Camera Localization and Building Reconstruction from Single Monocular Images

Ruisheng Wang
Centre for Intelligent Machines, McGill University
3480 University Street, Montreal, Quebec, Canada
ruisheng@cim.mcgill.ca

Frank P. Ferrie
Centre for Intelligent Machines, McGill University
3480 University Street, Montreal, Quebec, Canada
ferrie@cim.mcgill.ca

Abstract

This paper presents a new method for reconstructing rectilinear buildings from single images under the assumption of flat terrain. An intuition of the method is that, given an image composed of rectilinear buildings, the 3D buildings can be geometrically reconstructed by using the image only. The recovery algorithm is formulated in terms of two objective functions which are based on the equivalence between the vector normal to the interpretation plane in the image space and the vector normal to the rotated interpretation plane in the object space. These objective functions are minimized with respect to the camera pose, the building dimensions, locations and orientations to obtain estimates for the structure of the scene. The method potentially provides a solution for large-scale urban modelling using aerial images, and can be easily extended to deal with piecewise planar objects in a more general situation.

1. Introduction

1.1 Background

3D object reconstruction from images is a common problem in computer vision and photogrammetry. Traditional image-based modelling methods use multiple images to obtain 3D structure of the scene. In some cases, however, there is only one view of a scene available. 3D reconstruction from one single view is clearly an ill-posed problem, because a single projective image cannot provide 3D information without further assumptions. In this paper, we reconstruct rectilinear buildings up to a scale factor from single images under the assumption of flat terrain. The method is based on feature correspondence between pre-defined parameterized 3D models and corresponding image edges. Because no existing edge detection algorithms can provide only useful edges reliably from images of a common scene, human intervention is always needed. In our method, the image edges are digitized manually for the model-to-image correspondence, then the algorithm recovers the 3D geometry of the buildings automatically. The method consists of self-calibration and metric reconstruction. In the self-calibration stage, by choosing an object-centred coordinate system in the model space as shown in Fig. 1, the camera orientation can be solved first then the camera location and the model dimensions, given the correspondences between the model (building 1 in Fig. 1) and its corresponding image edges. The division of unknown parameters into two groups which are solved separately is the key for the 3D reconstruction using single images only. In the metric reconstruction stage, other buildings (e.g. building 2 in Fig. 1) are reconstructed based on the recovered camera pose and feature correspondence between the model (building 2 in Fig. 1) and its corresponding image edges. More buildings can be reconstructed using the same way.

The three main contributions of this paper are:
1. We demonstrate rectilinear buildings can be geometrically reconstructed with correct topological relationships by using single images only.
2. Not like other model-based approaches (e.g., [3]), our method does not require a model-to-image projection process, therefore is faster and more effective than existing methods.
3. The method potentially provides a solution for large-scale urban modelling using aerial images, and can be easily extended to deal with piecewise planar objects in a more general situation.

1.2 Related Work

In this section, we present the most important related work and discuss the main differences with ours. Liu et al. [6] solved for the camera rotation first and then the camera translation. They considered three camera rotation angles as obtained from a nominal orientation by small perturbations, e.g. 0 degrees. Based on this assumption, their algorithm only works if the three camera Euler rotation angles are less than 30 degrees. Kumar and Hanson [5] solved for the rotation and translation simultaneously by adapting an iterative technique formulated by Horn [4]. The initial estimates for translation and rotation are required to make the nonlinear algorithm converge. They also reported

U.S. Government work not protected by U.S. copyright
that the initial rotation estimates for some data sets must be within 40 degrees for all the three Euler angles representing the rotation. When initial estimates for rotation and translation are not available, they sampled rotation space, and each of the samples was used as an initial estimate for the rotation estimation by a method akin to Liu et al. [6]. The estimated rotation and translation based on the rotation samples are then used as initial estimates for solving the camera rotation and translation simultaneously. Taylor and Kriegman [1] estimated both the camera positions and the structure of the scene from multiple images. Based on a random initial estimate of rotation, the translation and model parameters are computed as initial inputs for the subsequent model-to-image fitting procedure. If the disparity between predicted edges and the observed edges is smaller than some preset threshold, the minimum is accepted as a feasible estimate. Debevec et al. [3] argued that if the algorithm begins at a random location in the parameter space, it stands little chance of converging to correct solution. They developed a method to directly compute a good initial estimate for the camera positions and model parameters, and then use those estimates as initial inputs for the subsequent model-to-image fitting process. A combination of vanishing points and model-to-image fitting approach can be found in the paper [10]. Multiple piecewise planar object reconstruction from single images is addressed through the constraints of connectivity and perspective symmetry in the paper [11].

Our approach builds on this line of work. Described is a two-step iterative scheme for recovering camera orientation that, unlike existing methods, does not require a good initial guess for the rotation. Instead, the good initial estimate for the rotation is computed directly by using coplanarity constraints. The camera translation and predefined model parameters are determined based on the calculated rotation through a linear least squares minimization. The 3D reconstruction of buildings is based on the recovered camera pose and the assumption of flat terrain. Unlike existing methods, our method does not require a model-to-image projection process, and is particularly suitable for oblique images with large shooting angles in urban environments.

2. The Method

2.2.1 Notation

Fig. 1 shows how a straight line segment, model edge 67, in a cube model (building 1) projects onto the image plane of a camera. The coordinates of two endpoints of the projected image edge 67 in the camera coordinate system can be represented as \((x_1, y_1, -f), (x_2, y_2, -f)\). The camera position relative to the object coordinate system is represented in terms of a rotation matrix \(R\) and a translation vector \(t\). The straight line 67 can be defined by a pair of vectors \((v, u)\) in the object coordinate system where \(v\) represents the direction of the line and \(u\) represents a point on the line. \(m\) is normal vector of the projection plane defined by the two lines \((C_6, C_7)\) and camera centre \(C\) in the camera coordinate system. The coplanar constraints derived in [1] are outlined in the following. The fundamental relation of the imaging geometry can be represented by the equation (1),

\[
m = R(v \times (u - t))
\]

Equation (1) is based on the fact that the 3D model lines
(e.g. line 67) in the camera coordinate system must lie on the projection plane formed by lines \((C_6, C_7)\) and camera centre \(C\).

\[
m^T R v = 0 \\
\]
\[
m^T R(u - t) = 0
\]

Equations (2) and (3) are deduced from equation (1), which shows that the determination of camera rotation \(R\) can be independent from the estimation of camera position \(t\) and model parameters. Note \(v\) becomes a known vector in the object-centered coordinate system which is parallel to the Y axis, while \(u\) can be represented by the model parameters. The normal vector \(m\) can be defined by the intersection of the projection plane \(C_{67}\) with the image plane as shown in Fig. 1 and represented in the equation (4),

\[
m_x t + m_y y - m_z f = 0
\]

where \(m_x, m_y, m_z\) are the coordinates of the normal vector in the camera coordinate system and \(f\) is the focal length of the camera; \(x\) and \(y\) are points on the image edge. Given an observed image of edge 67, the observed normal vector \(m'\) can be obtained by the equation (5),

\[
m' = (x_1, y_1, -f)^T \times (x_2, y_2, -f)^T
\]

Without losing generality, the location and orientation of the building 2 can be represented by a building vertex (e.g. vertex 3 \((X_3, Y_3)\) in the Fig. 1), a building orientation along the X axis (e.g. the \(\alpha\) in Fig. 1), and the building’s dimension of length, width, and height (e.g. \(L, W, H\) in Fig. 1). Those unknown parameters are solved in the metric reconstruction stage.

2.2 Recovery Algorithm

The recovery algorithm takes as input, a set of correspondences between edges in the models and edges in the image. The correspondences are performed manually. The algorithm then automatically recovers camera pose and model dimensions, consisting of self-calibration and metric reconstruction. In the first step, the focal length is estimated by locating two finite vanishing points associated with two perpendicular directions in 3D space. The camera pose and the model parameters are recovered with respect to an object-centred coordinate system. In the second step, the spatial relationship of buildings is represented by three intrinsic parameters (building length, width, and height) and three extrinsic parameters (a building vertex location and building orientation). Those parameters can be determined by using model-to-image correspondence and the recovered camera pose.

2.1.1 Self-Calibration

The self-calibration requires more than three line correspondences between the pre-defined model edges and the image edges, which consists of determination of the focal length, initial estimate of camera rotation, refinement of camera rotation, and determination of camera translation and model dimensions.

Determination of the focal length. Focal length information can be obtained from image EXIF tags. In the absence of EXIF tags, the focal length can be estimated as long as two finite vanishing points associated with two perpendicular directions in 3D space can be located in the image. For most man-made planar objects, it is not difficult to find two sets of lines in the images where the lines from each set in 3D space are parallel and two lines from different sets are perpendicular. Let’s denote \(X^\infty, Y^\infty\), as the vanishing points in two orthogonal directions. The focal length \(f[2]\) can be determined using the equation (6).

\[
f = \sqrt{-(x_{w1} * x_{w2} + y_{w1} * y_{w2})}
\]

Because the determination of the principal point using vanishing points is not so accurate [9], a feasible approximation that the principal point lays at the image centre is employed in this paper.

Initial estimate of camera rotation. The objective function of obtaining initial estimates for camera rotation is formulated according to the Equation (2) as shown in the Equation (7),

\[
O_i = \sum (m_i^T R v_i)^2
\]

where \(i\) is the number of the model edges, \(n\) is the total number of the employed model edges, \(m_i\) and \(v_i\) are the corresponding normal vector and direction of the model edge, \(R\) is 3x3 camera rotation matrix.

To minimize the objective function (7) that sums the extents to which the rotation \(R\) violates the constraints arising from Equation (2), we employ a three-loop computation for initial estimates of the camera rotation. The accuracy of initial estimates relies on the step of the loops. In this paper the step of the loops is set as 1, and the output can reach an accurate initial rotation estimate on the order of 1 degree. Choosing a large step may increase the search speed, but also decrease the accuracy of the initial rotation estimate. Actually, even if the initial estimates differ considerably from the correct solution, the algorithm
can still converge. Our implementation on a Pentium IV 1.6 Ghz computer with 1.5 Gb of memory takes 3-4 seconds to run the algorithm with a loop step of 1.

Refinement of camera rotation. Once initial camera rotation is obtained, a non-linear technique based on Gauss-Newton method is applied to the minimization problem. The direct calculation of Jacobian matrix of the objective function $O_i$ is complex. To simplify the linearization of $O_i$, we rewrite the rotation matrix $R$ as a multiplication of three sequential rotations, and compute the first derivative for each rotation angle. The Jacobian matrix of $O_i$ can then be formed as,

$$J_{\alpha \omega \kappa} = \begin{bmatrix} m_i^T R_{\omega \alpha} v_1 & m_i^T R_{\omega \kappa} v_1 & m_i^T R_{\kappa \alpha} v_1 \\ \vdots & \vdots & \vdots \\ m_i^T R_{\omega \alpha} v_n & m_i^T R_{\omega \kappa} v_n & m_i^T R_{\kappa \alpha} v_n \end{bmatrix},$$

where

$$R_{\omega} = R_{\omega} 0 0 1, \quad R_{\kappa} = \begin{bmatrix} 0 & 0 & -\cos \kappa \\ 0 & \cos \kappa & \sin \kappa \end{bmatrix} R,$$

$$R_{\kappa} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} R.$$

Given the three initial camera rotations obtained from the previous step, the Gauss-Newton algorithm computes accurate estimates of the camera rotations within 2-3 iterations.

Determination of Camera Translation and Model Dimensions. The objective function for determining camera translation and model dimensions is formulated according to Equation (3) as shown in Equation (8),

$$O_2 = \sum (m_i^T R (u_i - t))^2,$$

where $i$ is the number of the model edges, $n$ is the total number of the employed model edges, $m_i$ and $u_i$ are the corresponding normal vector and point on the model edge. In the case of rectilinear buildings, the minimization of the objective function $O_2$ is a constrained quadratic form minimization problem, and can be solved through a set of linear equations. It is also important to keep in mind that the resulting dimensions of the scene and camera translations are up to a scale factor.

2.1.2 Metric Reconstruction

The metric-reconstruction also requires more than three line correspondences between the pre-defined model edges and the image edges, which consists of initial estimate of building orientation, refinement of building orientation, and determination of building dimensions and location.

Initial estimate of building orientation. The three directions of model edges, $v_1$ (e.g. model edge 67 of the building 2 in Fig. 1), $v_2$ (e.g. model edge 78), and $v_3$ (e.g. model edge 27), can be represented as shown in Equation (9).

$$v_1 = \begin{bmatrix} -W \sin \alpha \\ W \cos \alpha \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} v_1, \quad v_3 = \begin{bmatrix} 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} H \end{bmatrix}$$

Since the value of building dimensions $(W, H)$ does not affect the building orientation, the only unknown parameter of building orientation is $\alpha$. At this stage, the camera orientation is known. The objective function of Equation (7) is employed to obtain an initial estimate for $\alpha$. The minimization of the objective function (7) sums the extents to which the model orientation $v$ violates the constraints arising from Equation (2).

Refinement of Building Orientation. Once initial building orientation is obtained, a non-linear technique based on the Gauss-Newton method is applied to the minimization problem. Based on Equation (10), the minimization is straightforward since there is only one unknown parameter $\alpha$,

$$-a_1^T W \sin \alpha + a_2^T W \cos \alpha = 0$$

where $a_1^T = m_i^T R$.

Given the initial building orientation obtained from the previous step, the Gauss-Newton algorithm computes the accurate building orientation within 2-3 iterations.

Determination of building dimensions and location. The building dimensions and location is determined by minimizing objective function $O_2$ as shown in Equation (8). In this stage, image norm vectors $m$ and camera rotation $R$ and translation $t$ are known. The unknowns are points on the model edges which are expressed as linear functions of building dimension and location, and can be solved through a set of linear equations. The same way can be employed to reconstruct more buildings.

3. Experimental results

This section describes a series of experiments that were
carried out in order to evaluate the effectiveness of proposed algorithm. Simulation is first used to systematically vary key parameters such as the camera parameters and measurement of the image segments; thereby enabling us to characterize the degradation of the algorithm in extreme situations, and to develop robust solutions for real world applications. Two examples from a real camera are shown to gauge practical results.

3.1. Simulation Experiments

The synthetic image data was generated with a virtual camera and two cubic building models as shown in Fig. 1. Table 1 shows information about the camera intrinsic and extrinsic parameters, assuming that the image centre lies at the centre of the image frame. Table 2 shows information about the building dimensions, locations, and orientations.

<table>
<thead>
<tr>
<th>Focal Length (m)</th>
<th>0.0798</th>
<th>Pixel size (μm)</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₀ (m)</td>
<td>-500.672</td>
<td>α (°)</td>
<td>50.346</td>
</tr>
<tr>
<td>Y₀ (m)</td>
<td>100.317</td>
<td>ψ (°)</td>
<td>3.582</td>
</tr>
<tr>
<td>Z₀ (m)</td>
<td>650.783</td>
<td>κ (°)</td>
<td>2.787</td>
</tr>
</tbody>
</table>

Table 1 Camera intrinsic and extrinsic parameters

<table>
<thead>
<tr>
<th></th>
<th>Building 1</th>
<th>Building 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (m)</td>
<td>40</td>
<td>26.413</td>
</tr>
<tr>
<td>Width (m)</td>
<td>20</td>
<td>20.927</td>
</tr>
<tr>
<td>Height (m)</td>
<td>30</td>
<td>22.315</td>
</tr>
<tr>
<td>Orientation along X axis (°)</td>
<td>0</td>
<td>α = 30.856</td>
</tr>
<tr>
<td>Location of a building model vertex (m)</td>
<td>X₃</td>
<td>100.512</td>
</tr>
<tr>
<td></td>
<td>Y₃</td>
<td>-200.217</td>
</tr>
</tbody>
</table>

Table 2 Building parameters of dimensions, locations and orientations

Inaccurate estimates of the focal length. The effects of incorrect estimates of the focal length on the estimation of the second building model were examined. The incorrect estimates of the focal length were obtained by multiplying the focal length by a scale as shown in Table 3 row 1. Entries in column with the focal length scale 1.0 correspond to the experiments with correct focal length.

<table>
<thead>
<tr>
<th></th>
<th>0.968</th>
<th>0.987</th>
<th>1.000</th>
<th>1.013</th>
<th>1.032</th>
</tr>
</thead>
<tbody>
<tr>
<td>L (m)</td>
<td>25.13</td>
<td>25.887</td>
<td>26.413</td>
<td>26.931</td>
<td>27.649</td>
</tr>
<tr>
<td>W (m)</td>
<td>18.65</td>
<td>20.069</td>
<td>20.927</td>
<td>21.680</td>
<td>22.575</td>
</tr>
<tr>
<td>H (m)</td>
<td>18.42</td>
<td>20.697</td>
<td>22.315</td>
<td>23.978</td>
<td>26.493</td>
</tr>
<tr>
<td>α (°)</td>
<td>27.02</td>
<td>29.362</td>
<td>30.856</td>
<td>32.270</td>
<td>34.210</td>
</tr>
<tr>
<td>X₃ (m)</td>
<td>100.42</td>
<td>100.42</td>
<td>100.512</td>
<td>100.674</td>
<td>101.01</td>
</tr>
<tr>
<td>Y₃ (m)</td>
<td>-184.5</td>
<td>-194.05</td>
<td>-200.217</td>
<td>-206.113</td>
<td>-214.29</td>
</tr>
</tbody>
</table>

Table 3 The parameter estimation of the second building model with different focal lengths

Table 3 shows that inaccurate focal length has a significant effect on the Y coordinates of the building locations, and moderate effect on building dimensions and orientations. This can be partially explained by the fact that incorrect focal length affects camera pose which results in large positioning errors in single 3D object points. The building dimensions are less affected because the differences of building vertices compensate the positioning errors. The building orientation is also less affected because the incorrect focal length has less effect on camera orientation, and the building orientation is only related to camera orientation and independent of camera translation.

Errors from image centre offsets. The influence of the incorrect image centre on the second building model was analyzed introducing an error of 10, 20, 30 pixels on x and y coordinates of the image centre. In Table 4, entries in column with 0 pixel offsets correspond to the experiments with correct image centre.

<table>
<thead>
<tr>
<th></th>
<th>0 (pixel)</th>
<th>10 (pixel)</th>
<th>20 (pixel)</th>
<th>30 (pixel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L (m)</td>
<td>26.413</td>
<td>26.407</td>
<td>26.400</td>
<td>26.394</td>
</tr>
<tr>
<td>W (m)</td>
<td>20.927</td>
<td>20.913</td>
<td>20.899</td>
<td>20.885</td>
</tr>
<tr>
<td>H (m)</td>
<td>22.315</td>
<td>22.167</td>
<td>22.021</td>
<td>21.877</td>
</tr>
<tr>
<td>α (°)</td>
<td>30.856</td>
<td>30.779</td>
<td>30.701</td>
<td>30.625</td>
</tr>
<tr>
<td>X₃ (m)</td>
<td>100.512</td>
<td>100.542</td>
<td>100.572</td>
<td>100.603</td>
</tr>
<tr>
<td>Y₃ (m)</td>
<td>-200.217</td>
<td>-200.076</td>
<td>-199.935</td>
<td>-199.795</td>
</tr>
</tbody>
</table>

Table 4 The parameter estimation of the second building model with different errors in the image centre

The results indicate that the algorithm is much less sensitive to camera centre bias than it is to errors in the
estimate of the focal length. This can be partially explained by the fact that camera centre bias has less effect on camera pose, and thereby less effect on the second building model.

**Errors from image noise.** This simulation experiment is designed to examine how the accuracy of the reconstruction is affected by the errors in measurement of the image segments. A uniformly distributed random image error is added to the endpoints of the image segments. Entries in columns with 0 random errors correspond to the experiments with the image segments without errors. Table 5 shows that the reconstruction errors increase as the random image errors are increased.

<table>
<thead>
<tr>
<th>0 (pixel)</th>
<th>5 (pixel)</th>
<th>10 (pixel)</th>
<th>15 (pixel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L (m)</td>
<td>26.413</td>
<td>26.433</td>
<td>25.961</td>
</tr>
<tr>
<td>W (m)</td>
<td>20.927</td>
<td>21.059</td>
<td>20.650</td>
</tr>
<tr>
<td>H (m)</td>
<td>22.315</td>
<td>22.407</td>
<td>21.821</td>
</tr>
<tr>
<td>α (°)</td>
<td>30.856</td>
<td>31.214</td>
<td>31.011</td>
</tr>
<tr>
<td>X₃ (m)</td>
<td>100.512</td>
<td>101.313</td>
<td>98.044</td>
</tr>
<tr>
<td>Y₃ (m)</td>
<td>-200.217</td>
<td>-201.372</td>
<td>-195.254</td>
</tr>
</tbody>
</table>

Table 5 The parameter estimation of the second building model with random errors in the endpoints of image segments

The results also indicate that the location of the second building model is relatively sensitive to random errors in the edge endpoints compared with its dimensions and orientation.

In summary, the algorithm is robust in terms of estimating dimensions and orientation of the second building model. The positions of the second building model are relatively sensitive to the noises. The same rule applies to other buildings if the image has.

### 3.2 Real Data Experiments

We take pictures using a Canon PowerShot SD750 digital camera. The first image shown in Figure 2 is a salt box (3072x2304 pixels) whose dimensions are measured using a ruler as listed in Table 6. The edges of the box were digitized as the black lines shown in the Fig. 2. Assuming the image centre to lie at the centre of the image frame, with the information from camera specification and EXIF tag, the results were quite accurate and reach the expected accuracy of millimeter level as can be seen from the Table 6. Note that the width is more accurate than the height, which coincides with the previous accuracy analysis [8]. Fig. 3 shows the recovered camera pose and wire frame of the salt box using MatLab.

The second building image shown in Fig. 4 is the Burnside Hall (3072x2304 pixels) at the downtown campus of McGill University, Montreal. The measured image edge features are those black lines digitized using mouse. Fig. 5 shows a 3D model of Burnside Hall in Google Earth, which shows the irregularity of the building. We use a rectilinear building model to model Burnside hall approximately which induces measurement errors, especially in the corners of the building. Besides, the occlusions caused by snow make the accurate measurement of the building top and bottom difficult. Under all of this noise, however, we still achieve reasonable results. The computed dimensions of Burnside Hall are 35.44, 34.92, and 53.33 meters respectively, as compared to the model dimensions obtained from DWG file of 35.44, 32.42, 50.00 meters. Fig.
6 shows the recovered camera pose and wire frame of the Burnside Hall using MatLab.

Figure 4 Burnside Hall at McGill campus

Figure 5 3D model of Burnside Hall from Google Earth

Figure 6 Visualization of the recovered camera pose and wire frame of the Burnside Hall using MatLab

4. Conclusions

This paper presented a method to recover 3D rectilinear building models from single monocular images. The method uses the correspondences between predefined 3D models and their corresponding 2D images to obtain camera pose as well as parameters of 3D building models. The camera orientation is first recovered followed by solving translation and the first building model dimensions. The direct computation of the initial estimate for camera rotation effectively solved problems in the previous approaches (e.g., [1]), and the determination of camera pose as well as the first building model dimensions are much simpler than the previous methods (e.g., [3]). Under the assumption of flat terrain, more 3D building models can be reconstructed based on recovered camera pose through model-to-image correspondence.

Simulation experiments were carried out in order to investigate how the accuracy of the algorithm would be affected as different parameters were varied. These experiments show that estimates of dimensions and orientation of the second building model are not sensitive to various noises. But, the position of the second building model is relatively sensitive to noises. Real data experiments show that the algorithm robustly and accurately estimates camera pose and building dimensions provided that accurate image measurements are available.

This paper deals with buildings without occlusions. Further work is suggested to deal with buildings with occlusions and selection of optimal views, to achieve the goal of large-scale urban modeling with minimum number of images.

ACKNOWLEDGEMENT

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and McGill fellowships.

REFERENCES


Geometry and Image-based Approach, SIGGRAPH, pp. 11-20, August 1996.


