工學博士學位論文

Joint Rectification and Stitching of Images Formulated as Camera Pose Estimation Problems

카메라 위치 예측을 이용한 영상 Rectification 및 Stitching 방법

2015 年 8 月

서울大學校 大學院

電氣컴퓨터工學部

安 宰 賢

ABSTRACT

This dissertation presents a study of image rectification and stitching problems formulated as camera pose estimation problems. There have been many approaches to the rectification and/or stitching of images for their importance in image processing and computer vision areas. This dissertation adds a new approach to these problems, which finds appropriate optimization problems whose solutions give camera pose parameters for the given problems. Specifically, the contribution of this dissertation is to develop (i) a new optimization problem that can handle image rectification and stitching in a unified framework through the pose estimation formulation, and (ii) a new approach to planar object rectification problem which is also formulated as an optimal homography estimation problem.

First, a unified framework for the image rectification and stitching problem is studied, which can handle both assumptions or conditions that (i) the optical center of camera is fixed or (ii) the camera captures a plane target. For this, the camera pose is modeled with six parameters (three for the rotation and three for the translation) and a cost function is developed that reflects the registration errors on a reference plane (image stitching results). The designed cost function is effectively minimized via the Levenberg-Marquardt algorithm. From the estimated camera poses, the relative camera motion is computed: when the optical center is moved (i.e., the camera motion is large), metric rectification is possible and thus provides rectified composites as well as camera poses are obtained.

Second, this dissertation presents a rectification method for planar objects using line segments which can be augmented to the previous problem for further rectification or performed independently to single images when there are planar objects in the image such as building facades or name cards. Based on the 2D Manhattan world assumption (i.e., the majority of line segments are aligned with principal axes), a cost function is formulated as an optimal homography estimation problem that makes the line segments horizontally or vertically straight. Since there are outliers in the line segment detection, an iterative optimization scheme for the robust estimation is also developed.

The application of the proposed methods is the stitching of many images of the same scene into a high resolution image along with its rectification. Also it can be applied to the rectification of building facades, documents, name cards, etc, which helps the optical character recognition (OCR) rates of texts in the scene and also to improve the recognition of buildings and visual qualities of scenery images. In addition, this dissertation finally presents an application of the proposed method for finding boundaries of document in videos for mobile device based application. This is a challenging problem due to perspective distortion, focus and motion blur, partial occlusion, and so on. For this, a cost function is formulated which comprises a data term (color distributions of the document and background), boundary term (alignment and contrast errors after the contour of the documents is rectified), and temporal term (temporal coherence in consecutive frames).

Key words: image stitching, image rectification, camera pose estimation, docu-

ment detection and segmentation

Student number: 2011-30241

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Image rectification and/or stitching of images are important problems in image processing and computer vision areas. Image stitching is a process that generates a high resolution and/or large field-of-view (FoV) image from the overlapping views of a scene. The image stitching algorithms are usually based on a (planar) homography model, which is a geometric relationship between two images when (i) the optical center of a camera is fixed (fixed-optical-center case) or (ii) the images contain the same plane (plane-target case) [5]. Based on these conditions, a huge number of algorithms have been developed to overcome the physical limitations of consumer cameras (such as small FoV and low resolution). Especially, the first condition is usually satisfied when users take pictures of distant targets, and this condition simplifies camera models. Therefore, many researchers focused on this case and a lot of practical systems were developed [6]. However, compared with the large amount of researches on the first case, the work on the second case (capturing the overlapping multiple views of the same plane) is rather limited and document oriented [7, 8]. Many methods addressing the second case were based on the sequential registration using pairwise homographies, which is a greedy algorithm that suffers from error propagations.

The image rectification problem can be interpreated as fronto-parallel view reconstruction and there have been many researches in this problem. One of application areas of image rectification is the optical character recognition (OCR), because conventional OCR engines were developed for the flat-bed scanned images and thus does not work well for camera-captured (experiencing perspective distortions) images. Therefore, many algorithms have been proposed to rectify input images based on the properties of documents [7, 9, 10]. Rectification is also importance in augmented reality (AR) and tracking applications. Since there is ambiguity in the rectification of general targets, they exploited inertia sensors [11] or simply assumed that the user provides a fronto-parallel views at the first frame [12]. Image rectification from a single image is also an important problem in many areas. There have been a lot of researches for the development of rectification algorithms that assume the existence of straight parallel lines from which vanishing points can be computed, or orthogonal structure known to exist in the scene.

1.2 Contributions

While the conventional methods have been developed assuming either of the above mentioned conditions (fixed-optical-center or plane-target), this dissertation is based on the observation that both conditions are actually very similar ones and hence it is desirable to develop a unified framework that can handle both cases simultaneously.

In this dissertation, a unified framework for the image rectification and stitching is presented, which is formulated as camera pose estimation problems. Specifically, the proposed framework can align images geometrically only with the estimated poses, which can be further improved by analyzing the camera poses. For example, when images are from planar surfaces, i.e., the camera motion is large, the proposed method can reconstruct the surface up to similarity, so that a full and fronto-parallel view image can be obtained. Otherwise, it may suffer from skews. In the case of fixed optical centers, metric reconstruction is impossible and thus a better viewpoint is selected by minimizing the overall distortions [13].

Second, in this dissertation, a new rectification method for the planar targets based on line segments is proposed. Using the basic assumption on single image calibration that the majority of line segments are aligned with principal directions, the proposed method develops a cost function whose objective is to find an optimal homography that makes the line segments horizontally or vertically straight. Unlike the previous works, the proposed method does not need vanishing points/lines estimation and segmentation of planar objects. Moreover, the proposed method has low computation time than the previous line-based method [3] that compares all line segments for computing orthogonal line-pairs.

In summary, the contributions of this dissertation are as follows:

- A new framework that generates image stitching results for two different conditions is proposed.
- It performs metric rectification when the camera motion is large.
- A new method for the rectification of an image with planar object is proposed in a related framework.

- Some interesting applications are presented, using the proposed rectification method.

1.3 Homography between the i-th image and $\boldsymbol{\pi}_E$

The notations used in this dissertation are as follows. The internal matrix of the i-th camera is denoted as

$$\mathrm{K_i} = \begin{bmatrix} f_i & 0 & 0 \\ 0 & f_i & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1.1}$$

where a pin-hole camera model is adopted and f_i is the focal length. Also, its pose (external parameters) is denoted as

$$\begin{bmatrix} \mathrm{R_i} & \mathbf{t}_i \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \in SE(3) \tag{1.2}$$

where $\mathrm{R_i} \in \mathrm{SO}(3)$ is the rotation matrix and $\mathbf{t}_i = [t_{ix}, t_{iy}, t_{iz}]^\top \in \Re^3$ is a translation vector. For the parametrization of a rotation matrix, an exponential representation is adopted:

$$\mathrm{R_i} = \exp\left([\boldsymbol{\theta}_\mathrm{i}]_\times\right) \tag{1.3}$$

for a vector $\boldsymbol{\theta}_i = [\theta_{ix}, \theta_{iy}, \theta_{iz}]^\top$. Finally, the reference plane is given by $\boldsymbol{\pi}_E = \left(\mathbf{n}^\top, d\right)^\top$. Without the loss of generality, we set $\mathbf{n} = [0, 0, 1]^\top$ and $d = 0$, i.e., the reference plane is z-plane in the world coordinate system.

The camera matrix of the i-th image [5] is given by

$$\mathrm{P_i} = \mathrm{K_i R_i} \left[\mathrm{I}_{3\times 3} | -\mathbf{t}_\mathrm{i}\right], \tag{1.4}$$

and the relationship between a point $[X, Y, 0]^\top$ on the surface and its corresponding point on the i-th image $\mathbf{p} = [u, v]^\top$ is given by

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathrm{K_i R_i} \left[\mathrm{I}_{3\times 3} | - \mathbf{t}_\mathrm{i}\right] \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} \tag{1.5}$$

$$= \mathrm{K_i} \begin{bmatrix} \mathrm{R_i}\mathbf{e}_1 & \mathrm{R_i}\mathbf{e}_2 & -\mathrm{R_i}\mathbf{t}_\mathrm{i} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{1.6}$$

for some λ. Thus, the homography $\mathrm{H_i}$ between the i-th image and the reference plane is

$$\mathrm{H_i} = \mathrm{K_i} \begin{bmatrix} \mathrm{R_i}\mathbf{e}_1 & \mathrm{R_i}\mathbf{e}_2 & -\mathrm{R_i}\mathbf{t}_\mathrm{i} \end{bmatrix} \tag{1.7}$$

$$= \mathrm{K_i} \begin{bmatrix} \mathrm{R_i}\mathbf{e}_1 & \mathrm{R_i}\mathbf{e}_2 & \mathrm{R_i}\mathbf{e}_3 - \mathrm{R_i}\mathbf{t}_\mathrm{i} - \mathrm{R_i}\mathbf{e}_3 \end{bmatrix} \tag{1.8}$$

$$= \mathrm{K_i R_i} \left\{ \mathrm{I} - \left(\mathbf{t}_\mathrm{i} + \mathbf{e}_3\right) \mathbf{e}_3^\top \right\} \tag{1.9}$$

where $\mathbf{e}_i$ $(i = 1, 2, 3)$ are the unit vectors representing each direction. By applying the Matrix Inversion Lemma to (1.9):

$$\mathrm{H_i^{-1}} = \left\{ \mathrm{I} - \frac{\left(\mathbf{t}_\mathrm{i} + \mathbf{e}_3\right) \mathbf{e}_3^\top}{\mathbf{e}_3^\top \mathbf{t}_\mathrm{i}} \right\} \mathrm{R_i^\top K_i^{-1}}. \tag{1.10}$$

1.4 Structure of the dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents the development of image rectification and stitching in a unified framework through the pose estimation formulation. And in Chapter 3, image rectification from a single image of planar targets is presented which is also formulated as an optimal homography estimation problem and its mobile-based application for document detection and segmentation in Chapter 4. The dissertation is concluded in Chapter 5.

Chapter 2

A unified framework for automatic image stitching and rectification

2.1 Related works

Image stitching algorithms are based on a (planar) homography model, which is a geometric relationship between two images when (i) the optical center of a camera is fixed or (ii) the images contain the same plane [5]. Based on these conditions, a huge number of algorithms have been proposed to overcome the physical limitations of consumer cameras, such as small FoV and low resolutions.

Between the above two conditions allowing image stitching, many papers focused on the first one (fixed-optical-center case) [6]; it is probably because this condition is (easily) satisfied when users take pictures of distant targets and this also simplifies camera models. Among them, the authors in [1] proposed a fully automatic system that discovers matching relationship between the images and recognizes panoramas automatically. In [1], each camera (corresponding to each image) is parameterized with four variables (three for the rotation matrix and one for the focal length) which are estimated by minimizing pairwise registration errors. This method was shown to be very robust to photometric variations and non-ideal effects (e.g., radial distortions, moving objects, slight violations of the assumption, and so on). Recently, real-time methods that progressively build panoramic maps from a video sequence were also developed [14, 15]. On the other hand, the research on the second condition (plane-target case) is rather limited; most of them were developed for some specific applications such as document processing [7, 8]. For example, a feature-based stitching algorithm in [7] rectifies the input images using the features of documents (text-lines) before geometric registrations, and therefore, it works only for text-abundant cases. Another problem with the conventional methods for the plane-target case [8] is that they are based on pairwise registration: registration errors between the images are accumulated by adding new images and the algorithm is likely to be stuck to local minima as shown in Fig. 2.1-(c). To address this problem, image mosaic methods [16–19] that deal with a combination of rotational and translational motions of cameras were developed. However, they cannot estimate the camera poses for metric rectification.

2.2 Proposed cost function and its optimization

Conventional image stitching methods have been developed assuming either of two conditions, and users have to know which condition is more appropriate. On the other hand, the proposed algorithm is able to handle both cases under same frame-

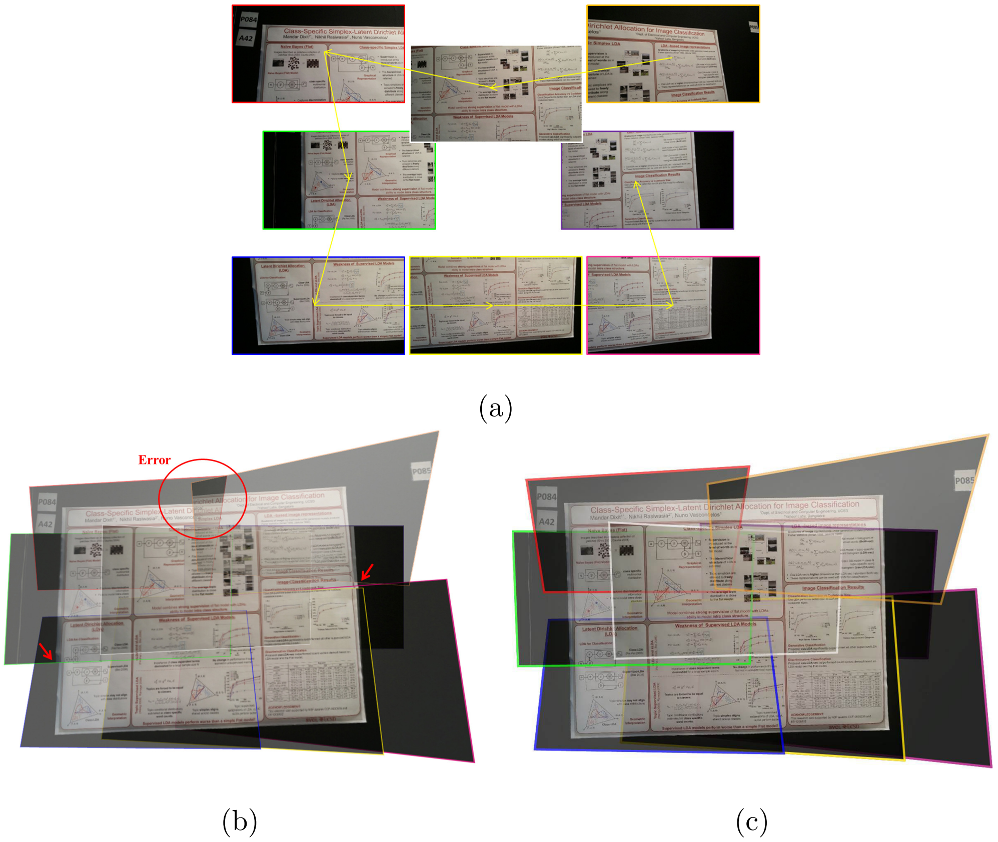


Figure 2.1: Comparison between the sequential registrations and the proposed method. (a) Eight input images capturing the same plane, (b) Image stitching result using the pairwise homographies. Note that the composite suffers from error accumulations, (c) Image stitching result using the proposed method. The proposed method tries to minimize the global registration errors, and it alleviates the error accumulation problem.

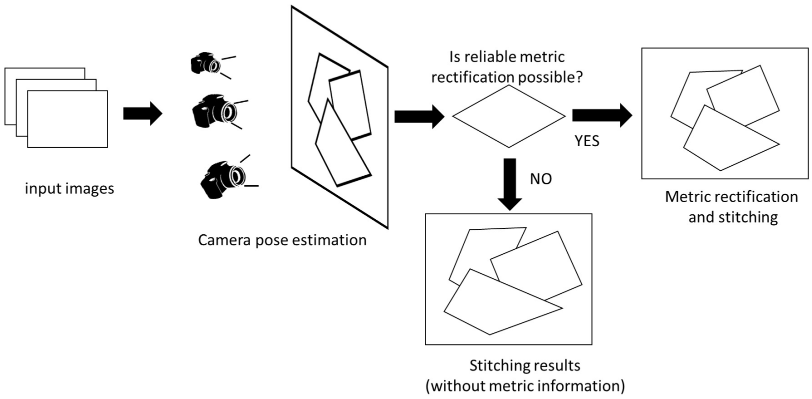

input images
Camera pose estimation
Is reliable metric rectification possible?
YES
NO
Metric rectification and stitching
Stitching results (without metric information)

Figure 2.2: Flowchart of the proposed algorithm.

work. Also, the proposed algorithm provides a general image stitching framework for the plane-target case (in a similar way as the conventional fixed-optical-center method [6]). The flowchart of the proposed method is shown in Fig. 2.2. Given the input images, the camera poses relative to a reference plane are computed by minimizing the proposed cost function which is based on the registration errors on the reference plane. After estimating the camera poses, the amount of camera motion compared with the scale of scene is computed. When this relative motion is large, the composition of the image on the reference plane is performed with the rectification of each image as illustrated in the first row of Fig. 2.3. Otherwise, (i.e., when the camera motion is small), since the metric rectification is impossible, visually pleasing results are composed instead, as shown in the second row of Fig. 2.3.

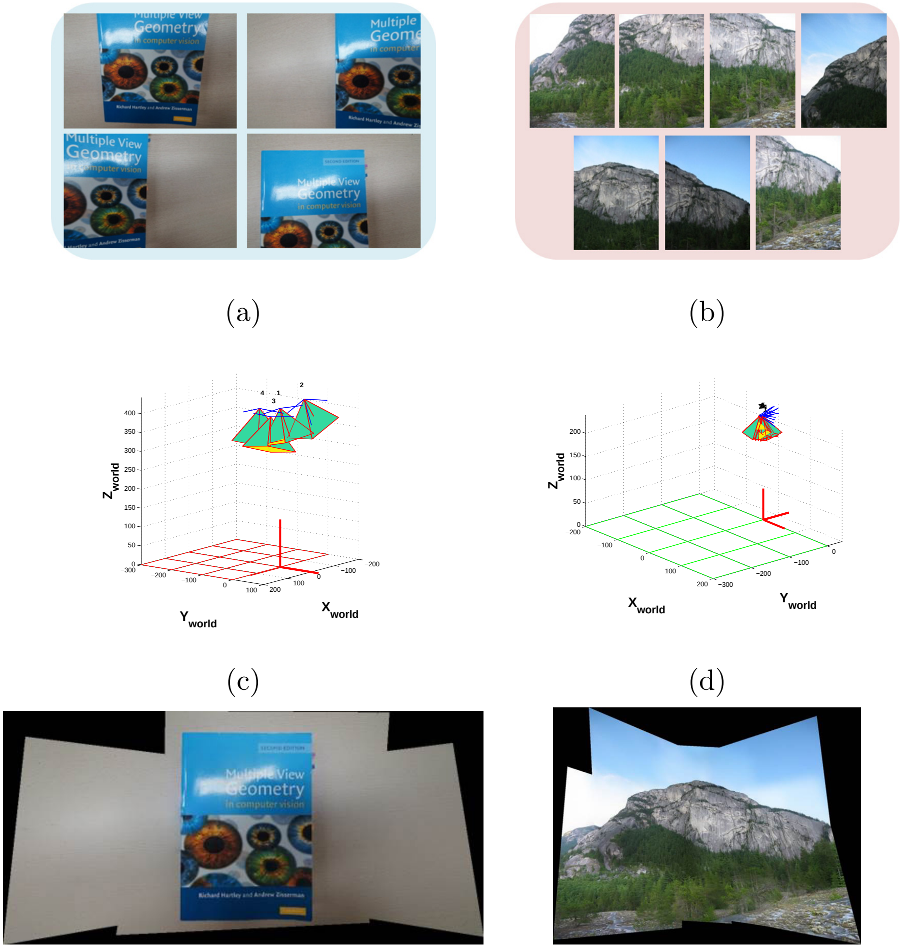


Figure 2.3: Results of the proposed algorithm. (a), (b) Input images, (c), (d) Estimated camera poses, (e), (f) Final results.

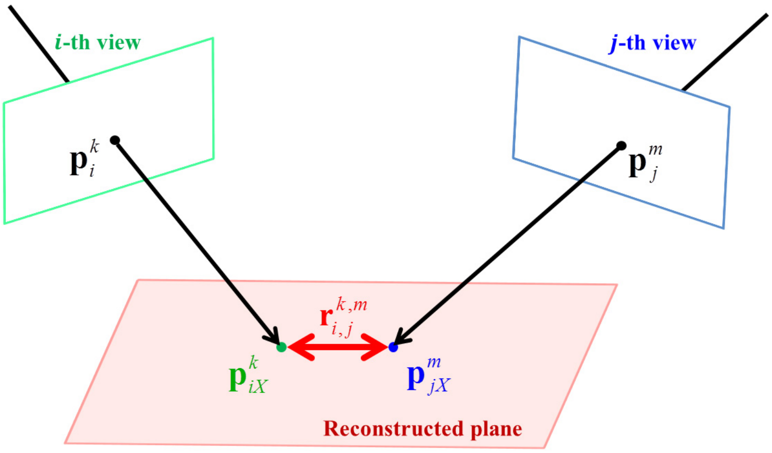


Figure 2.4: Illustration of notations in the proposed method.

2.2.1 Proposed cost function

For the registration of a set of images, this dissertation presents a new cost function reflecting the registration errors on a reference plane. To be precise, given a correspondence $\mathbf{p}_i^k \leftrightarrow \mathbf{p}_j^m$ between the i-th image and j-th image ($\mathbf{p}_i^k$ denotes the k-th feature in the i-image), the registration error $\mathbf{r}_{i,j}^{k,m}$ is given by

$$\mathbf{r}_{i,j}^{k,m} = \mathbf{p}_{iX}^k - \mathbf{p}_{jX}^m \tag{2.1}$$

where $\mathbf{p}_{iX}^k$ and $\mathbf{p}_{jX}^m$ are the projected points of $\mathbf{p}_i^k$ and $\mathbf{p}_j^m$ respectively, to the reference plane as illustrated in Fig. 2.4. That is,

$$\tilde{\mathbf{p}}_{iX}^k = \mathrm{H}_\mathrm{i}^{-1}\tilde{\mathbf{p}}_\mathrm{i}^\mathrm{k} \tag{2.2}$$

$$\tilde{\mathbf{p}}_{jX}^m = \mathrm{H}_\mathrm{j}^{-1}\tilde{\mathbf{p}}_\mathrm{j}^\mathrm{m} \tag{2.3}$$

where the tilde is used for the homogeneous representation of points.

The proposed cost function is given by the sum of registration errors for all the correspondences

$$e = \sum_{i=1}^{N} \sum_{j \in I(i)} \sum_{(k,m) \in F(i,j)} \left| \mathbf{r}_{i,j}^{k,m} \right|^2 \tag{2.4}$$

where N is the number of images, $I(i)$ is the set of images matched to the i-th image, and $F(i,j)$ is the set of correspondences between the i-th and j-th images. The correspondences are found by using the method in [1]: SIFT [20] features are extracted, correspondences are found by the nearest neighbor search, and inliers are selected using the random sample consensus (RANSAC) algorithm [21].

2.2.2 Optimization

The minimization of (2.4) is a non-linear least squares problem and the proposed method adopts the Levenberg-Marquardt algorithm [22]. For the implementation, the Jacobian of the registration error is derived analytically:

$$\frac{\partial \mathbf{r}_{i,j}^{k,m}}{\partial t} = \frac{\partial \mathbf{p}_{iX}^{k}}{\partial t} - \frac{\partial \mathbf{p}_{jX}^{m}}{\partial t}. \tag{2.5}$$

If the proposed method assumes that t is a parameter related to the i-th image (i.e., $t \in \{\theta_{ix}, \theta_{iy}, \theta_{iz}, t_{ix}, t_{iy}, t_{iz}\}$), then the first term in the right-hand side is denoted as

$$\frac{\partial \mathbf{p}_{iX}^{k}}{\partial t} = \frac{\partial \mathbf{p}_{iX}^{k}}{\partial \tilde{\mathbf{p}}_{iX}^{k}} \frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial t} \tag{2.6}$$

where

$$\frac{\partial \mathbf{p}_{iX}^{k}}{\partial \tilde{\mathbf{p}}_{iX}^{k}} = \frac{\partial \begin{bmatrix} x/z & y/z \end{bmatrix}}{\partial \begin{bmatrix} x & y & z \end{bmatrix}} = \begin{bmatrix} 1/z & 0 & -x/z^2 \\ 0 & 1/z & -y/z^2 \end{bmatrix}. \tag{2.7}$$

Since $\tilde{\mathbf{p}}_{iX}^{k} = \mathrm{H_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}}$ and $\mathrm{H_i^{-1}}$ can be obtained by (1.10), the proposed method can get $\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial t}$ for each parameter:

$$\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial \theta_{ix}} = \left\{\mathrm{I} - \frac{(\mathbf{t}_{\mathrm{i}} + \mathbf{e}_3)\,\mathbf{e}_3^{\top}}{\mathbf{e}_3^{\top}\mathbf{t}_{\mathrm{i}}}\right\} \frac{\partial \mathrm{R_i^{\top}}}{\partial \theta_{ix}} \mathrm{K_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}} \tag{2.8}$$

$$\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial \theta_{iy}} = \left\{\mathrm{I} - \frac{(\mathbf{t}_{\mathrm{i}} + \mathbf{e}_3)\,\mathbf{e}_3^{\top}}{\mathbf{e}_3^{\top}\mathbf{t}_{\mathrm{i}}}\right\} \frac{\partial \mathrm{R_i^{\top}}}{\partial \theta_{iy}} \mathrm{K_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}} \tag{2.9}$$

$$\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial \theta_{iz}} = \left\{\mathrm{I} - \frac{(\mathbf{t}_{\mathrm{i}} + \mathbf{e}_3)\,\mathbf{e}_3^{\top}}{\mathbf{e}_3^{\top}\mathbf{t}_{\mathrm{i}}}\right\} \frac{\partial \mathrm{R_i^{\top}}}{\partial \theta_{iz}} \mathrm{K_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}} \tag{2.10}$$

and

$$\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial t_{ix}} = -\frac{1}{t_{iz}} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathrm{R_i^{\top}K_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}} \tag{2.11}$$

$$\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial t_{iy}} = -\frac{1}{t_{iz}} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathrm{R_i^{\top}K_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}} \tag{2.12}$$

$$\frac{\partial \tilde{\mathbf{p}}_{iX}^{k}}{\partial t_{iz}} = \frac{1}{t_{iz}^2} \begin{bmatrix} 0 & 0 & t_{ix} \\ 0 & 0 & t_{iy} \\ 0 & 0 & 1 \end{bmatrix} \mathrm{R_i^{\top}K_i^{-1}}\tilde{\mathbf{p}}_{\mathrm{i}}^{\mathrm{k}}. \tag{2.13}$$

In the optimization, without the loss of generality, the proposed method assumes that the first camera is on the z-axis, i.e., $t_{1x} = t_{1y} = 0$.

2.2.3 Relation to the model in [1]

From (1.9) and (1.10), the proposed method can get pairwise homography between two views:

$$\mathrm{H_jH_i^{-1}} = \mathrm{K_jR_j}\left\{\mathrm{I} - \frac{(\mathbf{t}_{\mathrm{i}} - \mathbf{t}_{\mathrm{j}})\,\mathbf{e}_3^{\top}}{\mathbf{e}_3^{\top}\mathbf{t}_{\mathrm{i}}}\right\}\mathrm{R_i^{T}K_i^{-1}}, \tag{2.14}$$

which reduces to

$$\mathrm{K_j R_j R_i^\top K_i^{-1}} \tag{2.15}$$

when the optical center is fixed (i.e., $\mathbf{t}_i = \mathbf{t}_j$), which is the same model used in [1].

2.3 Post-processing

By minimizing the cost function in (2.4), the proposed method can estimate camera poses. With the camera pose, the proposed method first determines whether the camera motion is large or not. When the camera motion is large, the proposed method rectifies the images with the available structure information that can be obtained from the large camera motion. Conversely, when the camera motion is very small or the optical centers are fixed, the metric rectification is impossible and it would be better to provide stitching result without the metric rectification. In this section, this dissertation presents the criterion on this decision and our viewpoint selection method.

2.3.1 Classification of the conditions

From the estimated camera poses, the maximum amount of camera motion A_m is given by

$$A_m = \max_{1 \le i < j \le N} \left\| \mathbf{t}_i - \mathbf{t}_j \right\|_2 . \tag{2.16}$$

Because (2.16) is proportional to the scale of the scene, we also estimate the scale A_s:

$$A_s = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{t}_i^\top \mathbf{e}_3|, \tag{2.17}$$

which is the average distance between the cameras and the reference plane. Based on (2.16) and (2.17), our criterion is given by

$$\Gamma = \frac{A_m}{A_s}. \tag{2.18}$$

When $\Gamma < \tau$ (i.e., camera motion is small compared with the scale of the scene), metric rectification is not reliable and we adopt a view-point selection method [13]. Otherwise (i.e., $\Gamma > \tau$), we consider image on reference plane as metric rectification results.

2.3.2 Skew removal

For the skew removal, the proposed method computes a up-vector $\mathbf{u}$ based on the assumption that people do not severely twist the camera pose relative to the horizon when taking pictures [1, 2, 23]. This process is illustrated in Fig. 2.5 and it is formulated as a minimization problem:

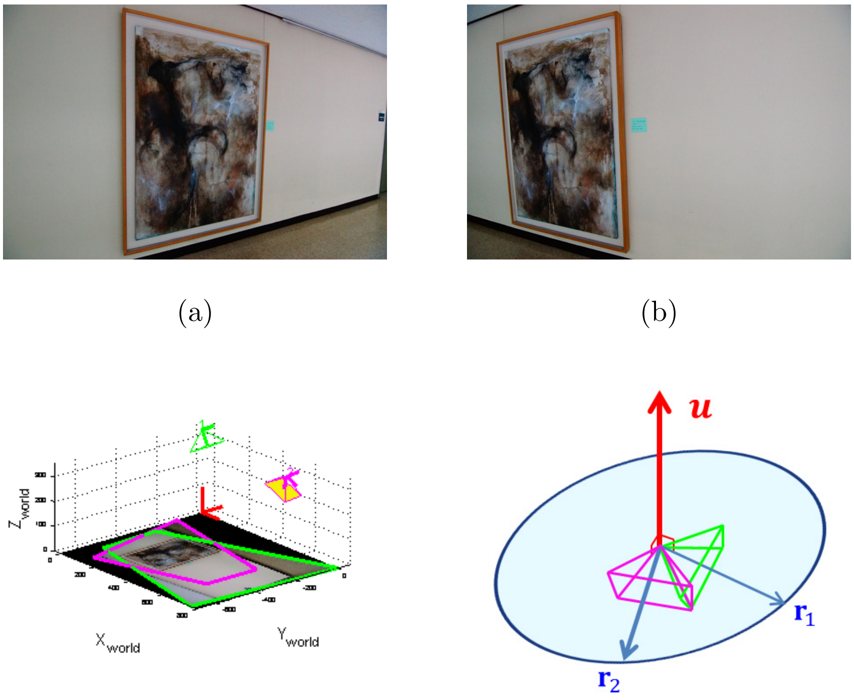
$$\mathbf{u} = \arg\min_{\|\mathbf{r}\|=1} \sum_{i=1}^{N} \left(\mathbf{r}^\top \mathbf{r}_i\right)^2 \tag{2.19}$$

$$= \arg\min_{\|\mathbf{r}\|=1} \mathbf{r}^\top \sum_{i=1}^{N} \left(\mathbf{r}_i \mathbf{r}_i^\top\right) \mathbf{r} \tag{2.20}$$

where $\mathbf{r}_i$ is the first row of the i-th rotation matrix (i.e., $\mathbf{e}_1^\top \mathrm{R_i}$). Therefore, $\mathbf{u}$ is the smallest eigenvector of the scatter matrix spanned by the x-directions of cameras as shown in Fig. 2.5-(c). Finally, the proposed method removes skews by applying the global rotation

$$\mathrm{R_g} = \left[\left[\mathbf{u}'\right]_\times \mathbf{e}_3, \mathbf{u}', \mathbf{e}_3\right] \tag{2.21}$$

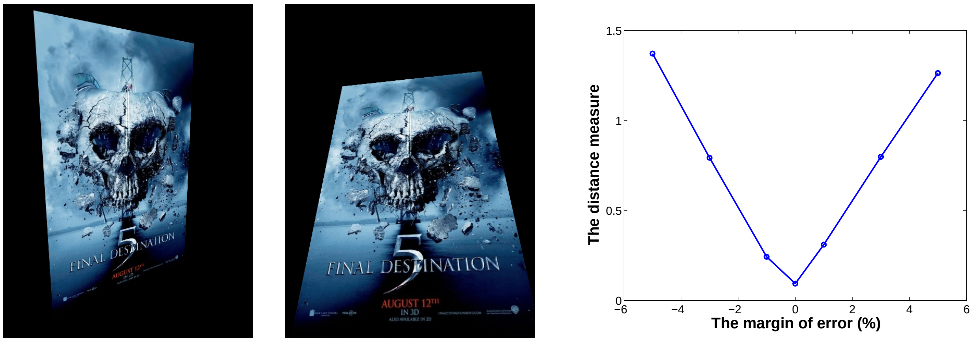
where $\mathbf{u}'$ is obtained by using the Gram-Schmidt orthonormalization.



(a) (b)

(c) (d)

Figure 2.5: Up-vector computation for skew removal in the composites. (a), (b) Two input images, (c) Visualization of estimated parameters, (d) Illustration of a up-vector **u**.



(a) (b) (c)

Figure 2.6: (a), (b) Synthesized image pair, (c) Average value of (2.22) for the 275 pairs of synthesized images according to the margin of error.

2.4 Experimental results

Metric rectification with unknown focal lengths $\{f_i\}$ can only be achieved with rich camera motions and low noise [24]. Therefore, rather than estimating these values, the focal length information in the exchangeable image file format (EXIF) is used, which is usually available on commercial cameras. In the experiments, the blending method [1, 25] is applied to reduce the photometric errors. The proposed method was implemented on a personal computer with an AMD Phenom(TM) II X6 1055T Processor running at 2.8 GHz and 4 GB RAM. The computation time depends on several factors such as the resolution of input images, the number of input images, and the amount of correspondences between images. However, it usually takes 1∼2 seconds in handling four 640 × 480 images.

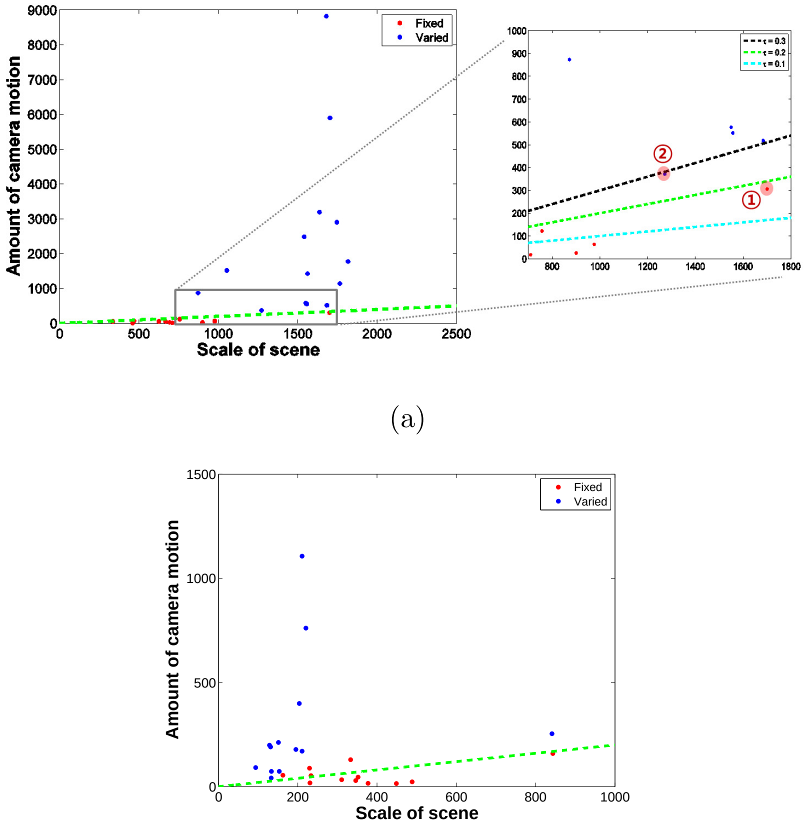
2.4.1 Quantitative evaluation on metric reconstruction performance

In this section, the metric rectification performance of our algorithm is evaluated. For this, synthetic pairs are built as shown in Fig. 2.6-(a) and (b): each image is synthesized with randomly chosen viewpoint and focal length. Given a pair, the homography in (1.10) with the proposed algorithm is estimated and it is compared with the ground truth. The distance measure between the ground truth homography H_{G} and the estimated one $\hat{\mathrm{H}}_E$ is given by

$$D = \min_{\mathbf{\Lambda}} \sqrt{\frac{1}{|R|} \sum_{\mathbf{x} \in R} \left| g\left(\mathrm{H}_{\mathrm{G}}^{-1} \tilde{\mathbf{x}}\right) - g\left(\Lambda \hat{\mathrm{H}}_{\mathrm{E}}^{-1} \tilde{\mathbf{x}}\right) \right|^2} \tag{2.22}$$

where R is a domain on which a given image is defined, $g(\cdot)$ converts a homogeneous vector to an inhomogeneous one (i.e., $g(\tilde{\mathbf{x}}) = \mathbf{x}$), and Λ is a similarity transform. Intuitively, the distance becomes small when the relationship between the estimated homography and the ground truth is given by a similarity transform. In order to find the optimal similarity transform, a robust method was adopted that finds the similarity transformation between two point patterns [26].

For the quantitative evaluation, 275 pairs are synthesized and the proposed method is evaluated. In order to evaluate the robustness to the errors in focal length, additional experiments are conducted by introducing the errors in focal length. To be specific, each pair is synthesized with focal length $f\left(1 + \frac{e_m}{100}\right)$ where the margin of error $e_m \in \{\pm 5, \pm 3, \pm 1, 0\}$, and try metric rectification with f. The results are summarized in Fig. 2.6-(c), which shows that our algorithm yields almost perfect metric reconstruction performance with the true focal length values, and its performance decreases as $|e_m|$ increases. However, registration errors are less than 1.5 pixels for a range of errors ($-5\% \leq e_m \leq +5\%$).



(a)

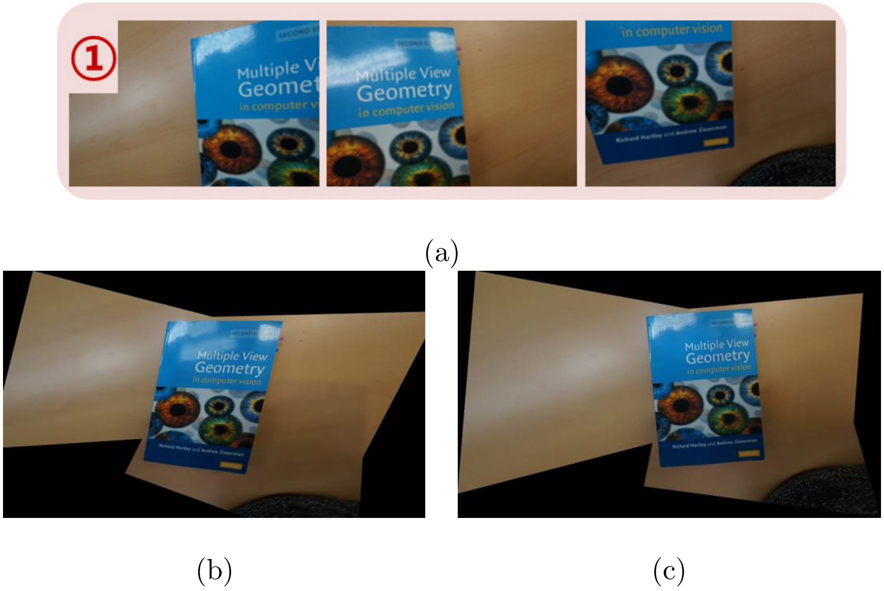
(b)

Figure 2.7: A scatter plot of the scale of scene A_s and the amount of camera motion A_m. (a) Scatter plot for full-size images, (b) Scatter plot for down-sampled images.

2.4.2 Determining the capturing environment

In this subsection, our criterion in (2.18) that classifies (i) fixed-optical-center case and (ii) plane-target case is tested. For this, a database was built that consists of 12 sets for the first case (fixed-optical-center) and 14 sets for the second (plane-target). The scatter plot of (2.16) and (2.17) is shown in Fig. 2.7-(a), which shows that two cases are well-separated with a proper threshold (i.e., the green dotted line represents $\Gamma = 0.2$). However, it should be noted that the classification is not a strict binary classification problem: the goal of the classification is to predict how reliable the metric rectification will be. For example, if a large threshold ($\tau = 0.3$) is chosen, our method will provide metric rectification results only when there is a large amount of motion. That is, we will have Fig. 2.8-(b) and 2.9-(b), rather than 2.8-(c) and 2.9-(c), for the relatively small amount of camera motions. On the other hand, if a small threshold ($\tau = 0.1$) is selected, some metric rectification results are likely to be un-reliable, because our method tries to perform metric rectification without enough structural information. Therefore, τ is set to 0.2, so that the proposed method provides visually pleasing results, however, we believe that the choice of τ can be changed according to the user preferences or application purposes.

Our criterion may depend on the quality of corresponding points (i.e., the number and/or distribution of corresponding points). In order to evaluate the robustness of our criterion, we conducted the experiments on down-sampled images (fewer correspondences and poor localization performances). The results are shown in Fig. 2.7-(b). As shown, there are three misclassifying pairs. However, the resolution of these three inputs are very low (about 150×100) and we can find that the criterion works robustly to other practical inputs.

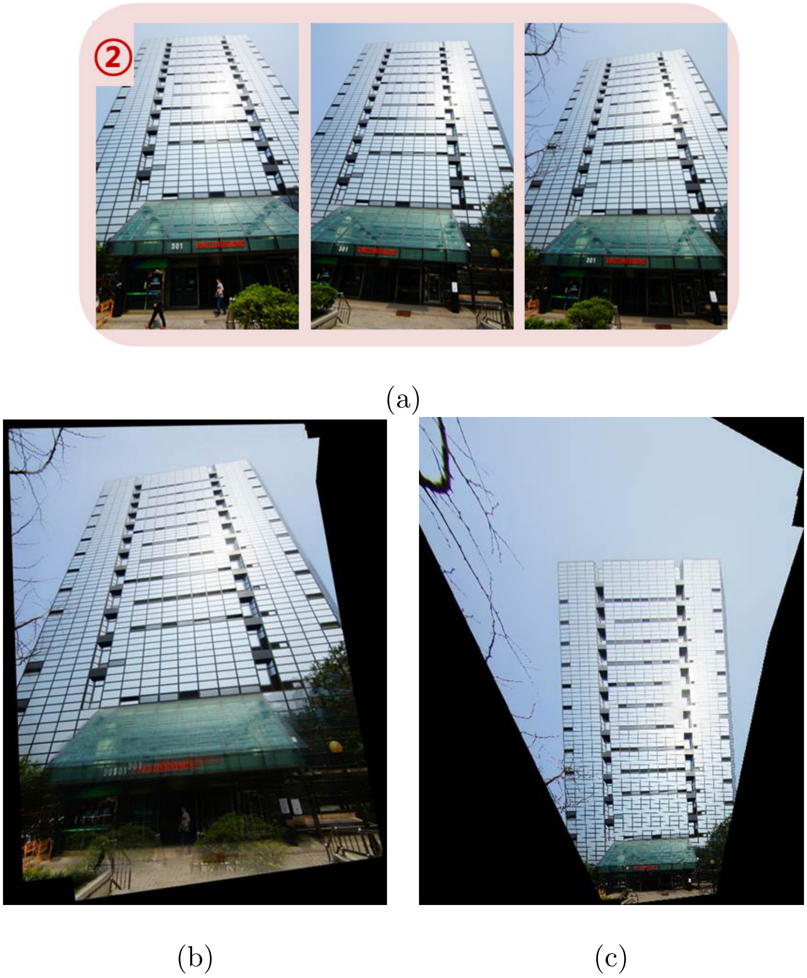


(a)

(b)

(c)

Figure 2.8: Case close to the decision boundary. (a) Input images, (b) Results for a large threshold ($\tau = 0.3$), (c) Results for the proposed threshold ($\tau = 0.1$).

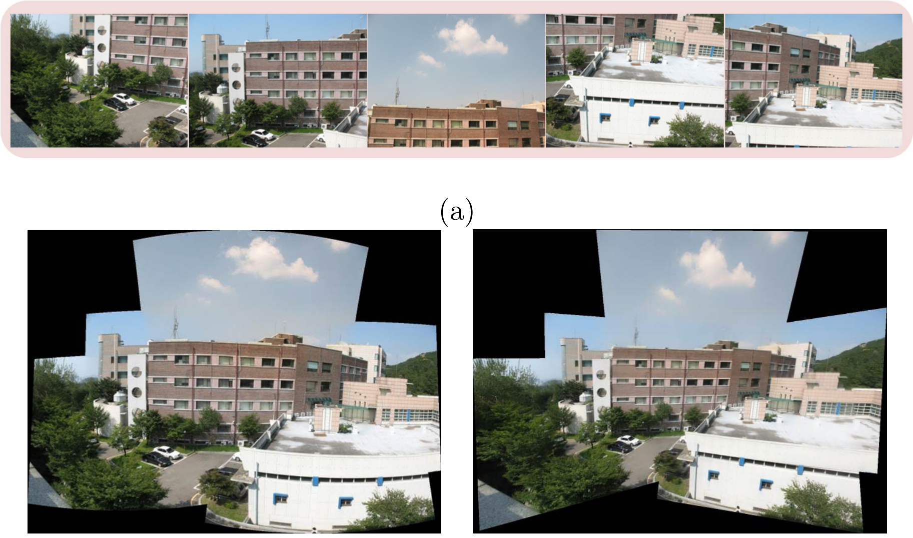


(a)

(b)

(c)

Figure 2.9: Case close to the decision boundary. (a) Input images, (b) Results for a large threshold ($\tau = 0.3$), (c) Results for the proposed threshold ($\tau = 0.1$).



(a)

(b)

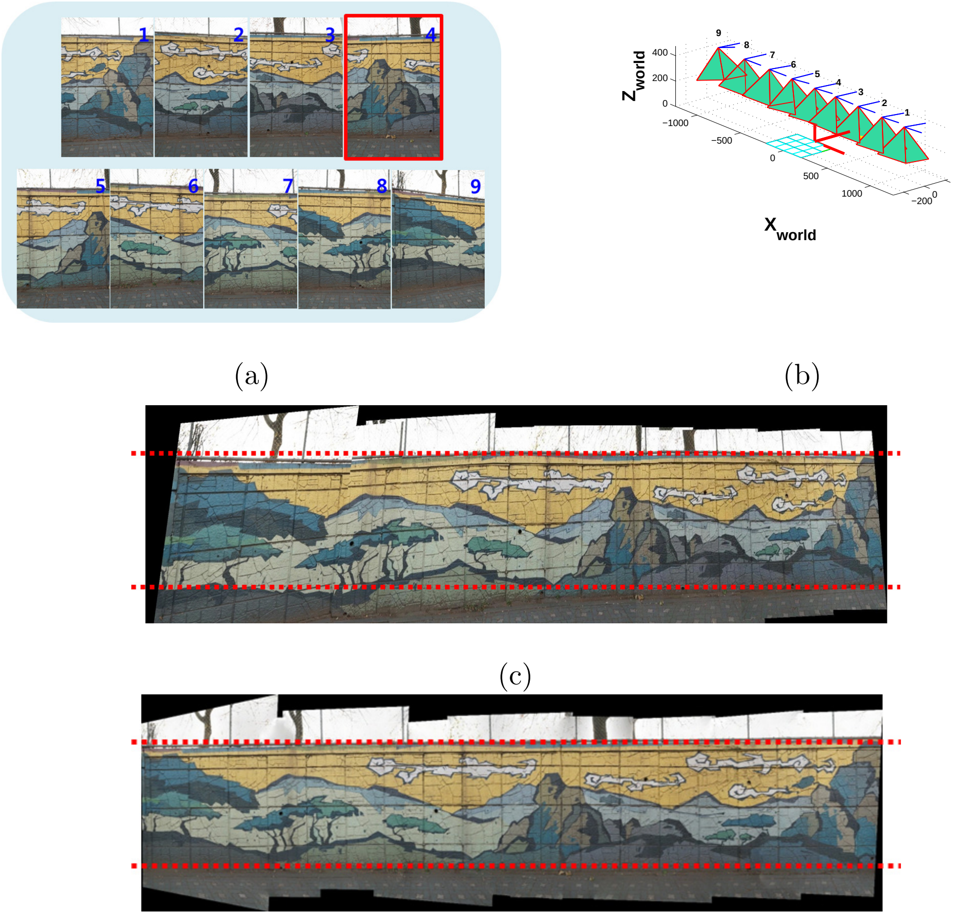
(c)

Figure 2.10: Comparison of our result with the conventional image stitching method. (a) Five input images, (b) Result of *Autostitch* [2], (c) Result of the proposed method.

2.4.3 Experiments on real images

When the camera center is fixed as shown in the second row of Fig. 2.3, our method addresses the same problem in [1] and thus yields similar results (See Sec. 2.2.3). Fig. 2.10-(a) shows five input images of fixed-optical-center cases. Fig. 2.10-(b) and (c) are image stitching results using *Autostitch* [1] and the proposed method respectively. Since our method adopts a viewpoint selection method in [13], the final composite is not the same. Note that *Autostitch* adopted the spherical projection and straight lines are not preserved.

However, in the case of plane-target, the conventional image stitching system such as [1] does not work. Therefore the proposed method is compared with a method based on the sequential registration using pairwise homographies [5]. Fig. 2.11-(a) shows nine input images from a moving camera, and Fig. 2.11-(b) shows estimated camera poses with respect to the reference plane. Fig. 2.11-(c) and (d) are image stitching results using the conventional approach and the proposed method respectively. In order to highlight the perspective distortions, the red horizontal dotted lines are overlaid. As can be seen, the proposed method generates a fronto-parallel view successfully.

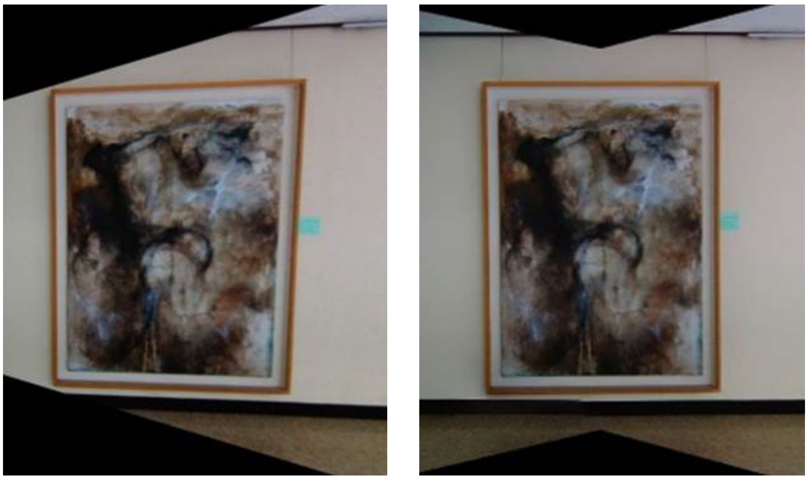


(a) (b)

(c)

(d)

Figure 2.11: Image stitching result on the images captured by a moving camera. (a) Nine input images, (b) Estimated camera poses, (c) Image stitching result that considers the fourth image as a reference, (d) Image stitching result using the proposed method.



(a) (b)

Figure 2.12: Comparison of our result with the conventional rectification method. (a) Result of a single image based rectification method [3], (b) Result of the proposed method.

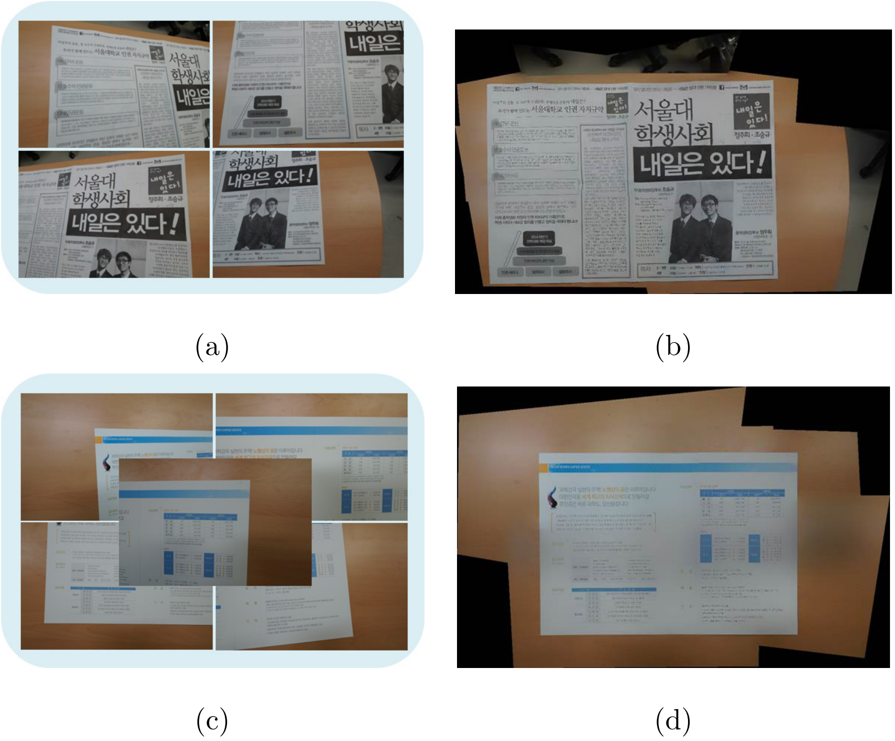
For the planar-target case, many methods were developed for the metric rectification. Among them, a single image based rectification method [3] finds the rectification transform by using the properties of line segments, and the rectification performance is compared with the method. Fig. 2.12-(a) shows the result by the conventional rectification method [3] where the input is the first image in Fig. 2.5-(a). Although this method can perform metric rectification, it cannot exploit the information from multiple views and thus yields less well rectified result compared to the proposed method as shown in Fig. 2.12-(b).

2.4.4 Applications to document image stitching and more results

In addition to image stitching, our method can be applied to document image processing [7, 8]. As shown in Fig. 2.13, the proposed method yields rectified results without any text-specific information. Another example (without blending) can be found in Fig. 2.1-(c). Full-resolution and more results can be found at `http://ispl.snu.ac.kr/jhahn/plane_stitching`.

2.5 Summary

In this section, a unified framework to the image stitching and rectification problem was proposed. While the conventional methods have been developed independently based on either case: (i) fixed-optical-center case or (ii) plane-target case, the proposed framework is able to handle both cases in the same framework. For this, six parameters of each camera was estimated by developing a cost function defined on the reconstructed plane, which is minimized via the Levenberg-Marquardt algorithm. From the estimated camera poses, the proposed method performed metric



(a)

(b)

(c)

(d)

Figure 2.13: Image stitching results for document images. (a), (c) Input images, (b), (d) Image stitching results of the proposed method.

rectification when there are enough camera motions (or differences), otherwise, it yielded stitching results without rectification like the conventional methods.

Chapter 3

Rectification of planar targets based on line segments

3.1 Related works

Since planar surfaces are common and important targets in many computer vision applications, robust and efficient rectification of surfaces has been a widely researched topic. Some examples where the 2D rectification algorithm plays a crucial role to the overall performance are object tracking, optical character recognition (OCR), and augmented reality [9–12,27,28]. In these applications, many of the algorithms exploited application-specific features, and hence one cannot be straightforwardly applied to other applications. Also, some of them are based on sophisticated segmentation or optimization schemes, which cannot be used in time-critical applications. In this section, in order to alleviate these limitations, an efficient and effective method is presented that can be applied to various kinds of planar objects.

3.1.1 Rectification of planar objects

The conditions for metric rectification of planar targets have been extensively studied in the literature [5, 29]. However, automatic rectification of planar targets is still a challenging problem. Specifically, among a variety of conditions for the metric rectification (e.g., known rectangles, angles, parallel curves, and so on), it is not easy to choose the right conditions for the given images and to select appropriate algorithm: some images have dominant rectangles, others have parallel curves, and so on. Therefore, rather than developing a general rectification algorithm, many researchers focused on individual cases. Probably, the most widely studied case is the rectification of rectangle target, which is a common and useful target in many images. For example, Hua & Liu [30] developed a rectification method for rectangle targets: they first segmented business card image patches by applying a segmentation method [31], and the segmented region is fitted to a quadrangle. Under the similar assumptions, four boundaries of rectangles were detected with Hough transform in [32]. Also, there are rectification algorithms for other conditions such as the existence of coplanar repeated patterns and parallel planar curves [33, 34]. They showed interesting theoretical results, however, these application areas are somewhat limited because such patterns are not commonly found in the images. Some researchers tried to circumvent this problem by using additional hardware or user interactions: Lee et al. [11] addressed the rectification problem with the help of inertia sensors and others simply assumed that the users provide a fronto-parallel view at the first frame [12].

3.1.2 Rectification based on self calibration

The rectification problem is also addressed in a different way: self calibration algorithms [3,4,35–40] yield the pose of the camera and the rectification transform of the planar targets can be computed from the calibration results. For instance, Zhang et al. [35] developed a camera calibration method based on the properties of low-rank textures, which can remove radial distortions as well as perspective distortions in the images (there is also a method for images of a plane [41]). However, these methods have some limitations in that they work only for low-rank textures, require user interaction and their computation cost are high. Recently, self-calibration of a camera in a (3D) Manhattan world received a lot of attention [3,4,36–40], which can also be used for the 2D metric rectification. Actually, one of these methods provides some successful results on planar targets [4]. However, these algorithms were developed for the 3D world and it may be sub-optimal for the 2D rectification problem. In the experimental section (Sec. 3.3), the proposed method will be compared with the state-of-the-art 3D self-calibration methods in terms of accuracy and computation cost for the 2D problems.

3.2 Proposed rectification model

The proposed method is based on the 2D Manhattan world assumption, where a profusion of lines are aligned either horizontally or vertically on the plane. Based on this assumption, a cost function is developed that evaluates the alignments of lines, and the rectification transform is found by minimizing the cost function. Since the proposed method is based on (a kind of) the Manhattan world assumption and the cost function is formulated with camera parameters, it can be considered a self-calibration method. Main difference from the conventional 3D calibration methods is that the proposed algorithm is focused on the alignment of lines (instead of the estimation of vanishing points), and hence the algorithm works robustly for the images with unclear vanishing points. As shown in Figs. 3.6-3.9, vanishing points are not clear features for many interesting 2D targets.

The most similar approach to the proposed work may be the orthogonality based method in [3], which finds the rectification transform by using the pairs of orthogonal line segments. However, the proposed cost function is based on line alignments rather than pairwise orthogonality, and the cost function can be minimized efficiently. Experimental results on a range of planar objects (including building facades, documents, and signposts) show that the proposed method is not only efficient but also yields robust rectification performance compared with other methods [3, 4] based on the 3D Manhattan world assumption.

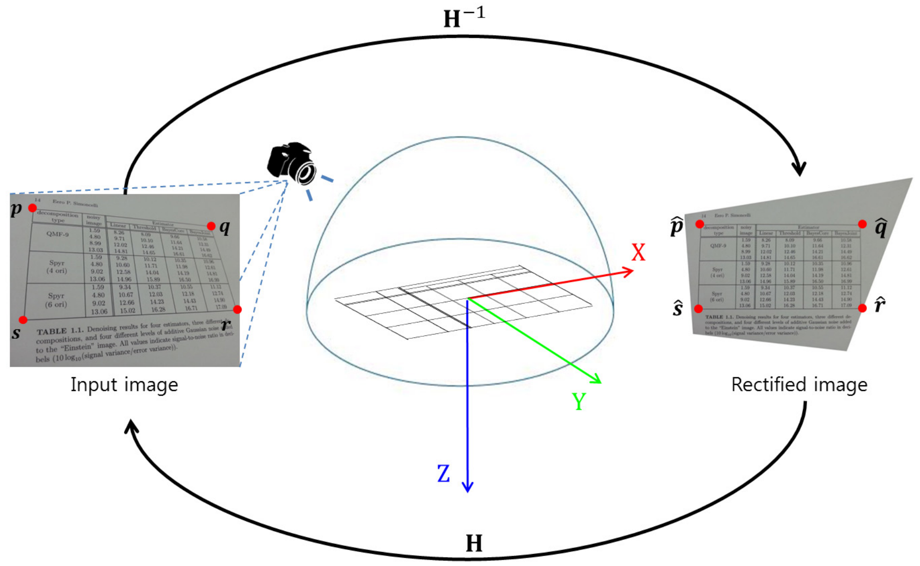


Figure 3.1: When the majority of line segments are aligned with principal axes, the camera-captured images can be rectified by finding a transform that makes the line segments horizontally or vertically aligned.

The 2D rectification problem is solved by estimating the internal and external parameters of a camera. For this, line segments in the image is first extracted and the homography is found that makes the majority of these line segments aligned with principal axes (see Fig. 3.1). For this, the translation vector $\mathbf{t}$ is set to

$$[0, 0, -\max(w, h)]^\top \tag{3.1}$$

without loss of generality, where w and h are the width and height of the input image respectively. Then, the metric rectification is equivalent to estimating f and $\boldsymbol{\theta} \in \Re^3$ from the detected line segments.

3.2.1 Optimization-based framework

Similar to other methods [3, 4], line segments are extracted using the LSD (line segment detector) method in [42]. Let us denote detected line segments as

$$\mathcal{L} = \{(\mathbf{u}, \mathbf{v})\}, \tag{3.2}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the homogeneous representation of two end-points of a line segment. Then, the rectification is formulated as an optimization problem defined on $\mathcal{L}$:

$$(\hat{\boldsymbol{\theta}}, \hat{f}) \;= \operatorname*{arg\,min}_{\boldsymbol{\theta}, f} \; E(\boldsymbol{\theta}, f; \mathcal{L}) + \lambda F(\boldsymbol{\theta}, f), \tag{3.3}$$

where the first term $E(\boldsymbol{\theta}, f; \mathcal{L})$ evaluates alignments to the principal axes when H^{-1} is applied to $\mathcal{L}$, the second term $F(\boldsymbol{\theta}, f)$ helps us to avoid trivial solutions by imposing the constraints on the focal length, and λ is a factor to control the balance between the two terms.

3.2.2 Cost function based on line segment alignments

The first term in (4.2) is defined as

$$E(\boldsymbol{\theta}, f; \mathcal{L}) = \sum_{(\mathbf{u},\mathbf{v})\in\mathcal{L}} w(\mathbf{u},\mathbf{v}) \times d_{\mu}^{2}(\mathrm{H}^{-1}\mathbf{u}, \mathrm{H}^{-1}\mathbf{v}), \tag{3.4}$$

where H^{-1} is given by (1.10) and $w(\cdot,\cdot)$ is the normalized weight (that sums to unity) of each segment. Since longer lines are more informative for estimating the rectification transform, the weights are defined as:

$$w(\mathbf{p},\mathbf{q}) \propto d^{2}(\mathbf{p},\mathbf{q}), \tag{3.5}$$

where $d(\cdot,\cdot)$ is a geometric distance between two homogeneous points ($\mathbf{p} = [p_1, p_2, p_3]^{\top}$ and $\mathbf{q} = [q_1, q_2, q_3]^{\top}$):

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1/p_3 - q_1/q_3)^2 + (p_2/p_3 - q_2/q_3)^2}. \tag{3.6}$$

In order to evaluate their alignments, the alignment error is also defined as

$$d_{\mu}(\mathbf{p},\mathbf{q}) = \min(|p_1/p_3 - q_1/q_3|, |p_2/p_3 - q_2/q_3|). \tag{3.7}$$

Even though (3.4) successfully measures the alignments of lines, its direct minimization may result in a trivial solution (by making the focal length f very large). Hence, a constraint is imposed that the focal length f should be similar to the image scale [4], i.e., the second term is designed as

$$F(\boldsymbol{\theta}, f) = \left(\frac{\max(a, f)}{\min(a, f)} - 1\right)^{2}, \tag{3.8}$$

where $a = \max(w, h)$.

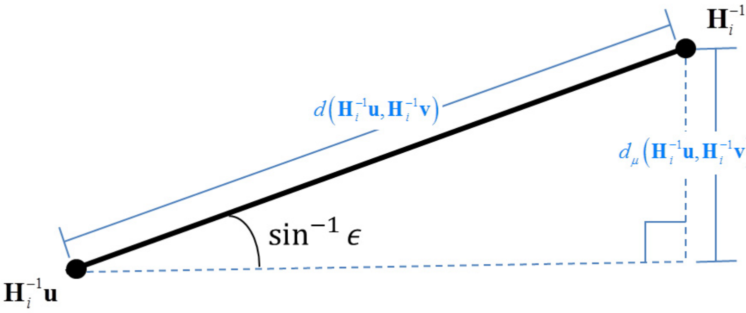

$\mathbf{H}_i^{-1}\mathbf{v}$
$d\left(\mathbf{H}_i^{-1}\mathbf{u}, \mathbf{H}_i^{-1}\mathbf{v}\right)$
$d_\mu\left(\mathbf{H}_i^{-1}\mathbf{u}, \mathbf{H}_i^{-1}\mathbf{v}\right)$
$\sin^{-1}\epsilon$
$\mathbf{H}_i^{-1}\mathbf{u}$

Figure 3.2: A Line segment with large angle, i.e., $\sin^{-1}\epsilon(\mathrm{H}^{-1}, \mathbf{u}, \mathbf{v}) \gg 0$, is considered an outlier.

3.2.3 Optimization

When the set of LSD results (4.1) does not contain outliers, the cost function (4.2) can be easily minimized via the Levenberg-Marquardt algorithm, because it consists of sum of square terms [22]. However, as shown in Fig. 3.3, LSD results usually contain lots of outliers, and hence an iterative approach is developed that incrementally filters out these outliers.

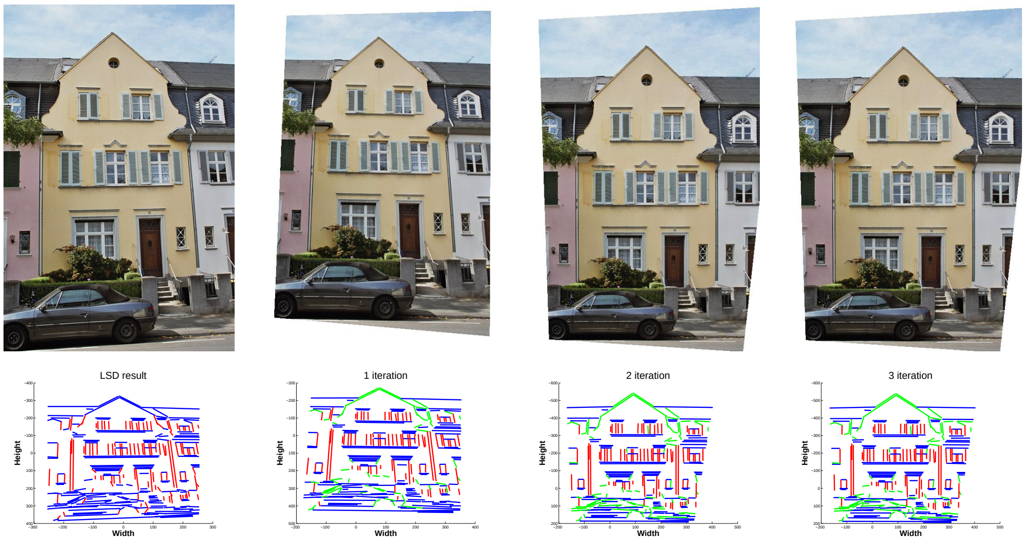
To be precise, for a current solution $\mathrm{H}_\mathrm{i}^{-1}$, a set of (potential) inliers is collected for the $(i + 1)$-th iteration, by finding the lines that have small $d_\mu(\mathrm{H}_\mathrm{i}^{-1}\mathbf{u}, \mathrm{H}_\mathrm{i}^{-1}\mathbf{v})$ compared to $d(\mathrm{H}_\mathrm{i}^{-1}\mathbf{u}, \mathrm{H}_\mathrm{i}^{-1}\mathbf{v})$ (refer to Fig. 3.2) as

$$\mathcal{L}_{i+1} = \left\{(\mathbf{u}, \mathbf{v}) \in \mathcal{L} | \; \epsilon(\mathrm{H}_\mathrm{i}^{-1}, \mathbf{u}, \mathbf{v}) < \tau_i\right\}, \tag{3.9}$$

where

$$\epsilon(\mathrm{H}_\mathrm{i}^{-1}, \mathbf{u}, \mathbf{v}) = \frac{d_\mu(\mathrm{H}_\mathrm{i}^{-1}\mathbf{u}, \mathrm{H}_\mathrm{i}^{-1}\mathbf{v})}{d(\mathrm{H}_\mathrm{i}^{-1}\mathbf{u}, \mathrm{H}_\mathrm{i}^{-1}\mathbf{v})}. \tag{3.10}$$

From $\mathcal{L}_{i+1}$, we can estimate H_{i+1}^{-1} by minimizing (4.2) defined on $\mathcal{L}_{i+1}$, with the Levenberg-Marquardt algorithm [22].



(a) (b) (c) (d)

Figure 3.3: Illustration of iterative approach. The first row shows current results (by applying H^{-1}) to the input image and the images in the 2nd row are corresponding LSD results. Blue and red are the lines close to the horizontal and vertical axes respectively, and green lines are outliers in the iteration. (a) Input, (b) After 1st iteration, (c) After 2nd iteration, (d) After 3rd iteration.

At initial stage, $\mathcal{L}_1 = \mathcal{L}$, and τ_i in (3.9) is selected so that only the minority of line segments are classified into outliers at each iteration:

$$\tau_i = \max\left(\sin\left(\frac{\pi}{60}\right), \min\left(\mu_i + k \times \sigma_i, \sin\left(\frac{\pi}{10}\right)\right)\right), \tag{3.11}$$

where μ_i and σ_i are the mean and standard deviation of (3.10) respectively on $\mathcal{L}_i$. The proposed method is iterated until the number of inliers becomes stable. Although the greedy approach does not guarantee the global optimum, it works well for a variety of inputs as will be demonstrated in the next section.

3.3 Experimental results

The proposed method is evaluated on a variety of images and compared it with two state-of-the-art techniques (i.e., upright adjustment [4] and SfAR [3]). Performances of the algorithms are evaluated in terms of the computation time and the amount of remaining geometrical distortions after the rectification [43]. In all the experiments, λ in (4.2) is set to 0.1, and k in (3.11) is set to 2.

3.3.1 Evaluation metrics

Geometric distortions are measured with four criteria: orthogonality, diagonal ratio, and length ratios for opposite sides. Let us assume that the rectification homography H^{-1} and four (manually annotated) corner points of an object in input image are given as $\mathbf{p}$, $\mathbf{q}$, $\mathbf{s}$, and $\mathbf{r}$ as illustrated in Fig. 3.1. Then, the diagonal ratio r_d is defined as the length ratio between two diagonals:

$$r_d = \max\left(\frac{d(\mathrm{H}^{-1}\mathbf{q}, \mathrm{H}^{-1}\mathbf{s})}{d(\mathrm{H}^{-1}\mathbf{p}, \mathrm{H}^{-1}\mathbf{r})}, \frac{d(\mathrm{H}^{-1}\mathbf{p}, \mathrm{H}^{-1}\mathbf{r})}{d(\mathrm{H}^{-1}\mathbf{q}, \mathrm{H}^{-1}\mathbf{s})}\right). \tag{3.12}$$

Also, two pairs of opposite sides must have the same length, and the ratio of two vertical sides is evaluated as

$$r_v = \max\left(\frac{d(\mathrm{H}^{-1}\mathbf{p}, \mathrm{H}^{-1}\mathbf{s})}{d(\mathrm{H}^{-1}\mathbf{q}, \mathrm{H}^{-1}\mathbf{r})}, \frac{d(\mathrm{H}^{-1}\mathbf{q}, \mathrm{H}^{-1}\mathbf{r})}{d(\mathrm{H}^{-1}\mathbf{p}, \mathrm{H}^{-1}\mathbf{s})}\right). \tag{3.13}$$

The ratio for the horizontal pair r_h is similarly defined. Finally, the orthogonality of four corners is measured. Specifically, for the upper left corner, the orthogonality is computed as

$$\theta_o = \cos^{-1}\left(\frac{\mathbf{l}_1^\top \mathrm{C}_\infty^* \mathbf{l}_2}{\sqrt{(\mathbf{l}_1^\top \mathrm{C}_\infty^* \mathbf{l}_1)(\mathbf{l}_2^\top \mathrm{C}_\infty^* \mathbf{l}_2)}}\right), \tag{3.14}$$

where $\mathrm{C}_\infty^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is the conic dual of the Euclidean coordinate system [5], and two lines are give by $\mathbf{l}_1 = \left(\mathrm{H}^{-1}\mathbf{p}\right) \times \left(\mathrm{H}^{-1}\mathbf{q}\right)$ and $\mathbf{l}_2 = \left(\mathrm{H}^{-1}\mathbf{s}\right) \times \left(\mathrm{H}^{-1}\mathbf{p}\right)$.

The deviation of these values from the ideal ones are used as evaluation metrics. Note that the ideal values for r_d, r_h, and r_v are 1, and the ideal value for the orthogonality is 90°, and $|r_d - 1|$, $|r_h - 1|$, $|r_v - 1|$, and $|\theta_o - 90°|$ are evaluated for the performance evaluation. When there are two or more planar targets in an input image, their average is computed as illustrated in Fig. 3.4.

3.3.2 Quantitative evaluation

The proposed method and two conventional methods [3, 4] are evaluated on four datasets.

- business card images (*Card*): 100 images
- document images including tables and/or figures (*Document*): 100 images

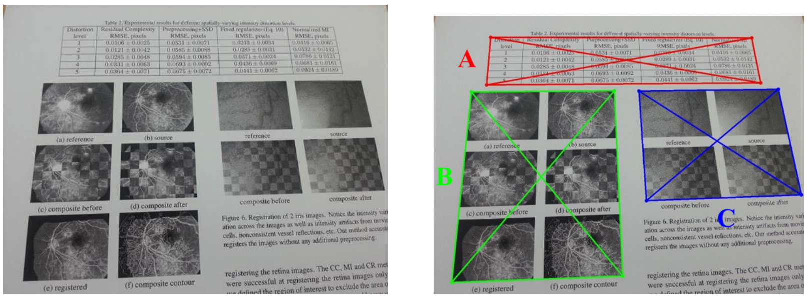


Table 2. Experimental results for different spatially-varying intensity distortion levels.

Distortion level	Residual Complexity RMSE, pixels	Preprocessing+SSD RMSE, pixels	Fixed regularizer (Eq. 10) RMSE, pixels	Normalized MI RMSE, pixels
1	0.0106 ± 0.0025	0.0531 ± 0.0071	0.0213 ± 0.0034	0.0416 ± 0.0065
2	0.0121 ± 0.0042	0.0585 ± 0.0088	0.0289 ± 0.0031	0.0532 ± 0.0142
3	0.0285 ± 0.0048	0.0594 ± 0.0085	0.0371 ± 0.0024	0.0786 ± 0.0121
4	0.0331 ± 0.0063	0.0693 ± 0.0092	0.0436 ± 0.0069	0.0681 ± 0.0161
5	0.0364 ± 0.0071	0.0675 ± 0.0072	0.0441 ± 0.0062	0.0924 ± 0.0189

(a) reference (b) source reference source

(c) composite before (d) composite after composite before composite after

Figure 6. Registration of 2 iris images. Notice the intensity var ation across the images as well as intensity artifacts from movi cells, nonconsistent vessel reflections, etc. Our method accurat registers the images without any additional preprocessing.

(e) registered (f) composite contour

registering the retina images. The CC, MI and CR met were successful at registering the retina images only we defined the region of interest to exclude the area o

(a) (b)

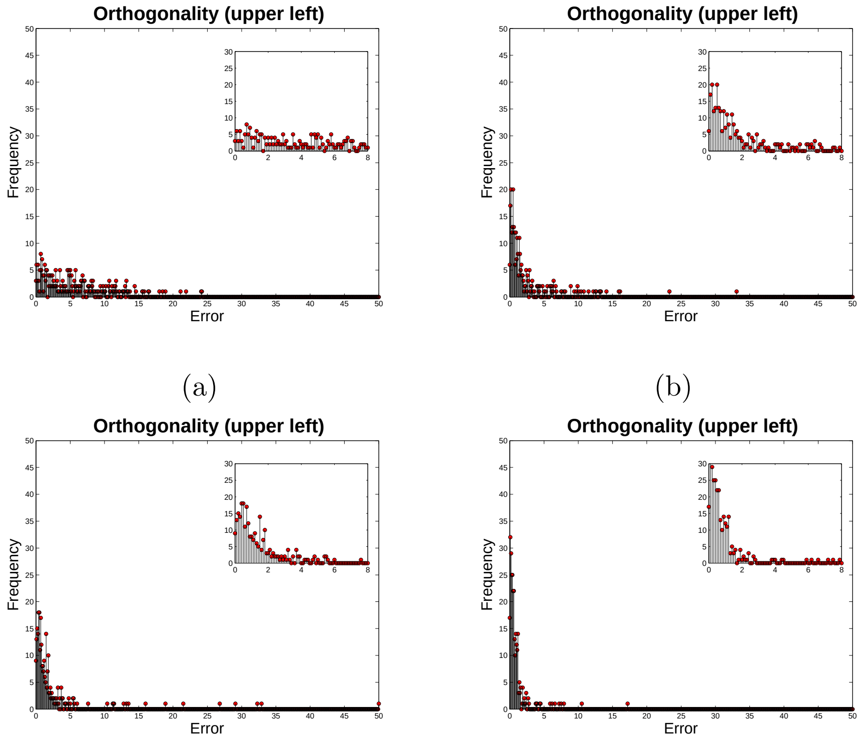
Figure 3.4: When there are two or more rectangles in an input image, the average of errors (i.e., A, B, and C rectangles) is computed.

- building facade images (*eTRIMs* [44]): 60 images
- signpost images (*Signpost*): 30 images

For the conventional methods [3,4], the source code or programs implemented by the authors are used. Since SfAR [3] requires the exact focal length, the focal length from the camera's EXIF data is used (for *Card* and *Document*) when the data is available. Otherwise, the automatic camera calibration method [4] is used for the estimation of focal length (for *eTRIMs* and *Signpost*). The other rectification methods discussed in Sec. 3.1.1 are not compared with the proposed method, because they are focused on finding individual cases (e.g., the presence of a single dominant rectangle).

Table 3.1: Performance comparison of the proposed method with upright adjustment [4] and SfAR [3]. The comparison is based on four values: mean and median as measurements of center, and standard deviation (SD) and interquartile range (IQR) as those of spread. The boldface represents the best result in each row.

Measure		Original	Upright [4]	SfAR [3]	Proposed
Orthogonality	Mean	6.0090	2.6052	2.4062	**0.9322**
	Median	4.6523	1.1117	0.9817	**0.5175**
	SD	4.8857	3.8343	5.4250	**1.6371**
	IQR	6.7114	2.2954	1.6388	**0.7575**
Diagonal ratio	Mean	0.0390	0.0180	0.0347	**0.0089**
	Median	0.0246	0.0065	0.0084	**0.0059**
	SD	0.0436	0.0346	0.1381	**0.0163**
	IQR	0.0441	0.0140	0.0126	**0.0115**
Vertical ratio	Mean	0.1062	0.0630	0.0902	**0.0156**
	Median	0.0742	0.0211	0.0207	**0.0088**
	SD	0.1028	0.1542	0.2928	**0.0301**
	IQR	0.1426	0.0509	0.0472	**0.0177**
Horizontal ratio	Mean	0.0979	0.0415	0.0552	**0.0117**
	Median	0.0729	0.0139	0.0148	**0.0048**
	SD	0.0884	0.0626	0.1809	**0.0291**
	IQR	0.1085	0.0413	0.0273	**0.0101**



(a)

(b)

(c)

(d)

Figure 3.5: Error histograms of orthogonality. (a) Original, (b) Upright [4], (c) SfAR [3], (d) The proposed method.

Evaluation results are summarized in Tab. 3.1, which shows that the proposed method yields the least errors. Since there are some failure cases, the standard deviation and interquartile range are also evaluated that can reflect the spread of values in a distribution. It can be seen that the proposed method shows the smallest values for all datasets. The error histogram for the orthogonality is illustrated in Fig. 3.5. As shown in the histogram, the proposed method shows sharp peak around zero, and the number of failure cases is smaller than others. The histograms of other measures also show similar patterns. In summary, the proposed method works robustly for the 2D rectification problem, because the line alignments are the salient features for 2D case and the proposed method exploits them.

3.3.3 Computation complexity

The proposed algorithm is implemented with C++, and the computation time is measured on a PC with AMD Phenom(tm) II X6 1055T Processor. The result is summarized in Table. 3.2, which shows that the proposed method is efficient. One might think this comparison is biased, because conventional methods were implemented with Matlab. However, since (a) LSD [42] takes about 90% of the computation time in the proposed method and (b) the LSD was also implemented with C-mex in the conventional methods [3, 4], it is believed that the proposed method is more efficient than conventional approaches.

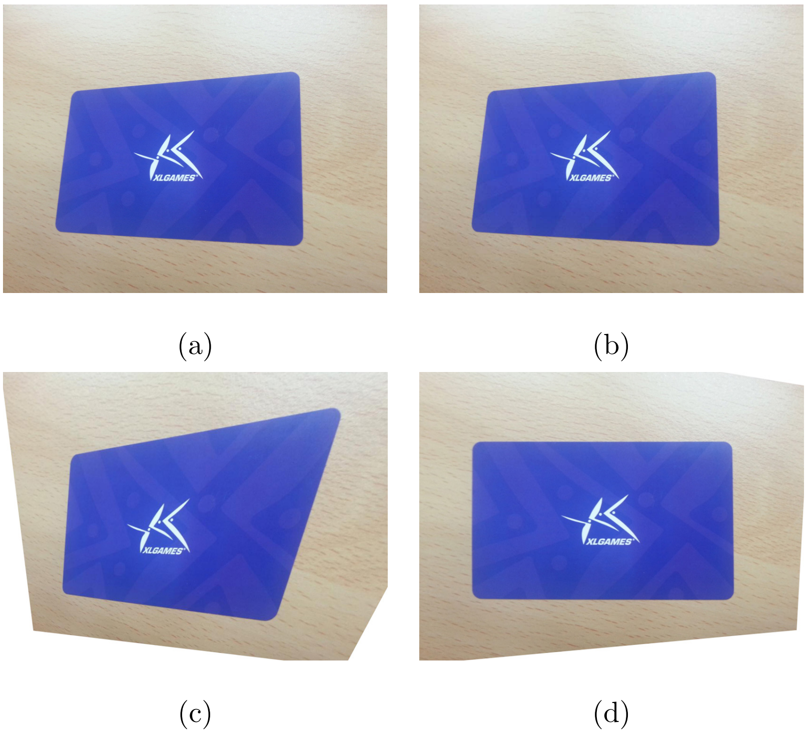
3.3.4 Qualitative comparisons and limitations

Figs. 3.6-3.9 and 3.10 show the rectification results of conventional methods [3, 4] and the proposed method for sample images of our dataset and the ones from [3, 4]. Because the proposed method depends on line segments, it sometimes fails when the

Table 3.2: The execution time (sec). In the case of [4], only a part of source code is available and its execution time is evaluated. The boldface shows the best result.

Method	*Card*	*Document*	*ETRIMs*	*Signpost*
Upright [4]	≥2.02	≥2.26	≥3.20	≥3.81
SfAR [3]	8.91	42.29	53.37	37.23
Ours (C++)	**0.13**	**0.17**	**0.18**	**0.21**

majority of the extracted line segments are not aligned with principal axes (see Fig. 3.11). Another limitation is that the proposed method may not perform the exact metric rectification when the focal length largely deviates from our assumption, i.e., (3.8). However, the exact focal length estimation is impossible from a planar rectangle target [5], and the proposed method yields reasonable results for a variety of inputs as shown in Figs. 3.6-3.9.



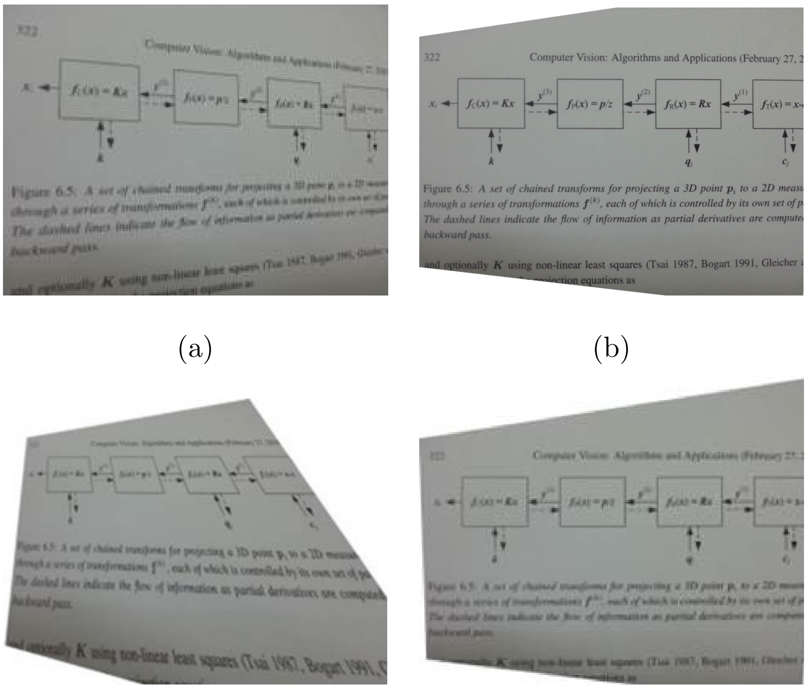
(a)

(b)

(c)

(d)

Figure 3.6: Sample results for *Card* dataset. (a) Input images, (b) Results of [4], (c) Result of [3], and (d) Results of the proposed method.



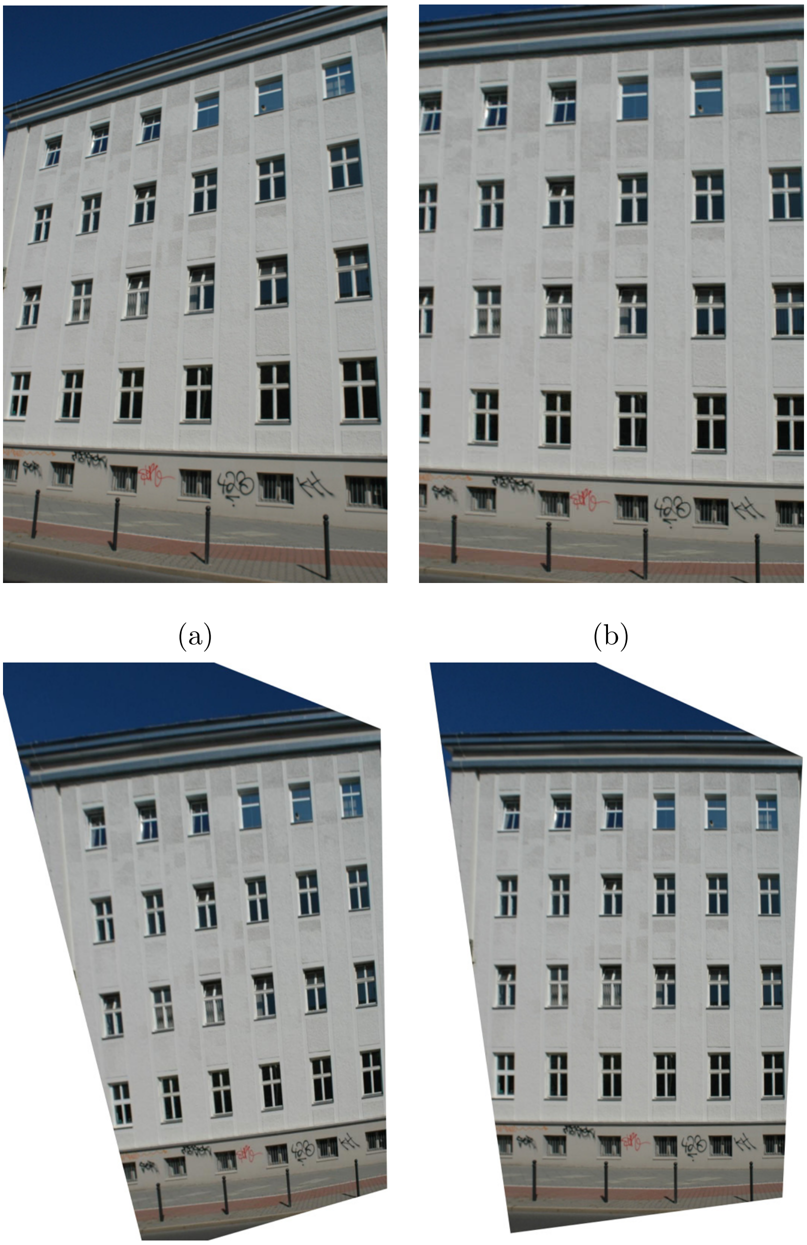
(a)

(b)

(c)

(d)

Figure 3.7: Sample results for *Document* dataset. (a) Input images, (b) Results of [4], (c) Result of [3], and (d) Results of the proposed method.



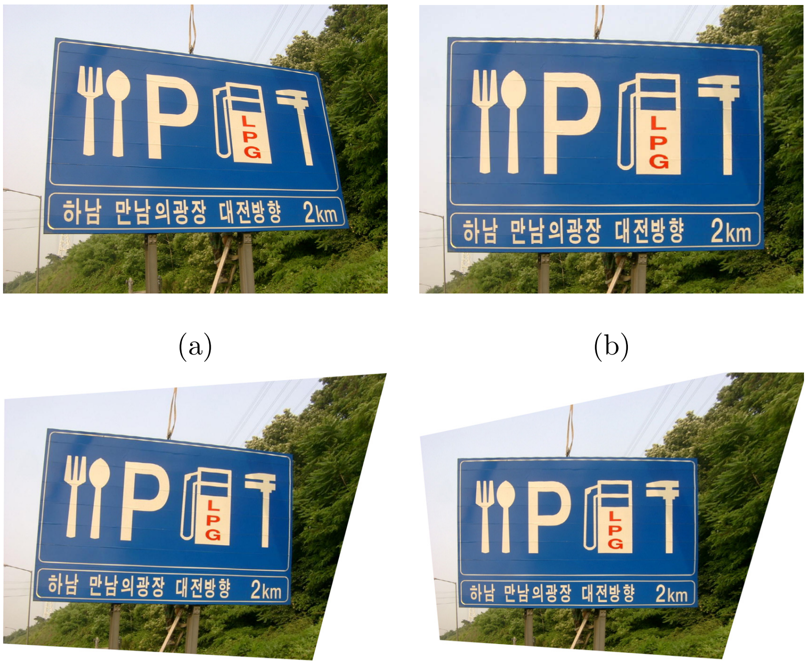
(a)

(b)

(c)

(d)

Figure 3.8: Sample results for *ETRIMs* dataset. (a) Input images, (b) Results of [4], (c) Result of [3], and (d) Results of the proposed method.



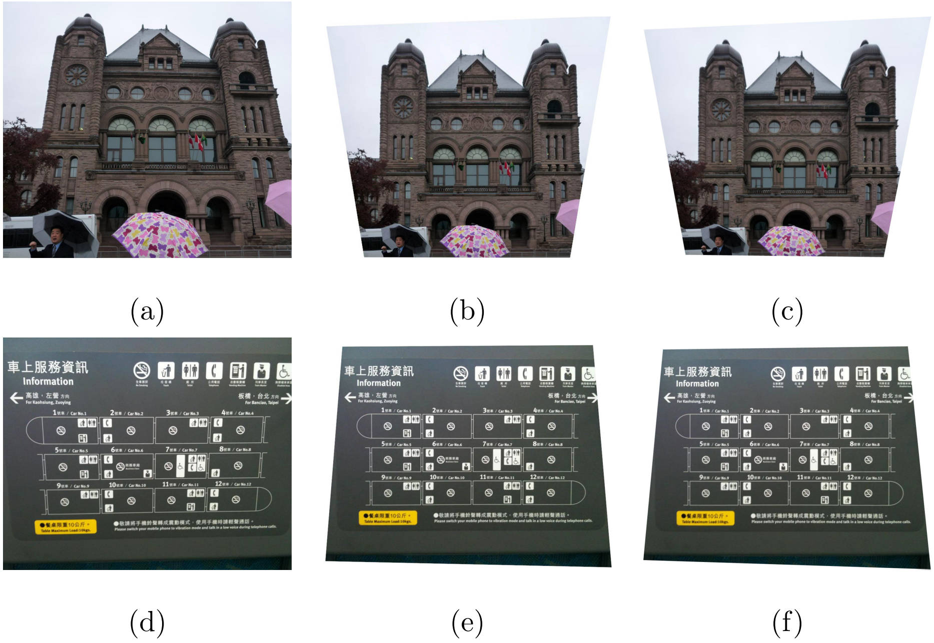
(a)

(b)

(c)

(d)

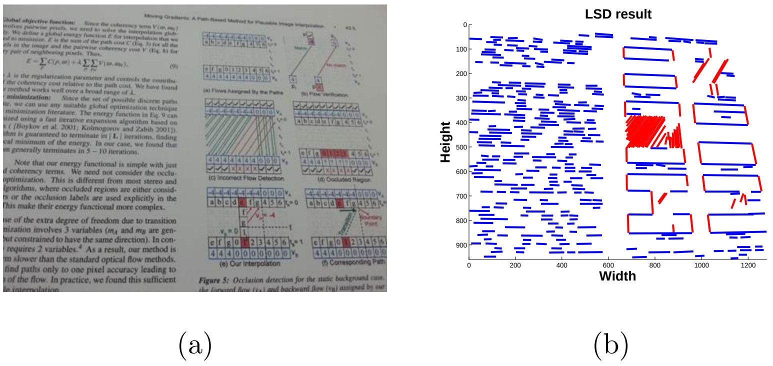
Figure 3.9: Sample results for *Signpost* dataset. (a) Input images, (b) Results of [4], (c) Result of [3], and (d) Results of the proposed method.



(a) (b) (c)

(d) (e) (f)

Figure 3.10: Examples of planar target in [4]. First column: input images. Second column: results of [4]. The last column: results of the proposed method.



(a) (b)

Figure 3.11: A failure example. (a) Input image, (b) LSD results.

3.4 Summary

In this section, an efficient and robust method for the rectification of planar targets based on line segments have been proposed. For the goal, the camera parameters were parameterized with four variables and a cost function defined on line segments was developed, which can be easily minimized via the Levenberg-Marquardt algorithm. According to the experiments, the proposed method showed better performance for the planar objects compared to the state-of-the-art self-calibration methods, as the proposed method was specialized for the 2D rectification. Hence, it is believed that the proposed method is a competent method for the rectification of name cards, signposts and documents where the rectification of 2D planar objects is the main goal.

Chapter 4

Application: Document capture system for mobile devices

4.1 Related works

Mobile devices have become a popular platform for document capture because they are portable, powerful and affordable. They have also embedded high-quality cameras and improved processing power, so that they are expected to play a critical role in many business applications such as document archival, ID scanning, check digitization. However, this is a challenging problem because document images captured by mobile devices often have perspective distortion, focus and motion blur, change of illumination, partial occlusions of the document pages, etc. The final goal of document capture is the OCR of text in the document and thus we need to alleviate these problems. For this, the most important step is to find the area of document in the image, i.e., the boundary of documents.

Hence, there have been several researches for finding the boundaries of documents on mobile devices. Rectangular document regions may be generally localized with the Hough transform [45, 46], but they require high computational cost and are susceptible to perspective distortion. To resolve this problem, the Hough transform applied to edge map after high frequency noise is filtered out [32] or the fast Hough transform [47] was used [48]. However, these methods need a user interaction for indicating a region of interest (ROI).

4.2 The proposed method

In this section, a cost function is derived for automatically finding the boundaries of documents in videos or actually from the live views of cameras before capturing a still scene. The goal of the proposed method is to extract promising candidates (i.e., line segments) for four sides of document and find the true boundary by evaluating likelihoods of possible configurations.

4.2.1 Notation

The frame sequence is numbered as $1, ..., t$, and the document region in frame t is defined Ω_t that is enclosed by a sequence of line segments Γ_t. The boundary component Γ_t is consisted of two horizontal lines (top $\mathbf{l}_T$ and bottom $\mathbf{l}_B$) and two vertical lines (left $\mathbf{l}_L$ and right $\mathbf{l}_R$) that are obtained by extracting line segments [42] and computing their orientations. Fig. 4.1 illustrates the notation used in this section.

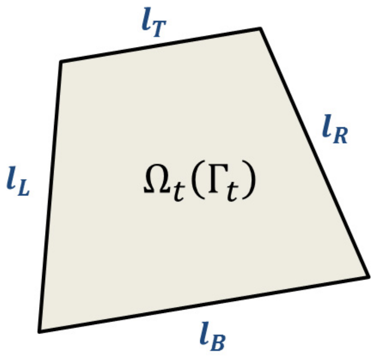


Figure 4.1: Notation used in this section. Γ_t represents four bounding line segments: $\mathbf{l}_T$, $\mathbf{l}_B$, $\mathbf{l}_L$ and $\mathbf{l}_R$. And $\Omega_t(\Gamma_t)$ is the document region enclosed by Γ_t.

4.2.2 Optimization-based framework

For the boundary detection of documents, candidates (line segments) for boundaries are first extracted. For the effective extraction, noise and textures are filtered out by applying morphological operations and the line segment detector in [42] is applied to a channel that shows the highest contrast among three (red, green, and blue) color channels. At any time t, the detected line segments is denoted as

$$\mathcal{L}_t = \{(\mathbf{u}_t, \mathbf{v}_t)\}, \tag{4.1}$$

where $\mathbf{u}_t$ and $\mathbf{v}_t$ are the homogeneous representation of two end-points of a line segment. Then, the document region detection is formulated as an optimization problem defined on $\mathcal{L}_t$:

$$\hat{\Omega}_t \ = \arg\min_{\Omega_t} E_c(\Omega_t; \mathcal{L}_t, \mathcal{C}_t) + E_b(\Omega_t; \mathcal{L}_t, \mathrm{H}_\mathrm{t}^{-1}) + \mathrm{E}_\mathrm{t}(\Omega_\mathrm{t}; \Omega_\mathrm{t-1}), \tag{4.2}$$

where the data term $E_c(\Omega_t; \mathcal{L}_t, \mathcal{C}_t)$ measures the closeness of document regions given their color distributions $\mathcal{C}_t$ and the boundary term $E_b(\Omega_t; \mathcal{L}_t, \mathrm{H}_\mathrm{t}^{-1})$ prefers boundaries

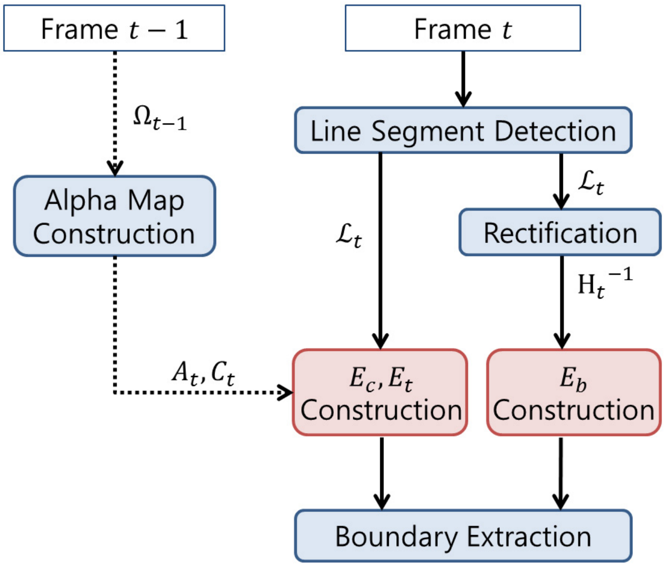
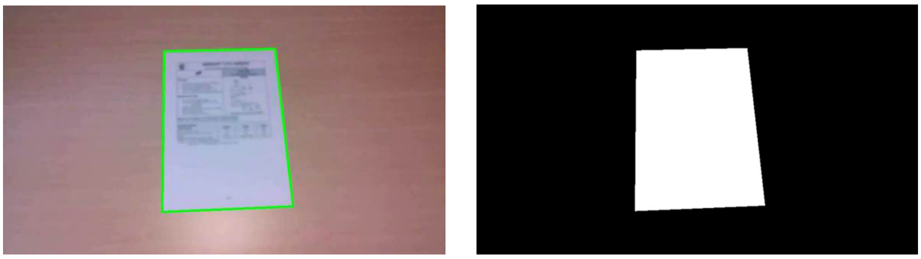


Figure 4.2: Block diagram of the proposed method.

that align to the principal axes and show high contrast. In boundary term, the transform $\mathrm{H}_\mathrm{t}^{-1}$ is exploited as illustrated Sec. 3. And the third term $E_t(\Omega_t; \Omega_{t-1})$ implies that the detected boundary component must be temporally similar. Fig. 4.2 shows a block diagram of the proposed method.



(a) (b)

Figure 4.3: Alpha map construction in frame t. (a) The green solid contour is Γ_{t-1} in frame $t-1$, (b) The alpha map $\mathcal{A}_t$ in frame t is constructed.

Cost function based on color distributions

A binary map is constructed for each frame, which indicates whether a pixel belongs to foreground ($\mathcal{F}$) or background ($\mathcal{B}$). For this, an alpha map $\mathcal{A}_t$ in the frame t is defined, which is a binary matrix with the size of input image. Note that this matrix can be constructed when the document region Ω_{t-1} at the previous frame $t-1$ is detected. The document is located locally between the consecutive frames, so at the current frame, the region $\Omega_{t-1}(\Gamma_{t-1})$ can be considered foreground (document). Precisely, the element of $\mathcal{A}_t$ at the pixel position $\mathbf{p}$, denoted as $\mathcal{A}_t(\mathbf{p})$, is given 1 when $\mathbf{p}$ belongs to foreground and 0 when it belongs to background. Fig. 4.3 illustrates the alpha map in frame t given $\Omega_{t-1}(\Gamma_{t-1})$.

From the alpha map, color histograms are constructed and let us denote them as

$$\mathcal{C}_t = \{h^R_{\mathcal{F}}, h^G_{\mathcal{F}}, h^B_{\mathcal{F}}, h^R_{\mathcal{B}}, h^G_{\mathcal{B}}, h^B_{\mathcal{B}}\}, \tag{4.3}$$

where $h^i_{\mathcal{F}}$ and $h^i_{\mathcal{B}}$ are color histograms corresponding to foreground and background ($\mathcal{B}$) regions for each color channel ($i = \{R, G, B\}$), respectively.

The first term in (4.2) is defined as

$$E_c(\Omega_t;\mathcal{L}_t,\mathcal{C}_t) = \sum_{\mathbf{p}\in\mathcal{F}} E_{color}(\text{“foreground”}) + \sum_{\mathbf{p}\in\mathcal{B}} E_{color}(\text{“background”}). \tag{4.4}$$

Specifically, the data cost is constructed as a negative of log-likelihoods of color histograms as

$$E_{color}(\text{“foreground”}) = -\ln Pr(\mathbf{p}|\mathcal{F}), \tag{4.5a}$$

$$E_{color}(\text{“background”}) = -\ln Pr(\mathbf{p}|\mathcal{B}), \tag{4.5b}$$

where

$$Pr(\mathbf{p}|\mathcal{F}) = h^R_{\mathcal{F}}(\mathbf{p}) \times h^G_{\mathcal{F}}(\mathbf{p}) \times h^B_{\mathcal{F}}(\mathbf{p}), \tag{4.6a}$$

$$Pr(\mathbf{p}|\mathcal{B}) = h^R_{\mathcal{B}}(\mathbf{p}) \times h^G_{\mathcal{B}}(\mathbf{p}) \times h^B_{\mathcal{B}}(\mathbf{p}). \tag{4.6b}$$

As illustrated in Fig. 4.4, if a pixel belongs to document regions, the corresponding energy value $E_{color}(\text{“foreground”})$ is small.

Cost function based on boundaries configurations

The boundary term in (4.2) is defined as

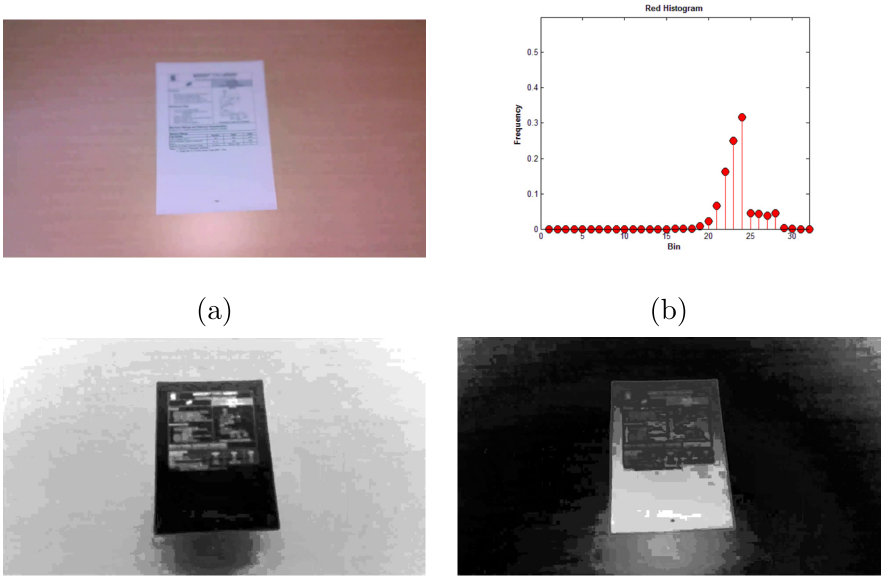
$$E_b(\Omega_t;\mathcal{L}_t,\mathrm{H}_\mathrm{t}^{-1}) = \lambda_1 \sum_{\mathbf{l}\in\Gamma_\mathrm{t}} \mathrm{E}_{\mathrm{alignment}}(\mathbf{l},\mathrm{H}_\mathrm{t}^{-1}) + \lambda_2 \sum_{\mathbf{l}\in\Gamma_\mathrm{t}} \mathrm{E}_{\mathrm{contrast}}(\mathbf{l},\mathrm{H}_\mathrm{t}^{-1}), \tag{4.7}$$

where $\mathrm{H}_\mathrm{t}^{-1}$ is the rectification transform. If the boundary component is right, the transformed boundary component under $\mathrm{H}_\mathrm{t}^{-1}$ has a small alignment error. In this regard, the measurement based on line segment alignment is defined as

$$E_{alignment}(\mathbf{l},\mathrm{H}_\mathrm{t}^{-1}) = \mathrm{d}_\mu^2(\mathrm{H}_\mathrm{t}^{-1}\mathbf{u},\mathrm{H}_\mathrm{t}^{-1}\mathbf{v}), \tag{4.8}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the homogeneous representation of two end-points of a line segment $\mathbf{l}$. As illustrated in Sec. 3, the alignment error is defined as

$$d_\mu(\mathbf{p},\mathbf{q}) = \min(|p_1/p_3 - q_1/q_3|, |p_2/p_3 - q_2/q_3|), \tag{4.9}$$



(a)

(b)

(c)

(d)

Figure 4.4: (a) Frame t, (b) Color histograms corresponding to foreground ($i = R$), (c) Energy map E_{color}("foreground"), (d) Energy map E_{color}("background").

where $\mathbf{p}$ and $\mathbf{q}$ are transformed points ($\mathbf{p} = [p_1, p_2, p_3]^\top$ and $\mathbf{q} = [q_1, q_2, q_3]^\top$). In addition to the alignment error, color contrast between the regions separated by line segment $\mathbf{l}$ is also measured. For this, color histograms in RGB space corresponding to regions are constructed and the histogram matching is used for measurement of similarity between regions. Fig. 4.5-(c) and (d) show two regions separated by the top horizontal line segment $\mathbf{l}_T$ and color histogram corresponding to the regions, respectively. Histogram intersection is normally used for evaluating how similar two histograms, so the cost function based on color contrast is defined as

$$E_{contrast}(\mathbf{l}, \mathrm{H}_\mathrm{t}^{-1}) = \frac{\sum_\mathrm{k} \min\left(\mathrm{h}_1\left(\mathrm{k}\right), \mathrm{h}_2\left(\mathrm{k}\right)\right)}{\min\left(\left|\mathrm{h}_1\right|, \left|\mathrm{h}_2\right|\right)}, \tag{4.10}$$

where h_1 and h_2 are two color histograms with k bins.

Cost function based on temporal coherence

The temporal term in (4.2) is defined as

$$E_t(\Omega_t; \Omega_{t-1}) = \lambda_3 \sum_{\mathbf{l} \in \Gamma_t, \mathbf{l\prime} \in \Gamma_{t-1}} E_{length}(\mathbf{l}, \mathbf{l\prime}) + \lambda_4 \sum_{\mathbf{p} \in \mathcal{P}_t, \mathbf{p\prime} \in \mathcal{P}_{t-1}} E_{location}(\mathbf{p}, \mathbf{p\prime}), \tag{4.11}$$

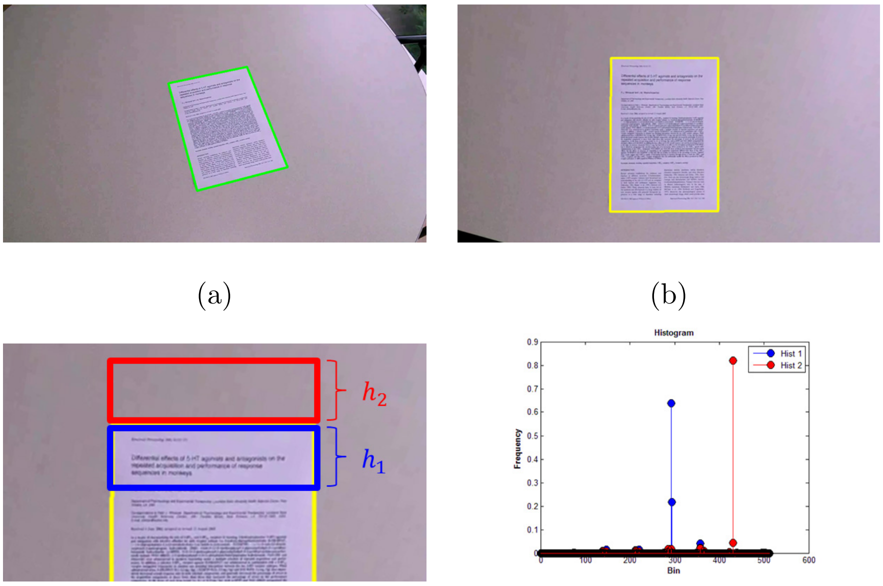
where

$$E_{length}(\mathbf{l}, \mathbf{l\prime}) = \left| \frac{d(\mathbf{u}, \mathbf{v})}{d(\mathbf{u\prime}, \mathbf{v\prime})} - 1 \right|, \tag{4.12a}$$

$$E_{location}(\mathbf{p}, \mathbf{p\prime}) = d(\mathbf{p}, \mathbf{p\prime}). \tag{4.12b}$$

In the above equations, $\mathcal{P}_t$ is a set of corners whose components are give by $\mathbf{l}_i \times \mathbf{l}_j, (\mathbf{l}_i, \mathbf{l}_j) \in \Gamma_t$. And $d(\cdot, \cdot)$ is a geometric distance between two points ($\mathbf{p} = [p_1, p_2, p_3]^\top$ and $\mathbf{q} = [q_1, q_2, q_3]^\top$):

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1/p_3 - q_1/q_3)^2 + (p_2/p_3 - q_2/q_3)^2}. \tag{4.13}$$



(a)

(b)

(c)

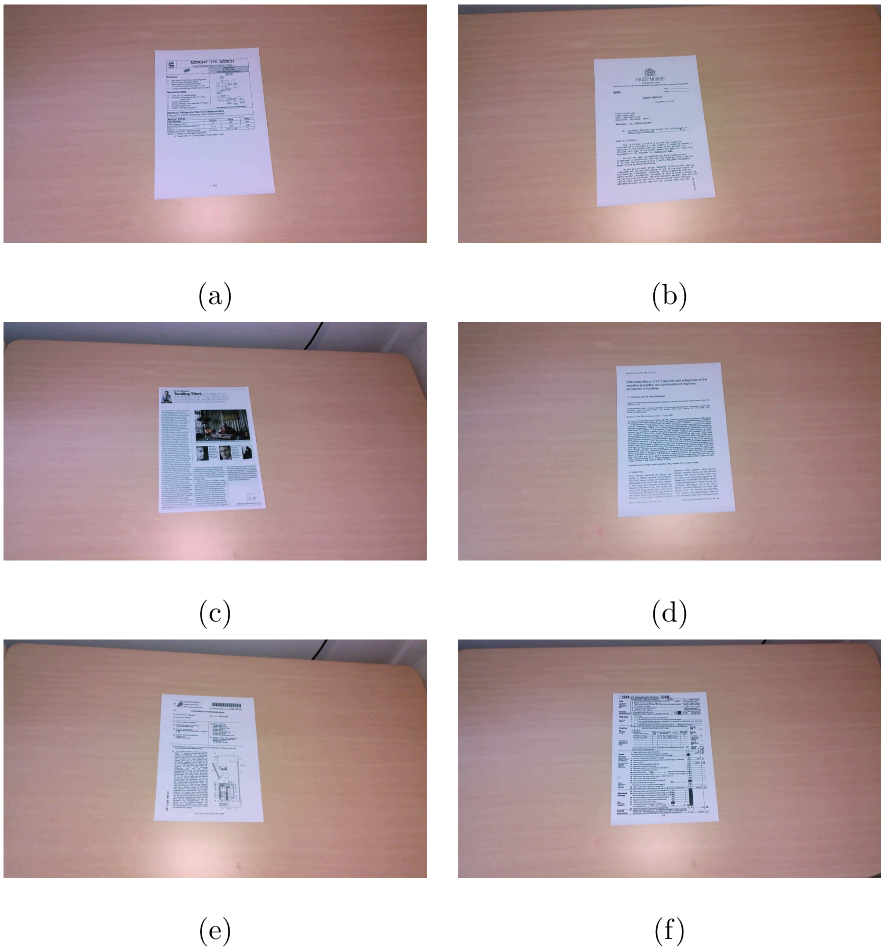
(d)

Figure 4.5: (a) Frame t (the green solid contour is the boundary component Γ), (b) Rectified frame (the yellow solid contour is the rectified boundary component), (c) Two regions separated by one horizontal line segment $\mathbf{l}_T$, (c) Two color histograms corresponding to upper and lower regions as shown in (c).

The defined temporal term implies that the length and location of boundary components should be similar between consecutive frames.

4.3 Experimental results

The proposed method is evaluated on a dataset from *ICDAR'2015 Competition* [49] which consists of six different document types and five document images are chosen per class. They provided small video clips of around 10 seconds for each of the 30 documents in five different background scenarios (i.e., the database consists of 150 video clips comprising around 25,000 frames). The videos were recorded using Full HD 1920×1080 resolution at variable frame-rate, but in our experiments, images are resized a half (960×540) for computation complexity. Fig. 4.6 and Fig. 4.7 show the examples of six different document types and five different background scenarios, respectively. As can be seen in Fig. 4.7, it is somewhat easy to distinguish between document and background in *Background 1* and *3*, but *Background 2*, *4* and *5* don't. Especially, images in *Background 5* suffer from partial occlusions of the document pages.



(a)

(b)

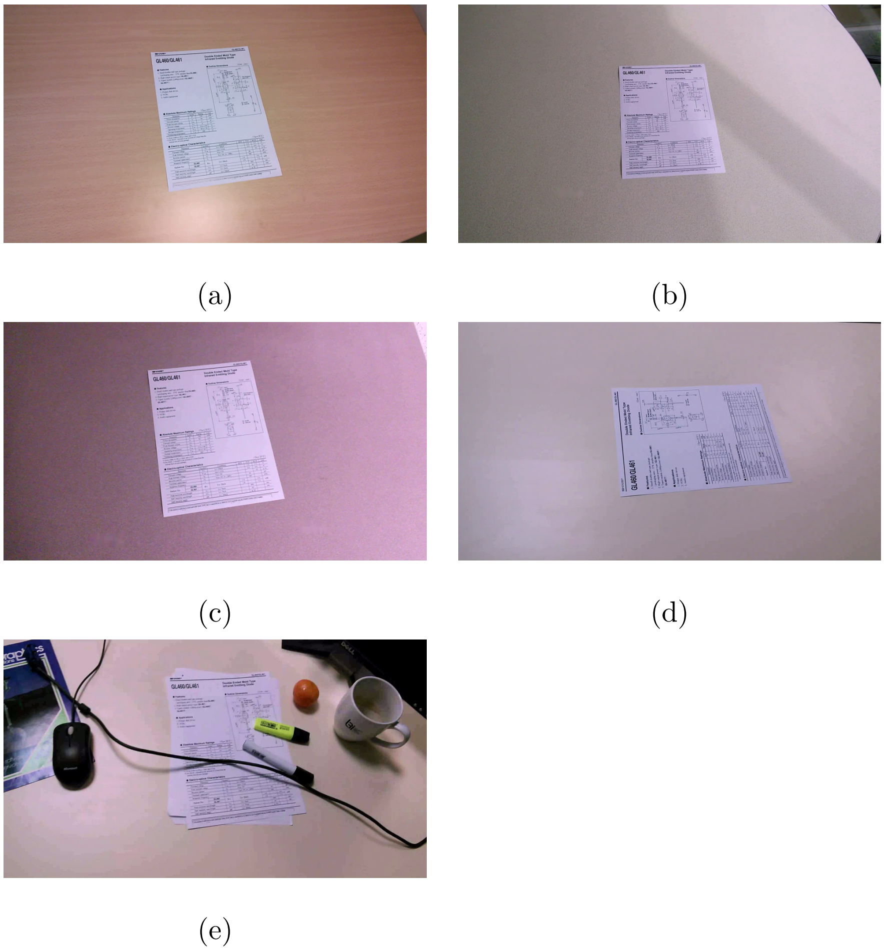
(c)

(d)

(e)

(f)

Figure 4.6: Six different document types. (a) Datasheet, (b) Letter, (c) Magazine, (d) Paper, (e) Patent, (f) Tax.



(a) (b) (c) (d) (e)

Figure 4.7: Five different background scenarios. (a) *Background 1*, (b) *Background 2*, (c) *Background 3*, (d) *Background 4*, (e) *Background 5*.

4.3.1 Initialization

At the first frame ($t = 1$), the alpha map $\mathcal{A}_1$ is constructed from a trained data:

$$\mathcal{A}_1(\mathbf{p}) = \begin{cases} 1 & \text{if } h^R(\mathbf{p}) + h^G(\mathbf{p}) + h^B(\mathbf{p}) > \tau_\alpha \\ 0 & \text{otherwise ,} \end{cases} \tag{4.14}$$

where h^i is color histogram for the document region of the trained data ($i = \{R, G, B\}$). The trained data is consisted of 3 video clips comprising around 600 frames. In all the experiments, the number of histogram bins is set to 32 for h^i, $h^i_{\mathcal{F}}$, $h^i_{\mathcal{B}}$ in (4.3) and (4.14) and 512($8 \times 8 \times 8$) for h_1, h_2 in (4.10). The parameters are set as $\lambda_1 = 0.05$, $\lambda_2 = 10$ for the boundary term and $\lambda_3 = 20$, $\lambda_4 = 0.05$ for the temporal term and $\tau_\alpha = 0.25$ for the initial alpha map.

4.3.2 Quantitative evaluation

For each frame $\mathbf{f}$, the four corner points of the document object are manually annotated for creating a ground-truth dataset. Let us denote the ground-truth and the result of proposed method as $\mathbf{G}$ and $\mathbf{S}$. Using the document size and its coordinate, the corrected quadrangles $\mathbf{G}\prime$ and $\mathbf{S}\prime$ are obtained by transforming $\mathbf{G}$ and $\mathbf{S}$ under the perspective transform. Then, the evaluation measure at frame $\mathbf{f}$ is defined as the Jaccard index (JI)

$$\text{JI}(\mathbf{f}) = \frac{\text{Area}(\mathbf{G}\prime \bigcap \mathbf{S}\prime)}{\text{Area}(\mathbf{G}\prime \bigcup \mathbf{S}\prime)}, \tag{4.15}$$

which measures the similarity between the set $\mathbf{G}\prime$ and $\mathbf{S}\prime$. If the result by using the proposed method coincides with the ground-truth, the Jaccard index is 1. Evaluation results are summarized in Tab. 4.1, which shows that the proposed method performs document detection and segmentation robustly and accurately in *Background 1, 2,*

Table 4.1: The mean of the Jaccard index (mJI) for six document types and five background scenarios.

Background (total number of frames)		*1* (6180)	*2* (6011)	*3* (5952)	*4* (4169)	*5* (2577)
Category	Datasheet	1.0000	0.9979	0.9998	0.9974	0.8200
	Letter	1.0000	0.9956	0.9932	0.9903	0.4967
	Magazine	0.9958	0.9999	0.9956	0.9681	0.5064
	Paper	1.0000	0.9979	0.9980	0.9974	0.5385
	Patent	0.9984	0.9735	0.9999	0.9895	0.6452
	Tax	0.9995	0.9582	0.9999	0.9791	0.6139
Average		**0.9989**	**0.9872**	**0.9977**	**0.9870**	**0.6035**

3, 4. However, the proposed method fails if the Manhattan assumption is broken like in *Background 5*.

4.3.3 Qualitative evaluation and limitations

Fig. 4.8-4.12 show the detection and segmentation results of the proposed method for videos in six different type and five background scenarios, respectively. Because the proposed method depends on line segments, it fails when the line segments are not extracted and/or the majority of the extracted line segments are not aligned with principal axes. Fig. 4.13 shows the example of failure cases for line detection in *Background 5*. The *Background 5* is a considered unsuitable scenario for document capture system because the major goal of the document capture is to replace personal scanner.

4.4 Summary

In this section, a document capture algorithm in video for mobile devices has been proposed. For finding the boundaries of documents, a cost function is developed which is composed of three terms: data term, boundary term, and temporal term. The data term exploits the color distributions of documents and backgrounds, and the boundary term prefers boundaries aligning the principal axes and showing high contrast. The temporal term is designed to consider the temporal coherence in consecutive frames. According to the experiments, the proposed method yields reasonable results in video for smartphone document capture system.

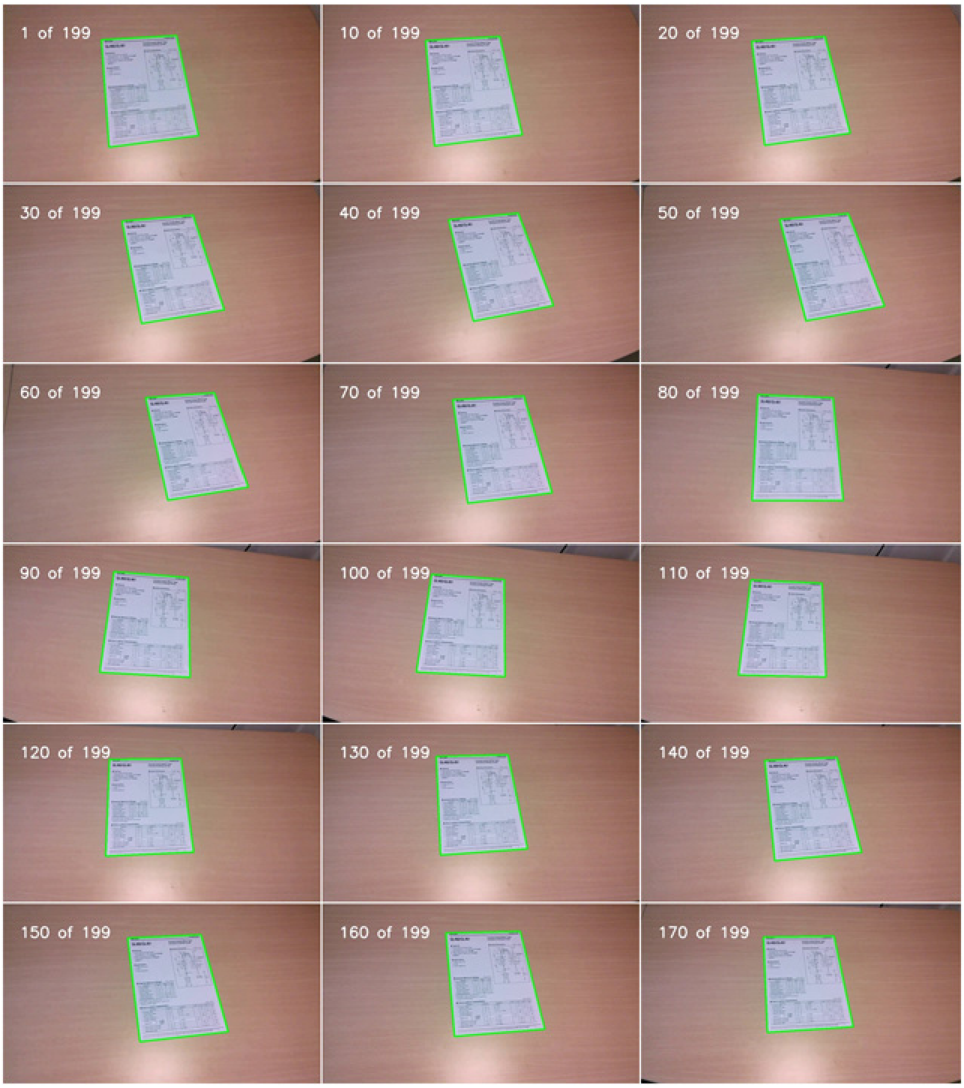


Figure 4.8: Results of the proposed method for a datasheet video in *Background 1*. The green solid contour is the detected boundaries and frame number is indicated at top-left side (The format is 't of total frame number').

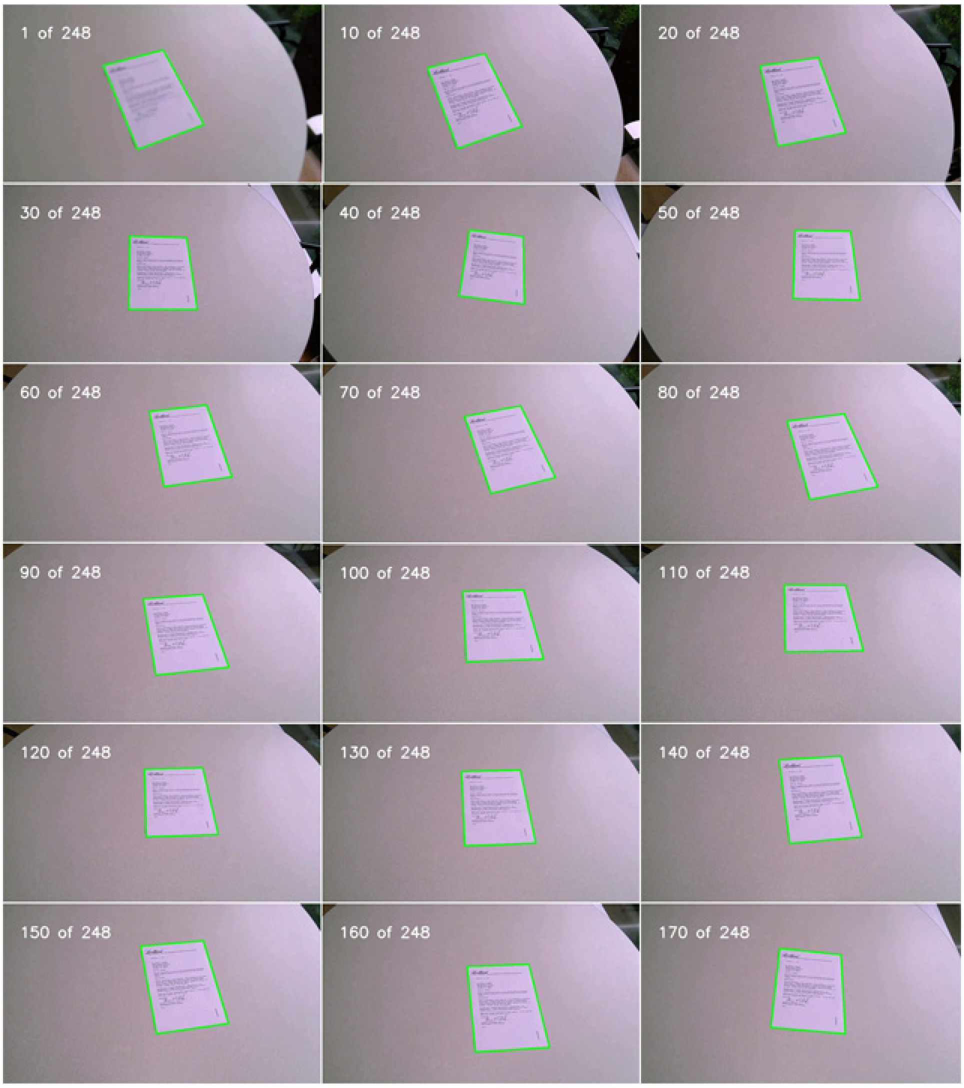


Figure 4.9: Results of the proposed method for a letter video in *Background 2*. The green solid contour is the detected boundaries and frame number is indicated at top-left side (The format is '*t* of total frame number').

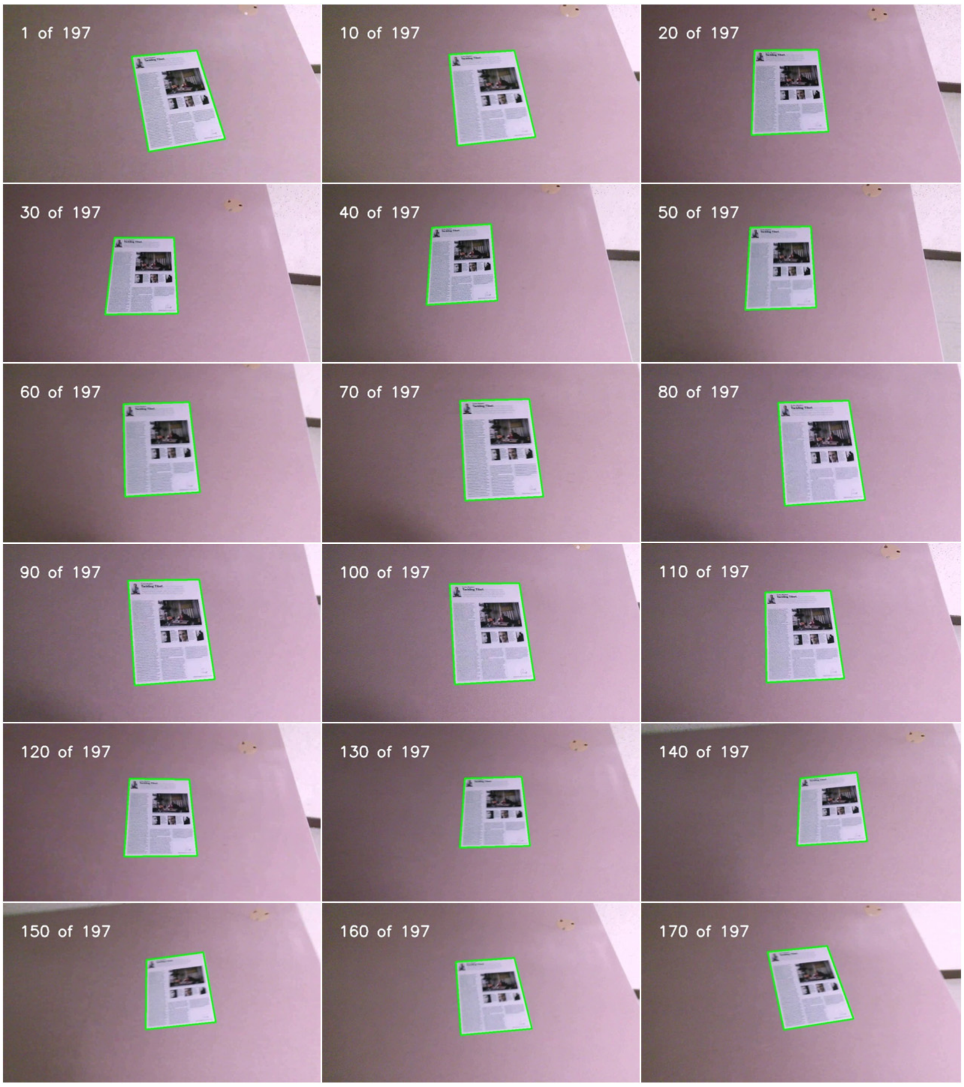


Figure 4.10: Results of the proposed method for a magazine video in *Background 3*. The green solid contour is the detected boundaries and frame number is indicated at top-left side (The format is 't of total frame number').

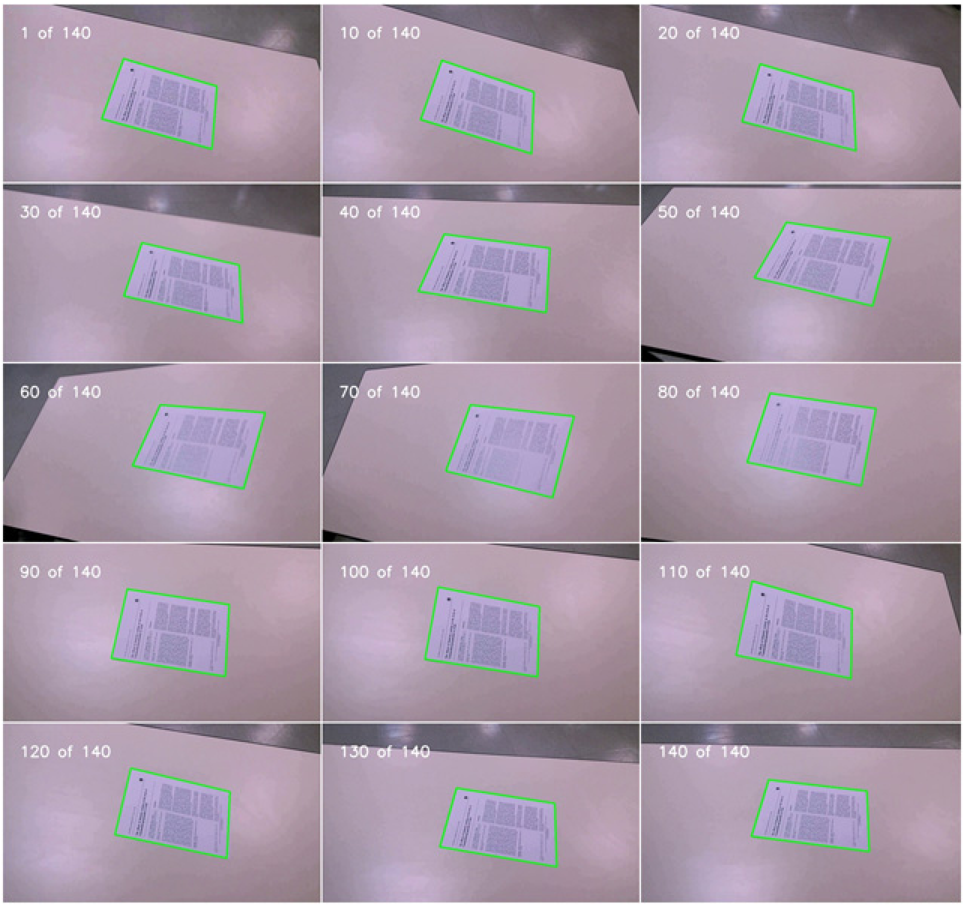


Figure 4.11: Results of the proposed method for a paper video in *Background 4*. The green solid contour is the detected boundaries and frame number is indicated at top-left side (The format is '*t* of total frame number').

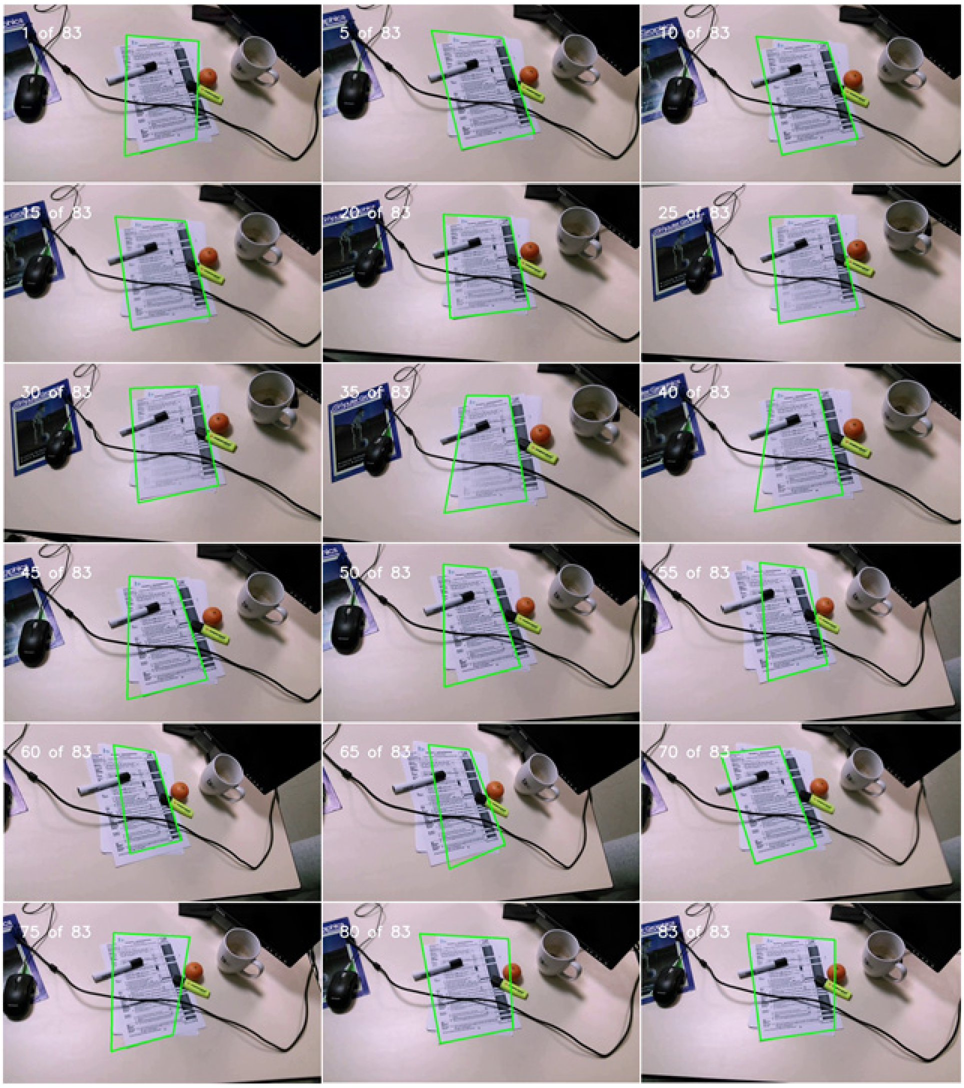


Figure 4.12: Results of the proposed method for a tax video *Background 5.* The green solid contour is the detected boundaries and frame number is indicated at top-left side (The format is 't of total frame number').

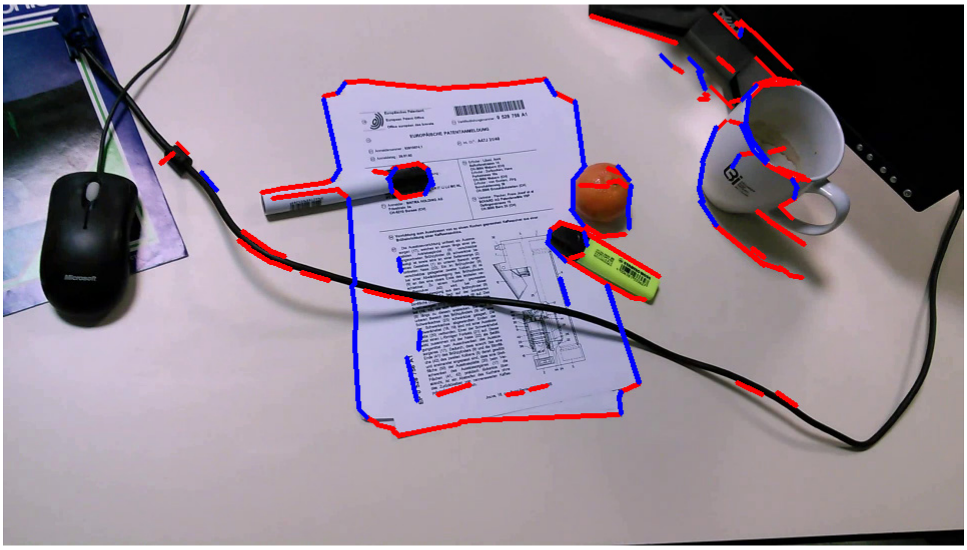


Figure 4.13: A failure example in *Background 5*.

Chapter 5

Conclusions and future works

In this dissertation, a unified framework for the image stitching and rectification is proposed, based on the camera pose estimation. For this purpose, appropriate optimization problems for the given problems have been constructed whose solutions give camera pose parameters. Unlike the conventional methods that have been developed assuming either of the conditions, i.e., (i) fixed-optical-center case or (ii) planar-target case, this dissertation noted that both conditions are actually very similar ones, and presented a unified framework that can handle both cases simultaneously has been proposed. Specifically, for given multiple images, the camera poses relative to a reference plane have been computed by minimizing the proposed cost function which is based on the registration errors on the reference plane. After estimating the camera poses, the amount of camera motion compared with the scale of scene has been computed. Then, if this relative motion is large, the composition of the image on the reference plane has been performed with the rectification of each image. Otherwise, (i.e., when the camera motion is small) since the metric rectification is impossible, visually pleasing results have been composed by selecting a viewpoint. Experimental results on synthetic and real images show that the proposed method successfully performs stitching and metric rectification. Also, this dissertation presents some applications with the proposed method, such as document image stitching and text-based augmented reality.

In the case of rectification problem, a simple, efficient, and robust approach to rectifying an image of planar targets has been proposed. This method is based on a basic assumption on single image calibration that the majority of line segments are aligned with the principal directions. Therefore, the proposed cost function has been formulated as an optimal homography estimation problem that makes the line segments horizontally or vertically straight. To be precise, the pose of the camera has been parameterized as four variables and the cost function which is defined on the reference plane has been developed. Unlike the conventional methods, the proposed method need not estimate vanishing points/lines and also need not the segmentation of planar object. Therefore, the proposed method has low computational time but successfully performs metric rectification than previous methods.

Finally, an additional example of applications is presented, which is a document capture algorithm for mobile devices. For this application, a new cost function has been developed by using color distribution and rectification model. To be precise, the data term exploits the color histogram of documents and backgrounds, and the boundary term reflects the alignment errors and the contrast of rectified boundaries. For the robustness of the whole system, the temporal term is designed to consider temporal coherence in consecutive frames. The experiments for various database videos shows that the proposed method successfully finds the boundary of documents.

Bibliography

[1] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.

[2] [Online]. Available: http://www.autostitch.net/

[3] A. Zaheer, M. Rashid, and S. Khan, "Shape from angle regularity," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, Oct. 2012, pp. 1–14.

[4] H. Lee, E. Shechtman, J. Wang, and S. Lee, "Automatic upright adjustment of photographs with robust camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 833–844, May 2014.

[5] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[6] R. Szeliski, "Image alignment and stitching: A tutorial," MSR-TR-2004-92, Microsoft Research, 2004, Tech. Rep., 2005.

[7] J. Liang, D. DeMenthon, and D. S. Doermann, "Camera-based document image mosaicing," in *Proceedings of International Conference on Pattern Recognition*, Aug. 2006, pp. 476–479.

[8] J. Hannuksela, P. Sangi, J. Heikkila, X. Liu, and D. Doermann, “Document image mosaicing with mobile phones,” in *Proceedgins of International Conference on Image Analysis and Processing*, Sep. 2007, pp. 575–582.

[9] P. Clark and M. Mirmehdi, “Estimating the orientation and recovery of text planes in a single image,” in *Proceedings of the 12th British Machine Vision Conference*, Sep. 2001, pp. 421–430.

[10] M. Pilu, “Extraction of illusory linear clues in perspectively skewed documents,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Dec. 2001, pp. 363–368.

[11] W. Lee, Y. Pack, and V. Lepetit, “Video-based *In Situ* tagging on mobile phones,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 10, pp. 1487–1496, Oct. 2011.

[12] C. Xu, B. Kuipers, and A. Murarka, “3d pose estimation for planes,” in *Proc. International Conference on Computer Vision Workshops (ICCV Workshops)*, Oct. 2009, pp. 673–680.

[13] H. I. Koo, B. S. Kim, and N. I. Cho, “A new method to find an optimal warping function in image stitching,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2009, pp. 1289–1292.

[14] D. Wagner, A. Mulloni, T. Langlotz, and D. Schmalstieg, “Real-time panoramic mapping and tracking on mobile phones,” in *Proceedings of the 2010 IEEE Virtual Reality Conference*, Mar. 2010, pp. 211–218.

[15] S. Lovegrove and A. J. Davison, “Real-time spherical mosaicing using whole image alignment,” in *Proceedings of the 11th European Conference on Computer Vision: Part III*, Sep. 2010, pp. 73–86.

[16] A. Mills and G. Dudek, “Image stitching with dynamic elements,” *Image Vis. Comput.*, vol. 27, no. 10, pp. 1593 – 1602, Sep. 2009.

[17] R. Garg and S. M. Seitz, “Dynamic mosaics,” in *3DIMPVT 2012*, Oct. 2012, pp. 65–72.

[18] L. Zeng, S. Zhang, J. Zhang, and Y. Zhang, “Dynamic image mosaic via sift and dynamic programming,” *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1271 – 1282, Jul. 2014.

[19] E. Zagrouba, W. Barhoumi, and S. Amri, “An efficient image-mosaicing method based on multifeature matching,” *Mach. Vis. Appl.*, vol. 20, no. 3, pp. 139–162, Feb. 2009.

[20] D. Lowe, “Distinctive image features from scale-incariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[21] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[22] J. Mor, “The levenberg-marquardt algorithm: Implementation and theory,” in *Numerical Analysis*, ser. Lecture Notes in Mathematics, G. Watson, Ed. Springer Berlin Heidelberg, 1978, pp. 105–116.

[23] M. Brown and D. Lowe, "Recognising panoramas," in *Proceedings of International Conference on Computer Vision*, Oct. 2003, pp. 1218–1225.

[24] A. Ruiz, P. E. Lopez-de Teruel, and L. Fernandez-Maimo, "Practical planar metric rectification," in *Proceedings of British Machine Vision Conference*, Sep. 2006, pp. 60.1–60.10.

[25] P. J. Burt, Edward, and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[26] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, Apr. 1991.

[27] J. M. Buenaposada and L. Baumela, "Real-time tracking and estimation of plane pose," in *Proceedings of International Conference on Pattern Recognition*, Aug. 2002, pp. 697–700.

[28] D. Cobzas, M. Jagersand, and P. Sturm, "3d ssd tracking with estimated 3d planes," *Image and Vision Computing*, vol. 27, no. 1-2, pp. 69–79, Jan. 2009.

[29] D. Liebowitz and A. Zisserman, "Metric rectification for perspective images of planes," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998. Proceedings*, Jun. 1998, pp. 482–488.

[30] G. Hua, Z. Liu, Z. Zhang, and Y. Wu, "Automatic business card scanning with a camera," in *IEEE International Conference on Image Processing*, Oct. 2006, pp. 373–376.

[31] ——, "Iterative local-global energy minimization for automatic extraction of objects of interest," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1701–1706, Oct. 2006.

[32] A. Hartl and G. Reitmayr, "Rectangular target extraction for mobile augmented reality applications," in *International Conference on Pattern Recognition*, Nov. 2012, pp. 81–84.

[33] J. Pritts, O. Chum, and J. Matas, "Detection, rectification and segmentation of coplanar repeated patterns," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 2973–2980.

[34] E. R. Corral-Soto and J. H. Elder, "Automatic single-view calibration and rectification from parallel planar curves," in *European Conference on Computer Vision*, Sep. 2014, pp. 813–827.

[35] Z. Zhang, Y. Matsushita, and Y. Ma, "Camera calibration with lens distortion from low-rank textures," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 2321–2328.

[36] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky, "Geometric image parsing in man-made environments," *Int. J. Comput. Vision*, vol. 97, no. 3, pp. 305–321, May 2012.

[37] F. Mirzaei and S. Roumeliotis, "Optimal estimation of vanishing points in a manhattan world," in *IEEE International Conference on Computer Vision*, Nov. 2011, pp. 2454–2461.

[38] J.-P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *IEEE International Conference on Computer Vision*, Sep. 2009, pp. 1250–1257.

[39] H. Lee, E. Shechtman, J. Wang, and S. Lee, "Automatic upright adjustment of photographs," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 877–884.

[40] A. Hanbury and H. Wildenauer, "Robust camera self-calibration from monocular images of manhattan worlds," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2831–2838.

[41] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "Tilt: Transform invariant low-rank textures," *International Journal of Computer Vision*, vol. 99, no. 1, pp. 1–24, Aug. 2012.

[42] R. von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 722–732, Apr. 2010.

[43] P. Monasse, J.-M. Morel, and Z. Tang, "Three-step image rectification," in *Proceedings of the British Machine Vision Conference*, Aug. 2010, pp. 89.1–10.

[44] F. Korč and W. Förstner, "eTRIMS Image Database for interpreting images of man-made scenes," Dept. of Photogrammetry, University of Bonn, Tech. Rep. TR-IGG-P-2009-01, Apr. 2009.

[45] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111 – 122, 1981.

[46] C. R. Jung and R. Schramm, “Rectangle detection based on a windowed hough transform,” in *Proceedings of the Computer Graphics and Image Processing, XVII Brazilian Symposium*, Oct. 2004, pp. 113–120.

[47] D. P. Nikolaev, S. M. Karpenko, I. P. Nikolaev, and P. P. Nikolayev, “Hough transform: underestimated tool in the computer vision field,” in *European Conference on Modelling and Simulation*, Jun. 2008, pp. 238–243.

[48] N. Skoryukina, D. P. Nikolaev, A. Sheshkus, and D. Polevoy, “Real time rectangular document detection on mobile devices,” in *Seventh International Conference on Machine Vision*, Nov. 2014, pp. 94 452A–1–94 452A–6.

[49] [Online]. Available: http://2015.icdar.org/program/competitions/

초록

본 논문은 카메라 위치 예측을 이용하여 영상을 rectification 및 stitching 하는 방법에 관한 것이다. 먼저, 기존의 영상 stitching 방법들은 다음의 두 가지 환경에 대하여 독립적으로 알고리즘이 개발되었다. 첫 번째는 카메라의 광학적 중심이 고정되어 있는 경우이고, 두 번째는 촬영된 물체가 평면일 경우이다. 본 논문에서는 이들을 하나의 프레임워크로 통합하여 해결하는 방법을 제안한다. 이를 위해 각 카메라를 6 개의 변수 (회전 행렬을 위한 3 개의 변수와 이동 벡터를 위한 3 개의 변수) 로 모델링하고 world coordinate 상에서 reconstructed 평면 (z-평면) 을 정의한다. 그리고 입력 영상과 reconstructed 평면 간의 관계를 이용하여 그 평면에서 정합 에러 (registration error) 를 반영하는 목적 함수를 설계한다. 그 후, LM (Levenberg-Marquardt) 알고리즘을 통해 목적 함수를 최소화하는 입력 카메라들의 변수들을 구하고 입력 영상이 어떠한 환경에 처해있는지 자동적으로 계산하여 stitching 및 rectification 된 영상을 생성한다.

두 번째로는 제안하는 방법을 이용하여 한 장의 영상을 rectification 하는 방법을 제안한다. 이를 위해서 영상에서 추출된 라인 세그먼트를 이용하고, 대다수의 라인 세그먼트들은 world coordinate 상에서 주 방향으로 정렬되어 있다는 가정 하에 이들을 수평 또는 수직 방향으로 정렬시키는 호모그래피를 찾고자 한다. 즉, 카메라의 위치를 4 개의 변수 (회전 행렬을 위한 3 개의 변수와 초첨 거리를 위한 1 개의 변수) 로 모델링하고 입력 영상의 라인 세그먼트를 reconstructed 평면 위로 투

영시키는 목적 함수를 설계한다. 또한, 추출된 라인 세그먼트 중에 주 방향으로 정렬되어 있지 않은 이상치를 거르기 위한 방법도 제안한다.

제안하는 방법들은 다양한 어플리케이션에 적용 가능하다. 본 논문에서는 제안하는 영상 rectification 방법을 이용하여, 동영상 기반에서 문서의 경계선을 검출하는 방법을 제안한다. 이를 위해서 문서와 배경의 색상 정보를 이용하여 데이터 항을 설계하고, 라인 세그먼트가 주 방향으로 정렬되어 있는 정도와 색상의 대조 정도를 이용하여 경계선에 대한 항을 설계한다. 또한, 인접한 프레임 간의 시간적 일치도를 위한 시간적 정보에 대한 항을 설계함으로써 동영상에서 문서의 경계선을 검출한다. 실험을 통하여 제안하는 방법이 다양한 동영상에 대하여 문서의 경계선을 잘 검출함을 보여준다.

주요어: 영상 stitching, 영상 rectification, 카메라 위치 예측, 문서 검출 및 segmentation

학번 : 2011-30241