3D Scene Reconstruction through a Fusion of Passive Video and Lidar Imagery

Prudhvi Gurram, Harvey Rhody, John Kerekes
Chester F. Carlson Center for Imaging Science Rochester Institute of Technology Rochester, NY, USA.
pxg0638@rit.edu, rhody@cis.rit.edu, kerekes@cis.rit.edu

Stephen Lach
Chester F. Carlson Center for Imaging Science Rochester Institute of Technology Rochester, NY, USA.
Air Force Institute of Technology 2950 Hobson Way WPAFB, OH, USA.
sfl1194@cis.rit.edu

Eli Saber
Department of Electrical Engineering and Chester F. Carlson Center for Imaging Science Rochester Institute of Technology Rochester, NY, USA.
esseee@rit.edu

Abstract

Geometric structure of a scene can be reconstructed using many methods. In recent years, two prominent approaches have been digital photogrammetric analysis using passive stereo imagery and feature extraction from lidar point clouds. In the first method, the traditional technique relies on finding common points in two or more 2D images that were acquired from different view perspectives. More recently, similar approaches have been proposed where stereo mosaics are built from aerial video using parallel ray interpolation, and surfaces are subsequently extracted from these mosaics using stereo geometry. Although the lidar data inherently contain 2.5 or 3 dimensional information, they also require processing to extract surfaces. In general, structure from stereo approaches work well when the scene surfaces are flat and have strong edges in the video frames. Lidar processing works well when the data is densely sampled. In this paper, we analyze and discuss the pros and cons of the two approaches. We also present three challenging situations that illustrate the benefits that could be derived from this data fusion: when one or more edges are not clearly visible in the video frames, when the lidar data sampling density is low, and when the object surface is not planar. Examples are provided from the processing of real airborne data gathered using a combination of lidar and passive imagery taken from separate aircraft platforms at different times.

1. Introduction

Research in the field of 3D scene reconstruction has gained momentum in recent years due to a wide range of military and civilian applications. Much advancement has been made in this front. There are many approaches in the market today to extract 3D models from aerial passive imagery. A good survey of some of the methods can be found in [3]. We can also find a large number of approaches (manual and semi-automated) dealing with lidar point cloud to extract 3D models in [3].

In this direction, we are trying to build a 3D model of a scene which conforms both geometrically as well as spectrally to the real world. For this purpose, accurate 3D geometry of the scene is required. We are trying to extract this information from multi-modal data sets like passive video and lidar point data. Once we have the 3D scene model, we are going to combine this model with the material properties derived from hyperspectral imagery to generate a realistic scene in RIT's DIRSIG (The Digital Imaging and Remote Sensing Image Generation) model [1]. A more detailed process can be found in [10].

Buildings and trees are two dominant classes of objects typically observed in urban scenes and are therefore the primary objects of interest for reconstruction of a scene. In this paper, we discuss the extraction and modeling of man-made buildings using passive video and lidar point data. We also compare the models extracted from these two sets of data. For this, first the terrain is extracted from the area of interest. The surfaces belonging to individual objects in this area, namely, the buildings and the trees, are identified next. Building surfaces and tree regions can be distin-

1550-5219/07 $25.00 © 2007 IEEE
DOI 10.1109/AIPR.2007.17

IEEE computer society

guished from each other using texture measures such as the entropy of height values within a small window. All the surfaces of a single building are then used to reconstruct a 3D CAD model of that building., while tree geometries are currently selected from a pre-constructed CAD library.

The rest of the paper is organized as follows: Section 2 describes the process of constructing a building model using passive video and lidar point cloud, Section 3 presents the results and Section 4 provides the conclusions and future goals.

2. Two approaches for extraction of 3D models from multi-modal data sets

We are using two approaches to extract 3D geometry of a building in a scene. In the case of passive video, stereo mosaics are built from the individual video frames and 3D coordinates are extracted from the stereo mosaics. A lidar point cloud is available in raw format. We have semi-automated the process of extraction and modeling of any building from this data. The approaches are explained in the following sub-sections.

2.1. Extraction of 3D models from passive video

Various methods are used to extract 3D information from visual imagery. The most popular and widely used method is stereo vision, which refers to the ability to infer information on the 3-D structure of a scene from images obtained at two or more viewpoints [17]. A stereo system usually has a stereo rig with two cameras placed at a particular angle with each other (to satisfy stereo geometry). But this hardware formerly required for a stereo system is no longer indispensable, as new algorithms can extract desired 3D information from two or more images of a scene photographed by a single camera. Every object in the scene requires two or more views for such techniques to be successfully applied on them. Hence, a large scene requires many images to cover the entire area.

To effectively deal with this problem, a video camera maybe attached to an aircraft and flown over the scene. However this results in hundreds of frames that need to be processed. Also, in order to extract 3D coordinates of a particular object, we must identify the frames in which that object is present. There is also an additional constraint that the object has to exhibit good disparity with respect to the baseline for fairly accurate results. Thus the frames need to be indexed and used for the retrieval of 3D data of any object. This is a very tedious process to implement for an extended scene due to the large number of frames and objects involved in the scene. To alleviate this issue, we have chosen to stitch the frames to form a mosaic which is easier to handle than individual frames. Two mosaics are built from these frames, in such a way that they form a stereo pair for the entire scene.

The process of stitching large number of video frames to form a mosaic is very simple once corresponding points are matched in successive frames. But such mosaics do not serve our purpose because they are not seamless, and the apparent motion parallax information between the frames is often lost in the mosaics. To address this, Zhu et al. proposed the Parallel Ray Interpolation for Stereo Mosaicing (PRISM) [18] algorithm to build seamless stereo mosaics. In this method, the mosaics are built by creating imaginary viewpoints between already existing viewpoints of the camera using ray interpolation, such that we are apparently looking at the scene at a particular angle (perspective does not change) at any (existing or imaginary) viewpoint (emulating a pushbroom camera). This technique converts the perspective-perspective view of the video frames to parallel-perspective view of the mosaics with parallel view in the dominant motion direction of the camera. By varying the angle of the parallel view in the algorithm, two mosaics are built: a left mosaic (forward looking in the direction of motion of the sensor) and a right mosaic (rear looking), which act as a stereo pair.

We have improved this technique by avoiding the artifacts generated during the triangulation process of the algorithm when there is large motion parallax between two surfaces in successive frames. The triangulation process was modified in such a way that none of the triangles would enclose regions from two different surfaces in any video frame and thus will not get warped in an undesired way on to the mosaic [6].

Along with the left and right mosaics, we also build a nadir mosaic (with the parallel view of the sensor looking straight down at the scene). The nadir mosaic is used to distinguish the visible surfaces from the occluded surfaces in the scene. For example, a vertical surface of an object will not be visible in the nadir mosaic but may be visible in the left or right mosaic. On the other hand, a slant surface will be visible in the nadir mosaic too.

The nadir mosaic is segmented to identify the different surfaces in the scene [15]. In this paper, we focus on modeling a building. First, all the surfaces belonging to a building are identified. For each segment, the boundary pixels (which form the edge of each surface) are found using morphological operations. Matches are then found for these boundary pixels in the left and right mosaics. Disparity between the corresponding points in the left and right mosaics is determined, and the elevation of the surface at each of these points is found as described in [18]. 3D points are available for each surface by combining the elevation of the boundary pixels with their 2D position in the nadir mosaic.

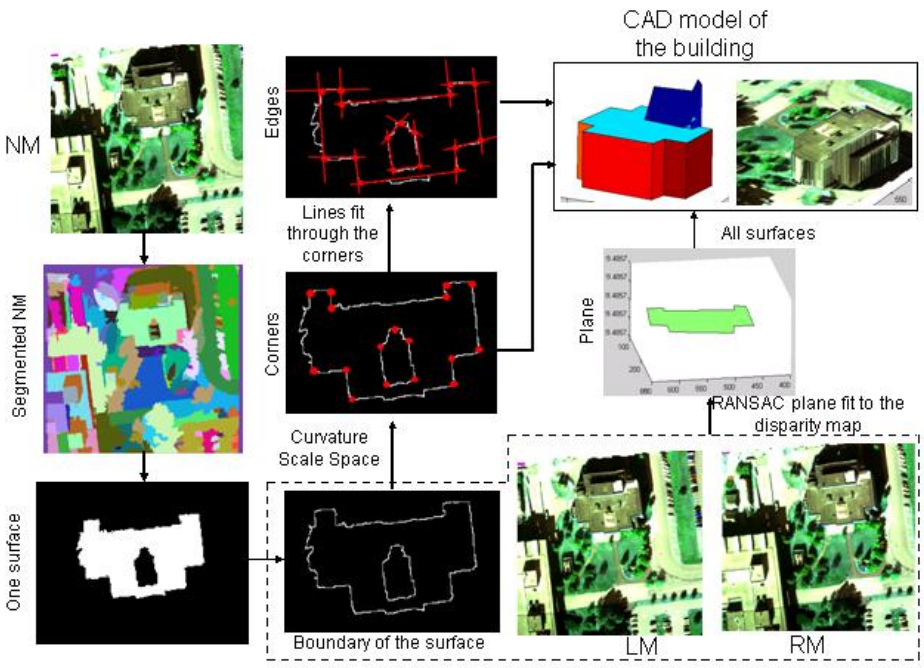

Figure 1. Building modeling from stereo mosaics

Planes are fit through these 3D points using the RANSAC (RANdom SAmple Consensus) algorithm [8]. RANSAC is used to make the process robust to noise which might arise due to improper segmentation and improper matching of points between the left and right mosaics and remove the outliers. The surface boundary pixels of each segment are also used to define the corners of that surface. Boundary pixels with high curvature are identified as corners as described in the Curvature Scale Space (CSS) algorithm [13], [7]. Since the boundary is derived from a segment, and not by gradient methods, it is continuous, and therefore we can avoid some of the steps described in [13]. The unwanted corners are removed by applying one more verification step using the mean squared distance between the boundary obtained by connecting the corners and actual boundary as a measure. Using the 2D corners and plane equation of the surface, we obtain 3D vertices of the surface. A CAD model can be generated with just the 3D vertices of every surface and the order of connectivity of these vertices. However, we still have an issue, as no information is available regarding the occluded (vertical) surfaces (shadows due to the view angle of the sensor). To address this, we make use of prismatic building assumption. Vertical surfaces can be identified using the information that they are not visible in the nadir mosaic and must exist between two surfaces which have a common edge but are at different heights [16]. For this purpose, edges are obtained by joining the corners of each surface. Once a vertical surface is identified, we drop a surface vertically down from the higher surface to the lower surface. We repeat this for all the surfaces to obtain the CAD model of a building. This entire process is depicted in Figure 1.

2.2. Extraction of 3D models from raw lidar point cloud

The method used to reconstruct buildings from lidar point cloud data also makes use of the prismatic building assumption. Dominant roof planes are identified by segmenting the original points using local point properties as feature vectors, and adjacent planes are then analyzed to determine the inner roof-segment boundaries. Since vertical surfaces of the building structure are not directly represented in the data, a separate methodology must be used to determine the outer roof boundary. Once the initial building model is completed, the geometry is refined through both a comparison of the model with the original point data and the introduction of geometric constraints. To this end, we have implemented the following steps in our approach, which is more fully detailed in [9]:

1. Determine initial exterior boundary estimate from lidar data using alpha shapes and line-fitting.

2. Segment the lidar range image such that each segment represents a planar face.

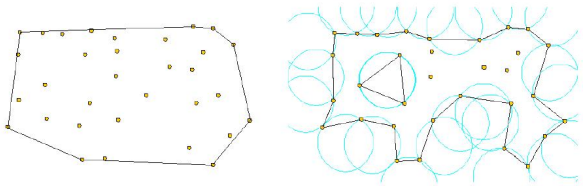

Figure 2. Shape of a collection of 2D points. Convex hull (left) and alpha shapes (right)

3. Determine internal boundaries through an intersection-of-planes approach

4. Determine vertices through intersections of inner and/or outer boundaries.

5. Refine the building model by introducing geometric constraints

6. Refine the lidar-derived model through a verification process using the original point data

Since many lidar datasets are obtained from near-nadir orientations, very few data points are available that lie on vertical surfaces. As such, it is often difficult to determine the planes corresponding to exterior walls from data points on these walls. This problem may be partially alleviated if we make the simplifying assumption that exterior walls are oriented directly under the outer boundary of a building object, a condition that is true in many building types. Therefore, in modeling the geometry of a given object on a given building layer, the first step is to determine the exterior roof boundary of that object.

Due to potential concavity in this boundary, simple shape descriptions such as the convex hull do not provide an adequate description of the outer roof shape. To this end, we have opted to use alpha shapes for the determination of our exterior roof boundaries. Like the convex hull, alpha shapes are simply another approach to formally describe the 'shape' of a set of spatial point data. Unlike the convex hull, alpha shapes are not limited to convex geometries, and may even represent holes inside the geometry.

As described in [5], we may think of alpha shapes as a family of shapes for a given point data set, where each shape is defined as the intersection of all closed discs with radius $1/\alpha$. In practice, is set such that $2\alpha^2$ is set to be 25% larger than the largest point-point spacing in the sampled lidar data. Figure 2 depicts the convex hull and one of the alpha shapes for a given set of 2D data points.

Once the outer boundary has been determined, inner plane edges and corresponding vertices still need to be defined. This is accomplished though a methodology similar to that presented in [14]. First, each data point is assigned a normal vector according to the plane best fitting the data in a lm^3 voxel centered on the point of interest. This plane is determined through a 3D Deming regression. In a manner similar to that presented in [14], the mean shift algorithm [4] is then used to segment the points into several groups, using the normal vectors and point locations as the defining features. Coplanar adjacent regions are then merged, and planes are fit to each data region. Planes from adjacent regions are then intersected with each other to determine candidate facet boundaries. Where the candidate boundaries match the actual data region boundaries well, those candidate boundaries are used to define the inner roof edges. In places where the candidate boundary does not match the actual data region boundaries, a breakline is assumed, and a piecewise linear boundary is fit to the data region boundary. Vertical walls are projected down from both breaklines and outer edge surfaces until they intersect another roof plane or the previously extracted terrain model.

After the initial roof structure has been coarsely modeled, we often refine this model by introducing geometric constraints. Although we typically only require the outer roof edges to lie along lines that are oriented at increments of 45 degrees relative to each other, additional constraints are possible. These include ensuring certain edges lie at a constant height, or that specific inner edges meet exactly at outer boundary corner points. A second refinement stage may also be performed, in which the distance between each original data point and the resultant model is calculated. In regions where the errors are large, we search for additional (and often undersampled) features such as window dormers. A reconstruction of these additional features is then attempted through an intersection of planes approach as well as through parametric modeling of common roof objects. When this additional feature extraction step fails and errors between the model and the original data remain high in a localized region, we then create facets directly from the point data using a Delaunay triangulation.

2.3. Need for the fusion of the two approaches

Both forms of data have limitations, because of which they can be used only upto a certain extent. Section 3 explains the limitations in one method and shows they can be compensated for by using the alternate approach. Hence our final goal is to fuse both the approaches to produce a fairly accurate 3D geometrical model of a scene.

3. Results

We present the results in the form of three challenging situations for us to use either the passive video or lidar to reconstruct the 3D model of a building. The passive video

and lidar data were collected from different airborne platforms at different times. Three-band video was collected using a sensor called Wildfire Airborne Sensing Program (WASP) Lite [2] which was developed at RIT. Three narrow band filters were used to capture the color information. The ground spot size of this sensor is about 0.25 m at a height of about 1000 ft. The overlap between successive frames varies between 90% and 98%. Lidar data was supplied by Leica Geosystems flying a commercial Optech sensor. The data contained approximately 6 $points/m^2$, roughly uniform in both the in- and cross-track dimensions. Multiple-return range and intensity data were provided.

3.1. Situation 1

In many cases, usually passive video has many shadow regions due to the viewing angle of the sensor and also due to the solar inclination angle. The shadow regions (occlusions) due to the viewing angle of the sensor can be taken care of, for many cases as shown in [16]. But the shadow regions due to solar inclination angle pose a much bigger problem as there is no visual information in such regions. Such regions when segmented as a part of the modeling process described in Section 2.1, will either merge into other shadow regions or provide no information even with their edges. In such cases, we have a part of the building missing in the reconstructed CAD model of RIT's Center for Imaging Science as shown in Figure 3. There is another problem with edges when there is noise in the images. One surface gets merged into another during segmentation due to noise and hence, the surfaces would have a different orientation and height compared to the true values as seen in Figure 3. But lidar data inherently has 3D information and does not have a problem with shadows, the surfaces along with their edges can be reconstructed fairly accurately depending on the sampling of the lidar data. On a side note, we can see in Figure 3, some of the projected structures on the top of the roof of the building could not be reconstructed using passive video as they do not show significant disparity compared to that of the roof. But adequately sampled lidar data can give us information about these small structures too.

3.2. Situation 2

Sometimes lidar data might be undersampled due to the constraints in the lidar data acquisition system hardware. In such cases, the outer boundary of a building found using the method described in Section 2.2 falls apart as shown in Figure 4. Even the segmentation to determine the inner boundaries of a building also falls apart and the building cannot be reconstructed. But the passive video, once segmented, can give good boundaries and hence help in the reconstruction of the CAD model.

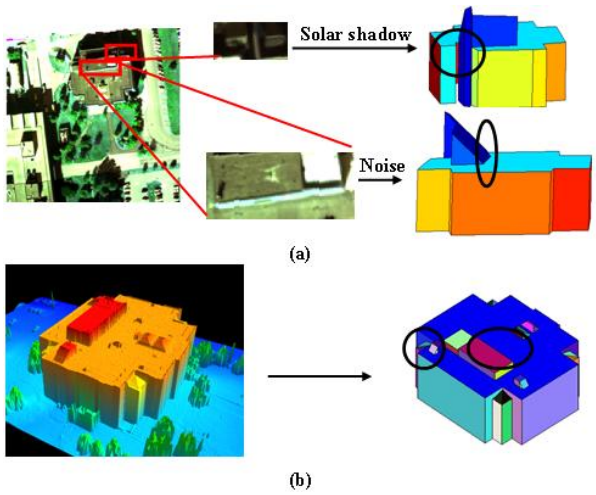

Figure 3. (a) Artifacts in 3D model of RITs Center for Imaging Science built from passive video due to solar shadows and noisy edges (b) Accurate 3D model of the same building built from lidar point cloud

3.3. Situation 3

The most difficult situation of all is the case of a textureless non-planar surface like a hemispherical dome. With passive video, as the surface does not have texture, it will not show any variation in disparity over the surface. And hence the surface rebuilt would look like a cylindrical structure rather than a hemispherical dome over a cylindrical structure. But in the case of lidar data, we already have 3D points on the surface. A spherical surface is fit to these 3D points using Levenberg-Marquardt non-linear least squares regression [11], [12]. The model is a good fit to the 3D data and fairly accurate. RIT Observatory which has a similar structure is shown in Figure 5.

4. Conclusion

We have shown in the results that passive video and lidar point cloud are complementary to each other for extraction and modeling of buildings in an urban scene. So if the two models are registered and fused together, they can mutually reinforce one another and an accurate scene model can be obtained. For this, there is a need to recognize which method performs better in what situation. Our next step is to develop some kind of "hypothesize and verify" process at each step for both the methods. Accordingly, we can attribute a confidence measure to each feature extracted from the two methods. If the models built from the two methods are registered using features with high confidence in both

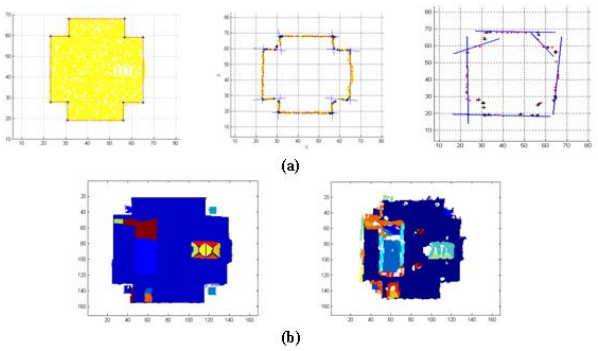

Figure 4. (a) Images showing the edges extraction using outer boundary analysis when lidar point cloud is adequately sampled, sampled at half the rate and sampled at quarter the rate. (b) Images showing failure of segmentation in inner boundary analysis due to undersampled lidar data

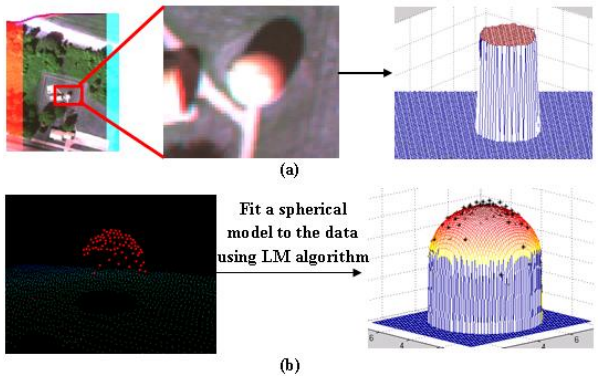

Figure 5. The hemispherical dome of RIT Observatory can be modeled by fitting a spherical surface to lidar point cloud (b) but can not be modeled from passive imagery (a)

the models, then the features with low confidence in one model can be corrected using the same feature in the second model.

5. Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the U.S. Air Force, the Department of Defense or the U.S. Government.

References

[1] The Digital Imaging and Remote Sensing Image Generation Model. <http://dirsig.cis.rit.edu/>.

[2] Wildfire Airborne Sensing Program (WASP) LT. <http://twiki.cis.rit.edu/bin/view/LIAS/>.

[3] C. Brenner. Building reconstruction from images and laser scanning. *International Journal of Applied Earth Observation and Geoinformation*, 6(3–4):187–198, 2005.

[4] D. Comaniciu. *Nonparametric Robust Methods for Computer Vision*. PhD Thesis, Rutgers University, New Jersey, 2000.

[5] H. Edelsbrunner and E. P. Mucke. *Three-dimensional Alpha Shapes*. Department of Computer Science, University of Illinois Urbana-Champaign, Illinois, 1992.

[6] P. Gurram, E. Saber, and H. Rhody. A novel triangulation method for building parallel-perspective stereo mosaics. In *Proceedings of SPIE/IS&T Electronic Imaging Symposium*, San Jose, CA, 2007.

[7] X. C. He and N. H. C. Yung. Curvature scale space corner detector with adaptive threshold and dynamic region of support. In *Proceedings of 17th International Conference on Pattern Recognition (ICPR' 04), IEEE*, 2004.

[8] P. D. Kovesi. MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[9] S. Lach. *Semi-Automated DIRSIG Scene Construction Using Multi-Modal Imagery, Proposal for Doctoral Research*. Rochester Institute of Technology, Rochester, NY, 2007.

[10] S. Lach and J. Kerekes. Multisource data processing for semi-automated radiometrically-correct scene simulation. In *Proceedings of IEEE Urban Remote Sensing Joint Event*, 2007.

[11] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

[12] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 11:431–441, 1963.

[13] F. Mokhtarian and R. Suomela. Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1376–1381, 1998.

[14] R. Mu. *Building Model Reconstruction from Lidar Data and Aerial Photographs*. PhD Thesis, The Ohio State University, Ohio, 2004.

[15] E. Saber, A. M. Tekalp, and G. Bozdagi. Fusion of color and edge information for improved segmentation and edge linking. *Image and Vision Computing*, 15:769 – 780, 1997.

[16] T. Schenk and B. Csatho. Fusing imagery and 3d point clouds for reconstructing visible surfaces of urban scenes. In *Proceedings of IEEE Urban Remote Sensing Joint Event*, 2007.

[17] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice-Hall, New Jersey, 1998.

[18] Z. Zhu, A. R. Hanson, and E. M. Riseman. Generalized parallel-perspective stereo mosaics from airborne video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):226–237, 2004.