Image-Based 3D Semantic Modeling of Building Facade*

Jun Yang[1] and Zhongke Shi[2]

[1] School of Mechanics, Civil Engineering and Architecture
Northwestern Polytechnical University
Xi'an 710072, China
junyang@nwpu.edu.cn
[2] School of Automation, Northwestern Polytechnical University
Xi'an 710072, China
zkeshi@nwpu.edu.cn

Abstract. 3D building modeling has many potential uses in the fields of construction, city planning and public security. An image-based 3D semantic modeling method of building facade is proposed in this paper. Dense point clouds are generated from inputting images by structure from motion and cluster based multi-view-stereo algorithms. Planar components are extracted from generated point clouds by random sample consensus and further recognized as structural components based on prior knowledge. Windows are detected through a multi-layer complementary strategy with binary image processing techniques. Experimental results from two building facades verify the proposed method.

1 Introduction

Various society fields demand three-dimensional (3D) building models. In the Architecture, Engineering and Construction (AEC) industry, semantically rich 3D building models are increasingly used throughout a building's life cycle, from design, through construction, and into facility management phase [1]. These models are generally known as building information models (BIMs). Recently, automatic generation of as-built BIMs is a hot issue being studied extensively, which is used not only for as-built documentation [2], but also for construction progress monitoring [3][4]. In urban planning domain, it is helpful to adopt 3D building models since analyzing in 3D world is much more efficient than on 2D maps. For public security, accurate 3D building models are indispensable to make strategies during emergency situations. Other fields, such as automobile navigation and virtual tourism also benefit from realistic 3D building models [5].

Automatic 3D building modeling are based on remote sensing technologies, such as laser scanning [1][5][8][9], photogrammetry [3][14] or the combination

* This work was supported by National Natural Science Foundation of China (No.51208425) and Research Foundation of Northwestern Polytechnical University (No.JCY20130127).

L.J. Chmielewski et al. (Eds.): ICCVG 2014, LNCS 8671, pp. 661–671, 2014.

of these two [6][7]. Laser scanners have been widely applied for its precise generation of dense 3D point clouds. Tang et al [8] outlined a scheme from laser scanning generated point clouds to as-built BIM, composed of three core operations: geometric modeling, object recognition and object relationship modeling. Xiong et al [1] focused on the creation of semantically rich 3D models of building interior. The proposed algorithm extracted planar patches from input point cloud, learned to label patches as walls, ceilings or floors and performed an analysis on surface openings, such as windows and doorways. Special reasoning were used to deal with occlusions and holes. Pu and Vosselman [5] presented a knowledge based approach for reconstruction of building facade models using terrestrial laser scanning data. Segmented plane surfaces were classified into various semantic features based on generic knowledge and a polyhedron facade model containing both detailed geometry and sematic meaning were generated. Martinez et al [9] proposed a novel facade contour detection method by converting point cloud data into a profile distribution function and looking for distribution peaks and valleys.

With its prevalence, laser scanner has limitations in reality due to expensive and fragile equipments, lack of portability, need of skilled operators [12] and a long preparation time for setting up. In addition, laser scanners only provide Cartesian coordinate information of the scanned scene. These featureless data without semantic information is especially challenging for high level 3D modeling [3]. On the contrary, photogrammetry offers a lower cost, lower skill, portable solution with abundant features such as color and texture. Different to laser scanning, a computation flow involving camera calibration based on multiple view geometry and cross view feature points matching are required to obtain 3D point clouds in photogrammetry. Compared to the millimeter accuracy of laser scanner, photogrammetry can only achieve centimeter level [2]. References [12][13] evaluated the accuracy of image-based 3D modeling and verified the serviceability of photogrammetry in 3D modeling.

Unlike the extensive studies of 3D modeling by laser scanning, there are not so many progresses on photogrammetry side. Golparvar-Fard et al [3] solved the photographer's locations, orientations and a sparse 3D geometric representation of the as-built site using daily progress photographs and superimposed the reconstructed scene over as-planned 4D models for progress estimation. Even generating a sparse 3D point cloud, the study was not about 3D modeling but geo-registration of daily photographs. Son and Kim [4] proposed an efficient, automated 3D structural component recognition and modeling method using color and 3D data acquired from a stereo vision system. However the structural component under consideration is just a simple steel structure specifically designed for experiments. Kim et al [14] designed a framework consisting of 3D photogrammetric data acquisition, refinement and concrete detection for progress measurement of buildings under construction. The study accomplished dense reconstruction of 3D point cloud using a commercial system. However 3D building modeling was not involved.

How photogrammetry performs in realistic 3D building model is still unknown. In this paper, we mean to demonstrate that a monocular handheld camera is capable of densely reconstructing and modeling building facade. Instead of only generating sparse point cloud, we propose to move to a dense reconstruction level, which can supply more features for further analysis. As for the 3D semantic modeling, we present a knowledge based semantic reasoning strategy, together with a novel window detection method.

The remaining of the paper is organized as follows: Section 2 explains the dense reconstruction workflow. Section 3 elaborates on the 3D semantic modeling procedure. Starting from plane segmentation, semantic structural components are recognized based on prior-knowledge, followed by a novel window detection procedure. Experimental results and conclusions are given in section 4 and 5 accordingly.

2 Dense 3D Point Clouds Generation

Most previous studies on image-based building reconstruction only generated sparse 3D point cloud, which is far more than enough for higher level semantic modeling. The paper proposes to move beyond sparse point cloud to dense reconstruction. The inputs are the images of building facade captured by uncalibrated handheld cameras from multiple points of view. And the outputs are dense reconstructed 3D point clouds.

The first step of reconstruction is feature points detection and matching. Robust feature detector which can result in reasonable dense feature points and are not sensitive to point-of-view, illumination change, scale change, etc, need to be used. The most frequently used detector is scale-invariant feature transform (SIFT)[15] for its invariance to scale and rotation and robustness to small affine or projective deformations and illumination changes. Once detected, feature points are matched by measuring the distance between their SIFT descriptors, which describe the intensity gradient over an image window centered at a feature point. Since false matches always exist and will have negative influence on the following reconstruction, a RANSAC (RANdom SAmple Consensus) procedure is evoked to remove false matches.

Starting from the initial successful feature matching between two images, the camera extrinsic and intrinsic parameters for each image and the 3D coordinates of each feature point are solved incrementally by repeatedly adding matched images,triangulating feature matches and bundle-adjusting the structure and motion.

The preceding Structure From Motion (SFM) procedure outputs a sparse 3D point cloud by computing the 3D position associated to each match. A Cluster based Multi-View-Stereo (CMVS) algorithm [22] is followed for dense point cloud generation. It decomposes the set of input images into clusters that have small overlap. Once clustered, a multi-view-stereo algorithm is applied to reconstruct dense 3D points. The resulting reconstructions are merged into a single dense point-based model.

3 3D Semantic Modeling

The geometry of a reconstruction 3D model can be described with boundary representation (B-Rep), constructive solid geometry (CSG) and spatial enumeration. Spatial enumeration decomposes the 3D space into a set of identical cells. CSG represents the target as a combination of certain fixed primitives using Boolean operators. B-Rep describe the 3D world by connected surface elements. Considering the nature of building facades, B-Rep is the most suitable method. B-Rep models are usually composed of two parts: topology and geometry (surfaces, curves and points). We model the building facade as several connected surfaces under certain topology. Three semantic components: wall, window and protrusion are considered in our model. The 3D reconstruction from images is first segmented into several planes. Then these planes are further recognized as semantic components based on prior knowledge.

3.1 Segmentation

Common buildings are usually polyhedrons consisted of multiple planes. Though other shape primitives, such as spheres or cylinders, may also exist in modern buildings, they are out of the paper's scope currently. The first step of 3D modeling is to segment planes from 3D reconstruction result. Three widely used segmentation methods are RANdom SAmple Consensus (RANSAC), Hough Transform and region growing. Hough transform is a voting scheme that extracts a parameterized shape primitive from a discretized parameter space. A primitive with a large number of parameters results in a high-dimensional discretized parameter domain, which causes memory issues. So the main application area of Hough transform remains in 2D. Region growing in 3D space usually takes normals at different points as features and starts growing at an initial seed region. So the selection of seed region has a strong impact on the segmentation result, especially for noisy data [11]. RANSAC is a robust estimator of parametric model even for data containing a high degree of noise and outliers, which serves our purpose best.

The principle of RANSAC is to search the best plane fitting the 3D data. It selects randomly three points and calculates the parameters of the corresponding plane. Then it detects all points in the original data belonging to the calculated plane, according to a given threshold. Afterwards, it repeats these procedures N times. In each iteration, it compares the obtained result with the last saved one. If the new result is better, then it replaces the saved result by the new one [19].

The basic RANSAC algorithm assumes that only one model can be fit to the data. In order to detect all the planar components in the 3D reconstruction result, an iterative strategy is applied. First, RANSAC is applied on the data returning the plane with the most inliers (referred as dominant plane). Then all inliers for this plane are removed the data. RANSAC is performed on the residue to find the next largest plane. The process terminates when no plane with a sufficient number of points can be found.

3.2 Semantic Components Recognition

Segmentation finds out planar components of building facades. The semantic roles of these components are to be recognized based on prior knowledge. A common building facade is usually composed of walls, windows and protrusions. Wall is the largest plane perpendicular to the ground. (Notice that ground plane is assumed as a user defined information in our proposed scheme. Naturally, we set up the coordinates system following the right hand rule, with plane XY being the ground plane and the dominant plane in point clouds parallel to plane YZ.) Windows are usually small rectangle planes embedded in the wall surface. Other structures, such as balconies or friezes, are a little outside the wall, are all sorted as protrusions.

It can be seen from the above description that the semantic components of the building facade can be recognized based on their geometric features. These features are listed as below:

Area: The area of a plane is a dominant feature in the semantic reasoning. Wall plane usually has the biggest area. A threshold can be set up to filter out too tiny planes for noise depression.

Position: The location of the plane, its orientation and its connection relationship with other components are also important for recognition. We use the centroid of the plane's convex hull for its location, the plane normal for its orientation and the intersection line of two planes as judgement of connection relationship.

Shape: As aforementioned, windows are usually in the shape of rectangle.

Due to light reflection, point cloud in the glass window area is usually much sparser compared to other parts. So the point cloud density is also a significant feature for semantic reasoning, especially for window detection, which will be illustrated later.

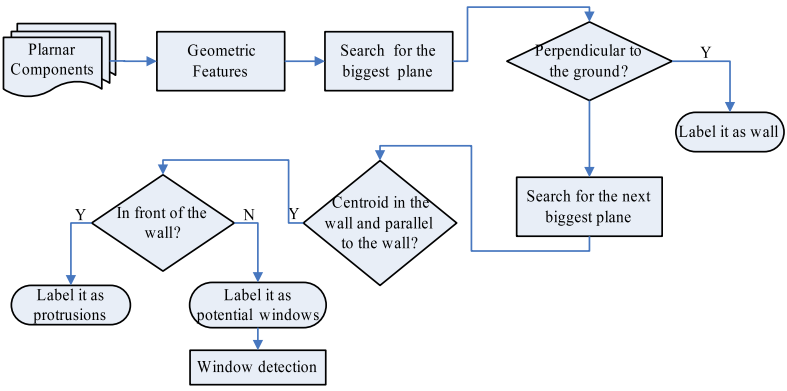


Fig. 1. Flowchart of semantic component recognition

In general, the semantic component recognition procedure is concluded in Fig. 1 . We follow a coarse to fine, bottom to up searching strategy, starting from the dominant wall plane and finally going to the windows and protrusions.

3.3 Window Detection

As a distinguishable component in building facade, window detection has been discussed several times in the literature. Most studies, no matter laser scanning-based or image-based, have used the low point cloud density in window area as a dominant feature. The rectangle shape, with two horizontal edges and two vertical edges, is often regarded as another accompanying feature for window detection.

Xiong et al [1] regarded windows as rectangular-shaped openings in laser scanned point cloud and used a support vector machine (SVM) classifier to detect partially occluded windows. Pu and Vosselman [5] extracted boundary points around holes in point cloud, grouped them and fitted a minimum bounding rectangle as detected window. Martinez et al [9] used the 3D point cloud density to distinguish points into multiple layers which corresponded to wall, windows, etc. Radopoulou et al [20] tested a depth-encoded Hough voting for window detection. Bohm et al [7] detected windows based on edge in laser scanning point cloud and integrated images for detail recovery, such as window crossbars. Bauer et al [21] applied a sweep based method to scan the density change in point cloud line by line and searched for windows. Machine learning based methods [1][20] require a labor-intensive training stage. Boundary or edge based methods [5][7] are sensitive to noise. Sweep based method [21] is inefficient.

Based on the previous semantic reasoning procedure, we have the following basic observations. Windows (intrusions) and protrusions both result in holes in the point clouds of the wall plane. On the other hand, though the glass itself does not appear as feature rich area in images due to reflection and hence corresponds to very sparse even none point cloud, windows may still relate to dense enough point cloud when they are covered by curtains or crossbars. So if the 3D points belonging to windows are detectable, they will serve as complementary information for window detection. Meanwhile, protrusions (if existing) will also be corrections for the hole detection result from the wall plane. Assuming holes in the wall plane, windows and protrusions are all rectangle shaped area in the 3D point clouds, we propose a multi-layer based window detection strategy as follows.

First, detect rectangle holes in the wall layer. The detected holes set is marked as Φ. Second, if existing a window layer (which is to say, the windows are detectable in the point cloud), detect the connected regions in rectangle shape and mark the detection result as Ω_1. Third, if existing a protrusion layer, do the same as the second step and ends up the set Ω_2. Both Ω_1 and Φ confirm the existence of windows and their information can be complementary. And Ω_2 will be the correction since it indicates not all the holes in wall layer are introduced by windows. So the final window detection result is:

$$Windows = \Phi \cup \Omega_1 - \Omega_2$$

For both hole detection in the wall layer or connected region detection in other layers, we propose an uniform method taking advantages of binary image processing. First, the point clouds are converted into a binary image through the following procedures. All the y, z coordinates of the 3D points are amplified by a certain factor, centralized by subtracting mean values of y, z accordingly, shifted by adding minimum values, and finally rounded up. Consequently, the converted y, z coordinates correspond to pixels with value 1 in the binary image. And the rest area is marked by value 0. A binary image is generated by now. Preprocessing steps include morphological open operation for small connected components removal and holes filling up. Then all objects in the binary image are traced with two properties measured. One is the *area*, which is the actual number of pixels in the object region; the other is the *extent*, which is defined by the ratio of pixels in the region to pixels in the total bounding box. If $area >= T_1$ and $extent >= T_2$, then the region is marked as a candidate window. T_1 is given by users according to experiences. A too small area can not be a window. In our experiment, it is around 100 pixels. T_2 is a key threshold to indicate the shape of the region. Technically, when $extent = 1$, the region is an exact rectangle. Considering the imperfection of the realistic situation, we usually set T_2 loosely, e.g. around 0.8. Notice all the above mentioned procedure is targeted at locating connect white regions in binary image. So for operation in the wall layer which has windows as holes (black area), its binary image needs to be inverted beforehand.

4 Experimental Results

Two groups of experiments have been set up for validation of the proposed scheme. Targets to be reconstructed are the teaching building and the office building on campus. The device used to capture images is a Canon IXUS 950 IS digital camera, which is a portable low-end product. The selection of the device is because in reality site inspectors usually prefer portable devices. The proposed system lays no restriction on the quality of images and means to take daily log photos as input. The only requirement is that the camera parameters, such as focal length, image resolution and zoom have to remain fixed during capturing. And at least a 50% overlap between continues images must be assured for stable feature matching. Notice that in the experiment we only focus on the reconstruction of a single building facade. A more complete reconstruction of the entire building will be explored in the future work.

4.1 The Teaching Building

The teaching building has embedded windows without any protrusions. As shown in Fig.2(a), 37 images are captured in resolution 3648 x 2736. The generated dense point clouds are shown in Fig.2(b). The plane segmentation results in two planes, which are further analyzed by their geometry features and recognized as wall layer and window layer. As in Fig.3, the magenta color represents the wall

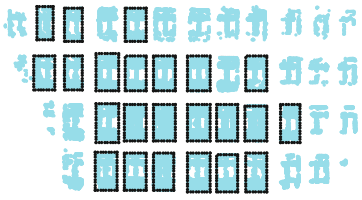
layer and the cyan color represents the window layer. Windows are detected on each layer separately as shown in Fig.3(a),3(b). After the combination procedure, windows are finally unified as in Fig.3(c). We can see that most windows are correctly located. A few windows on the point clouds edge are not stably detected because of the low point density there.

(a) Snapshots of the teaching building.

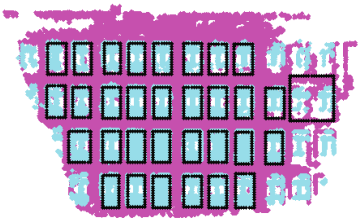
(b) Dense point clouds of the teaching building.

Fig. 2. The snapshots and point clouds of the teaching building

(a) Windows detected in the wall layer.



(b) Windows detected in the window layer.

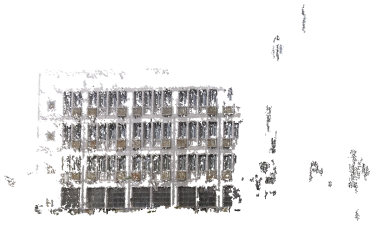


(c) Final windows after combination of results from all layers.

(d) legend

Fig. 3. The structural components recognition and window detection result

(a) Snapshots of the office building.



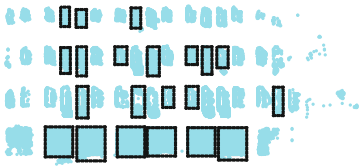
(b) Dense point clouds of the office building.

Fig. 4. The snapshots and point clouds of the office building

4.2 The Office Building

The office building is a regular four stories building. It has embedded windows and protrusions serving as air conditioners holders. As shown in Fig.4(a), we took 45 images in resolution 2048 x 1536. The generated dense point clouds are shown in Fig.4(b). Applying plane segmentation on the dense point clouds, results in three planes. These planes are further recognized as wall layer, window layer and protrusion layer. As in Fig.5, the magenta color represents the wall layer,

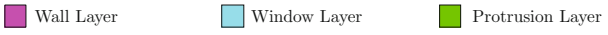
(a) Windows detected in the wall layer.



(b) Windows detected in the window layer.

(c) Windows detected in the protrusion layer.

(d) Final windows after combination of results from all layers.



(e) legend

Fig. 5. Structural components recognition and window detection results for the office building

the cyan color represents the window layer and the green color represents the protrusion layer. After segmentation, we detect windows on each layer separately as shown in Fig.5(a),5(b),5(c). It can be seen that the detection result is not very satisfying in a single layer. But after the combination procedure as shown in Fig.5(d), nearly all the windows are correctly located and those protrusion areas are successfully removed.

From the above experimental results, we can see that the proposed scheme works successfully for various scenarios. It can generates a 3D sematic model by segmenting the 3D point clouds and recognizing structural components correctly. The proposed multi-layer based window detection method can stably locate windows and reject the potential ambiguous protrusions. The undetected windows are mainly from low density area of the point cloud.

5 Conclusions

This paper proposed an image-based 3D semantic modeling scheme of building facade. 3D point clouds were generated from handheld camera captured images using SFM and CMVS. For semantic modeling, planar components were extracted from generated point cloud by Random Sample Consensus. Knowledge-based strategy was applied for semantic structural components recognition. The building facade was modelled as the combination of wall, windows and protrusions. Windows are detected through a multi-layer complementary strategy with binary image processing techniques. Experiment on two real building facades demonstrated the efficiency of the proposed scheme. Future work seeks to explore the color and texture features of 3D points clouds for higher level semantic modeling, e.g. building material recognition. A complete modeling of the entire building may also be included.

References

1. Xiong, X., Adan, A., Akinci, B., Huber, D.: Automatic creation of semantically rich 3D building models from laser scanner data. Automation in Construction 31, 325–337 (2013)
2. Klein, L., Li, N., Becerik-Gerber, B.: Imaged-based verification of as-built documentation of operational buildings. Automation in Construction 21, 161–171 (2012)
3. Goldparvar-Fard, M., Peña-Mora, F., Savarese, S.: D4AR-a 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication. Journal of Information Technology in Construction 14, 129–153 (2009)
4. Son, H., Kim, C.: 3D structural component recognition and modeling method using color and 3D data for construction progress monitoring. Automation in Construction 19, 844–854 (2010)
5. Pu, S., Vosselman, G.: Knowledge based reconstruction of building models from terrestrial laser scanning data. ISPRS Journal of Photogrammetry and Remote Sensing 64, 575–584 (2009)

6. Brilakis, I., Lourakis, M., Sacks, R., Savarese, S., Christodoulou, S., Teizer, J., Makhmalbaf, A.: Toward automated generation of parametric BIMs based on hybrid video and laser scanning data. Advanced Engineering Informatics 24, 456–465 (2010)
7. Bohm, J., Becker, S., Haala, N.: Model refinement by integrated processing of laser scanning and photogrammetry. In: Proceedings of 2nd International Workshop on 3D Virtual Reconstruction and Visualization of Complex Architectures (3D-Arch) (2007)
8. Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A.: Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. Automation in construction 19, 829–843 (2010)
9. Martinez, J., Soria-Medina, A., Arias, P., Buffara-Antunes, A.F.: Automatic processing of Terrestrial Laser Scanning data of building facades. Automation in Construction 22, 298–305 (2012)
10. Gonzalvez, P.R., Aguilera, D.G., Lahoz, J.G.: From point cloud to surface: Modeling structures in laser scanner point clouds. In: ISPRS Workshop on Laser Scanning, pp. 338–344 (2007)
11. Schnabel, R., Wessel, R., Wahl, R., Klein, R.: Shape recognition in 3d point-clouds. In: Proc. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision, vol. 2 (2008)
12. Bhatla, A., Choe, S.Y., Fierro, O., Leite, F.: Evaluation of accuracy of as-built 3D modeling from photos taken by handheld digital cameras. Automation in construction 28, 116–127 (2012)
13. Golparvar-Fard, M., Bohn, J., Teizer, J., Savarese, S., Pena-Mora, F.: Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. Automation in Construction 20, 1143–1155 (2011)
14. Kim, C., Son, H., Kim, C.: The effective acquisition and processing of 3D photogrammetric data from digital photogrammetry for construction progress measurement. In: ASCE International Workshop on Computing in Civil Engineering, pp. 178–185 (2011)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
16. Wu, C.: Towards Linear-time Incremental Structure from Motion
17. Labatut, P., Pons, J.-P., Keriven, R.: Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In: IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
18. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3121–3128 (2011)
19. Yang, M.Y., Förstner, W.: Plane detection in point cloud data. In: Proceedings of the 2nd Int. Conf. on machine control guidance, vol. 1, pp. 95–104 (2010)
20. Radopoulou, S.C., Sun, M., Dai, F., Brilakis, I., Savarese, S.: Testing of Depth-Encoded Hough Voting for Infrastructure Object Detection. In: ASCE International Workshop on Computing in Civil Engineering, pp. 309–316 (2012)
21. Bauer, J., Karner, K., Schindler, K., Klaus, A., Zach, C.: Segmentation of building models from dense 3D point-clouds. In: Proc. 27th Workshop of the Austrian Association for Pattern Recognition, pp. 253–258 (2003)
22. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1434–1441 (2010)