Multimed Tools Appl (2018) 77:27387–27404
https://doi.org/10.1007/s11042-018-5926-4

Interactive 3D building modeling method using panoramic image sequences and digital map

Hyungki Kim[1] · Soonhung Han[2]

Received: 8 August 2017 / Revised: 8 February 2018 / Accepted: 21 March 2018 /
Published online: 30 March 2018

Abstract This paper proposes a method of generating 3D building models with precise geospatial information and a photograph-based façade appearance from panoramic image sequences and digital maps. 3D building modeling research is actively being conducted in areas such as geographic information systems, virtual reality, and augmented reality. However, the generation of realistic 3D models from a ground-level viewpoint is still extremely costly in terms of labor of modeling experts, and collection of data. We have developed a method for 3D building modeling with high-resolution photograph-based appearance information using panoramic images captured at ground level with a mobile mapping system, and geospatial information obtained from a digital map. The proposed method includes 1) pre-processing for tilt correction and base 3D model generation, 2) geo-registration of panoramic images with minimal user input, and 3) building height and shape estimation. This paper presents the proposed method and the quantitative performance measure obtained from a developed test modeling system. In addition, modeling results from an experimental dataset are also presented.

Keywords 3D building modeling · Street view · Geo-registration · Shape estimation

1 Introduction

Recently, interest in 3D city models with geospatial information has been increasing. As a subset of the virtual globe or digital earth, a 3D city model represents the physical world in

✉ Hyungki Kim
diskhkme@gmail.com

Soonhung Han
shhan@kaist.ac.kr

[1] 3rd Aero Systems Division, ADD, Yuseong-gu, Daejeon 34168, South Korea

[2] Graduate Program of Ocean Systems, Department of Mechanical Engineering, KAIST, Yuseong-gu, Daejeon 34141, South Korea

digital form, and is a useful resource in 3D geographic information systems (GIS) [18], simulations with virtual reality techniques, and augmented reality with location-based services. Google Earth [7] is a well-known mirror world application that provides satellite/aerial imagery, and map information. Currently, Google is attempting to minimize the gap between Google Earth and the physical world, by developing 3D city models and 3D tree models.

The amount of similarity between a 3D model and the physical world depends on the levels of detail it has. For instance, CityGML [14] defines five levels-of-detail (LoDs) for 3D city models, and Yoo [30] classifies various 3D urban models from 2D maps and digital orthography to full volumetric computer aided-design (CAD) models based on geometrical details. The ultimate virtual globe should feature 3D models with varying levels of detail. However, 3D models with high detail are difficult to generate because of the high cost of data collection and processing. In practice, significant labor is required [17] to generate 3D models with high detail.

Recently, 3D city modeling approaches that use images collected at ground level as a texture—that is, for the appearance of the 3D models—have been gaining attention as another aspect of providing highly detailed 3D models. Constructing 3D models from ground-level images can provide higher resolution information compared to satellite/aerial imagery, which narrows the gap between the physical and digital worlds. For instance, 3D building models with high-resolution texture can give visual information including texts on signboards, or the locations of entrances—details rarely provided by 3D building models generated from satellite/aerial imagery. Depending on the application, lack of visual information could seriously hinder their usability [25].

The most easily accessible and widespread source of ground-level images is street view services, such as Google Street View, which is generally integrated with a map service. Providers of a street view service usually collect these ground-level images, as well as various sensor data, using a mobile mapping system (MMS), which mounts remote sensing devices and processors onto vehicles. However, because of the wide baseline of image sequences in current street view services, the amount of information on an individual building is limited. Consequently, automatic 3D reconstruction is difficult to achieve, solely using image data.

In this paper, we propose a method that uses wide-baseline panoramic image sequences from MMS with digital maps to generate 3D building models. For efficient modeling, we define the minimal user input required for geo-registration of panoramic images. Based on the geo-registration result, the shapes of buildings are estimated, and appearance information is captured from the images for a full 3D building model. For validation, we test our method using a dataset similar to that of current street view services.

The remainder of this paper is organized as follows. Section 2 presents researches conducted on 3D building modeling. Section 3.1 describes the overall process of our proposed 3D building modeling method. Sections 3.2 and 3.3 outline the pre-processing and geo-registration operations from the user-provided constraint, and image analysis with height estimation, respectively. Section 3.4 describes the shape estimation of buildings using adjacent images in detail. Section 3.5 illustrates a modeling and rendering method that uses the estimated information. Section 4 validates the proposed model: Section 4.1 shows the modeling system developed using the proposed method, and Section 4.2 presents modeling results and analysis. Finally, Section 5 concludes, and outlines plans for future works.

2 Related research

3D building or city modeling using images can be divided into satellite/aerial imagery-based approaches, and ground-level image-based approaches. Satellite/aerial imagery-based approaches have the advantage of efficiency, since they contain large amounts of information about buildings in a single image [2, 30–32]. Yoo [30] proposed a method of modeling and rendering 3D buildings from aerial orthoimages using a bi-layered displacement mapping technique, as well as a fusing approach using digital elevation map (DEM) data of 3D buildings. Zebedin [31, 32] proposed an approach that generates dense reconstructions with a large number of images for roof modeling. Baillard and Zisserman [2] argued that plane-based reconstruction is possible with a small number of images using a line-matching algorithm. However, as stated above, satellite/aerial imagery does not contain sufficient ground-level information; therefore, it is difficult to generate 3D building models with high-quality visual information. We focused on the 3D building modeling method that uses ground-level images to efficiently generate 3D building models of a large area with a high-resolution, photograph-based appearance.

Ground-level, image-based 3D building or city modeling research is actively being conducted. Image-based rendering approaches [3, 10] generally estimate 3D models using semantic information obtained from its pixel values, or geometrical relations from a single image. However, the modeling results cannot be directly adopted in applications such as GIS because the result model is not a complete 3D building model. Recent 3D modeling approaches frequently use large amounts of image data for 3D buildings [15], or cities [29]. They focus on generating high-quality 3D models with detailed geometry from dense reconstruction results, which can be obtained from estimated camera poses using stereo-matching algorithms. However, widely available image data, such as the street view service, has a limitation in the quantity of available information per building because of a wide baseline, which refers to the long distance between consecutive images. As authors of [1] addressed, the matching of wide-baseline panoramic images is a challenging problem. Therefore, the applicability of dense image-based reconstruction is still limited, as large quantities of new data have to be obtained from the physical world.

Research is also being conducted on image-based 3D modeling from a relatively small number of images. This approach compensates for the lack of data by relying on user input [5, 24], assumptions about the shape of the objects [28, 34], or the use of external data [11, 13, 19, 22]. Approaches using user input enable estimation, or the assignment of primitives or planes from the user, for 3D modeling. The detail of the resulting model is higher if the user provides more input during the modeling process; however, the time cost also increases. In addition, placing a 3D model onto a virtual globe platform induces further workload. Alternatively, approaches that use external data assume the existence of 3D CAD models [11, 19, 22], or geospatial information [7]. However, obtaining the 3D CAD model is difficult because existing models are very limited. Therefore, to the best of our knowledge, this approach is currently not practical for 3D city modeling. To obtain geospatial and geometrical information for 3D buildings, we propose a method of using a digital map with panoramic images. Unlike 3D CAD models of buildings, a digital map is widely available and consistently maintained by government agencies. A digital map provides the precise location of various artificial objects in the physical world as vector data using a geospatial coordinate system. In the case of buildings, 2D projected footprint vertices exist in the digital map. We propose a method of estimating the 3D geometry of buildings from a digital map and panoramic images, with minimal user intervention.

3 Proposed method

We propose an efficient 3D building modeling method that utilizes wide-baseline panoramic image sequences and digital maps to give a high-resolution, photograph-based appearance from a ground-level viewpoint. Our approach intends to overcome the limitation of completeness in 3D geometry reconstruction in wide-baseline panoramic image sequences—in current street view services, about 10 m image intervals—captured from MMS using information from digital maps. The approach requires a correlated analysis of geospatial information in digital maps and panoramic images captured by MMS. The error in the image location and direction from MMS is usually larger than the error in the digital map, and therefore should be corrected. To the best of our knowledge, fully automated geo-registration of wide-baseline panoramic image sequences has not been completely solved.

Therefore, we developed an efficient 3D building modeling method that uses user-provided constraints, along with information from image analysis and digital maps. In our assumptions about building shapes, we adopt a quasi-Manhattan world model in which buildings are described with walls perpendicular to the ground plane, and the complexity is that of the generally prismatic shapes seen in digital maps. With these assumptions, and minimal user-provided constraints, we can correct the error of geospatial location and direction in panoramic images. Moreover, building height information that is not contained in digital maps can be estimated. In addition, we show that detailed estimation of building geometry is possible by matching adjacent panoramic images. The resulting 3D building model has the same geospatial accuracy as the digital map, and delivers rich visual information to the user by using textures obtained from ground-level images.

3.1 Overall process

The overall modeling process we propose is illustrated in Fig. 1. The process is divided into the four major steps listed below; key parts of the process are steps 2 and 3. The resulting 3D building model can be generated using a base 3D model, a geo-registered image, a consensus building mask, and the metric height of the building obtained in the process.

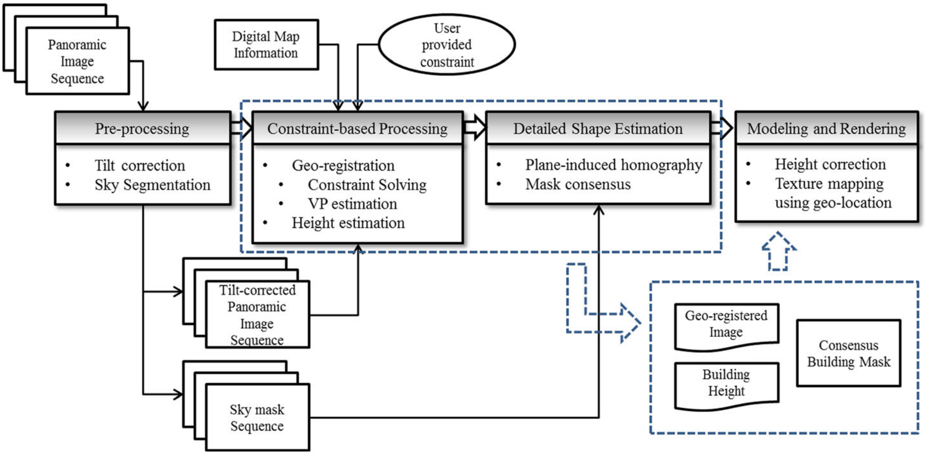

Fig. 1 Overall modeling process

(1) Pre-processing
(2) Constraint-based processing
(3) Detailed shape estimation
(4) Modeling and rendering

For source data, we use a panoramic image sequence and the footprint information of buildings in the digital map. Each panoramic image has corresponding latitude and longitude information, and the geospatial direction of the vehicle-when obtaining image. A building's footprint information in a digital map is a composite set of lines, defined by latitude and longitude of start and end points. We obtain tilt-corrected panoramic image sequences, and sky mask sequences from the tilt correction and sky segmentation pre-processes, respectively. In constraint-based processing, errors in location and direction in each panoramic image are corrected —called geo-registration in this paper—by solving constraints provided by the user, and a vanishing point (VP) estimated from the panoramic image. Based on the corrected location and direction, with line segments contributing to the VP, building height can be estimated. Detailed shapes can subsequently be estimated using matching information on adjacent panoramic images and sky masks. Finally, the information obtained is combined, and location-based texturing is applied for rapid visualization of the resulting model.

3.2 Pre-processing

Tilt correction and sky segmentation are carried out in the pre-processing stage. Both processes can be performed globally on the entire image without any prior knowledge about a scene.

Tilt correction is required to analyze a panoramic image and a digital map in the same degrees of freedom. As stated previously, MMS collects latitude, longitude, and direction data from sensor devices mounted on a vehicle. However, in real situations, because the vehicle is moving in 3D space, pitch and roll angles exist in the orientation of the vehicle, and the captured image reflects 6 degrees of freedom. Accordingly, a vertical edge in the 3D world can be projected as a curve in the panoramic image because of the pitch and roll angle. Therefore, like Ventura and Hollerer [26], we conducted tilt correction by estimating the pitch and roll angle using vertical lines in the upright panoramic image, and applying an inverse rotation matrix to the pixels of the entire image [12]. For lost pixel information according to the image rotation, we applied a nearest-neighbor pixel interpolation method. In the case of our dataset, the estimated pitch and roll angle is $\pm 4^{\circ}$ in most images. The result is illustrated in Fig. 2, where distortions of the vertical edges according to the pitch and roll angle are removed.

In sky segmentation, image pixels are classified as sky, or otherwise, via image processing. Subsequently, the information obtained is used in the detailed shape estimation process. In the upright panoramic image, we can assume that the sky is positioned in the upper part of the panoramic images, so a simple clustering algorithm can be adopted for sky segmentation. First, we apply mean shift segmentation [4] in the YUV color space and determine the dominant color in the upper part of the image—the sky. Then, additional small segments are merged if the clustered color is similar to that of the sky. As a result, we can segment the image into a sky region and others, as illustrated in Fig. 3, and store the result as a binary mask.

3.3 Constraint-based processing

In constraint-based processing, the geospatial location and direction of the image is recovered using a user-provided constraint and image analysis; in addition, the height of the building is

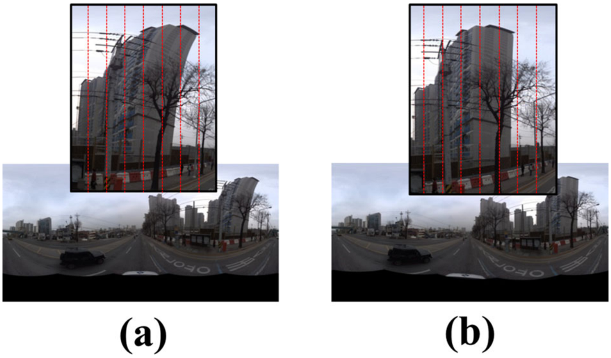

Fig. 2 **a** Source panoramic image; **b** tilt corrected panoramic image

estimated. In geo-registration, the estimated VP information is used, and in height estimation, line segments that contribute to the VP are used. In the following sections, it is assumed that the panoramic images are produced by equirectangular projection, and the pixel resolutions of the resulting images are *W* and *H*, representing width and height, respectively.

Based on the principal point $\boldsymbol{u_0} = (u_0, v_0)$, each pixel in an image coordinate can be uniquely converted to an azimuth angle θ, the polar angle ϕ, that corresponds to a unit vector $\overrightarrow{V} = (x, y, z)^T$ in a camera coordinate. For a pixel $\boldsymbol{u} = (u, v)$ in an image coordinate, Eq. (1) presents the conversion equation for θ, ϕ, x, y, z. Figure 4 shows the relationship between the image coordinate and camera coordinate of an equirectangular projected image. If the position of a camera in world coordinates is known, we can define the start and end points of a view ray in world coordinates that crosses a pixel.

$$\begin{gathered}
\theta^{img} = \frac{(u - u_0)}{W} \times 2\pi \\
\phi^{img} = \frac{(v - v_0)}{H} \times \pi \\
x = \cos\left(\phi^{img}\right)\cos\left(\theta^{img}\right) \\
y = \sin\left(\phi^{img}\right) \\
z = \cos\left(\phi^{img}\right)\sin\left(\theta^{img}\right)
\end{gathered} \tag{1}$$

3.3.1 Geo-registration method

Geo-registration is required to align the panoramic image with the external data—in our case, the digital map. The GPS/INS sensor data collected with the panoramic image describes the geospatial location and direction of the camera. In general, our dataset, and similar research papers, show that noise from sensors produces an error amounting to ±10 m in location, and ±5° in direction. This error is

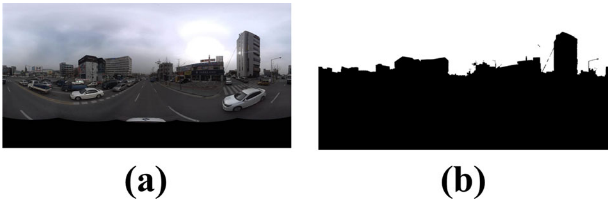

Fig. 3 **a** Source panoramic image; **b** binary mask resulting from sky segmentation

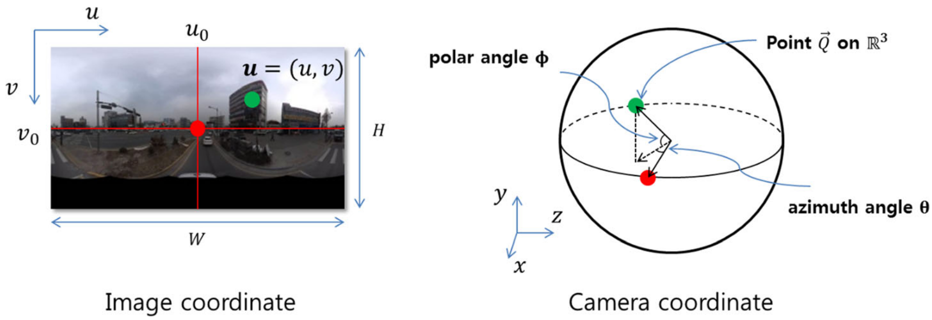

Fig. 4 Relationship between pixel position and unit vector in image and camera coordinate

the primary factor that induces mismatch between building information in the digital map, and the panoramic image. A case of mismatch is illustrated in Fig. 5a. Without geo-registration, image geometry is misaligned—as shown in Fig. 5b. Therefore, geo-registration should be carried out first, in order to generate a valid 3D building model.

We first define a constraint that is required to solve the geo-registration problem with minimal user input. Let the location of a panoramic image be **P** and the direction of the panoramic image be $\overrightarrow{\mathbf{D}}$ in a 3D geospatial coordinate, where the ground plane is defined as XZ-plane. If we represent an image as two-dimensional arrays, the location of a pixel can be described by (*u, v*), where *u* is the left-to-right index, and *v* is the top-to-bottom index of a pixel. By Eq. (1), the *u* index of a pixel represents a unique azimuth angle $\boldsymbol{\theta}^{img}$ relative to the image direction $\overrightarrow{\mathbf{D}}$. Therefore, if we know the image direction $\overrightarrow{\mathbf{D}}$ in a geospatial coordinate, we can convert *u* to the geospatial azimuth angle *θ*. Let the coordinates of the footprint vertices of a building be $\mathbf{F} = \{\mathbf{F_1}, \mathbf{F_2}, \ldots, \mathbf{F_p}\}$ in the XZ-plane. Without loss of generality, if two pixel left-to-right indices u_1 and u_3 correspond to $\mathbf{F_1}$ and $\mathbf{F_3}$, the geometrical relation of the image and digital map can be viewed as illustrated in Fig. 6, where the matching relation is described in Eqs. (2)–(3):

$$\mathrm{R_y}(\theta_{\mathrm{vp}})\overrightarrow{\mathrm{D}} = \mathrm{R_y}(\pm 90^{\circ})(\mathrm{F_n} - \mathrm{F_{n+1}}), \tag{2}$$

where $\mathrm{R_y}$ is the rotation matrix along the y-axis,

$$\angle \mathrm{F_1PF_3} = \theta_3 - \theta_1 \text{ and} \angle \left(\mathrm{P} + \overrightarrow{\mathrm{D}}\right)\mathrm{PF_1} = \theta_1. \tag{3}$$

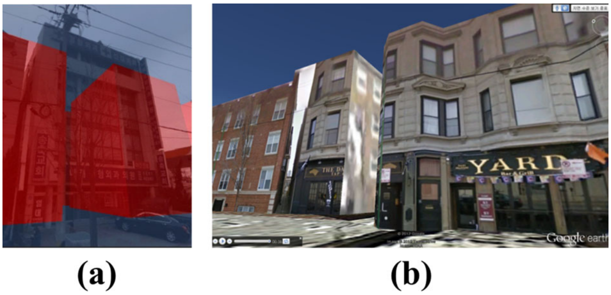

Fig. 5 **a** Mismatch between image location and digital map information; **b** example of misaligned model in Google Earth [7]

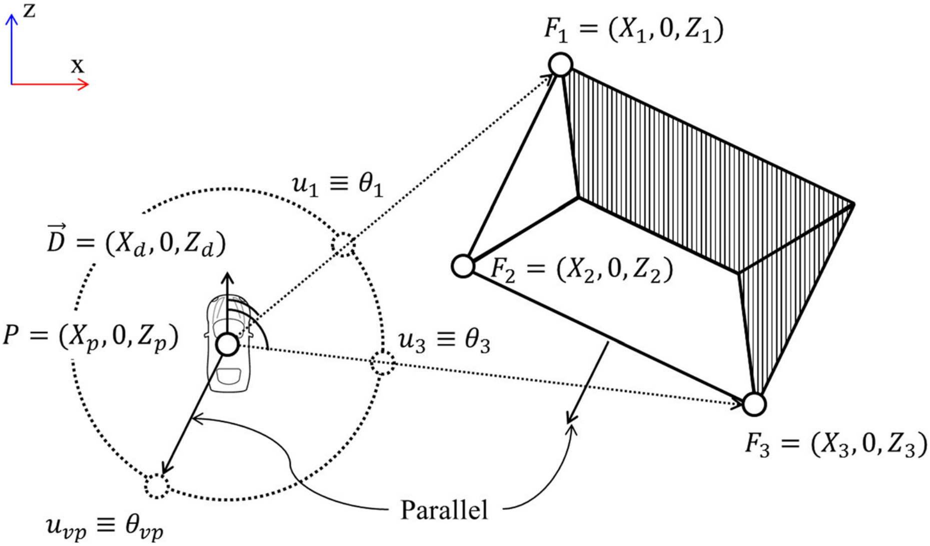

Fig. 6 Geometrical relation of image and digital map

Equation (2) implies that the normal of a wall estimated from the VP should coincide with the normal of a wall calculated from the digital map. The VP is a convergence point of the parallel lines in 3D space when projected onto the image plane, and the coordinate of a VP can be used to calculate the relative angle between the camera, and those lines in 3D. First, we detect the line segments from an image region between u_1 and u_3 in the panoramic image using LSD [6], and classify the vertical and horizontal line segments using the slopes of the line segments. The VP can be estimated using the RANSAC algorithm on the horizontal line segments—the formulation is described in Oh and Jung's research [20] for the panoramic image. If the estimated pixel left-to-right index is u_{vp} then, along with image direction $\overrightarrow{\mathbf{D}}$, the normal of a wall on the 3D geospatial coordinate can be estimated. The estimated normal of a wall should be matched with one of the walls that exist between $\mathbf{F_1}$ and $\mathbf{F_3}$. With the known angle error tolerance $\pm 5^{\circ}$, we can find two footprint vertices between $\mathbf{F_1}$ and $\mathbf{F_3}$ that correspond to the normal of a wall calculated from the VP. Therefore, the $\overrightarrow{\mathbf{D}}$ vector can be calculated from two points corresponding to u_1 and u_3 with $\mathbf{F_1}$ and $\mathbf{F_3}$. Eq. (3) defines the image location $\mathbf{P}$ from θ_1 and θ_3, which is uniquely defined with input u_1 and u_3. If $\overrightarrow{\mathbf{D}}$ is fixed, $\mathbf{P}$ can be calculated simply using trigonometry.

Figure 7 illustrates the geo-registration result. In the figure, footprint information is shown using procedurally generated base 3D geometry, colored as green—which will be described in Section 3.5. In Fig. 7a, the long solid and dotted line shows the left-to-right index of a pixel that the user provides, while the short solid and dotted line shows the selected footprint vertices. Note that even though the line appears to have an angle in 3D view, every line is perpendicular to the ground plane in 3D space. Figure 7b shows the geo-registration result.

3.3.2 Height estimation method

Height information is not widespread in an ordinary digital map. Estimation methods using satellite/aerial imagery are also being studied, but there are less studies conducted from

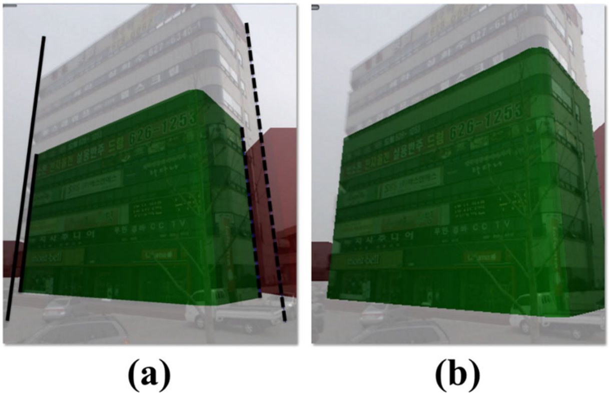

Fig. 7 **a** Two-point matching between image pixel and footprint vertices; **b** geo-registration result

ground-level images. We developed a method of height estimation by using line segments obtained from the VP estimation process.

As described in Section 3.3.1, line segments can be classified as vertical, or horizontal. Horizontal line segments can be further classified into line segments that contribute to the VP, and otherwise. Generally, a large portion of the line segments contributing to the VP are detected from the building walls. In height estimation, we cluster line segments that are parallel, and have similar height in 3D space, to find characteristic lines—if the contribution length of those lines is longer than the tolerance, proportional to the length of u_1 and u_3. Through this process, as illustrated in Fig. 8, we can find characteristic lines, such as Fig. 8b, from the line segments contributing to the VP, as shown in Fig. 8a. We assume that the highest characteristic line represents the height of a building. The metrical calculation of building height is possible using the relation shown in Fig. 9, where 2.5 m is the known height of the

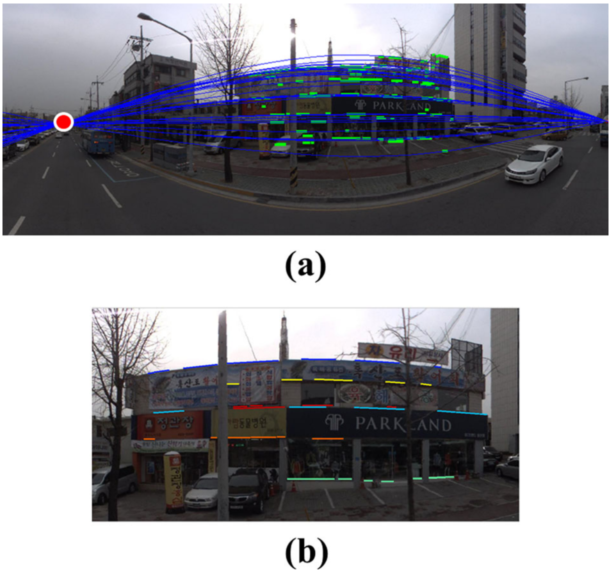

Fig. 8 **a** Horizontal line segment detected (green lines) and estimated vanishing point (red dot) with great circle arc of horizontal line segments (blue lines); **b** detected characteristic lines

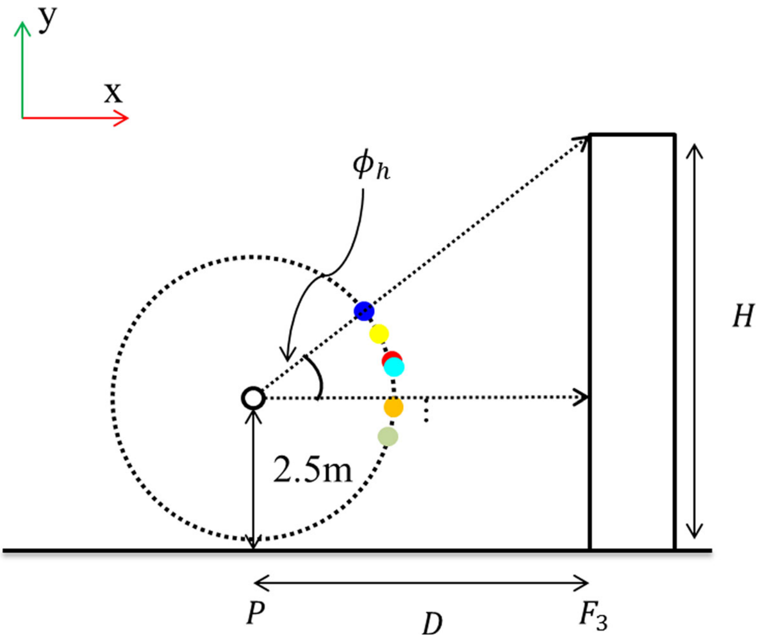

Fig. 9 Relation between height of a building and image location with characteristic lines

camera focal point, starting from the ground plane. Because we recovered the geospatial location of the camera, and the geospatial location of the target building wall, the metric height can be calculated using Eq. (4). Note that ϕ_h is calculated from the v coordinate of the characteristic line using Eq. (1).

$$H = 2.5 + D \times \tan(\phi_h) \text{ where } D = \|P - F_3\| \tag{4}$$

The estimated height can be used to generate and align 3D building geometry with the panoramic image. However, there are some limitations at this point. The following typical problem cases exist:

(1) Characteristic lines are detected above the building; e.g., electric cables.
(2) No characteristic lines are detected from the building because the upper part of the building is not a straight line.

In case (1), the metrical height is calculated higher than the real building; consequently, the pixels from the sky are also contained in the result model, which is a false positive. In case (2), we may lose some building pixel information, which is a false negative. Therefore, more processing is required using additional information.

3.4 Detailed shape estimation

We developed a detailed shape estimation method using adjacent panoramic images to handle both of the problem cases described above. An image from a single viewpoint is not robust to noise and loss of information. Hence, the solution to this problem is to use an additional image. However, as mentioned in Section 3, a panoramic image sequence has intervals of about 10 m; therefore, limitations exist in the amount of data obtainable for a single building, which makes stereo-matching-based 3D reconstruction difficult.

Our approach adopts plane-induced homography, which can be found frequently in buildings following the quasi-Manhattan world model. Plane-induced homography refers to the relationship between two images in the 3D plane that allows image-to-image transformation through four-point correspondence [9]. The process of matching and consensus of adjacent sky masks from plane-induced homography enables recovery of lost information, or removal of false positives.. By using homography, rectification or registration is possible if we can be sure of the existence of a 3D plane. Because we can assure the existence of a 3D plane—that is, a building wall—plane-induced homography is useful.

First, assume that the user provides a constraint on the *n*-th image for constraint-based processing. Subsequently, we can define the region of interest (ROI), which is the image region of the same building in the (*n-1*)-th and (*n* + *1*)-th image, using the geospatial location of the images. Because geo-registration is not applied in the (*n-1*)-th and the (*n* + *1*)-th image, the initially estimated ROI from the footprint information is erroneous. Therefore, we first find a point feature match in both image pairs using the SIFT [16] algorithm, and find the homography matrix that has the most inliers in the RANSAC process. We can subsequently map u_1 and u_3 to the (*n-1*)-th and (*n* + *1*)-th image using the homography matrix obtained. An image is selected as a pair of the *n*-th image if the image has more uniformly distributed point feature matches than the other, in the corrected ROI. As a result, we can estimate the homography matrix with the corrected ROI for the target building in the adjacent panoramic image, as illustrated in Fig. 10.

The next step is mask consensus, that defines the consensus region of the building wall belonging to the same building in each mask image. The consensus region is obtained via the application of image-to-image transformations to sky masks by the homography matrix. Through the consensus, false positive and false negative errors in the height estimation process

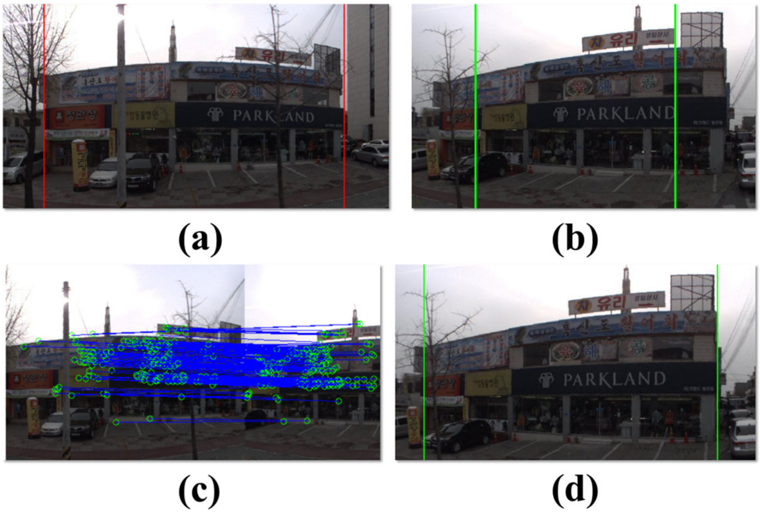

Fig. 10 **a** User-provided ROI in the *n*-th image; **b** estimated ROI in the (*n-1*)-th image before correction; **c** point feature matching result; **d** corrected ROI of target building in the (*n-1*)-th image

are compensated. An example of a result mask is illustrated in Fig. 11, which is stored as a binary mask. In Fig. 10, we can classify the building pixels that are estimated with the height estimation process (purple), and detailed shape estimation process (green).

3.5 Modeling and rendering

To generate 3D building models using the obtained information, we created a base 3D geometry from the footprint information in the digital map. As we described in Section 3.3.1, we can acquire the geospatial coordinates of vertices that describe the footprints of buildings. The base 3D geometry is generated by procedurally constructing a 3D mesh in altitude direction from a 2D profile that is a closed loop of footprint vertices. Note that when generating the base 3D geometry, the actual height of each building does not exist, so the height is identical for every building. As illustrated in Fig. 12, from digital map, the base 3D geometry can be generated while maintaining the geospatial locations.

Final modeling and rendering can be done by editing the base 3D geometry using the geo-registered panoramic image, result building mask, and estimated height. The geometry of the base 3D model is modified and texture-mapped using the obtained information to deliver a 3D building model with high-resolution, photograph-based texture. When the user provides the constraint, a panoramic image is geo-registered, and the height of the target building is estimated in metric units; thus, the height value is applied to the base 3D geometry first. Texture mapping is the process of assigning a color from the geo-registered panoramic image to the 3D building model. We developed a shader for texture mapping that calculates the vector from the panoramic image to the texture coordinate of the 3D model—based on the geospatial location—and assigns the image pixel color for each texture coordinate.

The result mask from the detailed shape estimation process is applied in a similar manner as the image-based rendering method, without modification of geometry. The shader looks up the binary mask and renders a transparent color, if a part is detected as the sky—white-colored part in Fig. 11. Thus, the 3D model is visualized only for the part that is classified as a building.

4 Experiments

In the experiments conducted, we developed a modeling system based on our proposed method, and performed 3D building modeling. The source panoramic image sequence was captured from MMS between latitudes and longitudes, [36.358933,127.432165] and [36.339823,127.436540]. Each panoramic image has a resolution 5400 × 2700 in width and height, with omnidirectional pixel information projected to a single image using equirectangular projection. The corresponding geospatial location and direction information

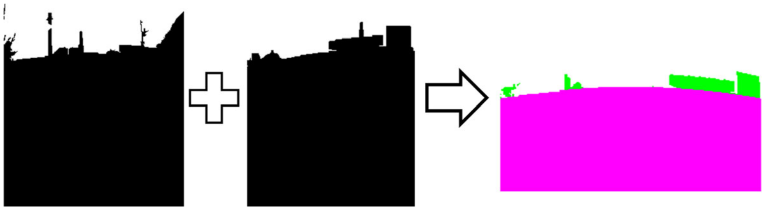

Fig. 11 Binary mask within ROI of the *n*-th and (*n-1*)-th images, and result mask for target building.

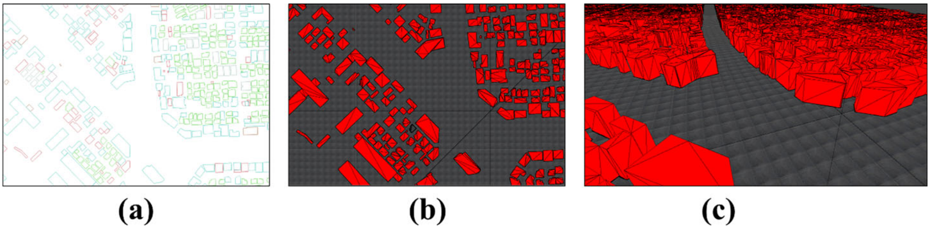

Fig. 12 **a** Source building footprint information in digital map; **b** generated base 3D geometry from top view; **c** from isometric view

was collected from a GPS/INS sensor, and the mean interval of the panoramic image sequence was 9.2462 m, based on the recorded values of the GPS sensor. We found that the baseline length is similar to that of current street view services—usually 10 m, according to related research. The scale of the digital map was 1:5000, known to have up to 1 m horizontal error.

4.1 Developed modeling system

The modeling system is designed to allow simultaneous and coherent observation of the digital map, 3D model, and panoramic image in the modeling process. The application view is divided into the model screen for the panoramic image and the 3D model, and the map screen for digital map information. The user can select the target building and two footprint vertices of a building in the map screen, while two left-to-right pixel indices corresponding to the footprint vertices are selected in the model screen via mouse input. The selection statuses are synchronized so the user can clearly recognize which building and footprint vertices are selected in both screens. Moreover, a camera pose indicator helps localize the current field of view of the model screen in geospatial context. The implemented application screen is illustrated in Fig. 13. Unity3D 4.5 was used as the development environment.

Average selection time is measured from three users after a brief introduction about the modeling system. Each user conducted 10 trials of selecting two target building vertices and two corresponding left-to-right pixel indices. Selection time starts when the user selects a target

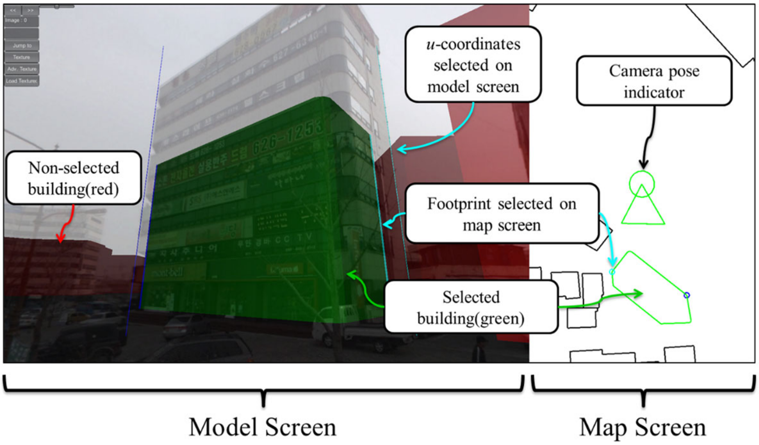
u-coordinates selected on model screen
Camera pose indicator
Non-selected building(red)
Footprint selected on map screen
Selected building(green)
Model Screen
Map Screen

Fig. 13 Implemented application screen

building to generate a 3D model in the map screen and ends when the finish button is clicked after all required input has been made. The measured average selection time is 8.14 s. After the user clicks the finish button, the modeling process begins. The processing time for modeling was measured: Geo-registration and height estimation takes an average of 0.43 s for 27 out of 30 trials; hence, the user can see the result almost immediately. If a detailed shape estimation is not required, the modeling process is finished at this stage. Otherwise, it takes an average of 8.80 s for 3 out of 30 trials for the detailed shape estimation process. Most of the time is spent projecting pixels identified as a "non-sky" into the mask image. Eventually, the average modeling time for an individual building is 17.27 s.

4.2 Modeling result and evaluation

From 28 panoramic image sequences, 31 building models were generated using the developed modeling system. Representative modeling results are displayed in Fig. 12. Every 3D building is modeled by four mouse inputs—two for the footprint vertices, and two for the image left-to-right pixel indices—and the resulting building model contains geospatial location and direction information, with accuracy identical to that of the digital map. Through detailed shape estimation, a 3D building model can be generated even when the upper part of the building is not shaped as a straight line, as illustrated in Fig. 14. The high-resolution visual information in the 3D model is provided by images captured at ground level. As illustrated in Fig. 15, the difference between the 3D building model using aerial imagery in the Vworld [27] system, and our proposed system, is clearly visible. The high-resolution appearance information is one of the essential factors for the presence and immersiveness of the virtual environment. Quantitatively, widely available satellite/aerial imagery has a resolution between 2 m and 50 cm, whereas our result model has a resolution between 6 and 1 cm. This is possible since the distances from the location of panoramic image and the modeled buildings are under 50 m, and

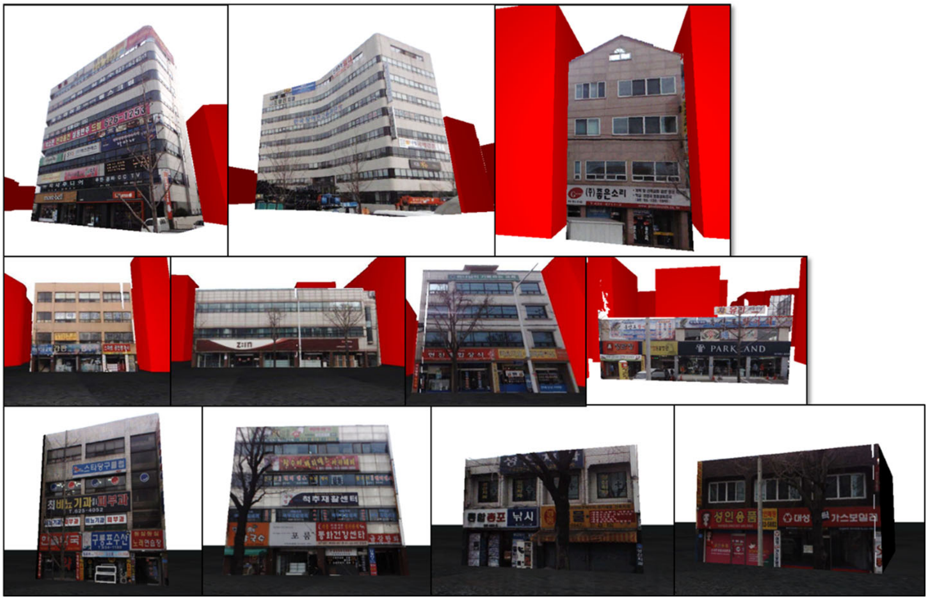

Fig. 14 Representative 3D building models generated using proposed method

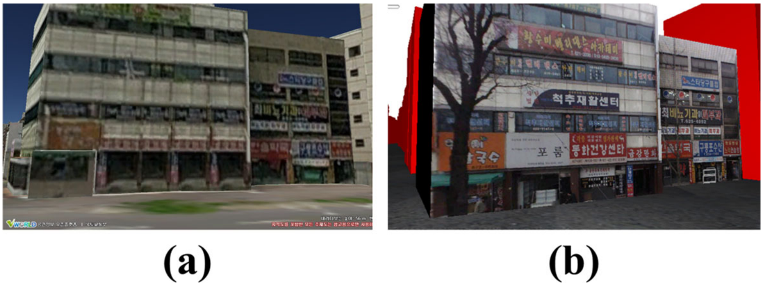

Fig. 15 **a** Existing 3D building model in the Vworld system; **b** resulting 3D building model using our proposed method

a 5400 × 2700 resolution image has one pixel of information for every 0.15°. This means that the resolution can be at least eight times higher, resulting in readable text on all signboards.

The geometrical shape is a complete prismatic mesh model without defects and noise, because it is generated using procedural modeling from footprint vertices. Existing 3D reconstruction methods based on stereo matching require extensive post-processing to obtain a resulting 3D model without defects and noise. These defects and noise can cause serious problems, typically in collision detection, which is usually mandatory for virtual reality applications, or simulations; therefore, the proposed model is also suitable for these applications. From the perspective of geometrical detail, traditional paper maps with 1:5000 scales have up to 2.5 m detail, whereas a digital map usually represents data under a centimeter unit in practice—comparatively, much more descriptive detail. In our experimental data, we discovered that the maximum geometrical resolution is up to 20 cm.

There is a limitation on the detailed shape estimation results due to our adoption of the image-based rendering approach for detailed shapes, instead of editing geometry. Intensive user input in can resolve this limitation, but we plan to extend our method to become a fully-automated process. We therefore decided to keep user input requirements to a minimum. Further, the method cannot guarantee the completeness of the modeling result if the shapes of buildings do not follow the quasi-Manhattan world model. In particular, further processing is required when there is more than one major building wall in an image of a building.

5 Conclusion

A 3D building modeling method that uses a panoramic image sequence and digital map was proposed. Through the proposed method, prismatic 3D building models with a photograph-based appearance can be generated by few user input commands for matching two left-to-right pixel indices in an image to build footprint vertices in the digital map. Geo-registration of designated and consecutive panoramic images is enabled by the defined relationship between building footprint information in the digital map, and panoramic images. In addition, height and detailed shape are estimated by characteristic line detection and plane-induced homography. The resulting 3D building model has rich visual information at the ground level that allows it to be highly realistic, and georeferenced.

The main contribution of this paper is the proposal and verification of a 3D modeling technique using wide-baseline panoramic images and digital maps. The resulting 3D building models contain geographically accurate information and have the advantage of a high-resolution appearance. Conventional aerial/satellite image-based techniques can be used for wide area modeling, but there are disadvantages—such as low resolution. Alternatively, the ground image-based reconstruction approach requires dense images per single object reconstruction, thus there are disadvantages in the limitations of the application area. In this paper, we proposed a method to solve these problems using the widely available panoramic image, street view image, and by addressing both the lack of geometric information inherent in wide-baseline, and registration problems caused by sensor noise.

For the issues addressed in the use of wide-baseline panoramic image sequences and digital maps for 3D building models—specifically, sensor error corresponding to the image location and direction—the method of geo-registration using minimal user input was proposed. To enable the method, we first argued that tilt correction is required to analyze the panoramic image and digital map data in the same 3D geospatial coordinate. We showed that VP estimation with four-point matching could solve the geo-registration problem and automate height estimation. To address the varying shapes of the upper parts of buildings, we suggested the image-based rendering method with detailed shape estimation using sky segmentation and two-view matching. However, further research is required for buildings with more complex shapes.

In the future, we plan to fully automate our method by adopting global geo-registration for wide-baseline panoramic images, such as [23], to remove significant errors before the modeling process; or applying semantic segmentation of urban scenes [33] for automatic contour detection, instead of using a user hint. In addition, for better quality of detailed shapes, application of a primitive fitting approach, such as [21], as a pre- or post-processing operation in the detailed shape estimation process, may be valid. Lastly, pattern recognition of building façades for detailed feature modeling can be directly combined with our framework for detailed 3D building models. We expect that our proposed method can be applied to update 3D geo-databases [8] with ground-level visual information.

References

1. Alexakis E, Tsironis V, Petsa E, Karras G (2016) Automatic adjustment of wide-base google street view panoramas. Remote Sens Spat Inf Sci XLI-B1:639–645
2. Baillard C, Zisserman A (1999) Automatic reconstruction of piecewise planar models from multiple views. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins 2, 559–565
3. Barinova O, Konushin V, Yakubenko A, Lee K, Lim H, Konushin A (2008) Fast automatic single-view 3-d reconstruction of urban scenes. In: ECCV 2008, 12-18 October, Marseille, France: Springer, 100–113
4. Comaniciu D, Meer P (2002) Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619
5. Debevec PE, Taylor CJ, Malik J (1996) Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 11–20
6. Gioi RG, Jakubowicz J, Morel J, Randall G (2012) LSD: a line segment detector. Image Process On Line 2:35–55
7. Google earth (2014) Available from: https://earth.google.com/. Accessed 2 June 2016
8. Guo H, Li X, Wang W, Lv Z, Wu C, Xu W (2016) An event-driven dynamic updating method for 3D geo-databases. Geo-spatial Inf Sci 19(2):140–147
9. Hartley R, Zisserman A (2004) Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge
10. Hoiem D, Efros AA, Hebert M (2005) Automatic photo pop-up. ACM Trans Graph 24(3):577–584
11. Kada M, Klinec D, Haala N (2005) Facade texturing for rendering 3D city models. In: ASPRS Conference 2005, 78–85

12. Kim H, Han S (2016) Geo-registration of wide-baseline panoramic image sequences using a digital map reference. Multimed Tools Appl 1–19
13. Kim H, Kang Y, Han S (2014) Automatic 3D City Modeling Using a Digital Map and Panoramic Images from a Mobile Mapping System. Math Probl Eng 1–10
14. Kolbe T, Gröger G, Plümer L (2005) CityGML – interoperable access to 3D city models. In: Proc. of the 1st International Symposium on Geo-information for Disaster Management, 21-23 March Netherlands: Springer, 883–899
15. Li X, Wu C, Zach C, Lazebnik S, Frahm J (2008) Modeling and recognition of landmark image collections using iconic scene graphs. In: ECCV 2008, 12-18 October, Marseille, France: Springer, 427–440
16. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
17. Lu Z, Guerrero P, Mitra NJ, Steed A (2016) Open3D: crowd-sourced distributed curation of city models. In Proceedings of the 21st International Conference on Web3D Technology, 87–94
18. Lv Z, Li X, Zhang B, Wang W, Zhu Y, Hu J, Feng S (2016) Managing big city information based on WebVRGIS. IEEE Access 4:407–415
19. Moslah O, Klee M, Grolleau A, Guitteny V, Couvet S, Philipp-Foliguet S (2008) Urban models texturing from un-calibrated photographs. In: IVCNZ, 26-28 November, Christchurch, New Zealand: IEEE, 1–6
20. Oh SH, Jung SK (2012) RANSAC-based Orthogonal Vanishing Point Estimation in the Equirectangular Images. J Korea MultimedSoc 15(12):1230–1441
21. Park J (2005) Interactive 3D reconstruction from multiple images: A primitive-based approach. Pattern Recogn Lett 26(16):2558–2571
22. Rau J, Teo T, Chen L, Tsai F, Hsiao K, Hsu W (2006) Integration of GPS, GIS and photogrammetry for texture mapping in photo-realistic city modeling. In PSIVT, 10-13 December, Hsinchu, Taiwan: Springer, 1283–1292
23. Sato T, Pajdla T, Yokoya N (2011) Epipolar geometry estimation for wide-baseline omnidirectional street view images. In: ICCV 2011, 6-13 November, Varcelona, Spain, 56–63
24. Sinha SN, Steedly D, Szeliski R, Agrawala M, Pollefeys M (2008) Interactive 3D architectural modeling from unordered photo collections. ACM Trans Graph 27(5):159 1–159:10
25. Tsai F, Lin H-C (2007) Polygon-based texture mapping for cyber city 3D building models. Int J Geogr Inf Sci 21(9):965–981
26. Ventura J, Hollerer T (2013) Structure and motion in urban environments using upright panoramas. Virtual Reality 17(2):147–156
27. Vworld (2014) Available from: www.vworld.kr/. Accessed 8 Aug 2017
28. Wang G, Tsui H, Hu Z (2005) Reconstruction of structured scenes from two uncalibrated images. Pattern Recogn Lett 26(2):207–220
29. Xiao J, Fang T, Zhao P, Lhuillier M, Quan L (2009) Image-based street-side city modelling. ACM Trans Graph 28(5):114 1–114:12
30. Yoo B (2013) Rapid three-dimensional urban model production using bilayered displacement mapping. Int J Geogr Inf Sci 27(1):24–46
31. Zebedin L, Klaus A, Gruber-Geymayer B, Karner K (2006) Towards 3D map generation from digital aerial images. ISPRS J Photogramm Remote Sens 60(6):413–427
32. Zebedin L, Bauer J, Karner K, Bischof H (2008) Fusion of feature- and area-based information for urban buildings modeling from aerial imagery, In: ECCV 2008, 12-18 October, Marseille, France: Springer, 873–886
33. Zhang C, Wang L, Yang R (2010) Semantic Segmentation of Urban Scenes Using Dense Depth Maps. EVVC 2010:708–721
34. Zheng X, Zhang X, Guo P (2011) Building modeling from a single image applied in urban reconstruction. In: Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry, Hongkong, Chana: ACM, 225–234

Hyungki Kim is a senior researcher in the 3rd aero systems division at Agency for Defense Development (ADD), Daejeon, Korea. He received his B.S. degree from the Department of Mathematical Science at KAIST in 2009 and M.S. degree from the Department of Mechanical Engineering at KAIST in 2011 and the Ph.D. degree in Department of Mechanical Engineering at KAIST in 2015. His current research interests include 3D urban modeling, computer vison, computer graphics, computer-aided design and real-time visualization.

Soonhung Han is a professor of Mechanical Engineering of the School of Mechanical, Aerospace & Systems Engineering of KAIST (www.kaist.edu). He is leading the Intelligent CAD laboratory (http://icad.kaist.ac.kr) of KAIST, and the STEP community of Korea (www.kstep.or.kr). His research interests include STEP (ISO standard for the exchange of product model data), VR for engineering design, and knowledge-based design systems. His domain of interests include shipbuilding and automotive. More information can be found from his personal web page at http://icad.kaist.ac.kr/~shhan.