A System for Building Detection from Aerial Images

R. Nevatia, C. Lin and A. Huertas*

Institute for Robotics and Intelligent Systems
Computer Science Department
University of Southern California
Los Angeles, California 90089-0273
e-mail:{nevatia,chungan,huertas}@usc.edu

Abstract

We describe a method for detecting rectilinear buildings and constructing their 3-D shape descriptions from a single aerial image of a general viewpoint. 2-D roof hypotheses are generated from linear features by perceptual grouping. Good hypotheses are selected and then verified by computing wall and shadow evidence for them, which also provide the height information for the buildings. A 3-D reasoning process resolves conflicts among hypotheses in 3-D space. Results from several images can be integrated at a high level. An interactive system allows efficient editing of results by making use of the analysis performed by the automatic system; it also allows for some initial preparation of the data to improve results of the automatic system. Some results and their evaluation are included.

1 Introduction

Detection and description of buildings from aerial images remains an active area of research; an excellent collection may be found in (Grun *et al*, 1995), some more recent work is described in (Fua, 1996; Henricsson *et al*, 1996; Weidner, 1996). Many different kinds of inputs, such as stereo images and range images, have been used. In this paper, we focus on the use of a single image. Lack of direct 3-D information makes use of a single image more difficult, but they are attractive due to the ease with which they can be obtained. It is also our experience that many of the processes involved in single image analysis are also required for multiple image analysis (Noronha & Nevatia, 1997). Our system is restricted to rectilinear shapes with flat roof but allows for *oblique* (*i.e*. non-nadir) views. It also allows for efficient human interaction where the results of the automated system can be improved with relatively few and simple interactions before and after automated processing.

Our basic approach is to use the geometric and projective constraints to make hypotheses for the presence of building roofs from the low-level features and to verify by using available 3-D cues. As, our system is restricted to rectilinear buildings with flat roofs, they project into compositions of parallelograms. We use shadow and wall evidence to verify

* This research was supported, in part, by the Advanced Research Projects Agency of the United States Department of Defense under grant No. F49620-95-1-0457, monitored by the Air Force Office of Scientific Research and in part by a subgrant from Purdue University under Army Research Office grant No. DAAH04-96-10444.

and reconstruct 3-D structures. The system also analyzes the 3-D structures to resolve conflicts among them. A summary of this approach and some results are given in section 2. In section 3, we describe how to integrate results from multiple images. We have also developed a methodology for efficient human interactions with this system, for the purposes of editing the results or to provide some guidance prior to automatic analysis. Many errors of the automated system can be corrected (or prevented) by relatively simple user interactions. These methods and some results are described in section 4.

2 Monocular Building Detection

This system consists of several layers. At first, linear edges are detected from the image. Next parallelogram hypotheses are formed that are consistent with the projective constraints given by the viewing geometry. Promising hypotheses are selected based on some 2-D and local 3-D evidence. The selected hypotheses are verified by searching for 3-D cues using wall and shadow evidence. The verified hypotheses are examined for mutual containment and overlap and a non-conflicting set is selected which provides 3-D building models. Each model is also assigned a confidence level, computed from combinations of lower-level evidence.The early stages of this process, including hypotheses formation, selection and verification by using wall and shadow evidence have been described previously (Lin & Nevatia, 1995) The current system uses an improved hypotheses generation system and various modifications have been made to the selection and verification steps, however, the general approach remains the same and we omit further discussion of them; details may be found in (Lin, 1996)

2.1 Containment and Overlap Analysis

The wall and shadow verification processes examine each hypothesis individually and do not analyze the relationships among them. Thus, some verified hypotheses might overlap with or contain others. At this stage, having knowledge of 3-D allows us to check that two inconsistent structures do not occupy the same 3-D space.

When one hypothesis is contained in the other, two cases can occur, as shown in Figure 1 (a) and (b). In the first case, a contained hypothesis does not share any side with the containing hypothesis; here the latter is likely to be a superstructure on top of the former. We also adjust the height of the superstructure to be relative to that of the base. In the second case, the two hypotheses share some common boundaries. If the two have different heights, we consider them to be in conflict and remove the one with the lower confidence. If they have the same height, and share boundaries, the containing hypotheses is removed unless there is strong wall and shadow evidence for its *non-shared* roof boundaries.

The overlap cases also fall in two cases. If the overlapping hypotheses have the same height, it is not considered a conflict and both are retained as shown in Figure 2 (a). When two roof hypotheses with different building heights overlap, they conflict in 3-D space and the one with weaker evidence is removed. Note that it is possible for two building hypotheses to have overlapping footprints even if the roof hypotheses don't overlap as shown in Figure 2 (b).

2.2 Building Interaction Analysis

When nearby buildings (or their parts) occlude another, they can affect the evaluation of wall and shadow evidence of the occluded objects. In Figure 3, a part of shadow evidence

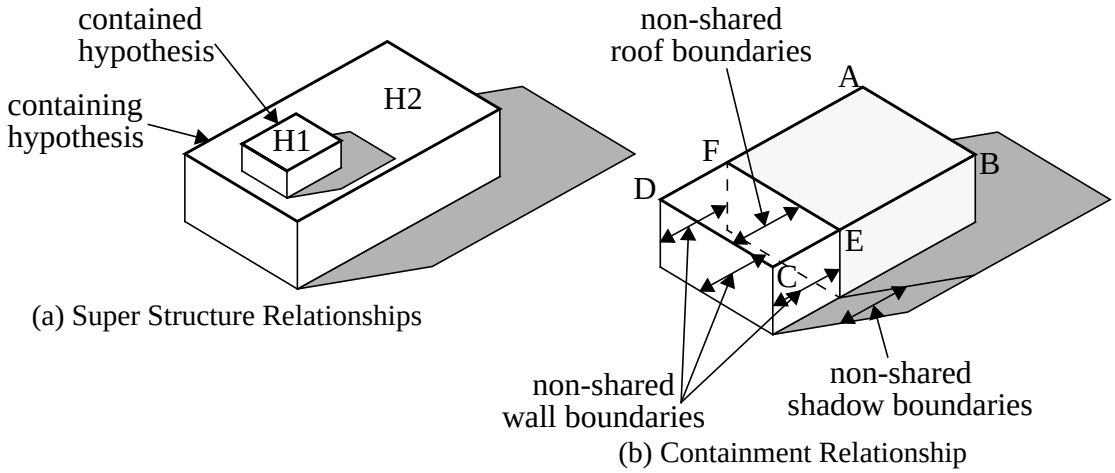


Figure 1 Containment analysis

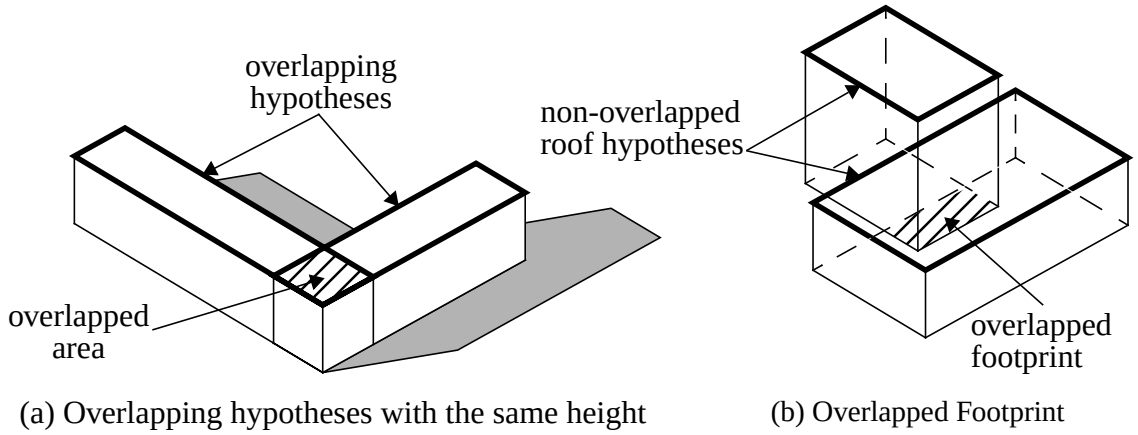


Figure 2 Overlap analysis

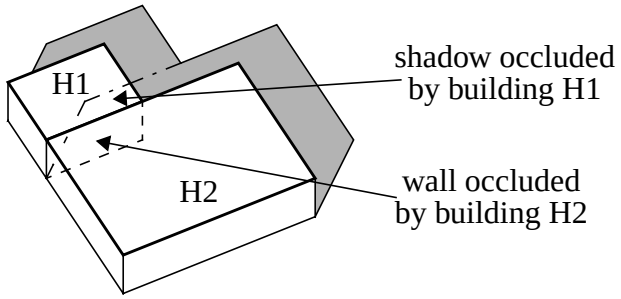
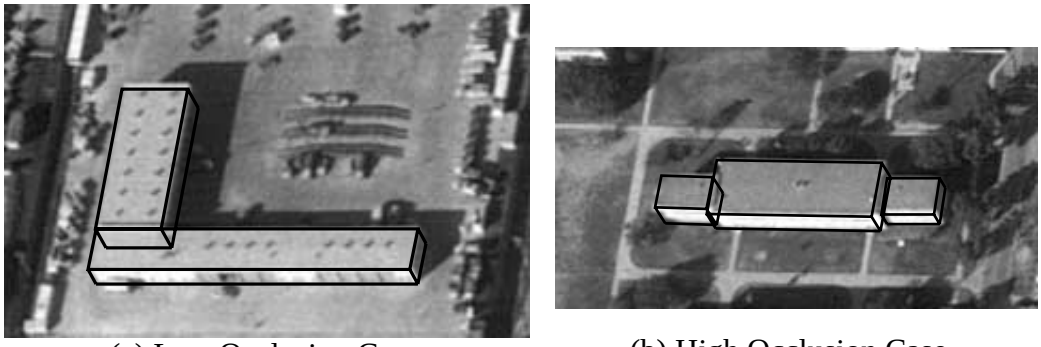


Figure 3 Evidence occluded by other buildings

of hypothesis H2 is not visible because of occlusion from H1, and a part of wall evidence of H1 is blocked by H2. Say that there is enough evidence to support H2 but not H1. However, once H2 has been verified, the interaction analysis process reevaluates H1 by examining the evidence in the non-occluded area and verifies. Verification of H1 causes confidence of H2 to be increased in turn. In general, this process may need to be iterated until no changes occur.

Figure 4 (a) shows the detected wire frames of the two verified hypotheses for an example. Both were detected initially, but the confidences of both parts were increased after interaction analysis. Figure 4 (b) shows an example where only two of the three structures of the building in the scene are detected initially; the left part is not verified as its wall and shadow boundaries are occluded by the middle part. After analyzing the occlusion, the left structure is recovered and the confidence of the middle structure is increased also, because it is occluded partly by the right structure.



(a) Low Occlusion Case (b) High Occlusion Case

Figure 4 Results on two examples

Figure 5 shows the results for a larger window (of an image from Fort Hood, Texas), containing several buildings in a complex scene viewed obliquely. As can be seen, most buildings are detected accurately. Only one has an obvious height error. No false positives are detected. Two buildings are not detected. The one in the bottom left area is not detected because of severe occlusions by nearby trees. The other is the bright building with two wings; mutual occlusions between the two parts cause both to be not verified. The two C-shaped buildings are detected but the descriptions are not accurate. The middle parts of the C-shaped buildings are not hypothesized, because there is no other evidence besides a pair of parallel lines. A part of the building in the top middle area is not detected due to occlusion and low height. There are also some structures attached to the four buildings on the left side of this image that are not detected, largely because of their low height.

It takes 877.58 seconds (14.62 minutes) to process image in Figure 5 on a SUN Sparcstation 20 (using the RCDE environment with all code being written in COMMONLISP). The most time consuming process, at 63% of the total, is that of parallelogram formation. The "higher-level" processes of hypothesis selection, verification and 3-D analysis take only a small fraction of the total time. The execution times are generally linearly proportional to the number of lines that are found in an image.

There are many ways to measure the quality of the results. Following (McGlone & Shufelt, 1994; Shufelt & McKeown, 1993), we use the following five measurements: Detection Percentage (100*TP/(TP+TN)); Branch Factor (100*FP/(TP+FP); Correct Building Pixels Percentage; Incorrect Building Pixels Percentage and Correct Non-Building Pixels Percentage. The first two measurements are calculated by making a comparison of the manually detected buildings and the automated results, where TP (True Positive) is a building detected by both a person and the program, FP (False Positive) is a building detected by the program but not a person, and TN (True Negative) is a building detected by a person but not the program. A building is considered detected if a part of the building is detected. The accuracy of shape is determined by counting correct building and non-building pixels. These quality measurements are rather consistent for most of the images

Figure 5 Results with multiple buildings in an oblique image of a complex scene complex scene

processed. Average approximate values over several examples are: Detection rate, 70%; Branch Factor,6%; Correct Buildings Pixels, 70%; Incorrect Building Pixels, 8%; and Correct Non-Building Pixels, 99%.

Another method of evaluation is to examine the number of true and false positives as a function of the hypothesis confidence; Figure 6 shows results for 12 windows, each containing several buildings. It should be noted that there are no false alarms for high confidence values, thus a clear choice is available between higher detection rates and lower false alarms.

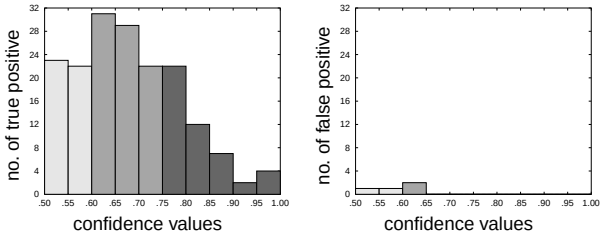


Figure 6 Distribution of true and false positives

3 Integration of Results from Multiple Views

Results from several views can be integrated to get more robust results. Some structures may be more reliably detected in some views depending on conditions, such as the viewing direction, the illumination direction, and the building orientation. The approach is not to perform complete stereo analysis but to merge the higher level structures only. Hypotheses in one view are projected into the other views (knowledge of relative camera geom-

etry is assumed) and verified as any other hypotheses. If a building is correctly detected in one view, supporting evidence for it should be found in other views. On the other hand, if an incorrect hypothesis has been made, it should be unlikely to find much supporting evidence from other views. Based on this observation, a better decision can be made by integrating all evidence from all available views. A building could be verified individually in more than one view resulting in multiple hypotheses for the same structure. An overlap analysis is performed and the hypothesis with the highest combined confidence is retained. A set of 3-D models is created from the list of retained hypotheses which can be projected into any view for visualization. The situation when none of the hypotheses from any of the views is correct is not handled.

Figure 7 shows an example of integrating the results from two views of a building. The building is composed of three structures. The main structure in the middle is detected in the left image only and the right wing in the right image only. The left wing is detected in both images. After integration all three parts are verified and shown reprojected in the two views in Figure 7. Similar improvements are obtained for other examples, such as the one shown in Figure 5, but are not included for lack of space.

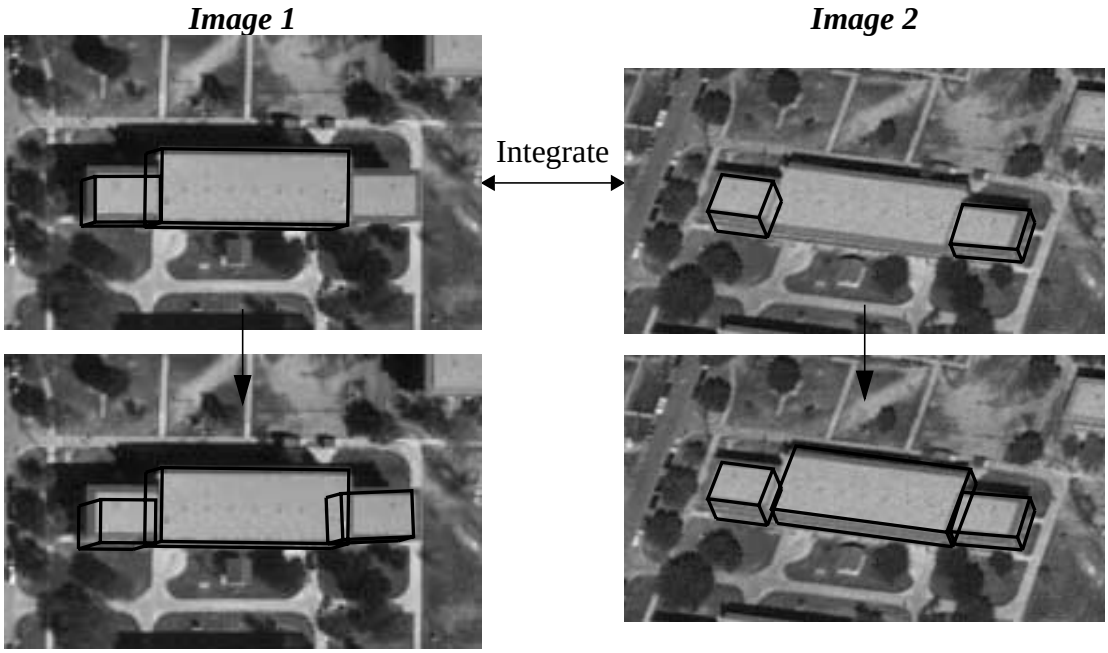


Figure 7 Integration of results from multiple views

4 Interactive Editing and Preparation

While the automatic system performs well under many conditions, there are also several situations that cause it to fail to find a building or to find a correct description of the building. An interactive system has been developed to correct these errors. Many interactive systems for building detection have been developed in the past (Heller *et al*, 1996, Neuwnschwander *et al*, 1994). One different aspect of our approach is to use the partial results of the automatic analysis to reduce the required interactions from the user.

There are two classes of interaction possible in this system. The first is a qualitative (or initial) interaction, the second is a quantitative (or corrective) interaction. The input for

the qualitative step is simply an indication of the problem, such as a missing building and its approximate location (indicated by positioning the cursor somewhere inside the roof area). This causes the automatic system to re-examine all of the roof hypotheses generated earlier by the system and select the one with the highest score. In many cases, just this interaction results in a correct building to be detected; it was not previously output because its score was too low. A version of this system also allows to qualitatively specify the probable cause of failure (such as a dark area) which can be used in selecting the best hypotheses (see (Heuel & Nevatia, 1995) for details)

If the building detected by qualitative interaction is not correct (in dimensions, location or orientation), quantitative, corrective interactions are needed. Two ways of correcting the hypothesis are available. The user can choose to associate extracted edges and corners with a part of the building model. For example, a roof-side of the building can be specified by an edge extracted in the image. Then this edge is added to the current hypothesis (by replacing the nearest edge of the current hypothesis). Such interactions are facilitated by mouse-sensitive features of the RCDE (Strat, *et al.*, 1992). After each corrective interaction, the system forms a new parallelogram hypothesis and looks for new edges, shadow and wall evidence to support the new hypothesis. Therefore, it is possible that, after a manual correction of a roof-boundary, the wrong building height is also corrected automatically.

The user can also adjust the roof-parallelogram by dragging sides with the mouse, rotating or translating the whole model. Changes can only be made within the constraints of the building model, for example opposite sides remain parallel The extraction of a ground corner or edge (shadow corner or edge) determines the building height. These interactions are similar to a completely manual system.

We find that, in conjunction with the automatic system, relatively few and simple user interactions yield correct models. In order to complete the building detection task for the example of Figure 5 (14 buildings made of 29 rectangular structures), the following user interactions were required: two of the detected structures required 1 quantitative correction; fifteen qualitative interactions were required to select hypotheses for structures not detected; of these, 2, 4, 8 and 1 structures needed 0, 1, 2 and 3 quantitative interactions respectively.

Figure 8 shows several other building models (processed in four separate windows as shown). For this example, 38 of the structures (a rectangular building or a rectangular part) required no interactions. 27 structures were detected but required some corrective interactions (20 required one, 4 required two, and 3 required three interactions). 10 undetected structures were correctly detected with just one qualitative interaction. Remaining 29 undetected structures required 1 qualitative and 1, 2 or 3 quantitative corrections (13 required one, 11 required two, five required three). In nearly all of the cases where corrective interactions are required, only corrections of the sides and height are necessary.

Initial Preparation

The performance of this system can be improved by providing the automated system with some information *prior* to its computations. In normal operation, a user would need to select images and image windows to be processed anyway. It is a relatively simple task for the user to also provide an indication of where the desired buildings are by simply point-

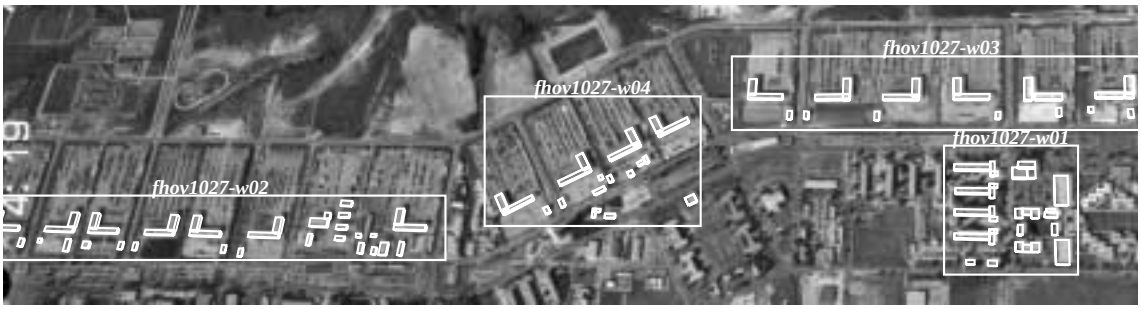


Figure 8 Edited results for four windows from a large image.

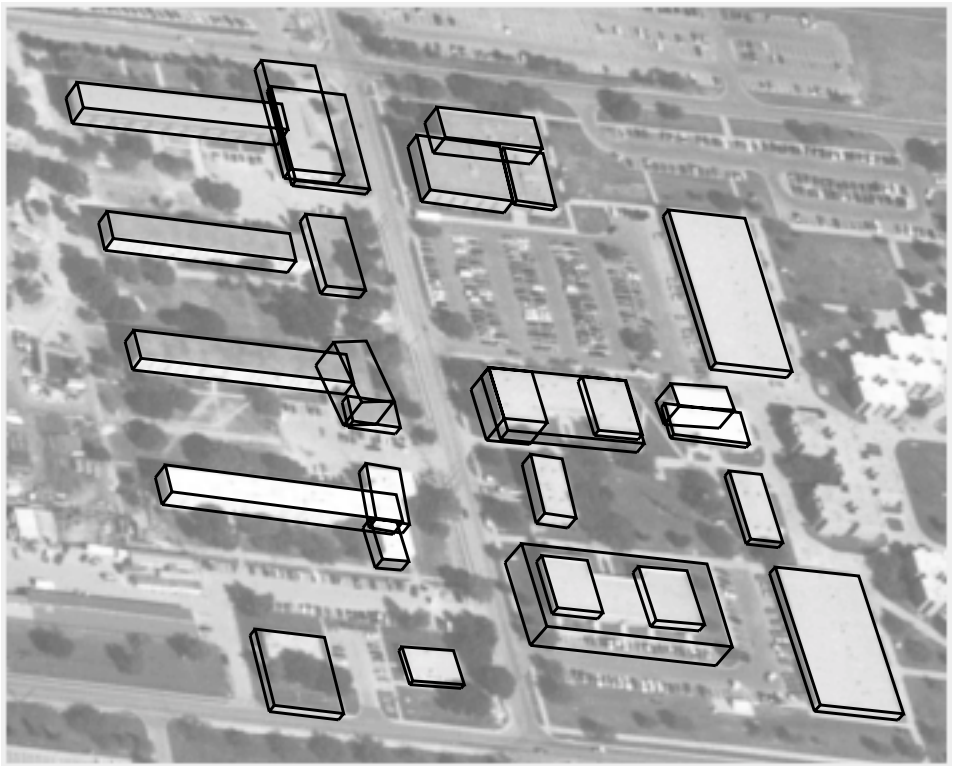
ing and clicking somewhere in the interior of such buildings and could be considered to be part of the preparation of the image data to be processed. Such input is viewed by the system in the same way as a qualitative input later, *i.e.* a building hypotheses with the highest score is always selected. Also, no buildings are output in areas not indicated by the user. This simple input greatly improves the performance of the automatic system, increasing its detection rate while reducing or eliminating the false alarms.

Automatic results obtained by selecting the locations in the image of the 29 roof components is shown in Figure 9 (a). All roof components are detected but 14 require quantitative corrections. Eleven of these required 1 correction and the other 3 required two corrections. Figure 9 (b) shows the completed model. For this example, the number of structures requiring interaction is the same with initial preparation or without (as in Figure 5). However, the former case requires fewer corrections and takes about half as much time.

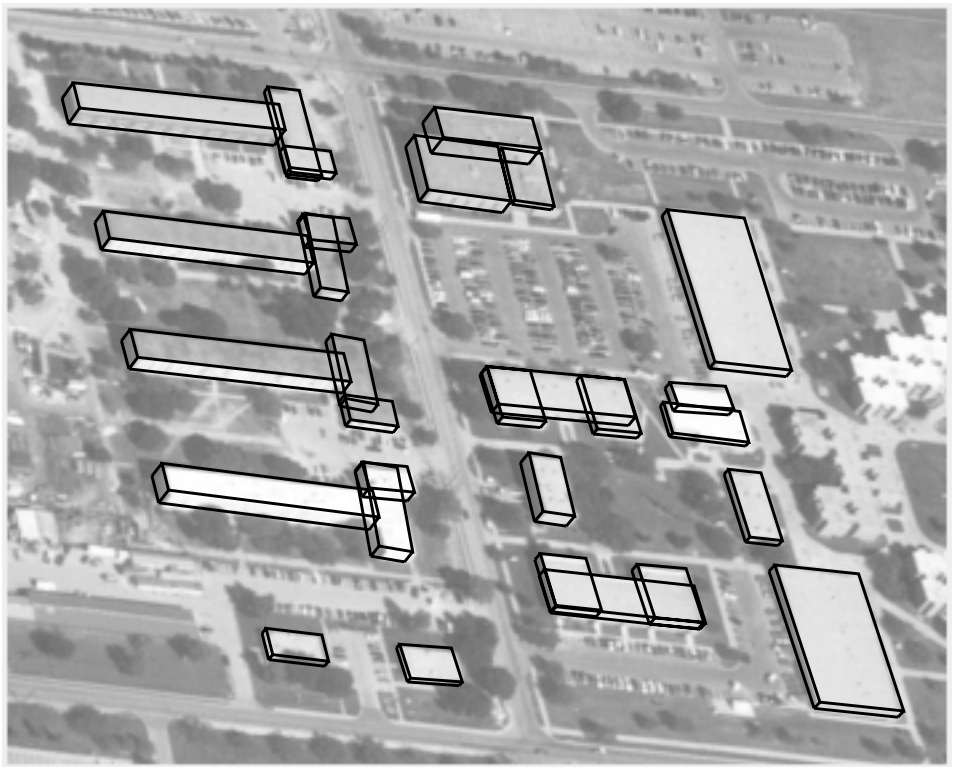
We have attempted a preliminary quantitative evaluation of this approach by comparing to the time required to construct building models in a given window by using traditional modeling tools, such as those supplied with the RCDE (Heller *et al*, 1996). For the interactive system, we only include the time needed for initial and editing steps but not the computation time for the automated step (as it can be executed *off-line* and does not require user's attention). The results on three windows from the Ft. Hood image data set are summarized in Table 1. t_m, t_i, and t_e denote time in minutes for manual, interactive and editing processes respectively. The L and I shape data are not shown due to lack of space but are similar to those shown in Figure 8. The "complex" window is the one shown in Figure 9. These results compare very favorably with the manual process that would be needed for the same task. As shown in the table, the speed-ups range from a factor of about 7 to about 11, the lower number being for more complex shapes where more user interactions are required. These results are preliminary and have not been tested on large data sets with different kinds of operators (all times are for A. Huertas). Nonetheless, we believe that the indicated speedups are significant and offer potential for use in a practical system.

Table 1: Time Comparison (time in minutes)

Image Description	# of Buildings	# of Boxes	t_m	t_i	t_e	# of Boxes edited	$\frac{t_m}{t_i + t_e}$
L-shape	8	12	8	0.2	0.5	2	11.4
I-shape	19	35	28	0.6	2.5	4	9.0
Complex	14	29	75	0.4	10	14	7.2



(a) Automated results with initial preparation



(b) Completed 3-D model

Figure 9 Results obtained with initial preparation and user interaction.

5 Conclusion

We have summarized our approach to automated building detection and description using a single intensity image, to integrating results of several such images, and of designing interactive tools for preparing data and editing results. The range of shapes to which these techniques can be applied remains limited but we believe that they cover a useful and significant subset. The system has been ported to some user laboratories for further testing and evaluation.

Acknowledgments

Stephan Heuel, of the University of Bonn, developed the original version of the interactive system as a visiting researcher in our laboratory. Bill Bremner, of Lockheed Martin Corporation, suggested user provided interactions before automatic processing. Jim Pearson, of GDE Systems Inc., suggested the methodology for comparing time performance of interactive and manual systems.

References

Fua P. (1996) *Model-Based Optimization: Accurate and Consistent Site Modeling*, in Proceedings of the 18th SPRS Congress, Comm. III, WG 2, Vienna, Austria, pp. 222-233.

Grun A., O. Kubler, P. Agouris (1995) editors, *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, Virkhauser Verlag, Basel, pp. 199-210.

Henricsson O., F. Bignone, W. Willuhn, F. Ade, O. Kubler, E. Baltsavias, S. Mason, A. Grun (1996) *Project AMOBE: Strategies, Current Status and Future Work*, in Proceedings of the 18th SPRS Congress, Comm. III, WG 2, Vienna, Austria, pp. 321-330.

Heller A., P. Fua, C. Connolly, J. Sargent (1996) *The Site-Model Construction Component of the RADIUS Testbed System*, Proceedings of the DARPA Image Understanding Workshop, Palm Springs, California, pp. 345-355.

Heuel S., R. Nevatia (1995) *Including Interaction in an Automated Modeling System*, Proceedings of the IEEE Symposium on Computer VIsion, Coral Gables, Florida, pp. 383-388.

Lin C., A. Huertas and R. Nevatia (1995) *Detection of Buildings from Monocular Images*, in A. Grun, O. Kubler, P. Agouris, editors, Automatic Extraction of Man-Made Objects from Aerial and Space Images, Virkhauser Verlag, Basel, pp. 125-134.

Lin C. (1996) *Perception of 3-D Objects from an Intensity Image using Simple Geometric Models* Ph.D. Dissertation, Computer Science Department, University of Southern California.

McGlone J., and J. Shufelt (1994) *Projective and Object Space Geometry for Monocular Building Extraction* IEEE Proceedings of Computer Vision and Pattern Recognition, 54-61.

Noronha S., R. Nevatia (1997) *Building Detection and Description from Multiple Aerial Images* to appear in Proceedings of IEEE Computer Vision and Pattern Recognition Conference, San Juan, Puerto Rico.

Shufelt J., D. McKeown (1993) *Fusion of Monocular Cues to Detect Man-Made Structures in Aerial Imagery* Computer Vision, Graphics and Image Processing, 57(3): 307-330.

Strat T., L. Quam, J. Mundy, R. Welty, W. Bremner, M. Horwedel, D. Hackett, A. Hoogs (1992) *The RADIUS Common Development Environment* Proceedings of the 1992 DARPA Image Understanding Workshop, San Diego, California, 215-226.

Weidner U. (1996) *An Approach to Building Extraction from Digital Surface Models* Proceedings of the 18th SPRS Congress, Comm. III, WG 2, Vienna, Austria, pp. 924-929.