Abstract

3D models of urban sites with geometry and facade textures are needed for many planning and visualization applications. Approximate 3D wireframe model can be derived from aerial images but detailed textures must be obtained from ground level images. Integrating such views with the 3D models is difficult as only small parts of buildings may be visible in a single view. We describe a method that uses two or three vanishing points, and three 3D to 2D line correspondences to estimate the rotational and translational parameters of the ground level cameras. The valid set of multiple combinations of 3D to 2D line pairs is chosen by a hypotheses generation and evaluation. Some experimental results are presented.

1. Introduction

Accurate 3D models of buildings in a city are needed for a variety of applications, such as fly-through rendering and simulation for mission planning. Many methods for extraction of such 3D models from aerial images have been developed in recent years [1, 2]. Aerial images provide wide area coverage, however, the walls of buildings are either not visible or seen at a very low resolution. We need to utilize images taken from the ground to get accurate facade textures.

Some research has been done for 3D reconstruction of architectural models by using 3D to 2D line correspondences [3, 4]. However, these approaches are applicable only for one or two buildings, rather than an entire urban site, as they require laborious user interactions for a large number of models.

Automatically obtainable image features such as vanishing points that indicate the external and some internal parameters of camera are used to reduce user interactions for the correspondences. Previous methods describe techniques to calibrate the camera rotational parameters if there are two or three orthogonal vanishing points [5, 6, 7, 8, 9].

Stamos and Allen [10] use high precision range data from ground views and fit them into building facade planes through camera calibration with three orthogonal vanishing points. Even though their method is an automatic approach, it has been applied to only one or two buildings at a time. Coorg and Teller [11] construct a large set of 3D building models by using spherical mosaics produced from calibrated ground view cameras with a GPS device. Their method can be applied to model a relatively large site area, but the models are limited to simple shape buildings and the method does not capture the roof structure.

In our previous work [12, 13], we describe methods for automatically integrating the ground views with models obtained from aerial views. We estimate the global position and orientation of the camera used for taking the ground view images; then we attach facade textures to 3D building model. A disadvantage of this method is that a 3D to 2D line segment correspondence is used to estimate the world position of the camera, the result can be sensitive to the image noise of the 2D line segment. In other words, the length of the 2D line segment can be wrong due to incorrect edge detection or self-occlusions of buildings as depicted in Figure 1. This causes errors in calibration of the position of the ground view camera.

This drawback can be overcome by using three line correspondences instead of a single line segment. When three 3D to 2D line correspondences are not parallel, they
form one virtual line segment by intersecting three planes formed by a pair of 3D and 2D lines. The camera position computed by using this virtual line segment is robust regardless of the image positions of 2D line correspondences.

In this paper, we propose an additional method to the previous work [12, 13] for estimating correct position of the ground view camera after computing the orientation of the camera from automatically extracted three vanishing points.

2. Approach

In our approach, complex 3D building models are obtained from aerial images by using a previous approach [2] as shown in Figure 2.

![Figure 2](image)

Figure 2. 3D building models from an aerial view.

We integrate ground view images to 3D complex buildings in an urban area by using projective geometry [12, 13]. The rotational parameters of a ground view camera by using vanishing points obtained from global and local feature extraction method are robust. However, the position of the camera from a 3D to 2D line segment correspondence is sensitive to the length of image line segment, which can be wrong due to poor performances of edge detection or line fitting process.

We firstly use line correspondences instead of line segment correspondences to get robust results. A 3D to 2D line correspondence forms a 3D plane passing through the camera center. The intersection among three planes from three 3D to 2D line correspondences gives the world position of the camera center. Since we have multiple combinations of three line correspondences, we use a generation and evaluation of hypotheses approach to choose the best pairs of line correspondences.

Section 3 explains how to obtain orientational pose estimation of 3D building models automatically by using vanishing points, followed by estimating correct position of the camera. Some results and conclusion are discussed in section 4.

3. Pose Estimation

In this section, we briefly discuss how to recover the external rotational camera parameters from the vanishing points, not necessarily orthogonal in our previous work [12, 13]. Vanishing points from a set of parallel image lines have been shown to be useful for estimating camera parameters in previous work [5, 6, 7, 8, 9]. Given vanishing points and a 3D building model, the rotation of ground view camera can be recovered. Vanishing points can also be used for extracting 2D line correspondence candidates to compute the location of the camera, which will be explained later.

3.1. 2D feature extraction

Vanishing points can be extracted from a group of parallel lines in an image. Many vanishing point extraction methods have used the concept of a Gaussian sphere [6], which is one of global feature extraction methods such as the Hough transform. Gaussian sphere approaches, however, are applicable only for one simple structured building that has one or two surfaces, as the global approach does not consider the proximity of lines in the image and line clusters from multiple surfaces of buildings cannot be segmented.

We use a hybrid method of Gaussian sphere approach and image-based approach. From a ground view image like in Figure 3 (a), each line segment is transformed into a unit vector on Gaussian sphere to get the dominant line directions as shown in Figure 3 (b) and (c). We select the most accumulated cell on Gaussian sphere to get vanishing points. The extracted vanishing points are used to compute the rotation of the camera as described in Section 3.2.

Once vanishing points are extracted, we cluster image lines according to their directional proximity to the vanishing points as shown in Figure 3 (d). We segment building facades (See Figure 3(e)) and compute 2D line correspondence candidates as shown in Figure 3(f). The extracted 2D line candidates are used to compute the position of the ground view camera as described in 3.3. More details about how to extract the 2D features can be found in [13].

3.2. Estimation of camera rotation

To compute the rotation of the camera, we need three orthogonal vanishing points to recover the camera orientation and its focal length and principal points [5, 9]. However, in some cases, only non-orthogonal vanishing points, or only two orthogonal vanishing points can be found in the ground view image. We analyze case-by-case situations to infer three orthogonal vanishing points by using the known angle of 3D building roof corners. For more details, see [12].
3.3. Estimation of camera position

Given the external rotation matrix and two point correspondences, the exact position of the camera can be computed. A 3D to 2D line segment match gives two points correspondences, but is not reliable. Instead, we use three 3D to 2D line correspondences. As shown in Figure 4, one 3D to 2D line correspondence forms a plane passing through the center of projection. If three such planes are given, the position of the camera can be found by intersecting the three planes.

Given an image line and the center of projection, we derive a 3D plane whose normal is \( m \), and the relation with a 3D point, \( P \) projected onto the line is expressed by

\[
m R_{cw}(P - t) = 0
\]

(1)

where \( R_{cw} \) is a rotation between the world coordinate and the camera coordinate and \( t \) is the world position of the camera. With three correspondences between 2D lines and 3D points, we compute the world position of the camera using least squares solution of the following equation:

\[
\begin{bmatrix}
(R_{cw}m_1) \\
(R_{cw}m_2) \\
(R_{cw}m_3)
\end{bmatrix}
\begin{bmatrix}
t_x \\
t_y \\
t_z
\end{bmatrix}
= \begin{bmatrix}
(R_{cw}m_1)^T \cdot P_1 \\
(R_{cw}m_2)^T \cdot P_2 \\
(R_{cw}m_3)^T \cdot P_3
\end{bmatrix}
\]

(2)

where \( m_1, m_2 \), and \( m_3 \) are plane normal vectors, \((t_x, t_y, t_z)^T\) is the 3D camera position, and \( P_1, P_2, \) and \( P_3 \) are three 3D points from three 3D lines.

3.4. Generation and evaluation of hypotheses

Since there exist multiple combinations of three 3D to 2D line correspondences, a selection mechanism is required. In this paper, we use a hypotheses generation and evaluation approach to choose valid line correspondences.

To generate visible 3D line candidates under the unknown camera position, we first hypothesize a coarse position of the camera relative to the 3D building. We divide the location space into four areas according to an origin vector set among 3D building roof boundary lines as depicted in Figure 5. We generate visible 3D line candidates by assuming the position of the camera is located in one of four candidate areas. We collect all visible lines from each hypothesized position.

Before we pick line correspondences, we classify the visible 3D model lines and 2D image line candidates into two \((x, y)\) major directions so that we can keep the valid groups of 3D and 2D lines. We select three visible 3D lines from the 3D building model in each hypothesized direction and 2D ones from 2D line correspondence candidates obtained in Section 3.1. The position of the ground view camera is computed by intersecting the three planes formed by three 3D to 2D line correspondences by Equation (2).
To generate all possible hypotheses would require computation time of the order of \(O(l_3^3l_2^3)\), where \(l_3\) is the number of 3D line candidates and \(l_2\) is the number of 2D line candidates. To reduce this complexity, we follow an approach similar to RANSAC [14]. We randomly select a triple of line correspondence pairs and evaluate the resulting hypothesis.

Given the rotation of the camera (as described in Section 3.2.), we hypothesize a camera model by using a set of three line correspondence candidates, which are used to compute the position of a hypothesized camera. Under this hypothesized camera model, we project all visible 3D lines into the ground view image and compare with image features (the 2D line candidates as in Figure 3 (f)), which are obtained by using the method explained in Section 3.1. Then, we collect supporting image evidence for the projected lines and measure the coverage of the image evidence over the projected lines.

Supporting image evidence consists of line segments, \(l_i\), in the image that support a hypothesis. Each line segment \(l_i\) should satisfy the following criteria:

- \(l_i\) must form a small acute angle (of less than a threshold angle) with one of the sides of the hypothesis
- The perpendicular distance of the midpoint of \(l_i\) from one of the sides of the hypothesis should be small (less than threshold pixels)
- A major part (more than a half) of \(l_i\) should overlap with the closest side of the hypothesis

The contribution of a line \(l_i\) in the set of evidence lines to the score is weighted by the ratio of the length that overlaps with the nearest side of the hypothesis (in the image) to the perimeter of the roof. This automatically weights longer sides (and evidence supporting them) more than it does shorter sides. Figure 6 shows examples of supporting and non-supporting image evidences.

![Figure 6. Image supporting evidences.](image)

When the evaluation value of a hypothesized camera model exceeds a threshold value, we consider the corresponding line pairs to be valid line correspondences. Given the valid line correspondences, we use a least square error fitting method to compute the camera position using the intersection of the collected multiple planes.

4. Experimental Results and Discussion

Figure 7 shows 3D building models and facade textures obtained from aerial view image only. As shown in Figure 7 (b), wall facades have low resolution textures while roofs have high resolution textures. It suggests that we should integrate aerial view image as well as ground view image to support high resolution rendering application.

![Figure 7. 3D building models augmented with textures from aerial view only.](image)

We generate 26 3D line candidates from the 3D wire frame building models in Figure 7 and 31 2D line candidates (18 inliers and 13 outliers) from the ground view image as shown Figure 3 (f). Full combination of three sets of line correspondences requires \(3C_{806}(=31\times 26)\) (86,943,220) iterations to find the correct position of the camera. The following equation, from [15] can be used to estimate \(N\), the required number of samples in RANSAC approach:

\[
N = \log(1-p) / \log(1-(1-e)^s)
\]  

where \(s\) is the number of lines required to compute the position of the camera (in our approach, \(s = 3\)). \(e\) is the pro-
portion of the outliers (1 - 18/806 = 0.98). $p$ is the probability that at least one of random samples of $s$ points is free from outliers (we set this number 0.99). For our example, we need at most 575,643 iterations to find the correct pair of line correspondences with the probability of 0.99. This is a factor of 151 improvement over the complete search. In our tests, we find that the method converges in many fewer steps in the range of 20,000 iterations with a set of three line correspondences that has greater evaluation score than the set threshold.

Figure 8 shows a result for the pose estimation of the ground view image and an integrated result of aerial and ground view images for 3D buildings in a university campus. Some facade textures of the buildings are obtained from the calibrated ground image as well as even the aerial view image since the aerial image is oblique. Eaves of the roof of a building can cause a problem as the vertical lines of roof boundary for a building from an aerial image do not match with the vertical lines from a ground view image. This recess gap from the roof boundary needs to be adjusted in order to obtain correct facade textures from the ground view images. We plan to include this in future work.

Acknowledgements

This research was supported in part by U.S. Army STRICOM under contract N61339-02-C-0053.

5. References


