

Knowledge-Based Building Detection and Description: 1997-1998

R. Nevatia and A. Huertas

Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, California 90098
<http://iris.usc.edu>

Abstract

An overview of a system for building detection and description using multiple views is given and its performance is evaluated. Use of cues from sensors such as IFSAR is included and helps improve the performance substantially. An option for user assisted operation is also described which is able to produce highly accurate models with relatively small efforts from a user.

1 Introduction and Overview

The goal of this project is to develop feature extraction methods for automated population of geospatial databases (APGD) with particular focus on buildings. The task includes detection, delineation and description of 3-D buildings. Buildings are objects of obvious importance in urban environments and accurate models for them are needed for a number of battlefield awareness tasks such as mission planning, mission rehearsal, tactical training and damage assessment. Other applications include intelligence analysis for site monitoring and change detection.

The problem of 3-D feature extraction has many sources of difficulties including those of segmentation, 3-D inference and shape description. Segmentation is difficult due to the presence of large number of objects such as roads, sidewalks, landscaping, trees and shadows near the buildings and due to presence of texture on building surfaces. 3-D information is not explicit in an intensity image; its inference from multiple images requires finding correct corresponding points or features in two or more images. Direct ranging techniques such as interferometric synthetic aperture radar (IFSAR

[Curlander & McDonough, 1991, Jakowatz et al., 1996]) can provide highly useful 3-D data though the data typically has areas of missing elements and may contain some points with grossly erroneous values. Once the objects have been segmented and 3-D shape recovered, the task of shape description still remains. This consists of forming complex shapes from simpler shapes that may be detected at earlier stages. For example, a building may have several wings, possibly of different heights, that may be detected as separate parts rather than one structure initially.

The approach used in this effort is to use a combination of tools: reconstruction and reasoning in 3-D, use of multiple sources of data and perceptual grouping. Context and domain knowledge guide the applications of these tools. Context comes from knowledge of camera parameters, geometry of objects to be detected and illumination conditions (primarily the sun position). Some knowledge of the approximate terrain is also utilized. The information from sensors of different modalities, such as IFSAR and EO (electro-optical), is fused not at pixel level but at higher feature levels. Our approach also allows for integration of information from multi-spectral images though this is not being pursued actively as part of the described effort. Other approaches recently reported [Hoepfner et al., 1997] use the IFSAR data to fit models of roof surfaces at regions of interest.

The described system is limited to buildings with rectilinear shapes. Most of our work has been on buildings with flat roofs but the system can also handle buildings with symmetrical slanted roofs (gables). It is assumed that camera models are given and approximated by orthographic projection locally, that the ground is flat with known height and that the sun position is given (computable from latitude, longitude and time-of-day). Multiple images are *not* assumed to have been taken at the same time. The multi-view system is described briefly in section 2.

* This research was supported in part by the Defense Advanced Research Projects Agency under contract DACA76-97-K-0001 and monitored by the Topographic Engineering Center of the U.S. Army.

Incorporation of cues from IFSAR is described in section 3 and a comparative system evaluation is given in section 4. A more detailed description of cue analysis is given in [Huertas et al., 1998].

A user can interact with the described system, either to edit the results of the automatic system or to provide cues for it. The aim is to make the user input efficient, requiring much less effort than would be necessary for conventional interactive systems which largely take care only of geometric computations and bookkeeping. The assisted system is described briefly in section 5 with more details given in [Li et al., 1998].

2 Multi-view System

A number of systems that use multiple views have been described in the literature [Jaynes et al., 1997; Collins et al., 1998; Roux & McKeown, 1994]. The system described in this paper derives from an earlier system described in [Noronha & Nevatia, 1997]; a block diagram is shown in Figure 1. The approach is basically one of hypothesize and verify. *Hypotheses* for potential roofs are made from fragmented lower level image features. The system is hierarchical and uses evidence from all the views in a non-preferential, order-independent way. Promising hypotheses are *selected* among these by using relatively inexpensive evidence from the rooftops only. The selected hypotheses are then *verified* by using more reliable global evidence. The verified hypotheses are then examined for overlap which may result in either elimination or in merging of them. Cues from a depth map (such as IFSAR or a DEM) can be incorporated at the hypotheses formation, selection or verification stages.

This system is designed for rectilinear buildings; complex buildings are decomposed into rectangular parts. Flat rooftops thus project to parallelograms in the images (the projection is nearly orthographic over the scale of a building), gables project to a pair of parallelograms sharing a side. Lines, junctions and parallel lines are the basic features used to form roof hypotheses. Flat roof hypotheses are formed a pair of parallel lines and *U* structures (*U*s represent three sides of a parallelogram). Gable hypotheses are formed from a *triple* of parallel lines. Closed hypotheses are formed from these features by using the best available image lines if any, else closures are synthesized from the ends of the parallel lines.

Three-D roof hypotheses could be inferred from the 2-D hypotheses by using line matches; however, the line matches are often not necessarily unique. Instead, an estimate for the heights of roof lines is made by conducting a search. For a flat roof only a

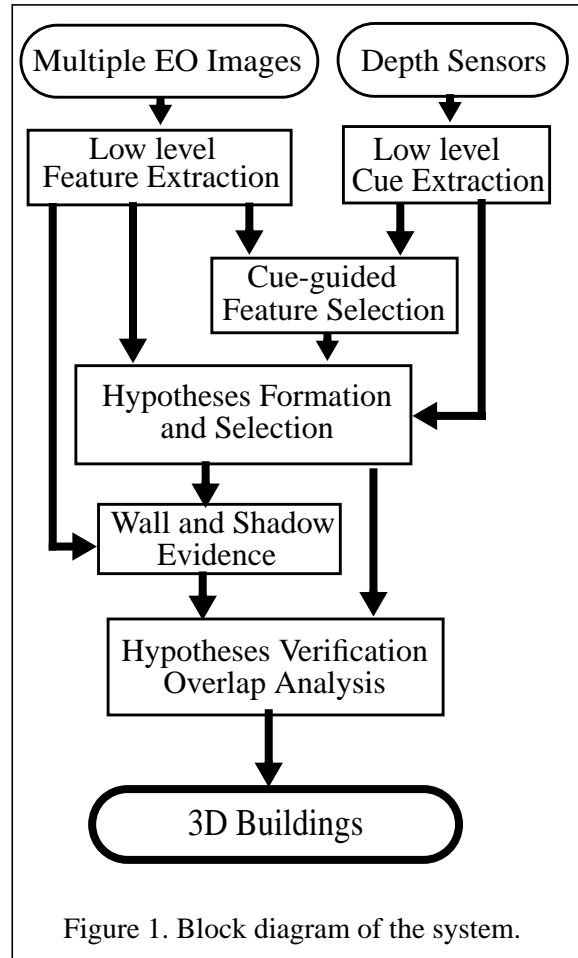


Figure 1. Block diagram of the system.

single height needs to be determined, for the symmetric gables we need to find two heights. For each height estimate, the corresponding 3-D hypotheses is projected in each other view and line evidence for each projection is computed. The evidence consists of the sum of the lengths of the supporting segments. The heights with the best evidence are selected.

The hypothesis formation process is rather liberal and a large number of hypotheses are typically formed at this stage. A smaller set is *selected* using the underlying image evidence for the roof hypotheses. Positive evidence comes from lines near the projected hypotheses, negative evidence comes from lines crossing the hypotheses. A coarse analysis is also applied to select among overlapping hypotheses. Currently, the selection process is applied to the flat roof cases only.

The next step is to *verify* whether the selected hypotheses have additional evidence for corresponding to being buildings. This evidence is collected from the roof, the walls and the shadows that should be cast by the building. Since the hypotheses are represented in 3D, deriving the projections of the

walls and shadows cast, and determining which of these elements are visible from the particular view point is possible. These in turn guide the search procedures that look in the various images for evidence of these elements among the features extracted from the image. A score is computed for each evidence element.

Each of the collected evidence parameters is composed of smaller pieces of evidence. A critical question is how to combine these small pieces of evidence to decide whether a building is present or not and how much confidence should be put in it. A variety of methods for this are available such as linear weighted sums of components, decision trees, certainty theory, neural networks and statistical classifiers. The results shown in this paper use a *decision tree* classifier. We are also investigating use of a Bayesian classification approach in a separately funded project.

After verification, several overlapping verified hypotheses may remain. Only one of the significantly overlapping hypotheses is selected. The overlap analysis procedure examines not only the evidence available for alternatives but also separately the evidence for components that are not common.

The system currently detects flat and gabled roofed buildings separately. These are shown in Figure 2 and Figure 3 respectively the flat roofed and gabled roofs detected for the McKenna MOUT site at Ft. Benning, GA, by using two stereo images.

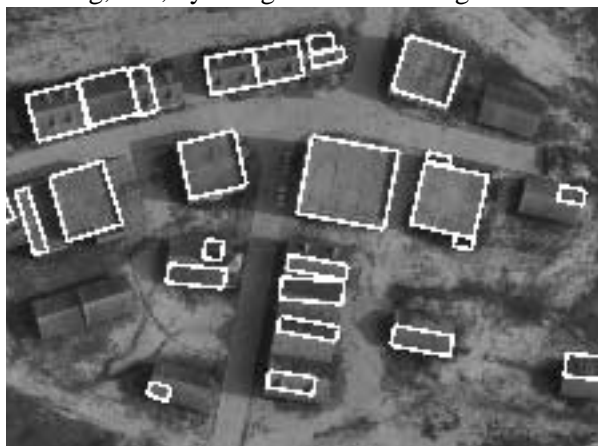


Figure 2. Flat roofed building components extracted using EO images only.

Figure 4 shows the combination of these two results after applying overlap analysis to eliminate conflicts. The results show that all the buildings are detected, at least in part, though not all are completely accurate. Some portions of gable roofs are detected as flat and vice-versa. The combination process that analyses the overlap between these is in a prelimi-

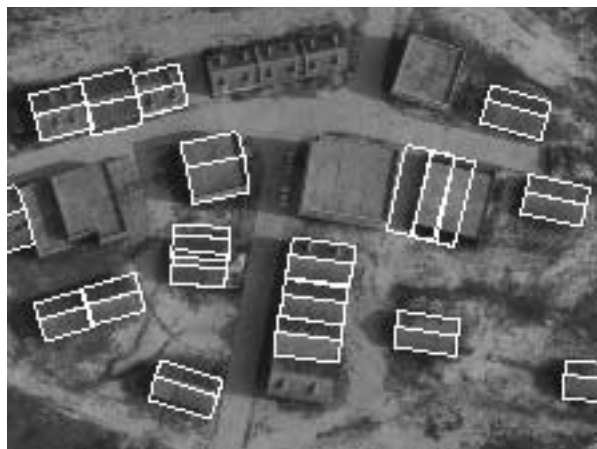


Figure 3. Gabled roofed building components extracted using EO images only.

nary stage of development. The roof on the lower right is detected as both a gable and a flat roof but the flat hypothesis dominates, incorrectly, in this case. The gabled portions on the upper center are not detected as gabled but as flat roofs. There are two small flat “false alarms” (*i.e.* buildings found where there are none) adjacent to the flat building in the center of the right part of the image. Another false alarm is present between this and the adjacent building. These are due to the walls that extend beyond the building sides.

3 Cues From IFSAR and other DEMs

The performance of the building detection and description system can be greatly improved if a source of direct range information is available. IFSAR data has started to become available in recent years. In addition to reflectivity information, also contains information of 3-D points in a scene.

However, the resolution of the IFSAR images is more limited and many wrong values are present due to the reflective properties of the surface material in the radar spectrum. Thus, it is preferable to use IFSAR for detection and panchromatic images for accurate delineation. Cues from IFSAR data can be used in a number of ways: in selecting areas to process where buildings may be present, in eliminating certain building hypotheses, and in adding confidence to the presence of buildings.

IFSAR data is given in the form of three images, called the **mag**, **dte** and **cor** images corresponding to the reflected magnitude, digital terrain elevation and phase correlation respectively. For certain kinds of sensors such as a searchlight mode Scandia sensor, cues for buildings can be derived from the **dte** data alone. A **dte** image for the Ft. Benning site is shown in Figure 5. Regions that may correspond to buildings, shown in Figure 6, are derived by con-

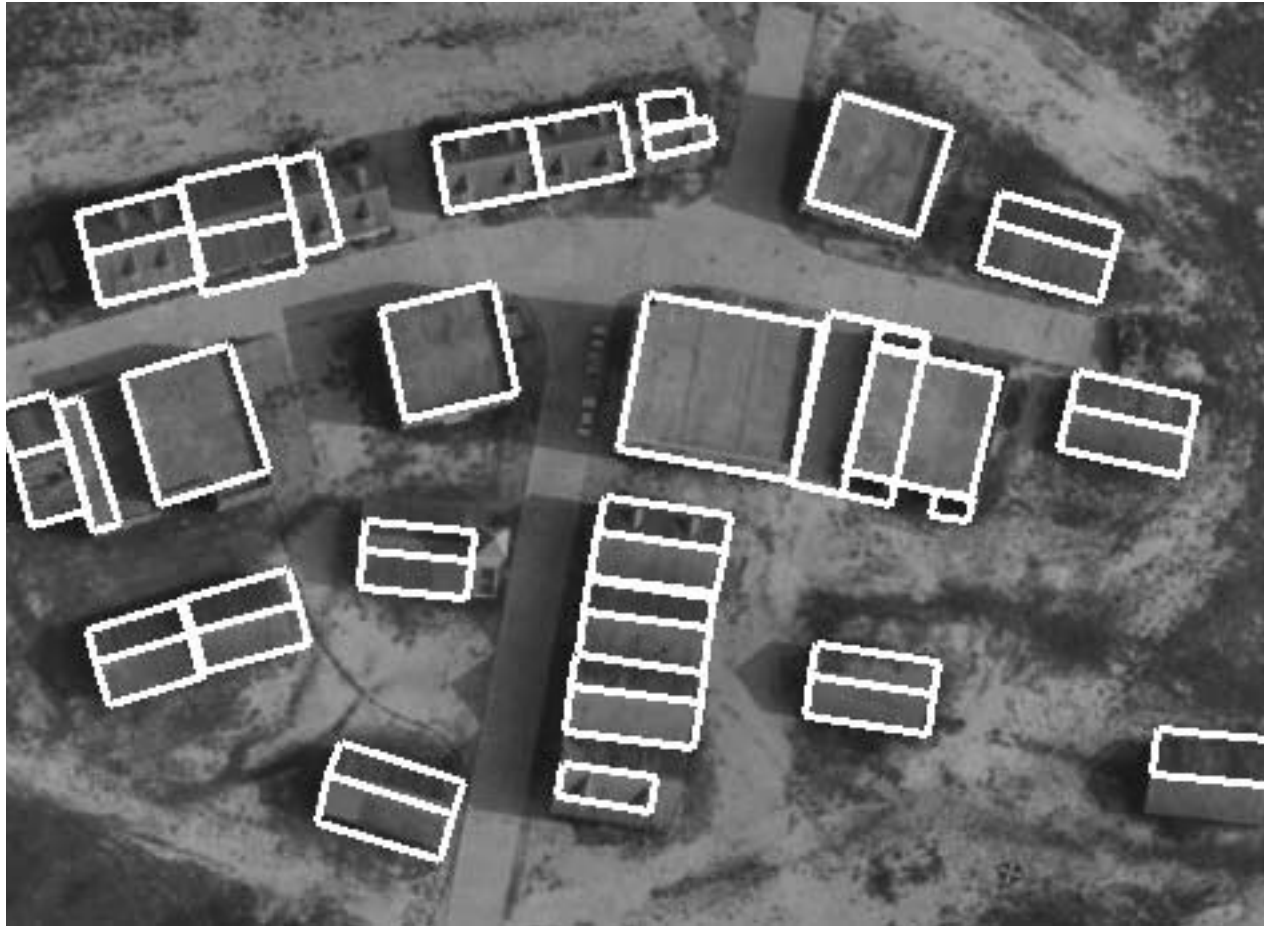


Figure 4. Final automatic result for MOUT site, using EO images only.

volving the image with a Laplacian-of-Gaussian filter that smooths the image and locates the object boundaries by the positive-valued regions bounded by the zero-crossings in the convolution output [Huertas et al., 1998].

The nearby trees on the South and West sides of the site are also well represented. For other kinds of sensors such as IFSARE, for which we have data over the Ft. Hood site, we have found it useful to use all three images (**mag**, **dte** and **cor**) for extracting the cues [Huertas et al., 1998].

Object cues are used in several ways and at different stages of the hypotheses formation and validation processes. Figure 7 shows the linear structures that are near the cue regions. By using lines near objects the system not only is more efficient as it processes a smaller number of features, but these, presumably, the more relevant features, lead to better hypotheses. We also use these cues to help select promising hypotheses, or conversely, to help disregard hypotheses that may not correspond to objects.

Just as poor hypotheses can be discarded because they lack IFSAR support, the ones that have a large



Figure 5. IFSAR derived DEM image.

support see their confidence increase during the verification stage.

Figure 8 and Figure 9 show the detected flat and gabled roofed buildings using the IFSAR cues.

Figure 10 shows the combined flat and gable verified hypotheses. This result shows no false alarms.

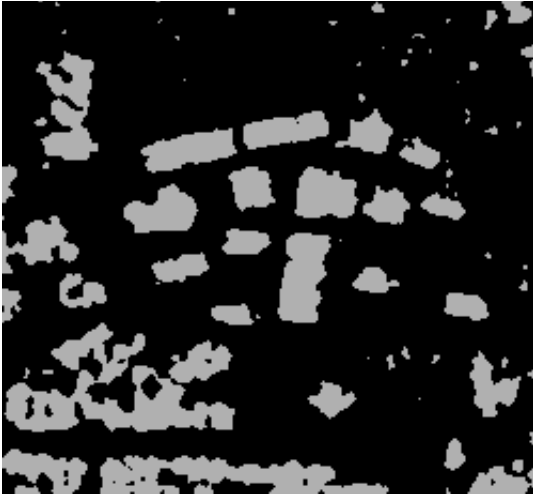


Figure 6. Cues extracted from IFSAR DEM.



Figure 7. Lines near IFSAR cues

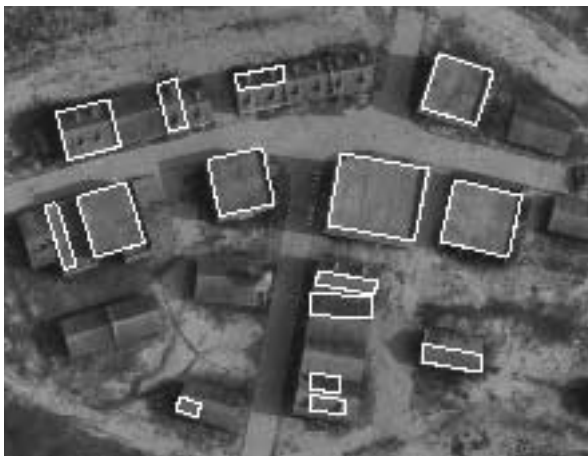


Figure 8. Flat roofed building components extracted using IFSAR cueing.

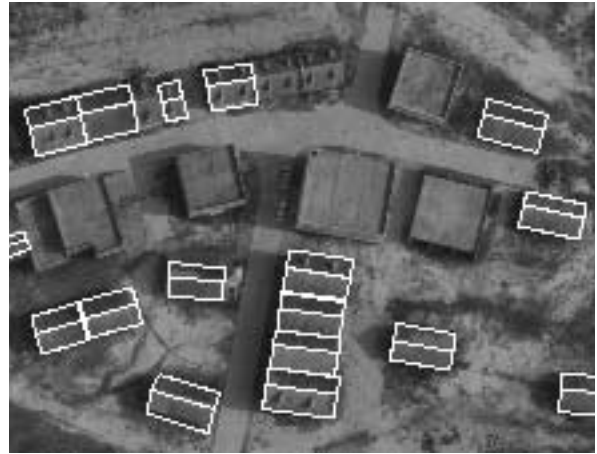


Figure 9. Gabled roof buildings extracted using IFSAR cueing. Also, the roofs of the gabled buildings are detected correctly. However, parts of the gabled buildings in the upper center have not been detected.

Table 1 gives a comparison of the number of features and final result counts with and without use of IFSAR cues. For additional and more formal evaluation of results, see next section.

Table 1: Automatic Processing Result

Feature	EO Only	With IFSAR
Line Segments	116400/18998	16400/18998
Linear Struc.	55827/6611	1758/2041
Flat Hypos	5218	3012
Selected Flat	329	192
Verified Flat	202	116
Final Flat	22	15
Gable Hypos	634	240
Selected Gab.	181	75
Verified Gab.	181	75
Final Gab.	19	17
Combined	41	32
Buildings	29 (3 false)	25 (0 false)

4 System Evaluation

Quantitative evaluation of system performance is important in determining its utility. We define some metrics below that are similar to those contained in various proposals though there is not yet a complete agreement on the most desirable ones [McKeown et al., 1997; Fischler et al., 1998]. All comparisons are made with a reference model which may be derived

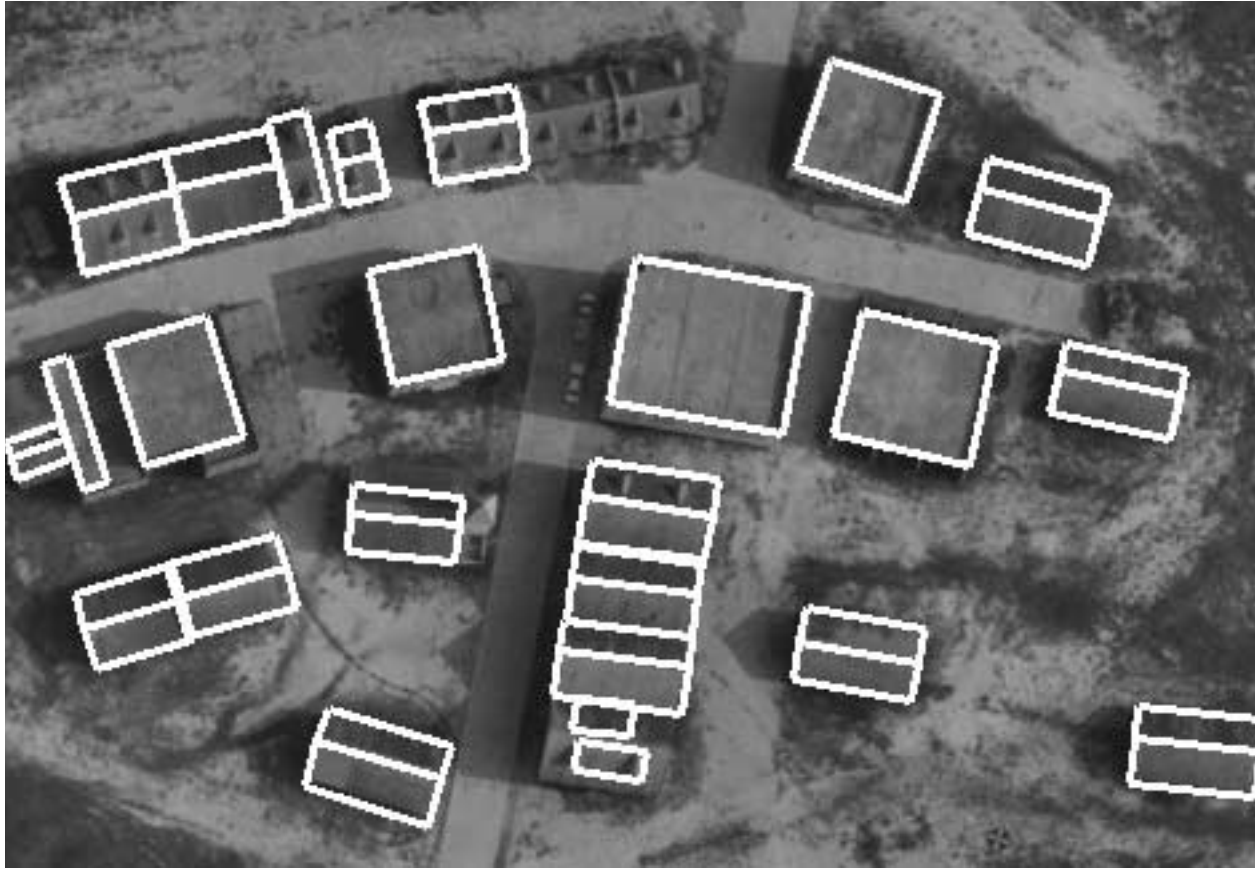


Figure 10. Final automatic result for MOUT site using IFSAR cues.

by a human operator or by measurements on the ground. The following terms are defined:

- **TP** (True Positive): a detected feature that is also in the reference.
- **FP** (False Positive): a detected feature that is not in the reference; also termed a *false alarm*.
- **FN** (False Negative), a feature in the reference that is not detected.

These quantities are combined to give:

$$\text{Detection Rate} = \frac{TP}{(TP + FN)}$$

$$\text{False Alarm Rate} = \frac{FP}{(TP + FP)}$$

Note that with these definitions, the detection rate is computed as a fraction of the reference features whereas the false alarm rate is computed as a fraction of the detected features.

In the definitions given above, a feature could be an object, an area element or a volume element. The disadvantage of using image pixels (*i.e.* the roof area) is that the numbers from a few large buildings may dominate a number of smaller buildings. If

they are to be computed on the basis of an entire object, then we need to define when we consider an object to have been detected. In our evaluations, we consider a building to have been detected if *any* part of it has been detected. The amount by which a building has been correctly detected is computed by the number of points inside that overlap with the reference. In our experiments, there are significant errors in camera models, so even correctly detected buildings can be displaced from the true positions. To compensate for this, we shift the building positions for maximal overlap with the reference and record the needed displacement as a location error.

The procedures to calculate and report performance evaluation figures of our systems are currently under development. We describe some preliminary results for the McKenna MOUT site, in terms of building objects and in terms of areas and volumes, with respect the model shown in Figure 11, for both the reference and the detected models. The gabled roofs are replaced by equivalent flat roofs formed by the four corners, so they can be evaluated by our current procedures.

Table 2 shows a summary of detection results for the McKenna MOUT site in terms of objects. Note

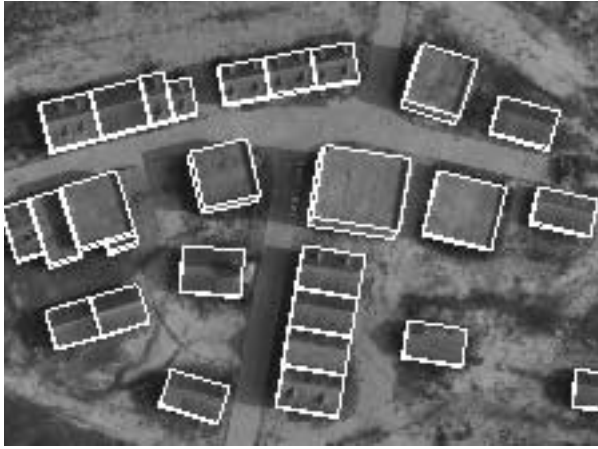


Figure 11. Reference buildings for evaluation.

that, as expected, false alarms disappear when IFSAR cues are available but the detection rate is also slightly lower.

Table 2: Components Evaluation

	Flat	Gable	Combined
Model	6	20	26
EO Total	22	19	41
TP	6	13	19
FP	2	1	3
FN	0	1	1
Detection Rate for EO Only			0.95
False Alarm Rate for EO Only			0.13
IFSAR Total	15	17	32
TP	6	16	22
FP	0	0	0
FN	0	2	2
Detection Rate for EO plus IFSAR			0.92
False Alarm Rate for EO plus IFSAR			0.00

Tables 3, 4, and 5 show area analyses for the results shown in Figure 4 and Figure 10. The rates shown in tables 3 and 4 correspond to detection rates for **reference** buildings and **detected** buildings respectively. The combined global rates for the scene are given in Table 5. Detection rate comes from Table 3, false alarm rate comes from Table 4, and the ground detection rate is the percent correct classification of ground pixels. When IFSAR cues are available, the performance can be improved significantly. A number of other analyses can be performed on these data. One can, for example, describe how the detection rates are affected by other factors such as building size and volume, scene density, available resolution, etc.

One convenient way to visualize the combined result is by a histogram of buildings which are detected to a certain accuracy. Figure 12 shows the detection rate vs. percent of **reference buildings** detected at that rate or below. Separate curves are shown for cases with and without IFSAR cueing. This graph corresponds to the figures shown in Table 3. Similarly, the area evaluation figures for the detected buildings, are shown in the graph in Figure 13.

Figure 14 and Figure 15 show similar graphs for the *volumetric* analyses for reference and detected buildings respectively. (The corresponding tables are omitted.)

Table 3: **Reference** Area Evaluation

Reference Building	Area	Detection Rate with EO only	Detection Rate with IFSAR
1	101.27	0.55	1.00
2	101.27	0.47	0.94
3	120.43	0.99	0.96
4	111.54	0.99	1.03
5	111.54	0.97	0.97
6	73.82	0.48	0.00
7	81.27	0.42	0.00
8	81.27	0.99	0.98
9	39.66	0.00	0.78
10	52.37	0.81	0.82
11	112.46	0.98	0.96
12	102.42	1.00	0.94
13	105.76	0.97	0.98
14	119.86	0.38	0.37
15	114.04	1.00	1.00
16	69.98	0.99	0.34
17	127.43	0.00	0.00
18	103.37	0.12	1.00
19	91.29	0.81	0.93
20	84.47	1.00	0.93
21	14.28	0.01	0.11
22	167.21	0.00	0.84
23	103.21	0.93	0.39
24	150.89	0.82	0.96
25	200.54	1.21	0.99
26	290.71	0.99	0.99
27	151.82	0.00	0.96

Table 4: **Detection Area Evaluation**

Detected Building	EO only/ with IFSAR	Detection Rate with EO only	Detection Rate with IFSAR
1	146.82/146.09	0.99	0.99
2	289.86/293.00	0.99	0.98
3	204.15/199.62	0.97	0.99
4	42.91/42.94	0.99	1.00
5	23.78/23.79,	1.00	1.00
6	21.20/21.21	0.98	0.99
7	110.95/110.94	0.98	0.98
8	159.53/147.77	0.99	0.98
9	113.44/113.66	1.00	0.99
10	111.10/111.15	0.92	0.94
11	115.06/116.94	0.94	0.94
12	124.29/124.52	0.98	0.98
13	83.36/83.32	0.96	0.96
14	87.51/87.27	0.91	0.91
15	103.2/101.55	0.93	0.94
16	124.9/123.43	0.82	0.83
17	131.46/147.5	0.95	0.98
18	117.00/114.82	0.94	0.96
19	115.71/111.75	0.75	0.91
20	39.67/36.32	0.88	0.85
21	96.87/97.76	0.83	0.89
22	74.12/99.91	0.93	0.95
23	67.91/90.63	0.82	0.89
24	48.04/40.17	1.00	0.91
25	13.57/24.84	0.93	0.95
26	23.14/na	0.01	na
27	18.01/na	1.00	na
28	17.11,/na	1.00	na
29	185.21/na	0.04	na

Table 5: **Combined Area Evaluation**

	EO Only	with IFSAR
Detection rate	0.8219	0.8341
False Alarm rate	0.1196	0.0407
Ground Detection rate	0.9814	0.9937

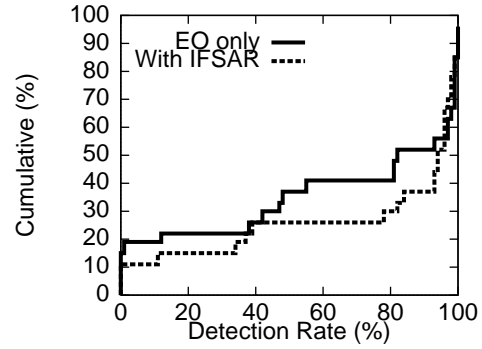


Figure 12. Evaluation curve for area analysis of reference buildings.

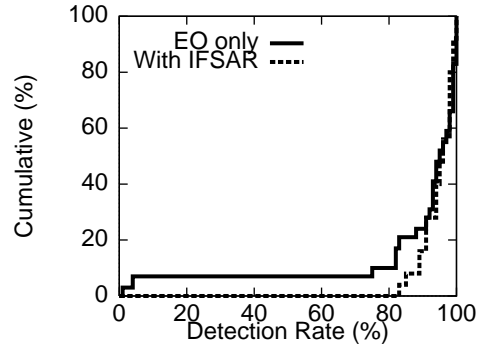


Figure 13. Evaluation curve for area analysis of detected buildings.

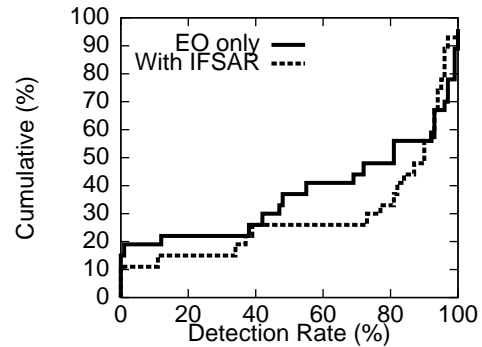


Figure 14. Evaluation curve for volumetric analysis of reference buildings.

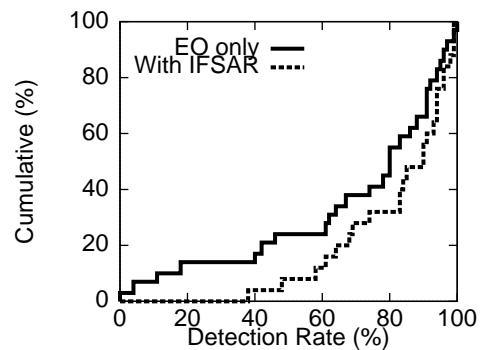


Figure 15. Evaluation curve for volumetric analyses of detected buildings.

4.1 Performance Curves

The performance evaluations given above apply to one setting of parameters of the system. Most systems, including the one described here, allow a trade-off between the false alarm and detection rates by adjusting a confidence threshold. This trade-off can be described by plotting the detection rate as a function of the false alarm rate; such curves have been called “receiver operator characteristics” (ROC) curves [Maloo et al., 1998]. We have not included such curves for this example.

5 Interactive Modeling

The results of the automatic system may be edited by a user or the user may provide some cues to the automatic system to improve its performance. The user interaction may occur either after a complete run of the automatic system or only after the stages where lines have been matched and junctions detected. User interactions aid the automatic system in forming new hypotheses. 3-D height computations are still performed automatically. The system requires a user to interact in one image only, even though a second view is displayed and used by the automatic system; this greatly reduces the effort required of the user.

A new building can be added by the user pointing to some corners of the building roof in one image; the pointing need not be precise as the precise corners are selected automatically from the image data. For flat roofs, between one and three clicks are required. For gabled roofs, between two and four clicks are required. The user given information is used to construct 2-D hypotheses from which 3-D structures are computed automatically. If the computed height is not correct, another single click is usually sufficient to correct it.

Building hypotheses, either derived automatically or by interactions in earlier stages, can also be edited by indicating a corner or side to be changed. More details of this system may be found in [Li et al., 1998].

Figure 16 shows the model for the entire Ft. Benning McKenna MOUT site constructed by this procedure (except for one building only partially visible in the window.) The distribution of the needed interactions is given in Table 6. The time measurements apply to user time *after* line and corner matches have been computed and do *not* include the initial set up times. As can be seen, it is possible to construct highly accurate models for a fairly complex site in a very short period of time.



Figure 16. User-assisted model of 26 structures takes 165 seconds to construct.

Table 6: Distribution of Interactions

Roof type	Clicks needed	Components Formed
Flat Roof Buildings	1	3
	2	0
	3	4
Gable Roof Buildings	2	9
	3	8
	4	8
TOTAL	15	26
Buildings requiring height correction = 1		
Total elapsed wall time = 165 seconds		

6 Acknowledgments

This paper describes the work of following students: S. Noronha (baseline multi-view system); Z. Kim (system upgrades and evaluation procedures); and J. Li (user assisted system).

References

- [Collins et al., 1998] R. Collins, C. Jaynes, Y. Cheng, X. Wang, F. Stolle, A. Hanson, and E. Riseman, "The ASCENDER System: Automatic Site Modeling from Multiple Aerial Images", To appear in *Computer Vision and Image Understanding Journal*, special issue on Building Detection and Reconstruction. R. Nevatia and A. Gruen, Editors, 1998.
- [Curlander & McDonough, 1991] J. Curlander and R. McDonough, "Synthetic Aperture Radar", Wiley Interscience, New York, 1991.
- [Fischler et al., 1998] M. Fischler, B. Bolles and A. Heller. "APGD Evaluation Metrics, Methodology, Rationale" SRI International, May 1998, <http://www.ai.sri.com/~apgd>
- [Hoepfner et al., 1997] K. Hoepfner, C. Jaynes, E. Riseman, A. Hanson and H. Schultz, "Site Modeling using IFSAR and Electro-Optical Images", *Proceedings of the DARPA Image Understanding Workshop*, New Orleans, LA, May 1997, pp 983-988.
- [Huertas et al., 1998] A. Huertas, Z Kim and R. Nevatia, "Use of Cues from Range Data for Building Modeling", *These proceedings*.
- [Jakowatz et al., 1996] C. Jakowatz, D. Wahl, P. Eichel, D Ghiglia and P. Thompson, "Spot-Light Mode Synthetic Aperture Radar: A Signal Processing Approach", Kluwer Academic, Boston, 1996.
- [Jaynes et al., 1997] C. Jaynes, M. Marengoni, A. Hanson, E. Riseman and H. Schultz, "Knowledge Directed Reconstruction from Multiple Aerial Images", *Proceedings of the DARPA Image Understanding Workshop*, New Orleans, LA, May 1997, pp 971-976.
- [Li et al., 1998] J. Li, S. Noronha and R. Nevatia, "User Assisted Modeling of Buildings", *These Proceedings*.
- [Maloof et al., 1998] M. Maloof, P. Langlay and R. Nevatia, "Generalizing over Aspect and Location for Rooftop Location", *IEEE Workshop on Applications of Computer Vision*. Princeton, NJ. October, 1998, to appear.
- [McKeown et al., 1997] D. McKeown, et. al. "Research in the Automated Analysis of Remotely Sensed Imagery: 1995:1996", *Proceedings DARPA Image Understanding Workshop*, New Orleans, LA, May 1997, pp. 779-812
- [Noronha & Nevatia, 1997] S. Noronha and R. Nevatia. "Detection and Description of Buildings from Multiple Aerial Images", *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, PR, June 1997, pp. 588-594.
- [Roux & McKeown, 1994] M. Roux and D. McKeown, "Feature Matching for Building Extraction from Multiple Views", *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 46-53.