Stereo Vision without the Scene-Smoothness Assumption: the Homography-Based Approach

by

Andrew L. Arengo

Department of Mechanical and Automation Engineering

The Chinese University of Hong Kong

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Philosophy

June 1998

Acknowledgments

I would like to thank my supervisor, Prof. Ronald Chung, for his valuable advices and suggestions. He has been a great supervisor, giving generous support and encouragement. For my friends, it's been a remarkable experience working with you all.

The courtesy of the Department of Computer Science of the University of Massachusetts at Amberest and the Institute of Robotics and Intelligent Systems (IRIS) of University of Southern California at Los Angeles in supplying the image data sets are much appreciated.

Contents

List Of Figures

簡介

立體視覺的目的是用兩組圖象找出圖象裡面的東西的三維結構，一般所採用的方法都是先找出兩組圖象中相符合的特徵跟估計它們的三維距離，然後用內插的方法得到一個原整的三維距離圖。以前常用的方法都是假設物件的表面圓滑性來將這個困難的問題簡化，進而解決問題。很可惜這假設並不是所有的圖象都適合，特別是一些有物件互相遮擋的場合。這論文描寫了一個新的方法包括特徵配對、分開物件跟外推這幾個步驟就可以推算出圖象裡面物件的三維距離。配對特徵的問題可以用一個矩陣來解決。只要有足夠數量的已知特徵配對就可以找出相關的矩陣。這些矩陣可以分化成代表不同物件表面的矩陣。對於每一個物件表面的已知矩陣，物件表面上的其他特徵可以用矩陣來做配對而且不像以前的方法一定要有分開物件的步驟。這方法並沒有假設表面圓滑，圖象裡面有物件互相遮擋的現象不會構成問題。

Abstract

The goal of stereo vision is to determine the three-dimensional depth of objects from a stereo pair of images. The usual approach is to first identify corresponding features between the two images and estimate their depths, then interpolate to obtain a complete depth map. Finding the corresponding features is regarded as the most difficult problem. In the classical approach, surface-smoothness assumption is used to turn an ill-posed problem to a well-posed one. However, the smoothness assumption is not valid in the presence of occlusions and orientation-discontinuities in the scene. In this thesis, a new approach integrating the feature matching, surface segmentation, and surface extrapolation processes in stereo vision for polyhedral objects or environment is described. Stereo correspondence problem can be tackled by constructing image-to-image mappings named homographies and does not require the surface-smoothness assumption. These mappings refer to mappings between the images which are induced by planar surfaces in 3D. The mappings can be captured by matrices referred to as homography matrices. From a small number of initial stereo correspondences, the matrices can be extracted. They are then clustered into different subspaces representing different planes in 3D world and hence achieving segmentation. For each 3D plane associated, the corresponding homography matrix can infer other correspondences and hence saving the processes of surface fitting and segmentation. Since scene-smoothness is not assumed, occlusions and orientation discontinuities in the scene are allowed.

Chapter 1

Introduction

Human vision is amazing as the three-dimensional (3D) shapes of objects presented can be recovered solely using the projections of these objects onto the two-dimensional (2D) retina. This information acquisition process is impressive as the same projection onto a 2D plane can come from infinitely many 3D objects. Computer vision (machine vision) in a way is the machine realization of human vision; it is the inferring of the structure and properties of the 3D world either from a single or multiple 2D images of the world.

Given a single image of a scene, there is no unique solution of the object leading to that image. Ways out of this can be the gathering of more data or images. Stereo vision is desirable as there is an extra view providing extra information and hence making this information inferring process easier. This thesis addresses the problem of inferring 3D information from stereo images. Typical applications of this depth from stereo process can be the remote sensing of autonomous robots in hostile environments to generate terrain maps and robot navigation.

1.1 Motivation and Objective

Stereo vision is a well-studied topic in computer vision. In conventional approaches, feature points are detected independently. The selected feature points are then used in finding their corresponding feature points in the other view, that is solving the correspondence problem. Among the most important clues in solving this correspondence problem are the epipolar and uniqueness constraints. Since these two constraints are not adequate to resolve the ambiguities, the correspondence problem is ill-posed. To turn this ill-posed problem into a well-posed one, most of the classical approaches rely on surface-continuity and feature-ordering heuristics. Surface-continuity heuristic assumes that the entire scene is continuous in 3D. Feature-ordering heuristic, or ordering heuristic in short, implies that corresponding feature points in stereo images appear in the same geometric order along corresponding epipolar lines.

However, these two heuristics are not applicable to all scenes. They are only valid within a surface patch. Assuming the whole scene is smooth will smooth out the depth-discontinuities and merge neighboring surfaces into one. Since the two cameras are in random positions and orientations, ordering heuristic might be invalid because of occluding views. Moreover, for feature points coming from two different surface patches, ordering heuristic is not applicable. Strong enforcement of these heuristics cannot infer true 3D information of a scene. As a consequence, in the presence of occlusions and depth-discontinuities, the depth estimates are erroneous.

Apart from giving erroneous depth estimates, surface segmentation is impossibles at the stage of matching correspondences as the entire 3D scene is assumed to be smooth. This segmentation process can only be done as an after-process in subsequent surface fitting process. Because of the poor depth

estimates near the occlusion boundaries and depth-discontinuities, results of this segmentation process is often undesirable.

Ignoring occlusions is undesirable as the occlusion boundaries are perhaps the places which convey the most crucial information about the scene. Moreover, repetitive patterns in the scene have made the heuristics unable to resolve the ambiguities. With such repetitive patterns all over the images, solution to the correspondence problem can be easily trapped to one which is only locally optimal to above heuristics and constraints. With all these difficulties, it would be desirable if stereo vision can handle occlusions and depth- discontinuities and separate the surfaces in the featured scene during the matching process. The question is "How can this be done?".

1.2 Approach of This Thesis and Contributions

Suppose we are living in a 3D world where all objects are polyhedral. With two cameras at random positions and orientations observing this polyhedral world, 3D objects can be seen on the images. As the world is made of objects with planar surfaces, images of these surfaces are seen on image planes. Because of this planar surface characteristic, there exists a unique mapping for each planar surface from one view to the other. Each unique mapping is characterized by a few parameters.

For a 3D planar surface, if there are enough stereo correspondences available, these characteristic parameters can be estimated. The estimated parameters capture the mapping of image points induced by a 3D plane from one view to the other one. Stereo correspondences from the corresponding plane can then predicted by these parameters. This stereo correspondence prediction

process is different from conventional approaches as no surface-smoothness assumption about the corresponding scene has been made, surface-smoothness and feature-ordering heuristics are not needed.

In the presence of occlusions and depth-discontinuities, conventional approaches give erroneous depth estimates. This is not the case in the adopted approach. No surface-smoothness assumption about the scene has been made, the stereo correspondences are predicted according to the mapping parameters of their corresponding inducing plane. The entire scene is not assumed to be smooth averts merging neighboring surfaces. Moreover, depth information at regions of occlusions is estimated according to the corresponding mapping parameters, this is an edge over the approaches of smoothing depth information of neighboring surface patches.

For each 3D plane having enough initial stereo correspondences available, the mapping parameters of each inducing plane can be estimated. This implies that surfaces in the featured scene are explicitly segmented in terms of mapping parameters at this matching process and no extra segmentation process is required. Unlike conventional approaches, surface segmentation process is no longer being performed as an after-process in subsequent surface fitting process. Surface fitting process is also not required as the mapping parameters are able to predict correspondence information of every point on that inducing plane.

1.3 Organization of This Thesis

Brief background information about the traditional approaches and surveys of related works are given in Chapter2. Theories of the adopted approach and how to formulate the problem are presented in Chapter3. Detailed descriptions

of implementation are also given. Chapter4 gives the experimental results and finally, summaries and conclusions are stated Chapter5.

Chapter 2

Previous Work

Stereo vision has been a well-studied topic. In inferring 3D information of a scene, epipolar and uniqueness constraint are two of the most important clues in solving the problem of feature correspondences. Because of the ambiguities in finding which is the correspondence, there are approaches making assumptions and hence using more heuristics and others make use of the distinct natures of matching features in solving this problem. For those using distinct features, these features can be grouped to a surface level and hence giving 3D information. In the approaches of using more heuristics, surface-continuity and feature-ordering heuristics are the most popular ones.

2.1 Using Grouped Features

This approach detects object surfaces based on geometric constraints from multiple viewpoints. This surface detection approach is particularly effective in the sense that it needs neither heuristic constraints nor a priori geometric information about the scene. In [1], [9], and [12], conceptual grouping of geometric properties have been implemented to make surfaces.

Mohan and Nevatia [9] proposed a novel system that matches not local features but rectangles across the stereo views. Venkateswar and Chellappa [12]

also proposed to group abstract features up to surfaces and objects from each view and to match those features. However, especially for cluttered scenes, monocular grouping of structural features up to surface level remains to be a challenging job in computer vision.

In [1], junctions are matched instead of lower order features. Junctions are less abstract features and easier to detect. Viewing the matched junctions as corners in 3D, it would then use collinearity and coplanarity among corner branches to extract boundaries of the planar structures in the scene. Even though this approach is reasonable but a huge computation time is needed to find the global minimum of the target function. Moreover, the extracted surfaces wouldn't be used in confirming the feature correspondences, boundaries, and inferring more useful information. The flow is restricted to one-way, from features to surfaces but not the other way round.

2.2 Applying Additional Heuristics

Because of using epipolar and uniqueness constraints is not adequate, attaching the surface to features by the regularization of some additional heuristics such as the surface-continuity and ordering heuristics are adopted by most of the researchers.

Maruya and Abe [8] proposed a system to reconstruct polyhedral structures using stereo vision, however, it makes use of the adjacency structure of polyhedron which is in a way similar to the surface-continuity and feature-ordering heuristics. Cochran and Medioni [2] use an area-based cross correlation along with ordering heuristic and a weak surface smoothness assumption to produce an initial disparity map. The disparity map was smoothed and unsupported points were removed by introducing edge information.

For conventional approaches, the usual paradigm for stereo algorithms using additional heuristics include the following steps:

- Features are located in each of the two images independently.
- Features from one image are matched with features from the other image. That is, for every feature in the left image corresponding to a certain point in the scene, at most a feature can be found in the right image such that it corresponds to the projection of the same scene point.
- Feature correspondences across two views must observe some constraints as well as additional heuristics. The constraints are: epipolar constraint and uniqueness constraint. Additional heuristics are: ordering heuristic and surface-continuity (smoothness) heuristic.
- The disparity between features is used, together with the parameters of the imaging geometry (i.e., relative separation and orientation of the cameras), to determine the distance to the corresponding point in the scene.
- The resulting depth points are often sparse whereas depth must be computed at every point in the scene. Thus the depth points are interpolated/extrapolated to obtain a surface, or a complete depth map.

Unfortunately, the additional heuristics that regularize the problem are not always true for every scene, they are only valid to some particular scenes. When there are occlusions and depth-discontinuities, which happen a lot in real images, using these two heuristics can only lead to erroneous results.

2.3 Homography and Related Works

Homography was first introduced to the vision community by Faugeras ([4] and [5]). It is a mapping between two images which is induced by a plane in 3D. Its essence is that the mapping can be captured by a simple matrix, and once the matrix is known it can be used to infer all other correspondences due to the inducing plane without an explicit recovery of that plane.

Since its introduction, homography has been used in several piece of work such as [7], [10], and [11]. However, so far the use has been limited to only three-view problems, namely the reprojection of an object from two known views to a third view, and the recognition of an object in a third view using two fixed views as reference.

Chapter 3

Theory and Problem Formulation

3.1 Overview of the Problems

A system that infers the 3D information from stereo images is desirable. This system should require minimum human intervention as well as capable of handling most of the man-made environments. In man-made environments, objects are mostly polyhedral. The major steps in the whole process are preprocessing, establishing correspondences, and recovering depth. A system which meets the requirements of inferring 3D information from stereo images of man-made environments, i.e., polyhedral world, should take the above steps into considerations carefully.

3.1.1 Preprocessing

In this stage, image locations satisfying certain well-defined feature characteristics are identified in each image. They have to be chosen carefully as the subsequent matching strategy shall make extensive use of these feature characteristics. Before choosing which kind of feature, matching strategies should be considered. There are two types of matching strategies, one is area-based and the other one is feature-based.

Since area-based matching attempts to match feature correspondences to intensity variations, this method is sensitive to photometric variations and hence unsuitable to stereo with large baseline. Even though the features are simple to extract, these simple features won't help to resolve correspondence ambiguities. Featured-based matching uses structural features rather than image intensities for stereo matching, it is hence more stable to photometric variations and handling non-textured scenes. Therefore, feature-based matching is more suitable in the approach proposed and structural features have to be carefully defined and chosen.

For the features themselves, the more distinct in nature the chosen features are, the less ambiguities they will cause in establishing correspondences (This is further explained in Section3.1.2). To be considered as non-primitive and high-level, features should be structural ones as well as containing information to resolve correspondence ambiguities. One of the distinct features is the L-junctions. It is a structural feature containing a corner point and two line segments that intersect making the corner point. L-junctions can be obtained through the process of edge detection, line fitting, and corner detection. Apart from providing a corner point for matching the corresponding epipolar geometry, two line segments can also be used for checking orientations along epipolar lines.

3.1.2 Establishing Correspondences

As mentioned earlier, a single 2D object in an image can be the projection of many 3D objects. With stereo images, the process of inferring 3D information can be potentially made easier because of the additional information provided by an extra view, 3D depth can be estimated by a simple triangulation process (shown in Figure3.1).

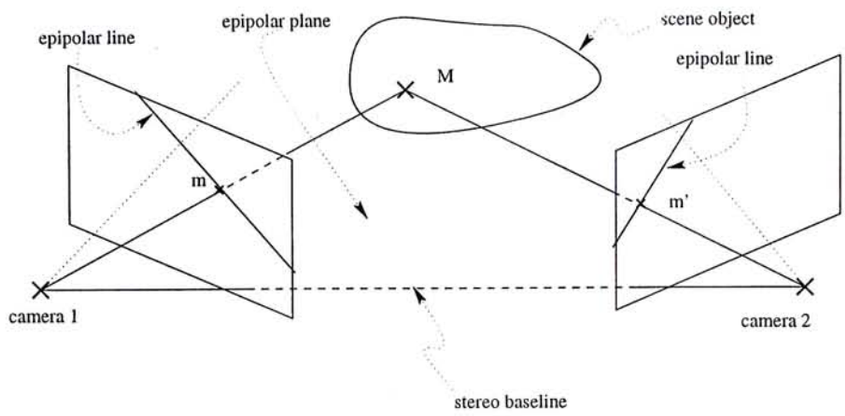


Figure 3.1: Recovering 3D depth using triangulation.

However, the determination of 3D locations of scene points through triangulation requires the establishment of correspondences between individual points in the two images such that each point in a pair of matched points is the image of the same object point. As illustrated in Figure3.2, when the problem of point correspondences between multiple points in two images is ambiguous, triangulation may lead to several different consistent interpretations of the scene. Because of this ambiguity, the practical difficulty with geometric stereo is the establishment of correspondences, that is, the paring up of points in the two images such that each point in a pair of points is the image of the same object point in 3D.

The problem of correspondence establishment is further confounded by the fact that, in general, some points in each image will have no corresponding point in the other image. First, clearly the two cameras will have different fields of view. Second, as illustrated in Figure3.3, objects in the scene may occlude differently in the two images. Occlusion causes certain regions in each image to have no conjugate region in the other image. As a consequence, the spatial position of points that are only visible from a single viewpoint cannot be recovered through triangulation.

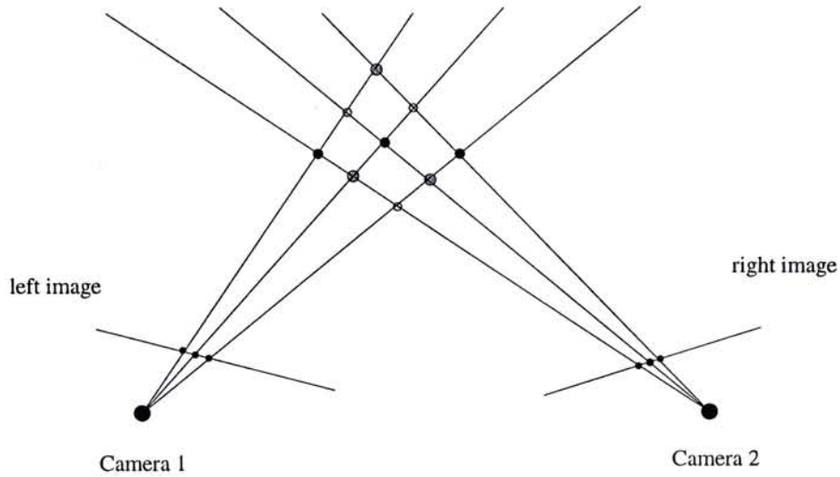


Figure 3.2: Ambiguous correspondence between image points with regard to scene recovery through triangulation.

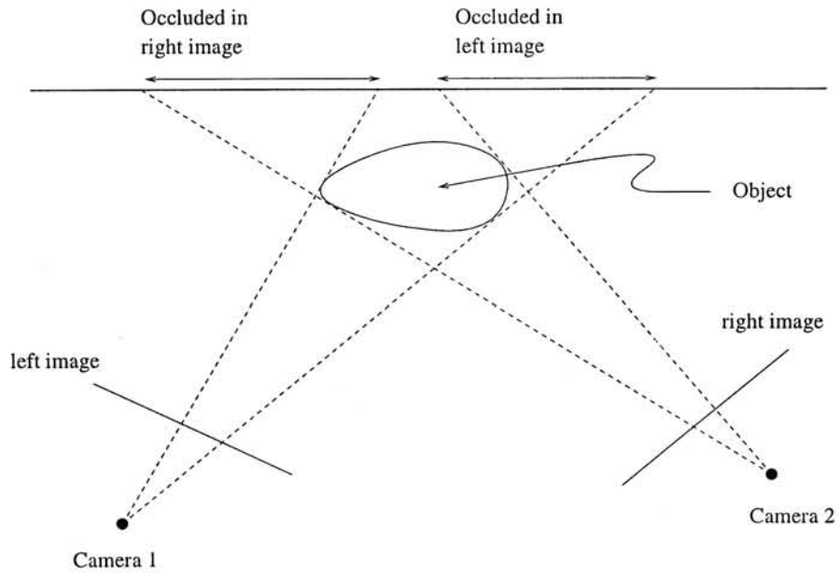


Figure 3.3: Occlusion in stereo.

To solve this correspondence problem, epipolar constraint can be imposed on the local matching search (Details of epipolar constraint is described in Section3.2.1.). However, this local searching process can lead to two or more candidate matches being judged as having almost equal possibility for getting matched. Or worse, an incorrect match point might satisfy the local matching

constraints (epipolar constraint and geometric property constraint) and get chosen as good match. Since 3D depth of a scene point is calculated from the relative disparity between two matched points, incorrect match pairs can lead to erroneous results. Thus, certain assumptions have to be made or imposing more heuristics in matching correspondences. The problem of solving correspondence problem is carefully studied in Section3.2.

3.1.3 Recovering 3D Depth

In parallel axis geometry, 3D position of scene points can be easily obtained from disparity values d computed. The disparity value d for each matched pair of point $\mathbf{m}(x_L, y_L)$ and $\mathbf{m}'(x_R, y_R)$ (see Figure3.1) is $d = x_L - x_R$. By triangulation process as shown in Figure3.1, world coordinate of a scene point $\mathbf{M}(x, y, z)$ can be calculated as

$$x = \frac{bx_L}{d}, \quad y = \frac{by_L}{d}, \quad z = \frac{bf}{d} \tag{3.1}$$

where b is the stereo baseline and f is the focal length of the camera.

For non-parallel stereo systems, 3D reconstruction of scene points requires a more general approach. In this approach, the relative orientation of the cameras is needed. 3D depth can be estimated by the following equation

$$\alpha\mathbf{m} - \beta R\mathbf{m}' = t \tag{3.2}$$

where R is the rotation matrix and t is the translation matrix. (α and β are two projective scalars.)

3.2 Solving the Correspondence Problem

3.2.1 Epipolar Constraint

For features found separately in two images, detecting conjugate pairs in stereo images has been an extremely challenging problem known as correspondence problem. To determine that two feature points, one in each image, form a conjugate pair, it is necessary for the pair to observe some rules. Two of the most important ones are the uniqueness and the epipolar constraint. Uniqueness constraint states that for a feature point in an image, it can at most find one feature point in the other image forming a conjugate pair. Epipolar constraint is described by Figure3.4. For two cameras in arbitrary position and orientation, the image points corresponding to a scene point must lie on the epipolar lines.

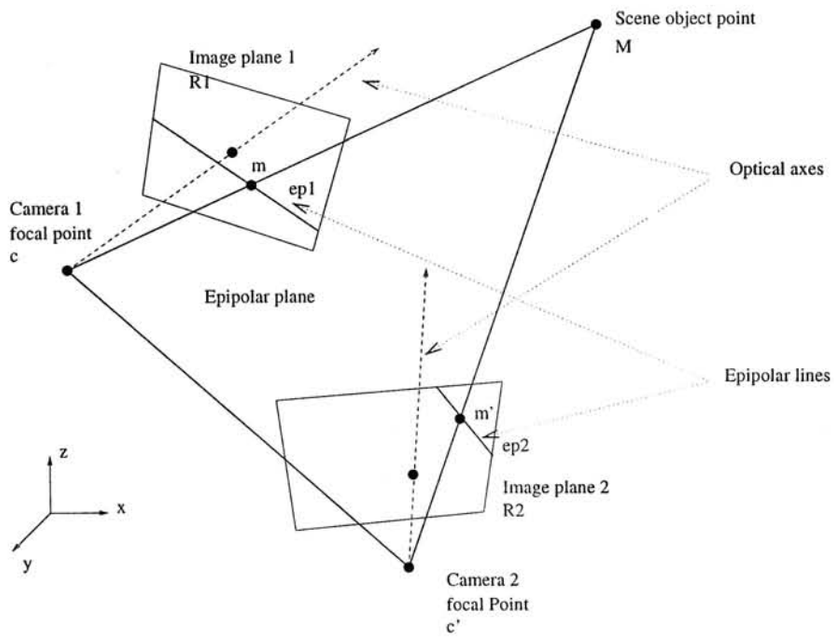


Figure 3.4: Epipolar constraint in stereo vision.

In this figure we see that, given **m** in image plane 1 (**R1**), all possible physical points **M** that may have produced $\mathbf{m}'$ are on the infinite half-line $\langle \mathbf{m}, \mathbf{c} \rangle$. As a direct sequence, all possible matches $\mathbf{m}'$ of **m** in image plane 2 (**R2**) are located on the image, through the second imaging system, of this infinite half-line. This image is an infinite half-line **ep2** going through the point $\mathbf{e_2}$, which is the intersection of the line $\langle \mathbf{c}, \mathbf{c}' \rangle$ with **R2**. $\mathbf{e_2}$ is called the epipole of the second camera with respect to the first, and the line **ep2** is called the epipolar line of point **m** in **R2** of the second camera. The corresponding constraint is that, given a point **m** in **R1**, its possible matches in **R2** all lie on a line. Therefore, search space dimension has been reduced from two dimensions to one. The epipolar constraint is symmetric, for a point $\mathbf{m}'$ in **R2**, its possible matches in **R1** all lie on a line **ep1** through epipole $\mathbf{e_1}$, which is the intersection of the line $\langle \mathbf{c}, \mathbf{c}' \rangle$ with **R1**.

Imposing epipolar constraint on the local matching search could not always guarantee a good match. Therefore, if certain assumptions can be made regarding the nature of the surfaces in the 3D scene, they could be used to determine the consistency of the disparities obtained as a result of the local matching, or guide the epipolar search so as to avoid inconsistent or false matching. An inherent assumption that is usually made about objects is that their surfaces are predominantly smooth.

3.2.2 Surface-Continuity and Feature-Ordering Heuristics

The basic idea of these two heuristics is that the world is mostly made up of objects with smooth surfaces. Smoothness in depth is expected to be the outcome of the smooth variation in disparities obtained from matching process. This is formulated in the form of a regional disparity continuity constraint. Also the contours on the scene surface project on each image as continuous

(or piecewise continuous) curves, which is the motivation behind the figural continuity constraint. Hence, physical features on objects that satisfy some form of the disparity continuity and figural continuity constraints. Assuming the world is mostly made up of objects with smooth surfaces, this means that the conjugate image points along corresponding epipolar lines have the same geometric order in each image, i.e., from left to right or from top to bottom. This is the feature-ordering heuristics and it is shown in Figure3.5.

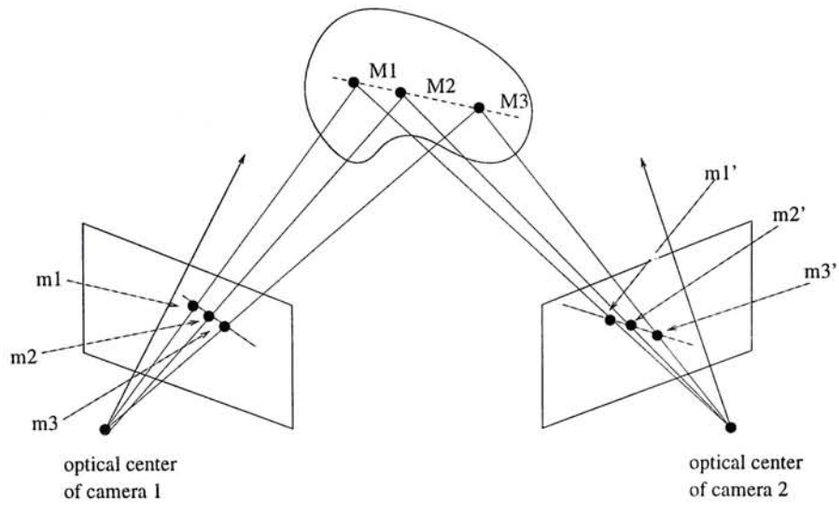


Figure 3.5: Surface-continuity assumption leading to same geometric ordering of feature points on two image planes.

However, this feature-ordering heuristic is violated whenever an object point is imaged from either side of another imaged object point that lies within the same epipolar plane as the first point. As shown in Figure3.6, different geometric orders are observed in different images because of the situation. Moreover, this heuristic is invalid when correspondences are coming from different objects in the scene.

As shown in Figure3.3, surface-continuity assumption is violated at occlusion boundaries. Ignoring occlusions is undesirable as the crucial information about the scene can no longer be recovered once they are ignored. Without

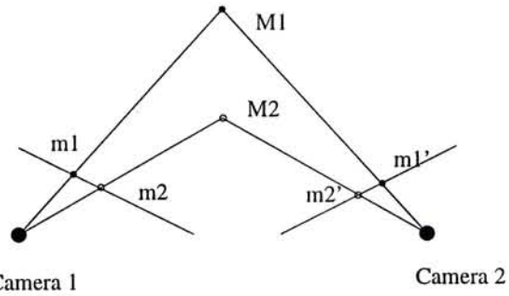


Figure 3.6: Violation of ordering assumption.

using these two additional heuristics, the problem of solving correspondences is an ill-posed problem. To turn this ill-posed problem to a well-posed one, apart from using distinct structural features as mentioned in Section3.1.1, global clues should also be used. The concept of homography introduced by Faugeras in [4] and [5], which states that image to image mappings of points coming from a 3D planar structure can be captured by a 3×3 generic matrix, can be used as a global clue in solving correspondence problem. Position of the conjugate point in the other view can also be predicted.

3.2.3 Using the Concept of Homography

The essence of homography is that image to image mapping can be captured by a single matrix, and once the matrix is known it can be used to infer all other correspondences due to the inducing 3D plane without recovering that plane explicitly. By using epipolar and uniqueness constraints, some initial correspondences can be found from distinct L-junction features across two views. From these initial L-junction matches, homography matrices can be estimated as each L-junction is equivalent to a three pair of point matches. The estimated homographies are already in segmented form, surface segmentation process which is important in conventional approaches is hence not

needed. Moreover, two-way information flow is allowed as estimated homography matrices can infer more point correspondences from polyhedral structures described by these matrices and hence confirm the matrices in turn. Occlusion boundaries don't cause problems to this concept as they do in approaches using surface-continuity and ordering heuristics. This concept is only a description of the orientation and position of the planes in 3D space that induce the matrices but not the information on these regions, a description on properties of 3D planes that induce them but not confining the plane in 3D space.

Despite the novel properties of homography concept, there are still a number of issues concerned:

- *Concept of Homography*

  How does the concept of homography capable of capturing image to image mappings of points coming from 3D planes induced by them?

- *Problem Formulation*

  How can the concept be implemented? Does the number of planes in the scene has impact on the way of formulating the problem?

- *Clustering*

  The optimization in the number of homography matrices. How can we achieve the goal of having minimal sets of homography matrices and each set containing maximal number of similar homographies? How can a optimal homography matrix be estimated to represent a set of similar matrices? What is meant by similar?

These issues are further elaborated in the coming sections.

3.3 Concept of Homography

Before we get into the concept of homography, let's look at a simple barycentric coordinate system. Based on the concept of barycentric coordinate system, we can then build the concept of homography.

3.3.1 Barycentric Coordinate System

For image coordinates of a projection of any point in 3D, $\mathbf{m}$, is related to the world coordinates of the point $[x, y, z]^T$, by

$$\begin{bmatrix} \omega\mathbf{m} \\ \omega \end{bmatrix} = {}^{\mathbf{r}}\mathbf{P}_{\mathbf{C}} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.3}$$

By referring to a barycentric coordinate system B defined with respect to four points P_0, P_1, P_2, and P_3 in 3D and $\mathbf{m_0}$, $\mathbf{m_1}$, $\mathbf{m_2}$, and $\mathbf{m_3}$ are the image coordinates of their projections respectively, the RHS of Equation3.3 can be expressed as

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha\mathbf{P_1} + \beta\mathbf{P_2} + \gamma\mathbf{P_3} + (1-\alpha-\beta-\gamma)\mathbf{P_0} \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \alpha(\mathbf{P_1}-\mathbf{P_0}) + \beta(\mathbf{P_2}-\mathbf{P_0}) + \gamma(\mathbf{P_3}-\mathbf{P_0}) + \mathbf{P_0} \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{P_1}-\mathbf{P_0}) & (\mathbf{P_2}-\mathbf{P_0}) & (\mathbf{P_3}-\mathbf{P_0}) & \mathbf{P_0} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 1 \end{bmatrix}$$

On substitution, the following is observed.

$$\begin{bmatrix} \omega\mathbf{m} \\ \omega \end{bmatrix} = {}^{\mathbf{r}}\mathbf{P_C} \begin{bmatrix} (\mathbf{P_1}-\mathbf{P_0}) & (\mathbf{P_2}-\mathbf{P_0}) & (\mathbf{P_3}-\mathbf{P_0}) & \mathbf{P_0} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \omega_1\mathbf{m_1} - \omega_0\mathbf{m_0} & \omega_2\mathbf{m_2} - \omega_0\mathbf{m_0} & \omega_3\mathbf{m_3} - \omega_0\mathbf{m_0} & \omega_0\mathbf{m_0} \\ \omega_1 - \omega_0 & \omega_2 - \omega_0 & \omega_3 - \omega_0 & \omega_0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 1 \end{bmatrix}$$

Therefore, a barycentric coordinate system B defined with respect to $\mathbf{P_i}$ in 3D and $\mathbf{m_i}$, the image coordinates of the projection of $\mathbf{P_i}$, for i=0,1,2, 3,

the image coordinates of the projection of any point, $\mathbf{m}$, and its barycentric coordinates with respect to B, $[\alpha, \beta, \gamma, (1-\alpha-\beta-\gamma)]^T$, are related by

$$\begin{bmatrix} \omega\mathbf{m} \\ \omega \end{bmatrix} = {}^{\mathrm{r}}\mathbf{P}_{\mathbf{B}} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 1 \end{bmatrix} \tag{3.4}$$

with

$${}^{\mathrm{r}}\mathbf{P}_{\mathbf{B}} = \begin{bmatrix} \omega_1\mathbf{m_1} - \omega_0\mathbf{m_0} & \omega_2\mathbf{m_2} - \omega_0\mathbf{m_0} & \omega_3\mathbf{m_3} - \omega_0\mathbf{m_0} & \omega_0\mathbf{m_0} \\ \omega_1 - \omega_0 & \omega_2 - \omega_0 & \omega_3 - \omega_0 & \omega_0 \end{bmatrix}$$

for some ω_0, ω_1, ω_2, and ω_3.

3.3.2 Image to Image Mapping of the Same Plane

Base on the results of Section3.3.1, we can have the following: given any plane Π in 3D on which a barycentric coordinate system B with respect to three points $\mathbf{P_0}$, $\mathbf{P_1}$, and $\mathbf{P_2}$ on Π is defined, the image coordinates of the projections of any point on Π to two views, $\mathbf{m}$ and $\mathbf{m}'$, are related by

$$\begin{bmatrix} \omega'\mathbf{m}' \\ \omega' \end{bmatrix} = ({}^{\mathrm{r}'}\mathbf{P}_{\mathbf{B}}{}^{\mathrm{r}}\mathbf{P}_{\mathbf{B}}^{-1}) \begin{bmatrix} \omega\mathbf{m} \\ \omega \end{bmatrix}$$

with

$${}^{\mathrm{r}}\mathbf{P}_{\mathbf{B}} = \begin{bmatrix} \omega_1\mathbf{m_1} - \omega_0\mathbf{m_0} & \omega_2\mathbf{m_2} - \omega_0\mathbf{m_0} & \omega_0\mathbf{m_0} \\ \omega_1 - \omega_0 & \omega_2 - \omega_0 & \omega_0 \end{bmatrix}$$

and

$${}^{\mathrm{r}'}\mathbf{P}_{\mathbf{B}} = \begin{bmatrix} \omega_1'\mathbf{m_1'} - \omega_0'\mathbf{m_0'} & \omega_2'\mathbf{m_2'} - \omega_0'\mathbf{m_0'} & \omega_0'\mathbf{m_0'} \\ \omega_1' - \omega_0' & \omega_2' - \omega_0' & \omega_0' \end{bmatrix}$$

foe some ω_i''s, where $\mathbf{m_i'}$ are the image coordinates of the projections of $\mathbf{P_i}$ to the two views, for all i=0, 1, 2. ($^{r'}\mathbf{P_B}^{r}\,\mathbf{P_B^{-1}}$) is a generic 3×3 matrix and it is the matrix of our interest, the homography matrix.

3.4 Problem Formulation

3.4.1 Preliminaries

Given image positions $\{p_i = [u_i, v_i]^T : i = 0, 1, ..., (Q-1)\}$ of Q features in one view and image positions $\{p'_j = [u'_j, v'_j]^T : j = 0, 1, ..., (Q'-1)\}$ of Q' features in the other, find as many pairings as possible between $\{p_i\}$ and $\{p'_j\}$ such that each pairing $\{(p_i, p'_j)\}$ are projected by the same feature in 3-D space. The entire space of all possible solutions is thus discrete and of size $C_Q^{Q'} \cdot Q!$ (assuming $Q' \geq Q$). The epipolar geometry of the stereo views are known in advance. For each feature in one view, the corresponding epipolar line in the other view generally contains several features which are all likely to be the correspondence.

As explained in [5], all pairs of images positions (p_i, p'_j) in two views which are projected from the same plane Π in 3-D satisfy

$$\begin{bmatrix} p'_j \\ 1 \end{bmatrix} \cong \mathbf{H_\Pi} \begin{bmatrix} p_i \\ 1 \end{bmatrix} \tag{3.5}$$

where $\cong$ denotes equality up to a scale and $\mathbf{H_\Pi}$ is a 3×3 nonzero matrix. $\mathbf{H_\Pi}$ characterizes the correspondences between the views due to Π, and is often referred to as the homography matrix induces by Π. It should be noted

that the epipoles $(\mathbf{e}, \mathbf{e}')$ also satisfy the above equation. This equation can be written as two linear equations w.r.t. the elements of $\mathbf{H}_{\Pi}$:

$$\begin{cases} u'_j = \frac{H_{11}u_i + H_{12}v_i + H_{13}}{H_{31}u_i + H_{32}v_i + H_{33}} \\ v'_j = \frac{H_{21}u_i + h_{22}v_i + H_{23}}{H_{31}u_i + H_{32}v_i + H_{33}} \end{cases} \tag{3.6}$$

3.4.2 Case of Single Planar Surface

If there is only a single planar surface Π in the scene, all correct stereo correspondences are captured by a constant 3×3 nonzero homography matrix $\mathbf{H}_{\Pi}$ under Equation3.5. Suppose there are altogether P feature points of Π which are visible in both views, combining the linear equations from all correct pairings, the following equation is obtained:

$$\mathbf{M}_{\Pi}\mathbf{h}_{\Pi} = \mathbf{0} \tag{3.7}$$

where $\mathbf{M}_{\Pi}$ is given by

$$
\begin{bmatrix}
\begin{bmatrix} p_0 \\ 1 \end{bmatrix}^T & 0^T & -\begin{bmatrix} p_0 \\ 1 \end{bmatrix}^T u'_k \\
0^T & \begin{bmatrix} p_0 \\ 1 \end{bmatrix}^T & -\begin{bmatrix} p_0 \\ 1 \end{bmatrix}^T v'_k \\
\vdots & \vdots & \vdots \\
\begin{bmatrix} p_i \\ 1 \end{bmatrix}^T & 0^T & -\begin{bmatrix} p_i \\ 1 \end{bmatrix}^T u'_j \\
0^T & \begin{bmatrix} p_i \\ 1 \end{bmatrix}^T & -\begin{bmatrix} p_i \\ 1 \end{bmatrix}^T v'_j \\
\vdots & \vdots & \vdots \\
\begin{bmatrix} p_{(P-1)} \\ 1 \end{bmatrix}^T & 0^T & -\begin{bmatrix} p_{(P-1)} \\ 1 \end{bmatrix}^T u'_l \\
0^T & \begin{bmatrix} p_{(P-1)} \\ 1 \end{bmatrix}^T & -\begin{bmatrix} p_{(P-1)} \\ 1 \end{bmatrix}^T v'_l
\end{bmatrix}
$$

Equation3.7 is a homogeneous system of $2P$ linear equations for the 9 unknowns in $\mathbf{H_\Pi}$, and under correct stereo correspondences non-trivial solution of $\mathbf{H_\Pi}$ should exist. The $2P \times 9$ matrix $\mathbf{M_\Pi}$ represents the matching between $\{p_i\}$ and $\{p'_j\}$ and is important.

The total number of features of Π which are visible in both images, P, has a significance. If $2P < 9$, the system of equations in Equation3.7 has to be under-determined, which means that any set of correspondences between $\{p_i\}$ and $\{p'_j\}$, correct or not, will result in the existence of the non-trivial solution of $\mathbf{H_\Pi}$. Equation3.7 is thus useless. If $2P \geq 9$, the system of equations can be just-determined or over-determined, which means no non-trivial solution generally exists other than the trivial solution. Equation3.7 then becomes useful, as it can tell the correct set of correspondences, which should have a non-trivial solution for $\mathbf{H_\Pi}$, apart from the other sets of correspondences.

Since epipoles already constitute one pair of point correspondence or two row vectors in $\mathbf{M}_{\Pi}$, the requirement of $2P \geq 9$ translates to a minimum of only three point correspondences. This is what is assumed throughout this thesis. That is, the approach in this thesis restricts the discussion only to surfaces with at least three point features or the equivalent visible in both images. Of course, to reduce the effect of noise in the final recovery of the surface, it would be desirable to have $2P \gg 9$.

To have non-trivial solution of $\mathbf{M}_{\Pi}$ under the correct pairing between $\{p_i\}$ and $\{p'_j\}$, the rank of $\mathbf{M}_{\Pi}$ should be such that

$$Rank(\mathbf{M}_{\Pi}) < \mathbf{9} \tag{3.8}$$

This is a distinct property of the correct solution to the correspondence problem. In fact if it is known that there is only one planar surface in the scene, the solution of $\mathbf{H}_{\Pi}$ should be unique up to a scaling factor, and the rank of $\mathbf{M}_{\Pi}$ should be exactly 8. Unless all correspondences are correct, such a rank property is generally not satisfied if $2P \geq 9$, and is unlikely to be satisfied if $2P \gg 9$.

Though inadequate to resolve all correspondence ambiguities, the epipolar constraint plus reasonable bounds of the disparity gradient are often enough to resolve the ambiguities of a few correspondences. One mechanism of solving the stereo correspondence problem is therefore as the following. Correspondences unique under epipolar constraint are first extracted. Should such initial correspondences exceed three point correspondences or the equivalent, they together with the epipoles allow 8 or more row vectors of $\mathbf{M}_{\Pi}$ to be available. Such initial row vectors are a subset of the row vectors of $\mathbf{M}_{\Pi}$, and the matrix they form in the same manner as $\mathbf{M}_{\Pi}$ is hereafter referred to as the initial

correspondence matrix ${}^{\mathbf{i}}\mathbf{M}_{\Pi}$. With ${}^{\mathbf{i}}\mathbf{M}_{\Pi}$, $\mathbf{H}_{\Pi}$ can be determined up to a scaling factor from the null-space of ${}^{\mathbf{i}}\mathbf{M}_{\Pi}$. Through Equation3.5, such an $\mathbf{H}_{\Pi}$ can then be used to infer other correspondences due to the same planar surface.

The initial correspondences need not be over three points. If surface boundary is visible even partially, junctions are often found along it. The correspondence of a single junction with two branches (L-junction) is already equivalent to three point correspondences: one correspondence over the location of the corner in 3D, and two correspondences over two line segments. Since the epipoles already constitute one point correspondence, a single junction can estimate $\mathbf{H}_{\Pi}$. Junctions are used instead of points as they are more distinct and more sparse and hence more likely to have unique correspondence under epipolar constraint.

Of course, with exactly three initial point correspondences (or equivalently one L-junction correspondence) the $\mathbf{H}_{\Pi}$ is at best a hypothesis only, as any three point correspondences, correct or not, would suggest with the epipoles an $\mathbf{H}_{\Pi}$. However, any single correspondence infer by the $\mathbf{H}_{\Pi}$ and in agreement with the image data would suffice to confirm the hypothesis. In view of this, the minimally three point features of a surface that have to be visible in both images, as required earlier, can be divided into two groups. One group is a set of at least three point correspondences (or the equivalent) which are unique under the epipolar constraint, and they serve to estimate the homography. The other group is the rest of the correspondences which are to be extrapolated by the homography, and which may serve to confirm the correctness of the homography if the first group consists of only three point correspondences or the equivalent.

3.4.3 Case of Multiple Planar Surfaces

Having a single planar surface in the scene is not realistic as there are usually more than one in real scenes. When extended to a more realistic case where there are multiple surfaces $\{\Pi\}$ in the scene, the solution mechanism in Section3.4.2 has a complication. It is not unreasonable to assume three initial point correspondences or one initial L-junction correspondence per surface to be available from the epipolar constraint. However, such initial correspondences are all mixed together, as the surfaces are not segmented in the images. In other words, the initial correspondences matrix ${}^{i}\mathbf{M}_{\Pi}$ for individual surface Π in the scene is not explicitly available. Rather, the row vectors of matrices $\{{}^{i}\mathbf{M}_{\Pi}\}$ for different surfaces $\{\Pi\}$ are available as row vectors in a single matrix ${}^{i}\mathbf{M}$, in no particular order. A mechanism is therefore needed to sort out which of the known correspondences or the row vector pairs are form which surfaces, i.e., to segment ${}^{i}\mathbf{M}$ into ${}^{i}\mathbf{M}_{\Pi}$'s due to different surfaces Π's. In the following section, a procedure serving that purpose is described.

3.5 Subspace Clustering

If the initial stereo correspondences are all unrelated, for example in the form of individual point correspondences, the homographies contained in them can be estimated using the following formulation. Each initial point correspondence corresponds to two row vectors in the initial correspondence matrix ${}^{i}\mathbf{M}$. Such 9×1 vectors are in a 9D vector space. Because of the rank property expressed in Equation3.8, vectors from the same surface Π, whatever the total number is, are contained in an 8D vector subspace, whose orthogonal subspace is the homography $\mathbf{H}_{\Pi}$ induced by Π. A subspace clustering procedure is thus needed to cluster the row vectors of ${}^{i}\mathbf{M}$ into different 8D subspaces of the 9D

vector space, with as few residue as possible left behind which correspond to either wrong correspondences or sets of inadequate correspondences from a few surfaces.

However, this work assumes the context of a polyhedral environment, and it can exploit the presence of L-junctions often found in its images. Each L-junction correspondence is the stereo correspondences of two line segments and a point or equivalently three points, and it is just enough to define with the epipole pair a homography. The problem then becomes how to group the homographies which are defined by the initial L-junction correspondences into different sets, each set being a collection of homographies which are alike enough (meaning that they are the same surface of the environment), and different sets having homographies different enough.

A simple solution to the above is the nearest-neighbor clustering algorithm [6], which is outlined below. To group a set of patterns $\{x_i\}$ into different clusters C_k's according to a particular inter-pattern distance measure $d(x_i, x_j)$ and a threshold t, the following can be used:

1. Set $i \leftarrow 1$ and $k \leftarrow 1$. Assign pattern x_i to cluster C_k.

2. Set $i \leftarrow i+1$. Find the nearest neighbor of x_i among the patterns already assigned to clusters. Let d_m denote the distance from x_i to its nearest neighbor which is in cluster C_m.

3. If $d_m < t$, then assign x_i to C_m. Otherwise, set $k \leftarrow k + 1$ and assign x_i to a new cluster C_k.

4. If every pattern has been assigned to a cluster, stop. Else, go to Step 2.

The algorithm requires a measure of inter-homography distance and a threshold to decide whether any two homographies should be treated as from the same surface or not. In the implemented system, the inter-homography

distance is defined not in the homography space, but in the stereo images directly, so as to put emphasis on the aspect of image-to-image mapping a homography represents. More precisely, to find the distance between two homographies, the image features defining one homography are pushed through the other homography and the mapping errors so resulted are noted. The inter-homography distance is defined as the maximum deviation between the actual position and the predicted position of image feature.

Once the initial homographies are clustered into sets of similar ones, a refined homography matrix is further defined for each set as a whole. Each set represents a number of L-junction correspondences, or a matrix ${}^{\mathbf{i}}\mathbf{M}$. A homography is then extracted from such stereo correspondences using the least-squares-error criterion: find $\mathbf{h}_{\Pi}$ of unit norm such that the cost function $E(\mathbf{h}_{\Pi}) = \|{}^{\mathbf{i}}\mathbf{M}_{\Pi}\mathbf{h}_{\Pi}\|^2$ is minimized. The solution to this is well-known. The optimal $\mathbf{h}_{\Pi}$ is the unit eigenvector of the square matrix ${}^{\mathbf{i}}\mathbf{M}_{\Pi}^{\mathbf{T}}{}^{\mathbf{i}}\mathbf{M}_{\Pi}$ associated with its smallest eigenvalue. Such homographies are then used to extrapolate correspondences of line segments in the images, and be confirmed by them if the extrapolations are supported by image measurements.

3.6 Overview of the Approach

For many environments made from planar surfaces, especially the indoor environments, when observed through stereo imaging can therefore be represented as a collection of homographies. Such a representation can be recovered using a mechanism as outlined below.

An overview of the recovery mechanism is shown in Figure 3.7. Line segments are first extracted from the images through edge detection and line fitting processes. L-junctions are then hypothesized from the line segments

through a corner detection process. L-junctions which have unique correspondences under the epipolar constraint are identified. The unique correspondences are then supplied to the subspace clustering process described in Section3.5, which extracts the homographies present in the stereo images. For any planar surface in the scene, as long as one L-junction correspondence over it is initially available, the associated homography can be estimated. This planar surface is not restricted to any category of orientation. Once there is enough information available, the associated parameters can be estimated. With more than one initial L-junction correspondence, the homography is even confirmed.

Through Equation3.5, the homographies can then serve as mappings to extrapolate correspondences of all other features in the two images. For any feature in one image, the identified homographies can be used in turn to predict its correspondence in the other image. For more and more members coming from a same plane Π are available as a consequence of the prediction process, parameters of a better matrix $\mathbf{H}_\Pi$ can be estimated. This newly extimated matrix is so-called the refined homography matrix of that corresponding plane. The more entries available, a better homography matrix can be estimated and this is the process of refining a homography matrix. Once the correspondence extrapolation and confirmation of homographies are completed, the representation of the environment as a number of homographies is available. The representation comes with a segmentation of the environment, i.e., different planar surfaces in the environment correspond to different matrices in the representation, and it is known which features in the images are contained in which matrix. This representation can be used to generate a dense 3D information map of the environment over the space around the detected line segments.

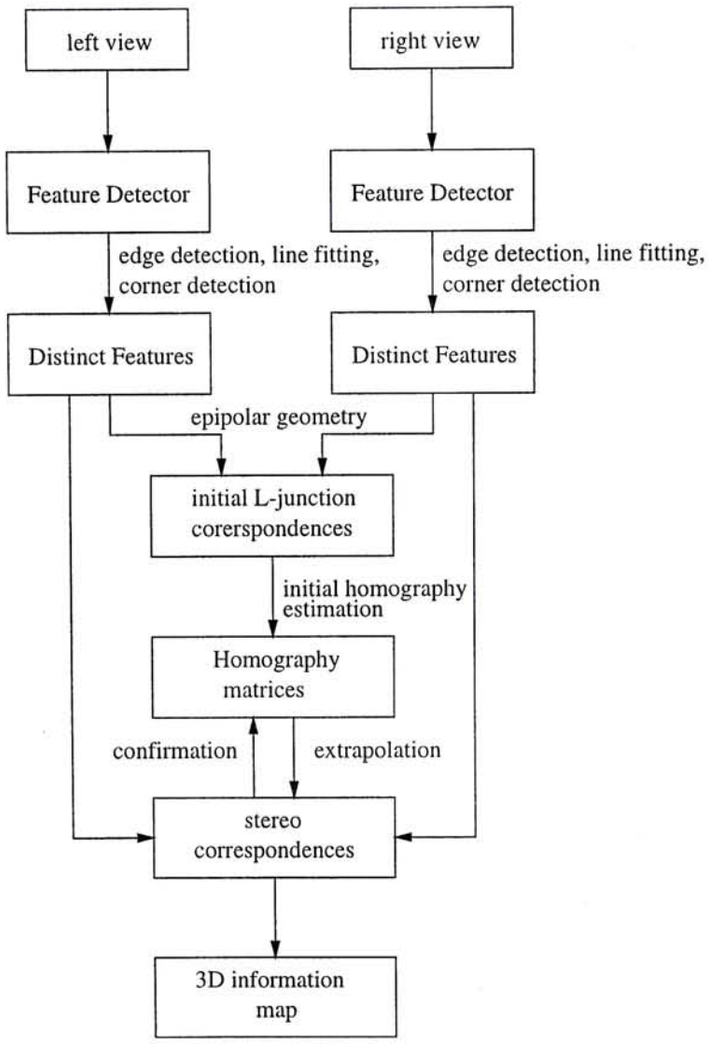


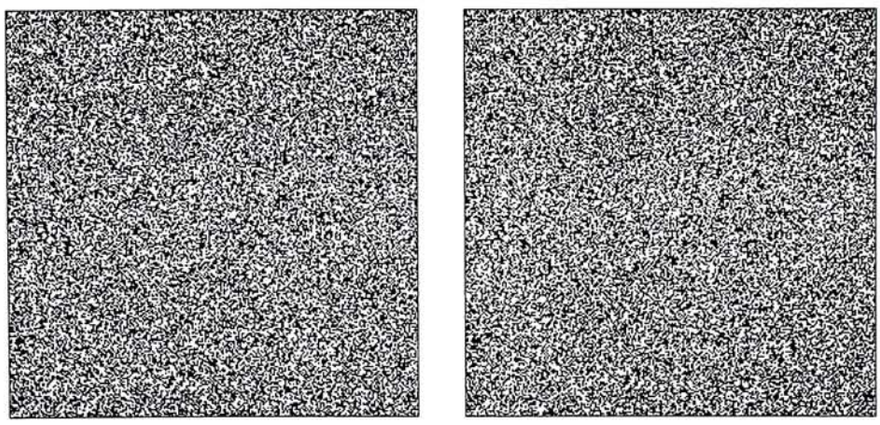
Figure 3.7: Overview of the recovery mechanism

Chapter 4

Experimental Results

4.1 Synthetic Images

A pair of random dot stereograms has been used. These synthetic images of dimension 256×256 were generated by a simple random dot generator. Noise is added to the stereograms by changing the intensity of pixels either from brightest to darkest or the other way round. The synthetic images have the advantage that the real disparity of the images are known and they can be compared with the results obtained by the approach adopted.



(a) image 1 (b) image 2

Figure 4.1: 256×256 synthetic images of a three layer cake.

The noisy stereograms (20% of the intensity values are changed) are shown in Figure4.1 of dimension 256×256, which are the top views a three layer cake. Since there is no feature on the cake, random features have to be added for correspondence matching. Since there are only black and white intensities on these two stereograms, the features added are simple point features. After sprinkling random feature points onto the stereo pair, for every feature point in one view is then checked if it has correspondence unique under the epipolar constraint and a certain allowable disparity range. Confidence is measured in terms of correlation values.

These unique correspondences are then supplied to the clustering process which extracts homographies in the stereograms. Unlike L-junction pairs, each point correspondence is not enough to determine a homography matrix. A different clustering process is hence required other than the one described in Section 3.5. For this process, it is reasonable to assume that there will be a dominant subspace giving a rough null space estimate. Based on this rough null space estimate, point pair entries with large deviations will be rejected hence giving a subspace with fewer noises. The process continues until acceptable amount of deviations are observed for remaining entries, these point pairs are hence declared as coming from a same 3D plane. The goal of this clustering process is to obtained a group of point pairs that gives the least median error in Equation 3.7. After undergoing the clustering process described, three homographies are obtained each corresponding to a layer of the cake.

Since no segmentation of the surfaces is required, we can use the homography matrices obtained to predict the right view position of a particular point in left view. By area-based matching measure, correlation value can be calculated by a particular predefined window. In this matching process, each point from the left view is supplied to three homography matrices for predicting its

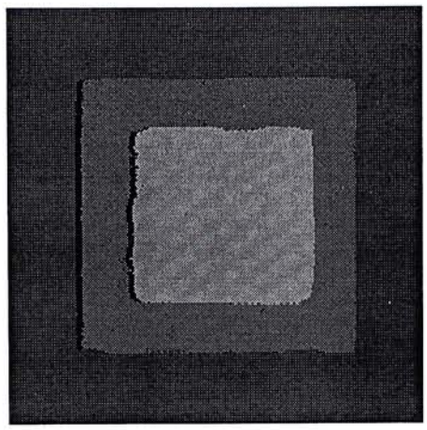
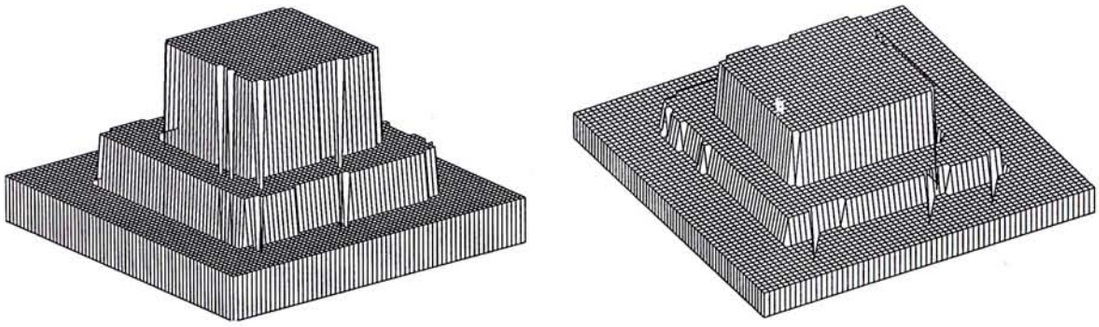


Figure 4.2: Gray level of the disparity map.

corresponding positions in the other view. For each predicted position, correlation value is calculated and the position with the highest measure will be declared as the best match and hence falling on the surface described by that homography matrix.

Depth is shown in Figure4.2. Disparities are in proportion to the gray level, darker scale implies smaller disparity value. The darkest part are the regions that no match can be found in the other view by using the area-based matching approach. Because of occlusion, there are two darker stripes in Figure4.2 corresponding to the occluded regions. These two regions can't be recovered as the similarity measure might be too low to pronounce a correspondence and there can't be a correspondence because of occlusion. From observing the boundaries of the matched pattern, we can see that the boundaries are not clear-cut. There are zig-zag patterns because of the matching strategy adopted. Larger matching windows will give fewer zig-zag boundaries while smaller windows will give smaller regions that are unable to declare matches have been found.

The disparity maps shown in Figure4.3 are of different views. Areas of unknown disparity are assigned with lowest value. These areas are the regions



(a) elevated view (b) oblique view

Figure 4.3: Disparity map of the reconstructed cake.

of no correspondence can be found and no disparity information to which vicinity disparity information can be used to interpolate from. The surface appears to have a hole from where the disparity value is missing and these missing holes are largely coming from the two occlusions. The viewpoints for displaying the reconstructed surface are elevated view and oblique view in Figure4.3. From these results, we can see that even stereograms without explicit object patterns, the adopted approach is capable of reconstructing objects. Even without using distinct features such as L-junctions, using simple point features are enough for this approach to infer 3D information.

4.2 Aerial Images

Results of aerial images are shown from Figure4.4 to Figure4.23. The stereo images were taken using two cameras at different positions. Positions and orientations of the cameras are not given. Although it is not difficult to calculate the parameters required as there are lots of work in that field, the epipolar geometry of these images are calibrated using a least square error approach. The focal length and baseline width are assumed to have a random value as they will only affect the depth information up to a particular scale. The epipolar

geometry calibration procedure was just manually picking some feature points and find the parameters required giving smallest error. The procedure was not intended to give a very precise description of the calibration parameters for all stereo pairs.

The T-shaped building in Figure4.4, square block building in Figure4.9, L-shaped building with two lower wings in Figure4.14, and the pentagon in Figure4.19 are all obtained from Institute for Robotics and Intelligence Systems (IRIS) of University of Southern California (USC) at Los Angeles.

The original stereo images are shown with features marked. These four sets of images are all aerial images of dimension 360×360 except the one for Pentagon is 512×512. For all these real image data sets, line segments are first extracted from the images using Canny edge-detector and the line-fitting subsystem in the Nevatia-Babu "Linear" package. A simple corner detector is then applied to the line segments, examining if any two line segments have their end-points nearby and have their orientations different enough, thereby proposing L-junctions between the line segments if they do. For each stereo pair, the L-junctions are then checked if any one of them have correspondences unique under the epipolar constraint and loose bounds of disparity gradient. The unique correspondences are then supplied to the subspace clustering process which extracts the homographies present in the stereo pair.

Since accurate measurements were not taken of the camera parameters, the results cannot be compared to the actual 3D information obtained. However, the disparity map can obtained by measuring the disparity difference of the corresponding points. By extrapolating the position of the points in one view to predict the position of the corresponding points in the other view using the homographies obtained, dense disparity map of the perceived objects can show the shape of the actual object perceived.

4.2.1 T-shape building

Figure4.4(a) and Figure4.4(b) shows one stereo pair of a T-shape building in a cluttered background, with the extracted L-junctions superimposed to them. It can be observed that not only the images are cluttered, but they also have a lot of repetitive patterns caused by roof boundaries, building's shadows, road edges, and lane markings. The stereo pair present difficulty to most of the stereo systems if the task of extracting 3D information as well segmenting the surfaces is to be achieved. In the new approach, with only a small number of initial L-junction correspondences which are unique under the epipolar constraint, as shown in Figure4.4(c) and Figure4.4(d), three homographies were identified. They are the ground level (shown in Figure4.5(a)), the roof of the building (shown in Figure4.5(b)), and the roof of the small rectangular structure on top of the building (shown in Figure4.5(c)).

The extracted homographies allow line segments in the images to be matched. The matched edges are shown in Figure4.6. Two different views of disparity map of the object perceived are shown in Figure4.8. For simplicity, the boundary (in T form) of the roof of this building is manually picked. As only one pair of L-junction match was found, the boundary of this small structure on top of the roof is considered to a triangular. From the disparity maps, we can see that the T-shaped building is vividly presented.

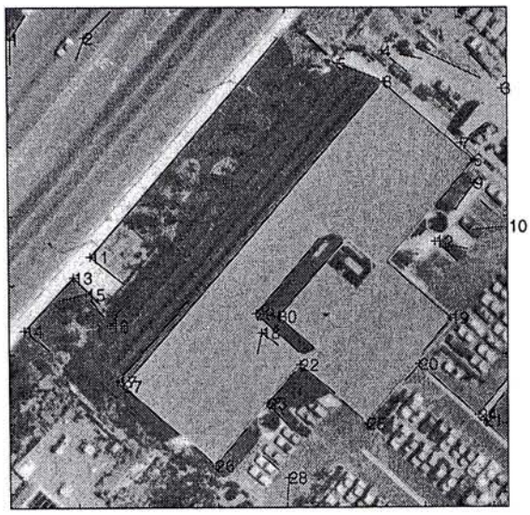
In Figure4.7, the position for a particular point in one view is shown in the other view by using the homographies obtained. By picking a random point on one of the surfaces, shown in Figure4.7(a), the predicted position is shown in Figure4.7(b), which is the other view. Since there are more than one homography matrices available, correlation value is calculated for each predicted position, i.e., three in this case. The one with highest correlation value will be chosen as correct position. By comparing these two images, we

can see the the positions of the matches predicted are very precise. Since the predicted position might have deviation to exact position, a process to refine the predicted positions is carried out. This can be a simple process of searching the neighborhood of the correct predicted position. The one that gives highest correlation value will be declared as the refined position. Even after refining the position of the predicted position by calculating correlation values, shown in Figure4.7(c), the performance is similar or at most as good as the ones without refinement shown in Figure4.7(b).

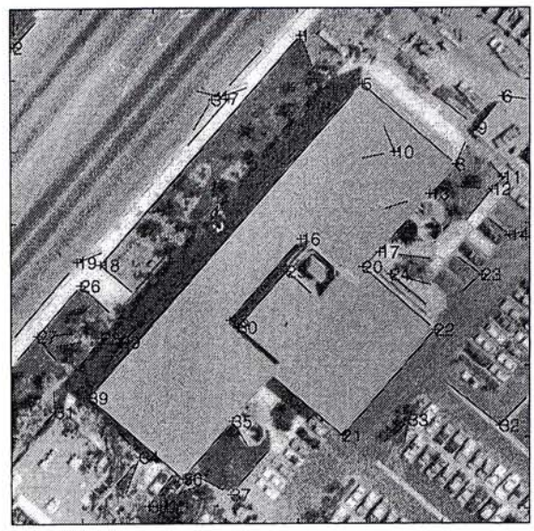
4.2.2 Rectangular Building

The images are about a rectangular building again in a cluttered background. Similar to the images in Figure4.4, the stereo pair also has a lot of repetitive patterns caused by roof boundaries, building's shadows, road edges, and lane markings. The images are shown in Figure4.9(a) and Figure4.9(b), with the extracted L-junctions superimposed to them. The initial L-junction correspondences unique under the epipolar constraint are shown in Figure4.9(c) and Figure4.9(d).

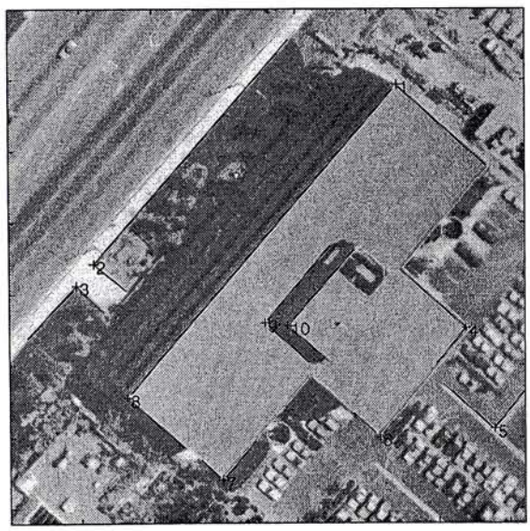
Three homographies were identified. They are the ground level (Figure4.10(a)), the roof top (Figure4.10(b)), and the square structure (Figure4.10(c)), whose initial correspondences are superimposed on each planes. The line segment correspondences extrapolated by the homographies are presented in Figure4.11. The position of a random point in one view is used in predicting its position in the other view by using the homographies obtained. Results of the points for prediction and predicted positions are shown in Figure4.12(a) and Figure4.12(b) respectively. The refined positions using correlation values are also shown in Figure4.12(c) for comparison. Relative disparity map of the reconstructed structures is shown in Figure4.13.



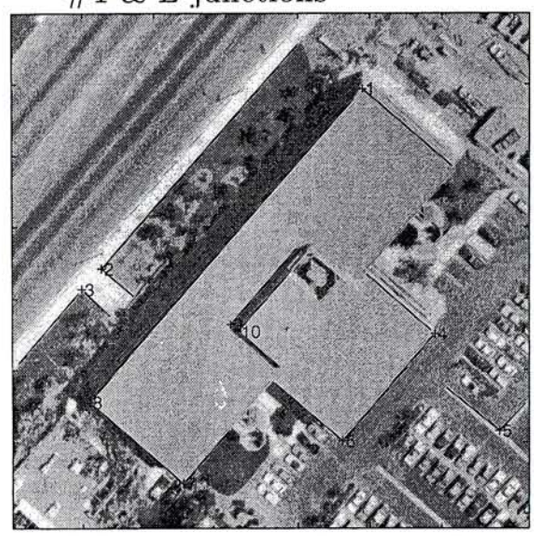
(a) left view of building #1 & L-junctions



(b) right view of building #1 & L-junctions



(c) left view of initial L-junction matches

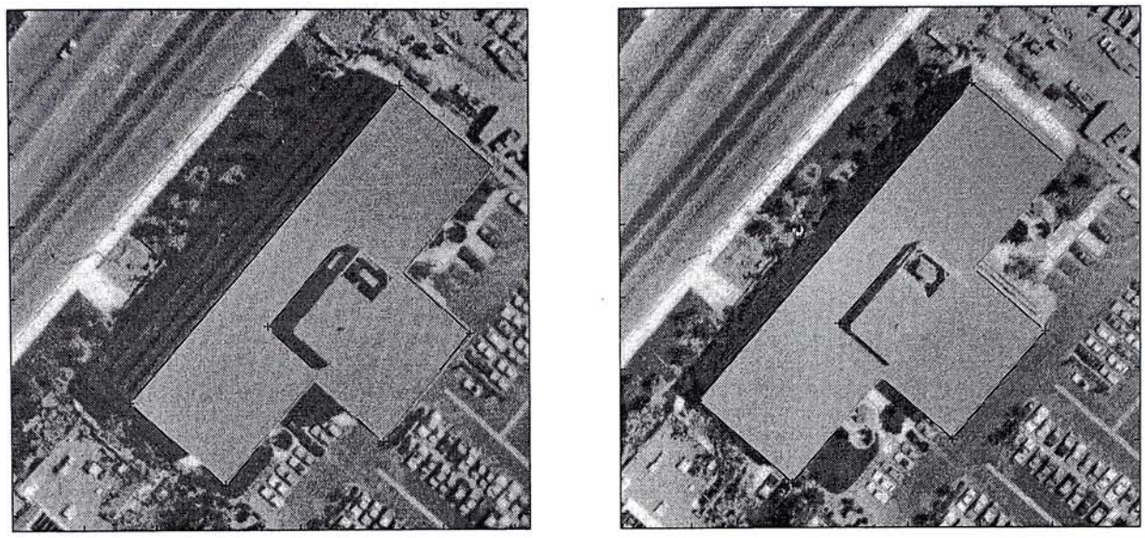


(d) right view of initial L-junction matches

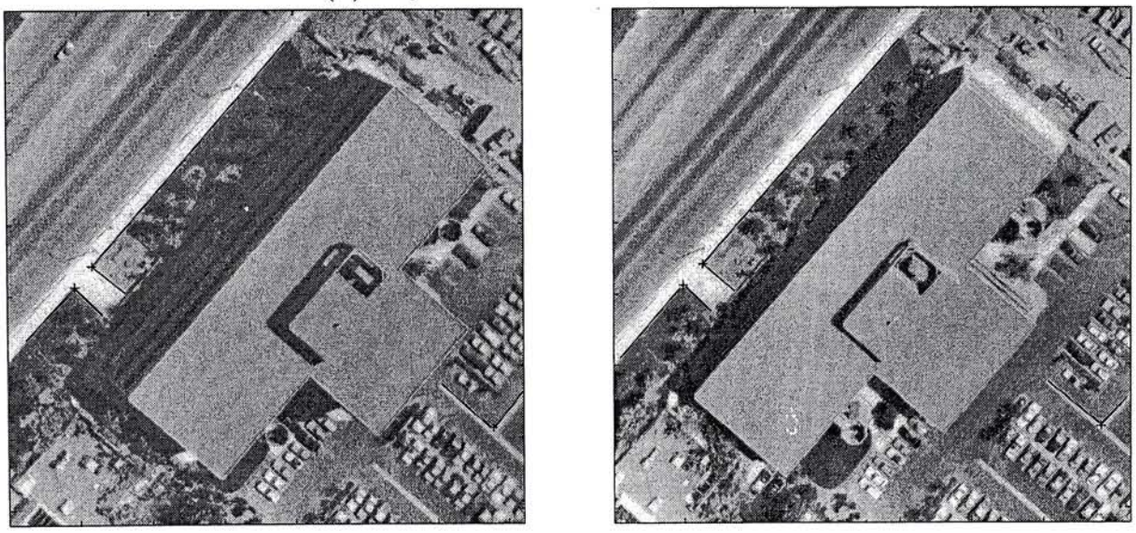
Figure 4.4: Stereo images of building #1 and preliminary processings.

4.2.3 3-layers Building

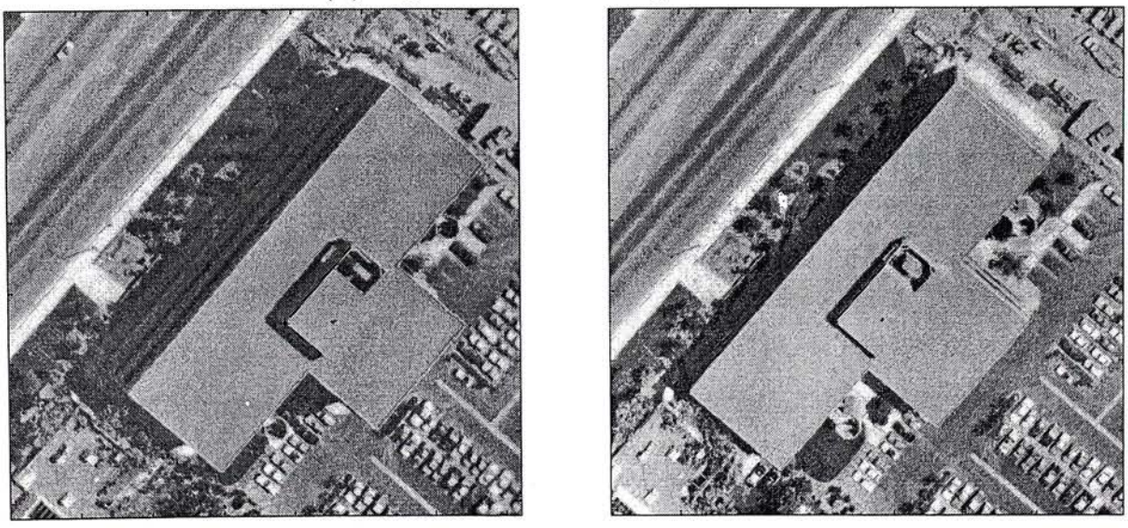
Similar to images in Figure4.4 and Figure4.9, the images are about a building in a cluttered background. This stereo pair has a lot of repetitive patterns caused by roof boundaries, building's shadows, road edges, and lane markings. Even though it might not be as cluttered as the two previous pair,



(a) L-junctions of 1st plane



(b) L-junctions of 2nd plane



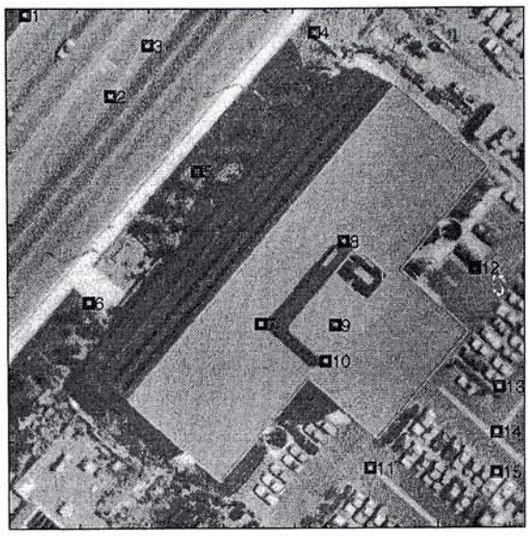
(c) L-junction of 3rd plane

Figure 4.5: Three extracted homographies.

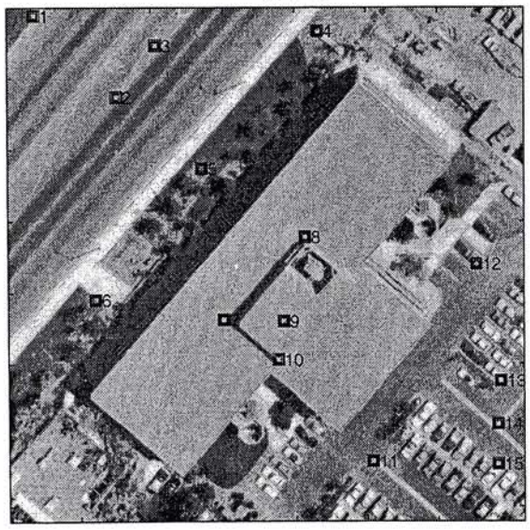
Figure 4.6: Edge matches found using homographies.

the image pair is more complex in terms of number of planes and structures. The image pair is shown in Figure4.14(a) and Figure4.14(b), with the extracted L-junctions superimposed to them. The initial L-junction correspondences unique under the epipolar constraint are shown in Figure4.14(c) and Figure4.14(d).

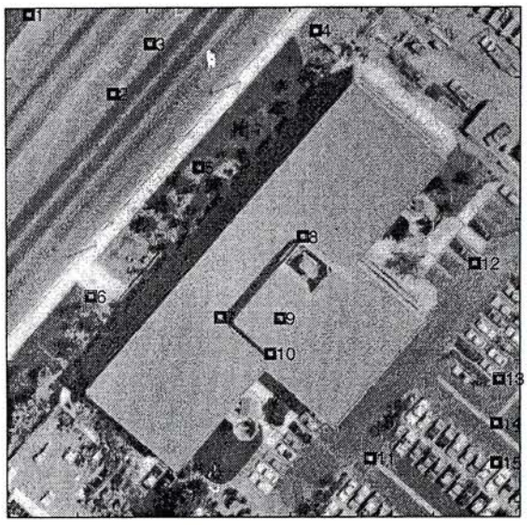
Four homographies were identified. They are the ground level (shown in Figure4.15(a)), the L-shaped roof top (shown in Figure4.15(b)), the roof top of the higher wing (shown in Figure4.15(c)), and the roof top of the lower wing (shown in Figure4.15(d)), whose initial correspondences are shown on each plane. Because of more complex nature of the building perceived, not all the desired structures were recovered. The two minute rectangular shape structures on the top of the L-shaped roof are missing. The negligence is due to L-junction can only be seen in one view but not the other view. Without the candidates of the corresponding L-junction in the other view, no initial match can be found and hence obtaining no homography matrix corresponding to these minute structures.



(a) position for prediction



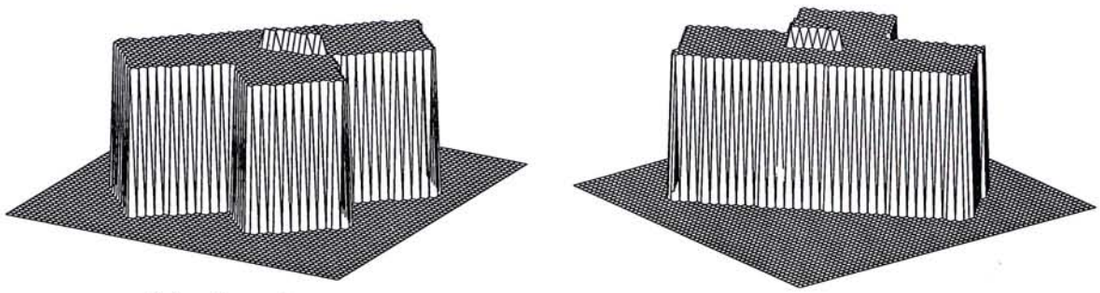
(b) predicted position



(c) refined position

Figure 4.7: Predicting positions by the homographies obtained.

The line segment correspondences extrapolated by the homographies are presented in Figure4.16. Most of the desired line segments matches on the roof tops are contained in the l-junction matches and hence no further line segment matches can be extrapolated. The positions predicted and refined positions are shown in Figure4.17. Again for display purpose, the boundary of the L-shaped roof top close boundary was manually picked. Relative disparity



(a) view 1 (b) view 2

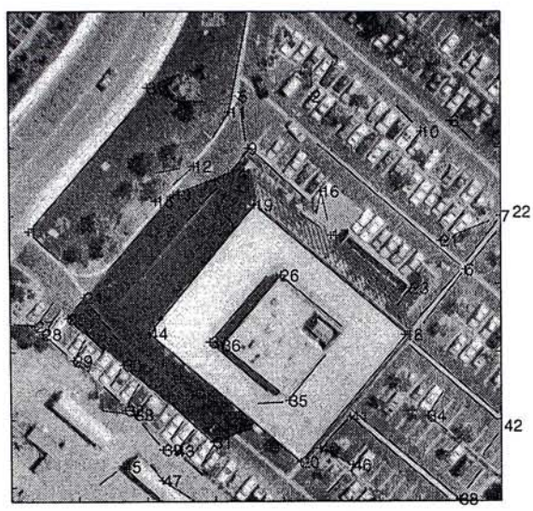
Figure 4.8: Relative disparity map of building #1.

maps of the reconstructed structures is shown in Figure4.18. Two different views are presented in Figure4.18(a) and Figure4.18(b) respectively.

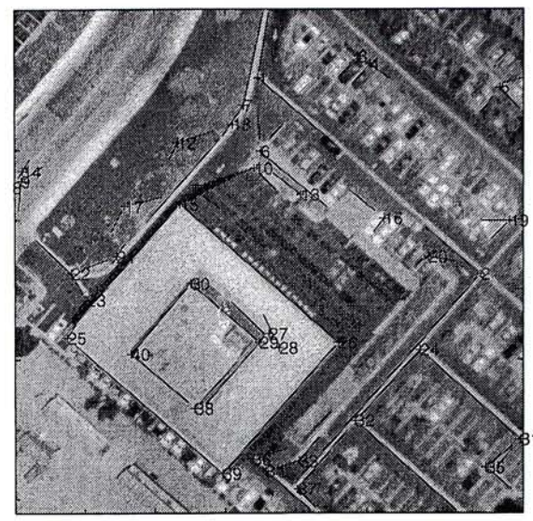
4.2.4 Pentagon

The object perceived in this stereo pair is the Pentagon of United States. Even though the image is easier in terms of number of planes when making comparison to the buildings in Figure4.4, Figure4.9, and Figure4.14, this stereo pair also has its difficulties. The stereo pair is richly textured, and there are a lot of repetitive patterns caused by roof boundary, lane markings, and road edges. The image pair is even more cluttered than the previous three as there are lots of repetitive patterns on the top of the roof and this adds complexity to all stereo systems.

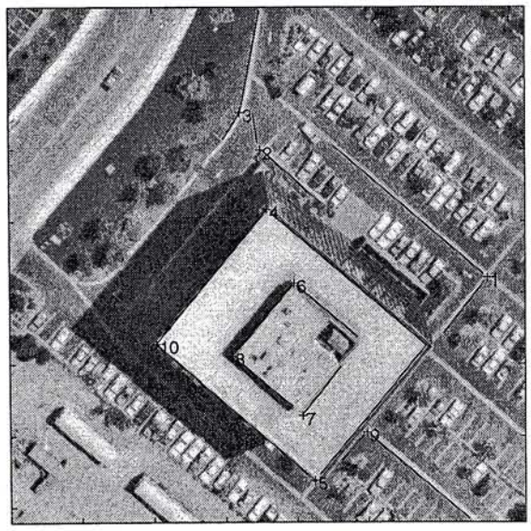
The image pair is shown in Figure4.19(a) and Figure4.19(b), with the extracted L-junctions superimposed to them. The initial L-junction correspondences unique under the epipolar constraint are shown in Figure4.19(c) and Figure4.19(d). Not all of the L-junctions on the roof top are matched, only part of them can find corresponding L-junctions in the other view.



(a) left view of building #2 & L-junctions



(b) right view of building #2 & L-junctions

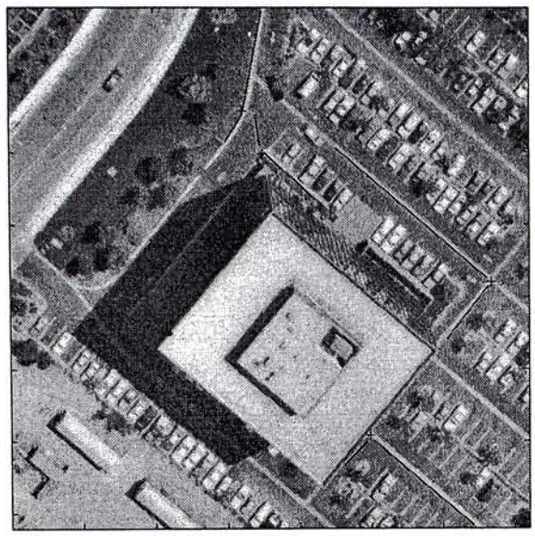


(c) left view of initial L-junction matches

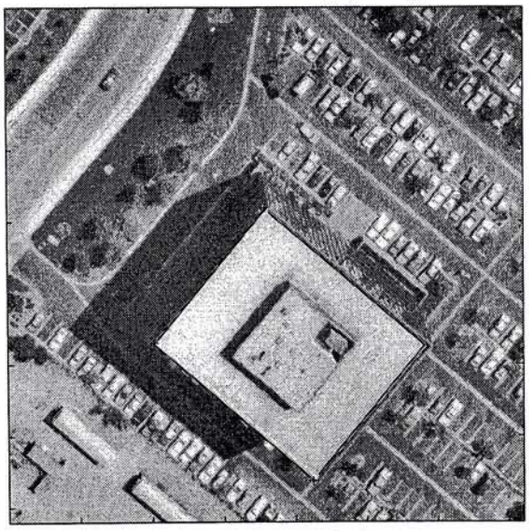
(d) right view of initial L-junction matches

Figure 4.9: Stereo images of building #2 and preliminary processings.

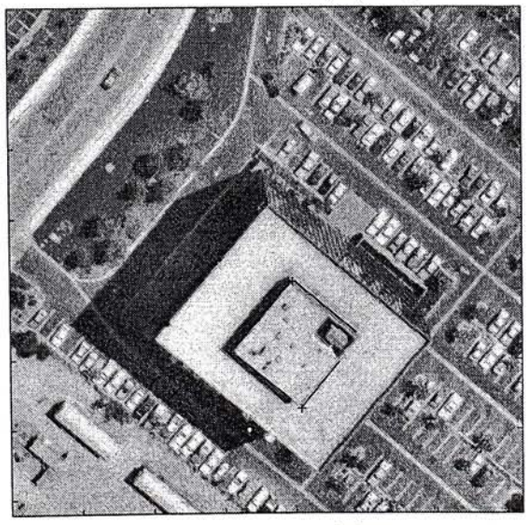
Two homographies were identified as expected, one for the ground level (shown in Figure4.20(a)) and other one is the richly texture roof top of Pentagon (shown in Figure4.20(b)), whose initial correspondences are displayed in each plane. The small difference in terms of the disparity values between two desired planes are difficult for other stereo systems to segment them successfully. In the approach adopted, clustering process is a grouping process




(a) L-junctions of 1st plane




(b) L-junctions of 2nd plane




(c) L-junctions of 3rd plane

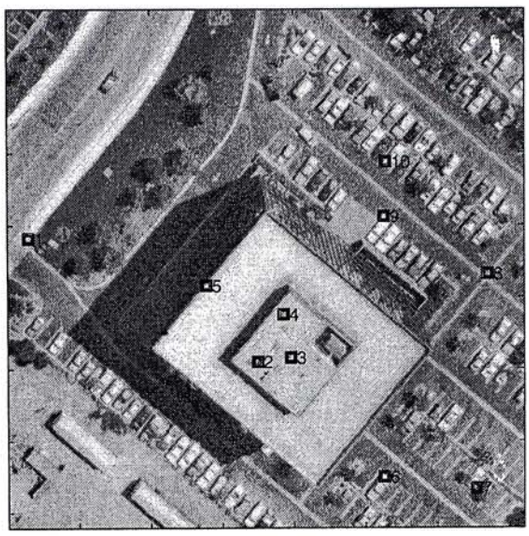
Figure 4.10: Three extracted homographies.

Figure 4.11: Edge matches found using homographies.

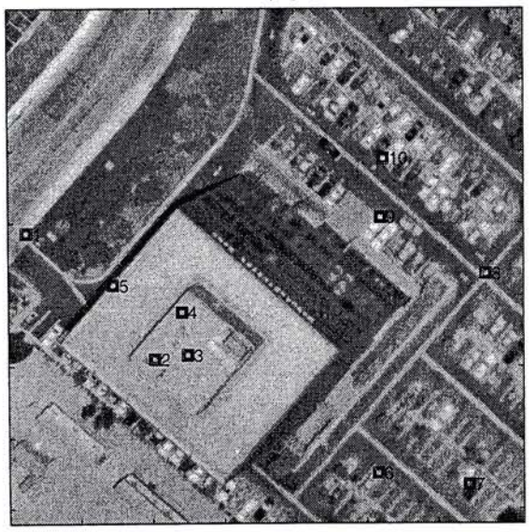
that gives the minimum number of groups and least error measure in terms of the best homography matrix for each group, these planes are segmented successfully.

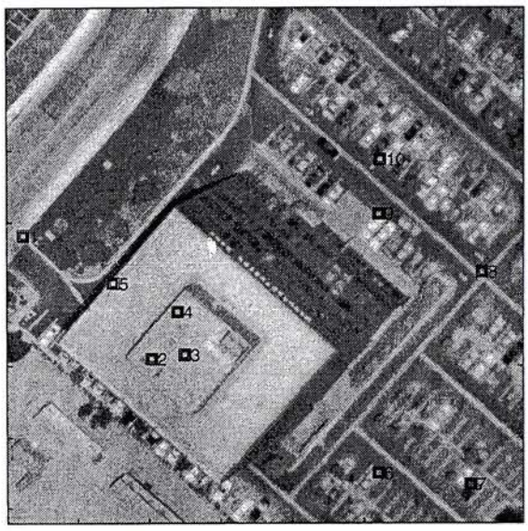
The line segment correspondences extrapolated by the homographies are presented in Figure4.21. The desired edge matches can be seen on the roof as well as the ground level. Most of the repetitive patterns on the roof can find their corresponding line segment matches in the other view as the corresponding homography matrix can predict their position to small deviation. The performance of the homography matrices obtained in position prediction is compared with the refined position in Figure4.22. Relative disparity map of the reconstructed structures is shown in Figure4.23.

Since the approach adopted didn't attempt to find the boundary of the surfaces corresponding to homography matrices obtained, manual picking of some of the surface boundaries for display purpose is not misleading. From the extrapolation results of the line segment matches, we can see that the positions are precisely given for corresponding line segments in the other view.



(a) position for prediction





(b) predicted position (c) refined position

Figure 4.12: Predicting positions by the homographies obtained.

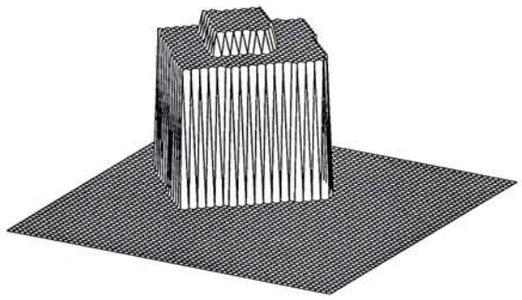
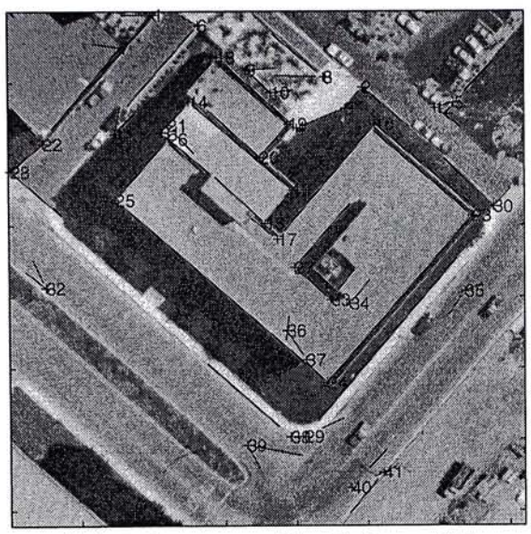
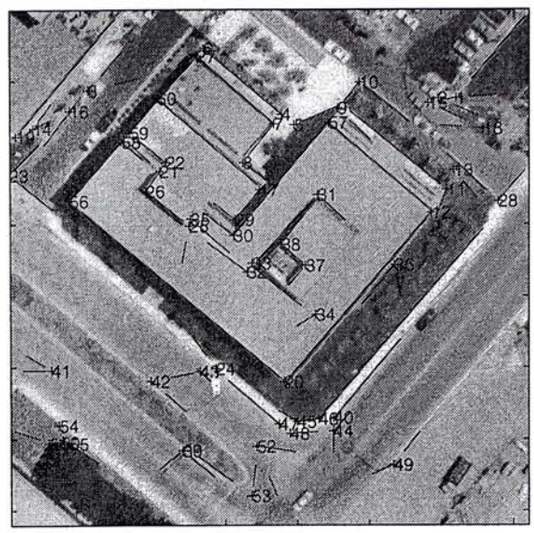


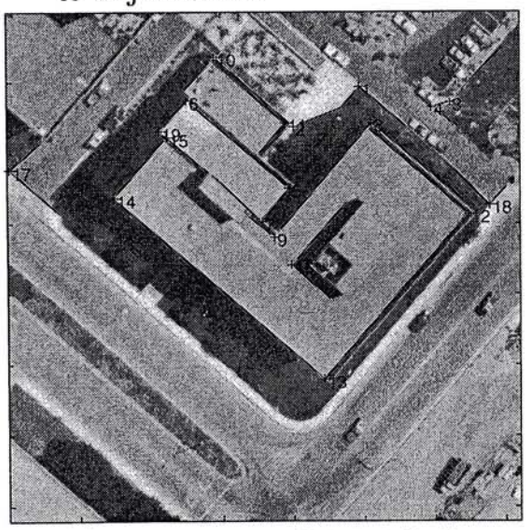
Figure 4.13: Relative disparity map of building #2.



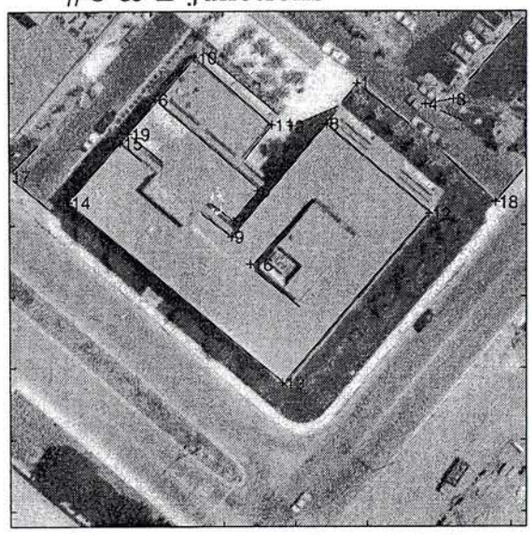
(a) left view of building #3 & L-junctions



(b) right view of building #3 & L-junctions



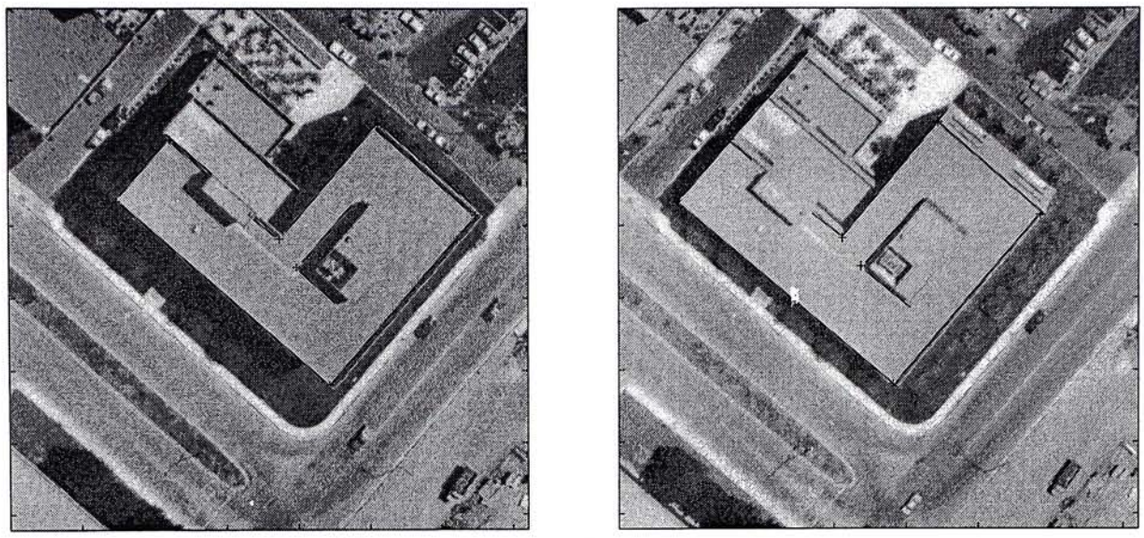
(c) left view of initial L-junction matches



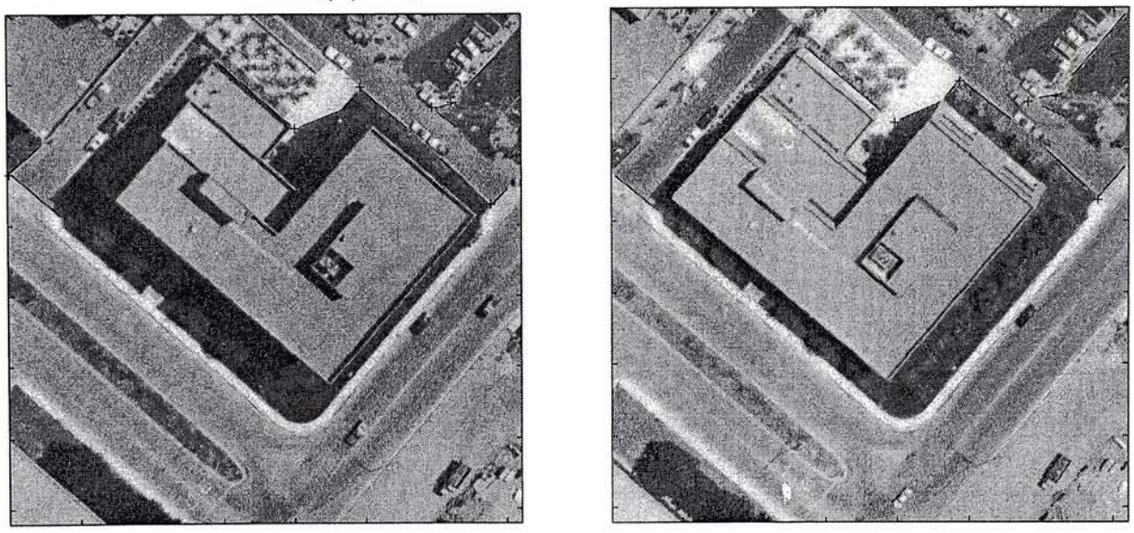
(d) right view of initial L-junction matches

Figure 4.14: Stereo images of building #3 and preliminary processings.

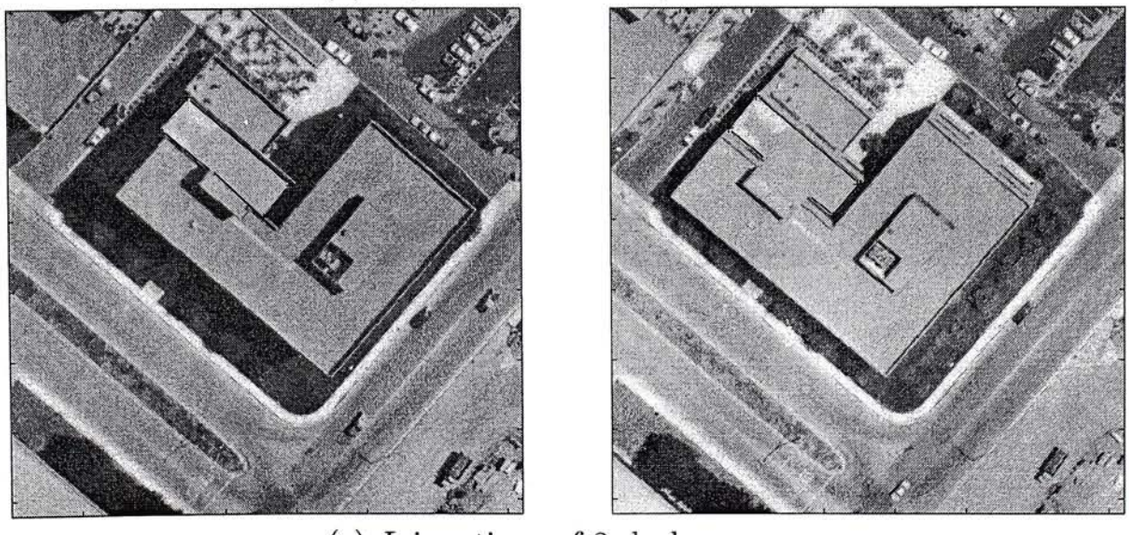
This is different from interpolation of local parameters to predict matches by other approaches. For the same cluttered images, their approach can give very erroneous results. The images shown in this section are all in top view, this does not imply that the approach can only give good results for this category of images. For other views of images with arbitrary camera orientations, the



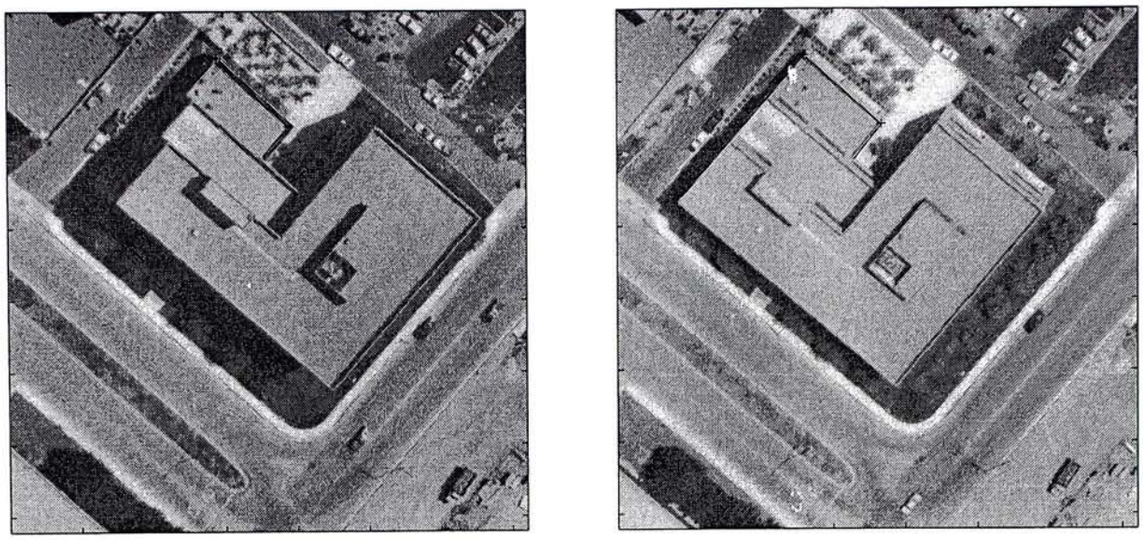
(a) L-junctions of 1st plane



(b) L-junctions of 2nd plane



(c) L-junctions of 3rd plane



(d) L-junctions of 4th plane

Figure 4.15: Four extracted homographies.

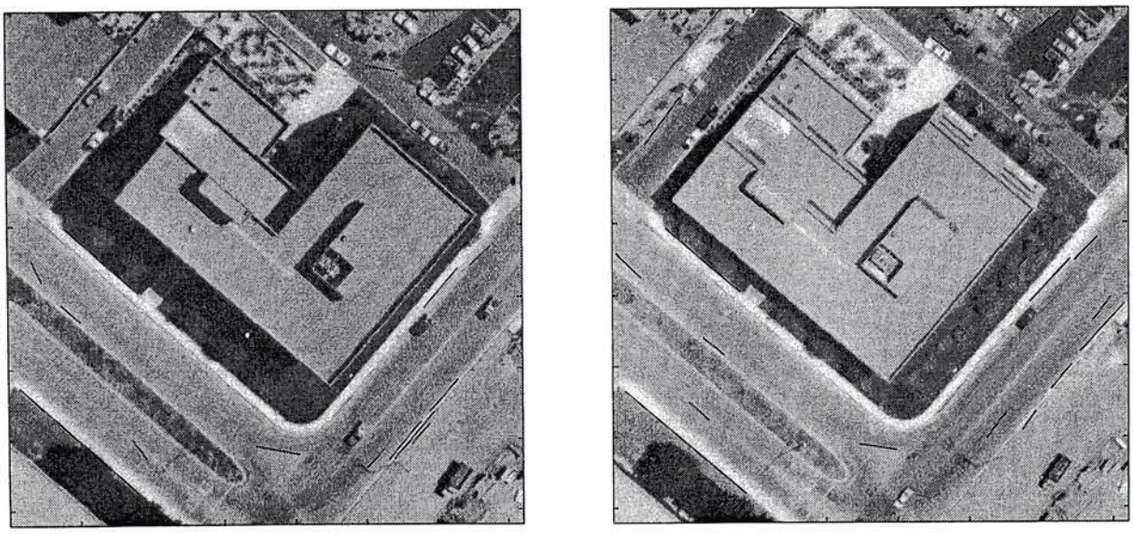
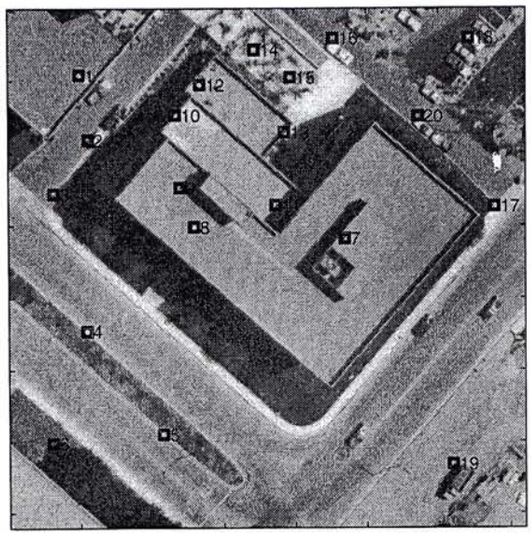
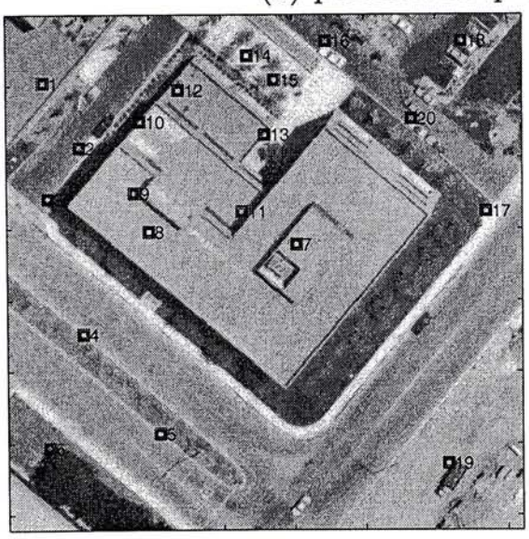


Figure 4.16: Edge matches found using homographies.

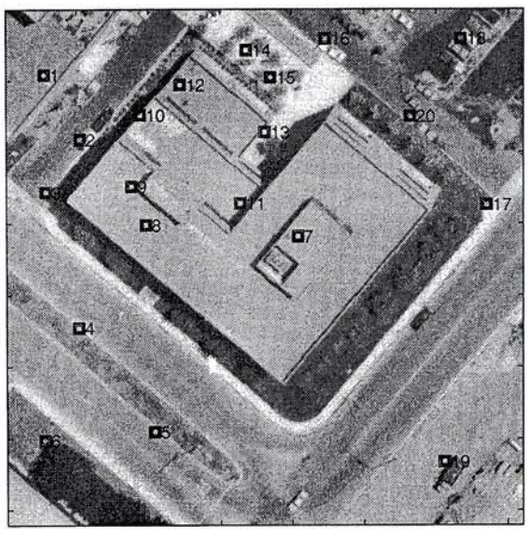
approach can also give good results. The results for other categories of images can be found in the coming section.



(a) position for prediction



(b) predicted position

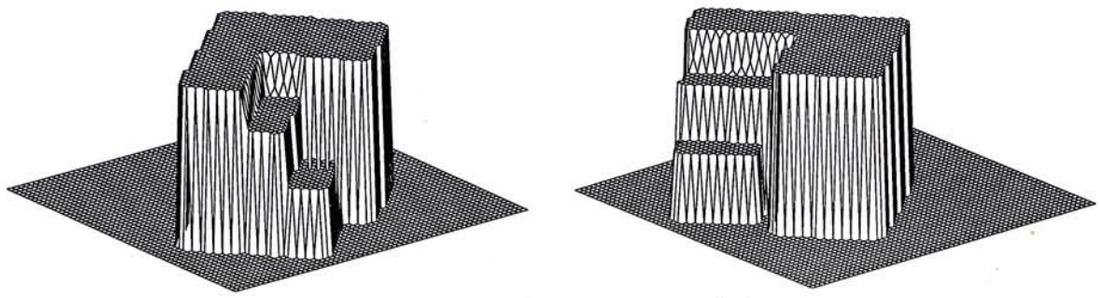


(c) refined position

Figure 4.17: Predicting positions by the homographies obtained.

4.3 Indoor Scenes

In this section, results of indoor images are shown. The results are shown from Figure4.24 to Figure4.30. These indoor images are obtained from the Department of Computer Science, University of Massachusetts at Amberest. These images were captured in an oblique view, different from the aerial view



(a) view (b) view 2

Figure 4.18: Relative disparity map of building #3.

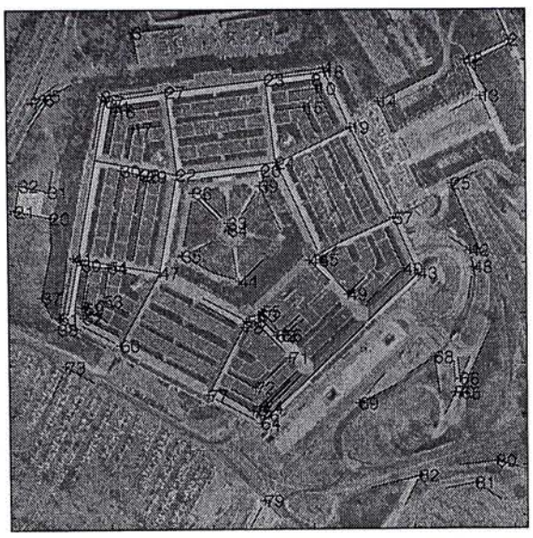
images in the Section4.2. Both image pairs are of dimension 484×512. Not all of the camera calibration parameters are known.

The image pair in Figure4.24(a) and Figure4.24(b) are the first two images of a stereo motion sequence taken at the University of Massachusetts at Amberest from an indoor run. The baseline width between the camera is approximately 20 inches. The image pair in Figure4.27(a) and Figure4.27(b) are the first two images of a single camera motion sequence of an indoor run at U-Mass. The motion between these two images are mostly translational.

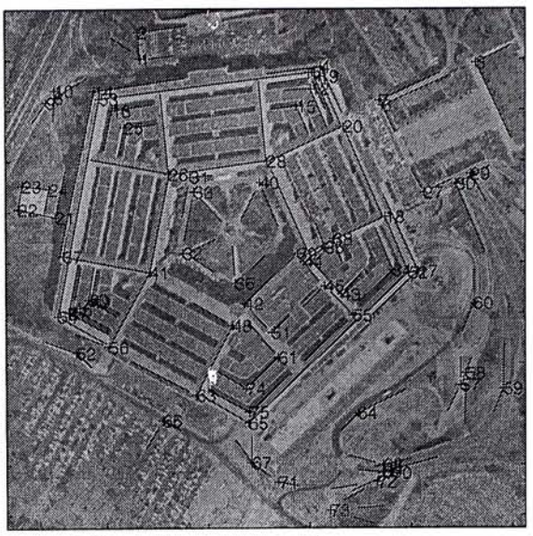
Again, the original images are shown with features marked. Line segments are extracted from the images using the same Canny edge-detector and the Line-fitting subsystem in the Nevatia-Babu "Linear" package. Corner detector is applied to the line segments to obtain L-junctions. For each image pair, initial L-junction matches are found and these initial matches are used in clustering process. As not all the camera parameters are given, epipolar geometries of two image pairs were estimated by manually picking point correspondences.

4.3.1 Stereo Motion Pair

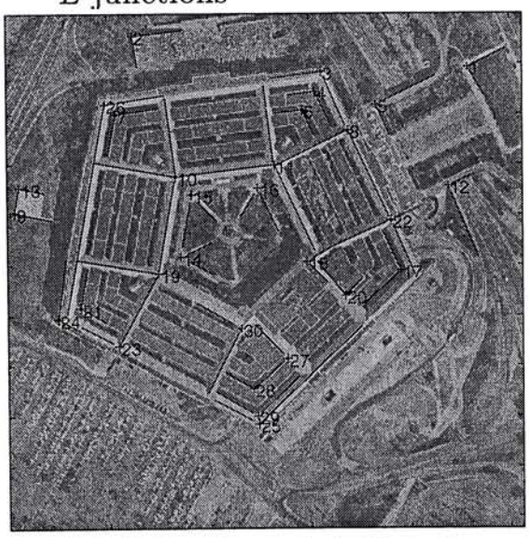
Figure 4.24 shows stereo images of a corridor with two walls, accompanied with the extracted L-junctions. Initial L-junction correspondences unique under the



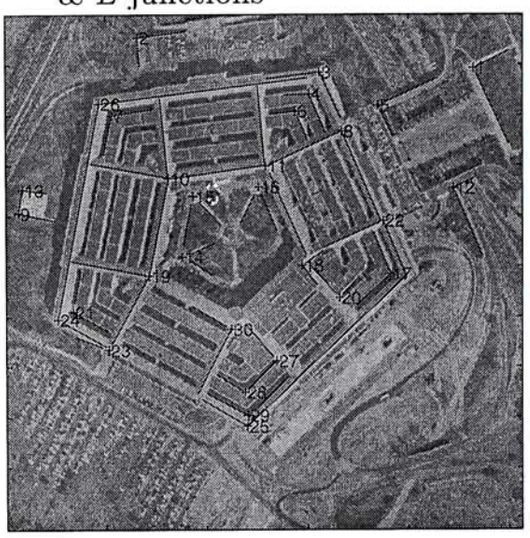
(a) left view of pentagon & L-junctions



(b) right view of pentagon & L-junctions



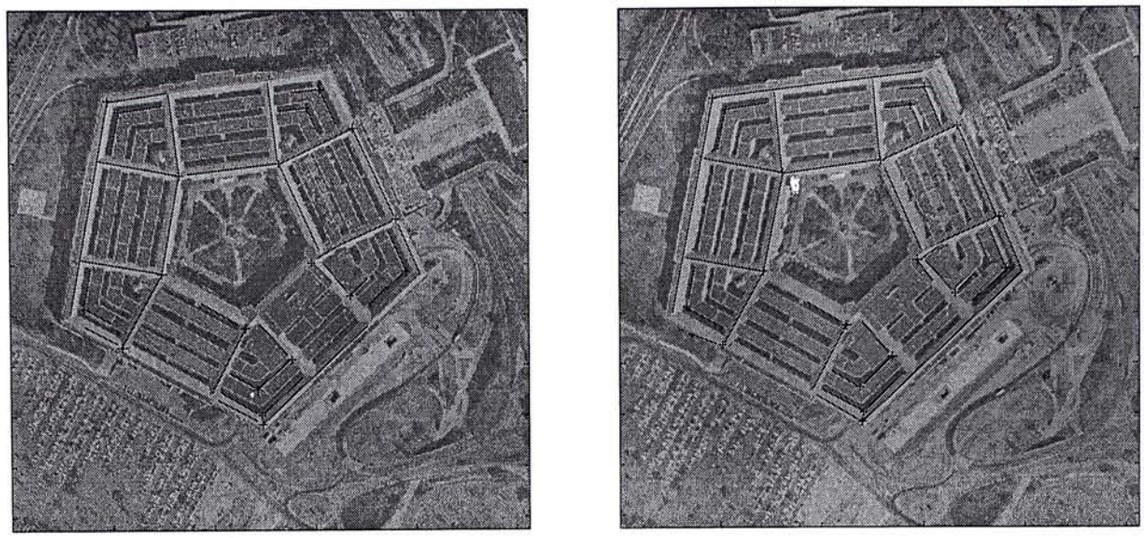
(c) left view of initial L-junction matches



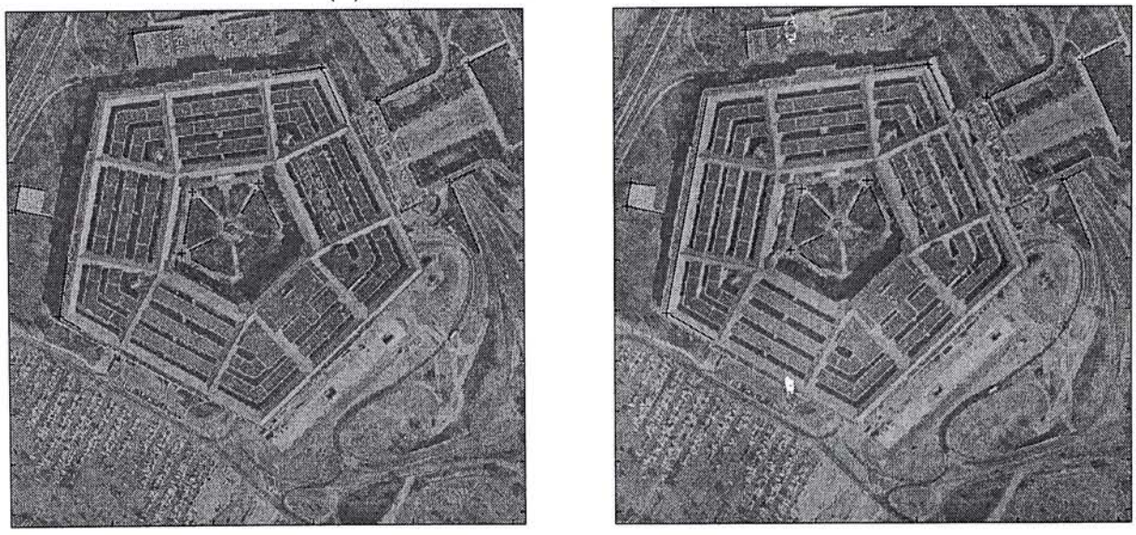
(d) right view of initial L-junction matches

Figure 4.19: Stereo images of pentagon and preliminary processings.

epipolar constraint allowed two homographies to be identified. As shown in Figure 4.24(c) and Figure 4.24(d), one is for the wall on the left and another one for the wall on the right. Careful inspection of the stereo pair can tell that there are altogether three vertical walls. However, the two on the left are so close to each other that they were indistinguishable under the thresholds used in the



(a) L-junctions of 1st plane



(b) L-junctions of 2nd plane

Figure 4.20: Two extracted homographies.

system. It is expected that if the scene is viewed at a closer range, the image resolution would allow the two to be separated. The homography matrices estimated are then used to extrapolate other stereo correspondences and be confirmed by them. The reconstruction result is illustrated with a reprojection of the environment from an oblique angle in Figure 4.26. The walls in this data set are neither parallel nor perpendicular to the image plane. The recovered

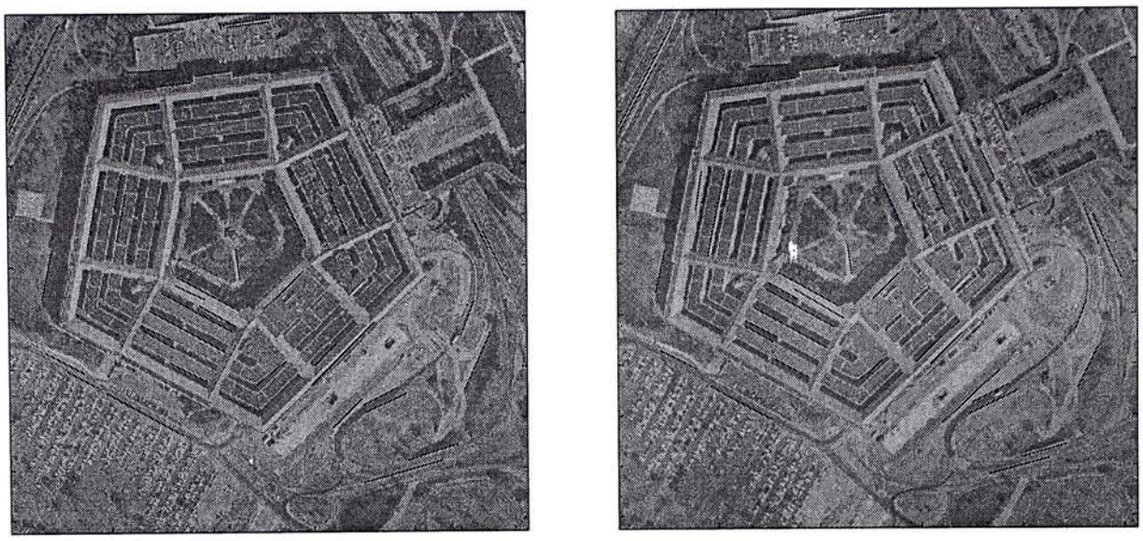
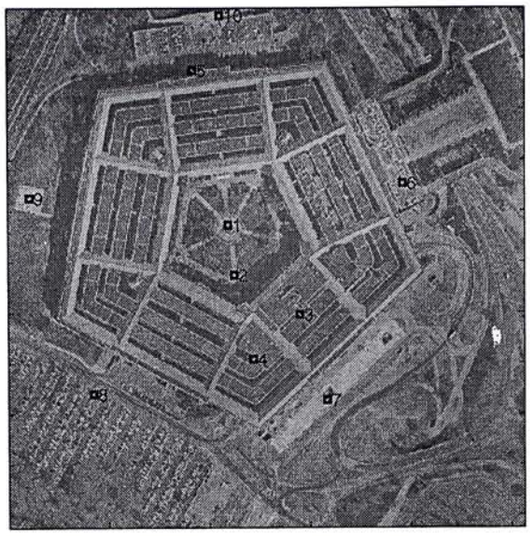


Figure 4.21: Edge matches found using homographies.

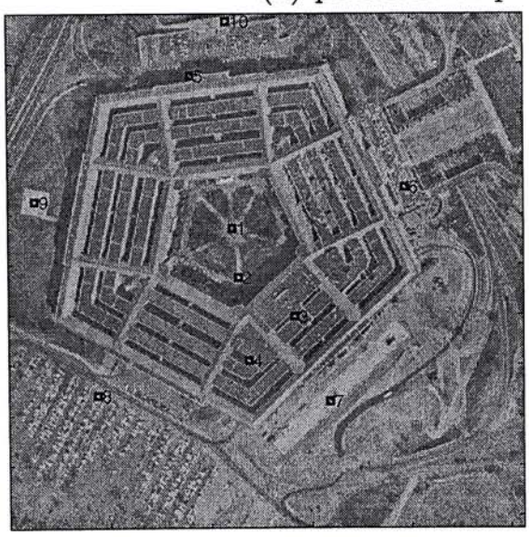
3D information shows that once there are initial matches available for a plane, it can be recovered, no restriction on a plane's 3D orientation.

4.3.2 Hallway Scene

Figures 4.27, 4.28, and 4.30 show results over another set of image data. The stereo images are those of a hallway. The hallway consists mainly of five surfaces, two horizontal surfaces being the ceiling and floor, two vertical surfaces being the left and right vertical wall, and one being the end of the hallway, i.e., the exit. This hallway is so sparsely featured and the contrast of the imaging is so weak that almost no feature can be found except on the surface of the exit. The L-junction correspondences on the four horizontal and vertical surfaces are shown in Figure 4.28(a) and Figure 4.28(b). Even with as few as one L-junction correspondence over each of these four surfaces, they are enough to estimate the corresponding homographies. The exit has more features detected, as shown in Figure 4.28(c) and Figure 4.28(d). All the five surfaces were recovered at the end of the clustering process. There are tiny



(a) position for prediction



(b) predicted position

(c) refined position

Figure 4.22: Predicting positions by the homographies obtained.

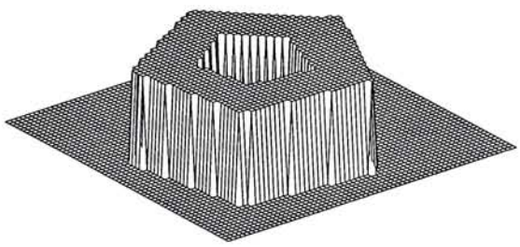
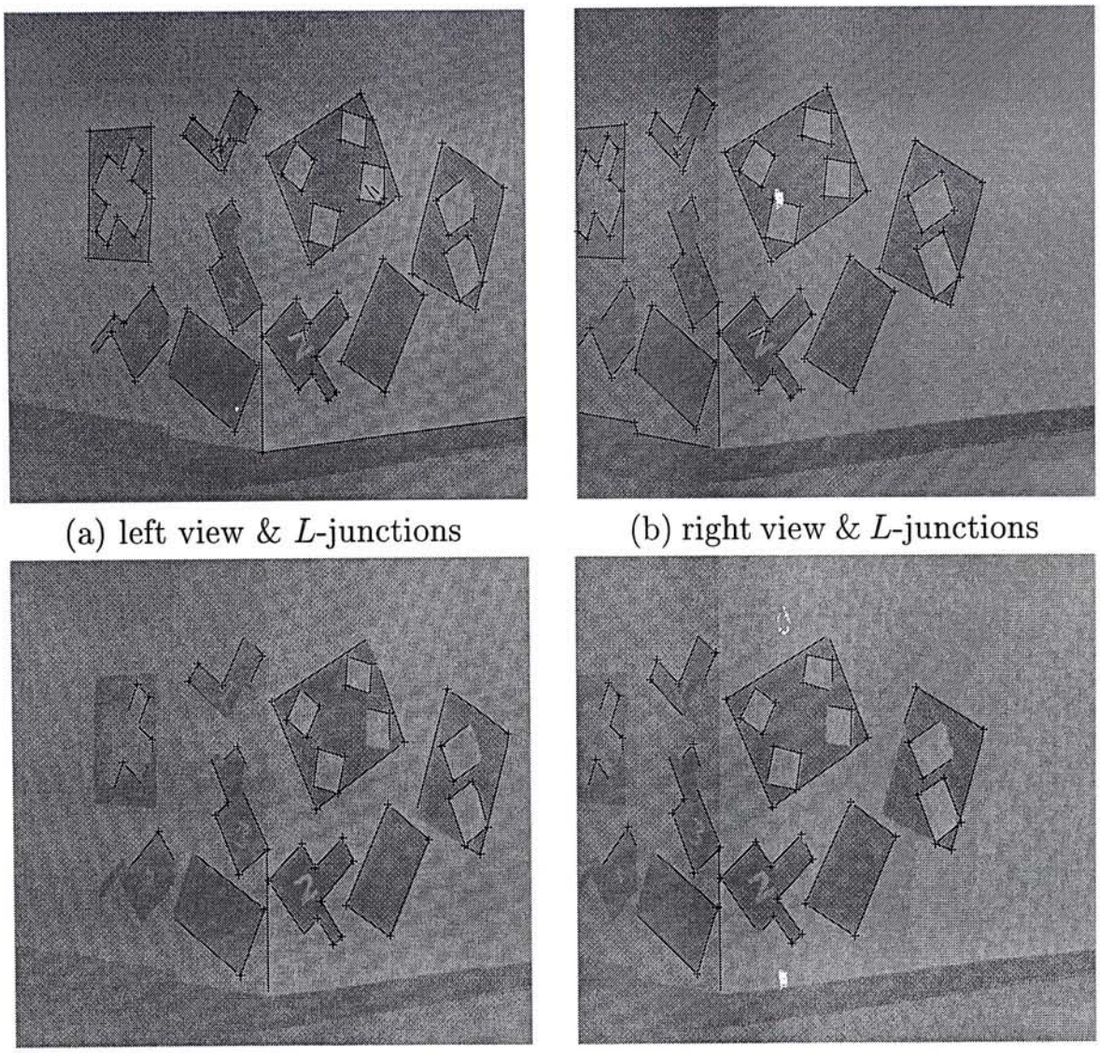


Figure 4.23: Relative disparity map of pentagon.



(a) left view & L-junctions

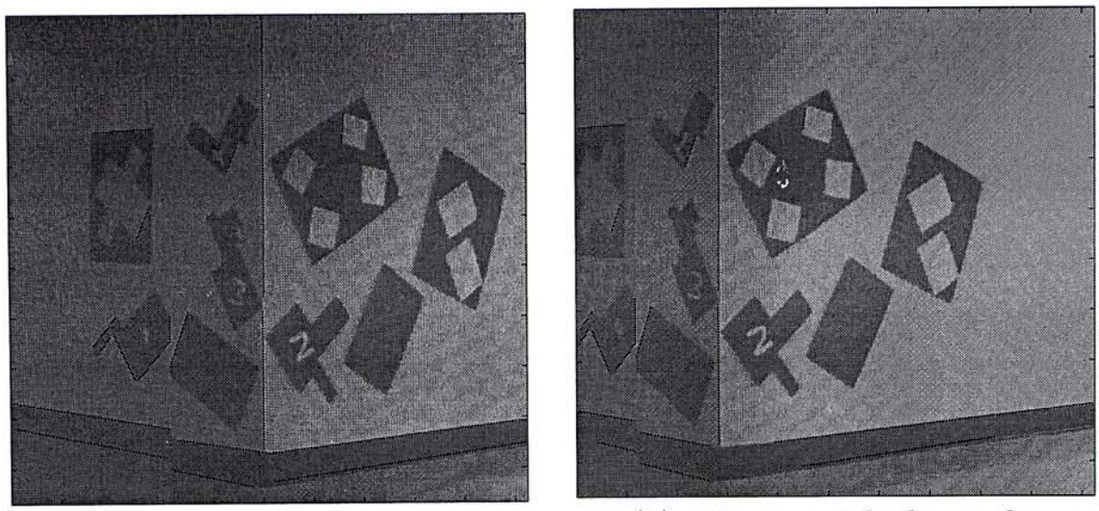
(b) right view & L-junctions

(c) left view of initial L-junction matches

(d) right view of initial L-junctions

Figure 4.24: Stereo images of a corridor and preliminary processings.

surfaces close and parallel to the surface of the exit, but under the thresholds of the implemented system they are indistinguishable. Again, it is expected that when the robot gets closer to the exit, the difference between their homographies and the homography of the exit will be significant enough for them to be isolated. This pair of stereo images is so sparsely featured that it presents great difficulty to generic stereo vision to recover a dense depth map of the



(a) edges matched on the left

(b) edges matched on the right

Figure 4.25: Edge matches found using homographis.

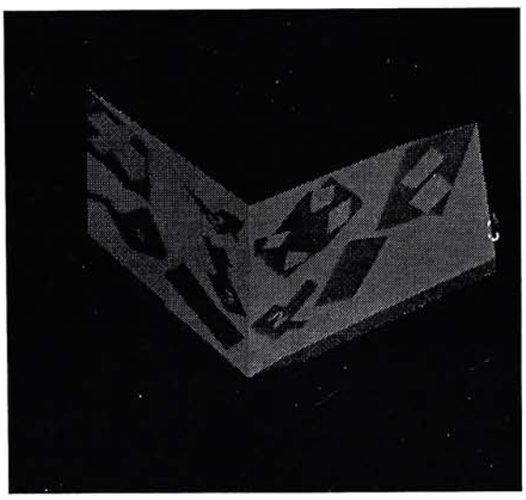
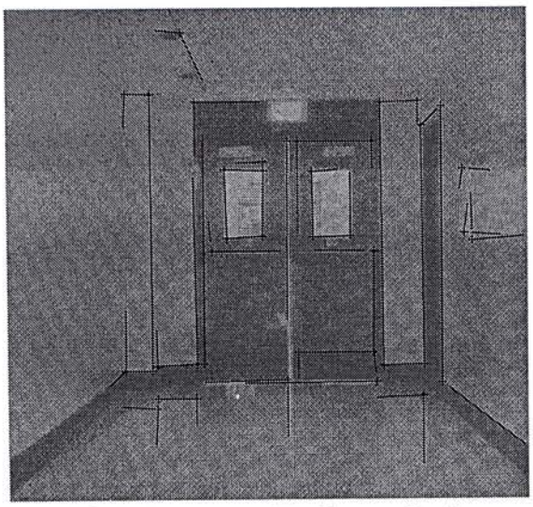
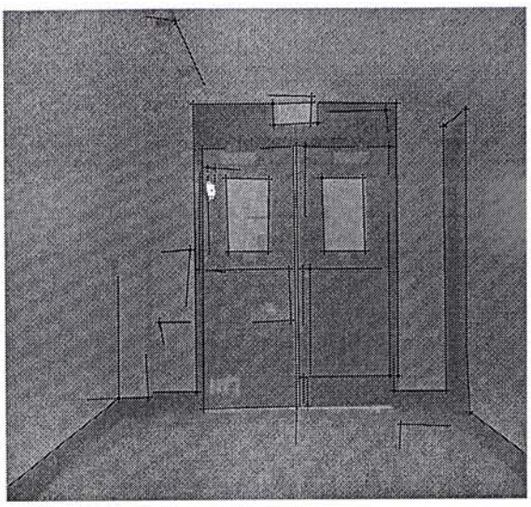


Figure 4.26: Reprojection of the reconstructed result

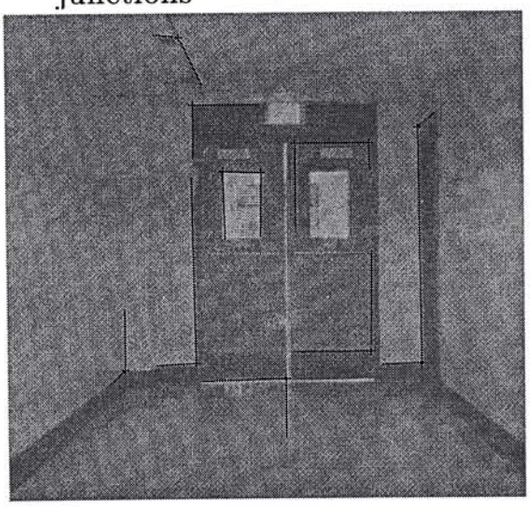
environment. Yet, with the proposed representation and the recovery mechanism, how many surfaces there are and where they are positioned can be estimated. To illustrate the performance of the system, a side view of the reconstructed environment is shown in Figure 4.30, which is a reprojection from



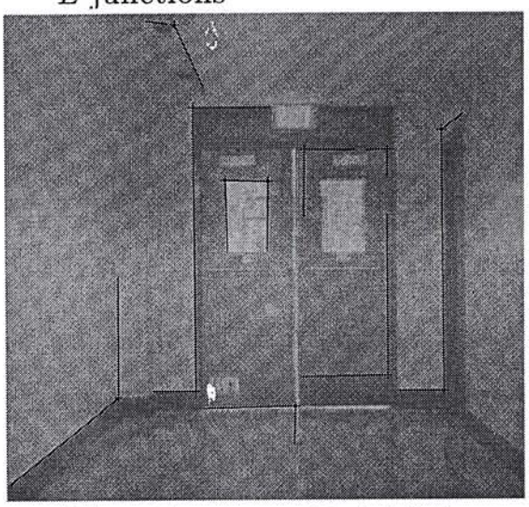
(a) left view of hallway & L-junctions



(b) right view of hallway & L-junctions



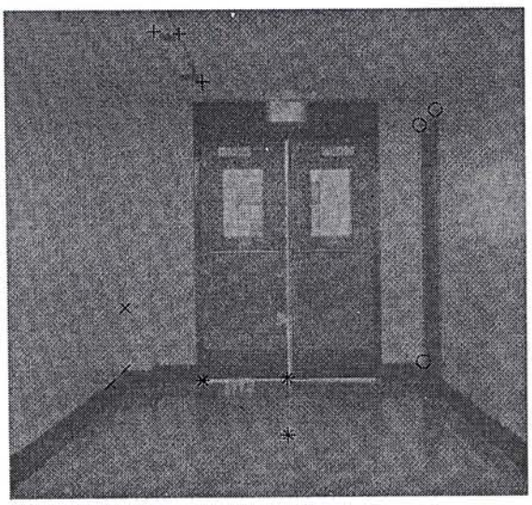
(c) left view of initial L-junction matches



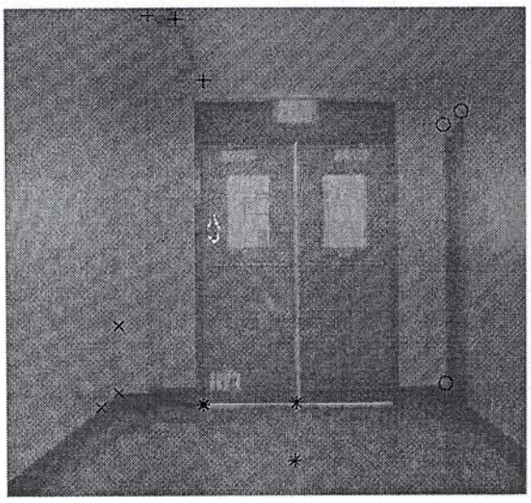
(d) right view of initial L-junctions

Figure 4.27: Stereo images of a hallway and preliminary processings.

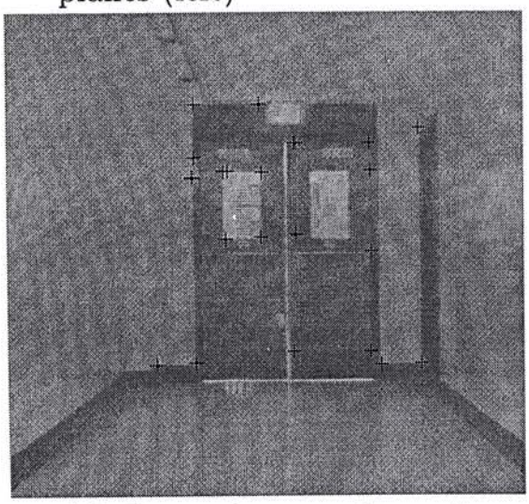
a small elevation angle. It must be noted that part of the exit surface and the floor are occluded by the left wall in this elevated view.



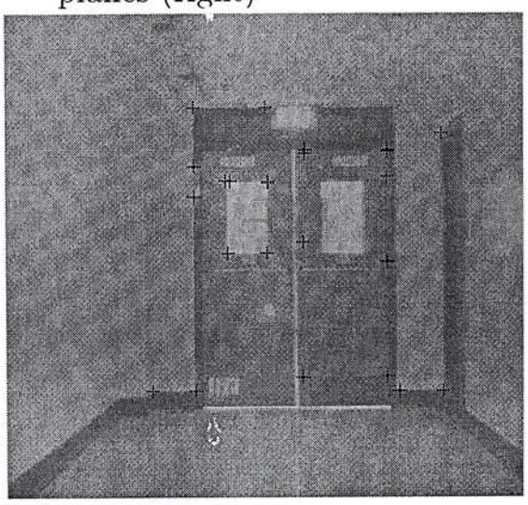
(a) L-junctions on the 4 planes (left)



(b) L-junctions on the 4 planes (right)

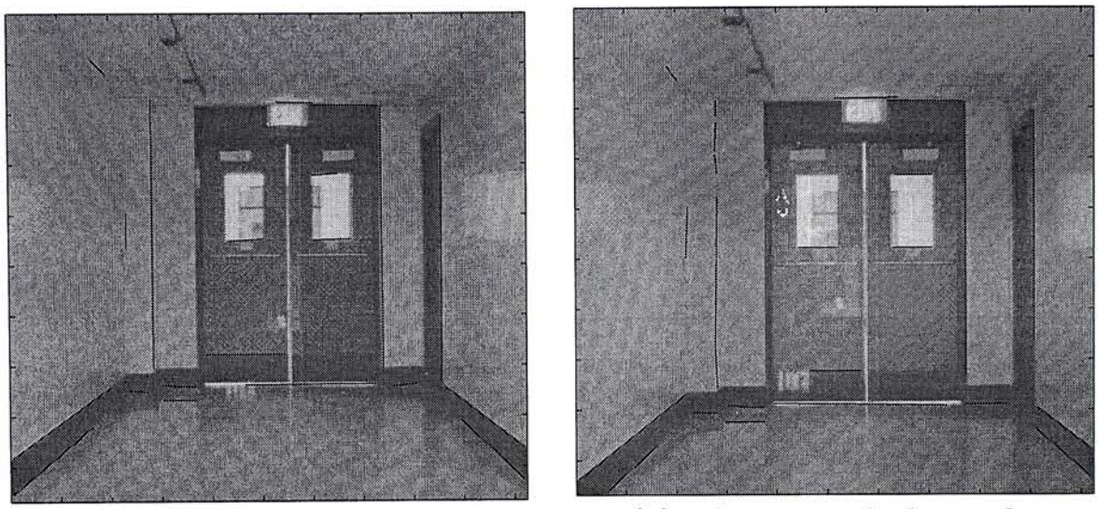


(c) L-junctions on the exit plane (left)



(d) L-junctions on the exit plane (right)

Figure 4.28: Five extracted homographies.



(a) edges matched on the left

(b) edges matched on the right

Figure 4.29: Edges matched found using homographies.

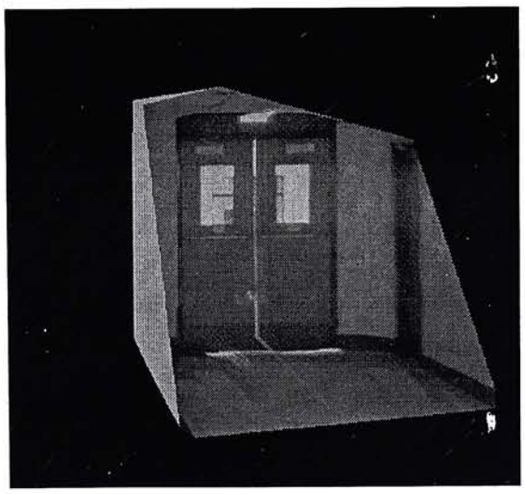


Figure 4.30: Side view of the reconstructed results (hallway).

Chapter 5

Summary and Conclusions

An approach of extracting polyhedral structures from a scene using binocular stereo vision has been presented. The approach does not require the surface-continuity and feature-ordering heuristics as used in the classical approach. It models the stereo correspondence problem as extracting homography matrices due to different planar surfaces in the scene. It estimates the homography matrices from a initial L-junction correspondences which are unique under the epipolar constraint. The homographies are then used to match other image features whose correspondences have ambiguity. The essence of the approach is that not only images features are matched, but a concept of surface segmentation of the scene is also constructed.

Once there are initial information available for 3D plane of any orientation, this plane can be recovered. From the indoor results of Chapter 4, planes being recovered are not restricted to particular orientations, i.e., if there is initial information available on plane which are not parallel and perpendicular to image plane, they can also be recovered. Even using less distinct features such as points, the framework still gives satisfactory results. The approach is not limited in recovering building structures. It can also be used in indoor robot navigation where most of the artificial environments are polyhedral. Industrial

applications such as the location of mechanical parts can also be implemented as the mechanical part are mostly polyhedral.

In this thesis, the targeted environments are polyhedral. Expanding the current approach to cover non-polyhedral objects either by polyhedral approximation of generic objects or formulating matrices that capturing the mapping of a generic plane from one image to another image is expected to be achieved in the foreseeable future. For this expansion, generic environments can be covered not just polyhedral ones. This will be the ultimate goal of this approach.

Reference List

[1] R. Chung and R. Nevatia, "Recovering Building Structures from Stereo", *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 64-73, Palm Spring, CA. USA, 1992.

[2] S. D. Cochranand G. Medioni, "3-D Surface Description from Binocular Stereo", *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 14(10):981-994, Oct. 1992

[3] O. D. Faugeras, E. Le Bras-Mehlman, and J. D. Bossonnat, "Representing Stereo Data With the Delaunay Triangulation", *Artificial Intelligence*, Vol. 44, pp. 863-868, 1991.

[4] O. Faugeras, "What Can be Seen in Three Dimensions with an Uncalibrated Stereo Rig?", *Proceedings of European Conference on Computer Vision*, pp. 563-578, Santa Margherita Ligure, Italy, May 1992.

[5] O. Faugeras, "Stratification of Three-Dimensional Vision: Projective, Affine, and Metric Representation", *Journal of the Optical Society of America - A*, 12(3):465-484, March 1995.

[6] S. Y. Lu and Y. S. Fu, "A Sentence-to-Sentence Clustering Procedure for Pattern Analysis", *IEEE Transactions on Systems, Man and Cybernetics*, 8(5):381-389, May 1978.

[7] Q. T. Luong and T. Vieville, "Canonic Representation for the Geometries of Multiple Projective Views", *Computer Vision and Image Understanding*, 64(2):194-229, Sept. 1996.

[8] M. Maruyama and S. Abe, "Acquiring a Polyhedral Structure Through Face Extraction and Verification", *Proceedings of International Conference on Pattern Recognition*, pp. 579-581, Rome, Italy, Nov. 1988.

[9] R. Mohan and R. Nevatia, "Using Perceptual Organization to Extract 3D Structures", *IEEE Transactions on Pattern analysis and Machine Intelligence*, 11(11):1121-1139, Nov. 1989.

[10] A. Shashua, "Projective Structure from Uncalibrated Images: Structure from Motion and Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):778-790, Aug. 1994.

[11] A. Shashua, "Algebraic Functions for Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779:789, August 1995.

[12] V. Venkateswar and R. Chellappa, "Hierarchical Stereo Matching Using Feature groupings", *Proceedings of the DARPA Image Understanding Workshop*, pages 427-436, San Diego,California, Jan. 1992.