Recovering Building Structures from Stereo *

Ronald C.-K. Chung and Ramakant Nevatia

Institute for Robotics and Intelligent Systems
School of Engineering
University of Southern California
Los Angeles, CA 90089-0273

E-mail: rchung@iris.usc.edu, nevatia@iris.usc.edu

Abstract

We address the problem of extracting polyhedral building structures from a stereo pair of aerial intensity images. We describe a system that computes a hierarchy of descriptions such as segments, junctions, and links between junctions from each view, and matches these features at the different levels. Such high level features not only help reduce correspondence ambiguity during stereo matching, but also allow us to infer surface boundaries even though the boundaries may be broken because of noise and weak contrast. We hypothesize surface boundaries by examining global information such as continuity and coplanarity of linked edges in 3-D, rather than merely by looking at local depth information. When the walls of the buildings are visible, we also exploit the relationship among adjacent surfaces in a polyhedral object to help confirm the different levels of descriptions. We give some experimental results for aerial images taken from overhead views and oblique views.

1 Introduction

A basic goal of computer vision is the capability of extracting structural descriptions of objects in an imaged scene. A typical application domain of such a capability is to detect architectural structures from aerial images. With stereo images not only can the problem be potentially made easier because of the additional information, three-dimensional (3-D) depth information about the scene can also be estimated quantitatively. In this paper we address the structural description problem in the context of extracting polyhedral building structures from a stereo pair of aerial images.

Use of stereo is common to recover 3-D information about a scene using multiple intensity images. While much work has been done on the correspondence problem to reconstruct depth estimates, few have addressed the problem of recovering structural descriptions from stereo. It is generally believed that dense depth map can be recovered through the coordinated effort of stereo matching of some primitives features and surface interpolation, from which depth and orientation discontinuties in the scene can be extracted from the local depth differences. Here we take a different point of view.

First, it is well known that correspondence ambiguity can be a problem in matching primitive features, especially when there are repetitive patterns in the scene. In urban scenes it is not uncommon to find the roof edges of the buildings accompanied by other parallel edges next to them, for example, those projected by pavement edges, lane markings on the roads, and even shadows cast by the buildings themselves. Features matched in stereo have to be distinct enough not to be confused by the parallel structures commonly occuring in urban scenes.

Second, it is usually assumed that depth and orientation discontinuities can be located either by setting some thresholds during the surface interpolation step, or by looking at the disparity differences between adjacent edge elements (*e.g.*, line processes [18]). However, if the scene is not densely textured, which is the case for many architectural scenes, even if we can assume the initial depth estimates are perfect, surface interpolation without knowing which side of an edge is the occluding surface would lead to erroneous result. In addition, the sparcity of features would present difficulties in adjusting the thresholds to detect discontinuities (points on a 1-D step function will be more "collinear" when they are farther apart). We conjecture that this is where localizing discontinuities from multiple intensity images is different from, and more difficult than, that from direct range data.

Our approach is to exploit the structural features to recover discontinuity information, without solely relying on local depth differences between adjacent features in the scene. One important observation is that a closed contour rarely occurs in 3-D, and if it does, it has to be either one of the three possibilities: (1) the boundary of a surface patch; (2) the boundary of a set of connected surface patches; (3) the boundary of a closed surface

*This research was supported in part by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under Contract No. F49620-90-C-0078, and in part by a subcontract from the Hughes Aircraft Company. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

0-8186-2840-5/92 $03.00 © 1992 IEEE

marking. Moreover, there are some intrinsic relationships among these possibilities. For example, a surface marking, as its name implies, is always contained in a surface patch (which can be the background) or a set of connected surface patches, and a set of connected surface patches usually compose a solid object. The key is how we can capture such relationships to infer where the actual surface boundaries are.

We therefore propose to hypothesize closed contours in 3-D as the surface boundaries, and when there are conflicts, to select among the hypotheses based on their individual merits and the relationships displayed among them.

However, edges projected by surface boundaries are usually far from perfect and they may be broken because of noise and weak contrast. The choice of the structural descriptions in each view that guide the stereo correspondence process and are also confirmed by it, should also allow us to infer the continuity of the surface boundaries. Let us also point out that such a continuity inference process is indeed unavoidable. Even if discontinuities can be identified by looking at local depth measurements, discontinuities so identified are also likely to be broken and sparse.

For generic surfaces, the only property we can make use of in capturing the continuity of surface boundaries is co-curvilinearity. It is easier if polyhedral structures can be assumed, as surface boundaries can then be broken down into merely *junctions* and *links*, where a link is the collection of collinear edges between two junctions. For polyhedral structures we can further constrain the number of surface boundary hypotheses by enforcing that the boundary of a hypothesized surface patch has to be planar in 3-D. This removes the possibility of extracting the boundary of a set of connected surfaces as in the case of generic structures.

Our approach can therefore be summarized as these:

1. we propose a stereo system that computes a hierarchy of descriptions such as *edges*, *line segments*, *junctions*, and *links* between junctions from each view, and matches these features at the different levels;
2. we hypothesize surface boundaries from such matched and confirmed structural descriptions based on *continuity* and *coplanarity*, and select among these hypotheses based on the the individual merits of the hypotheses and the relationships displayed among them.

Such approach requires checking collinearity and coplanarity of linked edges in 3-D whose depth information is estimated using the triangulation method, a process known to be error-prone. We will show that we can reduce the error made in inferring 3-D information from stereo correspondences at this stage by translating 3-D collinearity into 2-D collinearity in the two views, and coplanarity from the 3-D space into the disparity space, and working merely in the 2-D and the disparity domains.

We first outline in section 2 some previous work on stereo analysis, and the emphasis would be on those dealing with aerial images and recovering building structures. We describe how building structures can be extracted from stereo in section 3 and present some experimental results in section 4. Finally we give the conclusion in section 5.

2 Previous Work

To deal with the stereo correspondence problem, features of various abstraction in combination with various sets of constraints have been proposed for matching. A good survey for recent literatures can be found in [5]. Two principal approaches are generally used: area-based matching or feature-based matching.

Area-based matching attempts to match small windows from the left and right views by correlating their intensities or the derivatives of their intensities. Representative examples are systems of [11; 4]. Area-based matching has the advantages that the features being matched are simple to extract, and it delivers a relatively dense depth map. On the other hand, it requires presence of significant texture in the scene. As features being matched correspond directly to intensity variations, area-based matching is sensitive to photometric distortions and camera noises, and thus unsuitable for stereo images taken with long baselines. The presence of occluding boundaries in the correlation window will also confuse the correlation-based matcher, giving an erroneous depth estimate.

Feature-based matching uses structural features rather than image intensities for stereo matching, and hence is more stable toward photometric variations and capable of handling non-textured scenes. Most [9; 6] match zero-crossings or other types of edge elements. As low-level features contain relatively little information to resolve correspondence ambiguity, Medioni and Nevatia [15] have suggested using segments for stereo matching. Ganapathy [8] has also implemented a system for matching junctions in perfect line drawings of polyhedral objects.

Systems have also been developed specifically for architectural scenes, and task-specific knowledge is generally employed for stereo correspondence. For example, Liebes, Baker *et al.* [14; 1] have made use of orthogonal trihedral vertices (OTVs) for stereo correspondence. Herman and Kanade [10] incrementally update the 3-D model of a scene from successive views of the scene using stereo, and they assume the vanishing point of the vertical lines in the images is known beforehand.

Mohan and Nevatia [16] have built a system to recover buildings from aerial images. Yet the system extracts structural features all the way from edges to surfaces in each view independently without exploiting the existence of stereo views during the process, and it matches the structural features only at the highest level. It is also restricted to buildings of rectangular structures.

Fua and Hanson [7] have also built a system to reconstruct building structures from stereo. They model the roofs of buildings as rectilinear objects which have planar intensity distributions and are planar in 3-D. An objective function is thus defined for any roof hypothesis to represent how close is the hypothesis from the predefined building model. An initial roof hypothesis can therefore be gradually conformed to optimize the objective function using some energy-minimizing techniques. The initial hypothesis can be given by a human operator, or be extracted bottom-up from edges in the images. The system does extract surfaces in the scene, yet in semi-automatic mode it requires operator-guided cueing, and in automatic mode the hypotheses are generated from intensity regions or orthogonal edges which do not always have direct relationship with actual surface boundaries.

Hsieh *et al.* [13] have also built a system to combine edgel-based matching and area-based matching approaches for stereo matching, yet the output still falls short of explicitly recovering surface boundaries in the scene.

3 A 3-D Structural Description System

Aerial images can be taken from *overhead* views or from *oblique* views, where walls of buildings are visible in the latter but not in the former. We will first assume the images are taken from overhead views and present how buildings can be recovered from stereo. We will also show how the the process can be reinforced if images taken from oblique views are given.

An overview diagram of our stereo system is shown in Figure 1. It is basically composed of three interrelated modules: (1) **Structural Descriptions & Matching**: edges, line segments, junctions, and links are extracted and matched across stereo views; (2) **Figure Extraction**: surface boundaries are hypothesized from the confirmed links and junctions, and conflicts are resolved; (3) **Ground Extraction**: the ground level, assumed to be planar, is recovered as the plane that contains most of the matched features outside the recovered surface boundaries; surface boundaries extracted from the figure extraction module which are not above the ground level are in turn regarded as markings on the ground and discarded.

We will describe these modules in more details in the following sections. Notice that various levels of our stereo system require matching structural features of different levels in the two views, and constraints for the matching process can be formulated as unary and binary constraints among the possible matches. Unary constraints come from individual merits of an entity, whereas binary constraints relate a pair of entities. The constraints can be excitatory (positive) or inhibitory (negative), and can even be absolute such as enforcing mutual exclusion among some entities. We use a relaxation network described in appendix A to accomplish the tasks. In the rest of the thesis, when we describe the constraints for matching certain structural features, we will also indicate the nature of the constraints as being excitatory, inhibitory, or absolute.

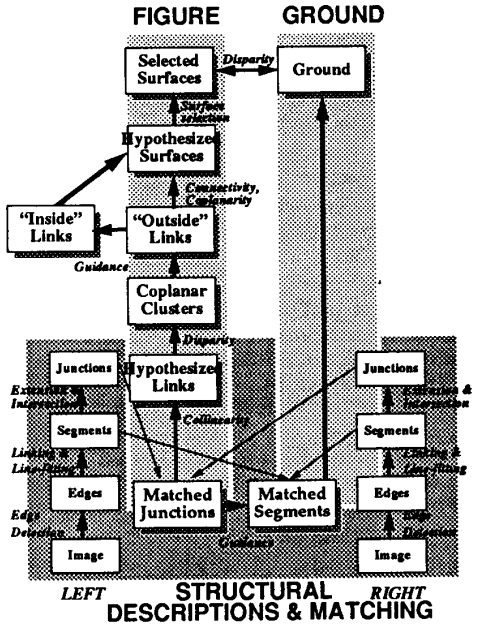


Figure 1: Overview of our stereo system.

3.1 Structural Description and Matching Module

Edges are extracted from each image using Canny's edge detector, and they are linked and fit to become chains of *line segments* using techniques of [17]. To make later processings easier, we extend each line segment from its two end-points for a limited length, and break any other line segments that cross over to the extensions. We then hypothesize an *L-junction* from each pair of line segments using the following criteria (Figure 2): (1) the two line segments are not parallel; (2) none of the line segments contain the point of intersection between them; (3) the paths from the end-points of the line segments to the point of intersection have to be free of other edges; (4) The distances from the end-points of the line segments to the point of intersection are within a threshold (60 pixels). We then say an L-junction is formed at the intersection of the two segments, and the segments are called *branches* of the junction.

Notice that if the images are taken from oblique views, at the projection of a multihedral vertex several L-junctions may be hypothesized at the same location by different combinations of the branches. As the junction-type and the number of branches of a junction may change across stereo views (*e.g.* , from "W" in the left view to "L" in the right view, see [2; 3]), and also what we are interested in is not just matching junctions but matching branches as well, here we

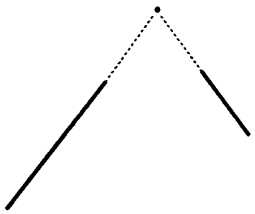


Figure 2: Hypothesis of an L-junction from two line segments.

treat each of these L-junctions separately in the matching phase. Such L-junctions will eventually hypothesize different faces of the vertex, and such information will be made use of in the surface boundary selection process described in section 3.2.

Such descriptions are extracted monocularly, and we match them across stereo views to confirm them and to maintain a consistent interpretation of the scene. Since junctions are more distinct and they capture information among the line segments, junctions, as well as the branches composing them, are first matched in a hierarchical manner using the following constraints.

Constraints for Matching Junctions:

1. Unary Constraints
   - *Epipolar constraint* (absolute): Two junctions are matchable only if they fall on corresponding epipolar lines.
   - *Hierarchical constraint* (absolute): Two junctions are matchable only if there exists at least one way to match all their component branches (as explained below in matching branches).
2. Binary Constraints
   - *Uniqueness constraint* (mutually exclusive): An L-junction can be matched with at most one L-junction. Junction-matches which share the same junction either in the left or in the right views are therefore mutually exclusive with one other.
   - *Ordering constraint* (inhibitory): If the component branch matches of two junction matches have order reversal along the epipolar lines, the junction matches desupport each other.
   - *Figural Continuity constraint* (excitatory): If junction l is matched with junction r, it is preferred that junctions with branches collinear with the branches of junctions l and junction r in the two views are also matched, if those junctions are indeed matchable (Figure 3).

As there are altogether two compromisable constraints, one predefined weight in the weighted sum of the constraints is necessary for the cost function to be optimized in the relaxation. The cost function is therefore:

$$E(V) = -\frac{1}{2}\sum_{ij} V_i V_j (\mathrm{FCC}_{ij} + w_{\mathrm{OC}}\mathrm{OC}_{ij})$$

where w_{OC} is kept constant at -2 in our system. The value is designed according to the relative importance of the constraints.

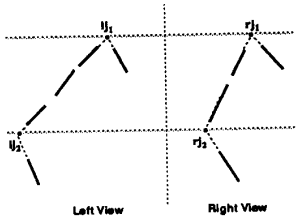


Figure 3: Figural Continuity constraint for matching junctions: matches (lj_1, rj_1), (lj_2, rj_2) support each other.

Constraints for Matching Branches of Junctions:

1. Unary Constraints
   - *Epipolar constraint* (absolute): Two branches are matchable only if they are either both pointing up or both pointing down across the epipolar lines. If the branches are almost parallel to the epipolar lines, they have to be pointing both toward the left or both toward the right.
2. Binary Constraints
   - *Uniqueness constraint* (mutually exclusive): One branch can be matched with at most one branch in the other view.
   - *Surface-orientation constraint* (mutually exclusive): The branch-matches associated with a junction-match have to be such that the cross-product of the two branches in a view must point to the same direction (either in or out of the image plane) as that of the corresponding branches in the other view (Figure 4), so as to ensure the same "face" of any physical surface in space is matched. This implies branches in the same epipolar band have to be matched in the order from left to right.

As there is no compromisable constraint, no weight in the weighted sum of the constraints is necessary for the cost function to be optimized in the relaxation.

The matched junctions and the corresponding branches are then used to guide the matching of the rest of the line segments. Line segments which are collinear with the matched branches in both views are extracted and regarded as correct matches. All these "correct" branch-matches and segment-matches are then used to lock on the segment matching process, in the sense that any other possible segment-matches which are mutually exclusive with them are discarded, and the rest of the

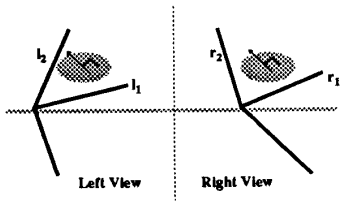


Figure 4: Surface-orientation constraint for matching branches of junctions: l_1 has to be matched with r_1, l_2 has to be matched with r_2, so that $l_1 \times l_2$ and $r_1 \times r_2$ point to the same direction (out of paper).

segment-matches go through the relaxation process by taking into account the information of the "correct" matches. The following shows the constraints used for matching line segments.

Constraints for Matching Line Segments:

1. Unary Constraints
   - *Epipolar constraint* (absolute): Two line segments are matchable only if there is overlap between the epipolar bands containing them.
2. Binary Constraints
   - *Uniqueness constraint* (mutually exclusive): A line segment cannot be matched with more than one line segment in the other view which have overlap in their epipolar extents.
   - *Ordering constraint* (excitatory): If line segment l is matched with line segment r, it is preferred that line segments on the left(right) of line segment l are matched with line segments on the left(right) of line segment r.
   - *Surface Continuity constraint* (excitatory): If line segment l is matched with line segment r, it is preferred that an immediate neighbor on the left(right) of line segment l is also matched with an immediate neighbor on the left(right) of line segment r, if they are indeed matchable.
   - *Figural Continuity constraint* (excitatory): If line segment l is matched with line segment r, it is preferred that line segments collinear with line segment l are also matched with line segments collinear with line segment r.

As there are altogether three compromisable constraints, two predefined weights in the weighted sum of the constraints are necessary for the cost function to be optimized in the relaxation. The cost function is therefore:

$$E(V) = -\frac{1}{2}\sum_{ij} V_i V_j (\mathrm{FCC}_{ij} + w_{\mathrm{scc}}\mathrm{SCC}_{ij} + w_{\mathrm{oc}}\mathrm{OC}_{ij})$$

where w_{scc} and w_{oc} are kept constant at 1 and 2 respectively in our system. The values are designed according to the relative importance of the constraints.

Such junctions and line segments confirmed in both views are then used for subsequent processes.

3.2 Figure Extraction and Ground Extraction Modules

We do not wish to rely on observing local depth measurements to recover surface boundaries, as we expect the scene may not be densely textured enough. Instead we use the properties *coplanarity* and *continuity* of the surface boundaries to hypothesize them. Junctions and line segments extracted and confirmed in the stereo views should allow us to capture such properties in 3-D.

Since each matched junction itself already defines a plane in 3-D, we can start with the matched junctions and cluster all matched junctions and segments into different sets such that entities in each cluster are all coplanar, and we extract possible surface boundaries from each cluster in turns. However, branches of junctions are usually short, rendering their 3-D information unreliable.

As a surface boundary can be broken down into junctions and links, we propose to first extract *links* between matched junctions to capture the collinearity information of the edges. A link is created between two junctions if the following criteria are satisfied (Figure 5): (1) there exists branches one from each junction that are collinear; (2) the corresponding branches of the corresponding junctions in the other view are also collinear; (3) there exists enough edgel evidence (50% of the distance between the junctions) to support the formation of the links in both views, *i.e.*, there are enough segments that lie between the junctions in both views.

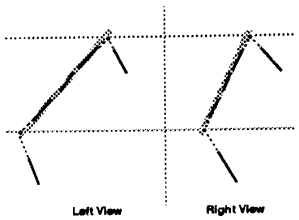


Figure 5: Hypothesis of a link-match from two junction-matches.

Since links are much longer and they capture information between connected junctions, their 3-D information are more reliable. Two links joining at a junction together with their correspondences in the other view define a plane in the 3-D space. We start with pairs of connected links to define planes in 3-D, and cluster all the links into different sets such that links in each cluster are all coplanar.

Surface boundaries can then be hypothesized by extracting sets of connected links in each cluster. However, computing all possible combinations of connected links can be exhaustive. In fact what we need from a set of coplanar links is not any inside closed contours, as those are formed from surface markings, but the contours composed from the "outsidemost" links which embed the rest of the links.

We use a simple algorithm to extract the outsidemost

links in each cluster. Once a center point is picked in the image plane, the two end points of any link will define a sector in the 2-D space with respect to the center point. The link will divide the sector into two zones, one outside and one inside with respect to the center point. A link will be an *outsidemost* link if the outside zone of its sector contains no other links in the cluster. Similarly, a link will be an *insidemost* link if the inside zone of its sector contains no other links in the cluster. To extract the outsidemost links in a cluster, we pick the end point of one of the links as the center point.

We start with only outsidemost links to extract surface boundary hypotheses based on continuity. When there are multiple links attached to the same end of a link, we trace through all possibilities to increase the robustness of the process. Each surface boundary hypothesis therefore contains at least one of the outsidemost links in the cluster.

There may be separate roofs in the scene that are coplanar, and we execute the above process recursively: for each cluster, we extract the outsidemost links, recover chains of links containing them, remove all links enclosed by the chains, and then we extract the outsidemost links from the rest of the links and execute the process again, until no more links are left in the cluster.

Edges are usually broken and sometimes totally missing in real images. Because of this, we do not expect all of the surface boundaries hypothesized from the above process to be closed. We use the following simple criteria to close an open surface:

1. If the branches of the junctions at the two open ends of the boundary are collinear with each other, and the edgel support across the opening exceeds a certain threshold (50% of the gap length), the opening is closed.
2. If the branches of the junctions at the two open ends are not collinear with each other but their extensions intersect on corresponding epipolar lines in the two views, then we look at how much edgel support there is along the paths from the two open ends to the point of intersection in each view. If the edgel support exceeds a certain threshold (50% of the total path length), the opening is also closed.

Such surface boundary hypotheses are likely to have conflicts among them, and a selection needs to be made. Ideally, the hypotheses should only be either actual surface boundaries or surface markings, since the possiblity of the boundaries of a set of connected surfaces have been ruled out by the planarity criterion. However, as we do allow gaps in the boundaries, some hypotheses may be constructed across collinear edges of different surfaces which happen to be coplanar.

To rule out surface markings from actual surface boundaries is simple: surface markings are contained in larger surface boundaries which are coplanar with them. As a result, among a set of coplanar surface hypotheses, the ones that enclose more area and are not contained by others are more likely to be actual surface boundaries. This is formulated as a weak constraint called the "Outsidemost-boundary constraint" outlined later in this section. It is more involved to distinguish between individual roofs and the false boundaries across different coplanar roofs. We basically rely on edgel support along the boundary of the hypotheses to make the decision.

As a summary, the constraints used to resolve conflicts among the hypothesized surface boundaries are given below.

Constraints for Selecting Surface Boundaries:

1. Unary Constraints
   - *Boundary-evidence constraint* (absolute): The fraction of edges detectable along the surface boundary has to exceed a certain threshold.
   - *Regularity constraint* (excitatory): A surface boundary hypothesis is considered more likely to be an actual surface boundary if: (1) it is made from two sets of parallel links (skew symmetrical); or (2) there are more than three junctions on the boundary, and there exists a circle containing all the junctions such that all the junctions are evenly distributed along the circumference of the circle (rotational symmetrical). To check this, we use the centroid of the 2-D positions of the junctions to hypothesize the center of the circle.
   - *Outsidemost-boundary constraint* (excitatory): A surface boundary hypothesis enclosing more area is considered more likely to be an actual surface boundary.
2. Binary Constraints
   - *Uniqueness constraint* (mutually exclusive): Surface boundary hypotheses are mutually exclusive with one another if: (1) there is overlap in their enclosed area, as we assume all surfaces are opaque; or (2) they share part of their boundaries and are coplanar with one another, as they should be absorbed into one single surface patch.

If the images are taken from oblique views, *i.e.*, the walls of the buildings are also visible, we have one more constraint: selected surfaces should compose feasible solid objects. This is formulated simply as a strong binary constraint added to the surface-selection process:

- *Solid-formation constraint* (excitatory): Two surface boundary hypotheses support each other if (Figure 6): (1) they share one and only one link; and (2) they are not coplanar with each another.

As there are altogether three compromisable constraints, two predefined weights in the weighted sum of the constraints are necessary for the cost function to be optimized in the relaxation. The cost function is

therefore:

$$E(V) = -\sum_i (V_i \mathrm{BEC}_i) - \frac{1}{2}\sum_{ij} V_i V_j (w_{\mathrm{rc}} \mathrm{RC}_{ij} + w_{\mathrm{sfc}} \mathrm{SFC}_{ij})$$

where w_{rc} and w_{sfc} are kept constant at 1 and 3 respectively in our system. The values are designed according to the relative importance of the constraints.

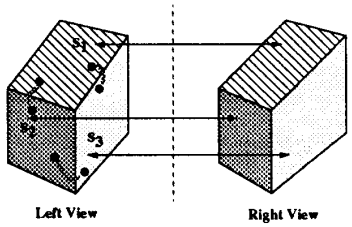


Figure 6: Solid-formation constraint for selecting surface boundaries: surface hypotheses s_1, s_2, and s_3 support each other pairwisely.

There are buildings whose roofs are not a single connected surface patch, but with "holes" in it. An example is the Pentagon building shown in Figure 8. An evidence of existence of a hole is that there are features which are inside the recovered surface boundary but are not coplanar with it, and the boundary of the hole is formed by the "insidemost" links coplanar with the recovered surface boundary.

For each recovered surface boundary, we therefore check if there are segment matches inside but not coplanar with the surface boundary. If there are, we first use an end point of one of the segment matches as the center point to extract the *insidemost* links coplanar with the surface boundary. Hole boundaries are then hypothesized based on the continuity of the insidemost links, and the one with smallest area is taken as the hole boundary.

Such an approach requires checking collinearity and coplanarity of linked edges in 3-D whose depth information is estimated using the triangulation method, a known error-prone process. The translation of 3-D collinearity into 2-D collinearity in the projection through any viewpoint is well-known. We also show in appendix B that coplanarity in the 3-D space is also preserved in the disparity space. Such properties allow us to reduce the error made in inferring 3-D information from stereo correspondences by working in the 2-D and the disparity domains.

We assume the ground level is planar, and we recover it as the plane containing most of the matched features outside the recovered roof boundaries. Surface boundaries extracted from the figure extraction module which are coplanar with the ground level are in turn regarded as markings on the ground and discarded.

4 Experimental Results

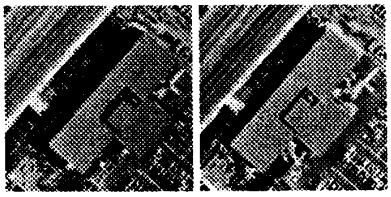
We have tested our system on five sets of images; three of them taken from overhead and oblique views are shown in Figures 7, 8, and 10 respectively. We extract *edges*, *line segments*, *L-junctions* from the images, and match them hierarchically across the stereo views. *Links* are then hypothesized from pairs of junctions whose branches are collinear. We cluster the links into different sets, such that links in a cluster are all coplanar. Outsidemost links are extracted from each cluster and *surface boundaries* are hypothesized based on the continuity of links. Boundary hypotheses which are open are closed if they display skew symmetries or rotational symmetries as defined in section 3.2. Surface boundaries are then selected from the hypotheses based on the individual merits of the hypotheses and the relationships displayed among them.

The Pentagon building shown in Figure 8 has been a popular example used by many stereo systems. However, most stereo systems merely recover a coarse depth map but not surface boundaries in the scene. The notion of the existence of a hole on the top of the building, in particular, has mostly been ignored. To recover the roof boundary of the Pentagon building is particularly difficult as there are a lot of coplanar features inside the roof. Hypothesizing surface boundaries from every possible combination of connected links would lead to a combinatorial explosion. Our system is capable of using the 3-D information from stereo to group all the links on the roof into one single cluster, and the outsidemost links of the cluster are extracted from it. Our system also recognized that the outsidemost links form a closed boundary which also displays rotational symmetry. The boundary is therefore taken as the boundary of a surface patch in the scene. In Figure 9 we also show the intermediate steps of how the hole on the roof of the building can be recovered.

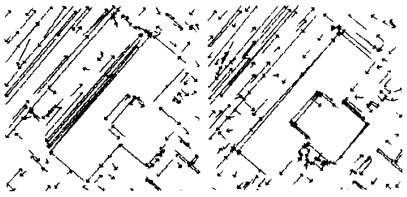
Since currently we do not have real image data in stereo pairs taken from oblique views, we have taken some stereo pictures ourselves in our laboratory for testing purposes. We put a toy building on top of a blow-up of a typical aerial picture and take the images with a camera mounted on a linear table. An example is shown in Figure 10. Such images are not completely realistic, yet they capture some of the basic characteristics of real images taken from oblique views. As in the above example, we show the intermediate steps and the performance of our stereo system. A major characteristic of oblique views is that multiple neighboring surfaces composing the same solid object may be visible, and their relationship among one another is exploited in our system in selecting among the surface boundary hypotheses.

5 Conclusion

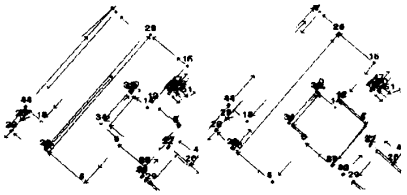
We have designed a stereo system that recovers surfaces from a stereo pair of intensity images. The system is geared toward recovering surfaces of building structures which display high degree of regularity. Experimental results for aerial images taken from overhead views and oblique views are also shown.



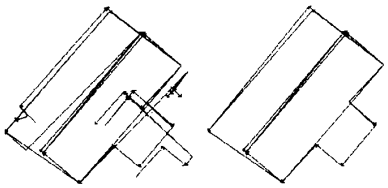
(a) left image (b) right image



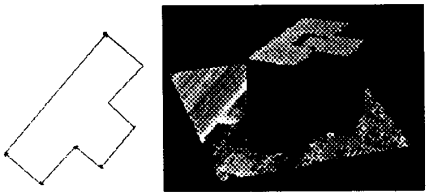
(c) left segments (d) right segments



(e) matched junctions (left) (f) matched junctions (right)

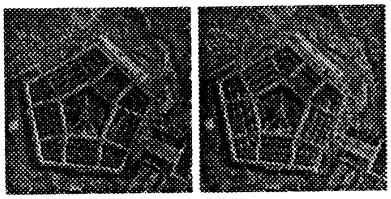


(g) links (left) (h) closed surfaces (left)

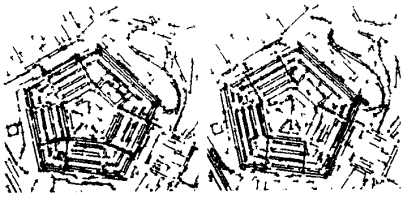


(i) selected surfaces (left) (j) 3-D Rendered View

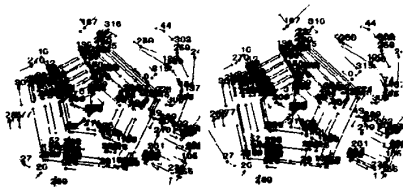
Figure 7: Results for the scene of b10.



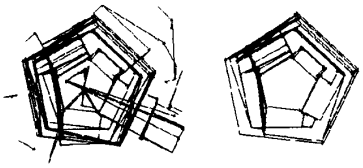
(a) left image (b) right image



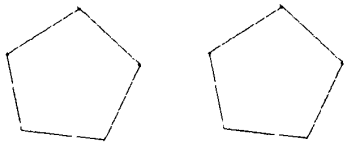
(c) left segments (d) right segments



(e) matched junctions (left) (f) matched junctions (right)



(g) links (left) (h) one coplanar cluster of links (left)



(i) outsidemost closed surfaces for the cluster (left) (j) selected surfaces (left)

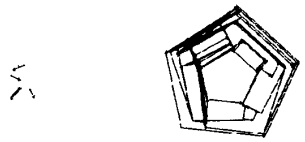
Figure 8: Results for the scene of Pentagon building.

Appendix

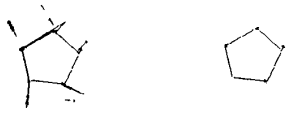
A A Relaxation Network for Constrained Optimization

A typical problem is, given a set of nodes $\{N_i\}$ whose values $\{V_i\}$ have some known interactions among one another, what are the values of the nodes that achieve the best compromise according to the constraints among them? The problem is especially difficult if each node can only take the value either 0 or 1, which is equivalent to the problem of making the best selection among the nodes.

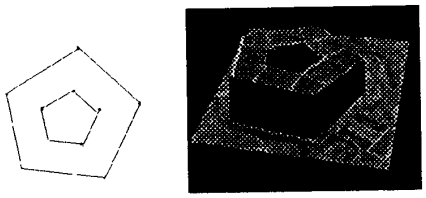
If all the interactions involve at most two nodes at a



(a) noncoplanar segments (left) (b) coplanar links (left)



(c) insidemost coplanar links (left) (d) insidemost surface boundary (left)

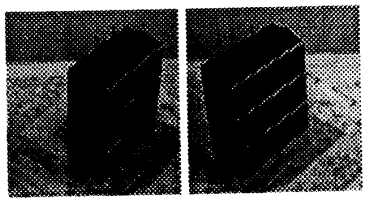


(e) selected surfaces and the holes (left) (f) 3-D Rendered View

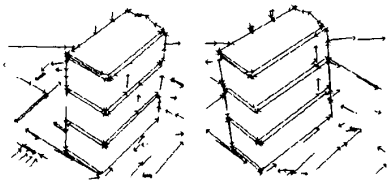
Figure 9: Results of recovering the hole on the roof of Pentagon building.

time, we can model the constraints as either *unary* or *binary*. Unary constraints come from individual merits of a node, while binary constraints relate a pair of nodes. We can further subclassify unary and binary constraints according to whether the constraint is absolute or compromisable: (1) **Unary Constraint**: (a) *Unary Absolute Constraint*: It is an unary constraint that has to be satisfied. (b) *Unary Excitatory or Inhibitory Constraint*: It is an unary constraint that represents how good or how bad a node is in a certain aspect. It can be positive (excitatory) or negative (inhibitory). (2) **Binary Constraint**: (a) *Binary Mutually Exclusive Constraint*: Two nodes are mutually exclusive if at most one of them can be selected at the same time. (b) *Binary Excitatory or Inhibitory Constraint*: It is a binary constraint that represents whether two nodes support (excitatory) or desupport (inhibitory) each other. It can be positive if it is excitatory, negative if it is inhibitory.

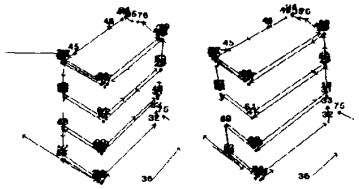
Such a problem can be formulated as an optimization problem: find $\{V_i\}$ such that the cost function $E(V) = -\sum_i (I_i V_i) - \frac{1}{2}\sum_{ij}(T_{ij} V_i V_j)$ is minimized with respect to $\{V_i\}$, where I_i is the total individual merit of the node N_i, and T_{ij} is the total binary constraint between the values of the nodes N_i and N_j. Typically, gradient



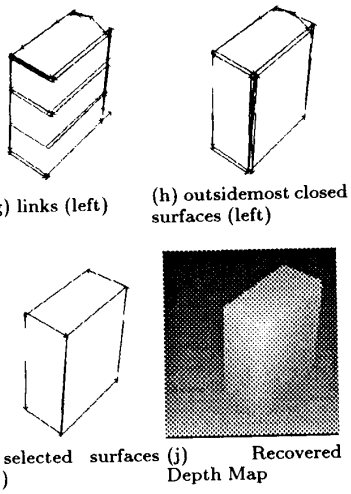
(a) left image (b) right image



(c) left segments (d) right segments



(e) matched junctions (left) (f) matched junctions (right)



(g) links (left) (h) outsidemost closed surfaces (left)

(i) selected surfaces (left) (j) Recovered Depth Map

Figure 10: Results for the oblique views of a hotel.

descent methods are used to get to the solution.

However, such formulation will return values $\{V_i\}$ of any magnitude, depending upon the unary and binary constraints among the nodes. If the problem is to achieve some *binary decisions* about the nodes, the idea is inapplicable. It is the contribution of Hopfield and Tank [12] to introduce a variable u_i for each node and a

sigmoid function $g(u)$ such that for all i, $V_i = g(u_i)$ to constrain V_i between the two specified values 0 and 1. Relaxation based on gradient descent methods is then done to get the best set of binary values $\{V_i\}$.

However, such a system may not be suitable for constrained optimization problems where absolute constraints among the nodes, such as some nodes being mutually exclusive with one another, are present. This can be exemplified by experiments done by others [19] on the travelling salesman problem. Since in our problem there are many instances in which mutually exclusive constraints have to be enforced among some entities, we make a simple modification so that the binary mutually exclusive constraints are guaranteed to be satisfied in the output of the relaxation. The transfer function from the input to the output of a node is modified to be a *winner-takes-all* function, *i.e.*, in every iteration only the node receiving maximum input among all its mutually exclusive competitors will have its output set to 1, while the rest of the competitors will have their outputs reset to 0. This relaxation network is used throughout our stereo system.

B Translation of Coplanarity from 3-D space to Disparity Space

The problem is to find out whether coplanarity in 3-D is preserved in the disparity space under perspective projection into stereo images. We assume the pinhole camera projection is the projection model, and the image planes of the stereo images are coplanar, *i.e.*, a parallel-axis epipolar geometry.

Taking the focal point of the left camera as the origin of the world coordinate system, for any point (x, y, z) in 3-D space, we have $u = (fx)/z, v = (fy)/z, D = (fB)/z$ where (u, v) is its image coordinates on the left view, D is the disparity, f is the common focal length of both cameras, and B is the baseline width.

Suppose a set of points $\{(x_i, y_i, z_i)\}$ in space are all coplanar, and suppose the plane that contains all the points is $ax + by + cz + d = 0$ for some a, b, c, d. Then for all i, $ax_i + by_i + cz_i + d = 0$.

It can be shown that for all i, $au_i + bv_i + \frac{d}{B}D_i + (cf) = 0$, *i.e.*, all the points $\{(u_i, v_i, D_i)\}$ in the disparity space are also coplanar and they lie on the plane $ax + by + (d/B)z + (cf) = 0$ in the disparity space.

References

[1] H. H. Baker, T. O. Binford, J. Malik, and J. Meller. Progress in stereo mapping. In *Proceedings of the DARPA Image Understanding Workshop*, pages 327–335, Arlington, Virginia, June 23 1983.

[2] C.-K. R. Chung and R. Nevatia. Use of monocular groupings and occlusion analysis in a hierarchical stereo system. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 50–56, Maui, Hawaii, June 1991.

[3] C.-K. R. Chung and R. Nevatia. Recovering LSHGCs and SHGCs from stereo. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 42–48, Champaign, Illinois, June 1992.

[4] S. D. Cochran and G. Medioni. Accurate surface description from binocular stereo. In *Proceedings of the Workshop in Interpretation of 3D Scenes*, pages 16–23, Austin, Texas, November 27–29 1989.

[5] U. R. Dhond and J. K. Aggarwal. Structure from stereo–A review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1489–1510, November/December 1989.

[6] M. Drumheller and T. Poggio. On parallel stereo. In *Proceedings of the IEEE Conference on Robotics and Automation*, pages 1439–1448, San Francisco, California, April 1986.

[7] P. Fua and A.J. Hanson. Objective functions for feature discrimination: Applications to semiautomated and automated feature extraction. In *Proceedings of the DARPA Image Understanding Workshop*, pages 676–694, May 1989.

[8] S. Ganapathy. *Reconstruction of scenes containing polyhedra from stereo pair of views.* PhD thesis, Stanford University, Stanford, California, 1976.

[9] W. E. L. Grimson. *From Images to Surfaces: A Computational Study of the Human Early Visual System.* MIT Press, Cambridge, Massachusetts, 1981.

[10] M. Herman and T. Kanade. The 3D MOSAIC scene understanding system: Incremental reconstruction of 3D scenes from complex images. Technical Report CMU-CS-84-102, Carnegie-Mellon University, Pittsburgh, PA, February 1984.

[11] W. Hoff and N. Ahuja. Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(2):121–136, February 1989.

[12] J.J. Hopfield and D.W. Tank. Neural networks and physical systems with emergent collective computational abilities. *Proceedings, National Academy of Science, USA*, 79:2554–2558, April 1982.

[13] Y. C. Hsieh, D. M. McKeown, and F. P. Perlant. Performance evaluation of scene representation and stereo matching for cartographic feature extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):214–238, February 1992.

[14] S. Liebes. Geometric constraints for interpreting images of common structural elements: Orthogonal trihedral vertices. In *Proceedings of the DARPA Image Understanding Workshop*, April 1981.

[15] G. Medioni and R. Nevatia. Segment-based stereo matching. *Computer Graphics and Image Processing*, 31(1):2–18, July 1985.

[16] 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, November 1989.

[17] R. Nevatia and K. R. Babu. Linear feature extraction and description. *Computer Graphics and Image Processing*, 13(3):257–269, July 1980.

[18] D. Terzopoulos. Regularization of inverse visual problems involving discontinuitites. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:413–424, 1986.

[19] V. Wilson, , and G. S. Pawley. On the stability of the TSP problem algorithm of Hopfield and Tank. *Biological Cybernetics*, 58:63–70, 1988.