Automatic Extraction of Buildings and Terrain from Aerial Images *

Robert T. Collins, Allen R. Hanson, Edward M. Riseman, and Howard Schultz

Department of Computer Science
Lederle Graduate Research Center
University of Massachusetts
Amherst, MA. USA 01003-4610

Abstract

A system has been developed to acquire, extend and refine 3D geometric site models from aerial imagery. The system hypothesizes potential building roofs in an image, automatically locates supporting geometric evidence in other images, and determines the precise shape and position of the new buildings via multi-image triangulation. Model-to-image registration techniques are applied to align new images with the site model, and model extension and refinement procedures are performed to acquire previously unseen buildings and improve the geometric accuracy of the existing 3D models. A correlation-based terrain recovery algorithm provides complementary information about the site, in the form of a digital elevation map.

1. Introduction

Acquisition of 3D geometric site models from aerial imagery is currently the subject of an intense research effort in the U.S., sparked in part by the ARPA/ORD RADIUS project (Gerson, 1992; Huertas, 1993; Collins, 1994; Roux, 1994). We have developed a set of image understanding modules to acquire, extend and refine 3D volumetric building models, and to provide a digital elevation map of the surrounding terrain. System features include model-directed processing, rigorous camera geometry, and fusion of information across multiple images for increased accuracy and reliability.

Site **model acquisition** involves processing a set of images to detect both man-made and natural features of interest, and to determine their 3D shape and placement in the scene. This paper focuses on algorithms for automatically extracting models of buildings (Section 2) and terrain (Section 3). The site models produced have obvious applications in areas such as surveying, surveillance and automated cartography. For example, acquired site models can be used for automated model-to-image registration of new images (Collins, 1993), allowing the model to be overlaid on the image to aid visual change detection and verification of expected scene features. Two other important site modeling tasks are **model extension,** updating the geometric site model by adding or removing features, and **model refinement**, iteratively refining the shape and placement of features as more views become available. Model extension and

* This work was funded by the RADIUS project under ARPA/Army TEC contract DACA76-92-C-0041 and ARPA/TACOM contract DAAE07-91-C-R035.

refinement are ongoing processes that are repeated whenever new images become available, each updated model becoming the current site model for the next iteration. Thus, over time, the site model is steadily improved to become more complete and more accurate.

2. Building Model Acquisition and Extension

This section focuses on algorithms for automatically extracting models of buildings in the site. To maintain a tractable goal for our research efforts, we have chosen initially to focus on a single generic class of buildings, namely flat-roofed, rectilinear structures. The simplest example of this class is a rectangular box-shape; however other examples include L-shapes, U-shapes, and indeed any arbitrary building shape such that pairs of adjacent roof edges are perpendicular and lie in a single plane.

2.1. Initial Model Acquisition

The building model acquisition process involves several subtasks: 1) line segment extraction, 2) building detection, 3) multi-image epipolar matching, 4) constrained, multi-image triangulation, and 5) projective intensity mapping. These algorithms will be presented by way of an experimental case study using images J1-J8 of the RADIUS model board 1 data set. Figure 1 shows a sample image from the data set. Each image contains approximately 1320 x 1035 pixels, with about 11 bits of gray level information per pixel. Unmodeled geometric and photometric distortions have been added to each image to simulate actual operating conditions. The scene is a 1:500 inch scale model of an industrial site. Ground truth measurements are available for roughly 110 points scattered throughout the model, which were used to determine the exterior orientation for each image. The residual resection error for each image is in the 2-3 pixel range, representing the level of unmodeled geometric distortion present in each image. This corresponds to a backprojection error of roughly 3-4.5 feet in (simulated) object space. This is a significant amount of error that presents a good test of system robustness.

Line Segment Extraction. To help bridge the huge representational gap between pixels and site models, feature extraction routines are applied to produce symbolic, geometric representations of potentially important image features. The algorithms for

Fig. 1: Sample image from the Radius model board 1 data set.

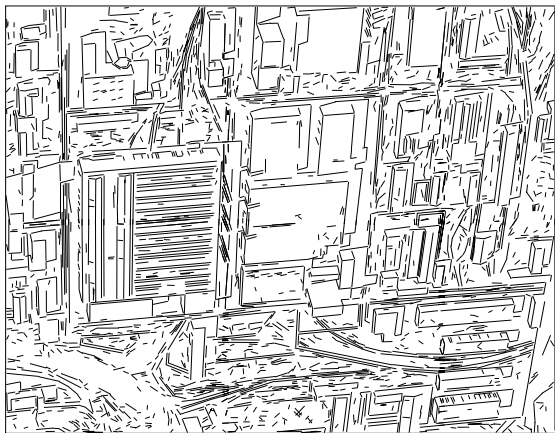

Fig. 2: Straight line segments produced by the Boldt algorithm.

acquiring building models rely on extracted straight line segments (Boldt, 1989). At the heart of the Boldt algorithm is a hierarchical grouping system inspired by the Gestalt laws of perceptual organization. Zero-crossings of the Laplacian of the intensity image provide an initial set of local intensity edges. Hierarchical grouping then proceeds iteratively; at each iteration edge pairs are linked and replaced by a single longer edge if their end points are close and their orientation and contrast values are similar. Filtering to keep line segments with a length of at least 10 pixels and a contrast of at least 15 gray levels produced roughly 2800 line segments per image. Figure 2 shows a representative set of lines extracted from the image shown in Figure 1.

Building Detection. The goal of automated building detection is to roughly delineate building boundaries that will later be verified in other images by epipolar feature matching and triangulated to create 3D geometric building models. The building detection algorithm is based on finding image polygons corresponding to the boundaries of flat, rectilinear rooftops in the scene (Jaynes, 1994). Briefly, possible roof corners are identified by line intersections. Perceptually compatible corner pairs are linked with surrounding line data, entered into a feature-relation graph, and weighted according to the amount of support they receive from the low-level image data. Potential building roof polygons appear as cycles in the graph; virtual corner features may be hypothesized to complete a cycle, if necessary. Rooftops are finally extracted by partitioning the feature-relation graph into a set of maximally weighted, independent cycles representing closed, high-confidence building roofs.

Figure 3 shows the results of building detection on image J3 of the model board 1 data set. The roof detector generated 40 polygonal rooftop hypotheses. Most of the hypothesized roofs are rectangular, but six are L-shaped. First, note that the overall performance is quite good for buildings entirely in view. Most of the major roof boundaries in the scene have been extracted, and in the central cluster of buildings (see area **A** in Fig. 3) the segmentation is nearly perfect.

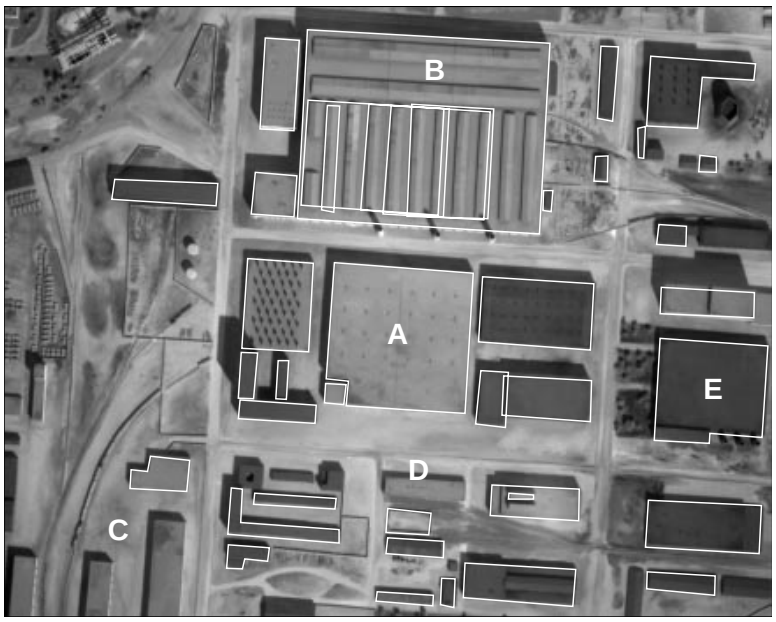

Fig. 3: Results of building detection on image J3.

There were some false positives, i.e. polygons extracted that do not in fact delineate the boundaries of a roof. The most obvious example is the set of overlapping polygonal rooftops detected over the large building with many parallel roof vents (area **B**) Note that the correct outer outline of this building roof is detected, however. There are also some false negatives, which are buildings that should have been detected, but weren't. The most prevalent example of

this is a set of buildings (area **C**) that are only partially in view at the edge of the image. Label **D** marks a false negative that is in full view. Two adjacent corners in the rooftop polygon were missed by the corner extraction algorithm. It should be stressed that even though a single image was used here for bottom-up hypotheses, buildings that are not extracted in one image will often be found easily in other images with different viewpoints and sun angles.

There are several cases that cannot be strictly classified as false positives or false negatives. Several split-level buildings appearing along the right edge of the image (area **E**) are outlined with single polygons rather than with one polygon per roof level. Some peaked roof buildings were also outlined, even though they do not conform to the generic assumptions underlying the system.

Multi-image Epipolar Matching. After detecting a potential rooftop in one image, corroborating geometric evidence is sought in other images (often taken from widely different viewpoints) via epipolar feature matching. The key problem in epipolar matching is disambiguation of multiple potential matches. One way to avoid ambiguity is to match higher-level structures that are more distinctive.

Rooftop polygons are matched by searching for each component line segment separately and then fusing the results. For each polygon segment from one image, an epipolar search area is formed in each of the other images, based on the known camera transformations and the assumption that the roof is flat. This quadrilateral search area is scanned for possible matching line segments, each potential match implying a different roof height in the scene. Results from each line search are combined in a 1-dimensional histogram, each match voting for a particular roof height, weighted by compatibility of the match in terms of expected line segment orientation and length. A single global histogram accumulates height votes from multiple images, and for multiple edges in a rooftop polygon. After all votes have been tallied, the histogram bucket containing the most votes yields an estimate of the roof height in the scene and a set of correspondences between rooftop edges and image line segments from multiple views.

Epipolar matching of a rooftop hypothesis is considered to have failed when, for any edge in the rooftop polygon, no line segment correspondences are found in any image. This criterion was chosen because the 3D line triangulation algorithm will fail to converge in this case. Based on this criterion, epipolar matching failed on eight rooftop polygons. Six were either peaked or multi-layer roofs that did not fit the generic flat-roofed building assumption, and the other two were building fragments with some sides shorter than the minimum length threshold on the line segment data. At this stage, six incorrect building hypotheses were removed by hand; detecting and removing such mistakes automatically is being actively investigated.

Multi-image Line Triangulation. Multi-image triangulation is performed to determine the precise size, shape, and position of a building in the local 3D site coordinate system. A nonlinear estimation algorithm has been developed for simultaneous multi-image, multi-line triangulation of 3D line structures. Object-space constraints are imposed for more reliable results. This algorithm is used for triangulating 3D rooftop polygons from the line segment correspondences determined by epipolar feature matching.

The parameters estimated for each rooftop edge are the Plücker coordinates of the algebraic 3D line coinciding with the edge - specific points of interest, like vertices of

the rooftop polygon, are computed as the intersections of these infinite algebraic lines. Plücker coordinates are a way of embedding the 4-dimensional manifold of 3D lines into $\boldsymbol{R}^6$. Although the Plücker representation requires 6 parameters to be estimated for each line rather than 4, it simplifies the representation of geometric constraints between lines. For the generic flat-roofed rectilinear building class being considered here, we specify a set of constraints to ensure that pairs of adjacent lines in a traversal around the polygon are perpendicular, that all lines are coplanar, and that all lines are perpendicular to the Z-axis of the local site coordinate system. An iterative, nonlinear least-squares procedure determines the Plücker coordinates for all lines simultaneously such that all the object-level constraints are satisfied and an objective "fit" function is minimized that measures how well each projected algebraic line aligns with the 2D image segments that correspond to it.

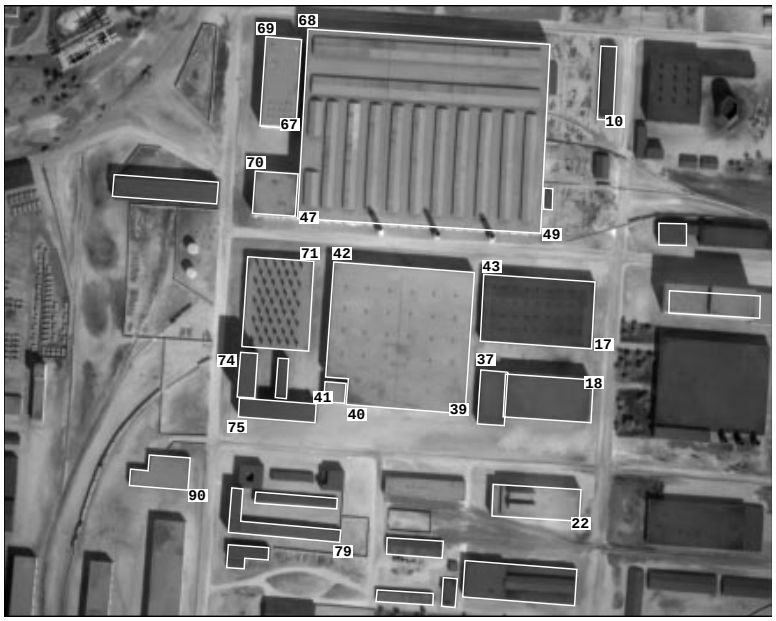

Fig. 4: The final verified and triangulated 3D rooftops.

After triangulation, each 3D rooftop polygon is extruded down to the ground to form a volumetric model. For the Model Board 1 site we represented the ground as a horizontal plane with Z-coordinate value determined from the ground truth measurements. More generally, we will soon be combining our symbolic building extraction routines with the digital terrain maps produced by the UMass Terrain Reconstruction System (Section 3). Outlines of the final set of triangulated rooftops are shown in Figure 4.

To evaluate the 3D accuracy of the triangulated building polygons, 21 roof vertices were identified where ground truth measurements are known (numbered vertices in Figure 4). The average Euclidean distance between triangulated polygon vertices and their ground truth locations is 4.31 feet, which is reasonable given the level of geometric distortion present in the images. The average horizontal distance error is 3.76 feet, while the average vertical error is only 1.61 feet. This is understandable, since all observed rooftop lines are considered simultaneously when estimating the building height (vertical position), whereas the horizontal position of a rooftop vertex is primarily affected only by its two adjacent edges.

Projective Intensity Mapping. Backprojection of image intensities onto polygonal building model faces enhances their visual realism and provides a convenient storage mechanism for later symbolic extraction of detailed surface structure. Planar projective transformations provide a locally valid mathematical description of how surface structure from a planar building facet maps into an image. By inverting this

transformation using known building position and camera transformations, intensity information from each image is backprojected to "paint" the walls and roof of the building model. Since multiple images are used, intensity information from all faces is available, even though they are not all visible from any single view. Multiple intensity maps for each polygonal building facet are combined using knowledge of the sun angle and camera viewpoint to remove visual artifacts caused by shadows and occlusion. The resulting intensity mapped site model can then be rendered to predict how the scene will appear from a new view (Figure 5).

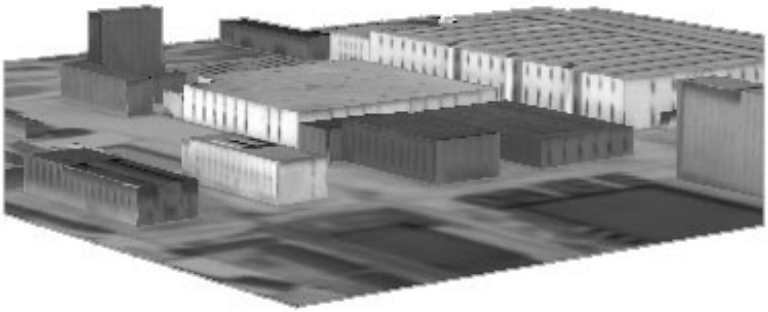

Fig. 5: Intensity mapped model rendered from a new view.

2.2. Site Model Extension

The goal of site model extension is to find unmodeled buildings in new images and add them into the site model database. The main difference between model extension and model acquisition is that the camera pose for each image can be determined via model-to-image registration using the current partial site model, whereas for initial model acquisition the pose must be supplied in some other way. Our approach to model-to-image registration involves two components: 1) model matching to determine correspondences between model features and image features, and 2) pose determination to determine the precise geometric relationship between the image and the scene.

The goal of **model matching** is to find the correspondence between 3D features in a site model and 2D features that have been extracted from an image; in our case this involves determining correspondences between edges in a 3D building wireframe and 2D extracted line segments from the image. To find this correspondence, we are using a model matching algorithm described in (Beveridge, 1992). The result of model matching is a set of correspondences between model edges and image line segments and an estimate of the transformation that brings the projected model into the best geometric alignment with the underlying image data.

The second aspect of model-to-image registration is precise **pose determination.** We are using a robust pose estimation procedure (based on a least median squares minimization procedure) described in (Kumar, 1994). The final results of pose determination are a set of camera pose parameters and a covariance matrix that estimates the accuracy of the solution.

Model Extension Example. The model extension process involves registering a current geometric site model with an incoming image, and then focusing on unmodeled areas to recover new buildings that have been recently built, that were previously unseen, or that for some other reason are not present in the site model database. We illustrate this process using the partial site model constructed in Section 2.1, and image J8 from the Radius Model Board 1 dataset.

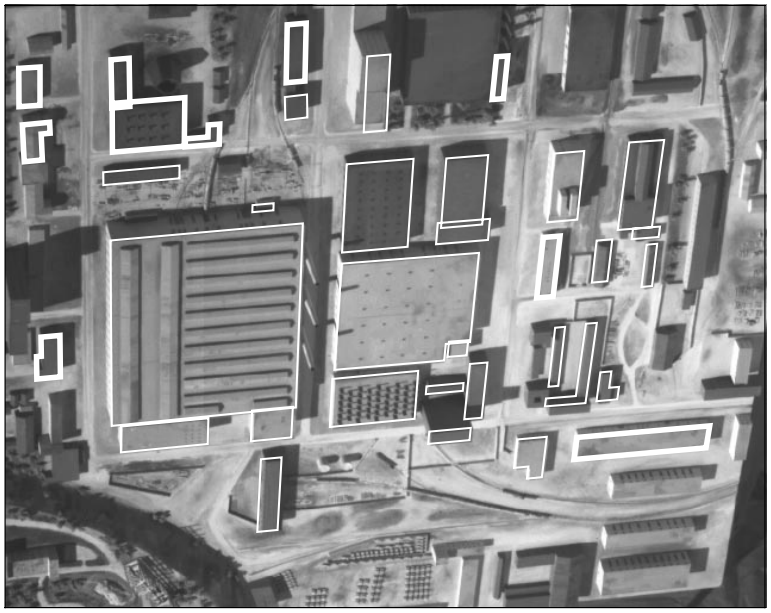

Fig. 6: Extended site model overlaid on J8 (Fig. 1).

Results of model-to-image registration of image J8 with the partial site model can be seen in Fig. 6, showing projected building rooftops from the site model (thin lines) overlaid on the image. Image areas containing buildings already in the site model were masked off, and the building rooftop detector was run on the unmodeled areas in the image, yielding 19 new rooftop hypotheses. The multi-image epipolar matching and constrained multi-image triangulation procedures from Section 2.1 were applied, again using images J1-J8, to verify these hypotheses and construct 3D volumetric building models. Only 10 hypotheses survived the verification and triangulation process. These were added to the site model database, to produce the extended model shown in Figure 6 (thick lines). The main reason for failure among building hypotheses that were not verified was that they represented buildings located at the periphery of the site, in an area which is not visible in very many of the eight views. If more images were used with greater site coverage, we expect that more of these buildings would be included in the site model.

3. Terrain Extraction

The geometric component of a site model consists not only of the building models and other cultural features but also an accurate model (digital elevation map or DEM) of the underlying terrain. The type of imagery of a site that can be expected in the RADIUS project can be characterized as being highly oblique, with widely separated views taken from (perhaps) different cameras at varying temporal intervals. In addition, the camera parameters (both extrinsic and intrinsic) may be unknown or only incompletely estimated. Even under the assumption of known camera parameters, these images present unique problems for correlation-based stereo reconstruction systems because of their oblique viewing geometry and the associated large base-to-height ratios[1].

When a disparity map is computed from widely separated images perspective distortion may result in a large number of false matches and poor reconstruction accuracy. For example, when the base-to-height ratio exceeds approximately 0.5, the performance of correlation-based matching algorithms begins to deteriorate, and when it becomes greater than 1, elevation errors caused by perspective distortion can become large. This

[1] For oblique geometries, the 'height' is the distance from the center of the camera baseline to a nominal point on the surface. In this case, the base to height ratio can vary considerably across the scene.

implies that the size of the correlation mask should be small to minimize the effects of perspective distortion. On the other hand, the size of the mask should be fairly large to provide increased robustness against random noise in the images. To develop algorithms that balance these competing factors we take advantage of the fact that pixels near the center of the correlation mask are less affected by perspective distortion.

Schultz (1994) has developed a correlation based stereo algorithm which incorporates several modifications to account for these effects. Briefly, these are:

(1) A weight is assigned to each element in the correlation mask that depends on its distance from the mask center. Gaussian weights are used with a variance that is either fixed or context dependent. By assigning position dependent weights, it is possible to place more emphasis on the central elements.

(2) The optimal match score is estimated from a series of match scores computed at subpixel disparity steps. By estimating the shape of the disparity function over a narrower disparity range, it is more likely that the match scores at the end of the intervals are statistically significant.

(3) An enhancement to standard multiresolution matching in which perspective distortion is iteratively removed from the images as the processing progresses from low to high resolution.

Fig. 7: ARPA/Martin Marietta UGV Demo C site reconstruction.

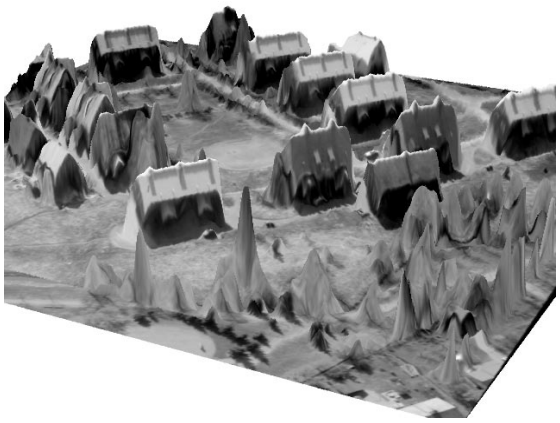

Fig. 8: ISPRS "FLAT" scene reconstruction.

At any resolution step, the disparity map is represented as the sum of an initial and incremental disparity map. The initial map (which is derived from the computed disparities at the previous level) is used to remove the perspective distortion at the current resolution level before the incremental disparity map is computed. This process of successive removal of perspective distortion makes it possible to match small features in images taken from widely varying viewpoints.

The resulting stereo analysis package has been tested on both synthetic and real images taken from widely disparate positions (that is, with high base to height ratios) and has been shown to be both accurate and robust in its reconstruction of elevation maps from two or more images (Schultz, 1994). Figure 7 shows a reconstruction of the ARPA Unmanned Ground Vehicle Demo C site at Martin-Marietta in Denver. The original

images were taken looking straight down, with a base-to-height ratio of 0.63. Figure 8 shows a portion of the reconstruction of the FLAT Test Dataset 3 of the ISPRS Working Group III/3 data set.

4. Summary and Future Work

A set of IU algorithms for automated site model acquisition and extension have been presented. The algorithms currently assume a generic class of flat roofed, rectilinear buildings. To acquire a new site model, an automated building detector is run on one image to hypothesize potential building rooftops. Supporting evidence is located in other images via epipolar line segment matching, and the precise 3D shape and location of each building is determine by multi-image triangulation. Projective mapping of image intensity information onto these polyhedral building models results in a realistic site model that can be rendered using virtual "fly-through"' graphics. To perform model extension, the acquired site model is registered to a new image, and model acquisition procedures are focused on previously unmodeled areas. In an operational scenario, this process would be repeated as new images become available, gradually accumulating evidence over time to make the site model database more complete and more accurate.

Several avenues for system improvement are open. One high priority is to add capabilities for detecting and triangulating peaked roof buildings. Another significant improvement would be extending the epipolar matching and triangulation portions of the system to analyze why a particular building roof hypothesis failed to be verified. There are many cases where the rooftop detector has outlined split-level buildings with a single roof polygon. This currently causes the subsequent epipolar verification procedure to fail, since all lines in the polygon are assumed to be at the same height. However, a careful analysis of the height histogram in these cases reveals it to be bimodal, meaning that some lines have been found to be at one height, while some occur at another. Automatic detection of these situations, followed by splitting of the rooftop hypothesis into two separate hypotheses, one for each roof level, would result in an improvement in system performance.

In the near future we plan to combine our symbolic building extraction procedures with the correlation-based terrain extraction system described in Section 3. The two techniques clearly complement each other: the terrain extraction system will be used to determine a digital elevation map upon which the volumetric building models will sit, and the symbolic building extraction procedures will be used to identify building occlusion boundaries where correlation-based terrain recovery can be expected to behave poorly. A tighter coupling of the two systems, where an initial digital elevation map is used to focus attention on distinctive humps that may be buildings, or where correlation-based terrain extraction techniques are applied to building rooftop regions to identify fine surface structure like roof vents and air conditioner units, may also be investigated.

Acknowledgments

This paper would not have been possible without the creativity, dedication, and hard work of the RADIUS team of graduate students: Yong Qing Cheng, Chris Jaynes, Frank Stolle, and Xiaoguang Wang. We would also like to acknowledge the software and technical support of Robert Heller and Jonathan Lim, the video wizardry of Fred Weiss, and the administrative support of Janet Turnbull and Laurie Waskiewicz.

References

Beveridge J.R., E.Riseman, "Hybrid Weak-Perspective and Full-Perspective Matching," *Proc. Computer Vision and Pattern Recognition*, Champaign, IL, 1992, pp. 432-438.

Boldt M., R. Weiss, E. Riseman, "Token-Based Extraction of Straight Lines," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 19, No. 6, 1989, pp. 1581-1594.

Collins R., A. Hanson, E. Riseman, Y. Cheng, "Model Matching and Extension for Automated 3D Site Modeling," *Proceedings of the ARPA Image Understanding Workshop,* Washington, DC, April 1993, pp. 197-203.

Gerson D. "RADIUS : The Government Viewpoint," *Proceedings of the DARPA Image Understanding Workshop,* San Diego, CA, January 1992, pp. 173-175.

Huertas A., C. Lin, R. Nevatia "Detection of Buildings from Monocular Views of Aerial Scenes using Perceptual Grouping and Shadows," *Proceedings of the ARPA Image Understanding Workshop,* Washington, DC, April 1993, pp. 253-260.

Jaynes C., F. Stolle, R. Collins, "Task Driven Perceptual Organization for Extraction of Rooftop Polygons," *Proceedings ARPA Image Understanding Workshop,* Monterey, CA, November 1994, pp. 359-365.

Kumar R., A. Hanson, "Robust Methods for Estimating Pose and Sensitivity Analysis," *CVGIP: Image Understanding,* Vol. 60, No. 3, November 1994, pp. 313-342.

Roux M., D. McKeown, "Feature Matching for Building Extraction from Multiple Views," *Proceedings ARPA Image Understanding Workshop,* Monterey, CA, November 1994, pp. 331-349.

Schultz H. "Terrain Reconstruction from Oblique Views," *Proceedings ARPA Image Understanding Workshop,* Monterey, CA, November 1994, pp. 1001-1008.