

**Table 1:**

Figure	Line Finding (s)	Matching (s)	Grouping (s)
Figure 9	67	547	265
Figure 10	70	586	247
Figure 11	59	473	181
Figure 12	34	187	83
Figure 13	32	173	79
Figure 14	22	142	65
Figure 15	25	163	72

ous systems.

The system that has performance most similar to our system is the one described in [2]. The results obtained by that system on the Radius Modelboard images are qualitatively and quantitatively comparable to the ones that we obtain on the “modelboard” images (results of their system on other data are not available to us). There are many theoretical differences in the two systems. We treat the views uniformly, making building hypotheses by utilizing information in all views in a non-preferential manner, instead of starting with roof hypotheses from one view. Our system does not rely on *a priori* knowledge of the orientation of the buildings, or on the assumption that the roofs are parallel to the ground, as their system does.

## 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] U.R. Dhond and J.K. Aggarwal, “Structure from stereo - A review”, IEEE Transactions on Systems, Man and Cybernetics, 19(5):1489-1510, 1989.
- [4] R.B. Irvin and D.M. McKeown, “Methods for exploiting the relationship between buildings and their shadows in Aerial Imagery”, IEEE Transactions on Systems, Man and Cybernetics, 19(6):1564-1575, 1989.
- [5] H.S. Lim and T.O. Binford, “Stereo correspondence: A hierarchical approach”, Proceedings of DARPA Image Understanding Workshop, 234-241, 1987.
- [6] C. Lin, A. Huertas and R. Nevatia, “Detection of Buildings Using Perceptual Grouping and Shadows”, IEEE Proceedings of Computer Vision and Pattern recognition, 62-69, 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.
- [10] A.P. Witkin and J.M. Tenenbaum, “On perceptual organization”, From Pixels to Predicates, 149-160, 1986.

ing hypotheses. This example shows some difficulties of the task. There are a number of vehicles between the buildings, that are aligned in an almost-rectangular formation. While there are no trees and grassy areas in this part of the scene, the contrast between the tops of the buildings and the ground is low, leading to detection of lines that are fairly fragmented.

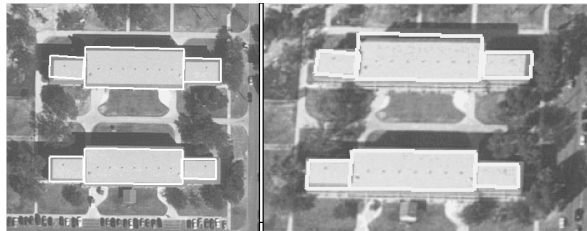


Figure 13a

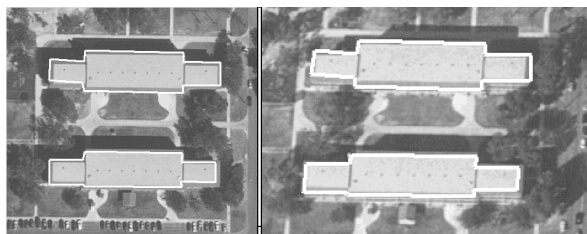


Figure 13a

Figure 13 Results on a section of Fort Hood

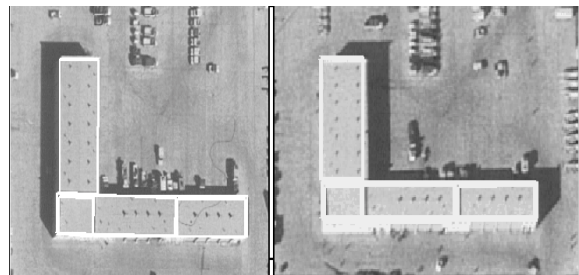


Figure 14a

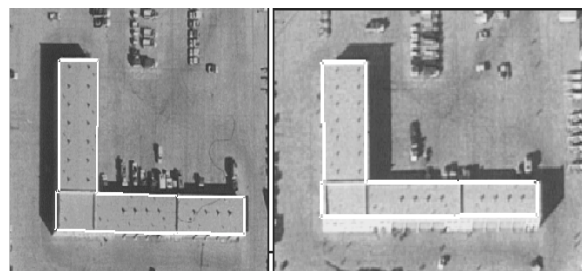


Figure 14b

Figure 14 Results on a section of Fort Hood

Figure 13, 14 and 15 provide instances of repeated application of the combination routine. Combination is

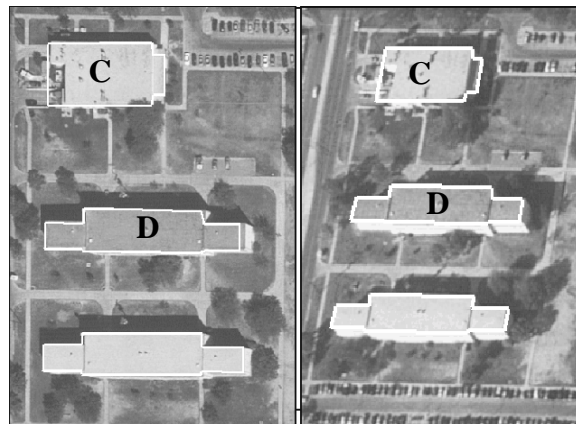


Figure 15 Results on a section of Fort Hood

performed based on proximity and overlap, till further combination is not possible, and Figure 12c shows the final results. Figure 13a shows verified rectangular building fragment hypotheses. Figure 13b shows the results after combination. Rectangular hypotheses are formed as a result of markings present on the rooftops of the buildings. Three rectangular hypotheses must be combined to form each building in this example. Figure 14a and Figure 14b show the rectangular building hypotheses, and the final building hypothesis respectively.

Figure 15 illustrates some failures of the system. The building labeled C in both views includes part of the ground on the left side, caused by the detection of a small protrusion from the building on that side. Figure 15 illustrates some failures of the system. The building labeled C in both views includes part of the ground on the left side, caused by the detection of a small protrusion from the building on that side. This problem may be solved by a more detailed analysis of the regions; this may require use of more sensitive feature detectors or other region analysis. The building labeled D in both views excludes a very narrow part of the building on the right, in the left view. This is as a result of a combination of inexact location of features (owing to errors in the location of the detected edges), and errors in the camera models, which lead to errors in matching. Better feature location would enable higher-confidence matching starting from the lowest level in the hierarchy of features. This would eliminate many incorrect matches, and allow the use of tighter tolerances in matching.

The results described above were obtained on a Sun Sparc station 20 with run times as indicated in **Table 1**:

We believe that the results we have obtained indicate the viability and potential of our approach though we anticipate further development of the system to handle more complex scenes. We believe that our hierarchical approach offers considerable advantages over previ-

cision 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.

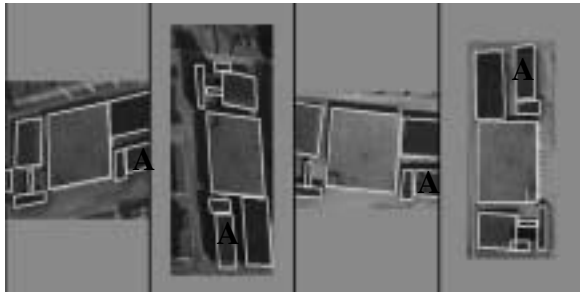


Figure 10 Results on a section of modelboard

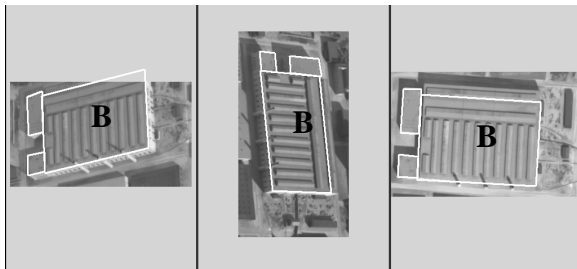


Figure 11 Results on a section of modelboard

## 8 Results and Conclusions

Our system has been tested on views of two different scenes so far. The first is that of a modelboard constructed for the RADIUS project. This scene is characterized by a dense array of buildings, aligned parallel to each other; however, it has no vehicles on roads or in parking lots and contains no vegetation. We have used up to four views of this scene in our experiments. One example was shown earlier in Figure 9. Other examples are illustrated in Figure 10 and Figure 11. We believe that the figures indicate good performance under fairly difficult conditions. Quantitatively, of the 27 buildings that the views we used covered, our system is able to detect 24 buildings. Three buildings are missed; the missed buildings are dark, and merge with their shadows. There is not enough boundary evidence to hypothesize the existence of these buildings. Of the 24 detected buildings, 21 seem to be detected accurately (by visual examination of the results). One building (labeled A in Figure 10) has its shadow included as well. It is a dark building that merges with its shadow in one of the views. Another building (labeled B in Figure 11) does not extend to the actual edge of the building. This is caused by numerous parallel lines competing to form the side of the building in all the views. A third building in the bottom right corner in Figure 9 is a multi-level building. The top level is delineated. The second level is discarded as it falls in the shadow of the top level in one view, has no shadow cast on its side in another view

(and hence cannot be verified from this view), and cast a shadow that is taken for the shadow of the top level in the third view.

The second scene is that of a military base in Fort Hood, Texas. This scene is more challenging than the modelboard, 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) in the scene. This causes height estimates to be less reliable and the selection process certain.



Figure 12a

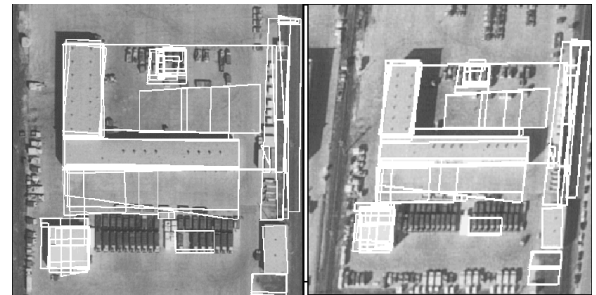


Figure 12b

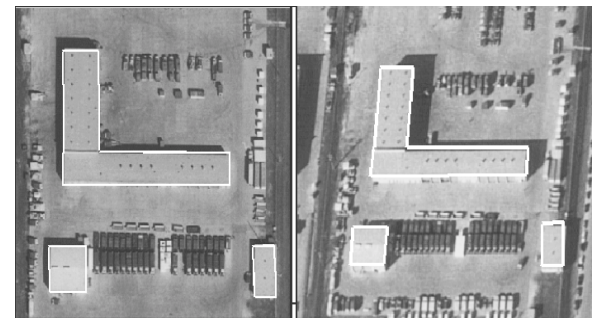


Figure 12c

Figure 12 Results on a section of Fort Hood

Figure 12 shows the results of a test on two of the Fort Hood views. Figure 12a shows the edges detected from both views, Figure 12b shows the selected build-

[6]. In our case, the analysis is made easier as we have an estimate of the height of the building *before* the need to search for shadows.

The search for shadows is carried out in a manner similar to that for the walls. Knowing the height of the building in 3D, and the direction of illumination, a search is performed to detect evidence of the predicted projection of the shadow. This includes the shadow cast by the roof that is detected, and the shadow cast by the vertical walls of the building. Occlusion of shadows by the building itself is taken into consideration when searching for shadows. The search for shadows in each view is carried out separately as the views are obtained with different sun positions. Hence shadows are strictly monocular cues.

The visible sides of the walls are dependent only on the 3D orientation of the building relative to the camera. The sides of the building for which the shadows are visible is dependent on the orientation of the building, and the direction of illumination. As these parameters (namely the viewpoint and the direction of illumination) are independent, it is possible that the shadows cast by the roof, and the vertical wall of the same side are visible on the same side of the building. In this case, the search is performed simultaneously. The shadow lines and the wall lines may be visible together, depending on whether the material of the building and the diffused light allow detection of the wall lines, which lie in the shadow area in this case.

The evidence of shadows and walls is accumulated for all the views, and a score is associated with the evidence detected. 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 numerical evidence for walls and shadows is compared against a threshold for acceptance or rejection. The threshold is set empirically from tests on sets of data from different scenes. Mathematically, if the wall evidence of a building,  $B_j$  in view  $i$  is  $wall_{ij}$ , and its shadows evidence is  $shadow_{ij}$ , then  $B_{ij}$  is verified iff

$$\sum_i (wall_{ij} + shadow_{ij}) \geq \text{Threshold (empirically set)}.$$

Figure 9 depicts the results obtained from the set of views shown in Figure 1. 13 of the 16 buildings have

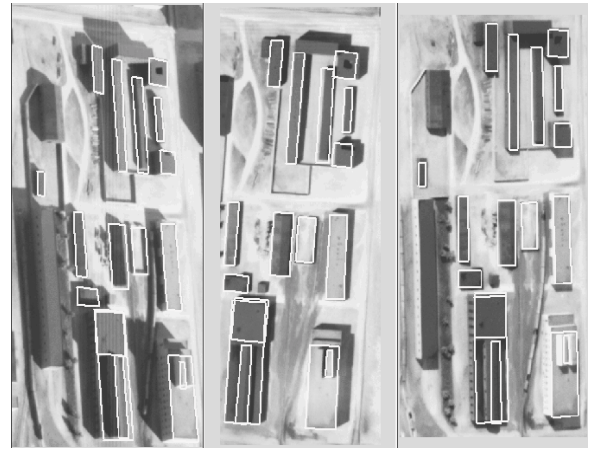


Figure 9 Results on the views in Figure 1

been detected. Some of the detected buildings have markings on the roof, and one is a compound building with a rectangular outline. A compound building in the lower right corner has the top (and dominant) level detected. The second (lower) level has an area that is smaller than the smallest expected buildings. The buildings that are not detected are dark, are those whose boundaries merge with their shadows. This makes detection difficult as U matches and parallel matches are not formed.

## 7 Combination of Rectangular Buildings

Some buildings are not rectangular, but can be decomposed into rectangular structures. 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. We consider the proximity and overlap criteria in the following paragraphs.

### 7.1 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.

### 7.2 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 de-

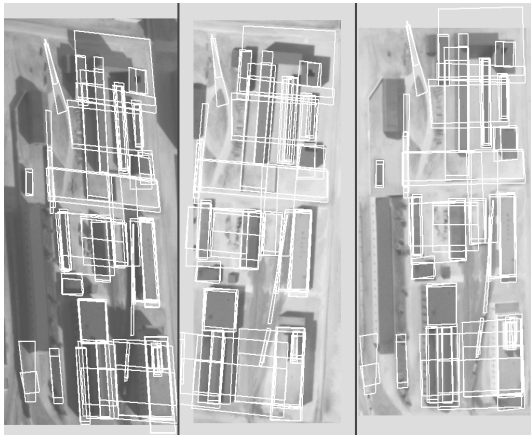


Figure 7 Selected parallelogram matches (building hypotheses)

rectangle is within the given height constraints. A check on the height constraint is necessary at this stage, even though the features forming the parallelogram match satisfy the constraint. This may be caused by the shifting of some individual parallelogram corners, as described in 4.5. If such shifting occurs, it is possible that the height constraint may not be satisfied.

### 5.2 Positive and Negative line evidence

If the height constraint is satisfied, a search is performed in each view for evidence supporting, or negating, the roof hypothesis. Lines that are found within a certain distance (which is a function of the image resolution) from the hypothesized parallelogram, and that differ in angle by not more than  $10^\circ$ , are considered positive evidence. Negative evidence consists of lines that cross the boundaries of the hypothesis. Figure 6 illustrates the concept of positive and negative evidence.

### 5.3 Orientation

The normal of the plane containing the roof hypothesis must make an angle of  $45^\circ$  or less, with the normal to the ground in 3D, to be considered for verification. The components of the parallelogram match making the roof hypothesis satisfy this constraint. However, owing to shifting of corners mentioned above, the roof hypothesis may not satisfy the constraint.

## 6 Verification of Building Hypotheses

So far, we have only used the evidence that comes from the component features of a single roof hypothesis. If we have indeed found a building roof, we should be able to find other features that come from the 3-D nature of a building. In our system, we look for evidence for walls and that of shadows cast by a hypothesized roof. In addition to the evidence of features supporting or negating roof hypothesis, we factor in statistical properties of the regions of the hypothesized roof and the shadows cast.

### 6.1 Wall evidence

In a view which is not nadir (and most views can be expected to be such), at least one and not more than two of the side walls of the buildings will be visible. These walls are assumed to be vertical. The verification for walls involves looking for the projections of the horizontal bottom of the wall (the interface of the vertical wall and the ground). At this point the hypothesized 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. Figure 8 illustrates this concept. 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 a function of the ratio of the length of the line coverage of the side to the length of the side.

As additional evidence for the presence of a wall, a

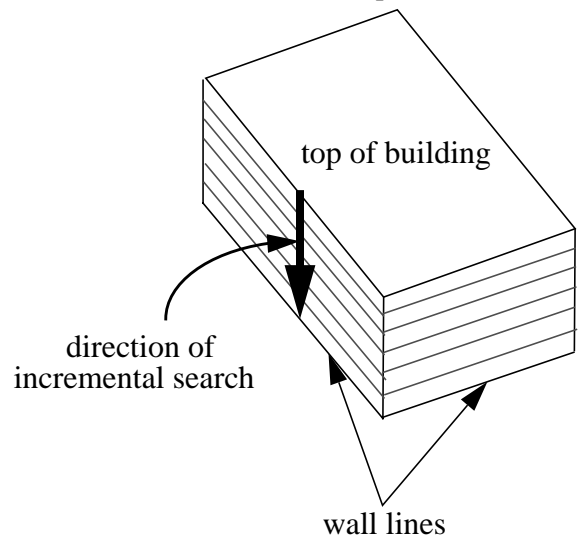


Figure 8 Search for walls

search for the projection of the visible vertical sides of the hypothesized building is performed. 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. Each vertical wall that is found increases the confidence of the hypothesis.

### 6.2 Shadow evidence

A 3-D building structure should cast shadows under suitable imaging conditions. We should normally possess knowledge of the direction of illumination from the sun, which in turn allows us to predict the location and orientation of shadows (on flat ground) from the 3-D hypotheses. If such shadows are found, our confidence in the hypothesis can be increased. Shadows have previously been used in monocular detection of buildings

## 4.5 Parallelograms

Formation of parallelogram matches is the basis for hypothesizing building roofs. To hypothesize a parallelogram match the minimal requirement is a **U** match (a match may be in 2 or more views), with additional evidence arising from one of the following four cases:

- an *opposing* **U** match from  $S_{um}$
- an *opposing* junction match from  $S_{jm}$
- an *opposing* **U** in a single view (from one of the  $S_{U_i}$ )
- an *opposing* junction in a single view (from one of the  $S_{J_i}$ )

Parallelogram matches are formed by running down the methods of formation in the order indicated. This order is not important, as the parallelogram matches are treated uniformly by the selection and verification mechanisms. Note that as the *opposing* features may not line up perfectly, we have imperfect parallelograms in 2D. This necessitates a *shifting* of the corners of the parallelograms in 2D, constrained by the requirement that they must be projections of a planar rectangle in 3D.

We define *opposing* **U** matches as follows. Suppose we have **U** matches  $U_{m1}$  and  $U_{n1}$  which belong to  $S_{um}$ . Denote their junction matches (real or virtual) as  $J_{m1}$  and  $J_{m2}$ ,  $J_{n1}$  and  $J_{n2}$  respectively.  $U_{m1}$  opposes  $U_{n1}$  iff exactly one of the following conditions is satisfied:

- there exist **U** matches ( $J_{m1}$ ,  $J_{n1}$ ) and ( $J_{m2}$ ,  $J_{n2}$ )
- there exist **U** matches ( $J_{m1}$ ,  $J_{n2}$ ) and ( $J_{m2}$ ,  $J_{n1}$ )

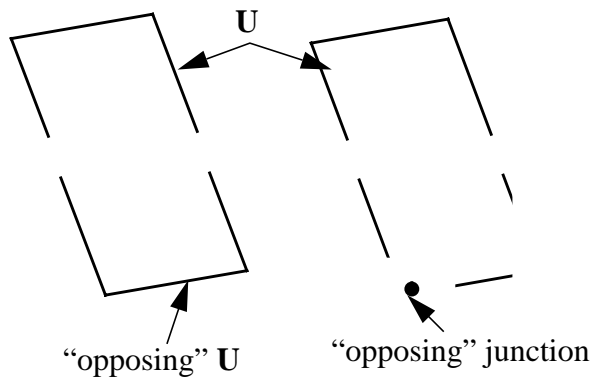


Figure 5 Opposing U and Opposing junction

The definitions of *opposing* in the other 3 cases are similar, and utilize the concept of alignment as defined in 4.4. Figure 5 depicts the formation of parallelogram matches in the last 2 cases outlined above.

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 views, we have chosen to ignore parallelogram evidence in one view alone, that is not

supported by any evidence detected in the other views. For example, it is possible to hypothesize a monocular parallelogram from 2 opposing **U**s in the same view, or from an opposing **U** and a junction, but we choose not to do so. Parallelogram matches across the views are the hypotheses of rectangular 3D buildings. The constraint used in parallelogram matching is:

### Planarity Constraint

In a manner similar to checking planarity of **U** matches, planarity of the parallelogram matches is tested. This is possible with parallelogram matches formed by opposing **U** matches, and those formed by a **U** match with an opposing junction match.

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. Decisions on the relative goodness of the hypotheses are delayed until the selection and verification of building hypotheses.

## 5 Selection of Roof Hypotheses

The parallelogram matches serve as roof hypotheses. They satisfy the constraints of being rectangular in 3D, and almost coplanar. In addition, the height and orientation with respect to the ground is known. We can use these parameters to distinguish among acceptable hypotheses and others, but height and orientation constraints are not sufficient because of inaccuracies of feature detection and camera models, and the possibility of erroneous matches giving rise to acceptable hypotheses accidentally. In addition to the 3D constraints that the roof hypothesis must satisfy, it must be acceptable in all the 2D views. This gives rise to the second constraint of the three described below. The hypotheses retained after selection are shown in Figure 7.

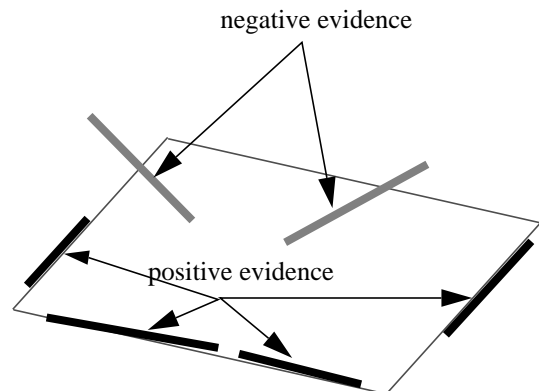


Figure 6 Positive and Negative line evidence

### 5.1 3D height

We check whether the height of the computed 3D

one of the following must hold: either there exist line matches  $(L_{im}, L_{kp})$  and  $(L_{in}, L_{kq})$  in  $S_{lm}$ , or there exist line matches  $(L_{im}, L_{kq})$  and  $(L_{in}, L_{kp})$  in  $S_{lm}$ .

### 3D Orthogonality Constraint

Given a junction match, we can compute the 3D angle between the lines forming it (from the knowledge of the matching lines). We require that this angle be between  $80^\circ$  and  $100^\circ$  in 3D.

### Trinocular Constraint

When there are more than 2 views available, the well-known trinocular constraint may be applied to the locations of the junctions.

### 4.3 Parallels

Next, we compute *parallel* pairs of lines and match the parallel pairs. Parallels are formed between pairs of lines,  $L_{ij}$  and  $L_{ik}$  in the same view<sub>i</sub>, that are separated by less than the maximum projected width of a building. While the task domain causes a large number of parallels in each view (two to three times the number of lines in that view), 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 view. A match is hypothesized if there is evidence in at least two views. When there is evidence in greater than 2 views, this forms a single parallel match in more than 2 views. The constraint used in matching is the parallel match constraint described below:

#### Parallel match constraint

Consider parallels  $P_{ik}$  with component segments  $L_{ik_1}$  and  $L_{ik_2}$  in view<sub>i</sub>, and  $P_{jl}$  with component segments  $L_{jl_1}$  and  $L_{jl_2}$  in view<sub>j</sub>. The parallel match constraint is satisfied for this pair of parallels iff exactly one of the following criteria is met:

- $(L_{ik_1}, L_{jl_1})$  and  $(L_{ik_2}, L_{jl_2})$  are elements of  $S_{lm}$
- $(L_{ik_2}, L_{jl_1})$  and  $(L_{ik_1}, L_{jl_2})$  are elements of  $S_{lm}$

In the case of parallels over more than 2 views, the parallel match constraint must be satisfied over parallels from every pair of views. Maximal parallel matches i.e. parallel matches that have the maximum number of parallels, are generated in order to ensure that duplicate parallel matches do not occur. Parallel matches over  $n$  views are represented as  $n$ -tuples. The set of parallel matches is denoted by  $S_{pm}$ . Note that as the line matches satisfy the quadrilateral constraint (they are constrained to being in a certain range in world  $z$  values) order reversal of the lines in the other views is automatically taken care of, if it should occur.

### 4.4 Us

Next we consider the formation of **U** structures. A **U** captures 3 sides of a parallelogram. **Us** are formed when two junctions are *aligned*. The definition of alignment is given below. **Us** are computed for each view<sub>i</sub>, to

form sets  $S_{U_i}$  ( $i = 1, 2 \dots \text{number\_of\_views}$ ). These sets are utilized in forming parallelogram matches as detailed in 4.5

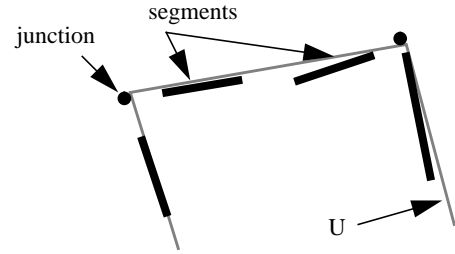


Figure 4 Aligned junctions

There are two ways in which the system hypothesizes **U** matches. The first way of hypothesizing **U** matches is by using 2 *aligned* junction matches. Junction matches  $Jm_p$  and  $Jm_q$  are *aligned*, if their component junctions are *aligned* in each view. *Alignment* of junctions is illustrated in Figure 4.

The second way that **U** matches are formed is by a parallel match with evidence of *closure* in at least one view. In this case we have two virtual junction matches hypothesized. These two virtual junction matches are hypothesized on the side of the **U** match where evidence of *closure* exists. The intersection of this *closure*, in each view (it is hypothesized in views where it does not exist), with the component parallel of the parallel match in that view, yields the virtual junctions that are components of the virtual junction matches.

Closure of a parallel is defined as follows. Let the line that extends further in the parallel be  $L_{ij}$ . Let the further endpoint of  $L_{ij}$  be  $P_{ij1}$ . Let  $L_{orth_{ij}}$  be the projection of the line in 3D that is orthogonal to the line in 3D that is parallel to the ground and projects to  $L_{ij}$ . The parallel is said to have closure iff there exist segments which form an angle of less than  $10^\circ$  with  $L_{orth_{ij}}$ , whose endpoints are at a distance of less than  $f(\text{resolution})$  from  $L_{orth_{ij}}$ , and whose perpendicular projections cover at least 50% of the distance between the parallel lines along  $L_{orth_{ij}}$ .

Denote the set of **U** matches by  $S_{um}$ . In the first case of **U** match formation (from two aligned junction matches), the following constraint should also be satisfied:

#### Planarity Constraint

Each junction match defines a plane in 3D. The planarity constraint checks that the planes of the junction matches forming a **U** match, are approximately coplanar (approximate coplanarity is defined to mean that the angle between the normals is less than  $10^\circ$ , and the distance is less than  $f(\text{resolution})$ , in each view).

place at various levels as well, and the results of matching at one stage are used for grouping at the higher levels. At each stage, some selections are made but the process is only intended to remove the hypotheses that become inviable with the increasing availability of context; at each stage, multiple hypotheses may remain even after selection.

Each hypothesis that is selected as being a candidate for being a roof based on the evidence formed by features in the multiple views is then *verified* by looking for supporting evidence from the visible walls and the shadows. Since we know the roof hypotheses in 3-D, we can predict the locations of the lines forming the wall boundaries as well as the shadows on ground (ground is assumed to be flat, though other kinds of known terrain could be included). Hypotheses with sufficient combined evidence form the output descriptions of our system. Our system does have the ability of providing confidence values for each object which may be useful for subsequent processes or humans that need to exploit the results. The confidence values are calculated based on the extent and accuracy of detected vertical walls and shadows cast by the roof, compared to their predicted locations.

## 4 Hierarchical Grouping and Matching of Features

In this section, we describe in detail the features used in our system including the methods for grouping and matching them. As described above, the system is hierarchical and uses evidence from all the views in a non-preferential, order-independent way.

### 4.1 Lines

Before matching, we first group line segments that are colinear. Segments are considered colinear if there is a *free path* from the end of one segment to the other i.e. no other segment blocks the line joining the two closest endpoints of the colinear segments, and if the angle between the segments is less than  $10^\circ$ . Colinearity is applicable to a set of greater than two segments as well. The above criterion must be met between every pair of neighboring segments.

After colinear grouping, the lines are tested for matches across the views by using the following quadrilateral constraint:

- Epipolar constraint

The match for a line segment in one view must lie at least partially within a quadrilateral owing to epipolar and 3D height constraints.

Each pair of lines that meet the quadrilateral constraint in any pair of views is determined to form a *line match* and included in the set of line matches that we will call  $S_{lm}$  and is passed to the higher levels for further processing.

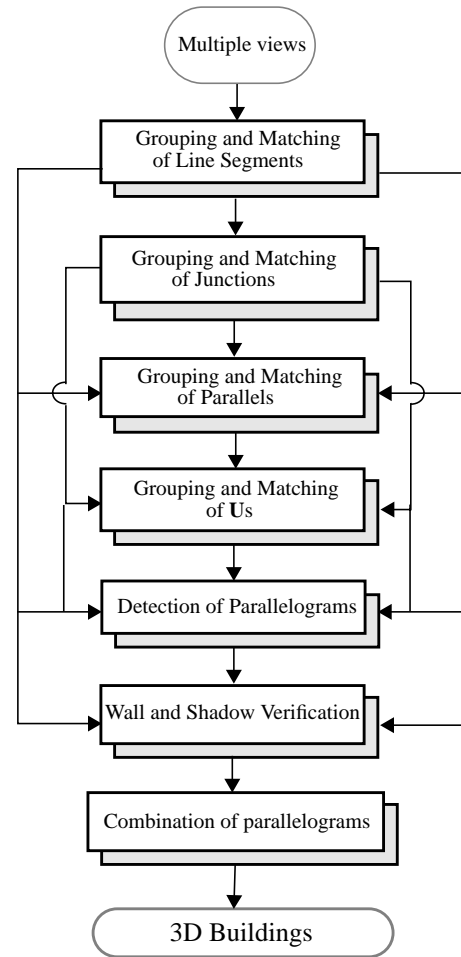


Figure 3 Block Diagram of the System

### 4.2 Junctions

Next, we consider matching of junctions formed by intersection of two lines. Consider a pair of lines  $L_{ik}$  ( $k = m, n$ ) in view<sub>*i*</sub>, with endpoints  $P_{ikl}$  ( $l = 1, 2$ ). Junction  $J_{ij}$  is formed at the intersection of  $L_{ik}$  ( $k = m, n$ ) iff the angle between  $L_{im}$  and  $L_{in}$  is greater than  $30^\circ$  and  $\min(\text{distance}(J_{ij}, P_{ik1}), \text{distance}(J_{ij}, P_{ik2})) \leq \text{length}(L_{ik})$  for ( $k = m, n$ ). Denote the set of junctions formed in view<sub>*i*</sub> by  $S_{J_i}$ . Junctions in the sets  $S_{J_i}$  ( $i = 1, 2 \dots \text{number\_of\_views}$ ) are then matched across the views to form a set  $S_{jm}$ , when the following constraints are satisfied:

#### Epipolar constraint

Given a junction  $J_{ij}$  in view<sub>*i*</sub>, its match in another view<sub>*l*</sub> must be within a certain segment (depending on the height range in 3D) of the epipolar line corresponding to  $J_{ij}$  in view<sub>*l*</sub>.

#### Line Match Constraint

If junction  $J_{ij}$ , formed by lines  $L_{im}$  and  $L_{in}$ , matches junction  $J_{kl}$ , formed by lines  $L_{kp}$  and  $L_{kq}$ , then exactly

to be present nearby in an urban scene where buildings are often parallel to each other, as are ancillary structures such as roads, sidewalks and landscaping.

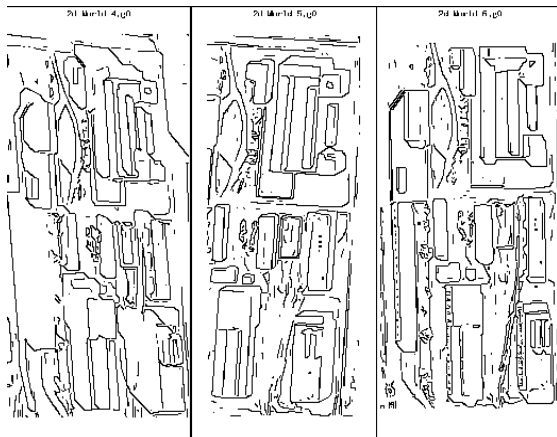


Figure 2 Detected lines from Figure 1

For such a problem, we suggest that the problems of matching and grouping (i.e. 3-D recovery and object segmentation) be not separated, but be solved simultaneously [5]. The difficulty with matching lower level features is that it is difficult to disambiguate the matches correctly; the difficulty at the higher levels is that the correct groupings may not be formed in the first place. We propose a hierarchical grouping scheme where lower level features are grouped into successively higher level features. At each level, the grouped structures are matched across the different views and only the consistent ones are retained. We attempt to recover roof structures first, as they form the dominant regions of the buildings in the projected images. However, final selection of roof hypotheses needs to take advantage of the context provided by the visible walls (which may be different in different views) and by shadows cast by them.

To simplify our task, we restrict the domain of buildings that we work with to rectilinear structures (i.e. those consisting of rectangular components). Further, we assume that the roofs are planar and that the walls are vertical. This allows us to make some predictions about the expected properties of the projected boundaries in the image as explained later in section 6. We also assume that the “camera models” are given, that is we can infer the epipolar geometry between the views and know the orientation with respect to a ground frame. Note that we do not require the different views be such that the epipolar lines are parallel, nor do we *rectify* the images to parallelize the epipolar lines.

In 2 we provide a brief review of previous work in this area. We give an overview of our approach in 3, with details following in 4 through 7. In 7, we present some results and conclusions.

## 2 Previous Work

Stereo analysis is a topic that has been studied for a long time in computer vision. It is not possible for us to give a detailed survey of previous stereo analysis work here; an excellent source of survey is [3]. In the following, we only comment on some previous work that has some close relations to our task or approach. A few researchers have advocated use of hierarchical and structural features in previous work, see for example [5]. The previous systems, however, assume views taken at the same time and were not oriented towards the task of building detection.

There have been some previous attempts more directly focused on detecting buildings from stereo images [1], [8], [9]. These systems, too, assume images taken at the same time. Largely, the previous systems match low-level features such as line and junctions and to attempt to infer buildings from the matches by some kind of tracing or grouping method. The system described in [8] matches higher level hypotheses (rectangles) but does not use stereo information to form the hypotheses themselves.

A recent system described in [2] does deal with the same kinds of imagery that we do (in fact, we use the same test data). However, the approach in this system is different in several ways. This system first uses a single view to determine roof outlines. Matches for these roof outlines are found in other views and heights are determined by peaks in a histogram of heights from different matches. This method has demonstrated very good results on one set of views. However, its performance may be critically dependent on the ability to generate good hypotheses from a single “seed” view (apparently a nadir view). This system also assumes that the orientation of the sides of the roofs in the image is known in advance.

## 3 Overview of the system

Our system, uses a hypothesis and verify paradigm. Roof hypotheses are formed by a hierarchical grouping and matching scheme and verified by using wall and shadow evidence. A block diagram is given in Figure 3. With the restrictions of rectilinearity in the shapes of the buildings our system is designed for, the roofs can be expected to project into parallelograms or a combination of them (we assume that projection is either truly orthographic or is approximately orthographic over the extent of a building; this is generally true of aerial images taken from a height substantially larger than the heights of the buildings). We form hypotheses for parallelograms in a hierarchical way, by forming lines, junctions, parallels, Us and finally the parallelograms themselves. Evidence from all the views is used to generate the groupings and the process is not dependent on the order in which the views are examined. Matching takes

# 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 described. The system uses a hierarchy of features for matching. Results of lower level are used for grouping at the higher levels are selected based on image evidence for them and are verified by using shadow and wall evidence.*

## 1 Introduction

Detection and description of buildings from aerial images is an important task in many areas of cartography, photo-interpretation and in other applications. The task is difficult due to a variety of reasons. Aerial images contain many structures, the images are typically quite complex, and the 3D structure is not explicitly given. Thus, we need to solve the usual problems of object segmentation (figure/ground separation) and of 3D recovery.

It is possible for us to recover the desired building structures from a single image and some automatic systems have been constructed for this task ([6], [7]). However, this is an extremely difficult task as only one view is available to resolve the ambiguities of segmentation, and 3D recovery must rely on shadows and projected lengths of vertical lines which may not be visible distinctly. For many applications, more than one view of the scene is available, which can simplify the task significantly. In this paper, we consider the case where two or more views are available but the views are not necessarily taken at the same time; hence, imaging conditions, including the sun position, the atmospheric conditions, and the environmental conditions, may be quite different.

Problems of segmentation and 3-D recovery are simplified by the presence of multiple views, but do not disappear completely. A simplistic view of multiple view processing would be that we could first recover a

dense 3D map by matching across the different views and then segment the desired structures in 3D. However, this is rarely possible in stereo processing and is particularly difficult for the problem being considered here. We cannot directly compute a dense 3D map of the scene as there are large homogeneous areas whose interiors can not be matched directly and we cannot match intensity values across images as they are not invariant with changing viewing conditions. Instead, what we can attempt to do is match features, such as object boundaries, that are invariant across the images. However, the set of such features will likely be sparse and fragmented, and we must group them [10], to infer coherent objects.

To illustrate the nature of the problem, consider three views of a scene shown in Figure 1. These views come from a modelboard, and are being used as standard test images by several researchers. Note that the sides of the buildings that are visible are not the same in all views, and that the shadows cast on the ground are quite different. Figure 2 shows the lines extracted from the Figure 1 images. Note that not all of these boundaries have correspondences in more than one view. Al-

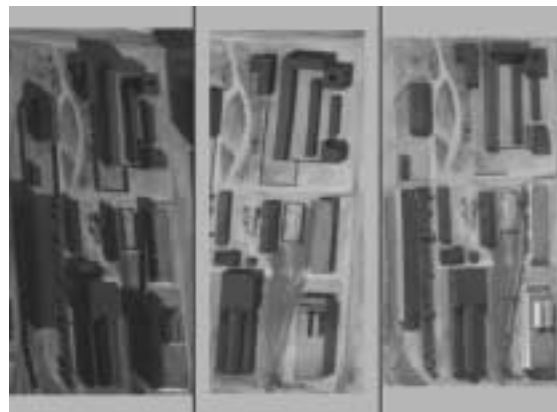


Figure 1 Three views of a scene

so, it is unlikely that we can find unambiguous matches even for those lines that do correspond just by looking at the lines individually, as many parallel lines are likely

---

\* 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.