types of geometric distortions. One is perspective distortion, which appears when the target document is not fronto-parallel to the camera's image plane. An example of this perspective distortion is shown in Figure 1.2(a). Due to the perspective distortion, lines of text and page boundary are no longer

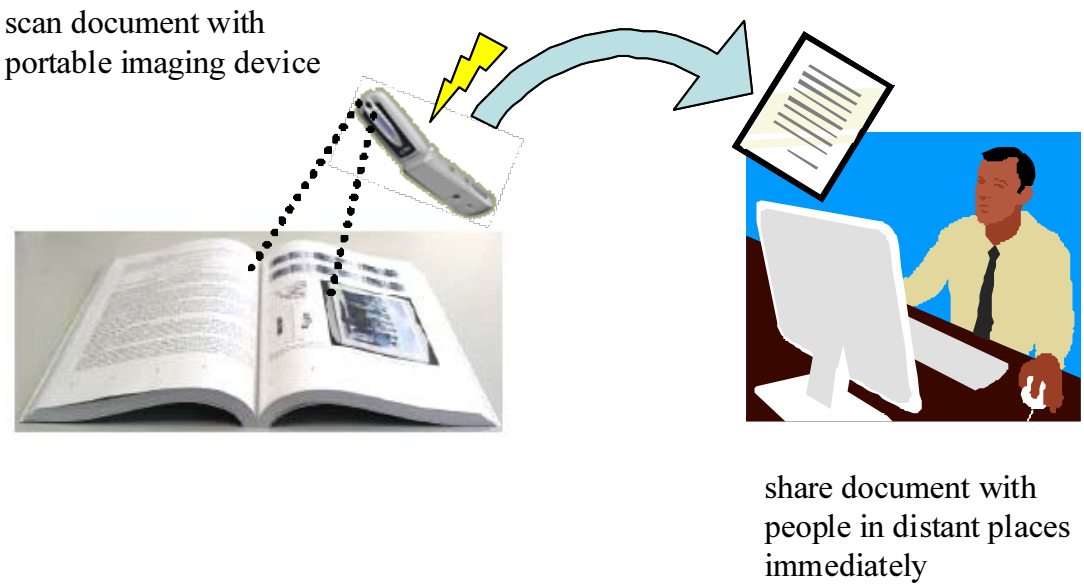
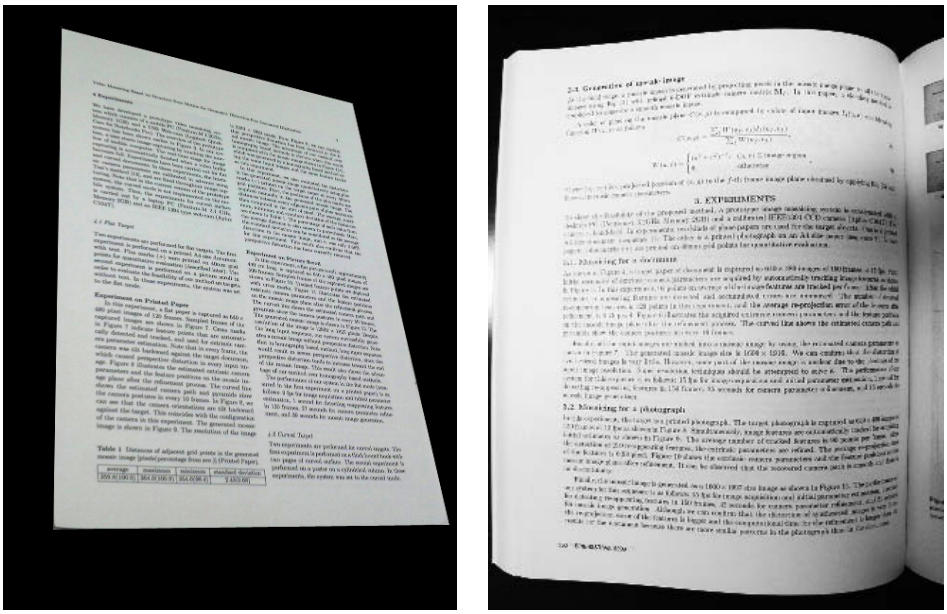


Figure 1.1. Concept of ubiquitous document digitization.

parallel in the captured image.

The other type of geometric distortion is curvature distortion, which is caused by projecting curved surface of the target document to the image plane of the camera. This type of distortion is commonly observed for a bound book, as shown in Figure 1.2(b). Distortion due to the curvature of the target is evident in the curved lines of text and contour.

Various methods have been proposed to overcome the problems in resolution and distortion. In this chapter, we first review previous works in camera-based document digitization. Then the positioning of this research against previous works is defined. Finally, the outline of this thesis is given.



(a) Perspective Distortion (b) Curvature Distortion

Figure 1.2. Geometric distortion in document images.

1.1. Previous Works in Camera-based Document Digitization

Previous document digitization methods can be classified into the following two types: methods for flat surface and those for curved surface. The first type of methods assumes the target document is on a plane. This is a common assumption which holds when the document is printed on a sheet of paper and is placed on a flat surface, e.g. desk, wall, etc. The other type of methods relaxes the above assumption of flat surface to curved surface. This kind of surface is often observed when the target is a thick bound book and is opened on a desk. In the following sections, both types of methods are reviewed and discussed.

1.1.1 Document Digitization for Flat Documents

In this section, we will have a review on camera-based document digitization methods for flat documents.

As described earlier, document images will undergo perspective distortion when the target document is not fronto-parallel to the camera's image plane. Various methods have been proposed to correct this perspective distortion. The most straightforward approach is to utilize the contour of the target to estimate the perspective. If the document can be assumed to be on a plane, the transformation from the document plane to the image plane can be modeled by a *homography*, which is defined by eight coefficients. The correction of perspective distortion can be accomplished once the eight unknowns are found. Since four pairs of corresponding points are sufficient to solve the eight unknowns, Jung et al. [JKH02] extract four corners on the rectangular boundary of the target to obtain four pairs of correspondences, and estimate the homography for distortion correction. This method, however, is only applicable when the rectangular boundary is captured in the image. It should also be noted that this method is very prone to error if the detection of four corners is unreliable.

Other methods utilize vanishing points to correct perspective distortion. As described in [Rot00], from vanishing points, which are the intersections of 3-D orthogonal lines, the perspective distortion can be recovered. In text documents, such orthogonal lines are abundantly obtained as parallel text lines, column edge lines, and page boundary lines. Thus, two vanishing points, one in the horizontal direction and the other in the vertical direction, can be robustly estimated. Myers et al. [MBLH01] first detect text lines in the image by connected component analysis after binarizing the image. Then, each text line is rotated and the horizontal projection profile is analyzed to find the top and baseline for each text line. Similarly, by analyzing the vertical projection profile, the vertical edge lines are found. Finally, the horizontal and vertical vanishing points are determined as the intersections of horizontal and vertical lines, respectively. Once the horizontal and vertical vanishing points are found, homography that maps the vanishing points back to infinity can be computed. By dewarping the image with this homography, perspective distortion is corrected. Although these methods are robust compared to the methods based

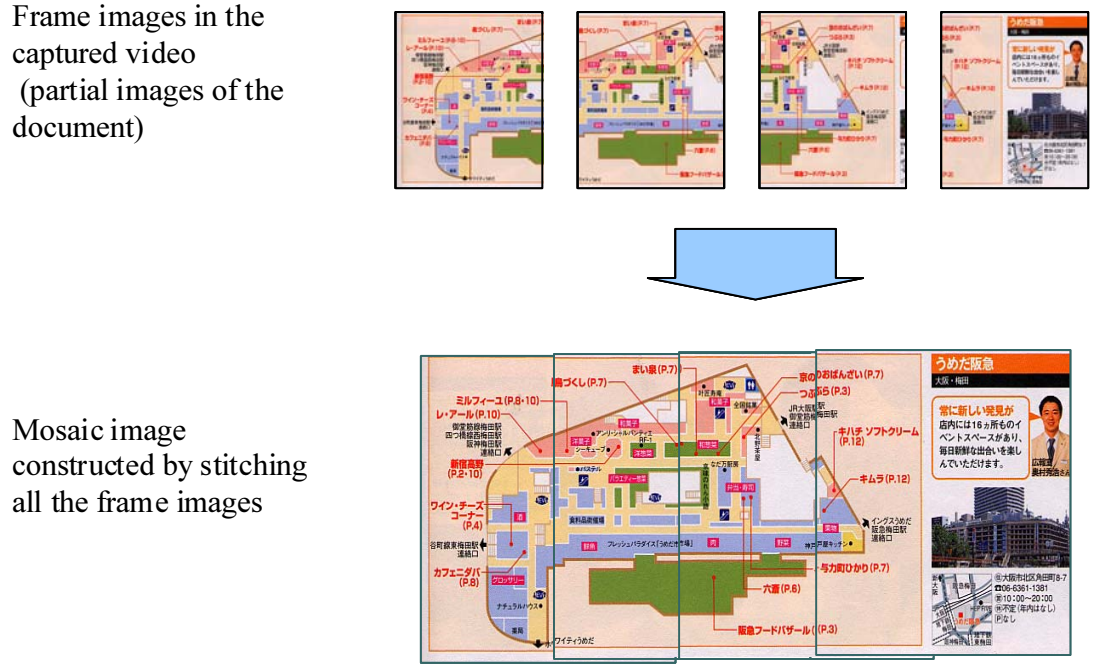


Figure 1.3. Concept of video mosaicing.

on four point correspondences, they are only applicable to documents with text lines. In case of documents with figures or photos, detection of vanishing points would be unreliable, and thus, these methods would fail in correcting perspective distortion.

So far, we have focused on methods for correcting perspective distortion. In the rest of this section, we will review methods to solve the resolution problem in camera-based document digitization. Various techniques have been proposed to solve the resolution problem. Among them, a video mosaicing technique is one of the most promising solutions. In video mosaicing, as shown in Figure 1.3, partial images of the document are captured as a video sequence, and multiple frame images are stitched seamlessly into one large, high resolution image called a *mosaic image*. Conventional methods usually carry out pairwise registration between two successive images, and construct a mosaic image by warping all the images to a reference frame (in general, the first frame). Szeliski [Sze94] has developed a method based on homography. In this

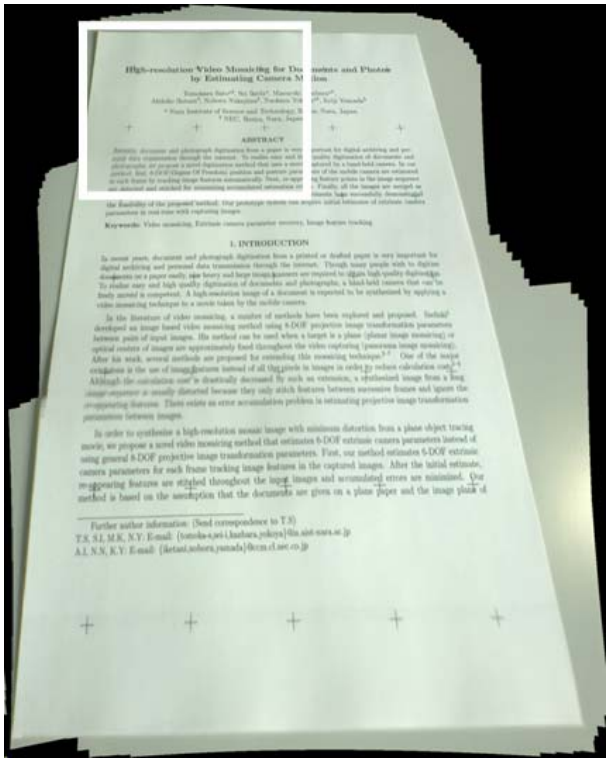


Figure 1.4. Mosaic image with perspective distortion.

method, for every pair of consecutive frames, the homography which minimizes the sum of squared differences between the two frames is estimated.

After his work, various extensions to this method have been proposed [CKH+98, TSTT00, HCBH00, LQST01, MJ02, KH03, BCR02, GS03]. One of the major extensions is the use of image features instead of all the pixels in images in order to reduce computational cost [CKH+98, TSTT00, HCBH00]. Some methods introduce a global optimization process of homographies after pairwise registration in order to reduce cumulative errors in the estimated homographies [MJ02, SHK98]. This process is sometimes called bundle adjustment.

Video mosaicing methods provide a simple way to construct a high resolution document image with a low resolution camera. The drawback in video mosaicing is that the user has to take special care in setting the camera's im-

age plane in the reference frame to be fronto-parallel to the document plane. Otherwise, the resultant mosaic image would suffer from the same type of perspective distortion as described earlier. Figure 1.4 shows an example of a mosaic image with perspective distortion. In this figure, the region corresponding to the reference frame is shown in a white rectangle. As can be seen, perspective distortion has been induced in the reference frame. Since all the other images are aligned to this reference frame, the same perspective distortion is observed all over the resultant mosaic image.

Combining video mosaicing with the above mentioned distortion correction methods can solve this problem. Liang et al. [LDD06] first correct perspective distortion in each frame image by utilizing vanishing points, and then synthesize these images to obtain perspective distortion-free mosaic image. This method, however, shares the same limitation as the above mentioned distortion correction methods in that it can only be applied to documents with text lines. It should also be noted that the detection of vanishing points, carried out in each frame individually, can be unstable since only a small number of partial text lines are present in each image.

1.1.2 Document Digitization for Curved Documents

In this section, camera-based document digitization methods for curved documents are reviewed.

As described earlier, the curvature of the document causes non-linear distortion in the captured image, as shown earlier in Figure 1.2(b). In the field of document analysis, various methods to remove this curvature distortion have been proposed. These methods can be classified into the following two types: methods based on 3-D shape measurement and those based on a priori knowledge on documents.

The first type of methods measures the surface shape or image deformation by special hardware equipment and restores the flattened image of the document. Brown and Seales [BS01] measure the shape of the document using a slit light projection device. Then, an elastic mesh model is fitted to the 3-D shape, and by pushing the mesh down to a plane, the flattened image of the page is obtained. Pilu [Pil01] fits an applicable surface model, where the distances

between mesh nodes are fixed, to the 3-D shape acquired by slit light projection device. This mesh model is flattened by an iterative process. Doncescu et al. [21] employ a light grid projector which is set up on a scanning table. They first project the light grid on a plane and record the position of each grid point as its initial position. Then, the same light grid is projected on a curved document. By morphing the captured image so that the grid points on the document is transformed to their initial positions, they correct the curvature distortion in the image. Note that in this method, the surface shape is not explicitly measured. Yamashita et al. [YKKM04] recover the range data of the surface by stereo vision system. After fitting NURBS surface to the range data, a flattened document image is restored by expanding the surface. Although these methods are applicable to arbitrarily curved surface, they require special and usually heavy hardware equipment, which makes them difficult to be applied for mobile solution.

The other type of methods restores a distortion-free image from a single input image using a priori knowledge on documents. The methods proposed by Cao et al. [CDL03] and Liang et al. [LDD05] assume the target document is composed of horizontal text lines, and extracts baseline for every text line using a morphological method. A distortion-free image of the target is generated by warping the captured image so that all the baselines are parallel to one another. Brown and Tsoi [BT04] assume the contour of the page is captured in the image, and warps the image so that the contour is transformed into a rectangle. Although these methods are capable of removing curvature distortion from a single input image and do not require any special hardware equipments, they can only be applied to targets which fulfill the underlying assumptions. Moreover, the resolution of the acquired image is limited to that of the camera itself.

Video mosaicing, described earlier in Section 1.1.1, might be considered as a good solution to acquire higher resolution images. However, video mosaicing methods, which are based on homography estimation, are only applicable when the target is a plane, or when the optical center of camera is approximately fixed throughout the video capturing. If the target is a curved surface, the above assumption no longer holds, thus will cause misalignment of images in

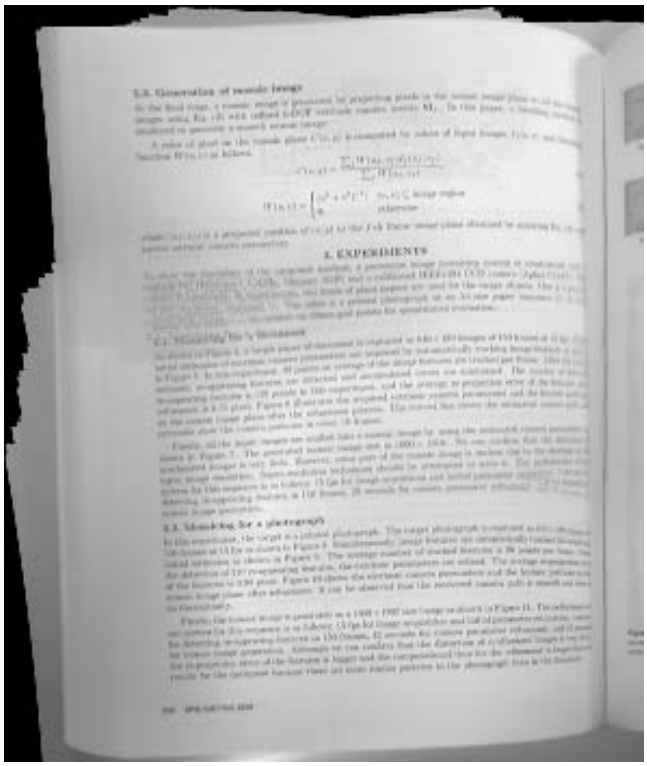


Figure 1.5. Mosaic image of curved surface by homography based method.

the resultant mosaic image. Figure 1.5 shows a mosaic image of a book shown in Figure 1.2(b) generated by a conventional homography based method. As can be seen, misalignment of images has caused distortion and blur in the mosaic image. Although there are some video mosaicing methods which can deal with curved surface, they require an active camera and a slit light projection device [GS03], or manual measurement of the surface shape [PBCP96].

1.2. Positioning of This Study

In the above sections, previous works in camera-based document digitization for flat and curved documents have been reviewed and discussed. Video mosaicing methods seem to be promising in that they enable document digitization with high resolution which goes beyond that of the camera itself. These

methods, however, are prone to perspective and curvature distortion, as shown in Figures 1.4 and 1.5. Although various methods have been proposed to correct these geometric distortions, they require special hardware equipments which make them difficult to be applied for mobile solution, or assume specific structures in documents, e.g. text lines, rectangular boundary, etc., which limits the type of documents they can be applied to. In summary, there has been no work proposed on video mosaicing that can deal with geometric distortion without using any special hardware equipments besides the camera.

The goal of this study is to develop a video mosaicing method which only requires a camera and is capable of generating a geometric distortion-free mosaic image. Details of the proposed method will be described in this thesis. This method is based on a structure-from-motion technique, which recovers extrinsic camera parameters as well as sparse 3-D scene geometry from an image sequence in real time. Using the 3-D geometry recovered by the algorithm, a mosaic image without perspective distortion is generated for a flat document. For a document composed of curved surfaces, mosaic images of virtually flattened pages are generated. Note that unlike previous distortion correction methods which make use of specific structures in documents, e.g. text lines or boundary of the document, the proposed method is capable of correcting geometric distortion regardless to the content of documents.

1.3. Outline of This Thesis

The rest of this thesis is structured as follows. This thesis first focuses on a flat document, and proposes a perspective distortion-free video mosaicing method for flat documents in Chapter 2. Experimental results on flat documents are shown and are quantitatively evaluated. Real-time implementation issues are also discussed. In Chapter 3, the above method for flat documents is extended to deal with curved documents. With this extension, a curvature distortion-free mosaic image can be generated for curved documents. Experiments are performed on curved documents to evaluate the feasibility of the method. Finally, conclusion of this thesis and future works are given in Chapter 4 .

Chapter 2

Video Mosaicing for Flat Document

2.1. Introduction

This chapter describes a perspective distortion-free video mosaicing method for flat documents. In this method, extrinsic camera parameters, instead of homographies, are estimated for each frame by applying a structure-from-motion technique [SKYT02] to the captured video. Using estimated extrinsic camera parameters, the method dewarps all the frame images and synthesizes them on a virtual rectified image plane to generate a perspective distortion-free mosaic image.

Two assumptions are made in this method. One is that the target document is planar. The other is that intrinsic camera parameters are known in advance, and remain fixed throughout image capturing.

In the following sections, first, the overview of the method is given(Section 2.2), and then each process composing the method is described in detail (Section 2.3, 2.4). After describing real-time implementation issue and the prototype system based on this method (Section 2.5), experimental results on flat documents using the prototype system is shown (Section 2.6). Finally, the conclusion of this chapter is given (Section 2.7).

2.2. Overview of the Method

The flow of the proposed method is given in Figure 2.1. As can be seen, the whole process is composed of two processes: extrinsic camera parameter estimation (A) and mosaic image generation (B).

In extrinsic camera parameter estimation (A), first, initial estimation of extrinsic camera parameter is carried out for each frame by tracking image features in the input video (a). Then, reappearing features are detected (b), and are utilized to globally optimize the estimated parameters to minimize the cumulative estimation error in the whole input sequence (c).

In mosaic image generation (B), all the images are projected and synthesized on a virtual rectified plane using the estimated extrinsic camera parameters. Finally, a super-resolved mosaic image is generated by an iterative back projection algorithm.

In the following sections, each process in the proposed method is described in detail.

2.3. Extrinsic Camera Parameter Estimation

This section describes the procedure to estimate extrinsic camera parameters. As shown in Figure 2.1, this process is composed of 3 processes: (a) initial estimation of extrinsic camera parameters by tracking image features, (b) detection of reappearing features and (c) refinement of the estimated camera parameters. Before describing these processes, first, extrinsic camera parameters and an error function to be minimized are defined. Then, processes (a) to (c) are described in detail.

2.3.1 Definition of Extrinsic Camera Parameter and Error Function

In this method, as shown in Figure 2.2, the extrinsic camera parameter $\mathbf{M}_f$ for the f-th frame is defined as a transformation between the coordinates on

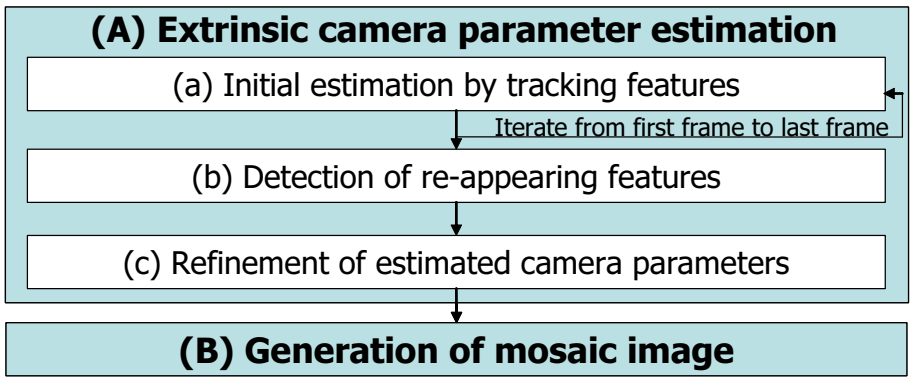


Figure 2.1. Flow diagram of video mosaicing for flat target.

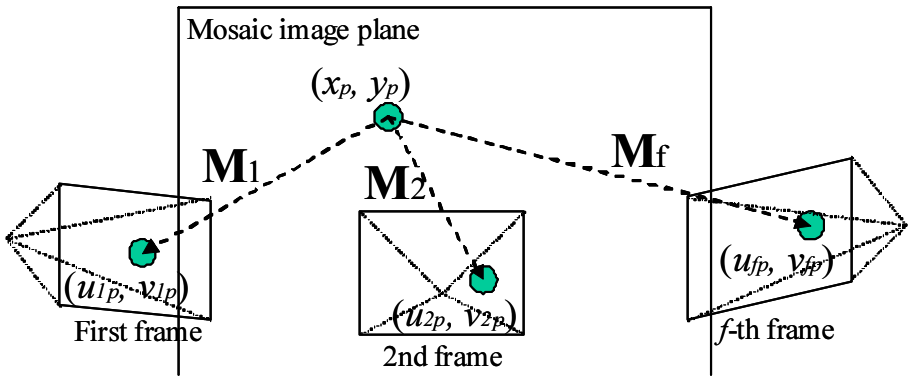


Figure 2.2. Mosaic image plane and camera.

the mosaic image plane and the f-th frame image plane. $\mathbf{M}_f$ is a simplified version of a standard extrinsic camera parameter $\mathbf{Mfull}_f$ commonly used in computer vision. In this section, we will see how the standard $\mathbf{Mfull}_f$ can be simplified to $\mathbf{M}_f$, and how lens distortion is dealt with in this definition. An error function used for the estimation of $\mathbf{M}_f$ is also described.

Generally, an extrinsic camera parameter $\mathbf{Mfull}_f$ for the the f-th frame that transforms the world coordinate system to the camera coordinate system

is given by a 3×4 matrix with 6 degree of freedom as follows:

$$\mathbf{Mfull}_f = \begin{pmatrix} c_1c_3 + s_1s_2s_3 & s_1c_2 & -c_1s_3 + s_1s_2c_3 & t_{1f} \\ -s_1c_3 + c_1s_2s_3 & c_1c_2 & s_1s_3 + c_1s_2c_3 & t_{2f} \\ c_2s_3 & -s_2 & c_2c_3 & t_{3f} \end{pmatrix}, \tag{2.1}$$

$$s_i = \sin(r_{if}),\ c_i = \cos(r_{if}) \quad (i = 1, 2, 3), \tag{2.2}$$

where (t_{1f}, t_{2f}, t_{3f}) are camera position parameters, and (r_{1f}, r_{2f}, r_{3f}) are camera posture parameters representing yaw, pitch, roll of a camera, respectively. If we consider an ideal camera with the focal length of 1 and without lens distortion, arbitrary 3-D point $\mathbf{S}_p = (x_p, y_p, z_p)$ is projected to $\hat{\mathbf{x}}_{fp} = (\hat{u}_{fp}, \hat{v}_{fp})$ on this ideal image coordinate by the following equation:

$$a \begin{pmatrix} \hat{u}_{fp} \\ \hat{v}_{fp} \\ 1 \end{pmatrix} = \mathbf{Mfull}_f \begin{pmatrix} x_p \\ y_p \\ z_p \\ 1 \end{pmatrix}, \tag{2.3}$$

where a is a parameter. Since the target is assumed to be a plane, without losing generality, we define the plane as $z = 0$, and set the z element of arbitrary 3-D positions on the target plane to 0. This degenerates 3-D coordinate $\mathbf{S}_p = (x_p, y_p, z_p)$ to 2-D coordinate (x_p, y_p), and simplifies the above equations as follows:

$$\mathbf{M}_f = \begin{pmatrix} c_1c_3 + s_1s_2s_3 & s_1c_2 & t_{1f} \\ -s_1c_3 + c_1s_2s_3 & c_1c_2 & t_{2f} \\ c_2s_3 & -s_2 & t_{3f} \end{pmatrix}, \tag{2.4}$$

$$a \begin{pmatrix} \hat{u}_{fp} \\ \hat{v}_{fp} \\ 1 \end{pmatrix} = \mathbf{M}_f \begin{pmatrix} x_p \\ y_p \\ 1 \end{pmatrix}. \tag{2.5}$$

The transformation matrix $\mathbf{M}_f$ in Eqs. (2.4) and (2.5) is the extrinsic camera parameter employed in this method, and is estimated for every frame in the input sequence.

So far, we have considered an ideal camera without lens distortion. In practice, however, we cannot neglect the effect of lens distortion, which induces displacement between the ideal image coordinate $\hat{\mathbf{x}}_{fp} = (\hat{u}_{fp}, \hat{v}_{fp})$ and the

corresponding coordinate $\mathbf{x}_{fp} = (u_{fp}, v_{fp})$ on the real image. According to Tsai's camera model [Tsa86], the relationship between these two coordinates is given by a set of intrinsic camera parameters as follows:

$$(\hat{u}_{fp}, \hat{v}_{fp}) = \left(R(u_{fp} - c_u)\frac{ccd_u}{scr_u F s_u}, R(v_{fp} - c_v)\frac{ccd_v}{scr_v F} \right), \tag{2.6}$$

$$R = (1 + k_1 r^2 + k_2 r^4), \tag{2.7}$$

$$r = \sqrt{\left(\frac{ccd_u}{scr_u s_u}(u_{fp} - c_u)\right)^2 + \left(\frac{ccd_v}{scr_v}(v_{fp} - c_v)\right)^2}, \tag{2.8}$$

where F is the focal length, s_u is the aspect ratio of the CCD sensor element, (c_u, c_v) is the position of optical center, (ccd_u, ccd_v) is the CCD size, (scr_u, scr_v) is the image resolution and (k_1, k_2) are distortion parameters.

In this method, the ideal image coordinate $\hat{\mathbf{x}}_{fp}$ is precalculated for every real image coordinate $\mathbf{x}_{fp}$by applying the above transformation before it is further transformed by Eq. (2.5). For simplicity, however, this transformation is omitted in the rest of this thesis.

Finally, an error function used for extrinsic camera parameter estimation is defined. In general, the projected position $\hat{\mathbf{x}}_{fp}$ of feature p to the f-th image frame does not coincide with the actually detected position $\mathbf{x}'_{fp} = (u'_{fp}, v'_{fp})$ on the ideal image coordinate, as shown in Figure 2.3, due to errors in feature detection, extrinsic camera parameter and 3-D feature position estimation. In this method, the squared distance between them, called reprojection error, is defined as follows for feature p on the f-th frame:

$$\begin{aligned} E_{fp} &= |\hat{\mathbf{x}}_{fp} - \mathbf{x}'_{fp}|^2 \\ &= \{(\hat{u}_{fp} - u'_{fp})^2 + (\hat{v}_{fp} - v'_{fp})^2\}. \end{aligned} \tag{2.9}$$

In the following sections, we will see how this reprojection error is employed to estimate $\mathbf{M}_f$ and $\mathbf{S}_p$.

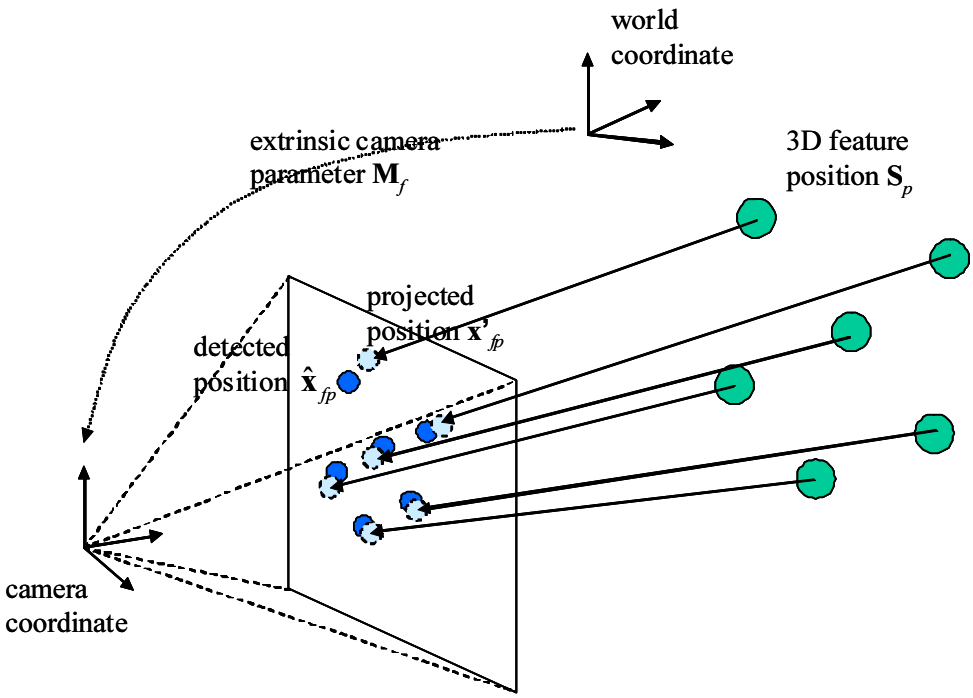


Figure 2.3. Definition of reprojection error.

2.3.2 Initial Estimation of Extrinsic Camera Parameters

Extrinsic camera parameter estimation starts from computing initial estimate of $\mathbf{M}_f$ for each frame. This is carried out by structure from motion method proposed in [SKYT02]. Given an initial value for $\mathbf{M}_1$, i.e. extrinsic camera parameter for the first frame, and (x_p, y_p) for every feature captured in the first frame, $\mathbf{M}_f$ for each frame, along with (x_p, y_p) for each feature on the target, are iteratively estimated.The flow of this process is as follows.

In the first frame ($f = 1$), as shown in Figure 2.4, it is assumed that the focal plane of the camera is approximately parallel to the target and that the viewpoint of the camera is a certain distance, say d, away from the mosaic image plane $z = 0$. Based on this assumption, the extrinsic camera parameter for the first frame is given as follows:

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -d \end{pmatrix}. \tag{2.10}$$

For each feature p in the first frame, which is detected by Harris corner detector [HS88], its position (x_p, y_p) on the mosaic image plane is given by the same assumption as follows:

$$\begin{pmatrix} x_p \\ y_p \end{pmatrix} = \begin{pmatrix} \hat{u}_{1p} d \\ \hat{v}_{1p} d \end{pmatrix}. \tag{2.11}$$

Note that these are only initial values, which will be corrected in the refinement process (Figure 2.1(c)).

In the subsequent frames ($f > 1$), $\mathbf{M}_f$ is estimated by iterating the following steps for each frame.

Tracking of image features: All the image features in the previous frame are tentatively tracked to the current frame using a standard template matching. In order to prevent drift problem in feature tracking, interest points detected by Harris corner detector [HS88] are used as candidate positions for matching.

After this tentative tracking, outliers of corresponding points are detected by the RANSAC approach [FB81]. For every feature detected as

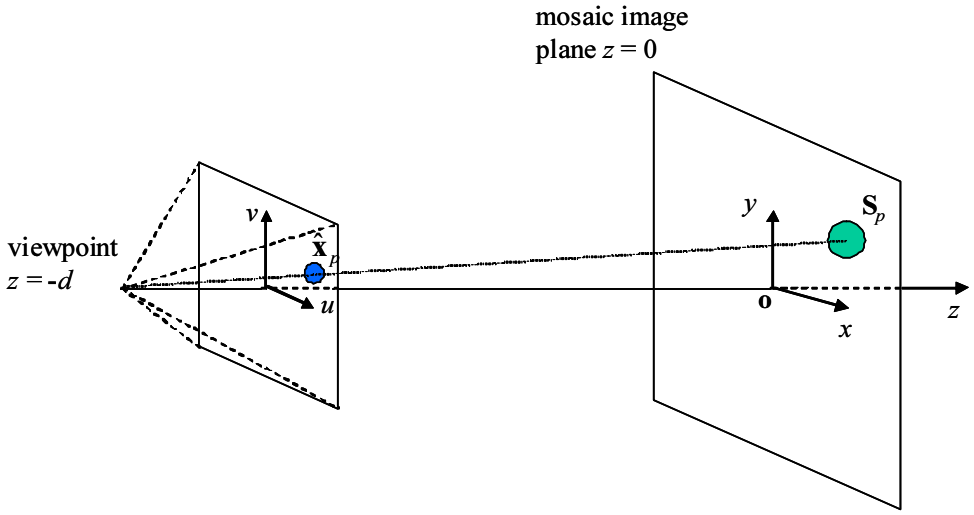


Figure 2.4. Initial assumption on the camera and the mosaic image plane.

an outlier, its corresponding point is re-searched in a limited searching area which is computed by a temporal camera parameter estimated using inlier points.

Extrinsic camera parameter estimation: In this step, extrinsic camera parameter $\mathbf{M}_f$ for the current frame f is estimated using the position (u'_{fp}, v'_{fp}) for each feature tracked in the current frame, and its corresponding position (x_p, y_p) in the mosaic plane. Let us consider the sum of reprojection errors for all the features tracked in the current frame, which is given as follows:

$$\begin{aligned} \sum_i E_{fi} &= \sum_i |\hat{\mathbf{x}}_{fi} - \mathbf{x}'_{fi}|^2 \\ &= \sum_i \{(\hat{u}_{fi} - u'_{fi})^2 + (\hat{v}_{fi} - v'_{fi})^2\}. \end{aligned} \tag{2.12}$$

Note that $(\hat{u}_{fp}, \hat{v}_{fp})$ is given by transferring (x_p, y_p) by Eq. 2.5 using $\mathbf{M}_f$. As described in Section 2.3.1, this error increases as the estimation error in $\mathbf{M}_f$ becomes larger. Thus, by minimizing this error function with respect to $\mathbf{M}_f$, the estimate for $\mathbf{M}_f$ can be computed. This computation, however, is a non-linear minimization problem, whose solution is

subject to local minima. In order to avoid local minima, first, approximated solution is computed as an initial estimate by linear algorithm [SKYT02]. Then, starting from this initial estimate, camera position (t_{1f}, t_{2f}, t_{3f}) and camera posture (r_{1f}, r_{2f}, r_{3f}) which minimize the error function are estimated by Levenberg-Marquardt algorithm. Finally, the extrinsic camera parameter $\mathbf{M}_f$ is obtained by the the estimated $(t_{1f}, t_{2f}, t_{3f}, r_{1f}, r_{2f}, r_{3f})$ and Eq. (2.4). Note that for the feature position (x_p, y_p) required in the above minimization, the estimated result in the previous iteration is used.

Estimation of feature position on mosaic plane: In this step, the position (x_p, y_p) of each feature p on the mosaic image plane is estimated. Let us consider for each feature p the reprojection error on every frame it appears, as shown in Figure 2.5. The sum of these reprojection errors is given as follows:

$$\begin{aligned} \sum_{i=1}^{f} E_{ip} &= \sum_{i=1}^{f} |\hat{\mathbf{x}}_{ip} - \mathbf{x}'_{ip}|^2 \\ &= \sum_{i=1}^{f} \{(\hat{u}_{ip} - u'_{ip})^2 + (\hat{v}_{ip} - v'_{ip})^2\}. \end{aligned} \tag{2.13}$$

Note that $(\hat{u}_{fp}, \hat{v}_{fp})$ is given by transferring (x_p, y_p) by Eq. 2.5 using $\mathbf{M}_f$. By minimizing this error function with respect to (x_p, y_p), the estimate for (x_p, y_p) is obtained. This non-linear minimization problem is also solved by Levenberg-Marquardt algorithm after the linear estimation of the approximated solution.

Deletion and Addition of features: Selecting good features for tracking is essential in obtaining accurate estimates of camera parameters. In this step, a set of image features to be tracked is automatically updated by testing multiple criteria for each feature. Here, the confidence measure for each feature is defined as an inverse of the variance of its reprojection error in each frame. If an image feature satisfies either of the following conditions, the feature is considered unreliable and is deleted from the set of features to be tracked.

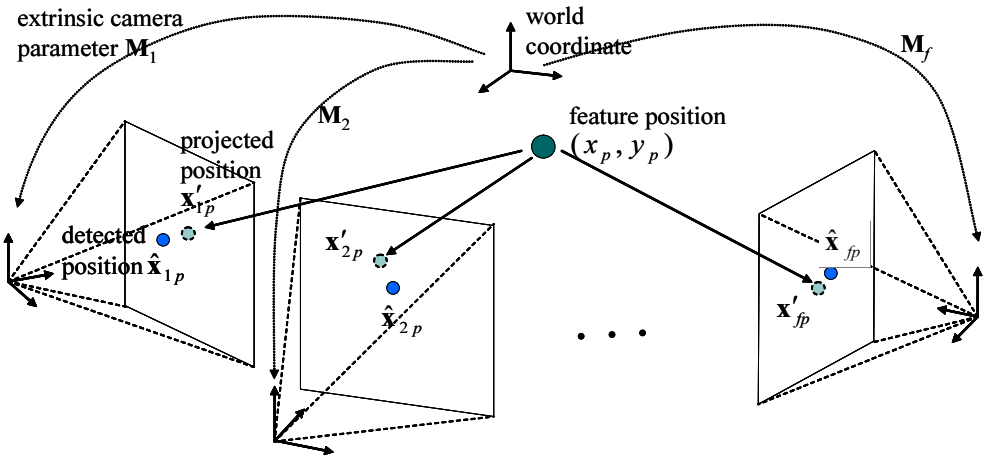


Figure 2.5. Sum of reprojection errors for feature position estimation.

- Confidence measure is under a given threshold.
- Matching error in template matching is more than a given threshold.

On the other hand, if a feature candidate detected by Harris corner detector [HS88] satisfies the all the conditions below, it is added to the set of features to be tracked.

- Confidence measure is over a given threshold.
- Matching error in template matching is less than a given threshold.

By iterating the above steps for each frame, extrinsic camera parameters $\mathbf{M}_f$ and feature positions $\mathbf{S}_p$ on the mosaic image plane are estimated.

2.3.3 Detection of Reappearing Features

Figure 2.6 (a) shows an example of camera path in video mosaicing. Due to this camera motion, most of the image features come into the image, move across toward the end of the image, and disappear. Some features, however, reappear

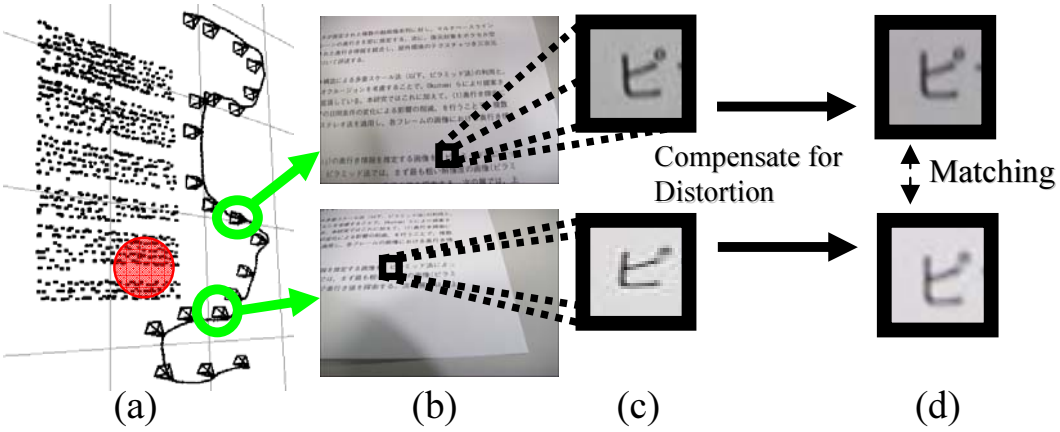


Figure 2.6. Detection of reappearing features. (a) camera path, posture and feature positions on mosaic image plane, (b) sampled frames of input image sequence, (c) templates of the same feature in different images, (d) templates projected to a mosaic image plane.

in the image as shown in Figure 2.6 (b). In this step, these reappearing features are detected, and distinct tracks belonging to the same reappearing feature are linked to form a single long chain. This will give tighter constraints among camera parameters in temporally distinct frames, and thus makes it possible to suppress cumulative errors in the parameter refinement step described later.

Reappearing features are detected by examining the similarity of the patterns among features belonging to distinct tracks. The problem here is that even if two patterns belong to the same feature on the target, they can have different appearance due to perspective distortion, as shown in Figure 2.6. To remove this effect, first, templates of all the features are projected to the mosaic image plane. Then, feature pairs whose distance in 3-D space is less than a given threshold are selected. For each feature, a set of multi-scale templates are generated, and in each scale, the similarity of the templates is evaluated with the normalized cross correlation function. If the correlation of templates in every scale is higher than a certain threshold, the feature pair is regarded as reappearing features, and tracks belonging to each feature are merged into a single track.

2.3.4 Refinement of Estimated Camera Parameters

Since the initial estimation of extrinsic camera parameters described in Section 2.3.2 is an iterative process, its result is subject to cumulative errors. In this process, the cumulative errors is minimized over the whole input images using bundle adjustment [TMHF99].

The estimation error E is given by the sum of reprojection errors as follows:

$$\begin{aligned} E &= \sum_f \sum_p E_{fp} \\ &= \sum_f \sum_p |\hat{\mathbf{x}}_{fp} - \mathbf{x}'_{fp}|^2 \\ &= \sum_f \sum_p \{(\hat{u}_{fp} - u'_{fp})^2 + (\hat{v}_{fp} - v'_{fp})^2\}. \end{aligned} \tag{2.14}$$

Note that $(\hat{u}_{fp}, \hat{v}_{fp})$ is given by transferring (x_p, y_p) by Eq. 2.5 using $\mathbf{M}_f$. E becomes larger as more error is accumulated in initial estimate. Moreover, if the assumption that the focal plane of the camera in the first frame is parallel to the target is violated, E becomes even larger. Thus, by minimizing the error function E with respect to the camera parameters $\mathbf{M}_f$ and the feature positions (x_p, y_p), cumulative errors are minimized, and in case where the focal plane of the camera in the first frame is not parallel to the target, the correct extrinsic camera parameters are estimated. Again, this is a non-linear minimization problem, which is iteratively solved by Levenberg-Marquardt algorithm. Here, the initial parameters for this optimization are given by the process described in 2.3.2.

Note that in this bundle adjustment process, reappearing features detected in the previous step are utilized to chain input images of non-successive frames. Reappearing feature detected in step (c) and its corresponding feature are treated as single feature to compute the error function E. Since these feature chains give strong geometric constraints for extrinsic camera parameters, accurate camera parameters can be acquired.

2.4. Generation of Super-resolved Mosaic Image

In the final process, a super-resolved mosaic image is generated using the estimated extrinsic camera parameters. Here, the iterative back projection algorithm [IP91] is employed to generate a super-resolved mosaic image.

First, an initial estimate of a super-resolved mosaic image $S^{(0)}$ is generated as follows. All the frame images are projected onto the mosaic image plane using Eq. (2.5) with the extrinsic camera parameters $\mathbf{M}_f$ estimated in the previous step, and blended. Here, the blended image is resampled on finer grid so that the size of a single pixel in the input image equals to $n \times n$ pixels ($n > 1$) in the mosaic image. This n is referred to as magnification ratio.

Then, the following process is iterated to estimate the super-resolved mosaic image. The flow of this process is shown in Figure 2.7. Starting with the initial estimate $S^{(0)}$, the imaging process is simulated using geometric transformation and point spread function (PSF) to obtain a set of low resolution images $\{I_f^{(0)}\}$, each of which corresponds to the observed input image $\{I_f\}$. If $S^{(0)}$ is the true super-resolved image, the simulated image $\{I_f^{(0)}\}$ must be identical to $\{I_f\}$. On the other hand, as the estimation error in $S^{(0)}$ becomes large, so does the difference between $\{I_f^{(0)}\}$ and $\{I_f\}$. Thus, the difference images $\{I_f - I_f^{(0)}\}$ are computed, and each value in the difference images is back-projected and added onto its corresponding pixels in $S^{(0)}$. This gives a new estimate of super-resolved mosaic image $S^{(1)}$. The above process is repeated iteratively until the super-resolved image $S^{(i)}$ converges.

It is known that there exists a certain limit on the increase in resolution obtained by the super-resolution technique [BK02]. An experiment on a resolution chart, later described in Section 2.6, has revealed that the resolution obtained by the super-resolution technique is limited by twice the resolution of the original image. Based on this experimental result, the magnification ratio n is set to 2.0 in the proposed method. As for the PSF, Capel [CZ00] has proved that PSF can be approximated by the Gaussian function. In the proposed method, the Gaussian function with standard deviation of 0.7 pixel is employed as the PSF.

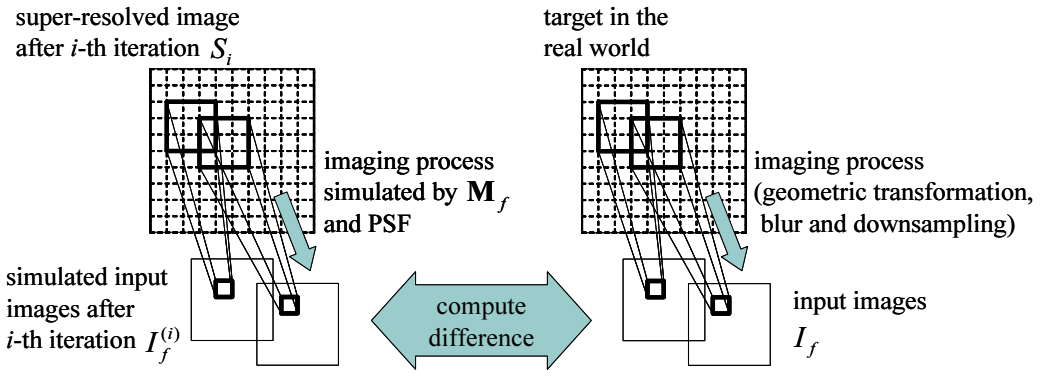


Figure 2.7. Flow of iterative back projection algorithm.

2.5. Prototype System

This section describes a prototype system based on the proposed method.

In video mosaicing, user operation for moving the camera over the target document plays an important role in accomplishing the task of document digitization efficiently and accurately. In this sense, video mosaicing system is, by nature, an interactive system which involves man and machine. This important fact, however, has not been focused in previous works of video mosaicing. In this section, first, we will have further discussion on this topic, and propose an ideal user interface for video mosaicing. Then, we will see how the proposed method and the proposed user interface are implemented in the prototype system.

2.5.1 User Interface for Video Mosaicing

In video mosaicing, the most efficient way to move the camera would be the one that enables to capture a set of input images which satisfies the following conditions:

- Input images exhaustively cover the target document.
- The region captured in each input image is mutually exclusive to one another.

Input images satisfying the former condition can be captured if the user memorizes which part of the target is already captured, and decides where the camera should be moved in the next frame to capture a new region on the target. This, however, becomes extremely hard when the number of the captured images gets larger.

The latter condition can be satisfied if the user moves the camera faster so that overlapping regions among the input images become as small as possible. Faster camera motion, however, degrades the quality of the mosaic image since the error of estimated camera parameters increases due to tracking errors and decrease in the number of frames where each feature is tracked. These two facts imply that there exists an optimal speed for camera motion in terms of accuracy and efficiency. Moving the camera with this optimal speed, however, requires a special training on video mosaicing, which is not a realistic solution.

In order to solve these problems, a novel user interface for video mosaicing is proposed. The proposed user interface is shown in Figure 2.8. During image acquisition, a preview of the mosaic image under construction is rendered in the right side window (Figure 2.8(3)). This preview is updated every frame in real time using captured images and the estimated camera parameters. The region corresponding to the current frame is highlighted in the mosaic image to help the user to recognize which part of the target is currently captured (Figure 2.8(4)). With this preview, the user can easily recognize which part of the document is still left to be captured, and figure out where to move the camera in the subsequent frames, thus can capture input images which exhaustively cover the whole target.

In addition to the preview image, a speed indicator which shows the current speed of the camera is shown at the bottom of the right window (Figure 2.8(5)). In this speed indicator, the current speed of the camera is shown by an arrow mark on a speed gauge, and the gauge is divided into 3 sections : "too slow", "best speed"(optimal), and "too fast". With this speed indicator, the user can move the camera with the optimal speed. Here, the optimal speed was determined by an experiment on synthetic data with ground truth. The detail of this experiment is described in Section 2.6.

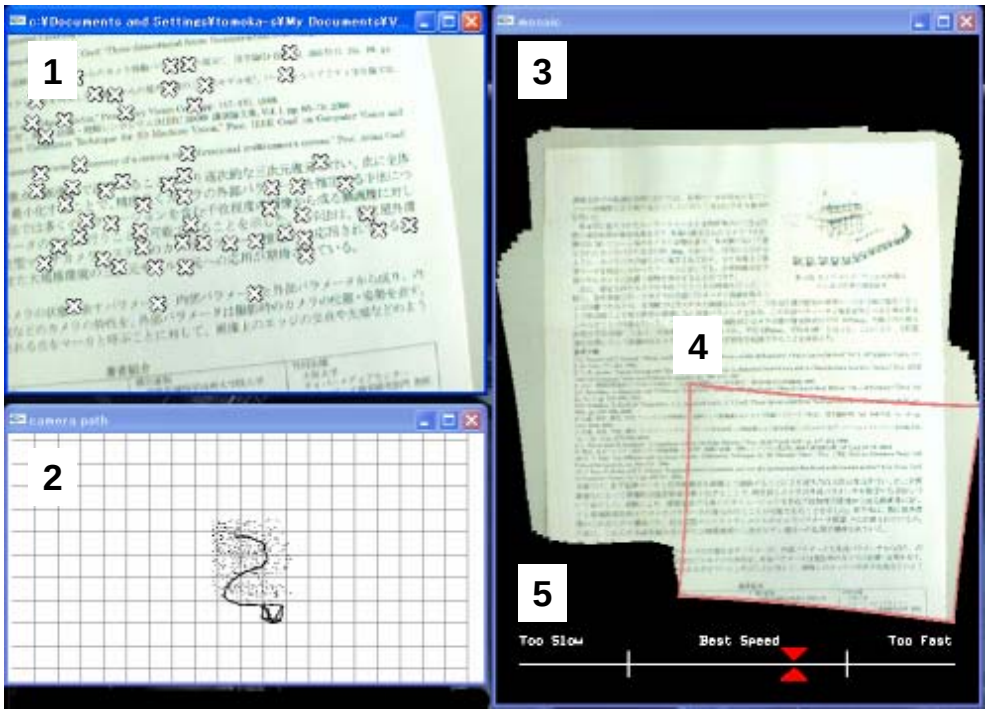


Figure 2.8. User interface for video mosaicing. 1: input image and tracked feature points. 2: estimated camera path and posture. 3: preview of generating a mosaic image. 4: capturing image area on mosaic image. 5: instruction for speed of camera motion.

The user can also take a look at currently captured images and estimated camera path in the left side windows (Figure 2.8(1), (2), respectively), if necessary.

2.5.2 Implementation of Video Mosaicing System

The most important factor of the proposed user interface is that real-time feedback is given to the user. To realize this in the prototype system, the processes of the proposed method is implemented in two stages: real-time stage and off-line stage, as is shown in Figure 2.9.

In the real-time stage (Figure 2.9 (1)), a user captures a target document using a handheld camera. In this stage, initial estimates of the extrinsic camera

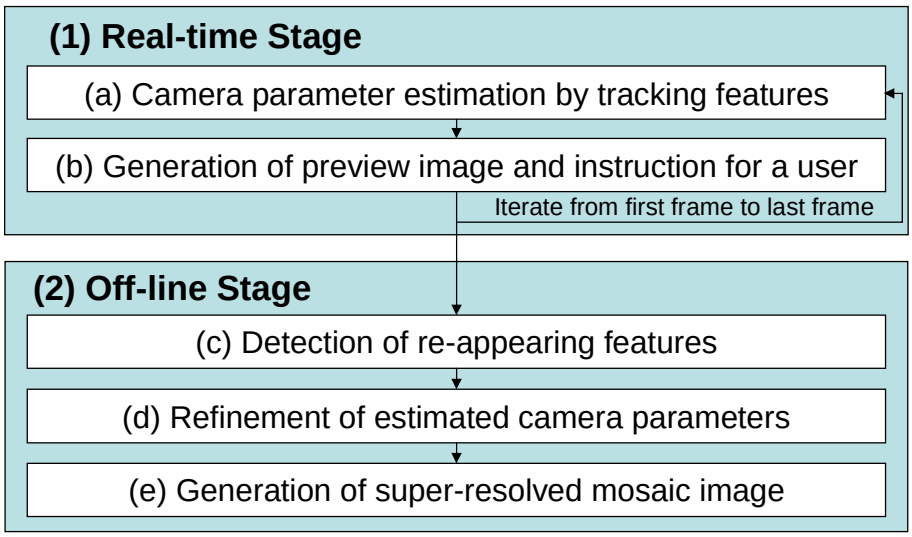


Figure 2.9. Two-stage implementation in prototype system.

parameters for each frame as well as the 2-D position of each image feature are estimated in real time by the process proposed in Section 2.3.2 (a). The system also renders the user interface described above (b). Here, the preview of the generated mosaic image is rendered as follows. For each frame, the captured image is resized into a lower resolution image and stored to texture memory. Every stored texture is warped to the mosaic image plane by texture mapping using Eq. (2.5) with the initial estimate of extrinsic camera parameters. Although cumulative errors are introduced in the initial estimate of the extrinsic camera parameters, their accuracy is sufficient for the purpose of rendering a coarse preview of the mosaic image.

In the off-line stage (Figure 2.9 (2)), reappearing features are detected (c), the estimated viewpoint and 2-D position of each feature are refined (d), and a super-resolved mosaic image without perspective distortion is generated (e).

The overview of the prototype system is shown in Figure 2.10. The system is composed of a laptop PC and a hand-held IEEE1394 CCD camera. The specifications of the system are shown in Table 2.1. Note that intrinsic parameters of the video camera are calibrated by Tsai's method [Tsa86] in advance, and they are fixed throughout the image capturing.

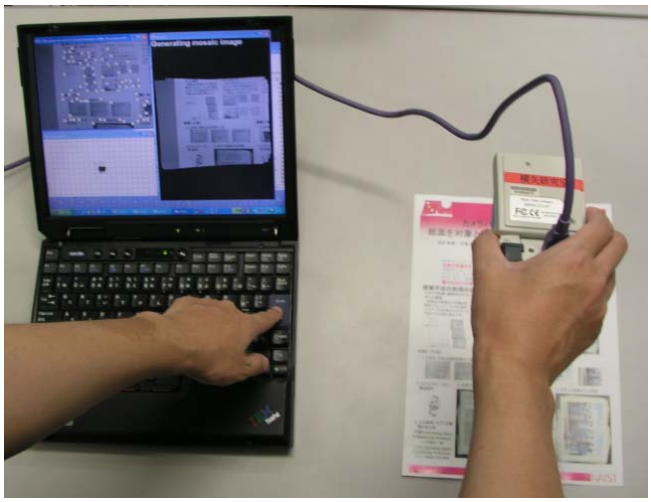


Figure 2.10. Overview of the prototype system.

Experiments performed by this prototype system is described in the following section.

Table 2.1. Specifications of a video mosaicing system for flat target.

Laptop PC	
CPU	Pentium-M 1.6GHz
Memory	1GB

IEEE1394 camera (Aplux C104T)	
Resolution	640×480 pixels
View angle	$31.7° \times 24.1°$
Maximum frame rate	15 frames/sec

2.6. Experiments

Experiments are performed by the prototype video mosaicing system to evaluate the feasibility of the proposed method.

First, an experiment on synthetic data with ground truth is performed to evaluate the accuracy of extrinsic camera parameter estimation, and to determine the optimal speed described in Section 2.5.1. Another experiment is performed on a resolution chart to evaluate the upper limit of resolution obtained by the super-resolution technique. The magnification ratio in the super-resolution process described in Section 2.4 is determined based on this upper limit.

Then, experiments are performed on three kinds of flat printed paper to generate super-resolved mosaic images. The first target is a printed A4 size document (sequence 1). The second target is a printed photograph on an A4 size paper (sequence 2). In these experiments, the distortion in the resultant mosaic images is evaluated quantitatively to see if the perspective distortion is removed by the proposed method. The last experiment is performed on a picture scroll, approximately 440 cm long (sequence 3).

2.6.1 Experiment on Synthetic Data

The purpose of this experiment is to see how the speed of the camera affects the accuracy of the camera parameter estimation. In this experiment, extrinsic camera parameters are estimated by applying the proposed method to synthetic data for various camera speeds, and the accuracy of the camera parameter estimation is evaluated by comparing the estimated parameters with the ground truth. By analyzing the result, the optimal speed for moving the camera is determined, and is utilized in rendering the user interface proposed in Section 2.5.1. The detail of this experiment is given below.

First, in order to generate synthetic feature tracks, a simulation is carried out with a virtual camera that has the same parameters as the camera shown in Table 2.1 and feature points randomly scattered on a virtual mosaic image plane as shown in Figure 2.11. In this simulation, the virtual camera is moved over the virtual mosaic image plane at constant speed. In each frame, feature

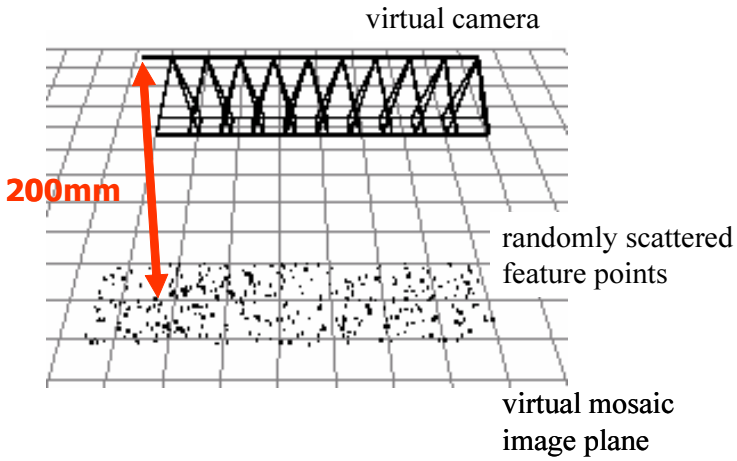


Figure 2.11. Synthetic data generated for simulation.

points on the mosaic plane are projected onto the virtual camera image. To simulate errors in feature detection and tracking, Gaussian noise is added to the projected positions. Finally, by quantitizing the positions to integer, synthetic feature tracks are obtained.

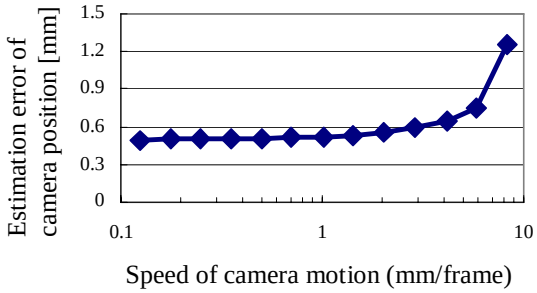
These synthetic data are generated for different camera speeds. For each camera speed, camera parameters are estimated by the proposed method and the errors in camera position and posture is evaluated by comparing the estimated parameters with the ground truth.

The configuration of this simulation is shown in Table 2.2. All these parameters are simulation of typical configuration in real experiments.

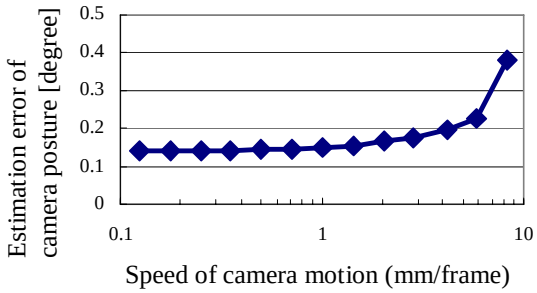
The estimation errors in camera position and posture for different camera speeds are shown in Figure 2.12 (a), (b), respectively. We can see that the estimation error monotonously increases as the speed of the camera gets

Table 2.2. Configuration of simulation.

Distance between camera and mosaic plane	200mm
Average number of detected feature points	90
Average reprojection error	0.8 pixel



(a) error in position



(b) error in posture

Figure 2.12. Evaluation of camera parameter estimation (simulation).

faster. More specifically, both errors for camera position and posture drastically increase when the camera speed is faster than 4 mm/frame. On the other hand, the errors are almost constant when the camera speed is slower than 2 mm/frame. According to these results, the optimal speed for the camera is determined to 2 mm/frame.

2.6.2 Experiment on Resolution Chart

In order to determine the upper limit of resolution obtained by super-resolution technique, an experiment is performed on a resolution chart. A resolution chart is a specially designed chart on which stripes of black and white lines with different intervals are printed. Blocks of stripes are aligned in horizontal or vertical direction such that each block of stripes has half the interval of that of the previous block. By analyzing if the stripes are resolved in the captured image, its resolution can be quantitatively measured.

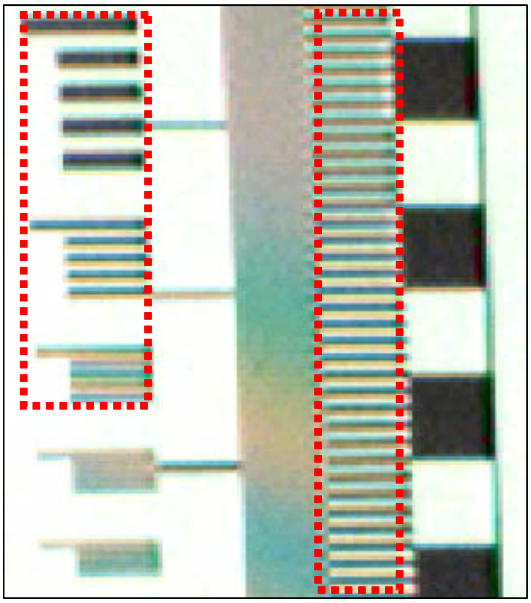
A sampled input image out of 50 captured images and its super-resolved image are shown in Figure 2.13 (a) and (b), respectively. Here, the magnification ratio n in the super-resolution process is tentatively set to 5.0. In each image, regions highlighted in red show the blocks where alternate black and white lines are observed, i.e. stripes which are resolved. On the hand, those without highlight are the stripes which cannot be resolved, resulting in solid gray regions. By comparing the blocks in the super-resolved image (Figure 2.13 (b)) with the input image (Figure 2.13 (a)), we can see that the stripes having half the interval of that in the input image has been resolved, but the stripes with shorter intervals remain unresolved. Thus, it can be confirmed that the resolution obtained by super-resolution process is limited by twice the resolution of the input image. Based on this experimental result, the magnification ratio n in the super-resolution process described in Section 2.4 is set to 2.0 in the prototype system.

2.6.3 Experiment on Document (Sequence 1)

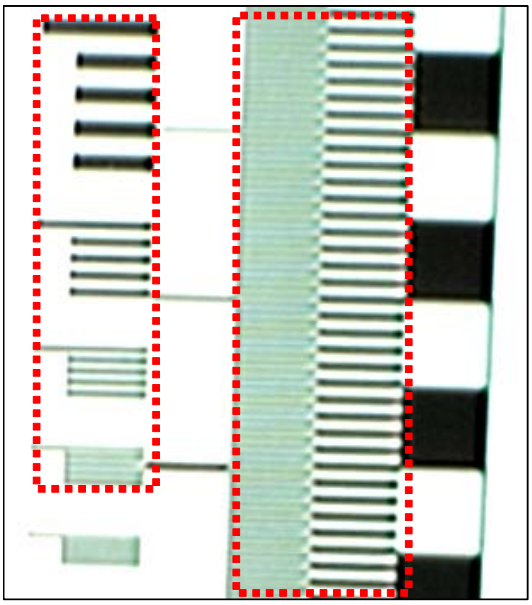
In this experiment, we will see how the proposed method successfully generates a super-resolved mosaic image for a target in the real world.

As shown in Figure 2.14, the target paper of document is captured as 640×480 images of 120 frames using the prototype system. Note that in every frame, the camera has been tilted backward against the target document, which has caused perspective distortion in every input image. Cross marks in Figure 2.15 indicate feature points which are automatically detected and tracked in the real-time stage. In this experiment, 111 image feature points are tracked per frame on average.

The result of the real-time stage is the initial estimate of extrinsic camera parameters and the feature positions on the mosaic image plane. The left figure in Figure 2.16 (a) shows this initial estimate. Here, the curved line shows the estimated camera path and pyramids show the camera postures in every 10 frames. The points show the estimated feature positions on the mosaic image plane. The right figure in Figure 2.16 (a) shows the the mosaic image generated by this initial estimate. In the mosaic image, misalignment of images is observed, especially at the upper-left part of the document contour, which



(a) Input original image



(b) Super-resolved image

Figure 2.13. Comparison of input image and super-resolved mosaic image (Resolution chart).

is depicted with red circle. This misalignment is due to two types of errors in the estimated parameters: cumulative errors introduced in the iterative process, and errors caused by violating the assumption that the focal plane of the camera in the first frame is parallel to the target.

In order to minimize these errors and to correct the misalignment in the mosaic image, reappearing features are detected and the estimated parameters are refined in the off-line stage. As described in Section 2.3.4, the refinement of parameters is a non-linear minimization problem, which is iteratively solved. Figure 2.16 (b) and (c) show the extrinsic camera parameters and feature positions, along with the mosaic image after 50th and 400th iteration, respectively. Red points on the mosaic image plane show the detected reappearing features. The total number of detected reappearing features in this experiment is 93. As can be seen, the extrinsic camera parameters are gradually refined and the misalignment in the mosaic image is corrected as the refinement process is iterated.

The refinement process converged after 838 iterations. The average reprojection error of the features before and after refinement are 0.83 pixel and 0.67 pixel, respectively. Figure 2.17 illustrates the acquired extrinsic camera parameters and the feature positions on the mosaic image plane after the refinement process. In Figure 2.17, we can see that the camera orientations are tilted backward against the target. This coincides with the configuration of the camera in this experiment.

Finally, a super-resolved mosaic image is generated by 3 iterations of back-projection as shown in Figure 2.18. The size of the image is 2533×2920. We can confirm that perspective distortion has been removed in the final mosaic image. The advantage of the proposed method over homography based methods is obvious when this result is compared with the mosaic image shown in Figure 1.4, which was generated by a homography based method using the same input images and the same feature tracks as this experiment. Quantitative evaluation of the distortion in the mosaic image will be shown in Section 2.6.6.

Close shots of an input image and the corresponding region in super-resolved mosaic image are compared in Figure 2.19. As can be seen, degraded texts in the input image are restored in the super-resolved mosaic image.

The performance of the prototype system for this sequence is as follows: 9 fps for image acquisition and initial parameter estimation, 1 second for detecting re-appearing features in 120 frames, 27 seconds for camera parameter refinement, and 240 seconds for generation of super-resolved mosaic image.

2.6.4 Experiment on Photograph (Sequence 2)

In this experiment, a photograph printed on a flat paper is chosen as a target. Since photographs gives less image features compared to documents full of texts, degradation in the accuracy of extrinsic camera parameter estimation and the quality of the resultant mosaic image is expected. The purpose of this experiment is to evaluate this degradation by comparing the result with that of the previous experiment on a document.

As shown in Figure 2.20, the target photograph is captured as 640×480 images of 150 frames. Cross marks in Figure 2.21 indicate feature points which are automatically detected and tracked in the real-time stage. In this experiment, the average number of tracked features is 62 points per frame, which is substantially less than that obtained for a document in the previous experiment.

Initial estimate of extrinsic camera parameters and the mosaic image generated by this initial estimate are shown in Figure 2.22 (a). Due to error in the initial estimate, misalignment of images is observed in the mosaic image, especially at the upper-left part of the contour, which is depicted with red circle. This initial estimate is refined in the off-line stage. Figure 2.22 (b) and (c) show the extrinsic camera parameters and the mosaic image after 10th and 100th iteration, respectively. Red points on the mosaic image plane show the detected reappearing features. The total number of detected reappearing features in this experiment is 134. As can be seen, the extrinsic camera parameters are gradually refined and the misalignment in the mosaic image is corrected as the refinement process is iterated.

The refinement process converged after 213 iterations. The average reprojection errors of the features before and after refinement are 0.94 pixel and 0.71 pixel, respectively. Figure 2.23 shows the extrinsic camera parameters and the feature positions on the mosaic image plane after refinement.

The super-resolved mosaic image after 3 iterations is shown in Figure 2.24. The size of the image is 2019×2758. An input image and the super-resolved mosaic image are compared in Figure 2.25. As can be seen, the detail structure of the glasses and the face has been restored in the super-resolved mosaic image.

Although degradation in the accuracy of extrinsic camera parameter estimation and the quality of the resultant mosaic image has been expected due to the less number of image features, there seems to be no noticeable difference in the quality of results obtained for this experiment on a photograph and for the previous experiment on a document. A further discussion on this comparison will be given based on quantitative evaluation in Section 2.6.6.

The performance of the system for this sequence is as follows: 9 fps for image acquisition and initial parameter estimation, 1 second for detecting reappearing features in 120 frames, 10 seconds for camera parameter refinement, and 220 seconds for super-resolution. Note that the computational time for camera parameter refinement is much shorter than that for the experiment on a document. This is due to the difference of camera configuration in the first frame. In this experiment, the camera posture in the first frame was set so as to satisfy the parallel assumption described in Section 2.3.2. This gives a better initial estimate, and thus helps the error function to be minimized more quickly in the refinement process.

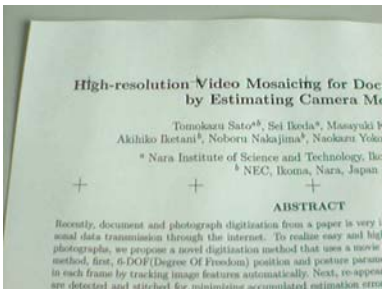
2.6.5 Experiment on Picture Scroll (Sequence 3)

One of the advantages of video mosaicing is that the field of view of the resultant mosaic image is unlimited, as far as one can afford the memory required. This is a huge breakthrough when it is compared to an ordinary camera whose field of view is limited by the lens and the focal length. In order to demonstrate this advantage of unlimited field of view, a challenge has been made to generate a mosaic image for a picture scroll, approximately 440 cm long.

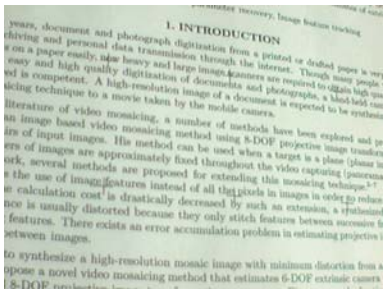
As shown in Figure 2.26, the target scroll is captured as 640×480 images of 500 frames. Tracked feature points are depicted with cross marks in Figure 2.27. Figure 2.28 illustrates the estimated extrinsic camera parameters and the feature positions on the mosaic image plane after the refinement process. The curved line shows the estimated camera path and pyramids show

the camera postures in every 50 frames.

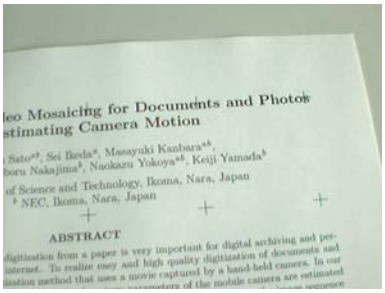
The super-resolved mosaic image is shown in Figure 2.29. The resolution of the image is 12800 $\times$ 1825 pixels. Note that this image resolution was determined by the limitation of the memory resource in the prototype system. Despite the huge target and the long input sequence, the system successfully generates a mosaic image without perspective distortion. It should be noted that in homography based method, long input sequence would result in severe perspective distortion, since the perspective distortion tends to increase toward the end of the mosaic image. This result also shows the advantage of the proposed method over homography based methods.



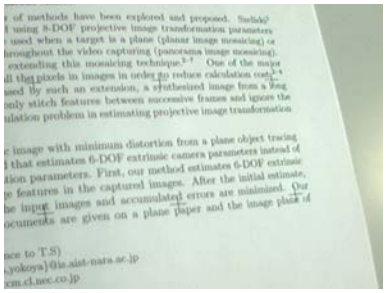
1st frame



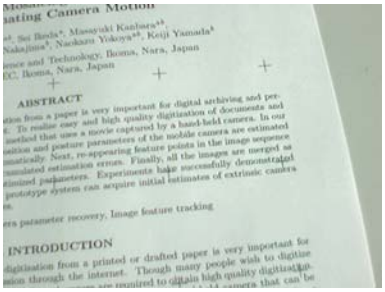
68th frame



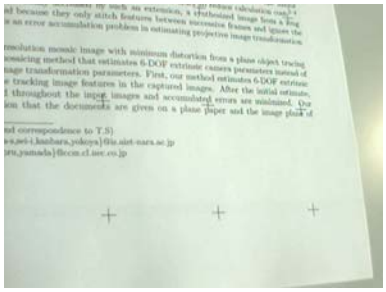
17th frame



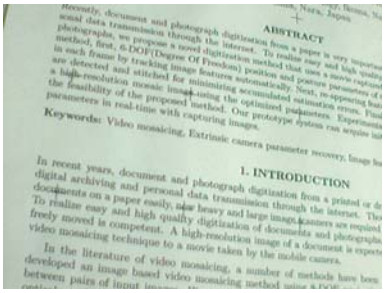
85th frame



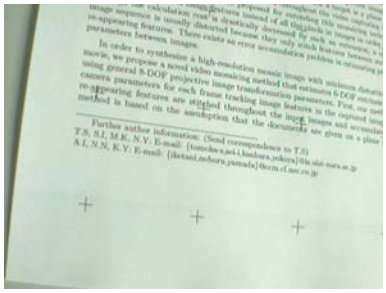
34th frame



102th frame

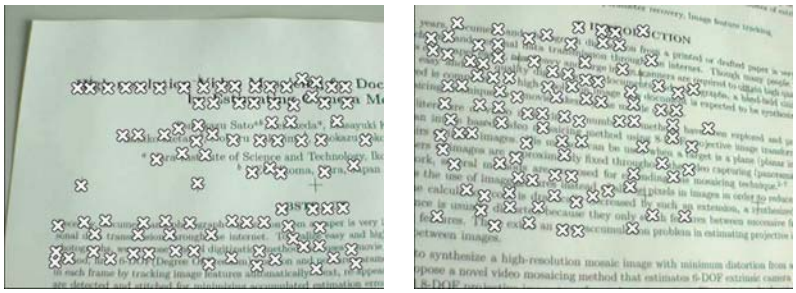


51th frame

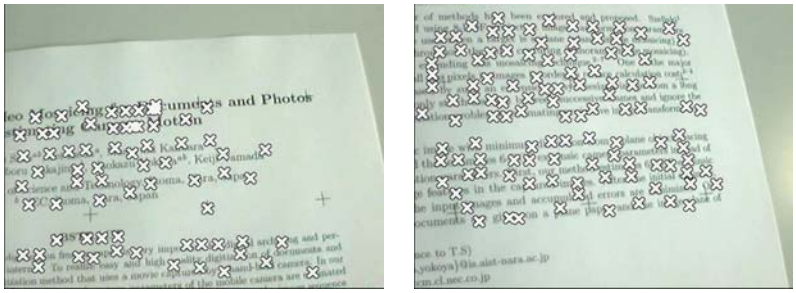


120th frame

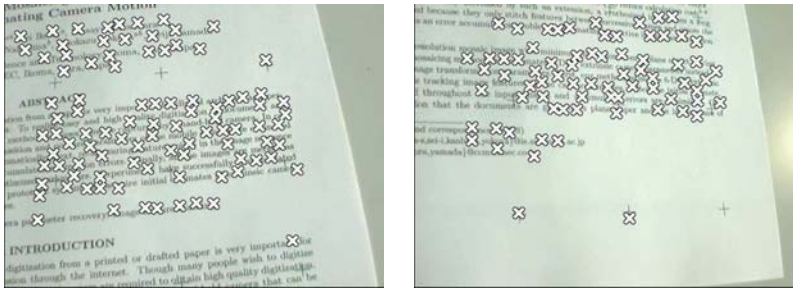
Figure 2.14. Sampled frames of input image sequence (Sequence 1).



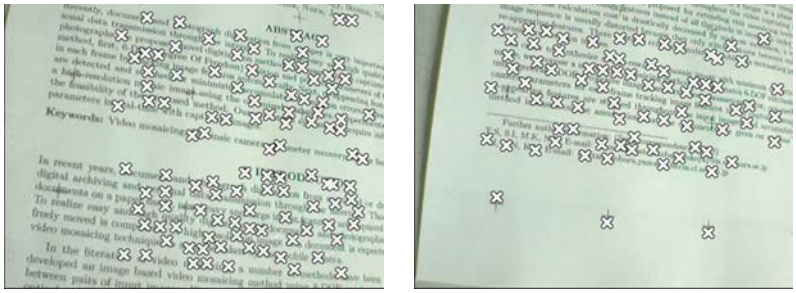
1st frame 68th frame



17th frame 85th frame

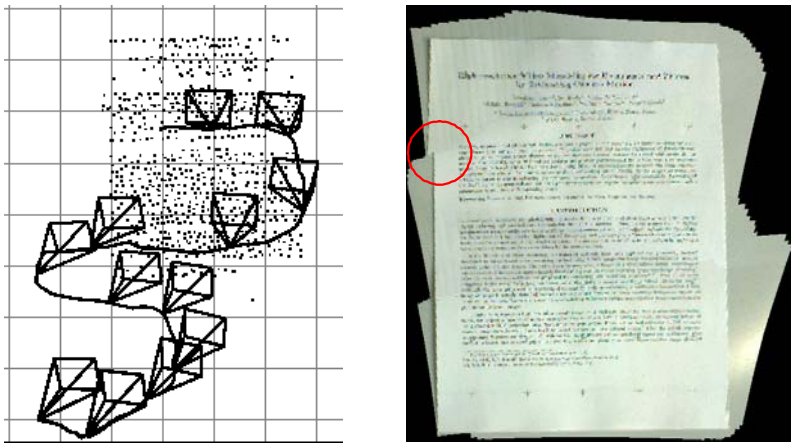


34th frame 102th frame

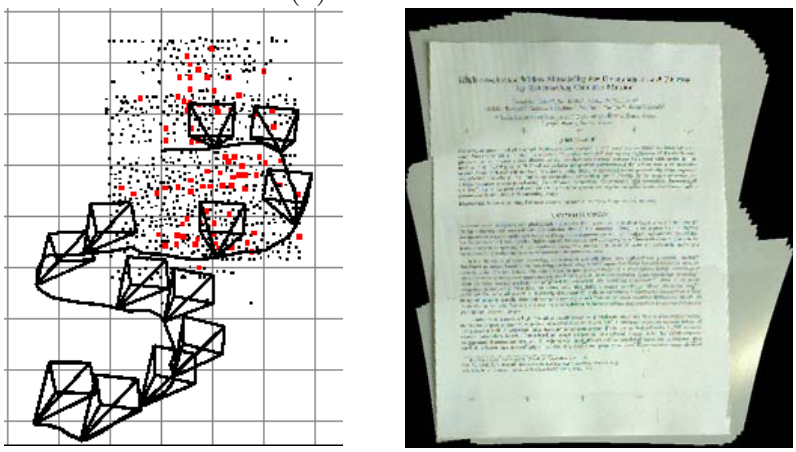


51th frame 120th frame

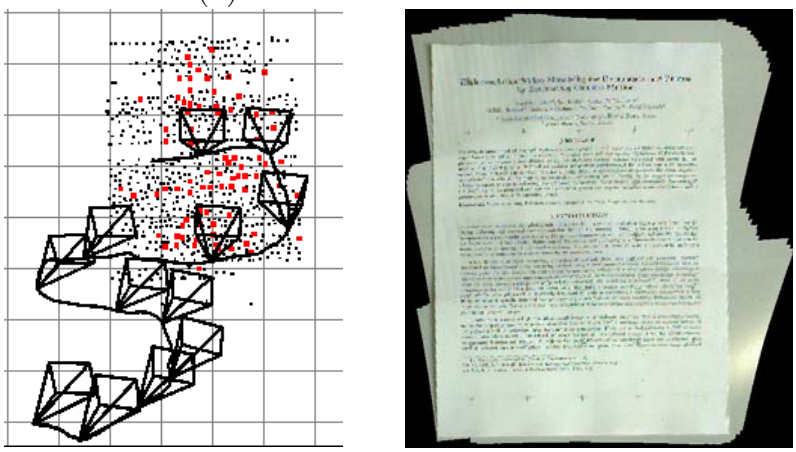
Figure 2.15. Tracked features in input image sequence (Sequence 1).



(a) initial result

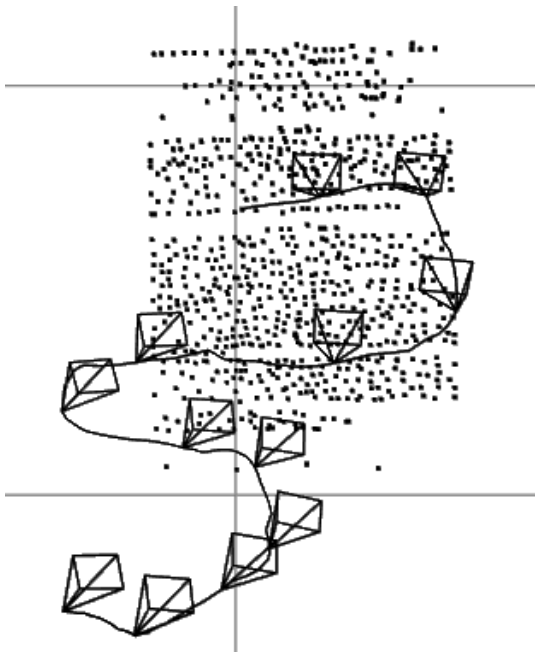


(b) result after 50th iteration

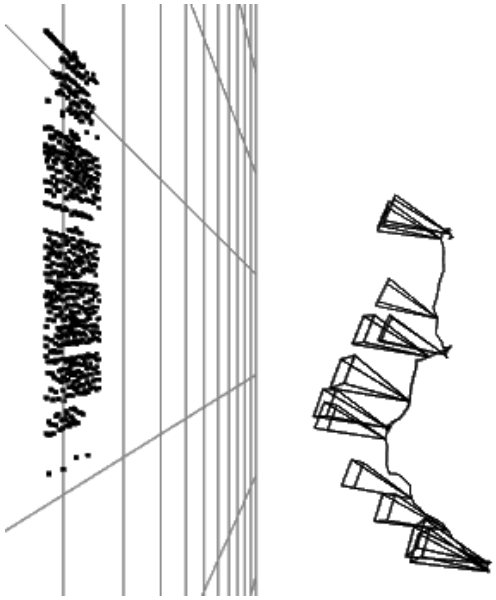


(c) result after 400 iteration

Figure 2.16. Extrinsic camera parameters and mosaic image under refinement (Sequence 1).

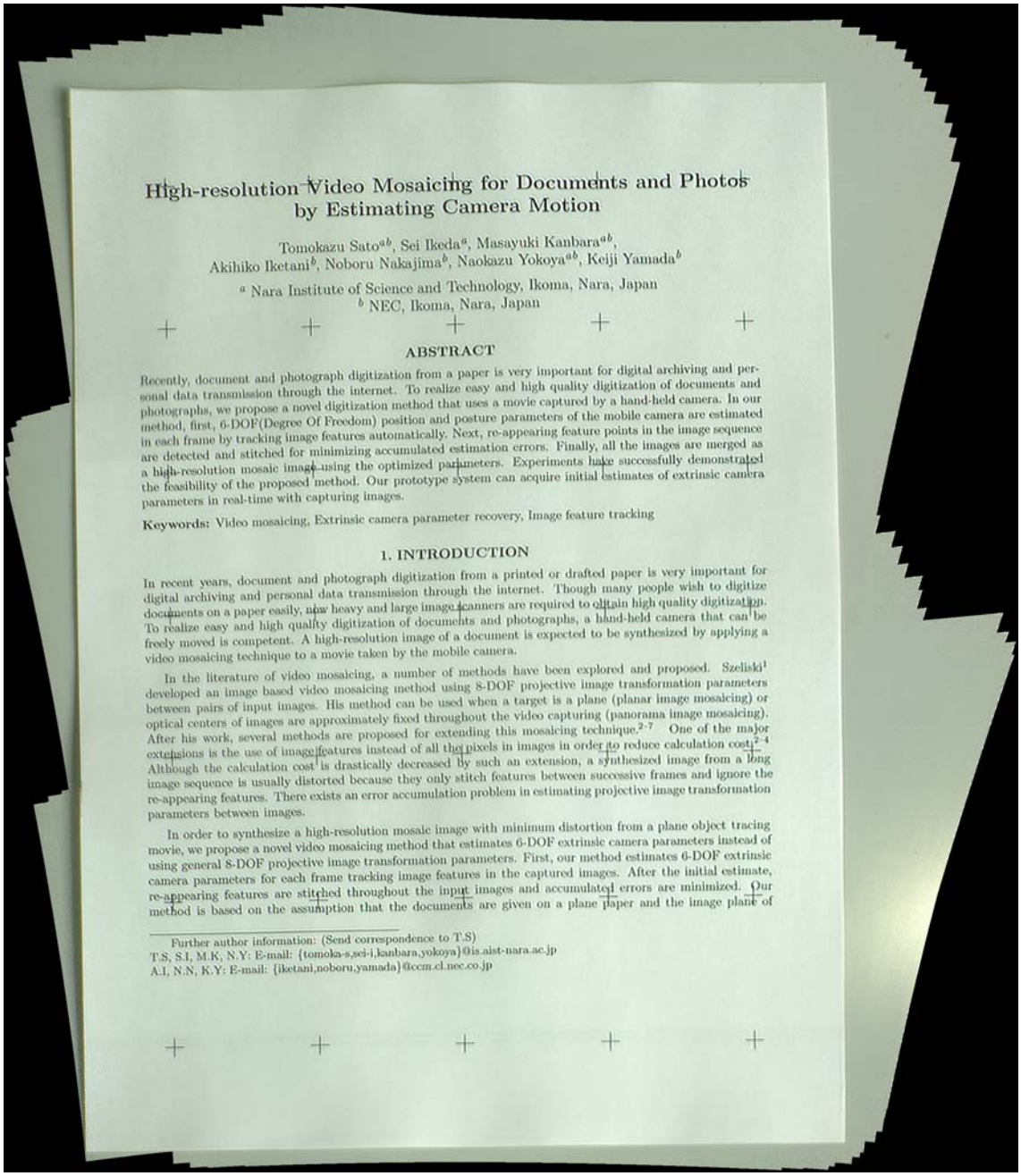


(a) top view



(b) side view

Figure 2.17. Estimated extrinsic camera parameters and feature positions after refinement (Sequence 1).



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

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

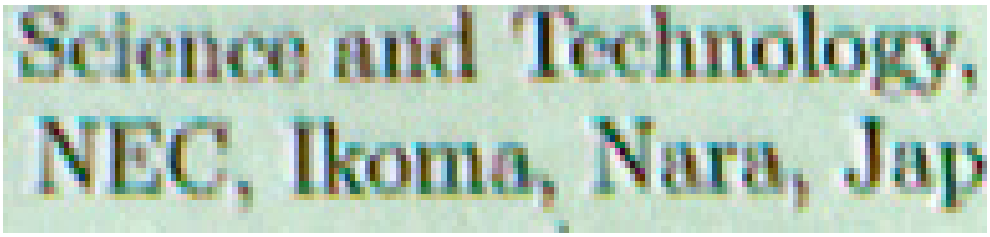
In recent years, document and photograph digitization from a printed or drafted paper is very important for digital archiving and personal data transmission through the internet. Though many people wish to digitize documents on a paper easily, now heavy and large image scanners are required to obtain 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.[1,2-4] Although the calculation cost[1] 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.

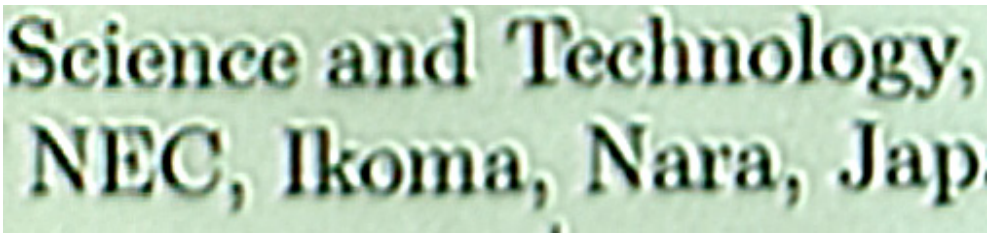
In order to synthesize a high-resolution mosaic image with minimum distortion from a plane object tracing movie, we propose a novel video mosaicing method that estimates 6-DOF extrinsic camera parameters instead of using general 8-DOF projective image transformation parameters. First, our method estimates 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 2.18. Generated super-resolved mosaic image (Sequence 1).



(a) Input original image



(b) Super-resolved image

Figure 2.19. Comparison of input image and super-resolved mosaic image (Sequence 1).

1st frame

84th frame

21th frame

105th frame

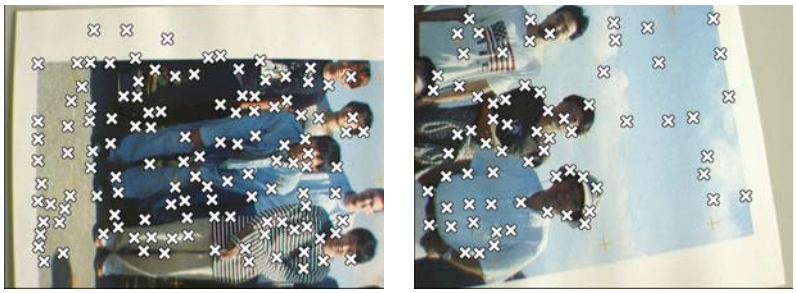
42th frame

126th frame

63th frame

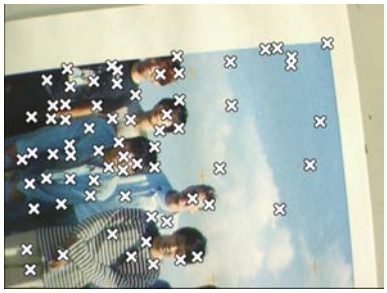
150th frame

Figure 2.20. Sampled frames of input image sequence (Sequence 2).

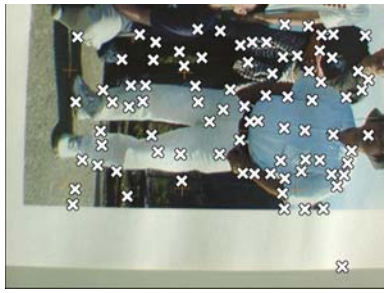


1st frame

84th frame



21th frame

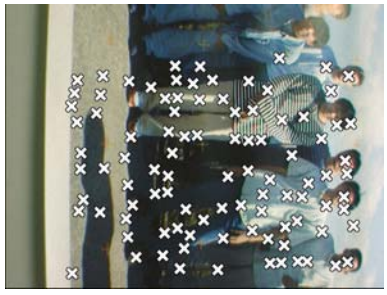


105th frame

42th frame

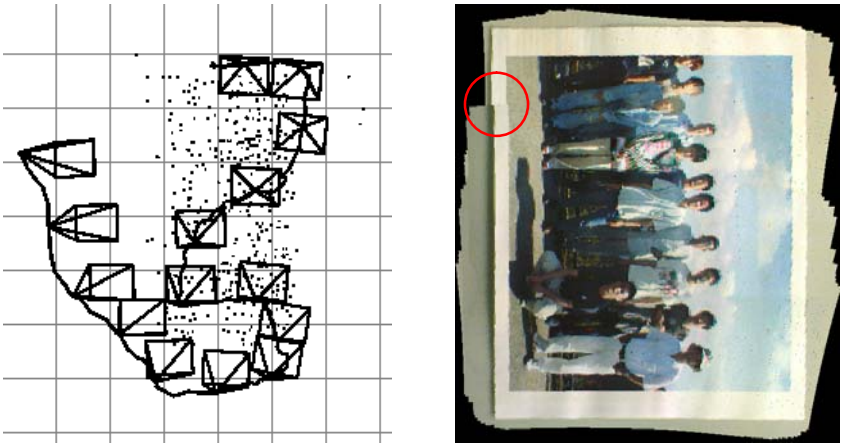
126th frame

63th frame

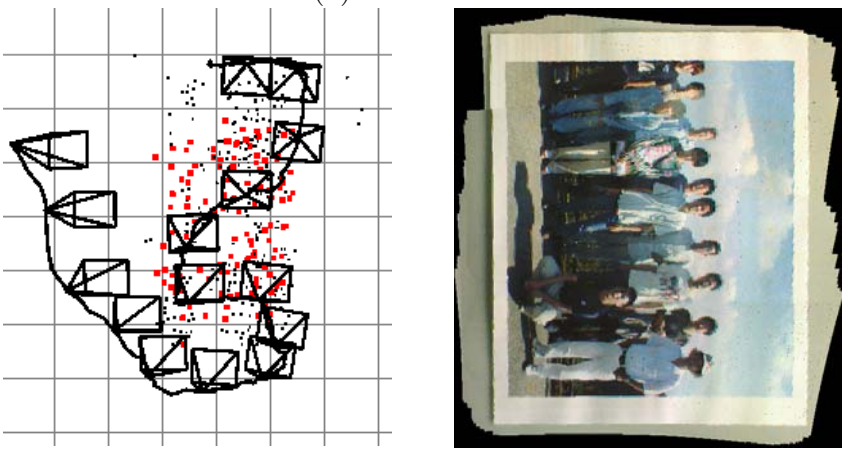


150th frame

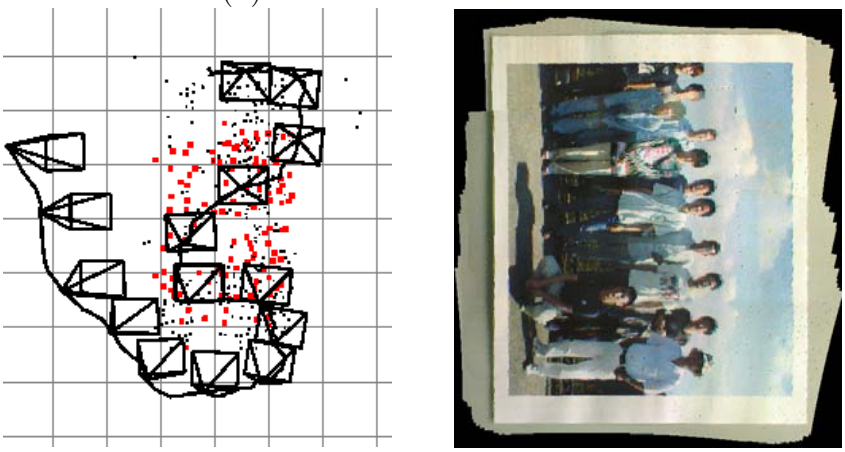
Figure 2.21. Tracked features in input image sequence (Sequence 2).



(a) initial result

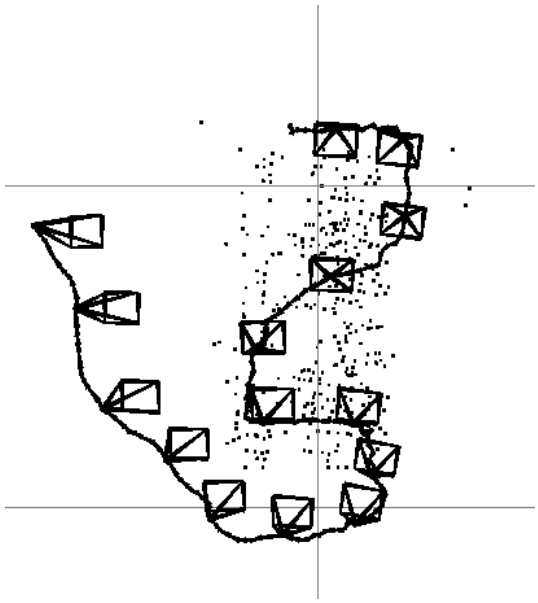


(b) result after 10th iteration

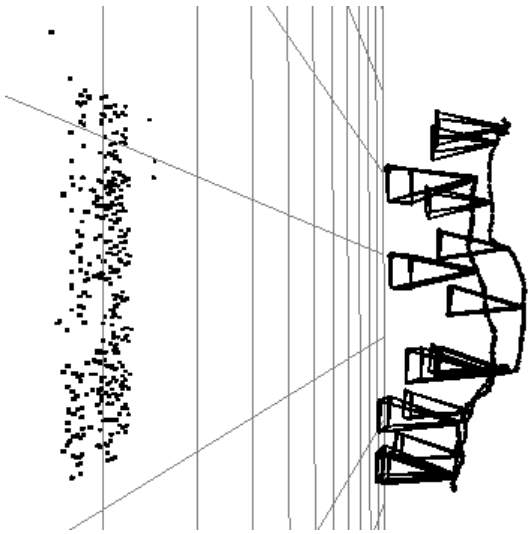


(c) result after 100 iteration

Figure 2.22. Extrinsic camera parameters and mosaic image under refinement (Sequence 2).



(a) top view



(b) side view

Figure 2.23. Estimated extrinsic camera parameters and feature positions after refinement (Sequence 2).

Figure 2.24. Generated super-resolved mosaic image (Sequence 2).

(a) Input original image

(b) Super-resolved image

Figure 2.25. Comparison of input image and super-resolved mosaic image (Sequence 2).

1st frame

280th frame

70th frame

350th frame

140th frame

420th frame

210th frame

490th frame

Figure 2.26. Sampled frames of input image sequence (Sequence 3).

Figure 2.27. Tracked features in input image sequence (Sequence 3).

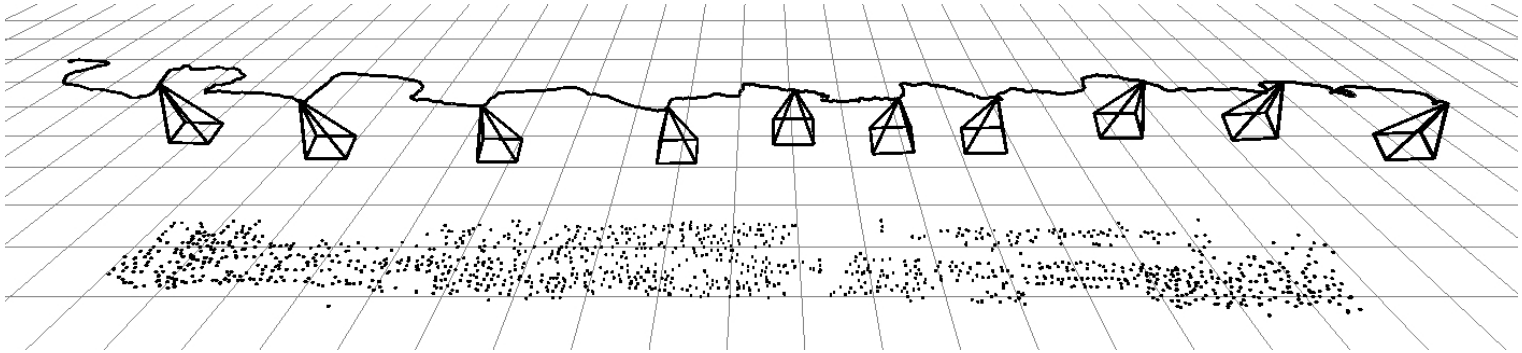


Figure 2.28. Estimated extrinsic camera parameters and feature positions (Sequence 3).

Figure 2.29. Generated mosaic image (Sequence 3).

2.6.6 Quantitative Evaluation of Distortion

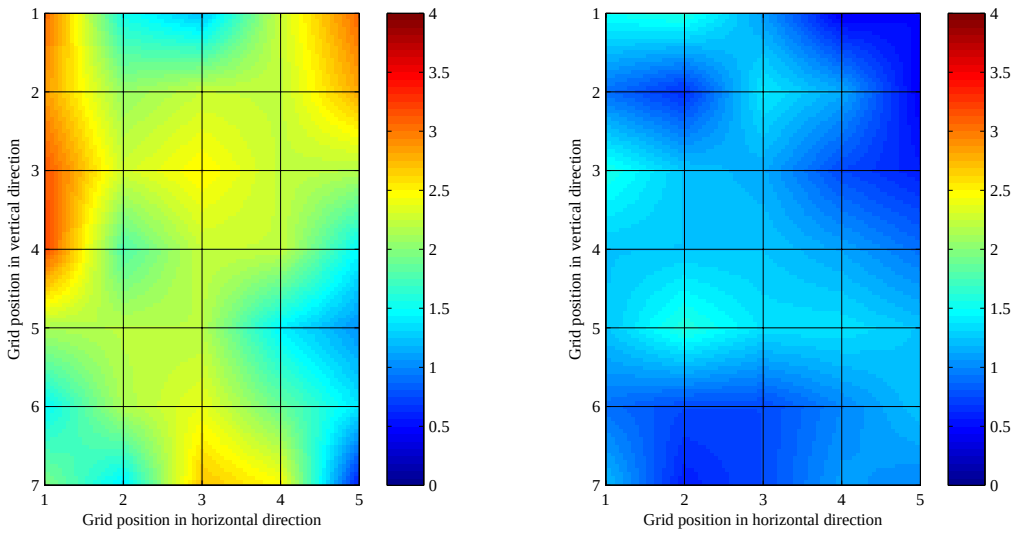
The feasibility of the proposed method is further evaluated by analyzing the distortion present in resultant mosaic images. In the experiments on a document and a photograph described in Sections 2.6.3 and 2.6.4, plus marks (+) have been printed on the target papers at every 40mm grid positions. By measuring the distances between every adjacent grid positions, the distortions in the generated mosaic images are quantitatively evaluated.

First, the positions of the plus marks on the generated mosaic image are acquired manually. Then, the distances between adjacent plus marks are computed in the unit of pixel. The average, maximum, minimum and standard deviation of the distances are shown in Table 2.3. The percentage of each value against the average distance is also shown in parenthesis. Here, the standard deviation can be considered as the average distortion in the mosaic image, and it was only 0.68% and 0.62 for a document and a photograph, respectively. These results confirm that perspective distortion in the mosaic images has been successfully removed by the proposed method. It should also be noted that, despite the less number of image features obtained in a photograph, the distortion in the mosaic image is comparable, or even smaller than that for the document.

Figure 2.30 (a) and (b) show the distribution of distortion on the mosaic image for the document and photograph, respectively. Here, each grid point is given the average of distortions measured between every adjacent grid point. Although the distortion is approximately constant in most of the grid points, there are some grid points with exceptionally large distortion. These grid points are found in the area where temporally distinct frames meet. Further

Table 2.3. Distances of adjacent grid points on generated mosaic images [pixels(percentage from average)]

target	average	maximum	minimum	standard deviation
Document	359.4(100.0)	364.3(101.2)	354.0(98.4)	2.46(0.68)
Photograph	328.6(100.0)	333.0(101.3)	325.0(98.9)	2.06(0.62)



(a) Document (Sequence 1) (b) Photograph (Sequence 2)

Figure 2.30. Distribution of distortion on mosaic image [pixels].

analysis on the experimental results reveals that some of the reappearing features have been left undetected in these areas, which reduces the effect of the parameter refinement process. It is expected that distortions on these grid points can be further reduced by improving reappearing feature detection.

2.7. Conclusion

In this chapter, a perspective distortion-free video mosaicing method for flat documents is described. In the proposed method, extrinsic camera parameters, instead of homographies, are estimated for each frame by applying structure from motion technique to the captured video. Using estimated extrinsic camera parameters, the method dewarps all the frame images and synthesizes them on a virtual rectified image plane to generate a super-resolved mosaic image without perspective distortion.

A novel user interface for video mosaicing is also proposed. In this user

interface, a preview of the mosaic image under construction is rendered in real time. In addition to the preview, a speed indicator to guide the user to move the camera with optimal speed is shown. With this user interface, camera motion which is not only efficient, but also gives accurate camera parameter estimation can be easily achieved by the user.

The proposed method has been implemented on the prototype system by two-stage implementation, along with the proposed user interface. Experiments on flat documents are performed using the prototype system. In each experiment, the mosaic image has been proved to be distortion-free by quantitative analysis on distortion. One of the advantages of video mosaicing is that the field of view of the resultant mosaic image is unlimited. This is a huge breakthrough when it is compared to an ordinary camera whose field of view is limited by the lens and the focal length. This advantage has also been demonstrated in an experiment on a picture scroll approximately 440 cm long.

The limitation of the method is that it can only be applied to flat documents. In the real world, however, there are many documents with curved surface, e.g. thick bound book. In the following chapter, the video mosaicing method is extended to deal with documents with curved surface.

Chapter 3

Video Mosaicing for Curved Document

3.1. Introduction

In the previous section, video mosaicing for flat document has been described. In this chapter, we extend the method to deal with documents with curved surface. The goal is to generate mosaic images of virtually flattened pages for documents with curved surface.

In case of curved surface, geometric distortion induced by the curvature of the target will be present in each input image. In order to generate mosaic images of virtually flattened pages, this curvature distortion in input images has to be corrected before blending the input images on a mosaic image plane. In this method, structure from motion technique is employed to the input images to estimate extrinsic camera parameters and feature positions, as is in the method for flat documents. This time, however, 3-D position is estimated for each feature, since features no longer lie on 2-D plane but on 3-D surface. After estimating the shape of the curved surface from the 3-D feature positions, the method dewarps all the frame images and synthesizes them on a mosaic image plane to generate a virtually flattened image of the curved surface.

Two assumptions are made in this method. One is that the shape of the target is a *generalized cylinder*, which is a surface swept by a straight line moving along an arbitrary curve. The other is that intrinsic camera parameters

are known in advance, and remain fixed throughout image capturing, as is assumed in the method for flat targets.

In the following sections, first, the overview of the methods is given(Section 3.2), and then we will see how the previous method for flat targets is extended for curved targets (Section 3.3). After describing the prototype system based on this method (Section 3.4), experimental results on curved documents is shown (Section 3.5). Finally, the conclusion of this chapter is given (Section 3.6).

3.2. Overview of the Method

The flow of the proposed method is given in Figure 3.1. As can be seen, the whole process is composed of three processes: initial 3-D reconstruction by feature tracking (A), parameter refinement and target shape estimation (B) and mosaic image generation (C).

In initial 3-D reconstruction process (A), estimation of extrinsic camera parameters along with 3-D feature positions on the target is carried out by tracking image features in the input video. In parameter refinement and target shape estimation process (B), reappearing features are detected (a), and the initial estimates of camera parameters and 3-D feature positions are refined by global optimization (b). Then, surface parameters are estimated by fitting a parameterized 3-D surface to estimated 3-D feature points (c). These three processes are iterated until convergence. Finally, a geometric distortion-free mosaic image is generated. A post-process is applied to remove the shade on the mosaic image (C).

The major differences between this method and the previous method for flat targets can be summarized as follows:

- Extension to deal with 3-D feature positions instead of 2-D positions (Figure 3.1(A),(a) and (b)).
- Processes specifically designed for curved surface (Figure 3.1(c) and (C)).

In the following sections, each process in the proposed method is described, focusing on the extensions described above.

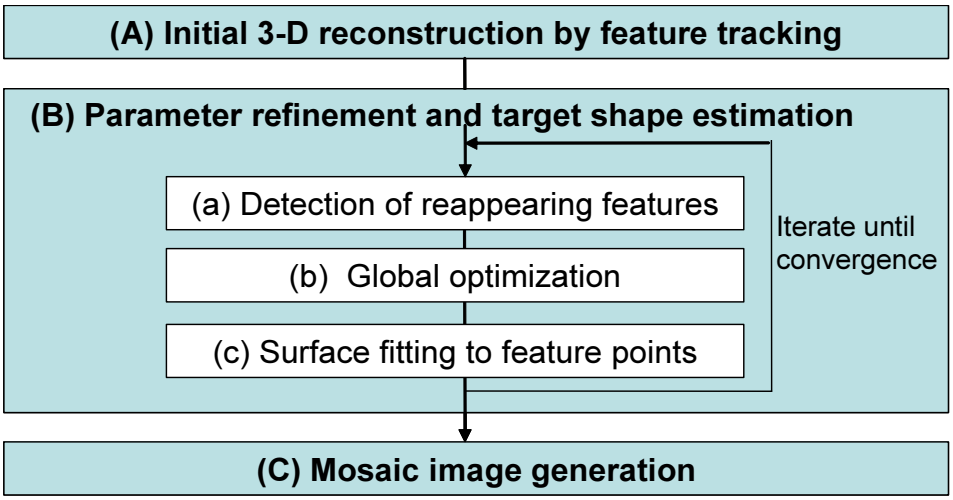


Figure 3.1. Flow diagram of video mosaicing for curved target.

3.3. Extension for Curved Target

3.3.1 Initial 3-D Reconstruction by Feature Tracking

This process is almost identical to the initial estimation of extrinsic camera parameter for flat targets described in Section 2.3.2, except that 3-D position is estimated for each feature and that standard extrinsic camera parameter $\mathbf{Mfull_f}$ is estimated for each frame.

Let us recall the general definition of extrinsic camera parameter in Eq. (2.1) and (2.2):

$$\mathbf{Mfull}_f = \begin{pmatrix} c_1c_3 + s_1s_2s_3 & s_1c_2 & -c_1s_3 + s_1s_2c_3 & t_{1f} \\ -s_1c_3 + c_1s_2s_3 & c_1c_2 & s_1s_3 + c_1s_2c_3 & t_{2f} \\ c_2s_3 & -s_2 & c_2c_3 & t_{3f} \end{pmatrix}, \tag{3.1}$$

$$s_i = \sin(r_{if}),\ c_i = \cos(r_{if}) \quad (i = 1, 2, 3), \tag{3.2}$$

where (t_{1f}, t_{2f}, t_{3f}) are camera position parameters, and (r_{1f}, r_{2f}, r_{3f}) are camera posture parameters representing yaw, pitch, roll of a camera, respectively. 3-D point $\mathbf{S}_p = (x_p, y_p, z_p)$ is projected to $\hat{\mathbf{x}}_{fp} = (\hat{u}_{fp}, \hat{v}_{fp})$ on the ideal image

coordinate by the following equation:

$$a \begin{pmatrix} \hat{u}_{fp} \\ \hat{v}_{fp} \\ 1 \end{pmatrix} = \mathbf{Mfull}_f \begin{pmatrix} x_p \\ y_p \\ z_p \\ 1 \end{pmatrix}, \tag{3.3}$$

where a is a parameter. In this process for curved targets, this 3-D coordinate $\mathbf{S}_p = (x_p, y_p, z_p)$ instead of 2-D coordinate (x_p, y_p) is estimated for each feature point on the target. Accordingly, $\mathbf{Mfull}_f$ instead of simplified $\mathbf{M}_f$ for flat target is estimated.

In the same manner as for flat targets, the process is composed of initialization step for the first frame ($f = 1$), and an iterative step for the subsequent frames ($f > 1$). In the first frame, assuming that the focal plane in the first frame is parallel to the target and that the distance between the target and the viewpoint of the camera is d, $\mathbf{Mfull}_f$ is initialized as follows:

$$\mathbf{Mfull}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -d \end{pmatrix}. \tag{3.4}$$

Based on the same assumption, 3-D position for each feature point p detected in the first frame is given as follows:

$$\begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix} = \begin{pmatrix} \hat{u}_{1p} d \\ \hat{v}_{1p} d \\ 0 \end{pmatrix}. \tag{3.5}$$

Note that these are only initial values, which will be corrected in the refinement process (Figure 3.1(d)).

Despite these differences, the process for subsequent frames ($f > 1$) is the same as the one for flat targets. Readers are encouraged to refer to Section 2.3.2 for the detail.

3.3.2 Detection of Reappearing Features

Reappearing features are detected by the process described in 2.3.3 with an extension to deal with curved surface. The flow of this process is as follows.

First, templates of all the features are projected to the fitted surface (described later) to remove the geometric distortion induced by the curvature of the target. Then, feature pairs whose distance in 3-D space is less than a given threshold are selected and tested with the normalized cross correlation function in multiple scales. If the correlation is higher than a certain threshold, the feature pair is regarded as reappearing features and tracks belonging to each feature are merged into a single track.

The difference between this process and the process for flat targets in Section 2.3.3 is that templates for features are projected to the fitted surface instead of the mosaic image plane. This surface, however, is unknown in the first iteration of steps (a) to (c), since it will be obtained in the later process (c) by fitting a parameterized 3-D surface to estimated 3-D feature points. Thus, this step will be skipped in the first iteration of steps (a) to (c).

3.3.3 Global Optimization

In order to remove cumulative errors in the initial estimates of extrinsic camera parameters and 3-D feature points, global optimization on these parameters is carried out by the same process described in Section 2.3.4.

Let us recall the estimation error E defined in Eq. 2.14:

$$\begin{aligned} E &= \sum_f \sum_p E_{fp} \\ &= \sum_f \sum_p |\hat{\mathbf{x}}_{fp} - \mathbf{x}'_{fp}|^2 \\ &= \sum_f \sum_p \{(\hat{u}_{fp} - u'_{fp})^2 + (\hat{v}_{fp} - v'_{fp})^2\}. \end{aligned} \tag{3.6}$$

Note that $(\hat{u}_{fp}, \hat{v}_{fp})$ is given by transferring (x_p, y_p, z_p) by Eq. 2.3 using $\mathbf{Mfull}_f$. This error function E is minimized with respect to the camera parameters $\mathbf{Mfull}_f$ and the feature positions (x_p, y_p, z_p) to obtain globally optimized estimates of these parameters.

3.3.4 Target Shape Estimation by Surface Fitting

In this step, assuming that the target is a generalized cylinder, the target shape is estimated using 3-D feature points optimized in the previous step (b). This

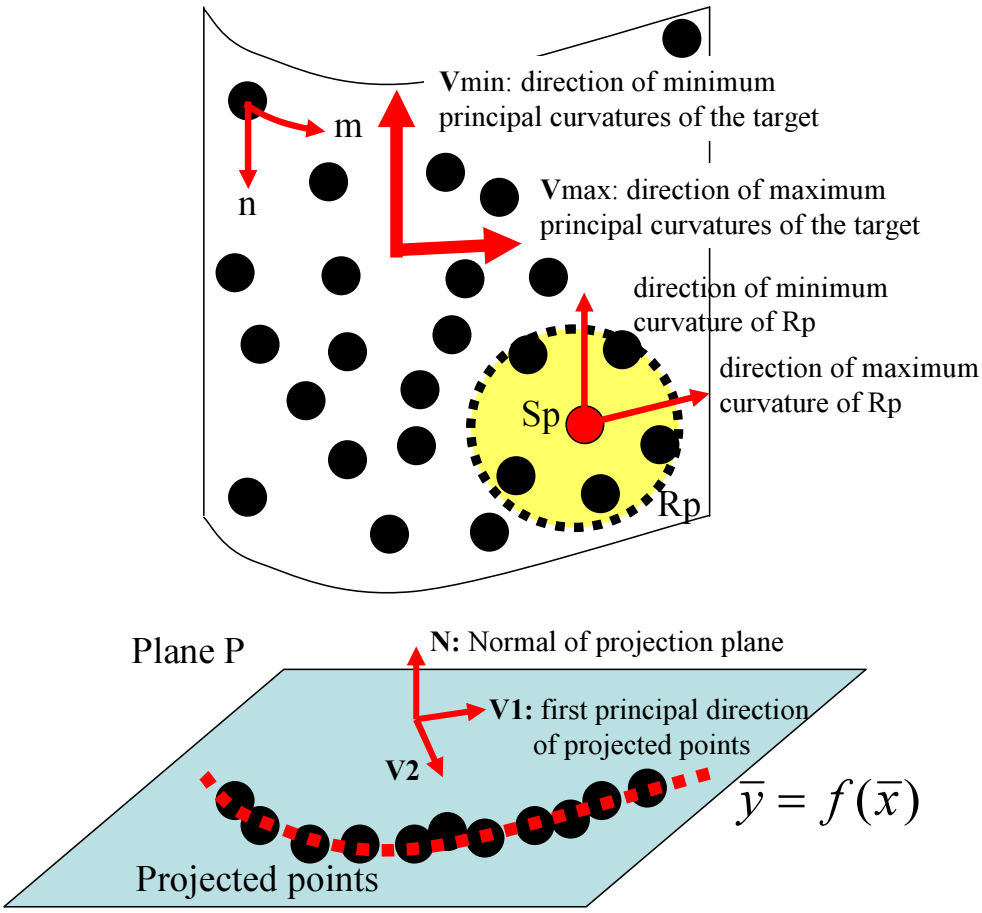


Figure 3.2. Target shape estimation by polynomial surface fitting.

step is a process specifically designed for curved targets.

First, as shown in Figure 3.2, the principal direction of curvature is computed from the 3-D point clouds. Next, 3-D position of each feature point is projected to a plane perpendicular to the direction of minimum principal curvatures. Finally, a polynomial equation of variable order is fitted to the projected 2-D coordinates, and the target shape is estimated. These steps will be described in detail.

Let us consider for each 3-D point $\mathbf{S}_p = (x_p, y_p, z_p)$ a point cloud $\mathbf{R}_p$ which consists of feature points lying within a certain distance from $\mathbf{S}_p$, as shown in Figure 3.2. First, the directions of maximum and minimum curvatures

are computed for each $\mathbf{R}_p$ using local quadratic surface fit. If the target is a generalized cylinder, as assumed in the proposed method, the minimum principal curvature must be 0, and its direction must be the same for all the feature points. In practice, however, there exists some fluctuation in the directions of minimum curvature, due to the estimation errors. Thus, a voting method is applied to eliminate outliers and to determine the dominant direction $\mathbf{V}_{min} = (v_{mx}, v_{my}, v_{mz})$ of minimum principal curvatures on the target.

Next, 3-D position $\mathbf{S}_p$ for each feature point is projected to a plane whose normal vector $\mathbf{N}$ coincides with $\mathbf{V}_{min}$; i.e. $P(x, y, z) = v_{mx}x + v_{my}y + v_{mz}z = 0$. The projected 2-D coordinate $(\bar{x}_p, \bar{y}_p)$ of $\mathbf{S}_p$ is given as follows:

$$\begin{pmatrix} \bar{x}_p \\ \bar{y}_p \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix} \mathbf{S}_p, \tag{3.7}$$

where $\mathbf{V}_1$ is a unit vector parallel to the principle axis of inertia of the projected 2-D coordinates $(\bar{x}, \bar{y})$, and $\mathbf{V}_2$ is a unit vector which is perpendicular to $\mathbf{V}_1$ and $\mathbf{V}_{min}$; i.e. $\mathbf{V}_2 = \mathbf{V}_1 \times \mathbf{N}$ (see Figure 3.2).

Finally, the target shape parameter $(a_0, a_1, \cdots, a_m)$ is obtained by fitting the following variable-order polynomial equation to the projected 2-D coordinates $(\bar{x}, \bar{y})$.

$$\bar{y} = f(\bar{x}) = \sum_{i=0}^{q} a_i \bar{x}^i. \tag{3.8}$$

Generally, higher q gives less residual between the 2-D coordinates and the fitted equation. However, too high q would cause over-fitting to the noise in the 2-D coordinates. This trade-off implies there exists an optimal order q in terms of residual and sensitivity to the noise. In the proposed method, the optimal order q is automatically determined by using geometric AIC [Kan98]. q which minimize the following criteria G-AIC is determined as the optimal order.

$$\text{G-AIC} = J + 2(N(m - r) + q + 1)\epsilon^2, \tag{3.9}$$

where J is the residual, N is the number of points, m is the dimension of observed data $(\bar{x}, \bar{y})$, which equals to 2, and r is the number of constraint equations in fitting Eq. (3.8) to the projected 2-D coordinates, which equals to N. ϵ, called *noise level*, is the average error of the estimated feature position

along $\bar{y}$ axis. Since ϵ is unknown, by assuming that the noise level of $(\bar{x}, \bar{y})$ is proportional to the distances between 3-D feature points and the camera, ϵ is approximated as follows:

$$\epsilon = Cl, \tag{3.10}$$

where l is the average of the depth of each feature point in camera coordinate of every frame, and C is a constant, which is empirically set to 0.007.

Here, the order q is independent of N, m, r and ϵ, thus the actual criteria to be minimized is given as follows:

$$G = J + 2q\epsilon^2. \tag{3.11}$$

The first term on the right hand-side is the residual, which monotonously decreases as the order q increases, and converges to 0 at the infinity. On the other hand, the second term monotonously increases in proportion to q. Thus, G, given as the sum of the two terms, is a function whose shape is convex downward. This is a favorable property for searching the global minimum.

In case of a target with multiple curved surfaces, first, the line where the normal vector of the locally fitted quadratic surface varies discontinuously is detected, and the whole target is divide with this line, as shown in Figure 3.3. Then the shape parameter is computed for each part of the target.

The shape of the target estimated in this step is used for generating a geometric distortion-free mosaic image in the next process, as well as for removing the geometric distortion in the reappearing feature detection process described in Section 3.3.2.

3.3.5 Mosaic Image Generation

Finally, a mosaic image is generated by using extrinsic camera parameters and surface shape parameters. In this step, first, a curvature distortion-free, or an *unwrapped* mosaic image is generated. Then, a post-process is applied to the mosaic image to remove the shade induced by the curved shape of the target.

Before describing the actual process for mosaic image generation, first, the relationship between the 2-D coordinate on the mosaic image and its corresponding 2-D coordinates on input images is defined. Let us consider a 2-D coordinate (m, n) on the unwrapped mosaic image as shown in Figure 3.4. Here,

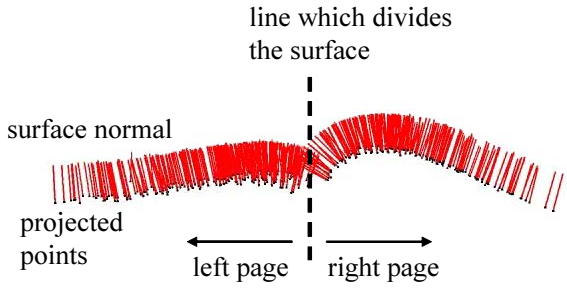


Figure 3.3. Detection of line where surface normal varies discontinuously.

the relation between (m, n) and its corresponding 3-D coordinate $(\bar{x}, f(\bar{x}), \bar{z})$ on the fitted surface is given as follows:

$$(m, n) = (\int_0^{\bar{x}} \sqrt{1 + \{\frac{d}{dx} f(x)\}^2} dx, \bar{z}). \tag{3.12}$$

The relationship between the 3-D coordinate $(\bar{x}, f(\bar{x}), \bar{z})$ on the fitted surface and its corresponding 2-D coordinate (u_f, v_f) on the f-th image plane is given by the following equation:

$$a \begin{pmatrix} u_f \\ v_f \\ 1 \end{pmatrix} = \mathbf{Mfull}_f \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{N} \end{pmatrix} \begin{pmatrix} \bar{x} \\ f(\bar{x}) \\ \bar{z} \end{pmatrix}, \tag{3.13}$$

where a is a parameter.

Now that we know the relationship between the 2-D coordinate (m, n) on the mosaic image and its corresponding 2-D coordinates (u_f, v_f) on input images, the mosaic image can be generated as follows. First, for each pixel (m, n) on the unwrapped mosaic image, the corresponding coordinate (u_f, v_f) in each input image is computed by Eq. (3.12) and (3.13). Then, the average of the pixel values at all the corresponding coordinates is computed, and is determined as the pixel value at (m, n) on the unwrapped mosaic image. Here, as shown in Figure 3.5, the resolution of the target captured in the input image decreases as the angle formed by the surface normal and the line from the target to the focal point of the camera, depicted by θ_f, increases. Thus, in the

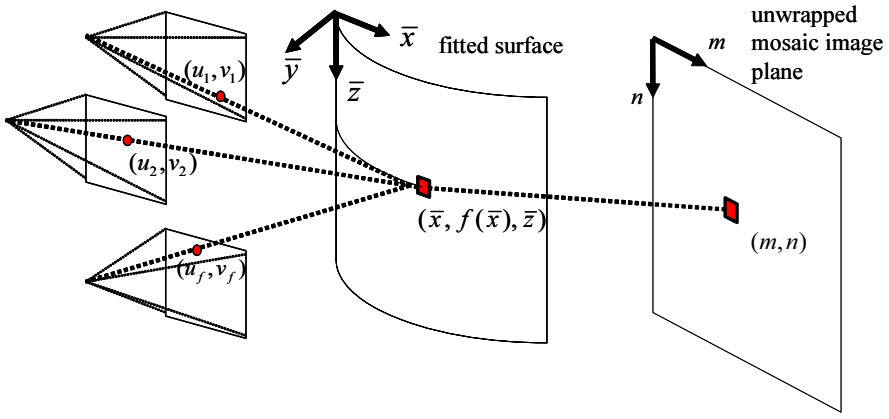


Figure 3.4. Relationship among coordinates on input image, fitted surface and mosaic image plane.

computation of the average of the pixel values, each pixel value is weighted according to θ_f. The concrete formulation of the computation of $I(m, n)$, the pixel value at (m, n) on the unwrapped mosaic image is given as follows:

$$I(m,n) = \frac{1}{\sum_f W_f} \sum_f W_f I_f(u_f, v_f), \tag{3.14}$$

$$W_f = K^{-\theta_f}, \tag{3.15}$$

where $I_f(u_f, v_f)$ is the corresponding pixel value at (u_f, v_f) in the f-th input image, K is a constant which is empirically set to 10^4, and θ_f is the angle formed by the surface normal at $(\bar{x}, f(\bar{x}), \bar{z})$ and the line from $(\bar{x}, f(\bar{x}), \bar{z})$ to the focal point in the f-th frame.

After the unwrapped mosaic image is generated by the above process, a post-process is applied to remove the shade on the mosaic image induced by the curved shape of the target. Here, the following assumptions are made: the target is a Lambertian surface with the background having the maximum intensity on the target, and the target is illuminated by a parallel light source.

As is described in Section 3.3.4, the vertical direction in the mosaic image coordinate (m, n) is defined to coincide with the direction of the minimum principle curvature of the target. Thus, under a parallel light source, the effect of shade is uniform for pixels having the same m coordinate on the mosaic

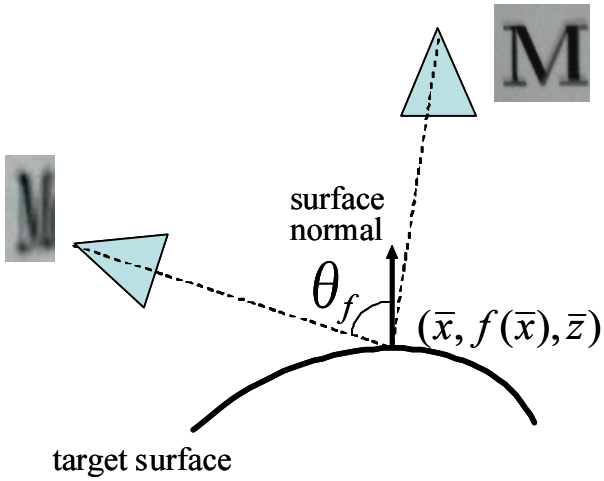


Figure 3.5. Angle formed by surface normal and camera orientation.

image. If we can assume that, in any column of the mosaic image, there exists at least one pixel belonging to the background, the true pixel value $I_{\text{new}}(m, n)$ without shade can be computed by the following equation:

$$I_{\text{new}}(m,n) = \frac{I_{\max} I(m,n)}{\max(I(m+u, n+v); \forall (u,v) \in W)}, \tag{3.16}$$

where W is a rectangular window whose height is larger than its width, e.g. a window with the size of 5×500 pixel, and $I_{\max}$ is the maximum possible intensity value of an image (typically 255).

3.4. Prototype System for Curved Document

This method for curved document has been developed on a prototype system. The system is composed of a laptop PC and a hand-held IEEE1394 CCD camera, whose intrinsic parameters are calibrated by Tsai's method [Tsa86] in advance, and are fixed during the image capturing. The specifications of the system are shown in Table 3.1.

To enable real-time processing, two-stage implementation is employed again. In the real-time stage, while the user captures a target document using a hand-held camera, initial 3-D reconstruction by feature tracking is carried out. The

rest of the processes, i.e. detection of reappearing features, global optimization, surface fitting and mosaic image generation are executed in the offline stage.

Experiments performed by this prototype system are described in the following section.

Table 3.1. Specifications of a video mosaicing system for curved target.

Laptop PC	
CPU	Pentium-M 2.1GHz
Memory	2GB
IEEE1394 camera (Aplux C104T)	
Resolution	640×480 pixels
View angle	$31.7^\circ \times 24.1^\circ$
Maximum frame rate	15 frames/sec

3.5. Experiments

In order to evaluate the feasibility of the proposed method, two experiments are performed by the prototype video mosaicing system. The first experiment is performed on a thick bound book with two pages of curved surface shown in Figure 3.6 (sequence 1). The distortion in the resultant mosaic images is evaluated quantitatively in the same manner with the experiments on flat targets in Section 2.6. The second experiment is performed on a label on a wine bottle shown in Figure 3.7 (sequence 2).

3.5.1 Experiment on Book (Sequence 1)

The target document captured in this experiment is shown in Figure 3.6. The target is a thick bound book composed of 2 pages of curved surfaces: the left page with texts and the right page with pictures and figures. Note that plus marks (+) are printed on both pages at every 40mm grid positions for quantitative evaluation. Since the right page gives less number of image features compared to the left page full of texts, degradation in the accuracy of 3-D reconstruction and the quality of the resultant mosaic image is expected. We will see how this different characteristic of pages affects the final result.

In this experiment, the whole target is captured as 640×480 pixel images of 300 frames. Sampled frames of the captured images are shown in Figure 3.8. Cross marks in Figure 3.9 indicate feature points which are automatically detected and are used for extrinsic camera parameter estimation. In this experiment, 94 image feature points are tracked per frame on average.

Initial estimate of 3-D reconstruction after the real-time stage and the preview of the unwrapped mosaic image generated by this initial estimate are shown in Figure 3.10 (a). Due to error in the initial estimate, the minimum principle curvature direction, which in ideal case coincides to the vertical direction of the page, is erroneously estimated. This has caused skew and rotation in the resultant mosaic image.

This initial estimate is refined in the off-line stage by iterating the sequence of reappearing feature detection, global optimization on estimated parameters and surface fitting, as shown in Figure 3.1. Figure 3.10 (b) and (c) show

the extrinsic camera parameters and the mosaic image after the 1st and 2nd iteration, respectively. By comparing the result of 3-D reconstruction after the 1st iteration (Figure 3.10 (b)) with that after the real-time stage (Figure 3.10 (a)), we can see that the 3-D reconstruction has been refined. This effect is obvious in the mosaic image after the 1st iteration (Figure 3.10 (b)), where the estimated minimum principle curvature direction is close to that of the vertical direction of the book. In the 2nd iteration, reappearing features are detected and are utilized to further optimize the 3-D reconstruction result, as shown in Figure 3.10 (c). Red points show the detected reappearing features. The total number of detected reappearing features in this experiment is 90.

The refinement process converged after 3 iterations. The 3-D reconstruction result obtained after the refinement process is shown in Figure 3.11. The curved line shows the camera path, pyramids show the camera postures in every 20 frames, and the point cloud shows 3-D positions of feature points. As can be seen, the point cloud coincides with the shape of the thick bound book. The average reprojection errors of the features before and after refinement are 0.88 pixel and 0.73 pixel, respectively.

Figure 3.12 shows the directions of the maximum principle curvature, minimum principle curvature and the surface normal estimated for the 3-D reconstruction result shown in Figure 3.11. The directions of the maximum principle curvature, minimum principle curvature and the surface normal are shown in cyan, magenta and yellow arrows, respectively. In this figure, the minimum curvature direction for each 3-D feature point is also shown in vector shooting out from the feature point. Here, red vectors are those which support the estimated minimum principle curvature direction, i.e. inliers. On the other hand, blue vectors are those rejected as outliers. The feature points without vectors are those having too small curvature, thus are neglected in estimating the minimum principle curvature direction.

Then, 3-D feature points are projected in the minimum principle curvature direction, and polynomial equations with different orders are fitted to their 2-D projected points. The optimal order of the fitted polynomial equation is automatically determined by geometric AIC [Kan98]. Figure 3.13 shows the fitted polynomial equations whose orders vary from 0 to 8 for the left and

right page, respectively. Geometric AIC computed for each order is plotted in Figure 3.14. In this experiment, the optimal orders of the fitted polynomial equation, which give the minimum geometric AIC, is 6 and 4 for the left and right pages, respectively. The estimated shape of the target given by the optimal orders is shown in Figure 3.15.

The unwrapped mosaic images before and after removing shade are shown in Figure 3.16 and 3.17, respectively. The resolution of the mosaic image is 3200×2192. As can be seen, the distortion on both pages has been removed in the resultant image. Since the right page gives less image features compared to the left page, degradation in the accuracy of extrinsic camera parameter estimation and the quality of the resultant mosaic image has been expected for the right page. In the resultant mosaic image, however, there is no noticeable difference in the quality between both pages. A further discussion on this comparison will be given based on quantitative evaluation in Section 3.5.3.

It should be noted that blurs can be observed in the boundary part between both pages. The reason for these blurs is that, in this area, the angle formed by the surface normal and the optical axis of the camera is nearly 90 degrees, which causes error in feature tracking and thus degrades the accuracy of 3-D reconstruction. This is why the super-resolution technique has been omitted in the proposed method. Utilizing other types of image feature, e.g. line and arc segments, besides feature points can be a promising solution to improve the accuracy of 3-D reconstruction, and to achieve the accuracy sufficient to apply super-resolution technique.

The performance of the system in this experiment is as follows: 22 seconds for initial 3-D reconstruction, 188 seconds for camera parameter refinement and surface fitting and 410 seconds for generating the final mosaic image.

3.5.2 Experiment on Label on Wine Bottle (Sequence 2)

The application of the proposed method is not only limited to documents. In this experiment, we will see how the proposed method can be applied to other types of targets.

Figure 3.6. Thick bound book with curved surface (Sequence 1).

Figure 3.7. Label on a wine bottle (Sequence 2).

The target in this experiment is a label on a wine bottle, as shown in Figure 3.7. This target is composed of drawings and few lines of texts. Some of them, e.g. the contour of the label and the text line which says "TYRRELL'S WINES", are curved by design. This makes it difficult for conventional distortion correction methods to perform successfully, since they assume that bunches of straight lines are present on the target.

The target is captured as 640×480 pixel images of 100 frames. Sampled frames of the captured images are shown in Figure 3.18. Cross marks in Figure 3.19 indicate feature points which are automatically detected and used for extrinsic camera parameter estimation. In this experiment, 33 image feature points are tracked per frame on average.

Initial estimate of 3-D reconstruction after the real-time stage and the pre-

view of the unwrapped mosaic image generated by this initial estimate are shown in Figure 3.20 (a). Due to error in the initial estimate, the minimum principle curvature direction, which in ideal case coincides to the vertical direction of the page, is erroneously estimated. This initial estimate is refined in the off-line stage by iterating the sequence of reappearing feature detection, global optimization on estimated parameters and surface fitting. Figure 3.20 (b) and (c) show the extrinsic camera parameters and the mosaic image after the 1st and 2nd iteration, respectively. As can be seen, the initial estimate of 3-D reconstruction is gradually refined, and the direction of minimum principle curvature is approaching to the correct direction after each iteration. Here, the detected reappearing features are depicted as red points. Although it is expected that no reappearing features will be detected since the camera moves in constant direction, 4 features have been detected as reappearing features. These are the features whose tracks have been split into distinct tracks due to failures in feature tracking. These tracks, however, have been merged into single tracks by reappearing feature detection process. This fact shows that reappearing feature detection process is also effective for recovering failures in feature tracking.

The refinement process converged after 3 iterations. The 3-D reconstruction result obtained after the refinement process is shown in Figure 3.21. The curved line shows the camera path, pyramids show the camera postures in every 10 frames, and the point cloud shows 3-D positions of feature points. As can be seen, the point cloud coincides with the shape of the bottle. The average reprojection errors of the features before and after refinement are 0.70 pixel and 0.58 pixel, respectively.

Figure 3.22 shows the directions of the maximum principle curvature, minimum principle curvature and the surface normal estimated for the 3-D reconstruction result shown in Figure 3.21. The directions of the maximum principle curvature, minimum principle curvature and the surface normal are shown in cyan, magenta and yellow arrows, respectively.

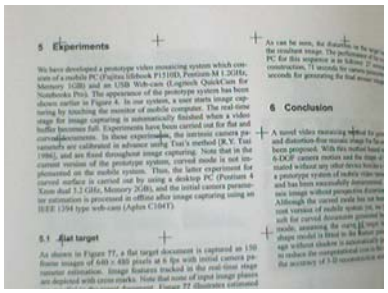
Then, 3-D feature points are projected in the minimum principle curvature direction, and polynomial equations with different orders are fitted to their 2-D projected points. The optimal order of the fitted polynomial equation is

automatically determined by geometric AIC [Kan98].

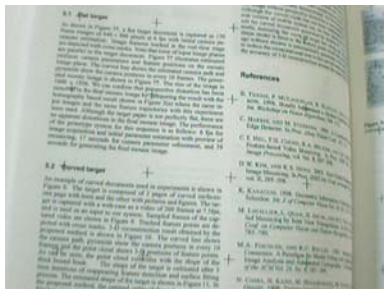
Figure 3.23 shows the polynomial equations fitted to the 2-D projection of 3-D feature points in the minimum principle curvature direction. The orders of the equation vary from 0 to 8. Geometric AIC computed for each order is plotted in Figure 3.24. In this experiment, the optimal order of the fitted polynomial equation, which gives the minimum geometric AIC, is 2. The estimated shape of the target given by this optimal order is shown in Figure 3.25.

The unwrapped mosaic images is shown in Figure 3.26. The resolution of the mosaic image is 3200×2343. The process for removing shade described in 3.3.5 has not been applied for this target, since specular reflection has been observed in the input images. This specular reflection, however, is not observed in the mosaic image. This is a side-effect of blending in the mosaic image generation, which averages out noise and specular reflection in the input images.

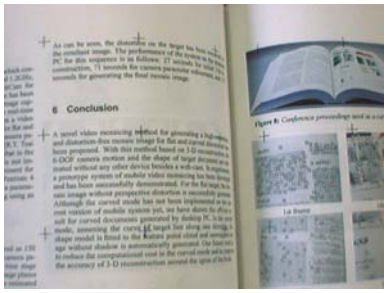
It should be noted that, although the target had few lines structures, some of which are not even straight, the distortion has been successfully removed by the method. This result shows the advantage of the proposed method over conventional distortion correction methods which rely on bunches of straight lines on a target. This result also shows the potential of the proposed method on targets other than documents.



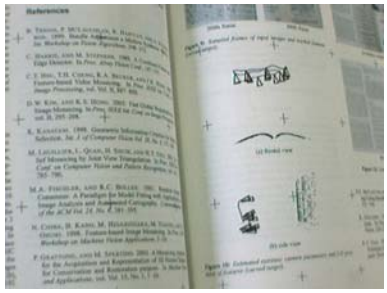
1st frame



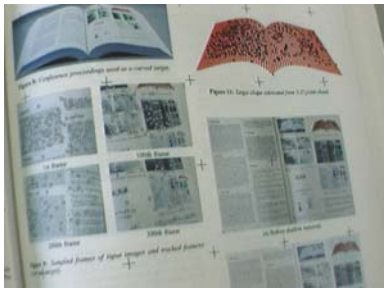
171st frame



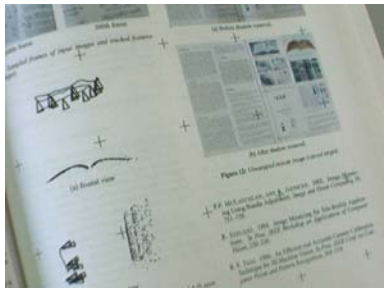
43rd frame



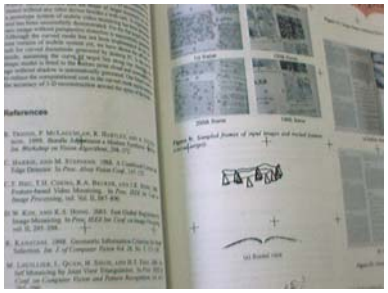
214th frame



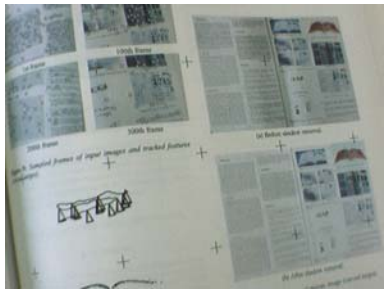
86th frame



257th frame

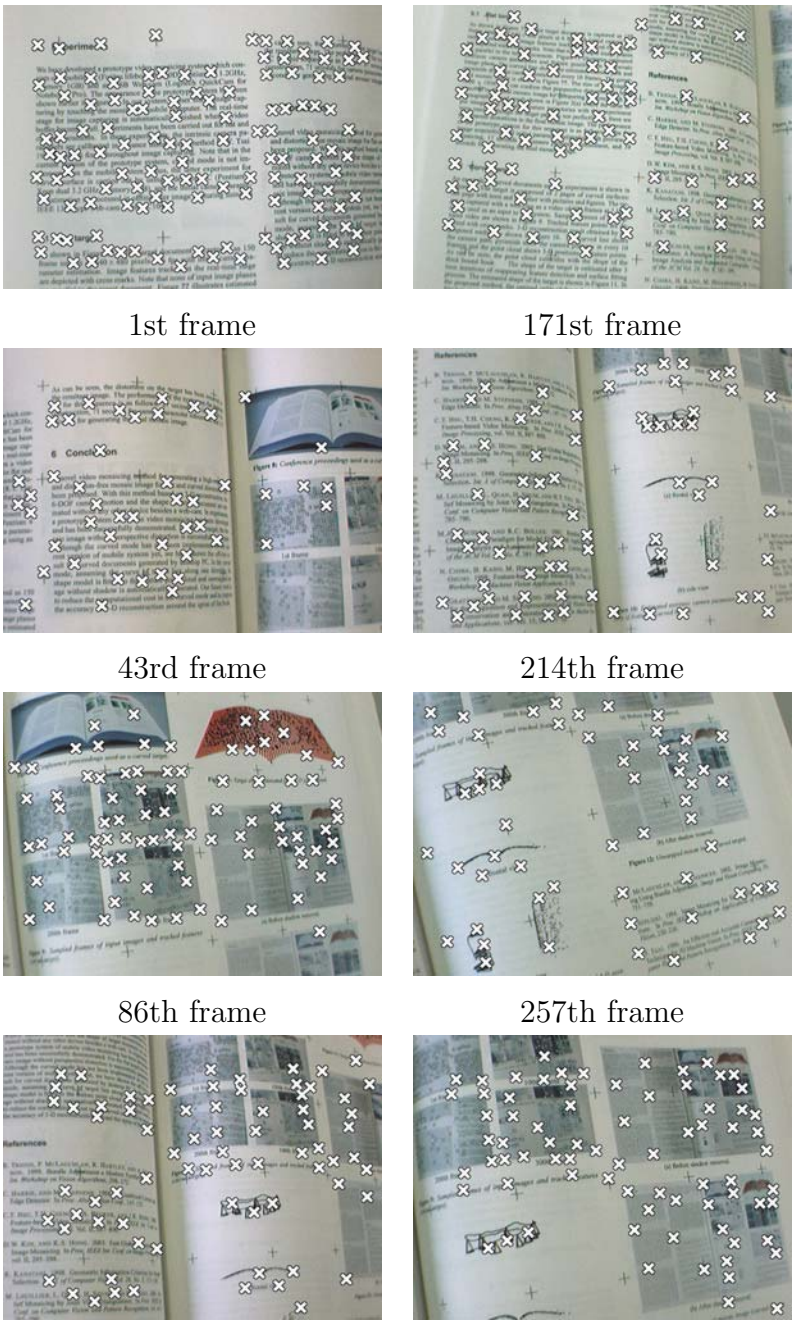


129th frame



300th frame

Figure 3.8. Sampled frames of input image sequence (Sequence 1).



1st frame 171st frame

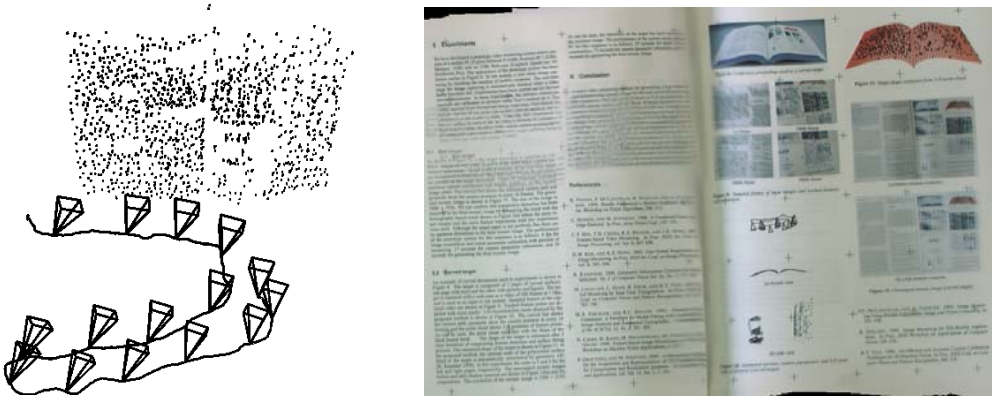
43rd frame 214th frame

86th frame 257th frame

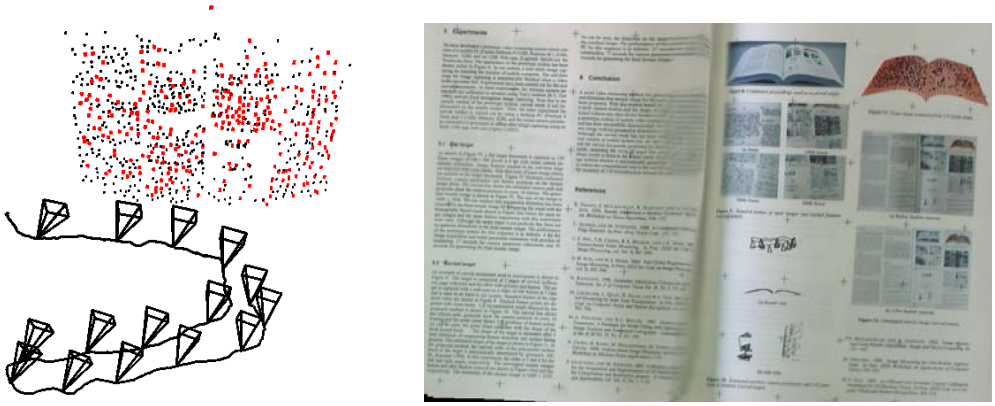
129th frame 300th frame

Figure 3.9. Tracked features in input image sequence (Sequence 1).

(a) initial result

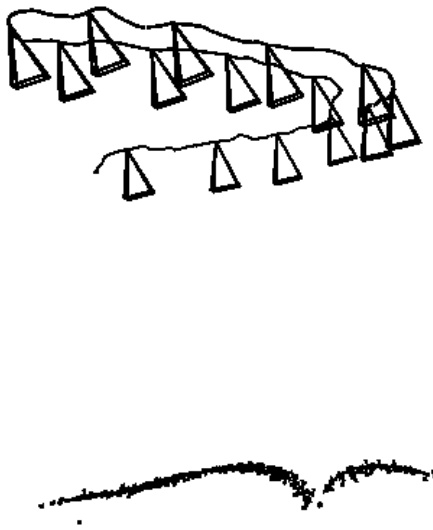


(b) result after 1st iteration

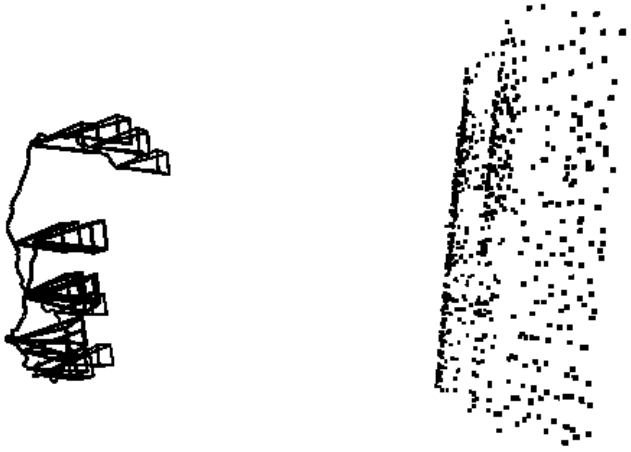


(c) result after 2nd iteration

Figure 3.10. Extrinsic camera parameters and mosaic image under refinement (Sequence 1).



(a) frontal view



(b) side view

Figure 3.11. Estimated extrinsic camera parameters and 3-D positions of features (Sequence 1).

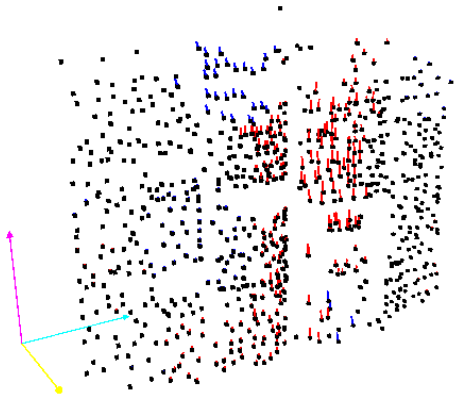
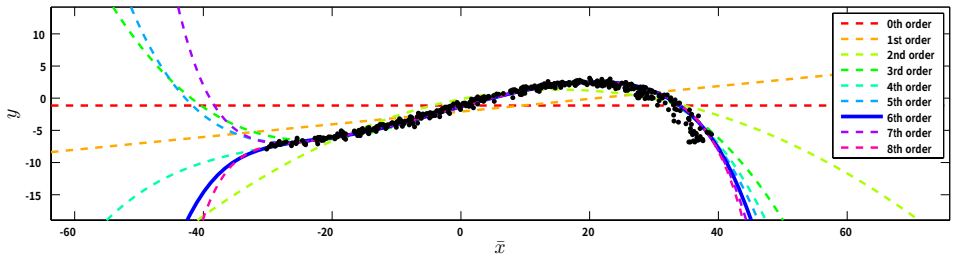
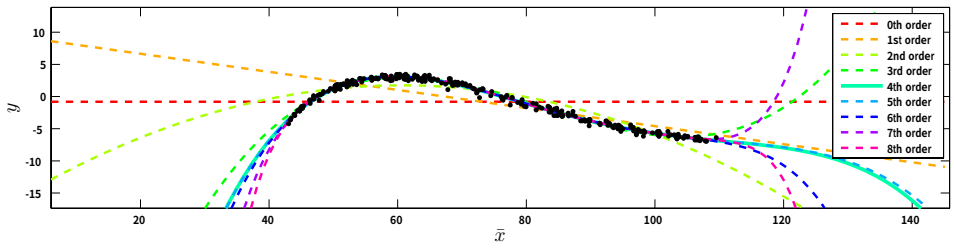


Figure 3.12. Estimated principle curvature directions (Sequence 1).

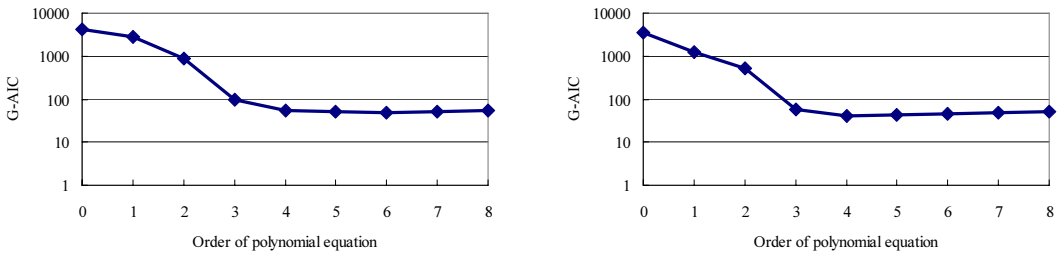


(a) Left page



(b) Right page

Figure 3.13. Polynomial equations fitted to projected 2-D coordinates of features (Sequence 1).



(a) Left page (b) Right page

Figure 3.14. Evaluation of G-AIC for fitted polynomial equations (Sequence 1).

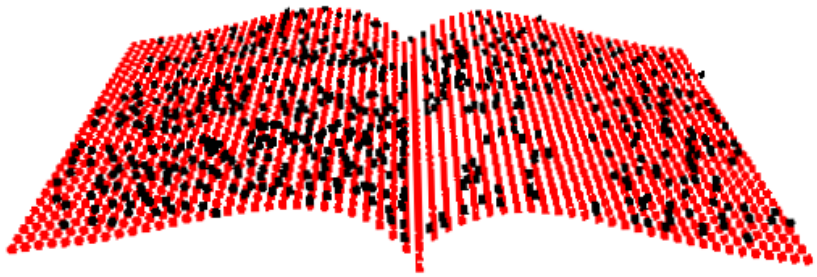
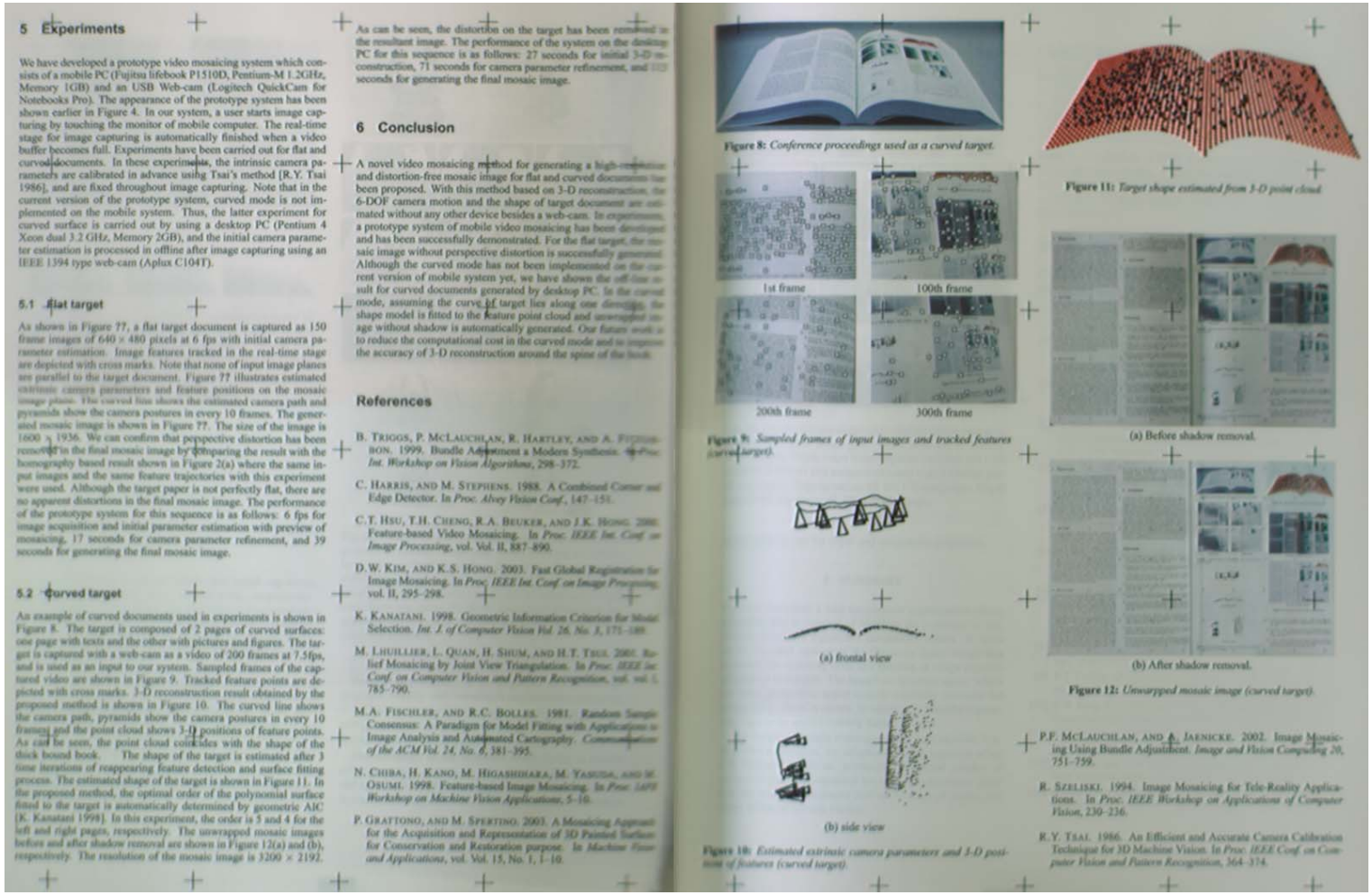


Figure 3.15. Target shape estimated from 3-D feature points (Sequence 1).



5 Experiments

We have developed a prototype video mosaicing system which consists of a mobile PC (Fujitsu lifebook P1510D, Pentium-M 1.2GHz, Memory 1GB) and an USB Web-cam (Logitech QuickCam for Notebooks Pro). The appearance of the prototype system has been shown earlier in Figure 4. In our system, a user starts image capturing by touching the monitor of mobile computer. The real-time stage for image capturing is automatically finished when a video buffer becomes full. Experiments have been carried out for flat and curved documents. In these experiments, the intrinsic camera parameters are calibrated in advance using Tsai's method [R.Y. Tsai 1986], and are fixed throughout image capturing. Note that in the current version of the prototype system, curved mode is not implemented on the mobile system. Thus, the latter experiment for curved surface is carried out by using a desktop PC (Pentium 4 Xeon dual 3.2 GHz, Memory 2GB), and the initial camera parameter estimation is processed in offline after image capturing using an IEEE 1394 type web-cam (Aplux C104T).

5.1 Flat target

As shown in Figure ??, a flat target document is captured as 150 frame images of 640 × 480 pixels at 6 fps with initial camera parameter estimation. Image features tracked in the real-time stage are depicted with cross marks. Note that none of input image planes are parallel to the target document. Figure ?? illustrates estimated extrinsic camera parameters and feature positions on the mosaic image plane. The curved line shows the estimated camera path and pyramids show the camera postures in every 10 frames. The generated mosaic image is shown in Figure ??. The size of the image is 1600 × 1936. We can confirm that perspective distortion has been removed in the final mosaic image by comparing the result with the homography based result shown in Figure 2(a) where the same input images and the same feature trajectories with this experiment were used. Although the target paper is not perfectly flat, there are no apparent distortions in the final mosaic image. The performance of the prototype system for this sequence is as follows: 6 fps for image acquisition and initial parameter estimation with preview of mosaicing, 17 seconds for camera parameter refinement, and 39 seconds for generating the final mosaic image.

5.2 Curved target

An example of curved documents used in experiments is shown in Figure 8. The target is composed of 2 pages of curved surfaces: one page with texts and the other with pictures and figures. The target is captured with a web-cam as a video of 200 frames at 7.5fps, and is used as an input to our system. Sampled frames of the captured video are shown in Figure 9. Tracked feature points are depicted with cross marks. 3-D reconstruction result obtained by the proposed method is shown in Figure 10. The curved line shows the camera path, pyramids show the camera postures in every 10 frames, and the point cloud shows 3-D positions of feature points. As can be seen, the point cloud coincides with the shape of the thick bound book. The shape of the target is estimated after 3 time iterations of reappearing feature detection and surface fitting process. The estimated shape of the target is shown in Figure 11. In the proposed method, the optimal order of the polynomial surface fitted to the target is automatically determined by geometric AIC [K. Kanatani 1998]. In this experiment, the order is 5 and 4 for the left and right pages, respectively. The unwrapped mosaic images before and after shadow removal are shown in Figure 12(a) and (b), respectively. The resolution of the mosaic image is 3200 × 2192.

As can be seen, the distortion on the target has been removed in the resultant image. The performance of the system on the desktop PC for this sequence is as follows: 27 seconds for initial 3-D reconstruction, 71 seconds for camera parameter refinement, and [illegible] seconds for generating the final mosaic image.

6 Conclusion

A novel video mosaicing method for generating a high-resolution and distortion-free mosaic image for flat and curved documents has been proposed. With this method based on 3-D reconstruction, the 6-DOF camera motion and the shape of target document are estimated without any other device besides a web-cam. In experiments, a prototype system of mobile video mosaicing has been developed and has been successfully demonstrated. For the flat target, the mosaic image without perspective distortion is successfully generated. Although the curved mode has not been implemented on the current version of mobile system yet, we have shown the off-line result for curved documents generated by desktop PC. In the curved mode, assuming the curve of target lies along one direction, the shape model is fitted to the feature point cloud and unwrapped image without shadow is automatically generated. Our future work is to reduce the computational cost in the curved mode and to improve the accuracy of 3-D reconstruction around the spine of the book.

References

B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. 1999. Bundle Adjustment a Modern Synthesis. In *Proc. Int. Workshop on Vision Algorithms*, 298–372.

C. Harris, and M. Stephens. 1988. A Combined Corner and Edge Detector. In *Proc. Alvey Vision Conf.*, 147–151.

C.T. Hsu, T.H. Cheng, R.A. Beuker, and J.K. Hong. 2000. Feature-based Video Mosaicing. In *Proc. IEEE Int. Conf. on Image Processing*, vol. Vol. II, 887–890.

D.W. Kim, and K.S. Hong. 2003. Fast Global Registration for Image Mosaicing. In *Proc. IEEE Int. Conf. on Image Processing*, vol. II, 295–298.

K. Kanatani. 1998. Geometric Information Criterion for Model Selection. *Int. J. of Computer Vision Vol. 26, No. 3*, 171–189.

M. Lhuillier, L. Quan, H. Shum, and H.T. Tsui. 2005. Relief Mosaicing by Joint View Triangulation. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. vol. I, 785–790.

M.A. Fischler, and R.C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM Vol. 24, No. 6*, 381–395.

N. Chiba, H. Kano, M. Higashihara, M. Yasuda, and M. Osumi. 1998. Feature-based Image Mosaicing. In *Proc. IAPR Workshop on Machine Vision Applications*, 5–10.

P. Grattono, and M. Spertino. 2003. A Mosaicing Approach for the Acquisition and Representation of 3D Painted Surfaces for Conservation and Restoration purpose. In *Machine Vision and Applications*, vol. Vol. 15, No. 1, 1–10.

P.F. McLauchlan, and A. Jaenicke. 2002. Image Mosaicing Using Bundle Adjustment. *Image and Vision Computing 20*, 751–759.

R. Szeliski. 1994. Image Mosaicing for Tele-Reality Applications. In *Proc. IEEE Workshop on Applications of Computer Vision*, 230–236.

R.Y. Tsai. 1986. An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 364–374.

Figure 8: *Conference proceedings used as a curved target.*

Figure 9: *Sampled frames of input images and tracked features (curved target).*

Figure 10: *Estimated extrinsic camera parameters and 3-D position of features (curved target).*

Figure 11: *Target shape estimated from 3-D point cloud.*

Figure 12: *Unwrapped mosaic image (curved target).*

Figure 3.16. Unwrapped mosaic image before shade correction(Sequence 1).

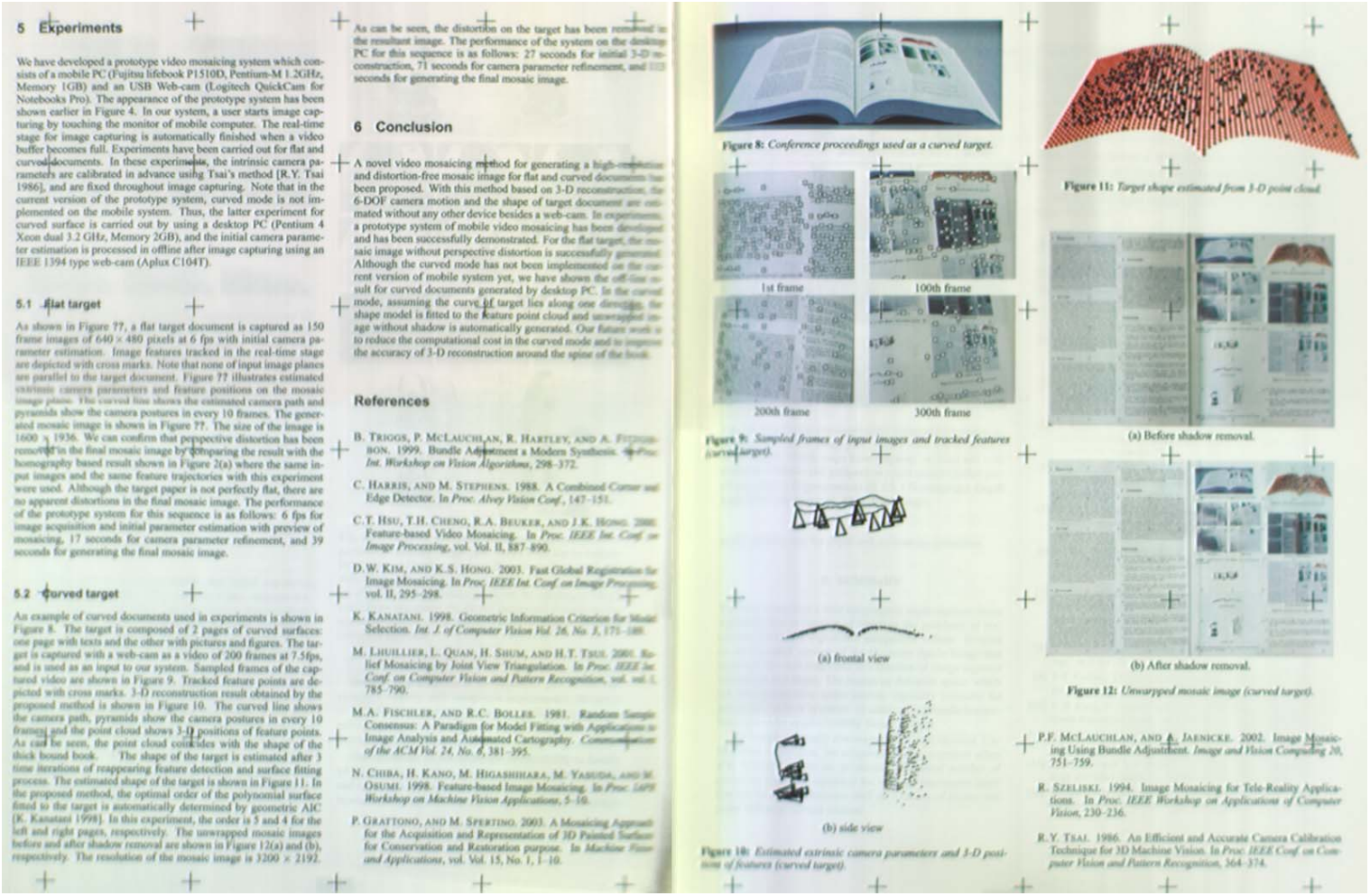


Figure 3.17. Unwrapped mosaic image after shade correction(Sequence 1).

Figure 3.18. Sampled frames of input image sequence (Sequence 2).

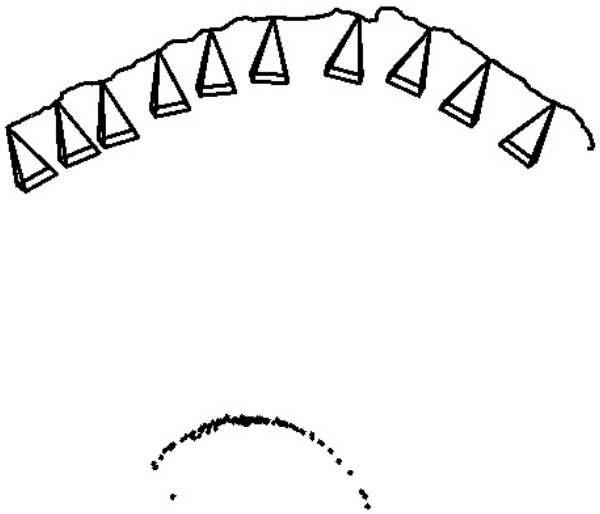
Figure 3.19. Tracked features in input image sequence (Sequence 2).

(a) initial result

(b) result after 1st iteration

(c) result after 2nd iteration

Figure 3.20. Extrinsic camera parameters and mosaic image under refinement (Sequence 2).



(a) frontal view

(b) side view

Figure 3.21. Estimated extrinsic camera parameters and 3-D positions of features (Sequence 2).

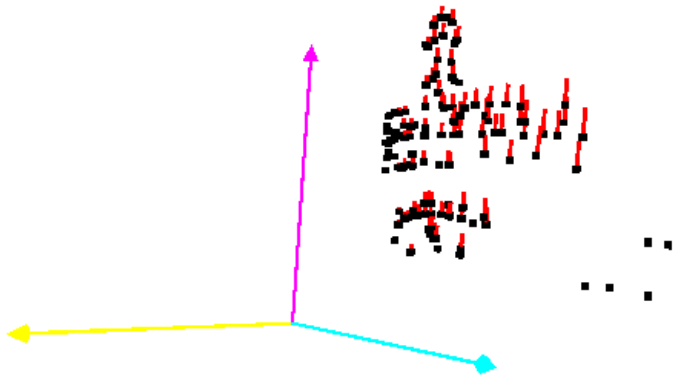


Figure 3.22. Estimated principle curvature directions (Sequence 2).

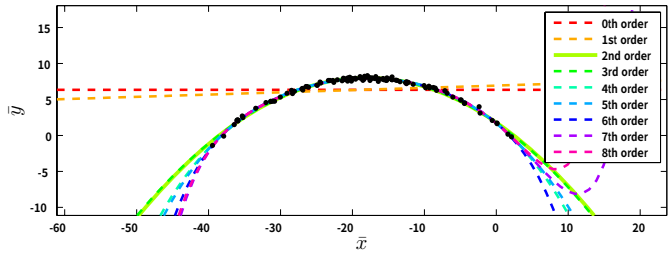


Figure 3.23. Polynomial equations fitted to projected 2-D coordinates of features (Sequence 2).

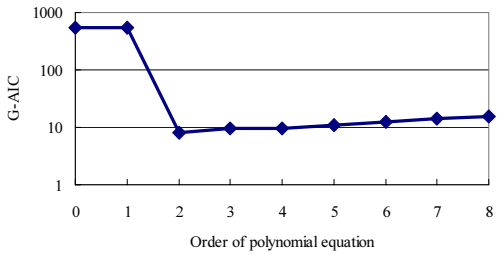


Figure 3.24. Evaluation of G-AIC for fitted polynomial equations (Sequence 2).

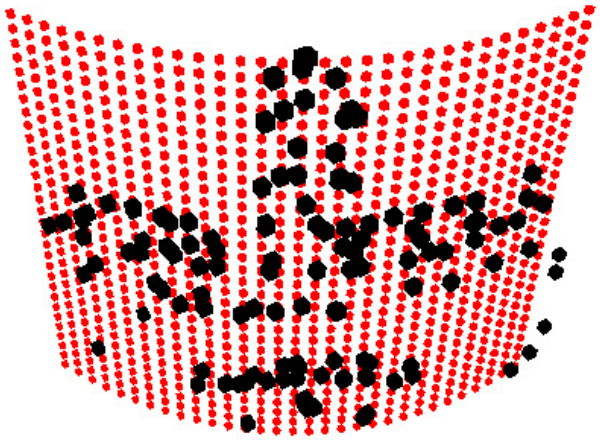


Figure 3.25. Target shape estimated from 3-D feature points (Sequence 2).

Figure 3.26. Unwrapped mosaic image (Sequence 2).

3.5.3 Quantitative Evaluation of Distortion

The feasibility of the proposed method is further evaluated by analyzing the distortion present in the resultant mosaic image. Here, the distortion in the mosaic image generated for a book (Figure 3.17 is evaluated using the same method as described in Section 2.6.6.

For each page, the distortion in the mosaic image is evaluated by the distances between every pair of adjacent plus marks (+). The average, maximum, minimum and standard deviation of the distances are shown in Table 3.2. The percentage of each value against the average distance is shown in parenthesis.

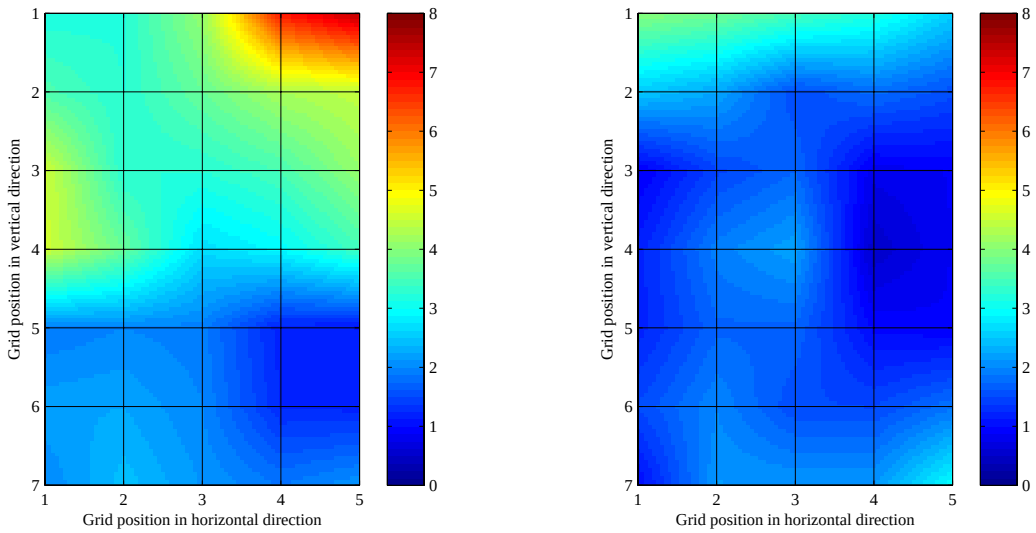
In this experiment, the average distortion is 1.11% and 0.81% for the left and right page, respectively. It should be noted that, despite the less number of image features obtained in the right page, the distortion in the mosaic image is comparable, or even smaller than that for the left page.

Figure 3.27 (a) and (b) show the distribution of distortion on the mosaic image for the left and right page, respectively. Each grid point is given the average of distortions measured between every adjacent grid point.

As can be seen, relatively large distortions are obtained on grid points in the upper right part of the left page. Close investigation on the target book has revealed that there is slight change in the curvature direction between this area and the rest of the page. This coincides with the fact that the curvature directions computed for feature points in this area are rejected as outliers in the minimum principle curvature direction estimation, as shown in Figure 3.12. This change in the curvature direction, which violates the assumption that the target shape is a generalized cylinder, has resulted in large distortions in this area. These distortions can be reduced by extending the proposed method

Table 3.2. Distances of adjacent grid points on generated mosaic image [pixels (percentage from average.)] (Sequence 1).

page	average	maximum	minimum	std. dev.
left	338.5(100.0)	348.0(102.8)	331.0(97.8)	3.77(1.11)
right	337.6(100.0)	345.0(102.2)	331.1(98.1)	2.75(0.81)



(a) left page (b) right page

Figure 3.27. Distribution of distortion on mosaic image [pixels] (Sequence 1).

to deal with general surfaces with arbitrary curvature. For example, fitting NURBS surface to 3-D feature points as proposed in [YKKM04], instead of fitting 2-D curve to the 2-D projection of feature points, can be one of the promising solutions. This will remain as a future work in this study.

3.6. Conclusion

In this chapter, a geometric distortion-free video mosaicing method for curved documents is described. In the proposed method, extrinsic camera parameters, along with 3-D feature positions are estimated by applying structure from motion technique to the captured video. By fitting a parameterized surface to the 3-D features, the shape of the curved document is estimated. Using the estimated shape and extrinsic camera parameters, all the frame images are dewarped and synthesized on a virtual plane to generate a mosaic image without curvature distortion.

A prototype system based on the proposed method has been developed, and is tested in experiments on documents with curved surface. Quantitative evaluation on distortion shows that the resultant mosaic image is distortion-free. The advantage of the proposed method over conventional distortion correction using text-lines has been shown in the experiment on a target with few texts.

The limitation of the proposed method is that, since it carries out 3-D reconstruction by feature points, the accuracy of 3-D reconstruction is degraded in the boundary between pages in a bound book, which results in blurs in this area. Utilizing other types of image feature, e.g. line and arc segments, besides feature points can be a promising solution for this problem. Considering photo consistency among pixels which fall onto the same coordinate on the mosaic image can further improve the accuracy of 3-D reconstruction. Another limitation of the method is that it can only be applied to a generalized cylinder, i.e. a surface whose curvature lies along one direction. One way to extend the method to deal with more complex, general type of surfaces, is to fit NURBS surface to 3-D feature points as proposed in [YKKM04], instead of fitting 2-D curve to the 2-D projection of feature points. These will remain as future works in this study.

Chapter 4

Conclusion

In this thesis, a novel video mosaicing method which is capable of generating a geometric distortion-free mosaic image has been presented.

In general, video mosaicing is prone to two types of distortion. One is perspective distortion, which appears when the target document is not fronto-parallel to the camera's image plane. The other is curvature distortion, which is caused by projecting curved surface of the target document to the image plane of the camera.

This thesis first focused on a flat document, and proposed a perspective distortion-free video mosaicing method for flat documents. In this method, extrinsic camera parameters are estimated for each frame by applying structure from motion technique to the captured video. Using estimated extrinsic camera parameters, the method dewarps all the frame images and synthesizes them on a virtual fronto-parallel plane to generate a super-resolved mosaic image without perspective distortion. A novel user interface to guide the user to capture video sequence which gives efficient and accurate camera parameter estimation has also been proposed. Experiments on flat documents have been performed using a prototype system. In each experiment, the mosaic image has been proved to be distortion-free by quantitative analysis on distortion.

Then, this method for flat documents was extended to deal with curved documents. This extended method generates a virtually flattened mosaic image of a curved surface. In this method, first, extrinsic camera parameters, along with 3-D feature positions are estimated by structure from motion. Then, by

fitting a parameterized surface to the 3-D features, the shape of the curved document is estimated. Using the estimated shape and extrinsic camera parameters, all the frame images are dewarped to remove curvature distortion and synthesized on a virtual plane to generate a mosaic image without curvature distortion. This method was also tested in experiments on documents with curved surface. Quantitative evaluation on distortion has shown that the resultant mosaic image is distortion-free.

The limitation of the method is that, since it carries out 3-D reconstruction by feature points, the accuracy of 3-D reconstruction is degraded in the boundary between pages in a bound book, which results in distortion in this area. Utilizing other types of image feature, e.g. line and arc segments, besides feature points can be a promising solution for this problem. Considering photo consistency among pixels which fall onto the same coordinate on the mosaic image can further improve the accuracy of 3-D reconstruction. Another limitation of the method is that it can only be applied to a generalized cylinder, i.e. a surface whose curvature lies along one direction. One way to extend the method to deal with more complex, general type of surfaces, is to fit NURBS surface to 3-D feature points as proposed in [YKKM04], instead of fitting 2-D curve to the 2-D projection of feature points. These will remain as future works in this study.

What has been consistent throughout this study is a strong will to overcome the limitation of conventional imaging devices, and to create new photographic functionalities by combining traditional imaging devices with a novel computer vision algorithm. The challenge made in the development of the first method for flat targets is to create a portable and cheaper alternative of flatbed scanners which can be used anytime, anywhere. The challenge made in the development of the second method for curved targets is to create a smart camera which automatically expands a curved surface into a flat plane, which no other imaging devices have ever achieved. These achievements made in this study not only bring camera and imaging industries an alternative means to boost the resolution and to give new functionalities to cameras, but also give us a glimpse of the future of digital document solution and ubiquitous computing, and how computer vision technologies can contribute to it.

References

[BCR02] U. Bhosle, S. Chaudhuri, and S.D. Roy. A Fast Method for Image Mosaicing Using Geometric Hashing. *IETE J. of Research, Special Issue on Visual Media Processing*, Vol. 48, No. 3-4, pp. 317–324, 2002.

[BK02] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 9, pp. 1167–1183, 2002.

[BS01] M.S. Brown and W.B. Seales. Document Restoration Using 3D Shape: A General Deskewing Algorithm for Arbitrarily Warped Documents. In *Proc. IEEE Int. Conf. on Computer Vision*, Vol. 2, pp. 367–374, 2001.

[BT04] M.S. Brown and Y.C. Tsoi. Undistorting Imaged Print Materials using Boundary Information. In *Proc. Asian Conf. on Computer Vision*, Vol. 1, pp. 551–556, 2004.

[CDL03] H. Cao, X. Ding, and C. Liu. A Cylindrical Surface Model to Rectify the Bound Document Image. In *Proc. IEEE Int. Conf. on Computer Vision*, Vol. 1, pp. 228–233, 2003.

[CKH+98] N. Chiba, H. Kano, M. Higashihara, M. Yasuda, and M. Osumi. Feature-based Image Mosaicing. In *Proc. IAPR Workshop on Machine Vision Applications*, pp. 5–10, 1998.

[CZ00] D. Capel and A. Zisserman. Super-resolution enhancement of text image sequence. In *Proc. IAPR Int. Conf. on Pattern Recognition*, Vol. 1, pp. 600–605, 2000.

[FB81] 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.

[GS03] P. Grattoni and M. Spertino. A Mosaicing Approach for the Acquisition and Representation of 3D Painted Surfaces for Conservation and Restoration Purpose. *Machine Vision and Applications*, Vol. 15, No. 1, pp. 1–10, 2003.

[HCBH00] C.T. Hsu, T.H. Cheng, R.A. Beuker, and J.K. Hong. Feature-based Video Mosaicing. In *Proc. IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 887–890, 2000.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, pp. 147–151, 1988.

[IP91] M. Irani and S. Peleg. Improving Resolution by Image Registration. *CVGIP: Graphical Models and Image Processing*, Vol. 53, No. 3, pp. 231–239, 1991.

[JKH02] K. Jung, K.I. Kim, and J.H. Han. Text extraction in real scene images on planar planes. In *Proc. IAPR Int. Conf. on Pattern Recognition*, pp. 469–472, 2002.

[Kan98] K. Kanatani. Geometric Information Criterion for Model Selection. *Int. J. of Computer Vision*, Vol. 26, No. 3, pp. 171–189, 1998.

[KH03] D.W. Kim and K.S. Hong. Fast Global Registration for Image Mosaicing. In *Proc. IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 295–298, 2003.

[LDD05] J. Liang, D. DeMenthon, and D. Doermann. Unwarping images of curved documents using global shape optimization. In *Proc. Int. Workshop on Camera-based Document Analysis and Recognition*, pp. 25–29, 2005.

[LDD06] J. Liang, D. DeMenthon, and D. Doermann. Camera-Based Document Image Mosaicing. In *Proc. IAPR Int. Conf. on Pattern Recognition*, pp. 476–479, 2006.

[LQST01] M. Lhuillier, L. Quan, H. Shum, and H.T. Tsui. Relief Mosaicing by Joint View Triangulation. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 785–790, 2001.

[MBLH01] G.K. Myers, R.C. Bolles, Q.T. Luong, and J.A. Herson. Recognition of text in 3-D scenes. In *Proc. Symposium on Document Image Understanding Technology*, pp. 85–99, 2001.

[MJ02] P.F. McLauchlan and A. Jaenicke. Image Mosaicing Using Bundle Adjustment. *Image and Vision Computing*, Vol. 20, pp. 751–759, 2002.

[PBCP96] W. Puech, A.G. Bors, J.M. Chassery, and I. Pitas. Mosaicing of Paintings on Curved Surfaces. In *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 44–49, 1996.

[Pil01] M. Pilu. Undoing Paper Curl Distortion Using Applicable Surfaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 1, pp. 67–72, 2001.

[Rot00] C. Rother. A new approach for vanishing point detection in architectural environments. In *Proc. British Machine Vision Conf.*, Vol. 1, pp. 382–391, 2000.

[SHK98] H. S. Sawhney, S. Hsu, and R. Kumar. Robust Video Mosaicing through Topology Inference and Local to Global Alignment. In *Proc. European Conference on Computer Vision*, Vol. 2, pp. 103–119, 1998.

[SKYT02] 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. In *Int. J. of Computer Vision*, Vol. 47, No. 1-3, pp. 119–129, 2002.

[Sze94] R. Szeliski. Image Mosaicing for Tele-Reality Applications. In *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 230–236, 1994.

[TMHF99] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment a modern synthesis. In *Proc. Int. Workshop on Vision Algorithms*, pp. 298–372, 1999.

[Tsa86] R.Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 364–374, 1986.

[TSTT00] S. Takeuchi, D. Shibuichi, N. Terashima, and H. Tominage. Adaptive Resolution Image Acquisition Using Image Mosaicing Technique from Video Sequence. In *Proc. IEEE Int. Conf. on Image Processing*, vol. 1, pp. 220–223, 2000.

[YKKM04] A. Yamashita, A. Kawarago, T. Kaneko, and K. Miura. Shape Reconstruction and Image Restoration for Non-flat Surfaces of Documents with a Stereo Vision System. In *Proc. IAPR Int. Conf. on Pattern Recognition*, Vol. 1, pp. 482–485, 2004.

List of Publications

Journal Papers

1. A. Iketani, A. Nagai, Y. Kuno, and Y. Shirai. Real-time surveillance system detecting persons in complex scenes. *Real-Time Imaging*, Vol. 7, No. 5, pp. 433-446, Oct. 2001.

2. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Super-Resolved Video Mosaicing for Documents by Camera Parameter Estimation. *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, Vol. J88-D-II, No. 8, pp. 1490-1498, Aug. 2005 (in Japanese) (Chapter 2).

3. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video Mosaicing for Curved Documents Based on Structure from Motion. *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, Vol. J90-D, No. 8, pp. 1900-1911, Aug. 2007 (in Japanese) (Chapter 2).

Technical Letters

1. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, N. Yokoya, and K. Yamada. High-resolution Video Mosaicing for Documents by Estimating Camera Parameters. *Information Technology Letters*, Vol. 2, pp. 163-164, Sep. 2003 (in Japanese) (Chapter 2).

2. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya.

Super-Resolved Video Mosaicing Based on Camera Parameter Estimation. *Information Technology Letters*, Vol. 3, pp. 165-168, Sep. 2004 (in Japanese) (Chapter 2).

International Conferences

1. A. Iketani, A. Nagai, Y. Kuno, and Y. Shirai. Detecting persons on changing background. In *Proc. IAPR Int. Conf. on Pattern Recognition*, pp. 74-76, Aug. 1998.

2. A. Iketani, Y. Kuno, N. Shimada, and Y. Shirai. Real-Time Surveillance System Detecting Persons in Complex Scenes. In *Proc. Int. Conf. on Image Analysis and Processing*, pp. 1112-1115, Sep. 1999.

3. T. Sato, S. Ikeda, M. Kanbara, A. Iketani, N. Nakajima, N. Yokoya, and K. Yamada. High-resolution video mosaicing for documents and photos by estimating camera motion. In *Proc. SPIE Electronic Imaging*, Vol. 5299, pp. 246-253, Jan. 2004 (Chapter 2).

4. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Super-resolved video mosaicing for documents by extrinsic camera parameter estimation. In *Proc. Int. Conf. on Computer Vision and Graphics*, pp. 327-336, Sep. 2004 (Chapter 2).

5. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video mosaicing for curved surface by 3D reconstruction using feature points. In *CD-ROM Proc. IEEE Int. Conf. on Computer Vision (ICCV2005)*, Demonstrations, Oct. 2005 (Chapter 3).

6. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Super-resolved video mosaicing for documents based on extrinsic camera parameter estimation. In *Proc. Asian Conf. on Computer Vision (ACCV2006)*, Vol. 2, pp. 101-110, Jan. 2006 (Chapter 2).

7. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Mobile video mosaicing system for flat and curved documents. In *Proc.*

Int. Workshop on Mobile Vision (IMV2006), pp. 78-92, May 2006 (Chapter 2, 3).

8. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video mosaicing for curved documents based on structure from motion. In *Proc. IAPR Int. Conf. on Pattern Recognition (ICPR2006)*, Vol. 4, pp. 391-396, Aug. 2006 (Chapter 3).

9. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video mosaicing for curved documents by structure from motion. In *ACM SIGGRAPH2006*, Sketches, Aug. 2006 (Chapter 3).

10. A. Iketani, T. Sato, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video mosaicing based on structure from motion for distortion-free document digitization. In *Proc. Asian Conf. on Computer Vision (ACCV2007)*, Vol. 2, pp. 73-84, Nov. 2007 (Chapter 3).

Domestic Conferences

1. A. Iketani and M. Hashimoto. Person Extraction and Shadow Elimination Using Background Subtraction of Color and Edges. In *Proc. the 2000 Institute of Electronics, Information and Communication Engineers, General Conference*, No. D-12-133, Mar. 2000 (in Japanese).

2. A. Iketani, N. Nakajima, R. Hiraike, and K. Yamada. Multiple Non-Rigid Object Tracking Using Backward Region Classification. In *Proc. the 2003 Institute of Electronics, Information and Communication Engineers, General Conference*, No. D-12-139, Mar. 2003 [IEICE Young Researcher's Award] (in Japanese).

3. A. Iketani, N. Nakajima, T. Sato, S. Ikeda, M. Kanbara, N. Yokoya, and K. Yamada. Video Mosaicing and Superresolution with Handheld Camera. In *Proc. Forum on Information Technology (FIT2003)*, Vol. 3, No. I-028, Sep. 2003 [FIT Young Researchers Award] (in Japanese) (Chapter 2).

4. A. Iketani, N. Nakajima, T. Sato, S. Ikeda, M. Kanbara, N. Yokoya, and K. Yamada. Video Mosaicing and Super Resolution for Documents by Camera Parameter Estimation. In *Technical Report of the Institute of Electronics, Information and Communication Engineers of Japan, Pattern Recognition and Media Understanding (PRMU)*, PRMU2003-223, pp. 49-54, Feb. 2004 [PRMU Award] (in Japanese) (Chapter 2).

5. A. Iketani, N. Nakajima, T. Sato, S. Ikeda, M. Kanbara, N. Yokoya, and K. Yamada. Super-Resolution Video Mosaicing for Documents by Camera Parameter Estimation. In *Proc. Meeting on Image Recognition and Understanding (MIRU2004)*, Vol. 1, pp. 505-510, July 2004 (in Japanese) (Chapter 2).

6. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Super-resolved Video Mosaicing for Plane Objects by Extrinsic Camera Parameter Estimation. In *Proc. Symposium on Pattern Measurement*, pp. 13-20, Nov. 2004 (in Japanese) (Chapter 2).

7. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video Mosaicing for Curved Surface by 3-D Reconstruction Using Feature Points. In *Proc. the 2005 Institute of Electronics, Information and Communication Engineers, General Conference*, No. D-12-12, Mar. 2005 (in Japanese) (Chapter 3).

8. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Super-Resolution Video Mosaicing for Documents by Camera Parameter Estimation. In *Proc. Meeting on Image Recognition and Understanding (MIRU2005)*, pp. 1638-1639, Jul. 2005 (in Japanese) (Chapter 2).

9. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video Mosaicing for Curved Surface Based on Structure from Motion. In *Technical Report of the Institute of Electronics, Information and Communication Engineers of Japan, Pattern Recognition and Media Understanding (PRMU)*, PRMU2005-203, pp. 7-12, Feb. 2006 (in Japanese) (Chapter 3).

10. T. Sato, A. Iketani, S. Ikeda, M. Kanbara, N. Nakajima, and N. Yokoya. Video Mosaicing for Curved Documents Based on Structure from Motion. In *Proc. Meeting on Image Recognition and Understanding (MIRU2006)*, pp. 98-105, Jul. 2006 (in Japanese) (Chapter 3).

Awards

1. Young Researcher's Award of Institute of Electronics, Information and Communication Engineers (IEICE), 2003.

2. Young Researcher Award of Forum on Information Technology (FIT), 2003.

3. Award of Pattern Recognition and Media Understanding (PRMU), 2004.