Large Scale Urban Modeling [1] [2]

Vaibhav Bajpai
Jacobs University Bremen gGmbH
School of Engineering and Science
Campus Ring 1, D-28759 Bremen, Germany
{v.bajpai@jacobs-university.de}

Abstract. The complexity of application demands and technology challenges have kept urban modeling an active research area for a long time. However recently large-scale 3D reconstruction with detailed textures and geometry of urban surfaces has become the focus of attention. To this end, Christian et al. present a system that uses a vehicle equipped with one camera and two 2D laser scanners. The camera captures images for textures, the horizontal laser scanner tracks the truck's motion, and the vertical scanner captures 3D building facade data. Similarly, Huijing and Ryosuke use a vehicle-borne sensor system with three single-row laser scanners and six line cameras. A navigation system fuses data from the GPS, inertial, and odometer sensors to synchronize the laser and image data and track the system position. Such vehicle-based scanner systems captures richly detailed building facade data catering to the need of the industry today.

1 Introduction

The demand for 3D models of urban environments has been growing at a rapid pace. A recent survey on city models revealed that 95 percent of the participants were most interested in 3D building data [3] emphasising the massive focus-shift from the use of aerial based imageries *alone* [4] towards exploiting ground-based survey techniques for a detailed large-scale 3D reconstruction of urban environments.

3D navigation systems have been major proponents for such a push. Many navigation systems use 2D maps and are unable to provide users with enough information on real world objects. Features and navigational landmarks are not easily available even though they are very useful and helpful for navigating in the real world. People recognize landmarks and find routes in the cities easier by using 3D models than using a symbolic 2D map which makes 3D visualization more relevant for (3D) navigation and due to the high visual correspondence between map objects and real world objects, it increases the navigational value of the 3D map. Similarly, 3D GIS systems requiring high resolution 3D city models to help visualize existing state of cities and to perform simulations of district development plan and 3D model projection systems for the restoration of historic buildings and experiencing them in an augmented virtualized environment have been also been gaining a lot of traction lately.

2 State of the Art

A variety of approaches currently exists on creating 3D models of urban environments. One of the approaches, photogrammetry is a cost-effective means of obtaining large-scale urban models using 2D images without a priori 3D data. Devebec's Facade system is a successful terrestrial system that lets users recover basic geometric models of the photographed scenes [5] as shown in figure 1. Facade exploits user-selected edge features to build a model from 3D primitives and then verifies its accuracy by projecting the model back into the original image. The system uses view-dependent texture mapping to render photorealistic novel views. Pollefeys proposed techniques to automatically recover 3D geometry using handheld video cameras [6]. Although these techniques extract useful building facade models from ground-based imagery, they don't scale well to large areas.

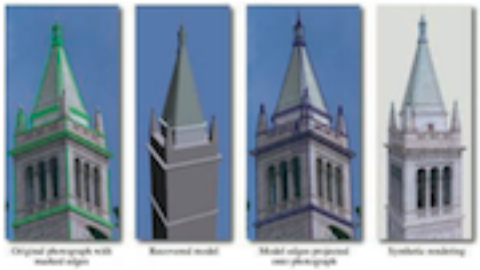

Figure1: Devebec's Facade System [5]

Figure2: MIT City Scanning Project [7]

Several efforts attempt to extract 3D models from panoramic images. Shum et al., for example, use an interactive system to extract 3D models from panoramic mosaics [8]. The MIT City Scanning Project has gathered calibrated (or pose) image data sets for a portion of the MIT campus and use the calibrated spherical mosaic images to reconstruct building models with textures [7] as shown in figure 2. Although, panoramic images are a convenient and economical data source, but integrating them with other data and scaling to large areas is difficult.

Extensive research has examined the use of single and multiple aerial images in urban modeling. Lin et al. used monocular aerial images from a general viewpoint to detect urban buildings and construct 3D shape descriptions of them [9]. Vanden and Frank use geometric and projective constraints to extract 3D models from a single image [10]. Noronha et al. used multiple aerial images and stereo algorithms to extract 3D models [11]. Aerial images often lack facade information, resulting in models with no visual realism. Integrating facade data is usually a manual process and requires additional sensor data.

Airborne Lidar technology combined with a GPS for aerial positioning provides accurate 3D information for structure roofs and most opaque surfaces greatly simpliflying large-scale urban modeling. You et al. developed an interactive system that models a variety of irregular building shapes [12]. The system resamples raw Lidar data into a regular grid, producing a mesh model through hole-filling and triangulation. Such a complex model construction however is semiautomatic, requiring a fair amount of operator intervention and resulting in painfully slow evolution of wide area models.

3 Approach

The reconstruction of urban environments using aerial imageries significantly lacks details of 3D objects such as facades of buildings which is becoming a prerequisites in the application scenarios today. The reconstruction efforts up until now have also been targeted towards small-scale and simple objects which do not fit well into the commercialized requirements of the industry. In addition, the previous efforts required significant manual intervention and the mobile scanning equipments were not exactly mobile but designed to capture in a stop and go fashion [13]. These processes not only resulted in high costs inhibiting broad use of the models, but also made it impossible to use them for applications where the goal is to monitor changes over time.

The papers propose an experimental setup that is capable of quickly scanning 3D and texture data while driving as shown in figure 3. The data is acquired continuously rather than in a stop-and-go fashion and is processed offline without any manual intervention to seamlessly obtain textured 3D models of the urban environments. Using a digital camera[1] or a number of line cameras[2] synchronized with other sensors a photorealistic view of the scene is obtained that is mapped on the geometrical models. Models of complicated and unexpected object geometries are reconstructed using a novel range-point classification technique that makes the effort vastly scalable.

Figure3: Approach [1]

In [1] in order to rapidly acquire 3D and texture data, fast 2D laser scanners and a digital camera were used instead of a 3D laser scanner. Since, a GPS was not used, the pose estimation problem was solved using an additional path computation step. To this end, Section 5 and Section 6 describe scan matching and path computation algorithms. Section 9 shows the results with future work in 10

In [2] in order to avoid occlusions at 45 degrees from using just two 2D laser scanners, three single-row laser range scanners and six line cameras were used. The system was equipped with a GPS, INS, Odometer based navigation system which solved the pose estimation problem and circumvented the need to perform path computation. A geometric surface model is then constructed using range point classification of point cloud data geo-referenced in world coordinate system and then texture mapped by projecting line images onto the surface model (also geo-referenced in world coordinate system). To this end, Section 7 describes the range point classification algorithms with texture mapping in 8. The results and future work follow in 9 and 10 respectively.

4 System Overview

Figure4: Truck with a Rack [1]

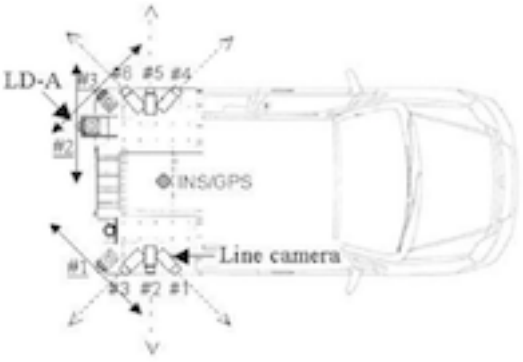

Figure5: Sensor System [2]

In [1] the hardware setup consists of 2 fast 2D laser scanners which are mounted with their scanning planes 90 degrees apart, a digital camera, a heading sensor and a true ground speed sensor mounted on top of a truck as show in figure 4. The 2D laser scanners are facing the same size of the street. The vertical laser scanner allows to acquire the complete geometry by capturing the shape of building facades as shown in figure 6. The digital camera synchronized with the other sensors is used to acquire the texture of the buildings to give it a photorealistic view and the horizontal laser scanner is used to accurately determine the position and orientation of the truck while in motion for localized pose estimation as shown in figure 7. The heading sensor is used to determine the orientation of motion.

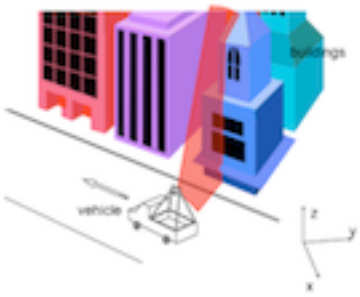

Figure6: Vertical Scanner [1]

Figure7: Horizontal Scanner [1]

In [2] the setup consists of three laser range scanners for measuring the object geometry along-with six line cameras to capture the object texture. The three range scanners are adjusted at 45 degrees to overcome occlusions experienced in [1]. The line cameras replace the digital camera of [1] and have a 180 degree field of view. The measurement vehicle is also equipped with a GPS/INS/Odometer based navigation system missing in [1] avoiding the additional path computation step. The navigation data is fused to each line image and range scans through linear interpolation using sensor's local clock. The sensor system on the car is shown in figure 5.

5 Scan Matching

The authors's approach to registration is based on solving point-to-line correspondences between two successive scans which is similar to [14] which solves the point-to-point least-squares problem and [15] which defines a linematch algorithm to extracts line segments from a scan and matches them with an a priori map of line segments. The lines of reference scans are extracted and matched to either the lines or points of the second scan. The block diagram of complete scan matching process is shown in figure 8

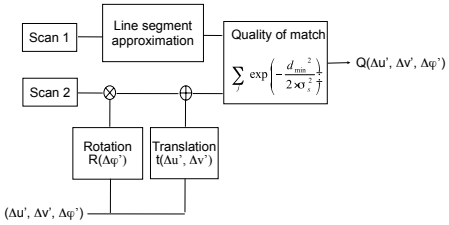

Figure8: Scan Matching Block Diagram [1]

The successive points in the reference scan are connected to form line, only if the difference between their depth values does not exceed some depth dependent threshold. This is a departure from a least-square fit method to find a straight line over multiple lines. However since the matching is done in an urban environment which consists of many curvilinear objects, the method rightly fits into the perspective. The line segment approximations of the reference scan can be used as a map to register the two successive scan. The two scans are registered by maximizing a quality function $Q = (\Delta u, \Delta v, \Delta\Phi)$ for a given displacement $\Delta u, \Delta v$ and rotation $\Delta\Phi$ using the following steps -

Step 1: Given a translation vector $\vec{t} = (\Delta u, \Delta v)$ and a 2x2 rotation matrix $R(\Delta\Phi)$, transform the points p_j to points p'_j according to equation 1

$$\vec{p'}_\mathrm{j}(\Delta u, \Delta v, \Delta\Phi) = R(\Delta\Phi).\vec{p}_\mathrm{j} + \vec{t} \tag{1}$$

Step 2: For each point $\vec{p'}_\mathrm{j}$ compute the euclidean distance $d(\vec{p'}_\mathrm{j}, l_\mathrm{i})$ to each line segment l_i and set $\mathrm{d_{min}}$ according to equation 2

$$d_{\min}(\vec{p'}_\mathrm{j}(\Delta u, \Delta v, \Delta\Phi)) = \min_i\{d(\vec{p'}_\mathrm{j}, l_\mathrm{i})\} \tag{2}$$

Step 3: Computer the quality function Q using robust least squares according to equation 3

$$Q(\Delta u, \Delta v, \Delta\Phi) = \sum_j exp\left[\frac{d_{\min}(\vec{p'}_\mathrm{j}(\Delta u, \Delta v, \Delta\Phi))^2}{2.{\sigma_\mathrm{s}}^2}\right] \tag{3}$$

where ${\sigma_\mathrm{s}}^2$ is the variance of the laser distance measurement.

Robust least square method was used because the point-to-line correspondences are erroneous and can have many outliers. The least square method is highly sensitive to outliers because of the extremely small tails of the Gaussian model compared to most real measurement error distributions [16] such as in this setup.

Step 4: Optimize Q to find the best match for parameters $(\Delta u, \Delta v, \Delta\Phi)$ by sampling the parameter space in course steps and then refining the search around the minimum with steepest descent.

6 Global Pose Estimation

In [1], since the authors do not use a GPS that provides a global pose estimate, there is a need to compute a traversed path by successively adding relative position estimates from the horizontal laser scans to effectively generate a 3D point cloud from the vertical scans. In addition, the authors have to assume that the environment is flat without significant altitude changes since the scans provide only 3 DOF (x, y, θ) and so the global pose can only be described in a 2D plane.

Starting with an initial global pose estimate (x_0, y_0, θ_0) and applying coordinate transformations using scan matching parameters $(\Delta u, \Delta v, \Delta\Phi)$, a new position $(x_{i+1}, y_{i+1}, \theta_{i+1})$ can be computed from 4

$$\begin{aligned} x_{i+1} &= x_i + \Delta u_i . \cos(\theta_i + \Delta\phi_i) - \Delta v_i . \sin(\theta_i + \Delta\phi_i) \\ y_{i+1} &= y_i + \Delta u_i . \sin(\theta_i + \Delta\phi_i) + \Delta v_i . \cos(\theta_i + \Delta\phi_i) \\ \theta_{i+1} &= \theta_i + \Delta\phi_i \end{aligned} \tag{4}$$

There is a tradeoff associated with this technique. With each iteration step in equation 4, errors start to accumulate and hence it is important to recover the path in least steps using a large subsampling factor but also ensure that the subsampling factor is small enough for maximum overlap between successive scans for an efficient match. Therefore, the authors decided to go with an adaptive subsampling factor adjusted to the driving speed of the truck.

The authors did not fuse results obtained from the heading sensor and the speedometer using a Kalman Filter arguing that the pose estimates obtained from these devices are much less reliable than the laser scans. In addition, the pose estimation error of the scan matching did not follow Gaussian white noise statistics. Therefore the authors opted to use these sensors only for consistency checks.

In [2] the authors use a GPS/INS/Odometer based navigation system to circumvent the need to perform the aforementioned to compute a global pose estimate. Instead they simply geo-reference the data sources from the local coordinate system of each sensor to a world coordinate system.

Using the navigational data, a transformation matrix T_{hg} is computed to transform range scan lines and line images from the local coordinate system of the GPS to the world coordinate system. The transformation matrix T_{lh} from local coordinate system of laser range scanner and T_{ch} from the local coordinate system of line camera are calculate using exterior calibration parameters.

A range point with a distance r and angle α is geo-referenced using 5

$$(x, y, z)^t = T_{hg} T_{lh} (-r \sin\alpha, 0, -r \cos\alpha)^t \tag{5}$$

The projection vector of image pixel h is geo-referenced using 6

$$\begin{aligned} (x, y, z)^t &= T_{hg} T_{ch} (0, -\cos\omega, \sin\omega)^t \\ \text{where } \omega &= 2x \arcsin((h - o)/2/f) \end{aligned} \tag{6}$$

7 Creating Geometric Models

In [2] the reconstruction of urban features is done in three levels: buildings, trees and ground surfaces. Buildings are further divided into vertical and non-vertical surfaces. To classify the range points into each levels, the scan lines are at first divided into meaningful line segments using [17]. The range points are then grouped into a hierarchical procedure based on the following rules -

1. points of vertical line segments are measurements of vertical buildings
2. points of low gradients and low elevation values are measurements of ground surfaces
3. points of small variance values and belonging to large clusters are measurements of trees

The above classification is only used as a starting point, since in practice a misclassification is quite common. Therefore the classification is also validated and refined using extracted geometric features.

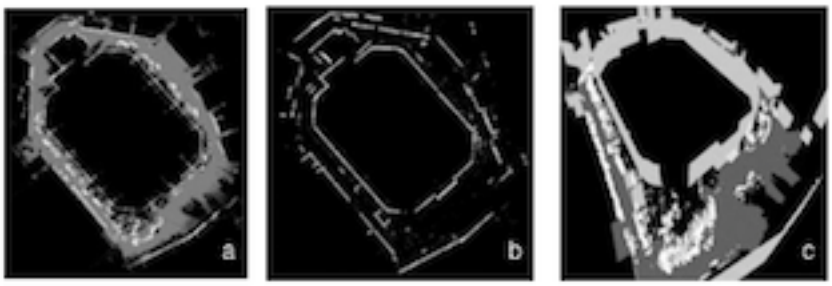

Figure9: (a) All Range Points, (b) Vertical Surfaces, (c) Geometrical Model [2]

The vertical and ground surface extraction is performed using z-images which reduces a 3D extraction problem into a 2D extraction problem [18]. The range points belonging to a group of vertical buildings are extracted using z-images of the range points and later recovered to geometric polygons as shown in figure 9. On the other hand, a triangulated irregular network constructed on an elevation map is used to represent the ground surface geometry. The elevation map contains minimal Z-value of range points projected onto a regularly tessellated horizontal plane as shown in figure 10

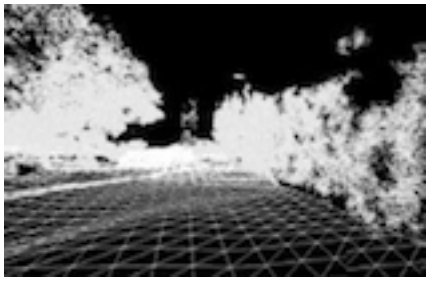

Figure10: Ground Surface in TIN Model [2]

Non-vertical surface extraction is performed using 3D USF segmentation [19] where the trees are modeled as triangular cells using the remaining set of range points that always yield large clusters. The small cluster of images features in z-images are removed.

8 Texture Mapping

In [2] distortions occur in line images due to the change in relative distance and direction of the objects from the line camera and are corrected by resampling the images on each planar surface to generate a texture of each building surface. For ground surfaces, each vertex of TIN model is projected along the direction of the range beams to the horizontal plane to create a connection between TIN model and texture image. The result of the texture mapping on the geometric model is shown in figure 11

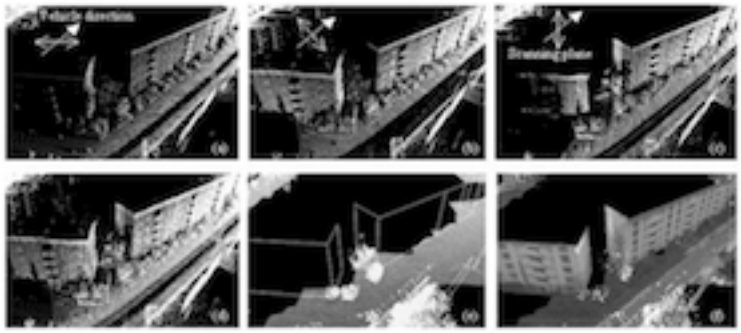

Figure11: Reducing occlusion using multiple laser range and line images: (a) range points by LDA1 (b) range points by LDA2 (c) range points by LDA3 (d) integrated range points (e) geometric model (f) textured model [2]

9 Results

In [2] each laser range scanner and line camera measured over 7000 scan lines and 30000 line images respectively. Multiple range scanners and line cameras helped reduce occlusions since a single range scanner can only capture building surfaces within a specific range of angles as illustrated in figure 11

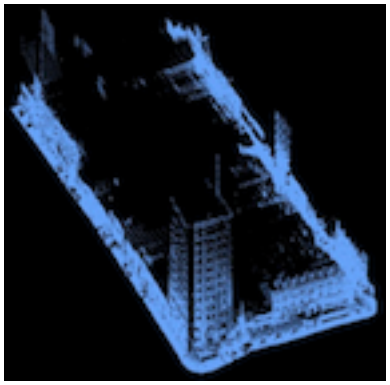

Figure12: Scan of the Entire Block [1]

Figure13: Texture Map [1]

The resulting point cloud of [1] is shown in figure 12. The post-processing process of the point cloud involving mesh generation, plane extraction and texture mapping as shown in figure 13 is not discussed in the current paper but in future works 10.

10 Future Work

In simultaneous localization and mapping, as the vector estimates of the relative movement, obtained from the path computation step are summed up, there is an inevitable accumulation of error over time, unless a global error correction step is performed using aerial images. Therefore, in [20] to improve the accuracy of localization of the truck and hence the resulting 3D models of the city, two different methods were developed and compared: the first method employed a correlation technique and the second method was based on Monte Carlo localization. Both techniques used digital road maps and aerial photographs in conjunction with laser scans.

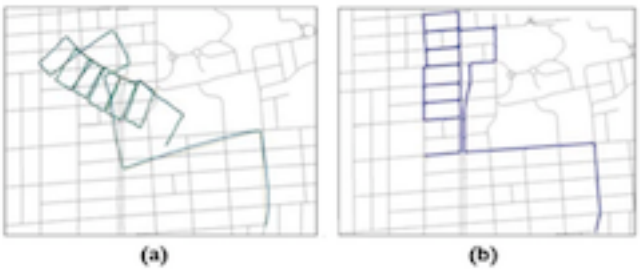

Figure14: (a) Original Path (b) Adjusted Path after MCL Correction [20]

The markov localization technique performed better since it represented position estimates not only by one single set of parameters, but instead by a probability distribution over the parameter space. A motion phase and a perception phase were performed iteratively where the motion phase flattened the probability distribution whereas the perception phase sharpened the position estimate. As this method propagated multiple hypotheses, it became capable of recovering from position errors and mismatches as shown in figure 14

Figure15: Background Layer [21]

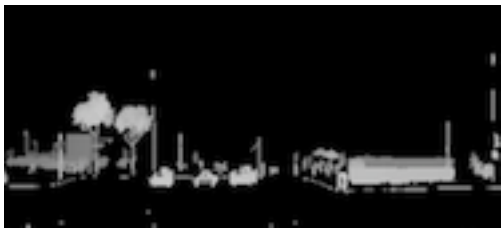

Figure16: Foreground Layer [21]

The laser scanning system in general has imperfections and non-idealities such as occlusions and reflections from glassed surfaces. Therefore in [21] the authors introduced several data processing algorithms that cope with these challenges to create a detailed 3D textured facade mesh to represent building walls at the highest level of detail.

The driven path was cut into easy-to-handle quasi-linear segments with approximately straight driving direction to ensure that there existed no two scan pairs with permuted horizontal order. The vertical scans were then transformed into a depth image using gray value as the depth parameter for easy interpretation and increased processing speed.
A representation based on multiple depth layers (foreground and background) was introduced to

Figure17: Before Processing [21]

Figure18: After Processing [21]

complete the geometry of the building facades from all viewpoints as shown in figure 1516. Large holes in the background layer, caused by occlusion from foreground layer objects, were then filled in by planar or horizontal interpolation and finally the images acquired from the digital camera were back projected on the mesh as shown in the figure 1718
In addition, since the facade models did not provide provide information about roofs or terrain shapes because they consisted of only surfaces visible from the ground level, the authors in [22] described an approach to register and fuse the detailed facade models with complementary airborne data as shown in figure 19

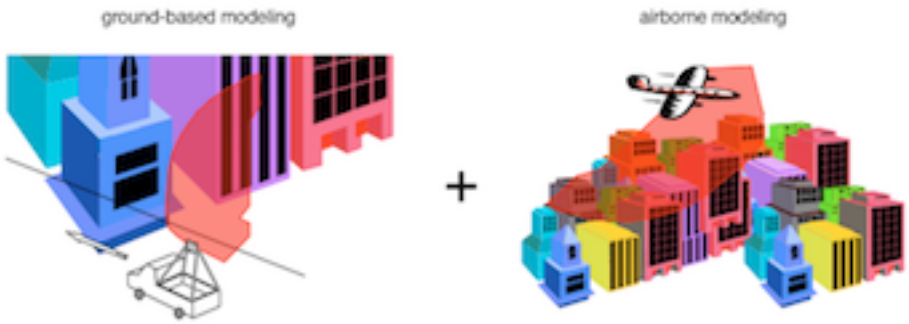

Figure19: Merging Aerial and Ground Views [1]

A digital surface map was generated from airborne laser scans and processed to prevent separation of rooftops into many small clustered regions. The cleaned up DSM was then directly transformed into a triangular mesh and texture mapped with aerial images semi-automatically. The resulting airborne-mesh was then used to compliment the roofs and terrain shapes in the detailed ground-level facade as shown in figure 20 21

Figure20: Ground-level View [22]

Figure21: Airborne View [22]

References

[1] CFruh and AZakhor. Fast 3d model generation in urban environments. *Multisensor Fusion and Integration for Intelligent Systems, 2001. MFI 2001. International Conference on*, pages 165–170, 2001.

[2] HZhao and RShibasaki. Reconstructing a textured cad model of an urban environment using vehicle-borne laser range scanners and line cameras. *Machine Vision and Applications*, 14(1):35–41, 2003.

[3] Wolfgang Förstner. 3d-city models: Automatic and semiautomatic acquisition methods. *Proc. Photogrammetric Week, Univ. of Stuttgart, Inst. for Photogrammetry*, 1999.

[4] E.Riseman R.Collins, A.Hanson and H.Schultz. Automatic extraction of buildings and terrain from aerial images. *International Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, pages 169–178, 1995.

[5] PaulE. Debevec, CamilloJ. Taylor, and Jitendra Malik. 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*, SIGGRAPH '96, pages 11–20, New York, NY, USA, 1996. ACM.

[6] M.Pollefeys, R.Koch, M.Vergauwen, and L.Van Gool. Hand-held acquisition of 3d models with a video camera. In *IEEE Proceedings of 2nd International Conference on 3D Digital Imaging and Modeling (3DIM'99)*, pages 14–23. Society Press, 1999.

[7] Seth Teller, Matthew Antone, Zachary Bodnar, Michael Bosse, Satyan Coorg, Manish Jethwa, and Neel Master. Calibrated, registered images of an extended urban area. In *International Journal of Computer Vision*, pages 93–107, 2001.

[8] Heung-Yeung Shum and Richard Szeliski. Stereo reconstruction from multiperspective panoramas. *Computer Vision, IEEE International Conference on*, 1:14, 1999.

[9] Chungan Lin and Ramakant Nevatia. Building detection and description from a single intensity image. In *Proceedings of Computer Vision and Image Understanding (CVIU)*, volume72, pages 101–121, 1998.

[10] FVan DenHeuvel. 3d reconstruction from a single image using geometric constraints. *ISPRS Journal of Photogrammetry and Remote Sensing*, 53(6):354–368, 1998.

[11] Sanjay Noronha and Ramakant Nevatia. Detection and modeling of buildings from multiple aerial images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:501–518, May 2001.

[12] Suya You, Jinhui Hu, Ulrich Neumann, and Pamela Fox. Urban site modeling from lidar. In *Proceedings of the 2003 international conference on Computational science and its applications: PartIII*, ICCSA'03, pages 579–588, Berlin, Heidelberg, 2003. Springer-Verlag.

[13] STeller. Automated urban model acquisition: Project rationale and status. *Image Understanding Workshop*, Jan 1998.

[14] Feng Lu and Evangelos Milios. Robot pose estimation in unknown environments by matching 2d range scans. *Journal of Intelligent and Robotic Systems*, 18:249–275, 1994.

[15] J.-S. Gutmann, Thilo Weigel, and B.Nebel. Fast, accurate, and robust self-localization in polygonal environments. In *Intelligent Robots and Systems, 1999. IROS '99. Proceedings. 1999 IEEE/RSJ International Conference on*, volume3, pages 1412 –1419 vol.3, 1999.

[16] Bill Triggs, PhilipF. McLauchlan, RichardI. Hartley, and AndrewW. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, 2000. Springer-Verlag.

[17] P.L. Rosin and G.A.W. West. Nonparametric segmentation of curves into various representations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(12):1140 –1153, December 1995.

[18] Huijing Zhao and Ryosuke Shibasaki. Reconstruction of textured urban 3d model by fusing ground-based laser range and ccd images(special issue on machine vision applications). *IEICE transactions on information and systems*, 83(7):1429–1440, 2000-07-25.

[19] A.Hoover, G.Jean-Baptiste, X.Jiang, P.J. Flynn, H.Bunke, D.B. Goldgof, K.Bowyer, D.W. Eggert, A.Fitzgibbon, and R.B. Fisher. An experimental comparison of range image segmentation algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(7):673 –689, July 1996.

[20] C.Fruh and A.Zakhor. 3d model generation for cities using aerial photographs and ground level laser scans. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001.

[21] C.Fruh and A.Zakhor. Data processing algorithms for generating textured 3d building facade meshes from laser scans and camera images. In *3D Data Processing, Visualization and Transmission 2002, Padua, Italy*, pages 834–847, June 2002.

[22] C.Frueh and A.Zakhor. Constructing 3d city models by merging ground-based and airborne views. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume2, pages II – 562–9 vol.2, 2003.