

We feel that the results we show, particularly in Figure 6, 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 [9]). 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] 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 9 Section 2 of the Motor Pool Area from Fort Hood, TX

**Table 1:**

| Section   | Detection Percentage<br>$t_p/(t_p + t_n)$ | Branching Factor<br>$f_p/(t_p + f_p)$ | 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%                      |

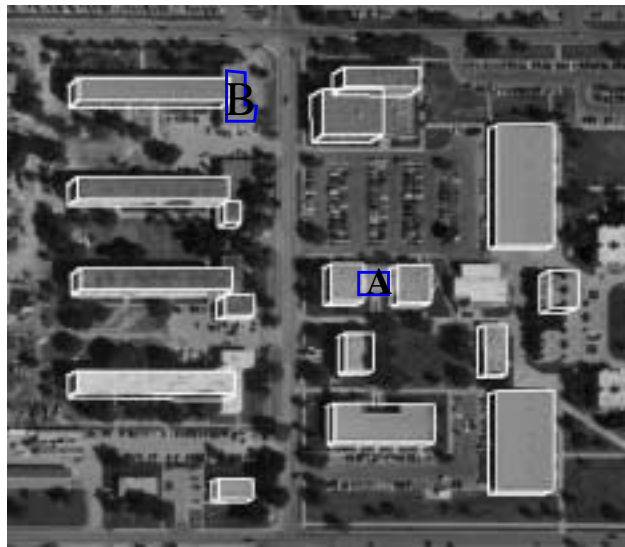


Figure 6 Results on a complex area of Fort Hood, TX



Figure 7 Section of the Motor Pool Area of Fort Hood, TX

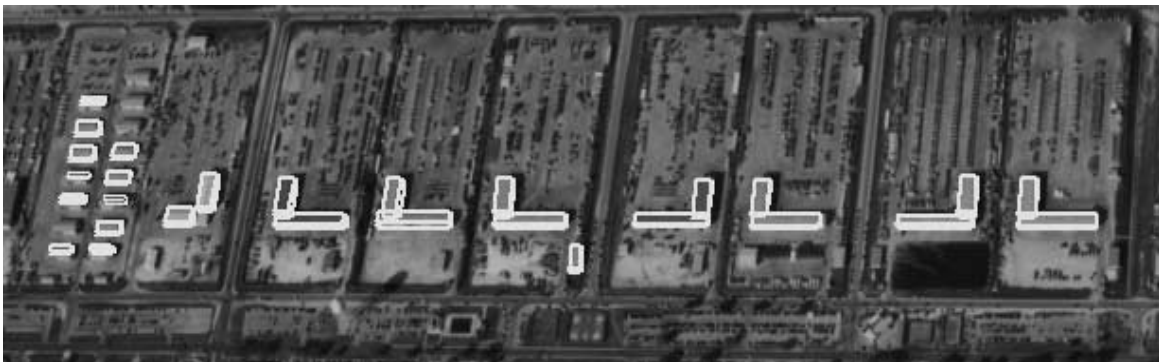


Figure 8 Section 1 of the Motor Pool Area from Fort Hood, TX

**Overlap:** Two hypotheses may partially overlap. The new hypothesis is obtained by taking the union of the areas of the hypotheses being combined. The combined hypothesis is verified with wall and shadow evidence, and a decision to accept the combination or not, is made, based on whether the confidence associated with the combination is higher or lower than the sum of the confidences of the individual hypotheses.

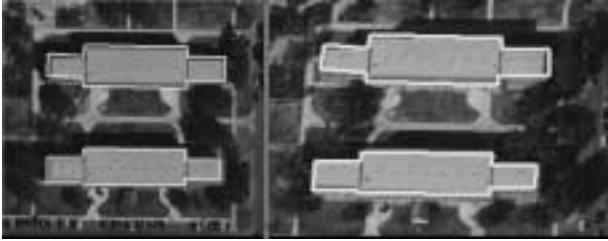


Figure 5 Results after combination from Figure 1

## 6 Results

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  $5^\circ$ ) 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 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 results on a fairly complex area. Most of the buildings are detected and modeled correctly, in spite 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 7, 8 and 9. Fig-

ures 8 and 9 are reproduced at low resolution to show the large sections. Figure 7 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. In spite 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 7 is inaccurately modeled. This error is caused by accidental background geometrical formations. Better registration would permit higher confidence 3D height estimates, facilitating better selection. In addition, a more sensitive analysis of wall and shadow evidence may yield an ability to better discriminate between configurations on the ground and buildings.

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 (detected existing buildings),  $t_n$  is the number of true negative (undetected existing buildings) hypotheses and  $f_p$  is the number of false positive (detected non-existent buildings) 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 8),  $t_p$  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 9)  $t_p$  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.

the hypothesis.

**Overlapping Gap evidence.** If there is a gap in one of the sides of a parallelogram hypothesis that overlaps with a gap in the side parallel to that side, the hypothesis is penalized.

Selection of hypotheses that satisfy the 3D height constraint is done by thresholding on a weighted sum of the evidence arising from the positive and negative line evidence, the junction evidence and the overlapping gap evidence. The results after the selection procedure are shown in Figure 4

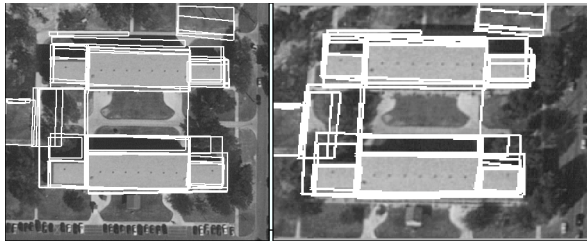


Figure 4 Selected hypotheses from Figure 1

#### 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.

**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 verification for walls involves looking for the bottom of the wall (where the vertical wall meets the ground). At this point the building's height is known through triangulation. Using the camera models, the projection of the vertical direction in 3D is computed. From the top of the wall to the bottom, a search for line evidence parallel to the side of the hypothesized building, is performed in incremental steps. Wall evidence is deemed to be found if there is evidence of parallel lines at the distance from the top of the building that is predictable from its height in 3D. The score associated with this evidence is the ratio of the length of the line coverage of the side to the length of the side.

Additional evidence from the projection of the visible vertical sides of the hypothesized building is obtained. If the predicted length of the projected vertical (obtained from the height of the building in 3D) is less than 5 pixels, it is considered unreliable, and not taken into consideration. A score of 0.25 is added for each vertical wall that is found.

**Shadow evidence:** When shadow lines are present, they

are used to boost the confidence of the hypothesis. This assumes we know the direction of illumination of the sun. The direction of illumination is easy to compute given which direction is North, the time of the day, and the time of the year.

The search for shadows is carried out in a manner similar to that for walls. Knowing the height of the building in 3D, and the direction of illumination, a search is done to detect evidence of the predicted projection of the shadow. This includes the shadow cast by the side of the roof that is detected, and the shadow cast by the vertical wall of the building. Occlusion of shadows by the building itself is taken into consideration when searching for shadows.

It is pertinent to point out that there is no assumption that the images were taken at the same time of the day. The images are assumed to have been obtained under very different imaging conditions, at varying times of the day. In essence, the shadows are monocular cues, and no effort is expended in searching for matches of shadows in other images.

The evidence of shadows and walls is accumulated for all the views, and a score is associated with the evidence found. This score is a function of the extent of coverage, against expected coverage, and the accuracy of the location of the evidence compared to the predicted location. However, the system does not take into account missing evidence because of occlusion by other detected structures, or because of the shape of the building. For instance, if the structure is L-shaped, the system might hypothesize the structure as a combination of two adjacent or two overlapping rectangles. In either case the two rectangular hypotheses may lack evidence in the common part, depending on the viewpoint and on the direction of illumination.

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.

**Proximity:** When two hypotheses have common boundaries, or common partial boundaries, they are candidates for combination, which is effected if the resulting hypothesis has sufficient wall and shadow evidence to support the combined hypothesis. The criterion used for deciding between combining and leaving the hypotheses separate, is whether the confidence associated with the wall and shadow evidence of the composite is greater or less when compared to the sum of the confidence values of the individual hypotheses. This combination is effected by deleting the common boundary, and retaining only the non-common boundaries of the two building hypotheses.

- 3D Orthogonality Constraint

The 3D lines determined by the matched line pairs, corresponding to the roof boundaries only, must be orthogonal

- 3D Height Constraint

The computed 3D height must lie between the terrain (ground) and the maximum allowed height.

- Trinocular Constraint

A match in two views is constrained in a third view by the standard trinocular constraint [3]. The junctions in the two views will cause epipolar lines in the third view, that intersect in a unique point, in general. It is not possible to use this constraint when the 3D points are in the trinocular plane. This corresponds to the epipolar lines in the third view coinciding.

### 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. While the task domain causes a large number of parallels in each image (two to three times the number of lines in that image), because of the alignment of buildings, roads, parking lots and shadows, the number of parallel matches is typically lower than the number of lines in any image. A match is hypothesized if there is evidence in at least two views. The following constraint is used.

- Parallel match constraint

Each line forming the parallel in one view should be a member of a line match with exactly one of the lines forming the parallel to be matched, in the other view.

### 2.4 U's

U's are formed by an alignment of two junctions, or by a 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). U matches are formed either by an alignment of two junction matches, or by a parallel match that has closure evidence in one or more views. The occurrence of U matches in two or more views is a higher confidence event than the matching of junctions or the matching of parallels. In the case of U match being formed from a parallel match coupled with evidence of closure, there is no constraint that we apply. In the case that a U match is formed from two aligned junction matches the constraint it must satisfy is:

- Planarity Constraint

The planes defined by the constituent junction matches should be approximately coplanar.

### 2.5 Parallelograms

Formation of parallelograms is the basis for hypothesizing buildings. To hypothesize a 3D rectangle the minimal requirement is a U match (a match may be in 2 or more images), with additional evidence arising from one of the following

four cases:

- An “opposing” U match
- An “opposing” U in a single image
- A parallel match

The existence of evidence to form a parallelogram match, is a strong indication that a rectangular 3D structure exists. As we focus on building detection and description from multiple images, we have chosen to ignore parallelogram evidence in one image alone, that is not supported by any evidence detected in the other images. Parallelogram matches across the images are the hypotheses of rectangular 3D buildings. The constraint used in parallelogram matching is:

- Planarity Constraint

The components of a parallelogram match (line matches and junction matches) must be coplanar in 3D

In the case of a U match with a single “opposing” U (or “opposing” junction), virtual matches for the U (or junction) are hypothesized, consistent with the planarity constraint. Parallelogram matches hypothesized in this way are accorded a lower initial confidence as there is no additional constraint to be met.

## 3 Selection of Building Hypotheses

The parallelogram matches serve as roof hypotheses, and are equivalent to having a 3D model of the buildings. They satisfy the constraints of being rectangular in 3D, and almost coplanar (they are coplanar if the registration is perfect and the images have infinite resolution). In addition, the height and orientation with respect to the ground is known. However, 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.

**3D height of the building.** The 3D height of the hypothesis should lie within the limits of the ground plane, and the maximum height

**Positive and Negative line evidence.** Lines that are found within a chosen distance (of 3m) in 3D, from the hypothesized parallelogram, and that differ in angle by not more than some angle ( $10^\circ$ ) are considered positive evidence. Negative evidence consists of lines that cross the boundaries of the hypothesis.

**Junction evidence.** Each junction that exists at the corner of a hypothesized parallelogram adds to the confidence of

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

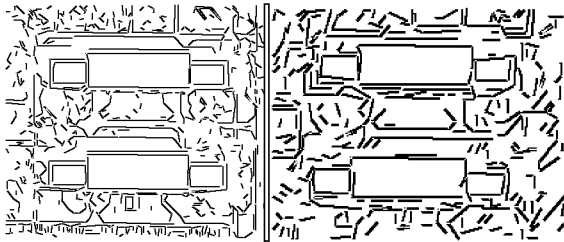


Figure 2 Lines detected in Figure 1

## 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 the views. The lines are grouped on the basis of colinearity, and are folded. Folding is the process by which lines with similar orientation are merged based on proximity. These lines are matched across all views. Multiple matches are retained at this stage. The constraint used in matching is the quadrilateral constraint described below; it uses the epipolar constraint and the restriction on the 3D height of a feature.

- Quadrilateral constraint

A line segment in a view (say View1), with endpoints, PointA and PointB, causes epipolar lines, LineA and LineB, respectively in another view (say View2). The restrictions on 3D height limit the extent of interest along LineA and LineB to SegmentA and SegmentB respectively. SegmentA and SegmentB will define a quadrilateral in View2, in general (a line segment, if they are colinear) that is searched for matches of the given line segment in View1. Each line that is found in View2 forms a line match with the given line in View1. This search is performed over all pairs of views.

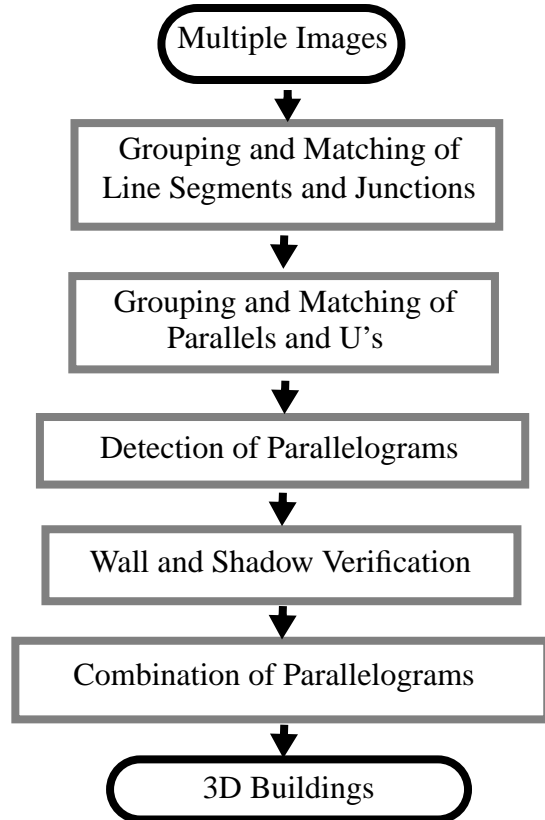


Figure 3 Block Diagram of the System

### 2.2 Junctions

Next, binary junctions (formed by the intersection of exactly two lines) are formed and matched. The constraints for junction matching are outlined below.

- Epipolar constraint  
Match of a junction (essentially a point) in one image, may occur only along the corresponding epipolar line in another image.
- Line match constraint  
Each line forming the junction in one view should be a member of a line match (as defined in Section 3.1) with exactly one of the lines forming the junction to which it is to be matched, in the other view.

# 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

(<http://iris.usc.edu/noronha/.netscape/results.html>)

Keywords: Stereo, Building Detection, Multiple Images

## 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 hypotheses 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

The task of detecting and describing buildings presents many challenges. 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. Some systems that use a single image, as input, have been developed ([5], [7]). In this work we use multiple images. 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], [9]) 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. 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

---

\*This research was supported in part by Contract No. 76-93-C-0014 from the Defense Advanced Research Projects Agency (DARPA) and monitored by the Topographic Engineering Research Center of the US Army, and by the US Air Force Office of Scientific Research.