Table 1:

Section	Detection Percentage $t_p/(t_p + t_n)$	Branching Factor $f_p/(t_p + f_p)$	Correct building pixels	Correct non-building pixels
Section 1	82.26%	0.08929	75.36%	99.13%
Section 2	78.13%	0.13333	71.84%	98.72%

We feel that the results we show, particularly in Figure 13, are on much more complex examples and that we have processed much larger number of buildings than reported in the previous literature ([5], [2], and [10]). We feel that our method has advantage over monocular systems such as ([5]) due to its ability to infer 3-D more directly and of having several views to validate the hypotheses. We feel that our system can perform more consistently over systems such as ([2]) which form hypotheses in one view and verify in others by not being dependent on a favored view.

Bibliography

[1] R.C.K. Chung and R. Nevatia, "Recovering building structures from Stereo", IEEE Proceedings of Workshop on Applications of Computer Vision, 64-73, Dec 1992.

[2] R. Collins, Y. Cheng, C. Jaynes, F. Stolle and X. Wang, "Task Driven Perceptual Organization for Extraction of Rooftop Polygons", Proceedings of International Conference on Computer Vision, 6:888-893, June 1995.

[3] M. Ito and A. Ishii, "Three-view stereo analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, 8:524-532, 1986.

[4] H.S. Lim and T.O. Binford, "Stereo correspondence: A hierarchical approach", Proceedings of DARPA Image Understanding Workshop, 234-241, 1987.

[5] C. Lin, and R. Nevatia, "3D Descriptions of Buildings from an Oblique View Aerial Image", IEEE International Symposium of Computer Vision, 377-382, 1995.

[6] X.C. Magnisalis and K.L. Boyer, "Hierarchical structural stereo matching with simultaneous autonomous camera calibration", Proceedings of International Conference on Pattern Recognition, 711-713, 1994.

[7] J. McGlone and J. Shufelt, "Projective and Object Space Geometry for Monocular Building Extraction", IEEE Proceedings of Computer Vision and Pattern Recognition, 54-61, 1994.

[8] R. Mohan and R. Nevatia, "Using perceptual organization to extract 3-D structures", IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(11), 1121-1139, Nov 1989.

[9] S. Noronha., R. Nevatia., "Detection and Description of Buildings from Multiple Aerial Images", Proceedings of Image Understanding Workshop 1996, Palm Springs, pp. 469-478.

[10] M. Roux and D.M. McKeown, "Feature matching for building extraction from multiple views", IEEE Proceedings of Computer Vision and Pattern Recognition, 46-53, 1994.

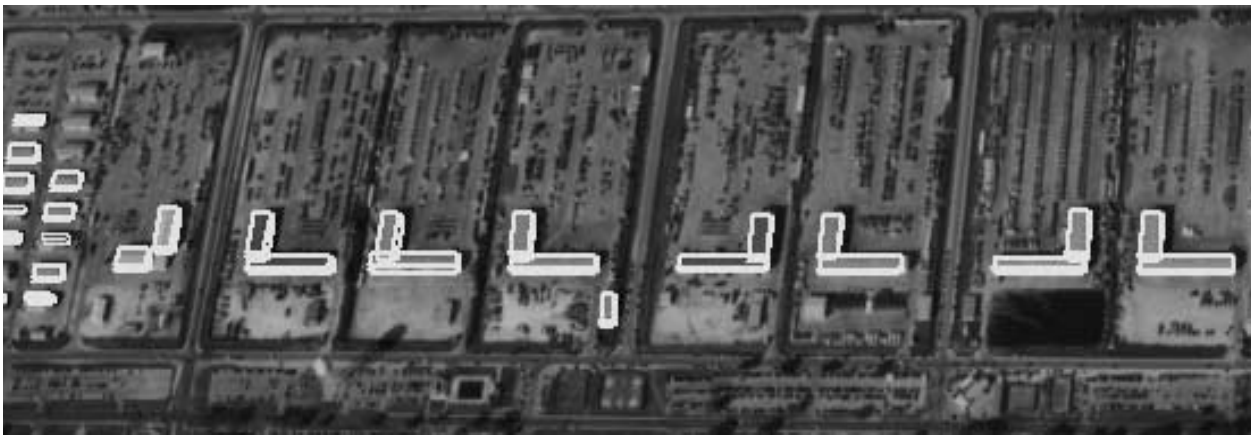

Figure 15 Section 1 of the Motor Pool Area from Fort Hood, TX

Figure 16 Section 2 of the Motor Pool Area from Fort Hood, TX

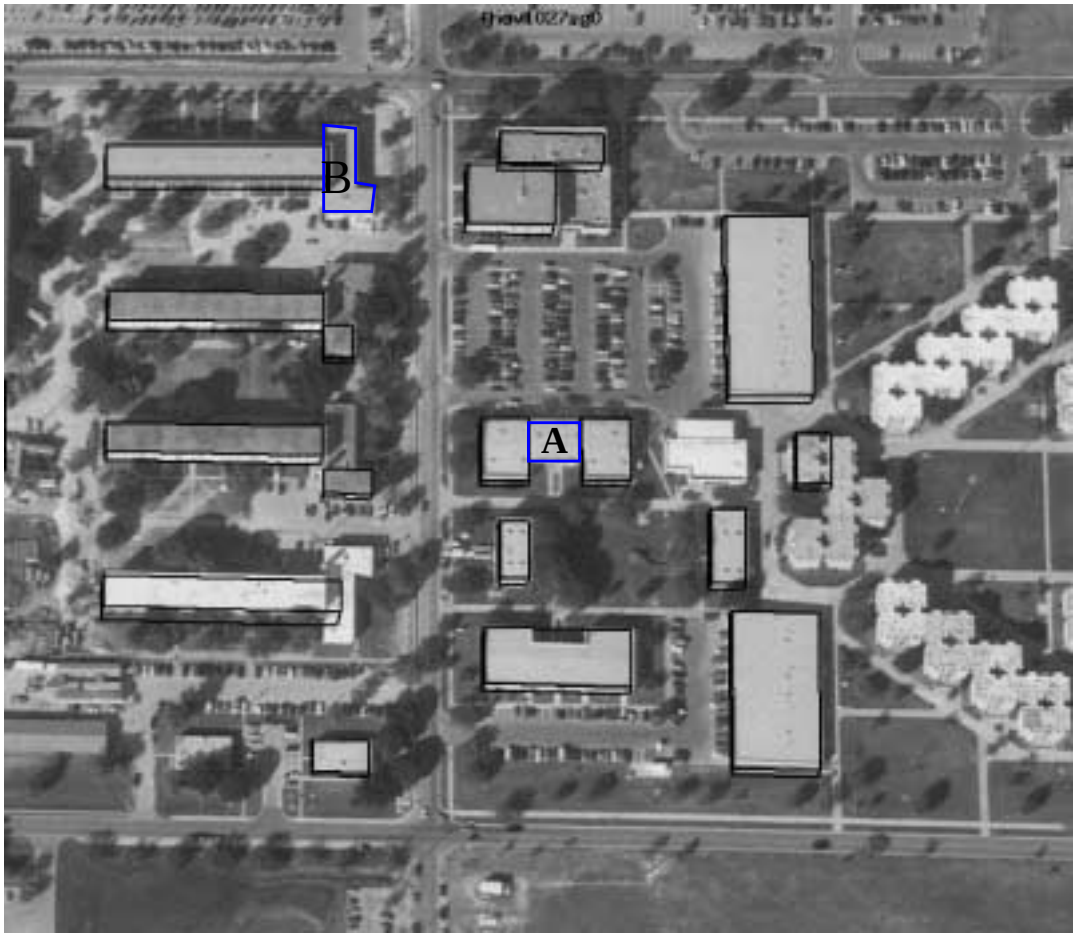

Figure 13 Results on a complex area of Fort Hood

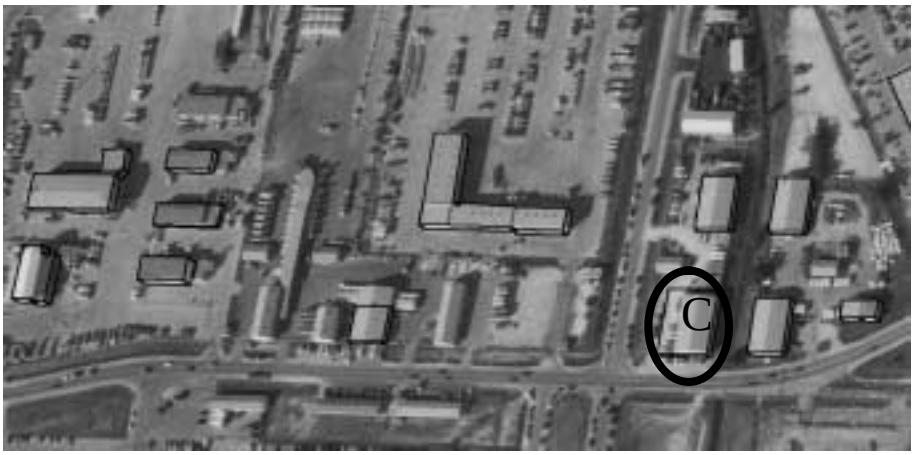

Figure 14 Section of the Motor Pool Area of Fort Hood, TX

Figure 9 Rows of small buildings

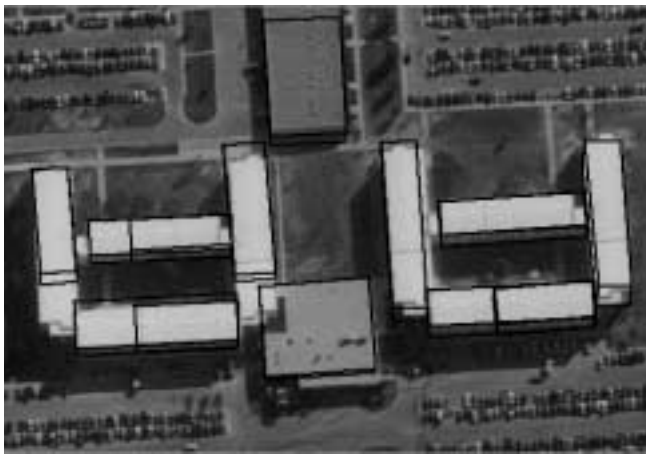

Figure 10 A-shaped buildings

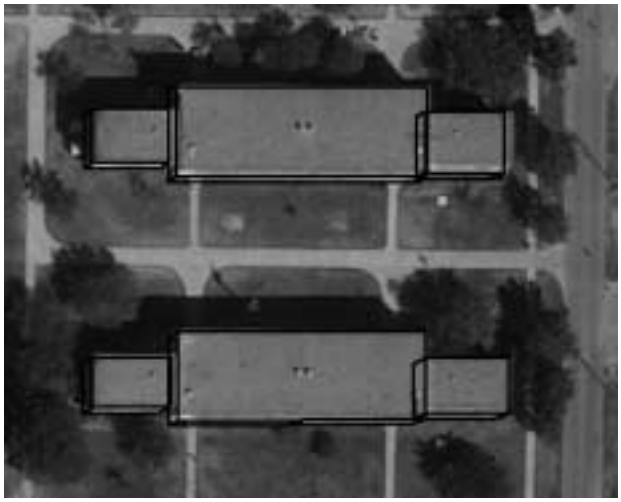

Figure 11 Buildings with wings

Figure 12 Gabled roof and multi-level buildings

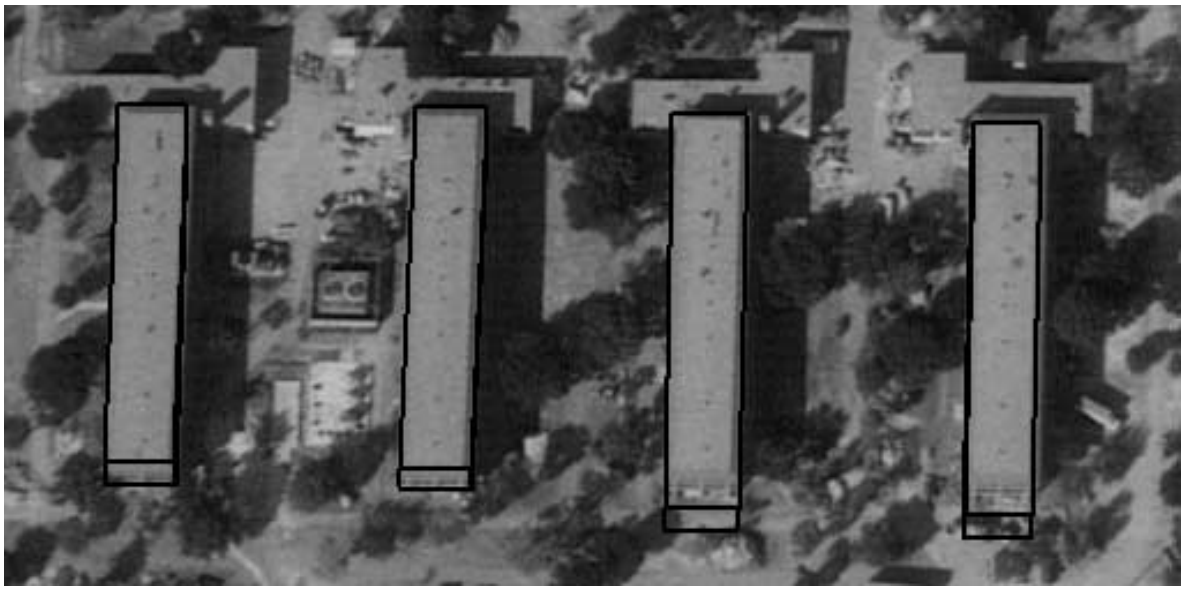

Figure 7 T-shaped buildings

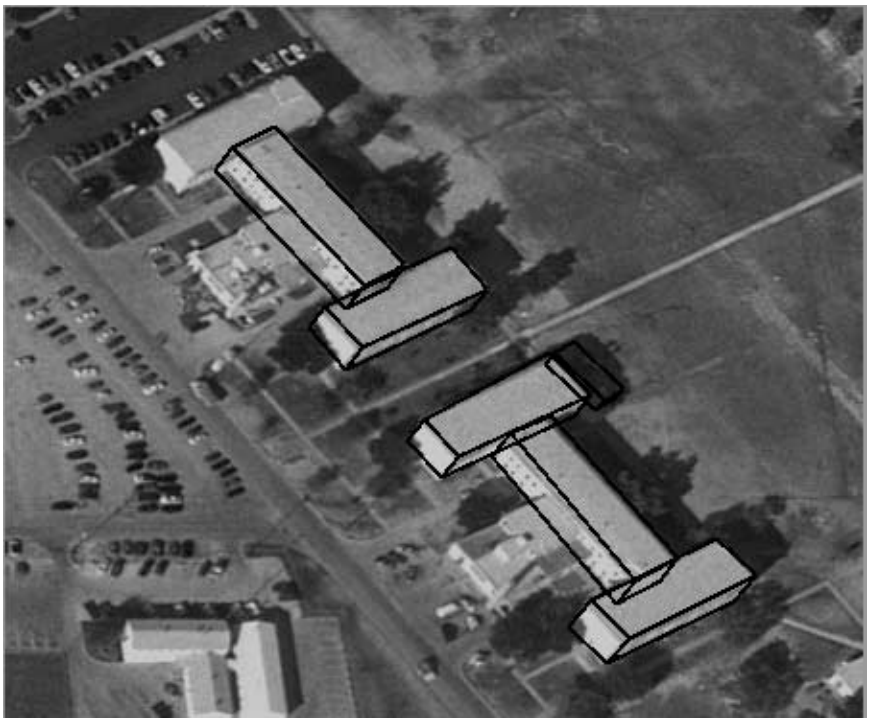

Figure 8 I-shaped buildings

Figure 1. Both buildings in this example are described. The example indicates how the combination routine may be recursively applied to combine rectangular building components into a rectilinear building.

Figure 6 shows L-shaped buildings, with markings on the roofs. It may be noted that the structural features are parallel, giving rise to a number of parallel structures that the program has to disambiguate between. Figure 7 shows T-shaped buildings. There is fairly dense vegetation close to the buildings, causing shadow evidence to be obfuscated in some areas. The small sections of the building near the top edge of the image are not verified because large sections are occluded by the shadows of the long parts of the building. Figure 8 shows the results obtained for fairly complex I-shaped buildings. It may be noted that the view depicted is fairly oblique. There is one false positive in this example. This is caused by accidental alignment of shadows and the side of the building. Figure 9 shows a number of small, low buildings. The height estimates for these buildings are extremely unreliable (the buildings are less than 5m high). Shadows and walls are extremely important in this case. Some of the small buildings were not verified because there was not enough edge evidence to support their selection. Figure 10 shows the performance on extremely complex A-shaped buildings. Many shadow and wall areas of rectangular components of these buildings are occluded by other parts of the buildings. This example proves that the system can work when significant amounts of evidence do not exist. Figure 11 shows two buildings with wings. This illustrates the working of the system with buildings with few features on the roof (which might create problems for area-based systems). Figure 12 shows the results on some multi-level and gabled-roof buildings. The multi-level building is not perceived as a multi-level building, because the difference in heights of the levels is too small to be detected in the presence of the registration errors. Figure 13 shows results on a fairly complex area. Most of the buildings are detected and modeled correctly, inspite of occlusion from vegetation and poor shadows. Some errors or deficiencies can also be seen. Some components of multi-wing buildings are not detected because of missing line evidence, such as the component labeled **A**. The building labeled **B** is an instance of a "true negative". This building is missed because of occlusion by shadows of the neighboring building, and the low height of the building itself.

We have processed large areas of the "Motor Pool Area" of the Fort Hood images as shown in Figures 14, 15 and 16. Figures 15 and 16 are reproduced at low resolution to show the large sections. Figure 14 shows a sub-section at the resolution at which data is processed. These results were obtained by using the depicted view with one other overlapping view. There are a number of multi-wing buildings, flanked by smaller rectangular buildings. The rooftops of these buildings are very similar photometrically, to the ground. None of these buildings is taller than 15m. Inspite of these difficulties, the system reliably finds the large buildings in areas where the sides of the buildings are not highly fragmented owing to the similar reflectance properties of the buildings and the ground near it. It performs less reliably when the epipolar lines are parallel to the sides of the buildings as matching these lines is harder than when the lines form a significant angle with the epipolar lines. For example, the building labeled **C,** in Figure 14 is inaccurately modeled. This error is caused by accidental background geometrical formations. Better registration would permit higher confidence 3D height estimates, facilitating better selection.

Evaluation of the system is performed using quantitative and qualitative criteria. A model is constructed by hand for comparison. A building is declared detected if its roof area overlaps more than 50% of a roof of a building in the supplied model. Quantitative measures of the performance of the system may be defined as follows: if t_p is the number of true positive hypotheses, t_n is the number of true negative hypotheses and f_p is the number of false positive hypotheses, then we define the detection percentage as $t_p/(t_p + t_n)$, and the branching factor as $f_p/(t_p + f_p)$. For one part of the site from the Motor Pool Area of Fort Hood, TX, (shown in Figure 15), t_p was 51, t_n was 11, and f_p was 5. For another part of the site, from the Motor Pool Area (shown in Figure 16) t_p was 25, t_n was 7 and f_p was 4. Measures of the number of pixels that are correctly labeled as building and non-building pixels are also useful. They are obtained by comparison with the supplied model.These measures are shown in **Table 1**.

We are unable to compare our results with those of other researchers directly as we do not have access to their software. We can compare to their previously published results, however, they may not be on the same data even when they may have used the Ft. Hood data set and the published results are necessarily outdated.

Figure 6 L-shaped buildings

parallel that has closure evidence near one of its ends. The presence of a U is a strong indication of a fragment of a parallelogram in 2D (implying a possible rectangular fragment in 3D). [9] contains details of the implementation, and the constraint applied at this stage.

2.5 Parallelograms

Formation of parallelograms is the basis for hypothesizing buildings. The existence of evidence to form a parallelogram match, is a strong indication that a rectangular 3D structure exists.

3 Selection of Building Hypotheses

The parallelogram matches serve as roof hypotheses, and are equivalent to having a 3D model of the buildings. Owing to the resolution of the images, and the large errors in triangulation from small errors in the images, additional processing needs to be done to distinguish which hypotheses are buildings or parts thereof, and which are rectangular areas on the ground. This necessitates a selection procedure. The selection procedure uses four criteria to decide which hypotheses should remain for verification, namely the 3D height of the building, positive and negative line evidence, and junction evidence. Details are given in [9]. The results after the selection procedure are shown in Figure 4

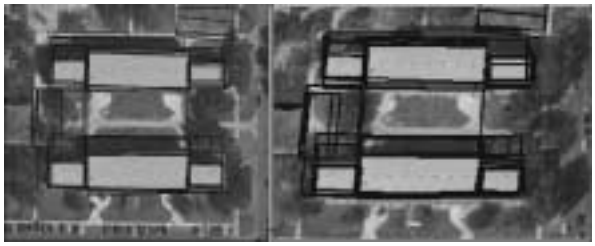

Figure 4 Selected hypotheses from Figure 1

4 Verification of Building Hypotheses

It may be noted that so far the evidence that was used was concerning the roof only. The presence of lighting causes shadows to be cast. When the view is oblique, some vertical sides of the walls of the building may be visible. These cues are used to verify the selected hypothesis, and further reduce the number of hypotheses based on the available evidence. The numerical evidence for the walls and shadows is accumulated for all the views, and the average is compared against a threshold. These monocular cues are extremely important when the registration has errors large enough to cause height estimates to be unreliable. This is the case with the Fort Hood images.

Wall evidence. In a view which is not nadir one or more of the side walls of the buildings should be visible. These walls are assumed to be vertical. The details of verification for walls are included in [9].

Shadow evidence: When shadow lines are present, they are used to boost the confidence of the hypothesis. In case of the Fort Hood images, shadow evidence is often a major factor in the verification of a selected hypothesis.

The evidence of shadows and walls is combined in a Bayesian manner, with a priori probability estimates obtained from the expected length of the shadow (wall) line).

The combined score from the wall evidence and shadow evidence is thresholded to obtain rectangular building (or building component, in the case of non-rectangular rectilinear buildings) hypotheses.

5 Combination of Rectangular Buildings

Rectilinear buildings can be decomposed into rectangular components. Verified rectangular hypotheses are examined for combination according to two mutually exclusive criteria: proximity, and overlap. The precondition for both criteria is that the hypotheses be of approximately the same height in 3D.

6 Results

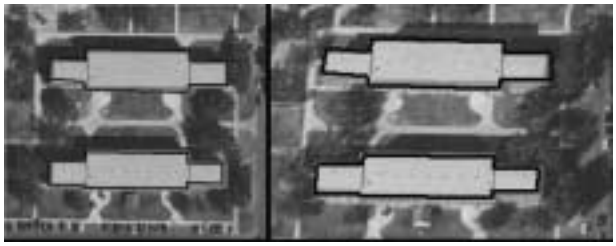

Figure 5 Results after combination from Figure 1

We have used images from Ft. Hood, Texas, for our experiments.This dataset was acquired for the U.S. government sponsored RADIUS program and has become a common "benchmark" for evaluation of building detection systems. This site is challenging, because it has non-rectangular buildings, vehicles are present on the roads and parking lots, and it has trees and grassy areas. Real lighting conditions cause shadows that are not necessarily the darkest areas in the images. Furthermore the acquisition geometry is such that the epipolar lines between many pairs of views are almost parallel (within 5^o) to one of the sides of the buildings (in at least one view) at the site. This causes height estimates to be less reliable and the selection process less certain.

Figure 5 shows the results obtained from the images in

Our approach is to first form hypotheses for building roofs as roofs project into larger areas and to verify the hypotheses by using evidence from cast shadows and visible walls (if any). As we consider only rectilinear buildings, a natural hierarchy for hypotheses formation is that of lines, junctions, parallels, Us (three sides) and parallelograms (roofs project into parallelograms since the imaging distance is large compared to the height of the buildings). Matching at one level is used to form group hypotheses at the next level. We maintain multiple matches at each level and resolve them only when sufficient information becomes available at the higher levels. We believe that this approach not only provides good results for the building detection task but also provides a model for more general conditions.

Our system has been tested on a number of real images. The details of our system and some results are shown in the following sections.

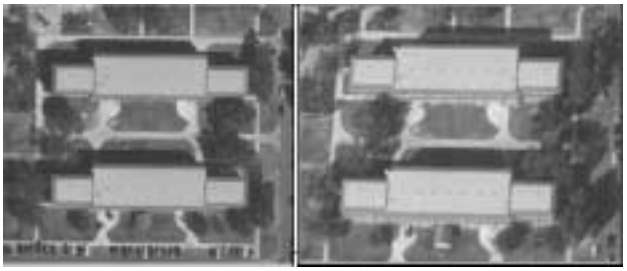

Figure 1 Two views of a scene

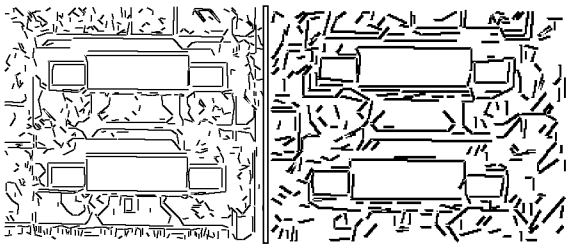

Figure 2 Lines detected in Figure 1

2 Hierarchical Grouping and Matching of Features

The system is built to handle two or more views non-preferentially. A hierarchy of features is used. Starting from the most primitive these are lines, junctions, parallels, U's and parallelograms (because the projection of a rectangle is a parallelogram in general, assuming negligible perspective effects), in each image. Grouping and matching is performed at each stage. Below we explain which features in the hierarchy were chosen for matching purposes.

2.1 Lines

Lines detected using the Canny edge-detector are matched across all views. Multiple matches are retained at this stage. The constraint used in matching is the quadrilateral constraint described in [9].

.

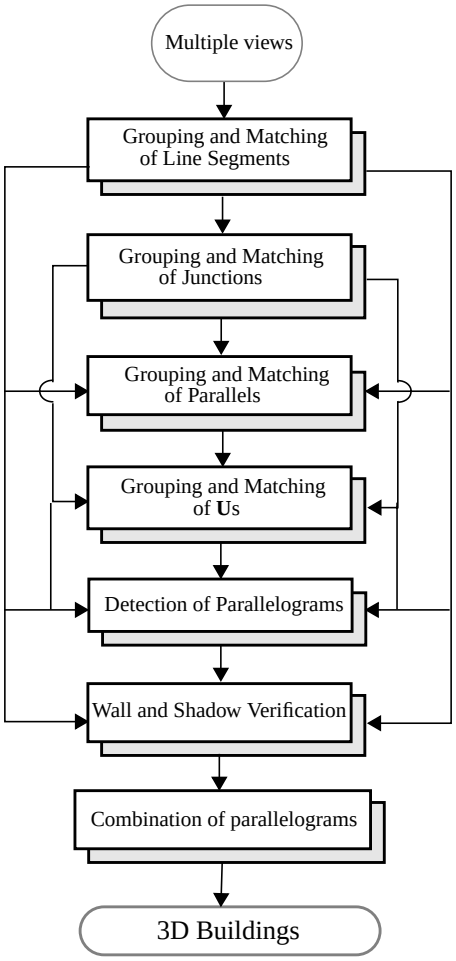

Figure 3 Block Diagram of the System

2.2 Junctions

Next, binary junctions (formed by the intersection of exactly two lines) are formed and matched. The constraints for junction matching are the epipolar constraint, the line-match constraint, the 3D Orthogonality constraint, the 3D height constraint and the trinocular constraint. These are outlined in [9].

2.3 Parallels

Next, we search for parallels and their matches. Parallels are formed between pairs of lines in the same view that are separated by the less than the maximum width of a building. A match is hypothesized if there is evidence in at least two views. The parallel match constraint described in [9] is used to remove parallel matches that cannot lead building hypotheses that are planar in 3D.

2.4 U's

U's are formed by an alignment of two junctions, or by a

Detection and Description of Buildings from Multiple Aerial Images

Sanjay Noronha and Ram Nevatia*
Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, California 90089-0273

Abstract

A method for detection and description of rectangular buildings from two or more registered aerial intensity images is proposed. The output is a 3D description of the buildings, with an associated confidence measure for each building. Hierarchical perceptual grouping and matching across views is employed to increase the robustness of the system. Verification of selected building hypothesxses is done using shadow and wall evidence of the buildings. The system is largely feature-based. Grouping and matching are performed in a hierarchical manner, utilizing primitives of increasing complexity, starting with line segments and junctions, and proceeding to higher level features. Binocular and trinocular epipolar constraints are used to reduce the search space for matching features.

1 Introduction

Detection and description of building structures from aerial images is becoming increasingly important for a number of applications such as map-making, change detection and databases for simulators. This problem also offers an opportunity of exploring the issues of object segmentation and 3-D shape inference in a limited setting but where significant challenges must be met. While this problem has been approached with use of just a single image ([5], [7]), multiple images of the same scene are often available. In this paper, we assume that two or more images are available though they may not be taken at the same time and so the imaging conditions may be quite different.

The task of detecting and describing buildings presents many challenges. In a single image, the object boundaries are typically highly fragmented due to low contrast, occlusion caused by nearby vegetation and by smaller structures on the roofs, and need to be grouped to yield the desired objects. In our work, we limit the buildings shapes to be rectilinear (i.e. rectangular or compositions of rectangular shapes) to aid the task of organization. However, many other structures such as roads, sidewalks and parking lots can also give rise to rectilinear organizations and need to be distinguished from the building structures. Availability of multiple images allows the possibility of doing some of this reasoning in 3-D, by making correspondences between the image features. This task too is difficult in the aerial image domain. Area correlation methods are likely to have difficulty as the viewpoints can be widely separated, the images are taken at different times and the building roofs have limited texture. In our system, we choose to match features instead.

Figure 1 shows two views of a scene. Figure 2 shows the line segments detected in the images in Figure 1. These views illustrate some of the difficulties that arise. A large number of line segments are detected but only a few correspond to boundaries of desired structures. Many of the lines are parallel to each other and hence difficult to match in the two views unambiguously, without higher level context. Also, many rectangular organizations of the features are possible if fragmentation is allowed. In addition, poor camera calibration prevents us from making highly accurate 3D position inferences, which complicate the task of higher level segmentation and description.

An important question in multiple image analysis is the level at which image features should be corresponded. Lower level features, such as edges, are easy to detect but are highly ambiguous. Higher level features, such as surfaces, are easily matched but hard to detect reliably in single images. Some systems have been constructed to match features such as junctions ([1], [10]) which are then used for grouping. Other systems have attempted to find candidates for roof boundaries and match them or verify them to get 3-D descriptions ([8], [2]). We feel that matching at only one level does not fully exploit the information available in the multiple images and that rather than deciding between grouping first and then matching, or matching first and then grouping, it is more advantageous to interleave the two processes so that local features are matched and then grouped to form higher level features in a hierarchical way. While hierarchical approaches have been suggested in the past (for example, [4], [6]) they have rarely been implemented for scenes of complexity considered here.

A block diagram of our approach is shown in Figure 3.

* This research was supported mostly by Contract No. DACA-76-93-C-0014 from the Advanced Research Projects Agency (ARPA) of the Department of Defense and monitored by the Topographic Engineering Research Center of the U.S. Army.