International Journal of Control, Automation and Systems 14(3) (2016) 814-826
http://dx.doi.org/10.1007/s12555-014-0203-4
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Automated Architectural Reconstruction Using Reference Planes under Convex Optimization

My-Ha Le, Hoang-Hon Trinh, Van-Dung Hoang, and Kang-Hyun Jo*

Abstract: In this paper, a method for the automated reconstruction of architectures from two views of a monocular camera is proposed. While this research topic has been studied over the last few decades, we contend that a satisfactory approach has not yet been devised. Here, a new method to solve the same problem with several points of novelty is proposed. First, reference planes are automatically detected using color, straight lines, and edge/vanishing points. This approach is quite robust and fast even when different views and complicated conditions are presented. Second, the camera pose and 3D points are accurately estimated by a two-view geometry constraint in the convex optimization approach. It has been demonstrated that camera rotations are appropriately estimated, while translations induce a significant error in short baseline images. To overcome this problem, we rely only on reference planes to estimate image homography instead of using the conventional camera pose estimation method. Thus, the problem associated with short baseline images is adequately addressed. The 3D points and translation are then simultaneously triangulated. Furthermore, both the homography and 3D point triangulation are computed via the convex optimization approach. The error of back-projection and measured points is minimized in L_∞-norm so as to overcome the local minima problem of the canonical L_2-norm method. Consequently, extremely accurate homography and point clouds can be achieved with this scheme. In addition, a robust plane fitting method is introduced to describe a scene. The corners are considered as properties of the plane in order to limit the boundary. Thus, it is necessary to find the exact corresponding corner positions by searching along the epipolar line in the second view. Finally, the texture of faces is mapped from 2D images to a 3D plane. The simulation results demonstrate the effectiveness of the proposed method for scenic images in an outdoor environment.

Keywords: 3D reconstruction, convex optimization, correspondence, planes detection, plane fitting, plane homography, sum of square error differences, two-view geometry.

1. INTRODUCTION

The automated architectural reconstruction of large-view scenes is one of the most important processes in virtual environments, scene planning, and the navigation of autonomous mobile robots. While some progress has been made in the field of 3D reconstruction over the last few years, there are still no methods that satisfy the requirement of robustness with the capability to produce structures with high accuracy. In addition, some approaches require a large amount of work to be performed by hand or with an apparatus, such as laser radar, airborne light detection, and ranging. Such schemes are usually expensive and require much more time for data acquisition.

Three-dimensional metric reconstruction from two uncalibrated views is one of the classical topics in 3D computer vision, and extensive work has been conducted in this field over the last few decades. Some typical articles related to vision research can readily be found [1, 2]. For a quick overview, it is important to separate the existing approaches into several groups. The first group comprises methods of upgrading projective metric reconstruction based on self-calibration. In such schemes, projective reconstruction is performed using an epipolar geometry and subsequently upgrading to metric reconstruction by finding the intrinsic parameters (camera calibration) via Kruppa's equations derived from the fundamental matrix [3, 4]. While these methods do not require previouse

Manuscript received May 20, 2014; revised November 5, 2014 and March 4, 2015; accepted June 11, 2015. Recommended by Associate Editor Gon-Woo Kim under the direction of Editor Euntai Kim. This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MOE) (2013R1A1A2009984). Also, we would like to express our thanks to Ho Chi Minh City University of Technology and Education.

My-Ha Le is with the Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology and Education, 01, Vo Van Ngan St., Thu Duc Dist., Ho Chi Minh City, Viet Nam (e-mail: halm@hcmute.edu.vn). Hoang-Hon Trinh is with the Faculty of Electrical and Electronic Engineering, Ho Chi Minh City University of Technology, 268 Ly Thuong Kiet Str., Dist. 10, Ho Chi Minh City, Viet Nam (e-mail: trinhhoanghon09@gmail.com). Van-Dung Hoang is with the Quang Binh University, 312 Ly Thuong Kiet Str., Dong Hoi City, Viet Nam (e-mail: dunghv@qbu.edu.vn). Kang-Hyun Jo is with the Graduate School of Electrical Engineering and Information Systems, University of Ulsan, Daehak road 93, Nam-gu, Ulsan 680-749, Korea (e-mail: acejo@ulsan.ac.kr).
* Corresponding author.

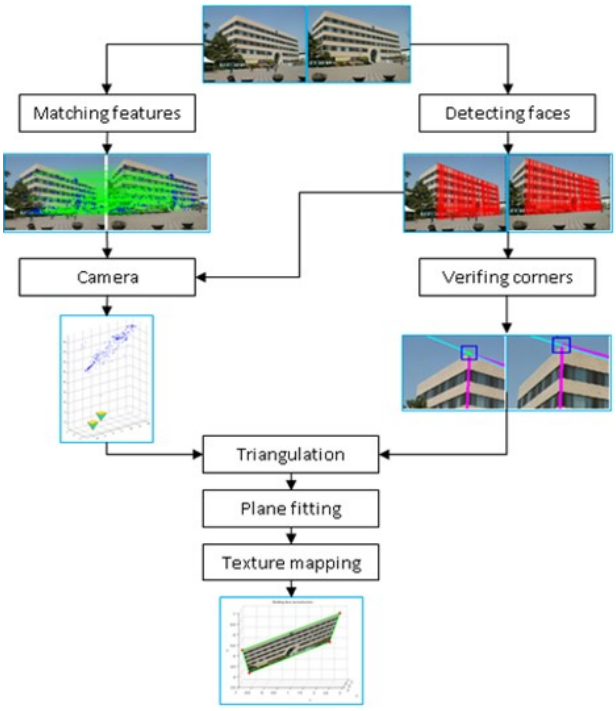


Fig. 1. Flowchart of the architectural reconstruction process.

knowledge of the scene, an epipolar geometry must be employed to estimate the fundamental matrix. In the normal camera motion case, the estimation is ordinary. However, in the critical configuration [5], the fundamental matrix cannot be estimated. In this paper, such a problem is solved as a short baseline problem. The second group of methods [6, 7] is based on scene knowledge or structure, including parallelism, orthogonality, and scenes with planes, in order to perform camera calibration.

Most of the research cited above is based on the linear constraint of 3D points and a camera. However, 3D triangulation using the L_2–norm usually becomes trapped in local minima. Thus, we relied on scene structure in a different way. Our work is in some ways similar to that outlined in [8], where the linear method is based on the algebraic cost-function. Here, a more global optimal solution is proposed using the convex optimization approach. The original idea was pioneered in [9], where the L_∞-norm was utilized to solve most common problems with a multiview geometry. In this paper, it is shown that the problems of automated reference plane detection and a small baseline can be solved absolutely via convex optimization.

Without using any additional devices (e.g., electromagnetic devices) for calibrated images from a single camera, our proposed method overcomes some of the critical configurations mentioned above. A flowchart of the proposed method is shown in Fig. 1. The generation of reference planes is the first requirement to compose image constraints: this is known as homography. Thus, a plane detection algorithm must be performed for an image pair as the initial step. In this paper, only a brief description of the method is presented, as a more detailed explanation is provided in our earlier work [10, 11]. Here, it may be asked why homography is needed but two-view geometric essential constraint for camera pose estimation. The answer is because of the short baseline image problem. It was shown in [12] that camera rotation is appropriately estimated, but the translation error increases significantly as the baseline decreases. This problem will lead to serious errors in the 3D structure. Therefore, in the modern 3D reconstruction community, the use of camera pose instead of only rotation is avoided [13]. Rotation can be achieved through the use of other sensors or scene properties (e.g., through homography). Here, a reference plane was utilized for this purpose. Specifically, at least four correspondence points inside the plane of an image pair must be extracted; these points obviously belong to the same plane in a real scene. The SIFT algorithm [14] is considered as an appropriate solution to this problem. It should be noted that only correct correspondence features inside the plane are extracted for homography computations. Thus, RANSAC-based [15] outlier removal of two feature sets will be performed. Although many methods can be utilized for the classical homography estimation problem (see [16]), we propose a new, more optimal approach. Some aspects of the scheme are similar to those of robust estimation using RANSAC, but we minimize the error in a different way through the use of the L_∞-norm under convex optimization. Such an approach sometimes yields the same accuracy rate, but it allows the local minimum problem of the L_2-norm in the general case to be avoided. The next step, 3D point triangulation, has been one of the classic topics in field of 3D vision over the last few decades. Most researchers have used the direct linear transform (DLT) method for triangulation. However, this scheme will produce a large error, especially when a lack of information exists (e.g., in a two-view geometry). In contrast, we approach the problem from the standpoint of convex optimization. Once the homography is obtained as mentioned above, scene structure and translation will be optimally triangulated. Under the assumption that the scene architecture consists of rectangular planes, point clouds must be described in terms of planar patches. Several researchers have proposed methods to solve this problem, and most of the schemes are based on the RANSAC algorithm. In this work, a similar method of RANSAC-based plane fitting is employed with some modifications. Usually, the plane will be limited by the convex hull of its own points. Here, the boundary of a plane is derived by its properties. The corners of a plane in two views allow the boundary to be determined. Therefore, the corresponding corner points must be extracted. This simple task can be performed by searching the sum of the square error along the epipolar line in the second view. Finally, the textures

of the 2D image are mapped to the reconstructed planes. One useful option in the practical application of this proposed method is that the true information of objects can be obtained if we know the displacement of the camera at adjacent moments or the baseline when a stereo system is used. For example, our proposed method concentrates on building an object automatically with face detection. Thus, 3D reconstruction with true building information, e.g., the distance from the camera to the building and the dimensions of the building, can be achieved automatically.

This paper is organized into 6 sections. A summary of the plane detection method is given in Section 2; a building face is used as one example to analyze the method. In Section 3, the homography estimation method based on a two-view geometry and the convex optimization approach is discussed. 3D point triangulation and camera translation are also investigated. Plane fitting and corner correspondence are explained in Section 4, while experimental details and obtained results are discussed in Section 5. Conclusions derived from the research are ultimately given in Section 6.

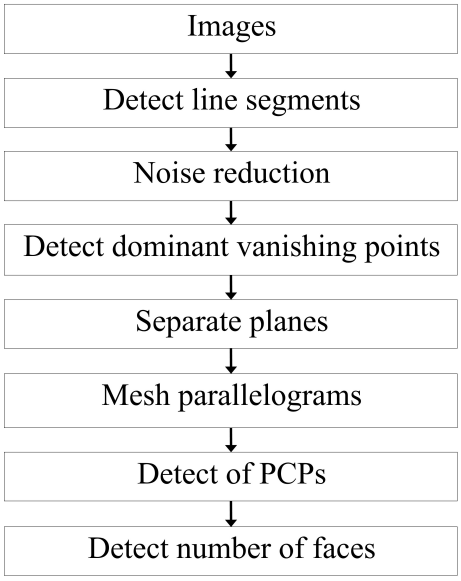


Fig. 2. Flowchart of the building face detection process.

2. REFERENCE PLANE DETECTION

Line segments and belongings in the appearance of a building are used as geometrical and physical properties, respectively. The geometrical properties are represented as principal component parts (PCPs) in the form of a set of doors, windows, walls, and so on. For the physical properties, the color, intensity, contrast, and texture of regions are used. The analysis process begins with the detection of straight line segments. We used the m-estimator sample consensus (MSAC) method [17] to group such parallel line segments, which have a common vanishing point. The detail explanation of applying MSAC Method for vanishing point detection could be found in our former research [11]. One dominant vanishing point was calculated for the vertical direction and a maximum of five dominant vanishing points were computed for the horizontal direction. A mesh of basic parallelograms is created by one of the horizontal groups and the vertical group. Each mesh represents one face of the building. The PCPs are formed by merging an area of basic parallelograms with similar colors. In addition, the PCPs are classified into doors, windows, and walls. Finally, the structure of the building is described as a system of hierarchical features. The building is represented by a number of faces, and each face is represented as a color histogram vector. The color histogram vectors are computed by the wall region of a face. A flowchart of this method is shown in Fig. 2.

2.1. Line segment detection

The first step in line segment detection is edge detection for an image. We used the edge detection function with the Canny edge detector algorithm. The function is run with a threshold that is automatically chosen. The threshold is a two-element vector in which the first element is the low threshold, and the second element is the high threshold. The value for thresh is relative to the highest value of the gradient magnitude of the image, the default value is [0.25 0.6]. A straight line segment is a part of an edge including a set of pixels which have a number of pixels larger than the given threshold (T_1) and all pixels are aligned. Therefore, if we draw a line through the ends, the distance from any pixel to this line is less than another given threshold (T_2). In our case $T_1 = 10$ pixels and $T_2 = \sqrt{2}$. The thresholds are figured out through the heuristics. It is correct for almost outdoor image database. The changes of these thresholds do not affect too much on the final results. When a different threshold is chosen the number of parallel line will change. In all cases, the vanishing points can be found out with a small number of parallel lines. Otherwise, when the number of parallel lines increase the vanishing points detection is more accurate but the processing time is high. Therefore, the building face still can be detected.

According to the threshold for RANSAC algorithm based plane fitting, the probability density function of the data points should be compute first. With that we can compute a confidence interval: 98% of the data points are located at a maximum distance D from their true position. Then we could use this D to set the distance threshold. The plane will be found out when the number of inlier point n get the maximum value.

2.2. Reducing the low contrast lines

Low contrast lines usually arise from such objects as an electrical line or the branch of a tree. Most low contrast

lines are not usually located on the edges of PCPs because two regions with a high contrast in color may be distinguished at these edges. Here, the intensity was based on two regions beside a line so that low contrast lines can be discarded.

2.3. MSAC-based detection of dominant vanishing points

The line segments are roughly separated into two groups. The vertical group contains line segments that create a maximum angle of 20? with the vertical axis. The remanent lines are treated as horizontal groups. For the fine separation stage, the robust MSAC method was used to estimate the vanishing point.

2.4. Horizontal vanishing point detection

Horizontal vanishing point detection is performed in a manner similar to that outlined in the previous section. In reality, a building is a prototypical structure where many faces and various colors appear in images. Therefore, it is necessary to separate the faces. A maximum of five dominant vanishing points is computed for the horizontal direction.

2.5. Separation of planes as the faces of building

The vertical segments are extended by their middle points and vertical vanishing point. The number of intersections of vertical lines and horizontal segments was used to detect and separate the planes as the faces of a building; the obtained results are shown in Fig. 3. Coarse face separation is performed according to the following rules:

1) If the same region contains two or more horizontal groups, then priority is given to the group with a larger number of segment lines.

2) If two or more horizontal groups are distributed along the vertical direction, then priority is given to the group with a lower order of dominant vanishing points. The second stage is the recovery stage. Some horizontal segments that are located close to the vanishing line of two groups are usually improperly grouped. Instead of belonging to lower order groups, some segments are in higher order groups and must be recovered. The recovery stage is performed from low to high order. The third stage involves finding the boundaries of the faces.

3. CAMERA REGISTRATION AND 3D POINT TRIANGULATION

The camera rotation and translation can be computed by the DLT method proposed in [18]. However, a problem occurs when the baseline of views is too small or close to zero. In [12], it was shown that the rotation could be appropriately estimated while the translation was affected by the baseline distance. The error of translation

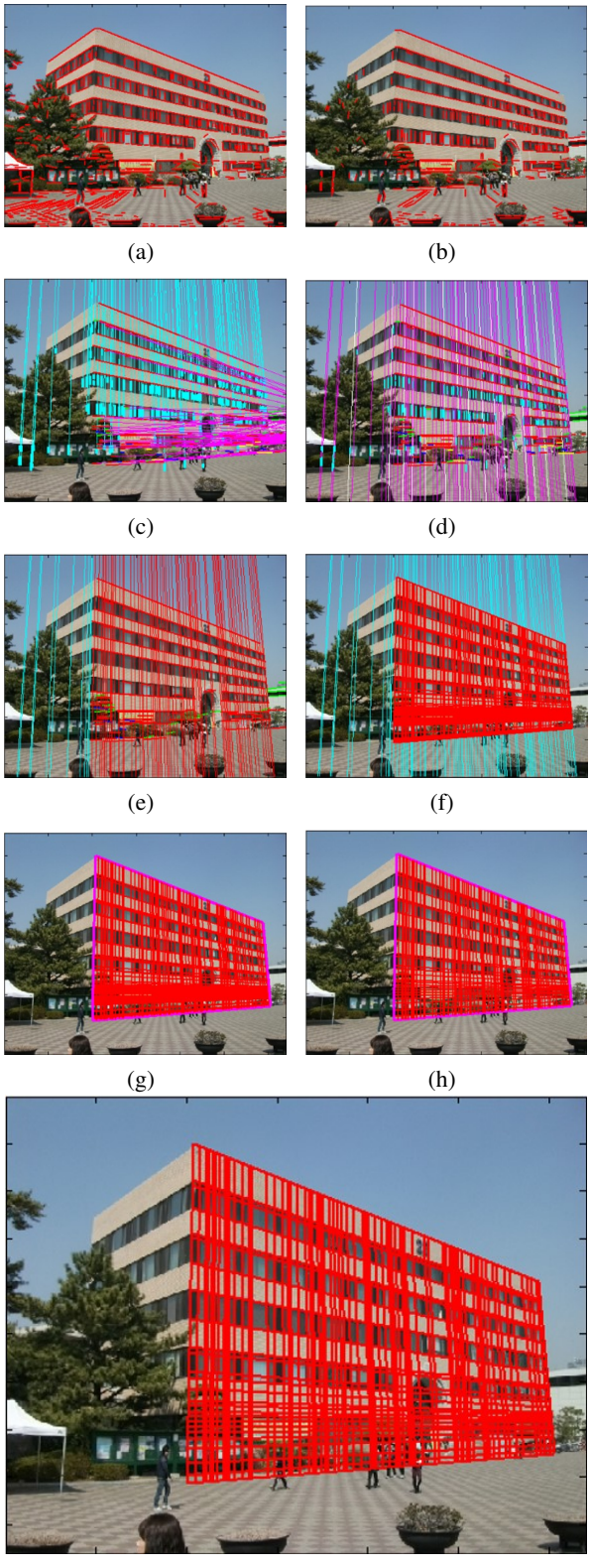


Fig. 3. Building detection results. (a) and (b) are line segment and low contrast lines reducing. (c) to (e) is detection of dominant vanishing points. (f) is roughly detected facets. (g) and (h) are boundary finding. (i) is final face detection.

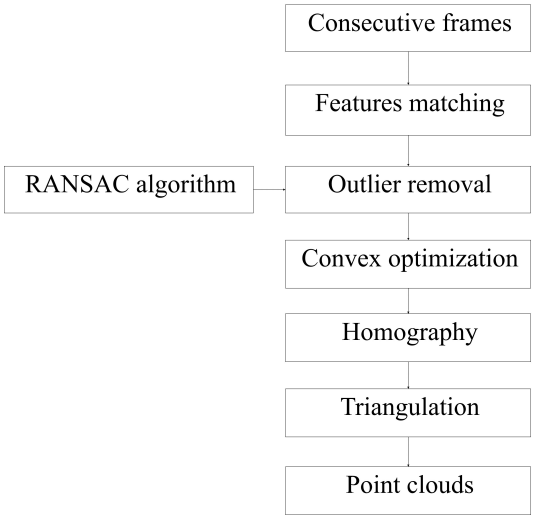


Fig. 4. Camera registration and 3D point triangulation scheme.

is increased significantly when the camera distance decreases. This problem will lead to poor accuracy in the triangulation step. In this work, the building faces are robustly detected in the previous step. Thus, the plane homography is considered as a robust constraint even if the camera distance is small or zero. This constraint can be computed by many different methods (see [16]). Here, an approach based on global optimization is used because it possesses certain advantages, as discussed in [9]. 3D points and translation will then be computed simultaneously. The step-by-step procedure is described as follows. Firstly, we extract invariant features inside the plane region of an image pair. Correct matching is achieved according to RANSAC-based outlier removal. Secondly, the bisection algorithm is performed to minimize the error of back-projection and measured points under the convex optimization approach. The problem of 3D point triangulation and translation estimation is also addressed in the same manner. In the subsection below, we will recall some principles of the pinhole camera model, the correspondence problem, and convex optimization theory applied to a multi-view geometry for the problem formulation and solution. The general arrangement of these steps is shown in Fig. 4.

3.1. Camera model

The projective geometry is used throughout this work to describe the perspective projection of a 3D scene onto 2D images [17]. This projection is expressed as follows:

$$x = PX, \tag{1}$$

where P is a 3×4 projection matrix that describes the perspective projection process, while $X = [X, Y, Z, 1]^T$ and $x = [x, y, 1]^T$ are vectors containing the homogeneous coordinates of the 3D world coordinates and 2D image coordinates, respectively. When ambiguity in the geometry is metric, i.e., Euclidean up to an unknown scale factor, the camera projection matrices can be expressed in the following form:

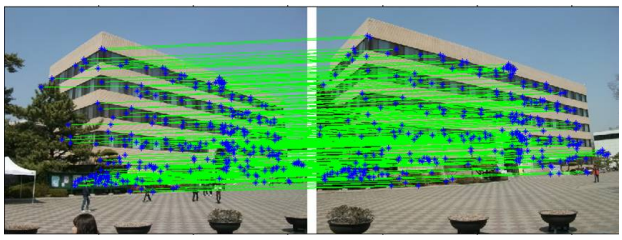


Fig. 5. SIFT feature extraction and matching after RANSAC-based outlier removal.

$$P = K[R| - Rt], \tag{2}$$

where t and R denote the translation and rotation of the camera, and K is an upper diagonal 3×3 matrix containing the intrinsic camera parameters. Here, K can be written as

$$K = \begin{bmatrix} f_x & s & u_x \\ 0 & f_y & u_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where f_x and f_y are the focal length divided by the horizontal and vertical pixel dimensions, s is a measure of the skew, and (u_x, u_y) is the principal point.

3.2. Feature extraction and matching

Many kinds of features have been considered for feature extraction and matching problems, including SIFT, Harris [19], SURF [20], and GHOL [21]. Among these, SIFT was first presented by David G Lowe in 1999 and was completely applied in a pattern recognition problem in 2004. This algorithm is quite invariant and robust for feature matching with scaling, rotation, or affine transformations. As such, we utilized SIFT feature points to find the corresponding points of image pairs. The SIFT algorithm performs the following tasks: scale-space extrema detection, accurate keypoint localization, orientation assignment, and keypoint descriptor. SIFT features and matching are applied for one image pair, as shown in Fig. 5. The result of a correspondence point will be used to compute the fundamental matrix described in the next step.

3.3. Two-view geometry

The epipolar constraint represented by a 3x3 matrix is called the fundamental matrix, F. The method based on two-view geometry theory was thoroughly examined in [17]. According to the theory, two image point u and u'

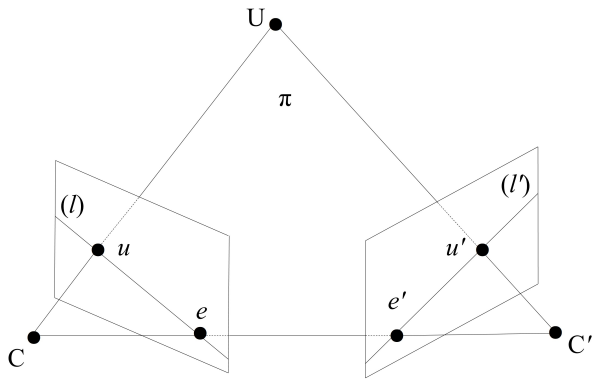


Fig. 6. Epipolar geometry principle.

are projected from a 3D point U observed by two cameras with optical centers C and C'. These five points form a common plane called the epipolar plane. The points e and e' are called the epipoles of the two cameras; epipole e' is the projection of the optical center C of the first camera in the image observed by the second camera and vice versa. If u and u' are projections of the same point, then u' must lie on the epipolar line associated with u, hence the epipolar constraint. The epipolar constraint plays an important role in stereo vision analysis. When the intrinsic parameters of the cameras are known, the epipolar constraint can be represented algebraically by a 3x3 matrix known as the essential matrix. Otherwise, the epipolar constraint represented by a 3x3 matrix is called the fundamental matrix, F. An illustration of the epipolar geometry principle is shown in Fig. 6.

The essential equations:

$$u'^T F u = 0, \tag{4}$$

$$l' = Fu, \quad l = F^T u', \tag{5}$$

$$Fe = 0, \quad F^T e' = 0. \tag{6}$$

3.4. Plane homography estimation

According to the explanation above, it is easy to realize that the constraint of two sets of points that belong to the same plane will be considered as the homography constraint, i.e., the fundamental matrix F becomes the homography matrix H. Now, a method based on optimization is performed to find a 3x3 matrix H. It is reformulated as a quasi-convex optimization problem. Some methods have been proposed from research outlined in [16]. One highly accurate method is known as back-projection error minimization. It is easy to see that solving the L_2-norm for error minimization is a difficult non-convex problem. Such an approach can yield local minima instead of a single global minimum when error is minimized in the L_∞-norm. In this section, the above problem will be formulated and solved using the bisection convex optimization method.

3.4.1 Problem formulation

Let u_i', $i = 1, \ldots, m$, denote a set of planar points represented by homogeneous plane coordinates, and u_i, $i = 1, \ldots, m$ denote the corresponding image features also represented by homogeneous coordinates. The two point sets are related by the relation $u_i \approx Hu_i'$, where H is represented by $H = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & 1 \end{bmatrix}$. The problem is to find a matrix H such that the projection of u_i' through H is nearest to u_i, i.e., the cost function is minimized:

$$\sum_{i=1}^{m} d(u_i, H(x)u_i')^2. \tag{7}$$

Here, $d(\cdot,\cdot)$ represents the geometric distance between two points in the image. In [9], it was noted that the L_2 -norm error of this cost function creates three local minima, whereas the L_∞-norm creates a single minimum. In this paper, we use a procedure similar to the RANSAC method in [16]. This leads to the following minimization problem:

$$\begin{aligned} &\min \max_i d(u_i, H(x)u_i') \\ &\text{subject to } \lambda_i(x) > 0, \quad i = 1,2,...,m. \end{aligned} \tag{8}$$

Here, $\lambda_i(x)$ is the depth of a point in image i. It is easy to realize that the square image distance is a rational function of x:

$$d(u, H(x)u')^2 = \frac{f_1(x)^2 + f_2(x)^2}{\lambda(x)^2}, \tag{9}$$

where $f_1(x)^2$, $f_2(x)^2$, and $\lambda(x)^2$ are affine functions in x with coefficients determined by u and u'.

Remark 1: The problem $\min \max_i d(u_i, H(x)u_i')$ has some convexity properties. Thus, this problem can be solved by a quasiconvex optimization method.

3.4.2 Bisection-based quasiconvex optimization solver

In this section, the bisection method will be applied to solve the problem mentioned above. Suppose that γ is an upper bound of the objective function in problem (8). According to the theory outlined in [22], this problem can be formulated as follows:

$$\begin{aligned} &\min \gamma \\ &\text{Subject to } \|f_{1i}(x), f_{2i}(x)\| \le \gamma \lambda_i(x) \\ &\qquad\qquad \lambda_i(x) > 0, \quad i = 1,2,...,m. \end{aligned} \tag{10}$$

If γ is unknown, Equation (10) can be re-written in second-order cone program (SOCP) feasibility problem

form:

$$\begin{aligned} &\text{Find } x \\ &\text{Subject to } \|f_{1i}(x), f_{2i}(x)\| \leq \gamma\lambda_i(x) \\ &\qquad\qquad \lambda_i(x) > 0, \;\; i = 1,2,...,m. \end{aligned} \tag{11}$$

If it is assumed that the optimal γ^* is lower than some threshold of γ_u pixels, then $\gamma^* \in [0, \gamma_u]$. Until now, typical convex feasibility problem solving has been applied. The detailed algorithm is presented below.

Algorithm 1: Bisection-based quasiconvex optimization solver

Given: *optimal value* $f_0^* \in [\gamma_l, \gamma_l]$ *and tolerance* $\varepsilon > 0$

Repeat

1. $\gamma := (\gamma_l + \gamma_u)/2$
2. *Solve the convex feasibility problem*
3. *If feasible* $\gamma_u := \gamma$, *else* $\gamma_l := \gamma$

Until $\gamma_u - \gamma_l \leq \varepsilon$

Remark 2: Note that if we define $H_1 = I$, then homography mapping from image 1 to image 2 is $H_{12} = HH_1 = H$. In this method, at least four coplanar points are needed to generate the homography matrix.

3.5. Simultaneous triangulation and translation estimation

Over the last two decades, many methods have been proposed for triangulation and motion estimation in the field of 3D computer vision. Among them, particular attention is given here to linear multi-view reconstruction and camera recovery using a reference plane [8]. We also present a similar approach of using a reference plane, but triangulation and pose estimation is based on global optimization. After the homography of views is obtained in the previous step, it is possible to estimate 3D points and camera translation simultaneously. Triangulation with a known geometry constraint (homography in this case) will also be reformulated as a quasi-convex optimization problem. Similar to the homography computation, some researchers have proposed methods using the L_∞-norm [9] or the L_∞-norm combined with the L_1-norm [23] instead of the L_2-norm to minimize the residual error associated with a measured feature and the back-projection of 3D points. It is easy to see that solving the L_2-norm for more than two cameras is a difficult non-convex problem. Such an approach can yield local minima instead of a single global minimum when error is minimized in L_∞-norm.

Let $P_i = \begin{bmatrix} H_i & t_i \end{bmatrix}, i = 1, 2, ..., m$ denote the m known homogaphy cameras, where t_i is an unknown vector (in this case $m = 2$). Here, $u_i$are the projection of point U in 3D space (both are expressed in homogeneous coordinates). The problem of finding the camera position and$U$ given the homography matrix and image points is known as triangulation and position estimation. In the ideal case (absence of noise), the triangulation is ordinary. In the case where noise is present, the back-projection of point

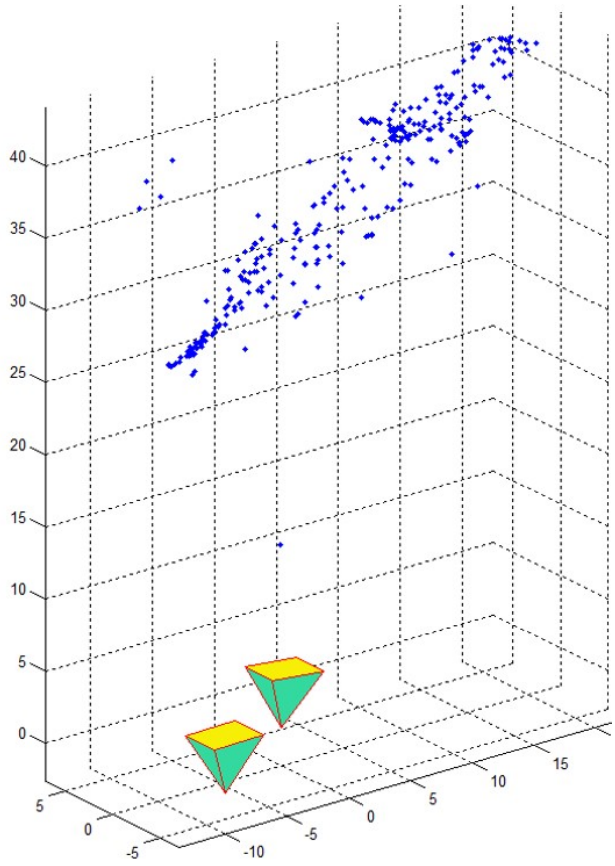


Fig. 7. Camera position with sparse 3D points of a scene.

U to the image plane does not coincide with u_i. Thus, we must find the camera position and point U such that its projection is nearest to u_i, i.e., the cost function is minimized:

$$\sum_{i=1}^{m} d(u_i, P_i U)^2. \tag{12}$$

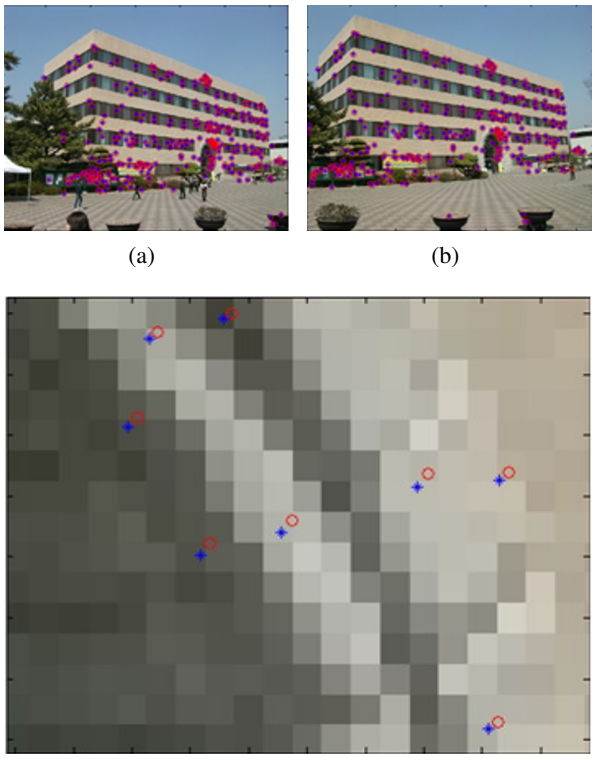
Here, $d(\cdot,\cdot)$ represents the geometric distance between two points in the image. In this paper, we use a procedure similar to that outlined in [16]. The known homography problem will now be described in detail. Considering the camera matrix P_i, we will try to solve the minimization problem:

$$\begin{aligned} &\min \max_i d(u_i, P_i U(x)) \\ &\text{Subject to } \lambda_i(x) > 0, \;\; i = 1,2,...,m, \end{aligned} \tag{13}$$

where $\lambda_i(x)$ is the depth of a point in image i. It is easy to see that the square image distance is a rational function of x:

$$d(u, PU(x))^2 = \frac{f_1(x)^2 + f_2(x)^2}{\lambda(x)^2}, \tag{14}$$

where $f_1(x)^2$, $f_2(x)^2$, and $\lambda(x)^2$ are affine functions in x with coefficients determined by u and P. The solution is



(a) (b) (c)

Fig. 8. The measured points and back-projection of 3D points: (a) and (b) are back-projection and measured points (blue *: measured points, red o: back-projection), and (c) is magnified view.

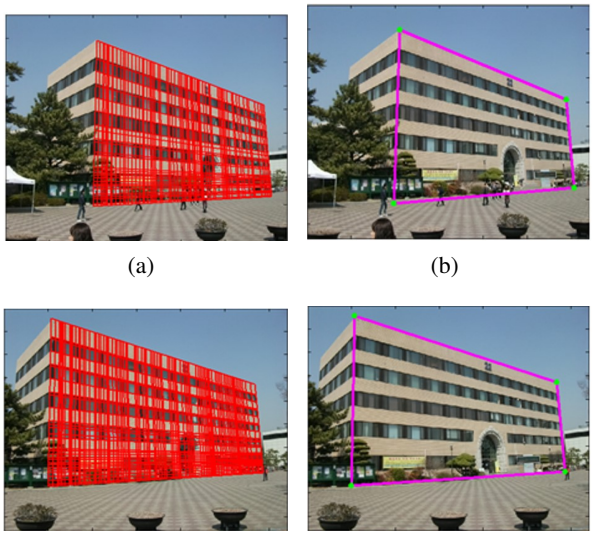
similar to the bisection method in (11). The simulation result of a point cloud and camera position is shown in Fig. 7. The back-projection error is also displayed in Figs. 8(a)-(c). This error can be used to check the accuracy of translation estimation and 3D point triangulation.

4. PLANE MODELING

Once the point cloud is generated, a way to present these scenery points should be devised. Usually, a plane is needed to fit these points because most of the scene structure is a plane or planar patches. The boundary of a plane may be a convex hull over a set of points. In contrast, we find the plane limits through the existing properties of the plane. When the corner of a plane is determined in 2D images, it can be used instead of a convex hull to circumscribe the plane. The texture is mapped from the original image to the determined planes. The manner in which to perform these tasks is described below.

4.1. Plane fitting

Several researchers have investigated plane fitting in previous studies. Here, particular attention is given to



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

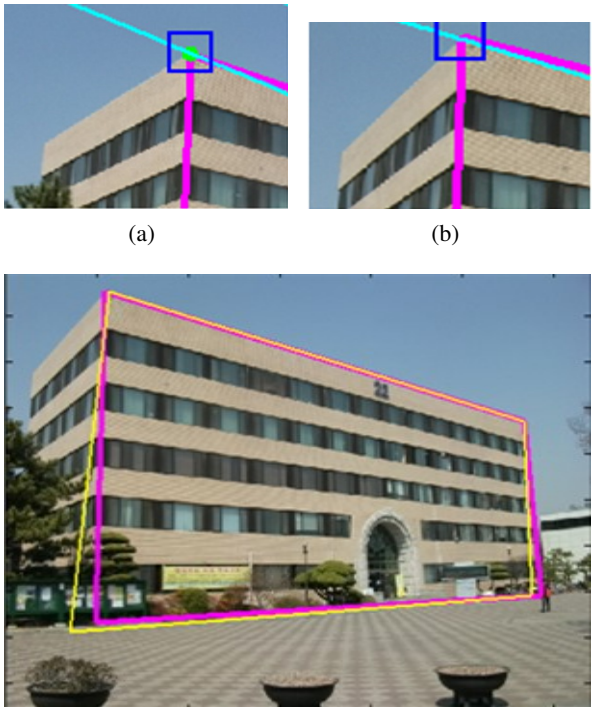
Fig. 9. The corner point extracted from the building faces: (a) and (c) are original face detection; (b) and (d) are corner points.

the RANSAC-based fitting method and plane sweeping scheme proposed in [24] and [25], respectively. Without using any additional scene assumptions, except that the scene can be represented by a planar patch, we also used the RANSAC-based approach in plane fitting in a manner similar to the canonical line fitting problem. First, a point in the set of points is randomly chosen. Then, the n closet points within a certain distance are also selected. This can be considered as the error associated with this method. The points inside the upper and lower plane are inliers. The RANSAC algorithm is used to reject the outliers and fit a plane to this subset.

Here, the process of randomly selecting the first point is repeated until the maximum number of iterations is reached or the number of remaining points is smaller than n. We can see that the size of the planar patches is dependent on the number of points n.

4.2. Corner point verification

In order to limit the boundary of the planar patches, some additional information must be known, i.e., the curve surrounding the plane must be determined. With the assumption that the object plane has a rectangular shape, connected lines through the corner points represent the boundary. Thus, the problem is simplified to find the corner point correspondence. Here, correlation-based similarity measurement methods are considered, and a simple sum of the square error differences (SSD) method is employed. We first need to extract the corner points in the first frame and find their matching points in the second frame automatically. To accomplish this task, the following steps are performed:



(a) (b)

(c)

Fig. 10. Corner point verification: (a) and (b) are epipolar lines for corner searching. (c) is verification of face of the second view (magenta: original face, yellow: verified face).

1) Pick up a corner point in the first frame. The corner are pick up by the crossing point of the margin horizontal and vertical line of detected building face.
2) Find the epipolar line in the second frame;
3) Use window sliding along the epipolar line and find the sum of the square error differences with respect to the one in first frame. The corresponding point is our matching point. A brief description of this method is given below. The formulation of SSD is expressed as:

$$D = \iint_R [I(R(x,y)) - I(R'(x',y'))]^2 w(x,y)dxdy, \quad (15)$$

where *I(R(x, y))* is the intensity at point *(x, y)* of a region *R* in the first image, and *w(x, y)* is a Gaussian weighting function. A point *(x', y')* in *R'* that gives a minimum SSD from *R* is considered a feature match of *(x, y)*. The results of corner point matching from two views are shown in Fig. 10.

5. EXPERIMENTS

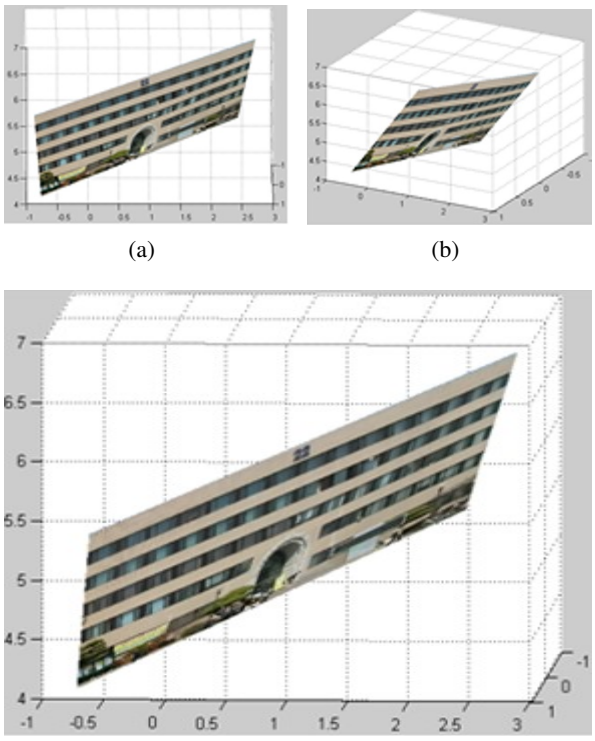
In this section, experiments conducted to evaluate the proposed method are described. The main objects are large-view scenes in an outdoor environment. The dataset images were acquired by a mini perspective camera, Fujifilm. The first experiment was carried out on the offline database. The building faces for reconstruction is assumed planar. The processing speed is neglected however the speed is presented as follow: The building face detection process in our experiment cost nearly 0.5 second on Intel(R) Core(TM) i5 CPU 750@2.67 GHz with 3 GB of RAM under Matlab environment with a MOSEK add-in toolbox [26]. The reconstruction and color rendering are cost nearly 0.2 second with image of size [640x480]. The accurate reconstructed points and camera pose are reflected in the back-projection error. The magnified figure shows that the error distance is almost less than one pixel. This excellent result is difficult to obtain even when bundle adjustment (L_2-norm minimization) is used in a two-view geometry. With high-accuracy point cloud reconstruction, points on the same plane in the 2D image are almost distributed on the plane in 3D space. This is one benefit of using RANSAC-based plane fitting in the next step. For comparison the accuracy of the proposed method and existing reference, author implemented standard linear algorithms and bundle adjustment which optimizes using L2-norm) to exactly the canonical data base. The data is available at http://www.robots.ox.ac.uk/ṽgg/data.html. In order to compare the sparse point cloud reconstruction method, the camera is calibrated beforehand. The Root Mean Squares (RMS) errors of the reprojected and corresponding feature points are given instead of the sum of squares errors. In the experiments, the RMS reprojection and L_∞ errors are measured. Table 1 shows experimental results of existing references and the proposed method. After solving these problems, the RMS errors for the Bundle Adjustment approache and L_∞ minimization (convex optimization) are 0.40495 pixels and 0.38783 pixels, respectively. The corresponding L_∞ errors are 0.89991 pixels and 0.85414 pixels. For a more clear visualization, we performed texture mapping from the original plane to the fitted planes. Figs. 11(a)-(c) show the different angles of view of the building faces, which are appropriate for determining the location of building faces in 3D space by the mapped texture.

In the second experiment, the same process is employed, but the number of faces is 2. The metric ambiguity of reconstruction can be realized if the corner of two faces as well as the rectangular shape of the faces is examined. In a real building, faces are truly rectangular and the angle of a corner between two faces is nearly 90 degrees. These structural properties are reconstructed quite well, as shown Fig. 12. In the last experiment, a more complicated scene is used to test the robustness of our algorithm. Two images with a size of 1600x1200 containing a scene with a building covered by complex and interlacing connection wires were fed into the system. Similar to the previous case, the faces of the building were detected robustly and quite fast. The results obtained for the reconstructed planes with texture are shown in Fig. 13.

The building detection rate was verified by two data sets. The first one is ZuBuD data set [27]; the second one

Table 1. Experimental results for the Bundle Adjustment approache and L_∞ minimization.

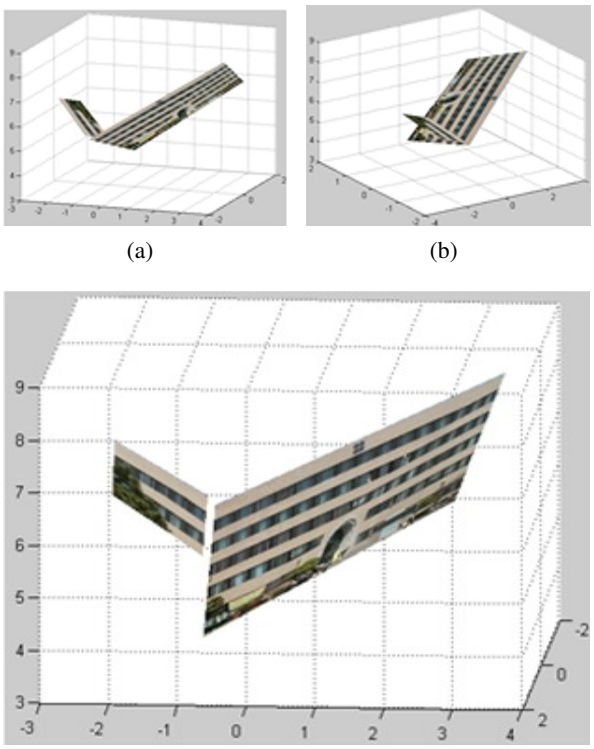
Methods	Bundle Adjustment	L_∞ Minimization
View 1 RMS	0.44343	0.43906
View 2 RMS	0.35416	0.38123
View 3 RMS	0.41216	0.33632
Total RMS	0.40495	0.38783
L-infinity error	0.89991	0.85414



(a) (b)

(c)

Fig. 11. Reconstruction results obtained for a one-face building: (a), (b) and (c) are different angle of views of the reconstructed face.



(a) (b)

(c)

Fig. 12. Reconstruction results obtained for a building with two faces: (a), (b) and (c) are different angle of views of the reconstructed faces.

is our data set. The ZuBud data set contains 201 buildings, each building is appeared five poses in the training data. Also there are 115 images for a test set. Totally, there are 1120 building images. Our data set comprises of 880 images with 680 building and 200 non-building images. The results are summarised in Table 2. Obviously, The proposed method is robust with color and intensity changes because of utilizing geometry features to detect the vanishing point and then the building faces. The pre-process of the proposed method is based on the edge detection. So the color and intensity changes do not affect the final results. There is no claim for cylindrical form building such as Baroque architecture but in the data set some of them can be detected correctly. The important point for accuracy detection of building face is the structure of parallel lines. When the number of dominant vanishing points can be found out then the building faces are also can be detected. The ideal for extending the planar patch as the element to mimic the complex shape is really interesting. This problems will be discussed in the further researches. The results of building detection under occlusion and intensity changes conditions are illustrated in Appendix A.

6. CONCLUSION

In this work, automated architecture reconstruction based on a reference plane under convex optimization from two-view images was presented. Some advantages of the scheme were realized and the method was mathematically and experimentally verified. The first advantage is that the proposed pipeline utilized only a homography constraint to avoid the serious error of the canonical pose estimation method. This also allows the significant error of translation in the small baseline problem to be overcome. The second benefit is that we avoid using a bundle adjustment, i.e., the L_2-norm in back-projection error minimization, to estimate homography. This method can lead to local minima. Instead, we employed convex optimization in our algorithm. By utilizing the L_∞-norm for mini-

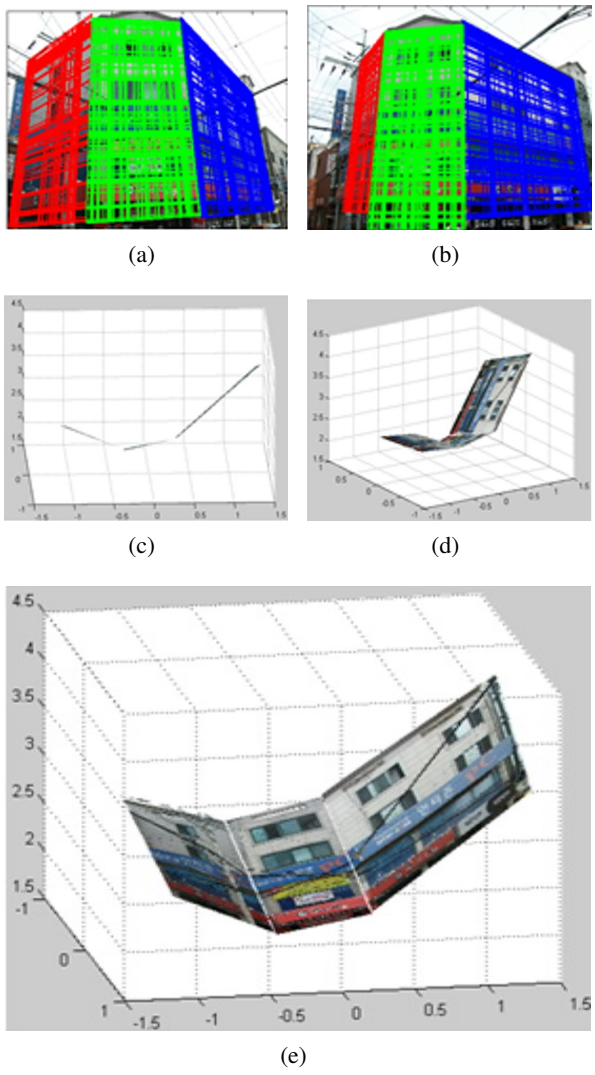


Fig. 13. Reconstruction results obtained for a building with three faces: (a) and (b) are face detection. (c)-(e) are different angle of views of the reconstructed faces.

Table 2. Summary of building detection.

Data sets	Type of images	Number of images	Corrections	Confusions
ZuBuD Data	Building images	1120	1117	3
UlBuD01 Data	Building images	680	660	20
UlBuD01 Data	Non-building Images	200	184	16

mization of the back-projection error, we can estimate the homography quite accurately. Furthermore, extreme exactness of the translation estimation and point clouds is achieved. One additional strong point is the robustness of reference plane detection even under a complex occlusion condition and in an outdoor environment with intensity changes. The iteration method of plane fitting is also a significant advantage. In future work, a general solution for high-accuracy multiple-view reconstruction will be developed using reference planes under global optimization. In addition, we will improve the method by upgrading to dense point clouds using a stereo rid or omni-directional camera. Trajectory estimation will also be considered. The final goal will be the application of this method to real scene modeling systems.

APPENDIX A

The results of building detection under occlusion and intensity changes are showed in Fig. 14. The non-planar buildings or cylindrical form building are detected in Fig. 15.

REFERENCES

[1] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 10, pp. 133-135, 1981.

[2] R. I. Hartley, "Estimation of relative camera positions for uncalibrated cameras," *Proc. of ECCV*, *Lecture Notes in Computer Science,* vol. 588, pp. 579-587, 1992. [click]

[3] S. J. Maybank and O. D. Faugeras, "A theory of self-calibration of a moving camera," *International Journal of Computer Vision*, vol. 8, no. 2, pp. 123-151, 1992. [click]

[4] Q.-T. Luong and O. D. Faugeras, "The fundamental matrix: theory, algorithms, and stability analysis," *International Journal of Computer Vision*, vol. 17, no. 1, pp. 43-75, 1996.

[5] R. I. Hartley and F. Kahl, "Critical configurations for projective reconstruction from multiple views," *International Journal of Computer Vision*, vol. 71, no. 1, pp. 5-47, 2007. [click]

[6] B. Caprile and V. Torre, "Using vanishing points for camera calibration," *International Journal of Computer Vision*, vol. 4, no. 2, pp. 127- 140, 1990. [click]

[7] B. Triggs, "Autocalibration from planar scenes," *Proc. of ECCV '92, Lecture Notes in Computer Science*, vol. 1, pp. 89-105, 1998.

[8] C. Rother and S. Carlsson, "Linear multi view reconstruction and camera recovery using a reference plane," *International Journal of Computer Vision*, vol. 49, no. 2, pp. 117-141, 2002.

[9] F. Kahl and R. Hartley, "Multiple view geometry under the L_∞-norm," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1603-1617, 2008. [click]

[10] H.-H. Trinh, D.-N. Kim, and K.-H. Jo, "Supervised training database for building recognition by using cross ratio invariance and SVD-based method," *International Journal of Applied Intelligence*, Vol. 32, no. 2, pp. 216-230, 2010.

(a)

(b)

(c)

Fig. 14. The examples of building detection: (a) building detection in the cases of multiple faces, multiple buildings and complex environment; (b) incorrect detection; (c) non-building images.

Fig. 15. Examples of detection of non-planar buildings.

[11] H.-H. Trinh and K.-H. Jo, "Image-based structural analysis of building using line segments and their geometrical vanishing points," *Proc. of the SICE-ICASE International Joint Conference*, pp. 566-571. 2006. [click]

[12] O. Enqvist, F. Kahl, and C. Olsson, "Non-Sequential Structure from Motion," *Proc. of the Eleventh Workshop on Omnidirectional Vision, Camera Networks and Non-classical Camera,* pp. 264-271, 2011. [click]

[13] R. Hartley and F. Schaffalitzky, "L_∞ minimization in geometric reconstruction problems," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 504-509, 2004. [click]

[14] D. Lowe, "Distinctive image features from scale-invariant interest points," *International Journal of Computer Vision*, Vol. 60, no. 2, pp. 91-110, 2004. [click]

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated car-tography," *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, 1981. [click]

[16] A. Agarwal, C. V. Jawahar, and P. J. Narayanan, "A survey of planar homography estimation techniques," *IIIT Technical Report IIIT/TR/2005/12*, 2005. [click]

[17] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision,* Cambridge University Press, Cambridge, 2004.

[18] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp.756-770, 2004. [click]

[19] C. Harris and M. Stephens, "A combined corner and edge detector," *Proc. of the 4th Alvey Vision Conference*, pp. 147-151, 1998.

[20] H. Bay, T. Tuytelaars, L. V. Gool, "SURF: speeched up robust features," *Proc. of ECCV*, Vol. 3951, pp. 404-417, 2006. [click]

[21] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, 2005. [click]

[22] S. Boyd and L. Vandenberghe, *Convex Optimization,* Cambridge University Press, Cambridge, 2004.

[23] A. Dalalyan and R. Keriven, "L_1-penalized robust estimation for a class of inverse problems arising in multiview geometry," *Proc. of the 23rd Annual Conference on Neural Information Processing Systems*, pp. 441-449, 2009.

[24] T. Werner and A. Zisserman, "New techniques for automated architectural reconstruction from photographs," *Proc. of ECCV*, vol. 2, pp. 514-555, 2002.

[25] R. T. Collins, "A space-sweep approach to true multi-image matching," *Proc. of Computer Vision and Pattern Recognition*, pp. 358-363, 1996. [click]

[26] "The MOSEK optimization toolbox for MATLAB manual," www.mosek.com

[27] H. Shao and L. V. Gool, "Zubud-zurich buildings database for image based recognition," *Swiss FI of Tech., Technical Report,* no. 260, 2003.

My-Ha Le received his B.E. and M.E. degrees from Department of Electrical and Electronic Engineering of Ho Chi Minh University of Technology, Viet Nam, in 2005 and 2008, respectively. Since 2007, he has been serving as a faculty member in Department of Electrical and Electronic Engineering, Ho Chi Minh City University of Technology and Education, Viet Nam. He received his Ph.D. degree from Electrical Engineering Department of University of Ulsan, Korea, in 2013. His research interests include 3D computer vision, pattern recognition, and vision based robotics.

Van-Dung Hoang received his bachelor of informatics from Hue University, Viet Nam in 2002, and master of computer sciences from Hanoi National University of Education, Vietnam in 2007. Since 2002, he has been serving as a lecturer in University of Quang Binh, Vietnam. He received Ph.D. degree from Electrical Engineering Department of University of Ulsan, Korea, in 2014. His research interests include pattern recognition, machine learning, computer vision, and vision based robotics.

Hoang-Hon Trinh was born in Dong Nai, Viet Nam, in 1973. He received his B.E. and M.E. degrees in Electrical-Electronic Engineering of Ho Chi Minh City University of Technology, Viet Nam, in 1997 and 2002 respectively. He received his Ph.D. degree from Electrical Engineering Department of University of Ulsan, Korea, in 2008. His research interests include computer vision, pattern recognition, understanding and reconstructing outdoor scenes, designing the outdoor mobile robot for civil and special applications.

Kang-Hyun Jo received his Ph.D. degree from Osaka University, Japan, in 1997. He joined the School of Electrical Eng., University of Ulsan right after having one year experience at ETRI as a post-doc research fellow. Dr. Jo has been active to serve for the societies for many years as directors of ICROS (Institute of Control, Robotics and Systems) and SICE (Society of Instrumentation and Control Engineers, Japan) as well as IEEE IES. He is currently contributing himself as an AE for a few journals, such as IJCAS (International Journal of Control, Automation and Systems), TCCI (Transactions on Computational Collective Intelligence) and IteN (IES Technical News, online publication of IEEE), TIE. He had involved in organizing many international conferences such as ICCAS, FCV, ICIC and IECON. He had visited for performing his research activity to Kyushu University, KIST and University of California Riverside. His research interest covers in a wide area where focuses on computer vision, robotics, and ambient intelligence.