Stereo Vision and Motion Analysis in Complement

by

Ho Pui-Kuen, Patrick

Master of Philosophy Thesis

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

September 15, 1998

Acknowledgments

I would like to thank my supervisor, Prof. Ronald Chung, for his guidance and support in the previous two years. He has not only taught me how to do research work, and also, the attitude of doing research work. ☺

This research was supported by Hong Kong Research Grants Council (RGC) under the 1997-8 Earmarked Grant for Research. It is also part of the project "A Next-generation Intelligent Robot with Creativity" under the Strategic Research Programme of the Chinese University of Hong Kong.

Contents

List Of Figures

List Of Tables

Abstract

Recovering three-dimensional (3D) information of a scene from its images is a fundamental problem in computer vision. There are two major multi-ocular cues for it, namely stereo vision and visual motion. Stereo vision, usually with a substantial baseline, can give accurate results, but feature correspondences across the images are difficult to establish. Visual motion has an easier correspondence problem because consecutive images are alike, but it requires a long image sequence for accurate 3D reconstruction. This thesis presents a new approach of combining the two cues with the objective of retaining their advantages and removing their disadvantages. It is shown that the image measurement data, if organized in a particular way in a matrix, can be decomposed into the 3D structure of the scene, the image projection parameters, the motion parameters, and the stereo geometry separately. With this, the approach can infer stereo correspondences from motion correspondences, requiring only linear time. It does not introduce smoothing in the recovered object structure and camera motion. The singular value decomposition (SVD) techniques and quaternion method are used to reduce the effect of noise in the image measurements. The approach offers the advantages of simple correspondence as well as accurate reconstruction, even with short image sequences. Performance of the approach is illustrated with experimental results on a variety of real images.

Keywords: Stereo-Motion, 3D Reconstruction, Affine Cameras

摘要

重建三維立體資料是計算機視覺中的基本問題。常見的處理方法有以下兩種：立體視像法和活動視像法。前者能計算出準確的物體形狀，惟視像上的對應點難以找尋。而後者的對應問題則較易解決，但它需要大量的圖像，才能得出準確的結果。本文嘗試結合以上兩種方法，去蕪存菁，使對應點既容易找尋而結構重建又準確。此方法乃將圖像特徵資料以矩陣形式列出，並進一步分解成立體結構、圖像投影基數、運動基數及立體相機基數等矩陣，從而以運動圖像對應點求得立體圖像對應點。此方法只需線性時間及較少圖像便能求得準確答案。另外，本文亦以奇異值分解法、四元數算法來減少圖像的噪聲。

關鍵詞：立體視像運動，三維數據重建，仿射相機

Chapter 1

Introduction

Recovering 3D structure of a scene is an important task in many automatic mobile systems, like navigation and robotic manipulation. For example, in a navigation system, a robot must have prior knowledge of the surrounding structure before performing tasks like path planning and localization. We can easily perceive the surrounding 3D structure with our eyes. In principle, a robot can recover 3D information with cameras as well. However, most existing algorithms are either too computational expensive or ineligible with noisy images, which make it difficult to apply to industrial applications.

1.1 Moviation of Problem

Depth information is lost during the projection of a scene to an image. How to recover three-dimensional (3D) information of a scene from two-dimensional (2D) images is a fundamental problem in computer vision. If more than one images are available the problem is potentially easier because of the additional information. There exists at least two vision cues that employ such a multi-ocular approach. One is visual motion, in which 3D structure is recovered from an image sequence acquired under a relative motion between the camera and the scene. The other is stereo vision, in which 3D structure is recovered from two widely separated views of the same scene.

Both the two multi-ocular cues require solutions of two subproblems: the *correspondence problem*, in which image features corresponding to the same

entities in 3D are to be matched across the image frames, and the *reconstruction problem*, in which 3D information is to be reconstructed from the correspondences.

In this thesis visual motion refers to the analysis of *densely* sampled image sequence. The motion cue has the advantage that the correspondence problem is relatively easy to solve, especially for distinct features like edges, junctions, or textured intensity windows, since there is a small limit on how far an image feature can move from one frame to the next one in the image sequence. However, it generally requires a long image sequence, up to hundreds of frames (for instance in [24]), for accurate 3D reconstruction. The reason is, 3D reconstruction from multi-ocular vision is based on the triangulation process, and the triangulation must be wide enough spatially for the sensitivity towards image position errors to be small enough. Requiring long image stream is an intrinsic weakness of the motion cue, for the longer the image sequence is, the more likely the often-required assumptions, like the assumption of a stationary scene, are violated.

In contrast, stereo vision has an easier reconstruction problem but a more difficult correspondence problem. It allows more accurate 3D reconstruction because the two views are widely separated. It has a more difficult correspondence problem because for each feature in one view the search distance for the correspondence in the other view is generally large, although prior knowledge of the spatial relationship of the viewpoints can reduce the originally 2D search to 1D search along the so-called epipolar lines [13].

Many algorithms have been proposed using these two cues. However, because of the difficulties encountered in solving the correspondence or reconstruction problem, most of them can be hardly applied to industrial applications. We note that the advantages and disadvantages of the two cues are contrary to each other. If the two cues are combined, is it possible to remove the disadvantages of two cues, such that a system with easy correspondence and reconstruction problem can be archived?

1.2 Our Approach and Summary of Contributions

This thesis presents an approach of combining stereo and motion analyses for 3D reconstruction, when a mobile platform with two fixed cameras is available to capture stereo pair of image streams. In contrast to previous stereo-motion work, this work emphasizes not on how to exploit the redundancy of the two vision cues to recover more accurate 3D information, but instead on how to couple the two cues tightly to make them complementary, so that their advantages are retained and their disadvantages removed. More precisely, the work aims at achieving simple correspondence and accurate reconstruction all at the same time, even with relatively short image streams. As can be expected, the approach relies on the motion cue in establishing feature correspondences, and on the stereo cue in estimating 3D information of the imaged scene. What is original in the work is a mechanism of relating stereo correspondences to motion correspondences and inferring the former directly from the latter. It turns out that the inference takes only linear time, and that the approach can recover 3D information even without requiring the camera motion be known to the process.

The emphasis of this research is to achieve a system with the following features:

- It recovers 3D information accurately even with relatively short image sequences; this is in principle possible since widely separated views are always in the stereo-motion data regardless of how short the sequences are.

- It does not require prior knowledge of the camera motion nor the assumption of a smooth motion; this frees the system from the effect of disturbances and uncertainty in the camera motion.

– Most importantly, the stereo and motion cues are integrated in a way that they are *complementary* to each other, so that both *simple correspondence* as well as *accurate reconstruction* are possible.

1.3 Organization of this Thesis

The organization of this thesis is as follows. In Chapter 2 related work is first outlined. In Chapter 3 the proposed stereo-motion approach for structure recovery is introduced. The approach has been tested with a large number of synthetic and real image data, and the experimental results are presented in Chapter 4. In Chapter 5, an approach that can recovery both object structure and camera motion is described. This approach utilizes data from all images to compute object structure, thus more accurate results can be obtained. Some more experimental results are presented in Chapter 6. A conclusion is then drawn in Chapter 7. Some background information related to our work can be found in the appendices. A portion of the work has been reported in [8].

Chapter 2

Previous Work

Much has been done on stereo vision; good surveys can be found in [4, 11]. Yet the technology is not robust enough to have been extensively used in industry and society, due mostly to the difficulty of solving the correspondence problem.

Visual motion has also been well-studied; classical references are listed in [14, 26]. The correspondence problem is much simpler than that in stereo vision, as consecutive images are "alike" and can be corresponded easily. Very good 3D reconstruction results have been obtained, for example in [24]. One drawback is that a long image sequence is required so as to have a wide enough triangulation for accurate 3D determination.

Below some recent work on motion analysis are outlined. It must be emphasized that the listed work are by no means complete; they are specifically mentioned because they are closely related to the work presented in this paper.

In an elegant work, Tomasi and Kanade [24] proposed a method for reconstructing 3D from an orthographically projected image sequence. It organizes the image measurements of object points in matrix form, and then factorizes it into shape and motion matrices through singular value decomposition (SVD). Later, Poelman and Kanade [18] extended the factorization method to the case of paraperspective projection, which produces more accurate results than the original method, especially when the object is offset from the camera principal axis. The factorization approach uses a large number of image measurements to counteract the noise sensitivity of structure-from-motion.

However, for accurate reconstruction, a long image sequence is needed. Extensive computational time is required for processing these images. Recently, Morita and Kanade [16] presented a sequential approach for the factorization method. This approach is similar to the orthographic one, except it does not handle the all image measurements in one time. The size of working matrix is reduced. Since the SVD works more efficiently with small matrices, the computational time is thus reduced. However, a long image sequence is still required, and much computational time is required for processing these images (e.g. tracking of features).

In [17], Okutomi and Kanade presented an approach that recover object structure from a set of stereo images with different baseline. Stereo matches across each stereo pair are established, and the sum of squared-difference (SSD) values for individual stereo pairs are calculated. The SSD values are added to produce SSD-in-inverse-distance function, which exhibits a unique minimum at correct matching position. This stereo matching approach can reduce false matches and increase precision of the recovered structure. However, the correspondence problem is still difficult to solve.

The motion cue under an unknown motion recovers the world only up to a scale factor. One way to remove this ambiguity is to use two cameras to take stereo pair of image sequences and to combine stereo and motion analyses. The redundancy in the image data – data for both stereo and motion cues – also has the potential of allowing 3D information to be recovered more accurately. In this connection, a few studies [25, 15, 29, 1, 28, 7] have looked into the so-called stereo-motion cue.

However, the focus of the above stereo-motion work is on exploiting the redundancy in the input data in recovering 3D information. How the two vision cues complement each other and what can be gained by combining them have not been explicitly addressed. This paper presents a framework that has an explicit mechanism to combine the advantages of the motion and stereo cues to provide simple correspondence as well as accurate reconstruction.

Chapter 3

Structure Recovery from Stereo-Motion Images

In this chapter the proposed framework of stereo-motion is presented. The framework can be outlined as the following. Motion correspondences are first established over the two image sequences separately. This can be achieved reliably as long as the image sequences are densely sampled. The trouble is, given the condition that the image sequences can be short, motion correspondences even correctly established may not admit an accurate Euclidean reconstruction of the scene. To make accurate Euclidean reconstruction possible, one possibility is to bring the two sets of separately established motion correspondences into some form of registration, so that in a way stereo correspondences are established across the two image streams.

On relating two sets of motion correspondences of the same scene, there are some work in the literature which may help. It was recently understood that 3D reconstruction from correspondences over multiple views does not necessarily have to be at the Euclidean level; it can be at the affine level or even the projective level. Faugeras provides an excellent tutorial of the concept in [6]. Affine and projective reconstructions are less specific than Euclidean reconstruction, and are thus less sensitive to errors. In other words, it is possible that motion correspondences over two image streams can be related more accurately by registering not the Euclidean reconstructions, but affine reconstructions, of the imaged scene.

The stereo-motion framework used in this work is motivated from the above idea. However, it does not even require explicit affine reconstructions from the two image streams. Instead, it infers stereo correspondences directly from motion correspondence through simple matrix manipulations. The inference mechanism is among the most important contributions of the work.

The stereo-motion uses a motion model modified from the one in [24]. The motion model is first described. Description of the stereo-motion framework then follows.

3.1 Motion Model

Tomasi and Kanade [24] have proposed an elegant discrete model for the motion cue. Below the model, with some variations, is described; the variations are to make the subsequent extension to the stereo-motion problem easier.

Suppose F image frames observing P points in space are available. Assume an affine model for the camera projection process. The image position $\mathbf{p}_{fp} = (u_{fp}, v_{fp})^{\mathrm{T}}$ of point p ($p = 0, 1, \ldots, (P-1)$) in image frame f ($f = 0, 1, \ldots, (F-1)$), is related to its 3D position $\mathbf{P}_p = (x_p, y_p, z_p)^{\mathrm{T}}$ (with reference to the last image frame: frame $(F-1)$), by

$$\mathbf{p}_{fp} = J_f \left[\begin{array}{ccc|c} & \mathbf{R}_f & & \mathbf{t}_f \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{c} \mathbf{P}_p \\ \hline 1 \end{array}\right]$$

where J_f is the affine projection matrix (a 2×4 matrix), and $(\mathbf{R}_f, \mathbf{t}_f)$ are the rotational and translational relationships between image frame f and the last image frame. $\mathbf{P}_p$ for all points p are the 3D structure to be reconstructed. Affine projection model is the general form of several commonly used linear projection model (e.g. orthographic, weak-perspective, and paraperspective projection models, see Appendix A.1 for more details). It closely approximates the camera projection in most cases.

Combining the above of all P object points in a single frame, one gets

$$\left[\begin{array}{ccc} \cdots & \mathbf{p}_{fp} & \cdots \end{array}\right] =$$

$$J_f \underbrace{\left[\begin{array}{c|c} \mathbf{R}_f & \mathbf{t}_f \\ \hline 0\ 0\ 0 & 1 \end{array}\right]}_{\mathbf{M}_f} \underbrace{\left[\begin{array}{ccc} & \mathbf{P}_p & \\ \cdots & 1 & \cdots \end{array}\right]}_{\mathbf{S}}$$

Combining the above of all F image frames, the following can be obtained:

$$\underbrace{\left[\begin{array}{ccc} & \vdots & \\ \cdots & \mathbf{p}_{fp} & \cdots \\ & \vdots & \end{array}\right]}_{\mathbf{W}} = \underbrace{\left[\begin{array}{ccc} \ddots & & \bigcirc \\ & J_f & \\ \bigcirc & & \ddots \end{array}\right]}_{\mathbf{J}} \underbrace{\left[\begin{array}{c} \vdots \\ \mathbf{M}_f \\ \vdots \end{array}\right]}_{\mathbf{M}} \mathbf{S}$$

Here $\mathbf{W}$ is a $2F \times P$ matrix representing the image measurements. Each row contains the u-coordinates or v-coordinates of points in a single frame, while each column contains the observations for a single point. $\mathbf{J}$ is a $2F \times 4F$ matrix representing the image projection process. $\mathbf{M}$ is a $4F \times 4$ matrix representing the entire camera motion. $\mathbf{S}$ is a $4 \times P$ matrix representing the 3D structure with reference to the last image frame. Since $\mathbf{W}$ can be factorized into matrices involving dimension four, $\mathbf{W}$ is of rank at most four (it is exactly four under general motion and general 3D structure).

A similar motion model has been applied successfully to recover 3D structure [24]. However, hundreds of frames are needed. If instead a stereo pair of cameras are available to acquire a stereo pair of image sequences, potentially even with relatively short image sequences the 3D structure can still be estimated accurately, as widely separated views are always in the data. The question is, how can the difficult stereo correspondence problem be solved with help from the solution of the simpler motion correspondence problem?

In the next subsection the motion model is extended to the stereo-motion problem. It is shown how the 3D structure, the motion, the stereo geometry, and the camera parameters can be decomposed from a matrix of the motion and stereo correspondences, and how the stated goals – simple correspondence as well as accurate reconstruction – can be achieved.

3.2 Stereo-Motion Model

As shown in Figure 3.1, let $(\bar{\mathbf{R}}, \bar{\mathbf{t}})$ be the rotational and translational relationships between the stereo cameras, in the sense that the 3D coordinates of any point with respect to the two camera coordinates frames, $\mathbf{P}$ and $\mathbf{P}'$, are related by

$$\begin{bmatrix} \mathbf{P}' \\ 1 \end{bmatrix} = \left[\begin{array}{ccc|c} & \bar{\mathbf{R}} & & \bar{\mathbf{t}} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix}$$

$$= \bar{\mathbf{M}} \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix}$$

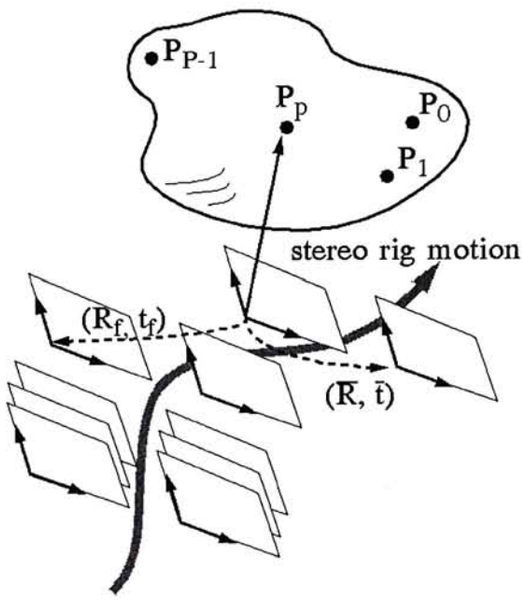


Figure 3.1 Stereo-motion image sequences from two cameras mounted on a mobile platform.

Applying the above motion model to the two cameras separately, one obtains two image measurement matrices for feature points in the two cameras respectively:

$$\mathbf{W} = \mathbf{JMS}$$

$$\mathbf{W}' = \mathbf{J}'\mathbf{M}'\mathbf{S}'$$

Here $\mathbf{W}', \mathbf{J}', \mathbf{S}'$ are matrices in the same form as $\mathbf{W}, \mathbf{J}, \mathbf{S}$, but $\mathbf{W}', \mathbf{J}', \mathbf{S}'$ are with respect to the second camera whereas $\mathbf{W}, \mathbf{J}, \mathbf{S}$ are with respect to the first camera.

Suppose stereo correspondences are established correctly across the two image sequences. This means feature points in $\mathbf{W}'$ and $\mathbf{S}'$ can be listed in the same left-to-right order of those in $\mathbf{W}$ and $\mathbf{S}$. Then $\mathbf{S}'$ is related to $\mathbf{S}$ by

$$\underbrace{\begin{bmatrix} \cdots & \mathbf{P}'_p & \cdots \\ & 1 & \end{bmatrix}}_{\mathbf{S}'} = \bar{\mathbf{M}} \underbrace{\begin{bmatrix} \cdots & \mathbf{P}_p & \cdots \\ & 1 & \end{bmatrix}}_{\mathbf{S}}$$

Due to rigidity of the stereo camera system, the relative motion between image frame f and the last image frame of the second camera, $\mathbf{M}'_f$, is related to the corresponding relative motion of the first camera, $\mathbf{M}_f$, by

$$\mathbf{M}'_f = \bar{\mathbf{M}} \mathbf{M}_f \bar{\mathbf{M}}^{-1}$$

as illustrated below:

$$\begin{array}{ccc} \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} & \xrightarrow{\mathbf{M}_f} & \mathbf{M}_f \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} \\ \downarrow \bar{\mathbf{M}} & & \downarrow \bar{\mathbf{M}} \\ \bar{\mathbf{M}} \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} & \xrightarrow{\mathbf{M}'_f} & \mathbf{M}'_f \bar{\mathbf{M}} \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix} = \bar{\mathbf{M}} \mathbf{M}_f \begin{bmatrix} \mathbf{P} \\ 1 \end{bmatrix}, \forall \mathbf{P} \end{array}$$

Combining the above of all F image frames, one has

$$\begin{aligned} \mathbf{M}' &= \begin{bmatrix} \vdots \\ \mathbf{M}'_f \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \bar{\mathbf{M}} \mathbf{M}_f \bar{\mathbf{M}}^{-1} \\ \vdots \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \ddots & & \bigcirc \\ & \bar{\mathbf{M}} & \\ \bigcirc & & \ddots \end{bmatrix}}_{\widetilde{\mathbf{M}}} \underbrace{\begin{bmatrix} \vdots \\ \mathbf{M}_f \\ \vdots \end{bmatrix}}_{\mathbf{M}} \bar{\mathbf{M}}^{-1} \end{aligned}$$

As a result, the image measurement matrix of the second camera can be written as $\mathbf{W}' = \mathbf{J}'(\widetilde{\mathbf{M}} \mathbf{M} \bar{\mathbf{M}}^{-1}) \bar{\mathbf{M}} \mathbf{S}$ or

$$\mathbf{W}' = \mathbf{J}' \widetilde{\mathbf{M}} \mathbf{M} \mathbf{S}$$

If $\mathbf{W}'$ is stacked under $\mathbf{W}$ such that each column of the resultant matrix corresponds to the same feature in space, a new image measurement matrix is then obtained for the stereo-motion cue:

$$\underbrace{\left[\begin{array}{c}\mathbf{W}\\\hline\mathbf{W}'\end{array}\right]}_{\mathcal{W}} = \left[\begin{array}{c}\mathbf{JMS}\\\hline\mathbf{J}'\mathbf{M}'\mathbf{S}'\end{array}\right] = \left[\begin{array}{c}\mathbf{JMS}\\\hline\mathbf{J}'\widetilde{\mathbf{M}}\mathbf{MS}\end{array}\right]$$

$$= \underbrace{\left[\begin{array}{c|c}\mathbf{J} & \bigcirc\\\hline\bigcirc & \mathbf{J}'\end{array}\right]}_{\mathcal{J}} \underbrace{\left[\begin{array}{c}\mathbf{I}_{4F}\\\hline\widetilde{\mathbf{M}}\end{array}\right]}_{\widetilde{\mathcal{M}}} \mathbf{MS} \tag{3.1}$$

Here $\mathcal{W}, \mathcal{J}, \widetilde{\mathcal{M}}$ are analogues of $\mathbf{W}, \mathbf{J}, \mathbf{M}$ (which are for single-camera motion) in the stereo-motion system, each containing information about the stereo-motion data. $\mathcal{W}$ is a $4F \times P$ matrix representing the image measurements over the stereo cameras (with the stereo correspondences correctly established). $\mathcal{J}$ is a $4F \times 8F$ matrix representing the image projection parameters of the stereo cameras. $\widetilde{\mathcal{M}}$ is a $8F \times 4F$ matrix representing the stereo geometry. The overall $(\mathcal{J}\widetilde{\mathcal{M}}\mathbf{M})$ is then a $4F \times 4$ matrix.

Since the factorization in Equation 3.1 involves matrices with dimension four, $\mathcal{W}$ in stereo-motion, like $\mathbf{W}$ or $\mathbf{W}'$ in single-camera motion, is of rank at most four. Such a property is unlikely to be satisfied accidentally, as $\mathcal{W}$ is $4F \times P$ large; it is however satisfied when $\mathcal{W}$ is constructed under fully correct stereo matching.

The rank property is important as it can help in establishing stereo correspondences. For example, stereo matching across the two image sequences should always keep $\texttt{Rank}(\mathcal{W}) \leq 4$. In general, any mismatch across the stereo views would raise the rank of $\mathcal{W}$ beyond four. Of course, with noise, even under correct stereo matching $\mathcal{W}$ may turn to be of full rank. But singular values of $\mathcal{W}$ can be obtained via the singular value decomposition (SVD) technique, and it can be checked if the singular values of $\mathcal{W}$ after the four most significant ones are close enough to zero or not. However, merely trying different combinations of stereo correspondences and testing if any one of them satisfies the

rank property is too passive a strategy and too inefficient a process. Below a more efficient mechanism of inferring stereo correspondences is described.

3.3 Inferring Stereo Correspondences

The proposed stereo-motion system proceeds in the following way. A stereo rig of cameras is constructed, and it undergoes a motion during which F pairs of images are taken from the cameras. Distinct feature points are then extracted from the first stereo pair of image sequences, and are tracked over the two sequences separately. The estimated motion correspondences are assumed to be mostly correct as the image frames are dense. With such motion correspondences image measurement matrices $\mathbf{W}^*$ and $\mathbf{W}'^*$ can be constructed. $\mathbf{W}^*$ and $\mathbf{W}'^*$ are in the same form as $\mathbf{W}$ and $\mathbf{W}'$, except that their columns are not properly ordered, i.e. no stereo correspondence is established yet. They may also have different number of columns, as feature points observable in one image sequence may not be observable in the other.

Establishing stereo correspondences across the two images sequences is equivalent to matching columns of $\mathbf{W}^*$ with columns of $\mathbf{W}'^*$ so as to form the matrix $\mathcal{W}$ with the matched columns. Such a problem turns out to be rather trivial to solve.

Suppose $\mathcal{W}$, $\mathbf{W}$ and $\mathbf{W}'$ are of rank four, its column space is only a 4D subspace in a $(4F)$-dimension vector space, and all columns in $\mathcal{W}$ are linear combinations of 4 independent vectors. Suppose the four basis vectors of $\mathcal{W}$ are known. For any column $\mathbf{h}$ of $\mathcal{W}$, there exists four scalars $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ such that

$$\mathbf{h} = \sum_{i=1}^{4} \alpha_i \mathbf{b}_i = \mathbf{B} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}$$

where $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4$ are the four basis vectors of the column space of $\mathcal{W}$, and $\mathbf{B}$ is a basis matrix equals to $[\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4]$. The $4F$ elements of $\mathbf{h}$ can be

separated into two groups: one belonging to $\mathbf{W}^*$, forming a sub-vector $\mathbf{h}_W$, and the other to $\mathbf{W}^{*\prime}$, forming another sub-vector $\mathbf{h}_{W'}$. As the sub-vector $\mathbf{h}_W$ is known from $\mathbf{W}^*$, the scalar breakdown of the above vector equation implies there exists $2F$ linear equations for the four unknowns $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. If $2F \geq 4$ the scalars can be determined uniquely, then the sub-vector $\mathbf{h}_{W'}$ in $\mathbf{W}^{*\prime}$ can be inferred.

To put it more precisely, one can divide $\mathbf{B}$ into two sub-matrices $\mathbf{B}_W$ and $\mathbf{B}_{W'}$ according to which of $\mathbf{W}^*$ and $\mathbf{W}^{*\prime}$ the elements belong to, and obtain

$$\mathbf{h}_W = \mathbf{B}_W \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}, \quad \mathbf{h}_{W'} = \mathbf{B}_{W'} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}$$

Using the least-squares-error method, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ can be estimated as:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \mathbf{B}_W^{\mathrm{I}} \, \mathbf{h}_W$$

where $\mathbf{B}_W^{\mathrm{I}}$ is the pseudo-inverse of $\mathbf{B}_W$ and is given by $\mathbf{B}_W^{\mathrm{I}} = (\mathbf{B}_W^{\mathrm{T}} \mathbf{B}_W)^{-1} \mathbf{B}_W^{\mathrm{T}}$. $\mathbf{h}_{W'}$ can then be estimated from the α_i's.

In summary, $\mathbf{h}_{W'}$ can be inferred from $\mathbf{b}_i$'s and $\mathbf{h}_W$ simply as

$$\mathbf{h}_{W'} = \mathbf{B}_{W'} \, (\mathbf{B}_W^{\mathrm{T}} \mathbf{B}_W)^{-1} \, \mathbf{B}_W^{\mathrm{T}} \, \mathbf{h}_W \tag{3.2}$$

That is, given any column of $\mathbf{W}^*$, the corresponding column in $\mathbf{W}'^*$ can be predicted, provided that the basis vectors of $\mathcal{W}$ are known. The basis vectors can be formed if four linearly independent columns of $\mathcal{W}$ are available, which are equivalent to a minimum of four features matched across any stereo pair in the image data. Such initial correspondences may be obtained by observing and matching stereoscopically very distinct features like corners in the images. Alternatively, the epipolar geometry of the stereo cameras can be pre-estimated in an off-line process, and the epipolar lines so obtained, plus a restricted range

of disparity, often result in a few unique stereo correspondences. The latter approach is used in this work.

If more than 4 matches are available, a more accurate basis can be determine by minimizing the least-squares-error. Suppose u matches are found using the epipolar constraint and the restricted disparity gradient, where $u \geq 4$. A sub-matrix of $\mathcal{W}$ (size $4F \times u$) can be formed from these initial matches. $\mathcal{W}_u$ can be decomposed through SVD into $\mathcal{W}_u = \mathbf{U}_u \Sigma_u \mathbf{V}_u^{\mathsf{T}}$, where $\mathbf{U}_u, \mathbf{V}_u$ are orthogonal matrices, and Σ_u a diagonal matrix. Under the least-squares-error criterion the first four columns of $\mathbf{U}_u$ then form an optimal basis of the above-mentioned 4D vector subspace [8][24].

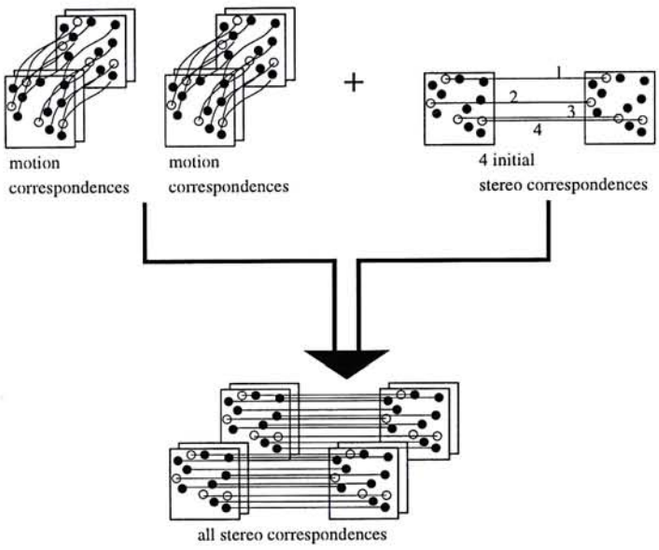


Figure 3.2 Input-output description of the correspondence-inference mechanism. Stereo correspondences can be estimated if four initial stereo matches and motion correspondences in both sequences are available.

If all the stereo correspondences are to be considered at the same time, one can view the inference problem in the following way. $\mathcal{W}$ can be written in the following form:

$$\mathcal{W} = \left[\begin{array}{c|c} \mathbf{W}_u & \mathbf{W}_v \\ \hline \mathbf{W}'_u & \mathbf{W}'_v \end{array}\right]$$

where the two sub-matrices $\mathbf{W}_u$ and $\mathbf{W}'_u$ on the left represent the initial stereo correspondences established by the epipolar constraint, and the sub-matrices

$\mathbf{W}_v$ and $\mathbf{W}'_v$ represents the remaining columns in $\mathbf{W}^*$ and $\mathbf{W}'^*$. The sub-matrix $\mathbf{W}'_v$ can be estimated from $\mathbf{W}_v$ by Equation 3.3

$$\mathbf{W'} = \mathbf{B}_{W'} (\mathbf{B}_W^{\mathsf{T}} \mathbf{B}_W)^{-1} \mathbf{B}_W^{\mathsf{T}} \mathbf{W} \tag{3.3}$$

Equation 3.3 can be easily derived from Equation 3.2. Without lost of generality, the matrix $\mathbf{W}$ can be set equal to $\mathbf{W}^*$, and then estimate $\mathbf{W}'_v$ from it. With noise, columns in the estimated $\mathbf{W}'_v$ may not be exactly those in $\mathbf{W}'^*$, but should be quite close to them. For each column in $\mathbf{W}'_v$, a column is then selected from $\mathbf{W}'^*$ that has the least-squares-error with it. If the least-squares-error is small and the correlation between the corresponding intensity windows is high, the selected column will take the place in $\mathbf{W}'_v$ to form $\mathcal{W}$. If not, the corresponding feature in $\mathbf{W}^*$ is regarded as having no correspondence in $\mathbf{W}'^*$ (this is possible as the feature may not be observable in both cameras) and ignored.

If the rank of some measurement matrices (i.e. $\mathbf{W}$, $\mathbf{W}'$, or $\mathcal{W}$) is less than 4, the procedure of inferring stereo matches may be a bit different. See Appendix B for details.

To increase the accuracy of inference process, an iterative approach is employed in this work. At the end of each iteration, the matrix B is updated using the newly acquired stereo matches through the SVD technique. The accuracy of matrix B is thus improved at each iteration, and so are the estimated values of $\mathbf{h}_{W'}$. This makes the inferred vectors be closer to the actual ones and inferring more stereo correspondences be possible.

An input-output description of the inference mechanism and an overview of the whole stereo-motion framework are summarized in Figures 3.2 and 3.3 respectively.

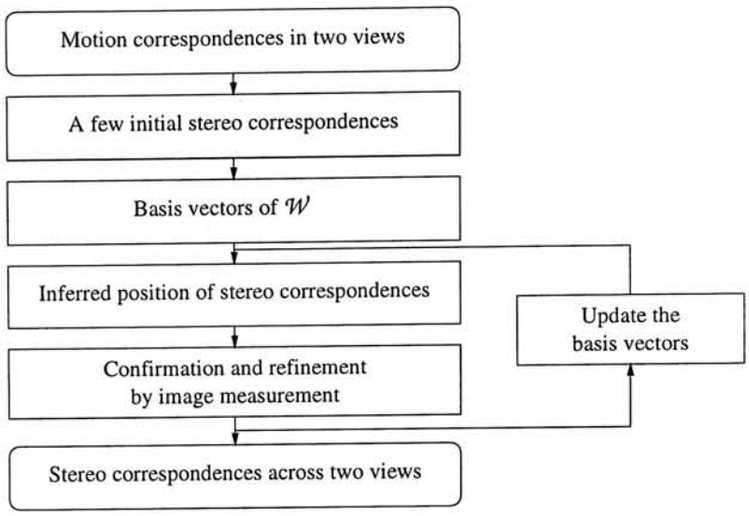


Figure 3.3 Overview of stereo-motion framework.

3.4 Determining 3D Structure from One Stereo Pair

Once stereo correspondences are established, the 3D positions of all the P tracked points can be determined via the triangulation geometry of any stereo pair in the image data. If $\mathbf{p}$ and $\mathbf{p}'$ are corresponding image positions in a stereo pair of images, then the 3D position of the corresponding scene point with respect to the camera coordinate frame of $\mathbf{p}'$ is:

$$\mathbf{P}' = \frac{(\bar{\mathbf{t}} \times \bar{\mathbf{R}} \begin{bmatrix} \mathbf{p} \\ f \end{bmatrix}) \cdot (\begin{bmatrix} \mathbf{p}' \\ f' \end{bmatrix} \times \bar{\mathbf{R}} \begin{bmatrix} \mathbf{p} \\ f \end{bmatrix})}{\| \begin{bmatrix} \mathbf{p}' \\ f' \end{bmatrix} \times \bar{\mathbf{R}} \begin{bmatrix} \mathbf{p} \\ f \end{bmatrix} \|^2} \begin{bmatrix} \mathbf{p}' \\ f' \end{bmatrix}$$

where f and f' are the focal lengths of the two cameras. Since the structure in $\mathbf{S}$ is referred from the last stereo frames, the last stereo pair is used for calculating the structure. The computed structure is accurate, as the triangulation is through a stereo system of long baseline. As evidenced by empirical results that are to be shown in Chapter 4, the reconstruction results are already accurate enough for many applications.

A more accurate 3D reconstruction method will be described in chapter 5, which uses all stereo pairs to optimize the results.

3.5 Computational Complexity of Inference Process

Here a simple analysis of the computational complexity that it takes to infer stereo correspondences from motion correspondences is given. Computations for extracting features from images and tracking the features along motion frames are not considered, as they are not the contribution of this work, and any feature extraction and motion correspondence systems in the literature are just the same to be used in this work.

Suppose P features are extracted and matched in the two image sequences, and suppose each image sequence is of F image frames. The system first uses the epipolar geometry to extract a few initial stereo matches, spending $O(P)$ time. Most likely a little over four initial stereo correspondences, say I of them, are obtained. The system then extracts the optimal basis of a 4D space (the column space of $\mathcal{W}$) from the initial stereo correspondences, by applying SVD to a $4F \times I$ matrix. The SVD process takes $O(FI^2)$ (if $I \leq 4F$) or $O(IF^2)$ (if $I > 4F$) time, which can be assumed constant and negligible since both I and F are most likely small. Once the basis is available, the system infers columns of $\mathcal{W}$ one by one using a linear algorithm. Since each column has $4F$ entries, it takes $O(F)$ time to infer a column. Since there are altogether P columns, it takes $O(PF)$ time to infer all columns of $\mathcal{W}$.

In total, it takes $O(P) + O(PF) = O(PF)$ time to infer stereo correspondences from motion correspondences. The inference process is therefore linear with respect to the total number of features in each image and the total number of image frames in each image sequence.

Chapter 4

Experimental Results

The 3D recovery framework proposed in chapter 3 has been implemented. The system has been tested with synthetic images and the performance of the inference mechanism is illustrated. It has also been tested with a variety of indoor and outdoor images to show the robustness of the algorithm.

4.1 Synthetic Images and Statistical Results

The purpose of the synthetic data experiment is to assure ground truth be available to check the validity of the proposed method. A scene with a synthetic sphere initially 4.5 m away from the stereo cameras was simulated. In the scene the stereo cameras were 60 cm apart, each with a focal length of 60 mm and a resolution of 500×500. The radius of the sphere was 15 cm long, with 450 dots randomly distributed on the spherical surface to represent the detectable features (Figure 4.1). In a duration of 15 image frames, the cameras moved along a seesaw-shape trajectory on a plane parallel to the image plane of the left camera at time zero (similar to the camera motion shown in Figure 6.2). The amplitude of the seesaw trajectory was 2 cm. The cameras translated 1 cm along the path over each frame, and rotated a small angle to keep the object always in sight.

The random dots on the sphere were imaged under full perspective projection, with Gaussian noise of zero mean and 1 pixel variance added to them. The dots were treated as feature points and tracked along the image streams.

Totally 206 points were successfully tracked in the two streams. The measurement matrices $\mathbf{W}^*$ and $\mathbf{W}'^*$ (size: 30×206) were then constructed. Under the epipolar constraint 36 unique matches were located across the two streams. The matrix basis was determined from these initial matches and other stereo correspondences were inferred and refined. Altogether 190 point correspondences were established in two iterations. The 3D coordinates of these points were then calculated using the triangulation geometry of the last stereo pair. The shape of the spherical surface was correctly recovered (Figure 4.2). The root-mean-square error of the recovered structure is about 5 mm in Euclidean distance, which is mainly due to additive Gaussian noise to the image positions.

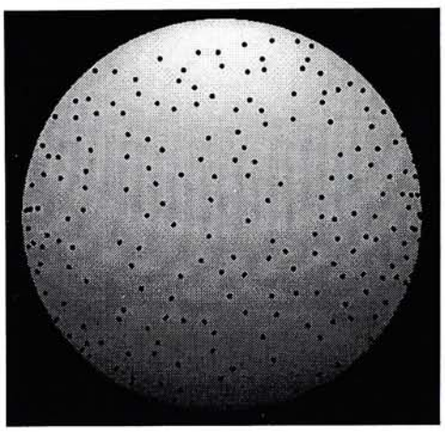
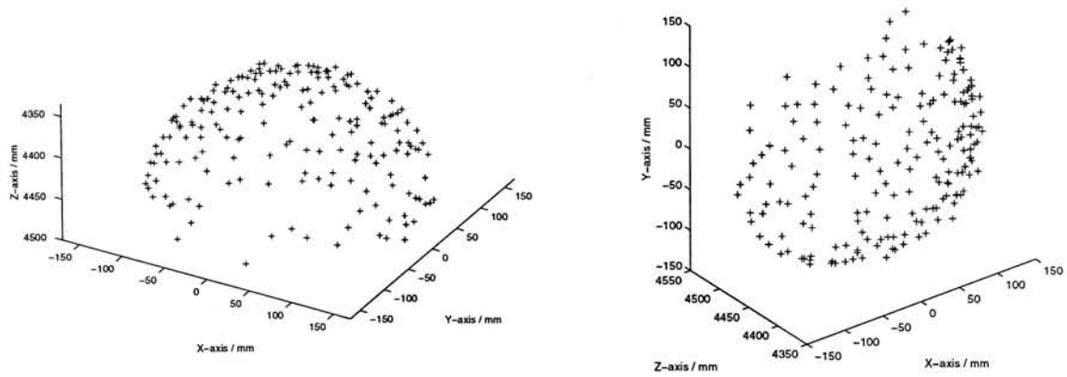


Figure 4.1 Synthetic image of a spherical surface (first image frame).



side view top view

Figure 4.2 3D structure recovered from synthetic image sequences (a spherical surface).

The $\mathbf{h}_{W'}$ estimated from Equation 3.2 is compared with its true value under different lengths of the image streams (Figure 4.3) The difference is measured in average Euclidean distance between the predicted and true image positions. The prediction error decreases as the total number of image frames increases, and is less than 5 pixels for image streams of more than 11 frames long. This shows that the extrapolation of stereo correspondences is satisfactory even with large measurement noise and relatively short image sequences.

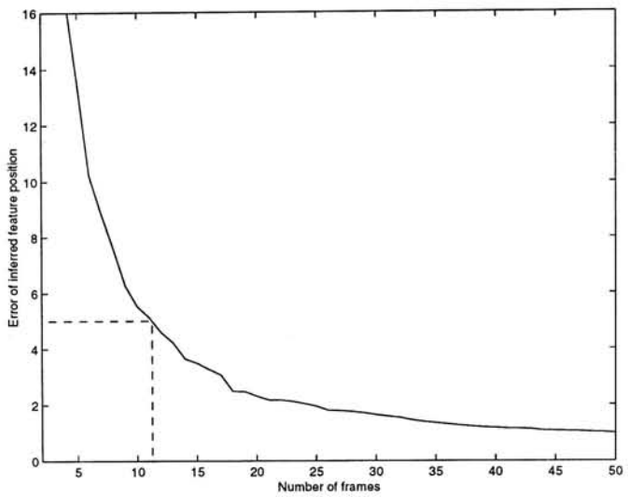


Figure 4.3 Deviation of inferred feature positions from their true values. The deviation is measured in average Euclidean distance between the predicted and true image positions.

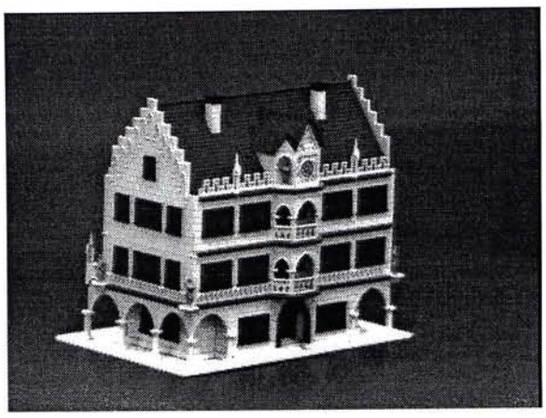
4.2 Real Image Sequences

The proposed method was also tested with various sets of indoor and outdoor image data. Three sets of image streams – *house model streams*, *oscilloscope and soda can streams*, and *bowl streams* were taken from a laboratory. The image streams of a *building* were taken outdoor with two hand-held camcorders.

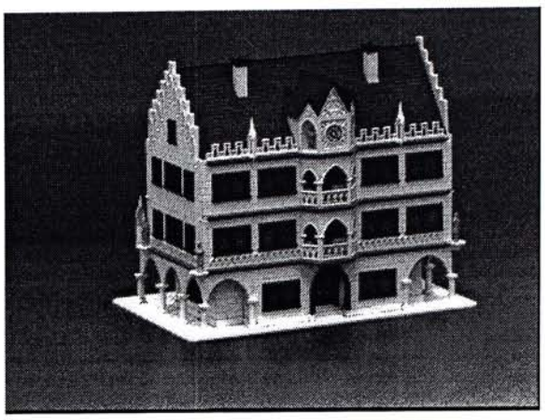
4.2.1 'House Model' Image Sequences

In one experiment, a house model was to be reconstructed. Figure 4.4 shows some images in the sequences. The images were taken from a laboratory using a stereo rig of two CCD cameras, both with 50 mm lens, which translated approximately 1 cm sideways and rotated a small angle over each image frame. The house model was about 2 m away from the cameras. The baseline of the cameras was about 58 cm. The camera optical axes were convergent, forming an angle of 17 degrees. The stereo cameras were first calibrated using the method described in [3]. Objects with orthogonal trihedral vertices were placed in front of the cameras and the relative geometry of the cameras was estimated off-linely from the image projections of the vertices. This method computes the stereo geometry from a set of non-linear equations, and can produce unique and accurate results.

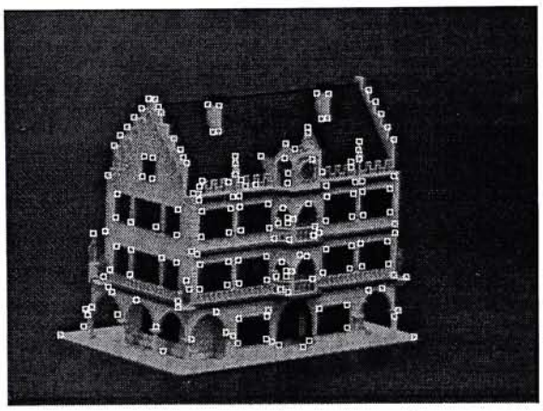
Each camera captured 9 images in the entire duration. Image features were selected and tracked individually in both sequences, using a publicly available feature tracker [21]. The feature tracker automatically selected and tracked 300 features throughout both image sequences. A total of 15 unique stereo matches were located using the epipolar constraint and a simple correlation-based matching algorithm, which evaluates match candidates by the mean square difference between the corresponding feature windows in two images. Other stereo correspondences of features were then predicted and refined. Altogether 167 features were successfully matched across the stereo image sequences in three iterations. The image positions of these matched features are overlaid on the last image pair in Figure 4.4. Figure 4.5 shows two different views of the computed 3D structure. It can be seen that the house model's shape was correctly recovered. Images of actual house model taken at the same orientation are shown in Figure 4.5 for comparison. The points on the frontal wall were recovered accurately, and all roughly lied on a plane. The position of the roof, the chimneys, the balcony, were all reconstructed correctly.



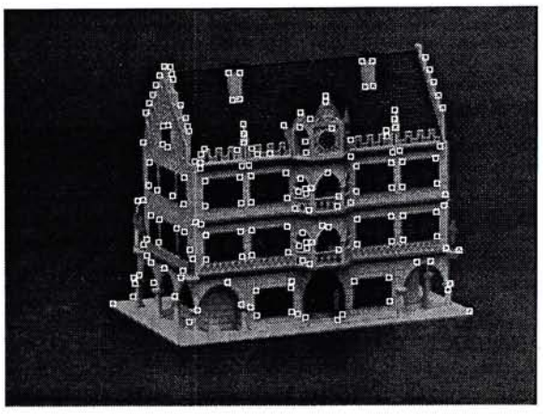
first image (camera 1)



first image (camera 2)



last image (camera 1)



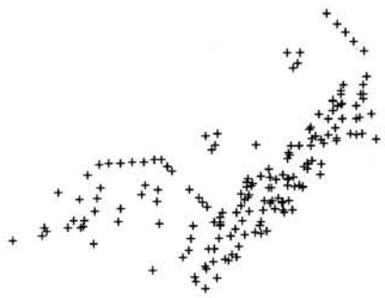
last image (camera 2)

Figure 4.4 The first and the last image pairs in the house model image sequences. Each sequence consists of 9 image pairs. The house model was about 2 m away from the cameras, which had a baseline of about 58 cm. The matched features are overlaid on the last image pairs, with the background darkened.

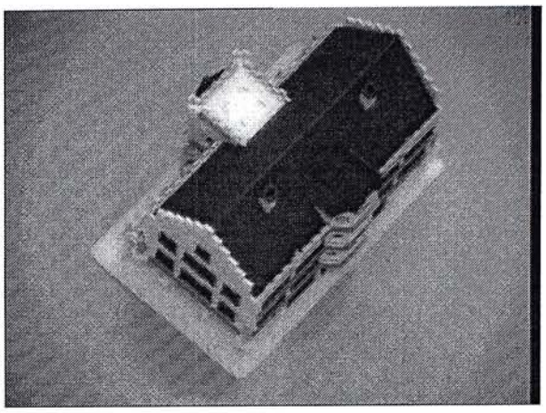
4.2.2 'Oscilloscope and Soda Can' Image Sequences

In another experiment, two disjointed objects, an oscilloscope and a soda can, were observed by the cameras (Figure 4.6). As the stereo-motion framework does not make any assumption on the object's shape, it is applicable to even disjointed objects, as long as the affine projection approximation is valid, i.e. the object's depth range is small compared with its average depth from the cameras.

The feature tracker established 300 motion correspondences in both sequences. A total of 19 initial stereo matches were found by the epipolar constraint. After three iterations, the system was able to infer 141 stereo correspondences. The overhead view of the reconstructed 3D structure is illustrated in Figure 4.7. The positions of the two objects are clearly separated. Points

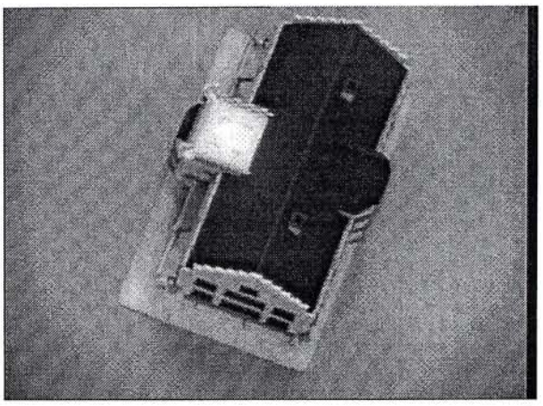


computed shape (first view)



actual shape (first view)

computed shape (second view)



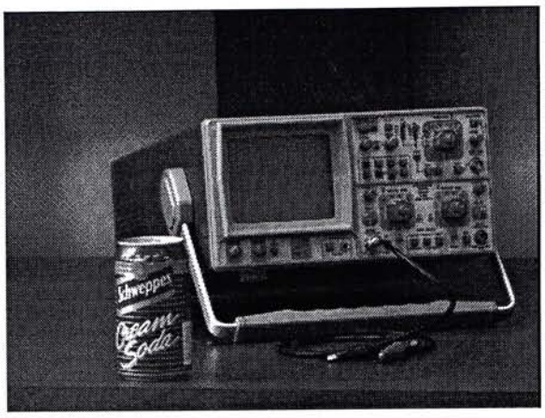
actual shape (second view)

Figure 4.5 Reconstruction results of the 'house model' images. Note that the frontal wall and the balcony was reconstructed accurately.

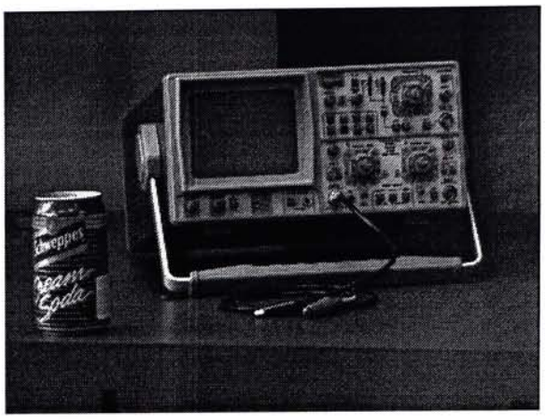
in the lower-left corner belong to the curved surface of the soda can. The five points in the center belong to the cable connected to the oscilloscope. Points in the upper-half belong to the frontal surface, a side-wall, and the stand of the oscilloscope. The points on the oscilloscope's frontal surface looks a bit scattered because the surface was inclined. No smoothing effect was introduced to points near the disjointed boundaries.

4.2.3 'Bowl' Image Sequences

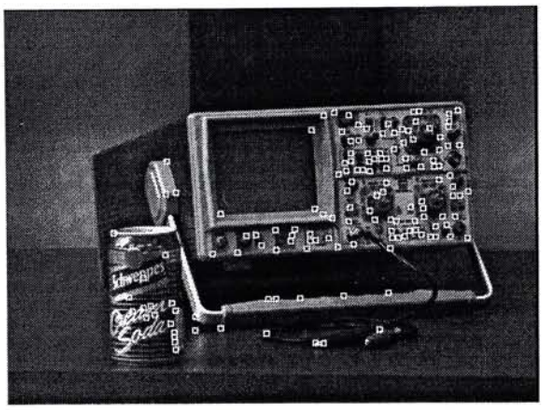
Figure 4.8 shows another set of image data. A set of 400 features were tracked in both sequences and 35 initial matches were established by the epipolar constraint. In total, 210 stereo correspondences were found in three iterations. Figure 4.9 shows two different views of the bowl surface reconstructed



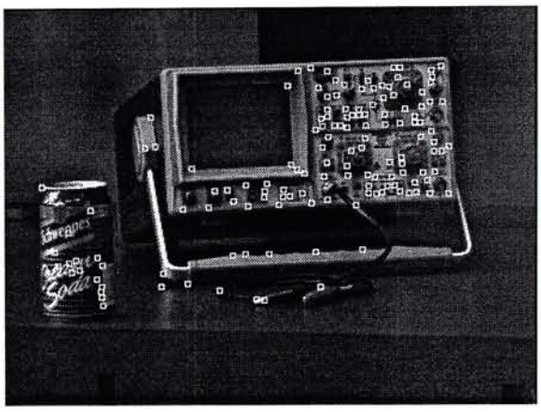
first image (camera 1)



first image (camera 2)



last image (camera 1)



last image (camera 2)

Figure 4.6 The first and the last image pairs in the 'oscilloscope and soda can' image sequences. Each sequence consists of 9 frames. The objects were about 4 m from the cameras. The separation between cameras was about 60 cm. The stereo-rig motion was similar to that in the house model experiment.

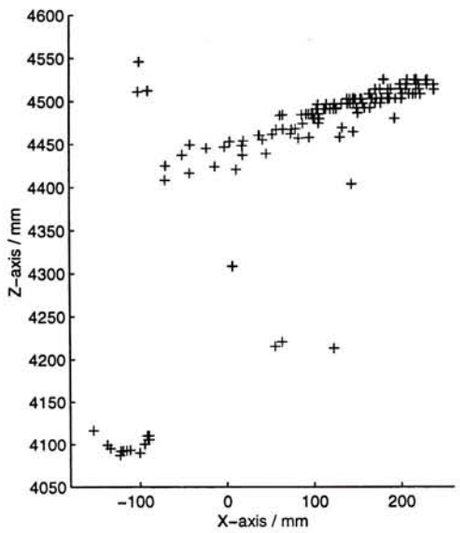


Figure 4.7 Top view of 3D structure recovered from the 'oscilloscope and soda can' image sequences.

from image date. To make reprojection possible, the object surface was approximated by a set of triangular patches, which were generated from the 3D object points by the Delaunay Triangulation method [20]. The pixel values of the last image pairs were then mapped onto the corresponding patches in the new image (details can be found in Appendix C). It can be seen that the reconstructed surface is smooth, and very little distortion is observed in the views. This shows that the 3D reconstruction is accurate, though relatively short image sequences were used in the experiment.

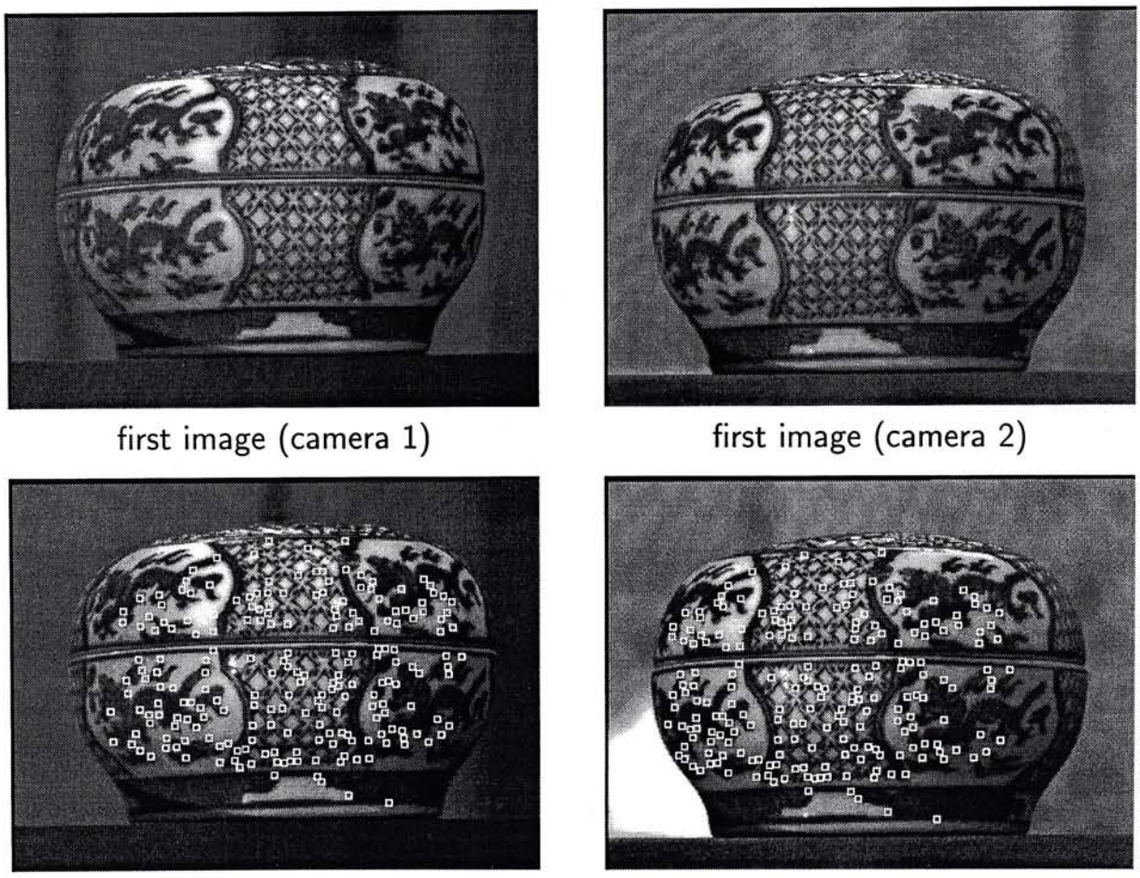
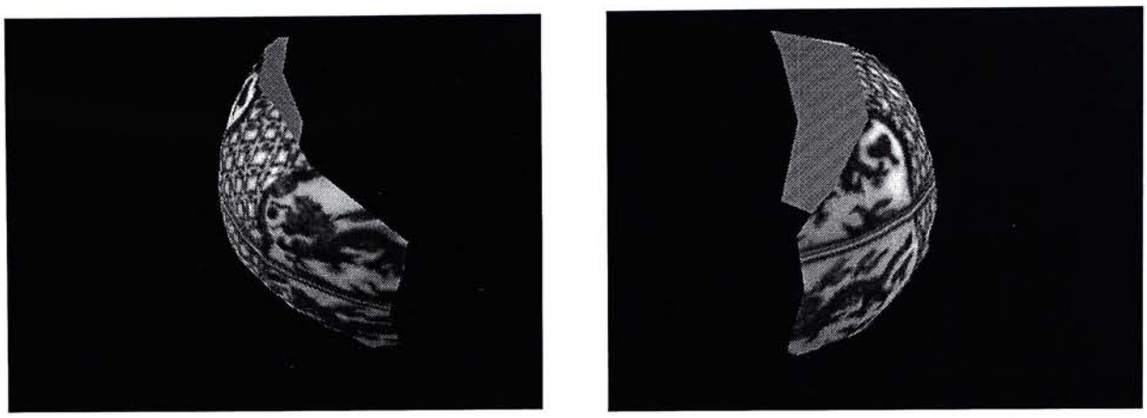


Figure 4.8 The first and the last image pairs in the 'bowl' image sequences, which consist of 9 pairs of images. The bowl was about 1.2 m from the cameras, which had a baseline of about 40 cm. The stereo-rig motion was similar to that of the house model sequences.



first view second view

Figure 4.9 Two different views of the reconstructed bowl surface. The bowl surface was approximated by a set of triangular patches. The pattern on the surface was then rendered from the last image pair.

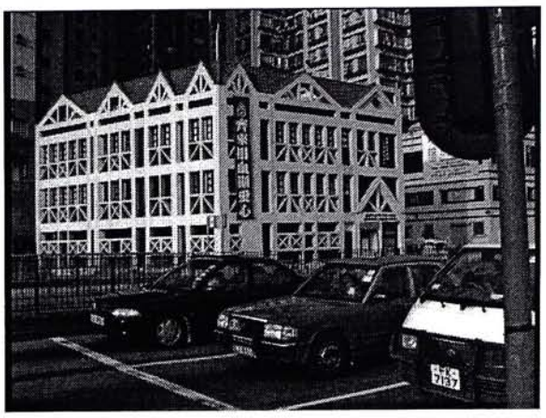
4.2.4 'Building' Image Sequences

Figure 4.10 shows the images of an outdoor scene. The image streams were taken with two hand-held digital camcorders mounted on a metal bar. A person holding the camcorders walked sideways slowly in front of a Red-Cross building. Images were extracted from the two video streams every 0.25 seconds. Twenty frames were obtained in both sequences in a duration of 5 seconds.

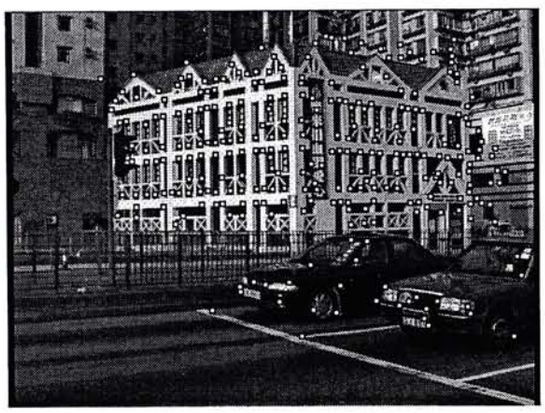
The camcorder motion was jerky, and image features moved as much as 52 pixels between successive images. The feature tracker produced many false motion correspondences in this case. A two passes method was employed to solve this problem. First, a few very distinct and non-repeative features were observed by the feature tracker. The tracker could established their motion correspondences correctly. The displacements of other features were then estimated from these feature points. Over 500 features were successfully tracked in the sequences.

A total of 49 initial matches were found. The system was able to infer 282 stereo correspondences in three iterations. Over 80% of them were on the Red-Cross building. The inference was accurate. The deviation between the inferred position and the true one was less than 1 pixel for most features on the building. The top-view of the recovered 3D structure is shown in Figure 4.11. It can be seen that the position and the size of the Red-Cross building and

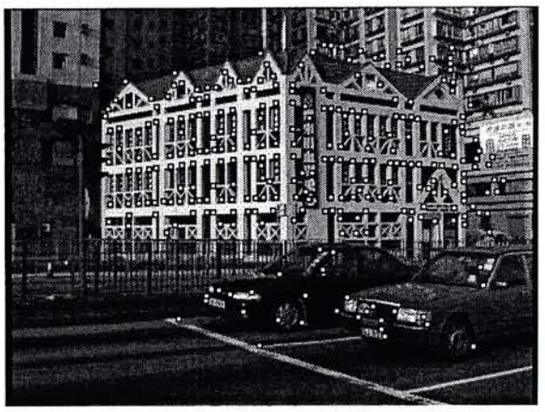
first image (camera 1)



first image (camera 2)



last image (camera 1)



last image (camera 2)

Figure 4.10 The first and the last image pairs in the outdoor image streams. The streams consist of 20 pairs of images, which were taken with two hand-held camcorders. The separation between the cameras was about 60 cm.

the cars are correctly recovered. In the reconstruction, the angle between the two major walls of the building is about 115 degree, which is consistent with the physical measurement (about 113 degrees). There are a number of other buildings in the background, which were all far away (about 50 - 60 m) from the cameras. Due to the limited resolution of cameras for such long distance, the computed 3D shape appears to be a bit noisy.

4.3 Computational Time of Experiments

Table 4.1 shows the computational time requirements in inferring stereo correspondences from motion correspondences when the inference mechanism was run on an UltraSPARC-II machine. It should be noted that the time

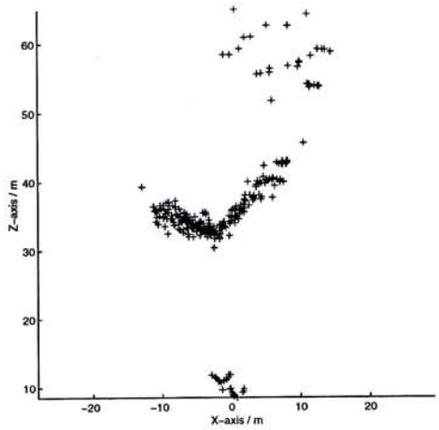


Figure 4.11 Top view of computed 3D structure from the outdoor images. The points at the bottom belong to the two cars and the road. The large cluster in the center belongs to the Red-Cross building. The points at the top belong to the buildings in the background.

for detecting features in the images and tracking them along the motion sequences is not included; those are not the focuses of this research, and any feature detection and motion tracking algorithm will fit just as well to the proposed stereo-motion framework. It can be seen that the computational time is roughly proportional to the number of features to infer when the number of image frames is fixed.

Experiment	No. of Stereo Pairs	No. of Point Corr. Established	No. of Iterations	Computational Time (sec)
House model	9	167	3	1.67
Oscilloscope & soda can	9	141	3	1.60
Bowl	9	210	3	2.32

Table 4.1 Computation time requirement of the Correspondence Inference Mechanism.

Chapter 5

Determining Motion and Structure from All Stereo Pairs

In Chapter 3, the object structure was calculated via the triangulation geometry of last image pair only. In fact, one can compute the structure from any stereo pair in the sequences. Let $\mathbf{P}_{fp}$ be the 3D position of the object point p ($p = 1, 2, \ldots, P$) as determined by stereo pair f ($f = 1, 2, \ldots, F$) and with respect to the reference camera coordinate frame there (the reference coordinate frame of a stereo pair is taken as the camera coordinate frame of the first camera or camera 1 in this work). Note that $\mathbf{P}_{fp}$'s for all P points and all F stereo pairs are obtainable by applying the triangulation process to the stereo pairs separately. However, $\mathbf{P}_{fp}$'s from different stereo pairs may not be consistent with one another because of the different image noise and disturbances. We are to integrate the information from all stereo pairs to determine a set of 3D positions $\mathbf{P}_p$'s (with respect to the last stereo pair, i.e., stereo pair F here) which best fit all the image data. Such $\mathbf{P}_p$'s are the optimal structure we want for the scene. However, to recover $\mathbf{P}_p$'s, we have to register the $\mathbf{P}_{fp}$'s from different stereo pairs by putting them under the same coordinate system for reference. This means we have to estimate the rigid transformations between the stereo pairs, which are actually the platform motion we also want.

5.1 Determining Motion and Structure

There are a number of ways to formulate the problem. One particular formulation is the following. As the scene is assumed stationary, camera motion can be determined by comparing the relative object position at different instant. The relationship between the structure and motion can represented by

$$\mathbf{S}_f = \underbrace{\left[\begin{array}{c|c} \mathbf{R}_f & \mathbf{t}_f \\ \hline 0\ 0\ 0 & 1 \end{array}\right]}_{\mathbf{M}_f} \mathbf{S}^*$$

where $\mathbf{S}_f$ is the 3D structure computed from the f-th stereo pair, $\mathbf{S}^*$ is the optimal object structure to determine (with respect to the last stereo pair), and $\mathbf{M}_f$ represents the rigid transformation between the f-th stereo pair and the last stereo pair. Again, note that $\mathbf{S}_f$'s for all stereo pairs are available once stereo correspondences are obtained. Seeing that $\mathbf{M}_f$'s are readily recoverable from the above equation if $\mathbf{S}^*$ is known, we adopt the following iterative scheme for estimating the platform motion $\mathbf{M}_f$'s and the optimal structure $\mathbf{S}^*$ in an alternate fashion.

Initially, the optimal structure $\mathbf{S}^*$ is set equal to $\mathbf{S}_F$. The points sets $\mathbf{S}_f$ and $\mathbf{S}^*$ are then translated so that their geometric centers are located at the origin of the coordinate system. Using a unit quaternion representation for $\mathbf{R}_f$, which assures the orthonormal property of the rotation matrix, we determine the quaternion by minimizing the squared distances between the two sets of points [12] using a least-squares method which has a closed-form solution [9] (see Appendix E for more information about the quaternion representation). The rotation matrix $\mathbf{R}_f$ is then reconstructed from the quaternion vector. Once $\mathbf{R}_f$ is determined, the translation vector $\mathbf{t}_f$ can be determined from

$$\mathbf{t}_f = \mathbf{C}_f - \mathbf{R}_f\mathbf{C}^*$$

where $\mathbf{C}^*$ and $\mathbf{C}_f$ are the centroids of $\mathbf{S}^*$ and $\mathbf{S}_f$ respectively. This way we can estimate $\mathbf{M}_f$ for all stereo pair f when the optimal structure $\mathbf{S}^*$ is assumed to be $\mathbf{S}_F$.

In the next iteration we first determine a more accurate estimate for $\mathbf{S}^*$ by combining all structures $\mathbf{S}_f$ from different stereo pairs. To do this, all $\mathbf{S}_f$ have to be referred from the same coordinate system, which can be done by pre-multiplying them with the corresponding motion matrix $\mathbf{M}_f$ just estimated. We refine the optimal structure as the one with the least-median-of-square (LMedS) error from all these sets of points [27], i.e.

$$\mathbf{S}^* = \mathbf{LMedS}\left(\mathbf{M}_1\mathbf{S}_1, \mathbf{M}_2\mathbf{S}_2, \ldots, \mathbf{M}_F\mathbf{S}_F\right)$$

The least-median-of-square estimator finds a structure that yields the smallest median of squared error computed for the entire data set. Unlike the least-square estimator, the LMedS approach is very robust towards outliers due to bad localization. However, it does not have a closed-form solution and must be obtained by an iterative method. With the refined value of $\mathbf{S}^*$, the motion matrices $\mathbf{M}_f$'s for all stereo pairs are once again estimated using the quaternion method described above.

The motion-and-structure-recovery procedure is repeated, having the optimal structure $\mathbf{S}^*$ and the motion matrices $\mathbf{M}_f$'s refined alternately, until the values of both $\mathbf{S}^*$ and $\mathbf{M}_f$'s are stable enough. In our experiments, typically less than three iterations are enough to generate accurate results. We conjecture that this is because the initial estimate of $\mathbf{S}^*$, $\mathbf{S}_F$, is already quite close to the true value already. An overview of the procedures is shown in Figure 5.1.

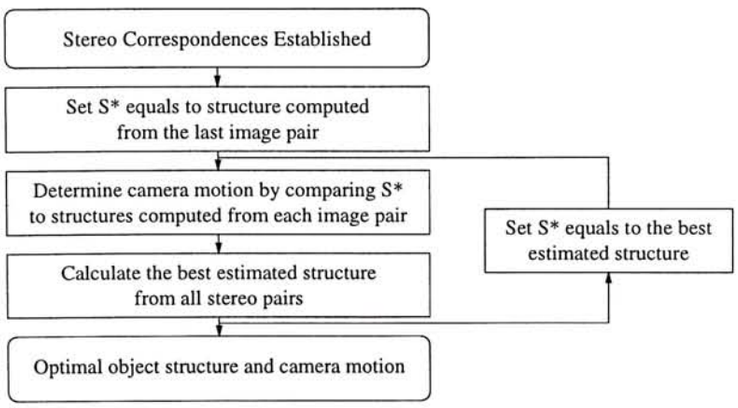


Figure 5.1 Procedure of recovering structure and motion from all images.

5.2 Identifying Incorrect Motion Correspondences

The above framework assumes that prior motion correspondences are correctly established in the two image streams. However, feature tracker does give incorrect results occasionally, especially when the image feature is not very distinct or when there are repetitive patterns in the neighborhood. If the motion correspondences over a feature are wrong, the inferred position of the stereo correspondence will also be incorrect, and the system may establish a false stereo correspondence for this feature. So features with faulty motion correspondences should be discarded.

With the use of all stereo pairs faulty motion correspondences in isolated image frames can be easily identified. If the motion correspondence over a feature in a particular image frame is faulty, the stereo correspondence in the corresponding stereo pair will not be consistent with the stereo correspondences over the same feature in the other stereo pairs. To put it more precisely, the 3D reconstructions from the different stereo pairs simply would not align, and the variance of the 3D positions would be large. In our system, we simply discard features which have large variance of 3D positions.

Chapter 6

More Experiments

Two methods are presented for computing 3D information from stereo correspondences. The first one (one-pair-method) uses one pair of images to calculate object structure, while the second one (all-pairs-method) uses all image pairs to determine both motion and structure. In this chapter, the latter method is tested with synthetic and real image streams. Then the results of the two methods are compared in terms of their accuracy and efficiency.

6.1 'Synthetic Cube' Images

The algorithm was tested with images of a synthetic cube (Figure 6.1). In the simulated scene, a cube of 20 cm high was initially 1.2 m away from the cameras. The cameras had a baseline of 46 cm. They moved along a seesaw-shape trajectory similar to the one in Section 4.1. The cameras translated 5 mm along the path over each frame, and in total 15 pairs of images were captured. The cube had 200 random dots on its three visible faces. Gaussian noise of zero mean and 0.5 pixel variance was added to the image position of feature points. These points were then quantized onto the 500×500 image planes.

A total of 181 stereo correspondences were established. The camera motion was then computed by comparing the 3D structure calculated from different stereo pair. The plot in Figure 6.2 compares the computed motion (solid line) with the actual value (dashed line). The arrows represent the orientation

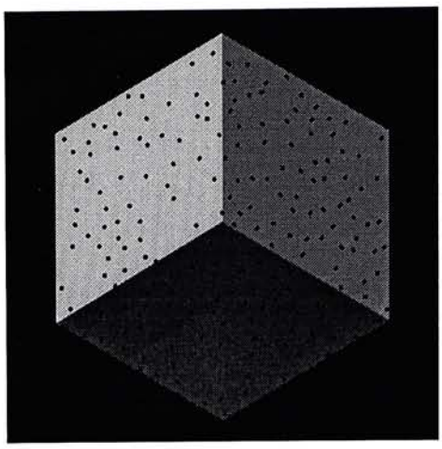


Figure 6.1 First image frame of synthetic data (a cube).

of camera optical axes. The recovered motion was quite close to the true motion. The computed structure is shown in Figure 6.3. The actual shape of cube is drawn in dashed line for comparison. Since the cameras baseline was relatively large, very accurate results were obtained. The root-mean-square distance error was only 0.2 mm.

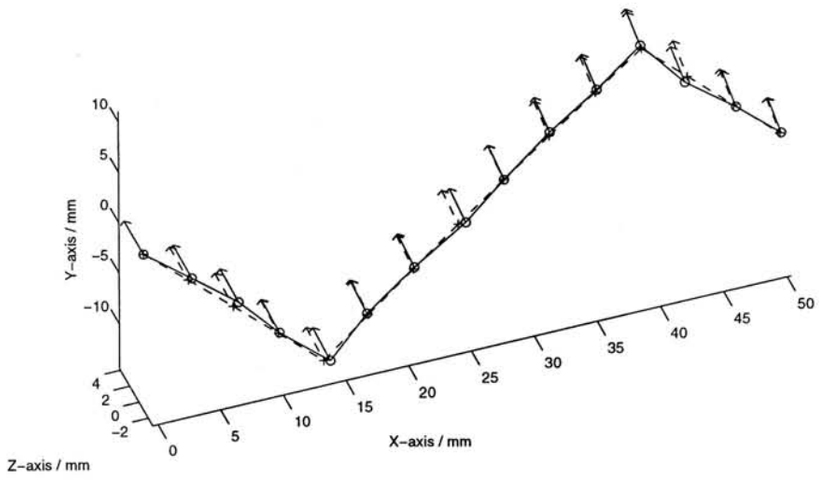


Figure 6.2 Motion recovered from the synthetic cube streams (solid line). The actual motion is shown in dashed line.

6.2 'Snack Bag' Image Sequences

In the next experiment, the algorithm was tested with image sequences of a snack bag (Figure 6.4). The images were taken with similar experimental setup

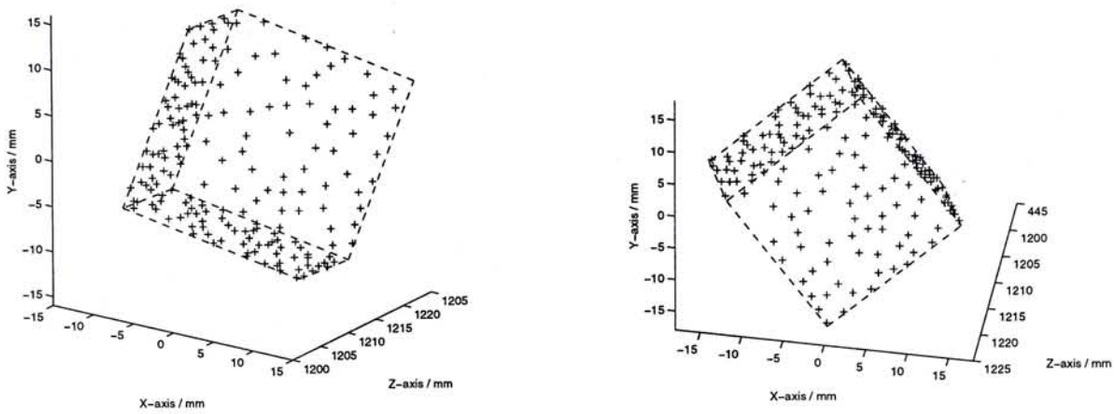
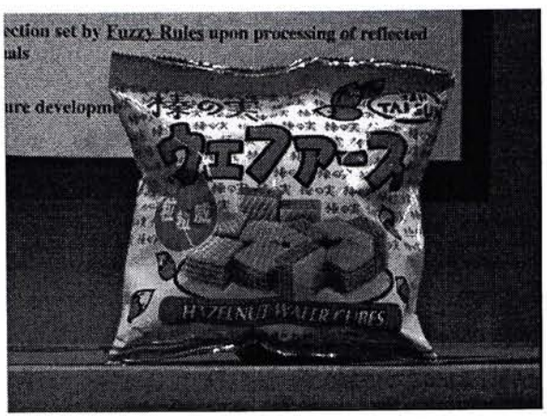


Figure 6.3 3D structure recovered from synthetic image sequences (a cube).

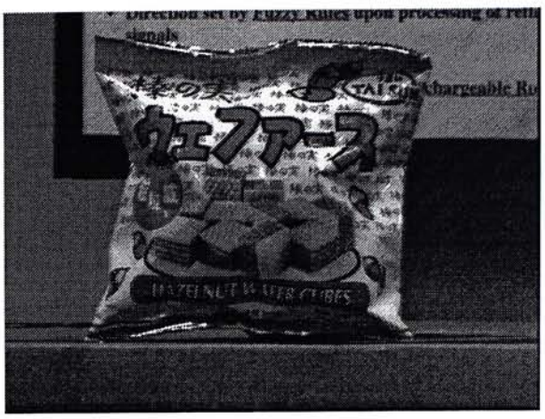
as the indoor image sequences in Section 4.2. The position and orientation of cameras with respect to the mobile platform was determined by moving the platform and observing the resulting motion of the cameras from a calibration object [22]. Then the cameras moved along a smooth trajectory in front of a snack bag, and captured 9 pairs of images in the entire duration. A total of 17 initial stereo matches were found from 450 motion correspondences, and the system established 116 stereo correspondences after three iterations.

The computed camera motion is shown in Figure 6.5. The direction and magnitude of the motion was roughly correct. However, the recovered motion was a bit noisy. The reason is that for small camera motion, effects of camera rotation and translation can be confused with each other, especially for distant objects. For example, a small rotation about the vertical axis and a small translation along the horizontal axis can generate very similar changes in an image.

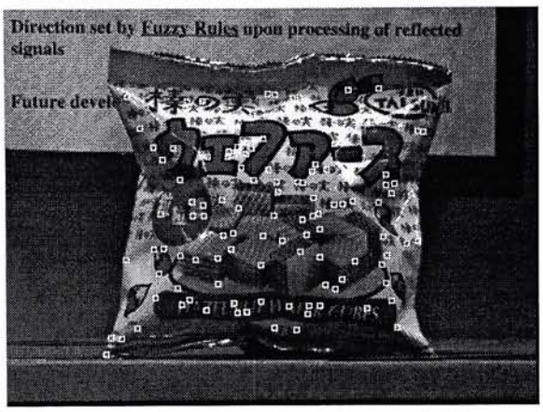
Several features were poorly tracked, and as a result the deviations of their computed 3D positions were rather large (Figure 6.6(a)). These features were discarded from the measurement matrix, because results obtained from these features would not be reliable. The object structure was then reconstructed from the remaining 107 features. The reconstruction was accurate, as can be seen in the reprojected images shown in Figure 6.7.



first image (camera 1)



first image (camera 2)



last image (camera 1)

last image (camera 2)

Figure 6.4 The first and the last image pairs in the 'snack bag' image sequences, which consist of 9 pairs of images. The bag was about 3 m from the cameras, which had a baseline of about 54 cm. The stereo-rig motion was similar to that of the house model sequences.

6.3 Comparison of the Two 3D Recovery Methods

To illustrate the improvement made by utilizing all stereo pairs, both one-pair-method and all-pairs-method were tested with the same synthetic image sequences used in Section 4.1. The 3D structure resulted from the two methods are shown in Figure 6.8. The two structures look quite alike and it is difficult to tell which one is more accurate. Figure 6.9 shows the shaded images of the spherical surfaces reconstructed from the two methods. The difference between the two results is obvious in these images. The surface recovered from the one-pair-method was rather rough; the error was due to the additive Gaussian noise in the image data. The all-pairs-method is more robust toward

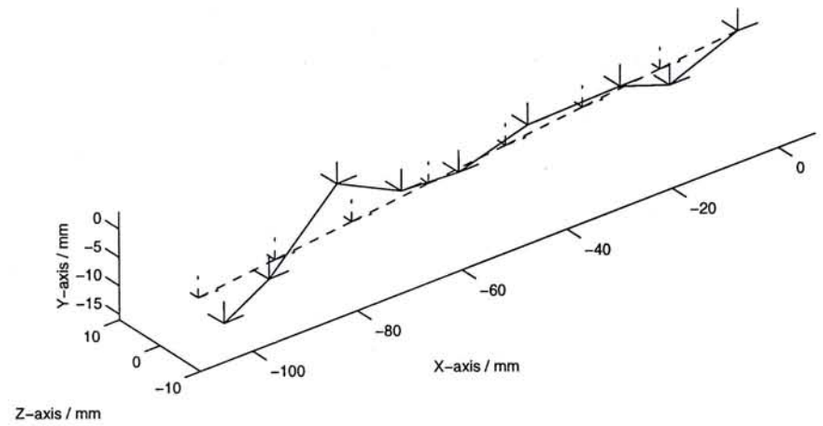


Figure 6.5 Motion recovered from the 'snack bag' image streams. The solid line represents the recovered motion, and the dashed one represents the true motion. The three orthogonal lines represents the position of the first camera's coordinate system.

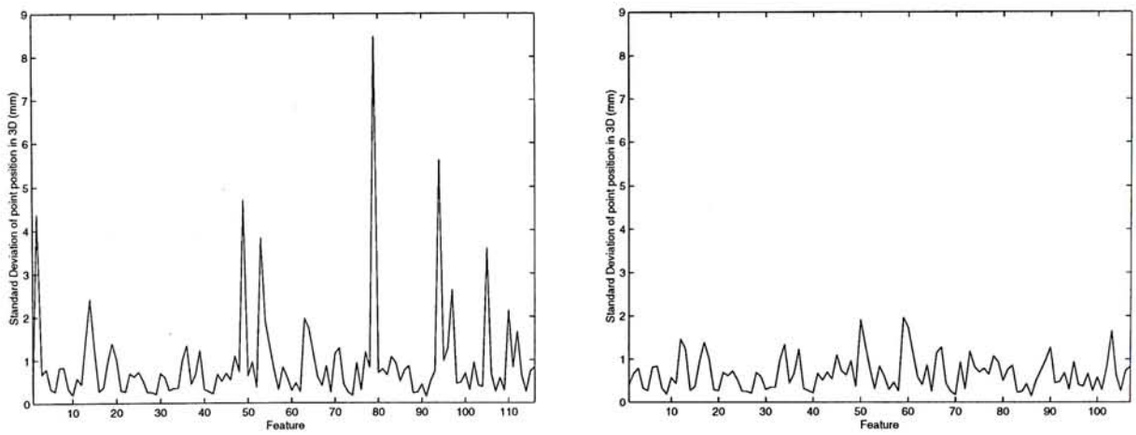
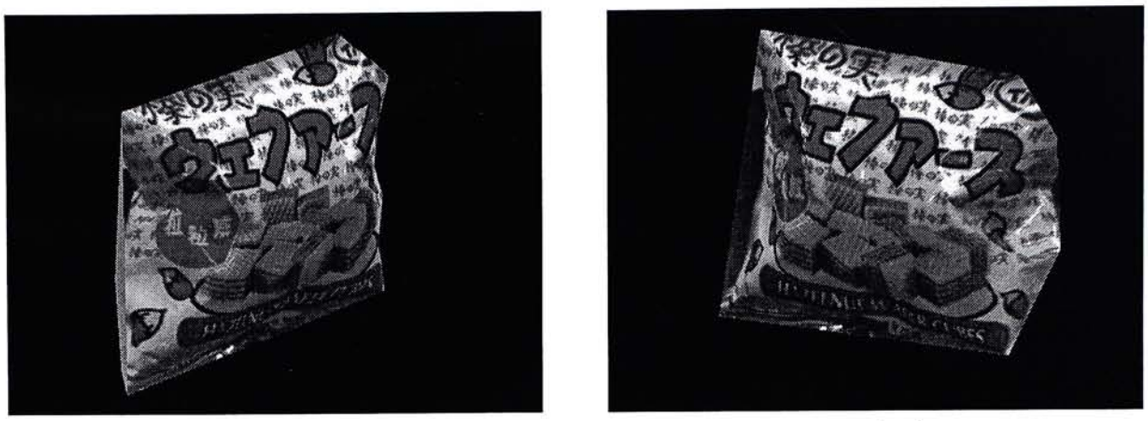


Figure 6.6 Standard deviation of 3D point position (a) before filtering, (b) after discarding unwanted features.

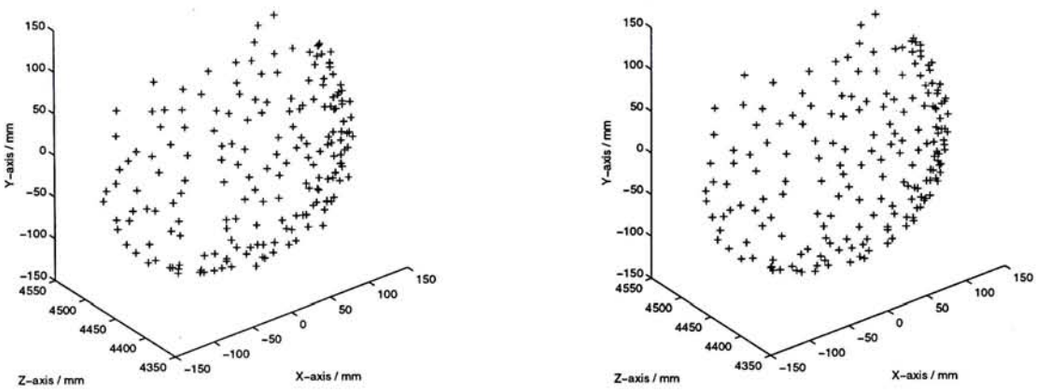
noise and produced a more accurate result. The root-mean-square error is 1.7 mm for the all-pairs-method, which is only one third of the error from the one-pair-method.

Although the all-pairs-method produces more accurate results, its computational time is much longer. For the 'snack bag' experiment in Section 6.2, the all-pairs-method spent 54 seconds to compute the results, while the one-pair-method only required 1.7 seconds. The reason is that the all-pairs-method



first view second view

Figure 6.7 Two different views of the reconstructed bag surface.



one-pair-method all-pairs-method

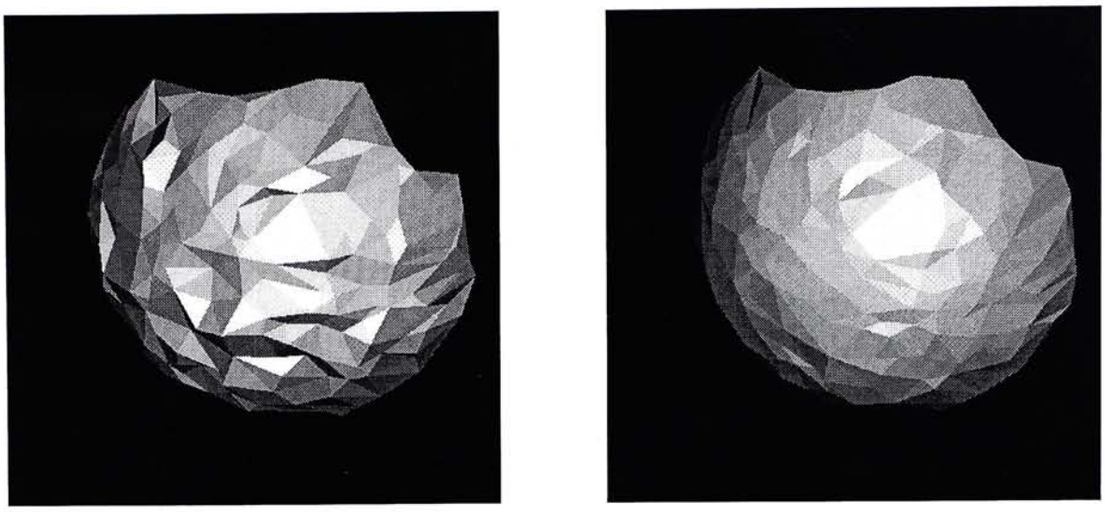
Figure 6.8 3D structure of synthetic sphere recovered from the two methods.

has to spend much time on calculating the quaternion vectors and the least-median-square-fit solution.

Experiment	No. of stereo pairs	No. of stereo correspondences	Computational time (one pair)	Computational time (all pairs)
House model	9	167	1.67 sec.	53.95 sec.
Oscilloscope & soda can	9	141	1.60 sec.	51.73 sec.
Bowl	9	210	2.32 sec.	69.45 sec.

Table 6.1 Computation time of the two structure recovery methods.

The one-pair-method is very efficient. It is suitable for applications that require real-time operation, like navigation and robotic manipulation. The all-pairs-method is less efficient, but it produces more accurate results. Besides,



one-pair-method all-pairs-method

Figure 6.9 Comparison of structures recovered by two different methods.

it is capable of locating features with faulty motion correspondences, which reduce the likelihood of establishing false matches. It is particularly suitable for applications which require high accuracy in the results and which allow off-line processing; an example application is scene reprojection.

Chapter 7

Conclusion

A framework of combining visual motion and stereo vision for reconstructing 3D structure has been described. The framework offers the advantages of both vision cues: simple correspondence, and accurate 3D reconstruction. It can compute 3D structure without the necessity of determining camera motion in the interim. The SVD technique is used to optimize the accuracy of the results in the presence of noise. The most important contribution of the work, among others, is a linear mechanism of inferring stereo correspondences directly from motion correspondences.

Experiments show that the framework does not require long image sequences in the input for reasonably accurate 3D reconstruction. Typically, say for objects of about 2 m away from the cameras, 9 image frames over 9 cm travel of each camera are needed. This is an advantage, as with shorter image sequences the assumption that the scene has no independent motion in the duration becomes more valid. The problem of occlusions and disocclusions as the cameras move is also less severe with shorter image sequences.

The affine camera model is used when stereo correspondences are predicted from motion correspondences. It is an adequate model for many practical cases when the scene is not too close to the cameras. Satisfactory output of the implemented system shows that the predicted stereo correspondences are reasonably close to the true positions. Nevertheless, a more accurate camera model would predict the stereo correspondences more accurately. Future work

will include how to replace the the affine camera model in the framework with the full perspective camera model.

Appendix A

Basic Concepts in Computer Vision

In this chapter, some background knowledge related to our stereo-motion framework is briefly described. Detailed description of these topics can be found in most computer vision textbooks.

A.1 Camera Projection Model

Several camera projection models are mentioned in this thesis. In this section, the projection models and the difference among them are described [10][27].

Full Perspective Projection Model

Figure A.1 shows the basic model for the projection of points in a scene onto an image plane. The line passing through an object point and it projection on image plane is called the line of sight. In full perspective projection model, the line of sight always passes through the camera center, i.e. the center of camera lens.

The 3D coordinates of an object point $[X, Y, Z]^T$ and its corresponding 2D image point $[u, v]^T$ are related by

$$u = -\frac{Xf}{Z} \qquad\qquad v = -\frac{Yf}{Z}$$

where f is the focal length of camera.

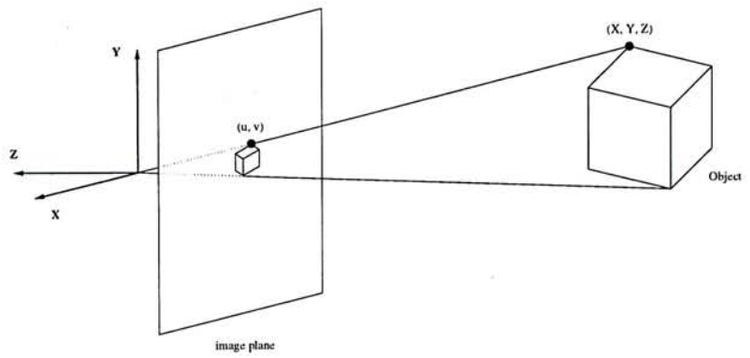


Figure A.1 Full perspective projection model

Throughout this thesis, a right-hand coordinate system is used, with z-axis pointing away from the scene. So the z-coordinates of object points are always negative.

Orthographic Projection Model

The full perspective projection model is a nonlinear mapping. This makes many vision problems difficult to solve. In case if the object distance is very large compared with its size, and the focal length is long enough, the image projection can be approximated as if $f \longrightarrow \infty$ and $z \longrightarrow \infty$. This projection model ignores completely the depth dimension. All lines of sight are parallel to z-axis. The equation of orthographic projection model is

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

This model is very simple, and is suitable for application that does not required the absolute size of object. However, it produces unreasonable results in some situatin. For example, the projection of two identical objects at different depth will be the same.

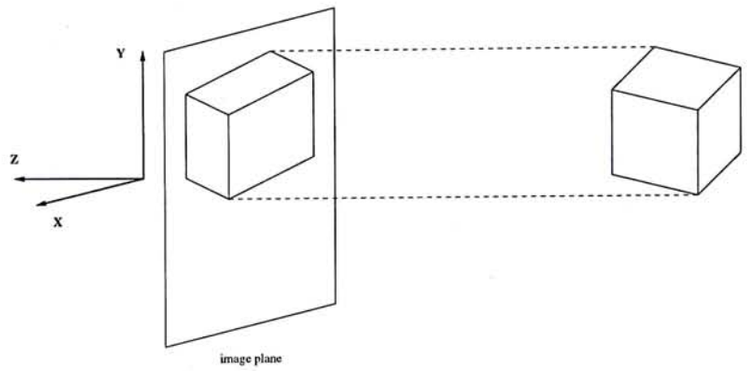


Figure A.2 Orthographic projection model

Weak Perspective Projection Model

When the object distance is large compared with its size, the nonlinear term $\frac{-f}{Z}$ in the perspective model can be replaced by a constant scaling factor s. Then the projection equation becomes linear

$$\begin{bmatrix} u \\ v \end{bmatrix} = s \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Usually s is set equal to $\frac{-f}{Z_c}$, where Z_c is the z-coordinate of object centroid. Unlike the orthographic model, this model takes the depth information of object into account and produces much better results. Physically, the weak perspective projection can be understood as a two-step projection. The first step is a parallel projection of object points onto a plane parallel to the image plane and passing through the object's centroid (average depth plane), where all projection lines are perpendicular to the average depth plane. The second step is a full perspective projection of that plane onto the image plane (Figure A.3).

Paraperspective Projection Model

The weak perspective projection model assumes the object is close to the optical axis. This causes a significant approximation error if the object is

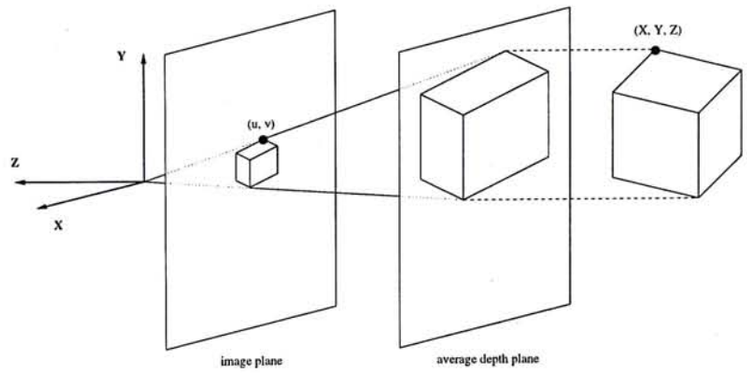


Figure A.3 Weak perspective projection model. The average depth plane is drawn in front for clarity.

distant from the optical axis. Paraperspective projection is similar to weak perspective projection, but it accounts for the position effect of object. The paraperspective projection involves 2 steps. In the first step, object points are projected onto the average depth plane, with all projection rays parallel to the line joining the object's centroid and camera center. Then the image on this plane is projected onto the image plane through full perspective projection. The equation for paraperspective projection is

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{-f}{Z_c} \begin{bmatrix} 1 & 0 & -\frac{X_c}{Z_c} & X_c \\ 0 & 1 & -\frac{Y_c}{Z_c} & Y_c \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

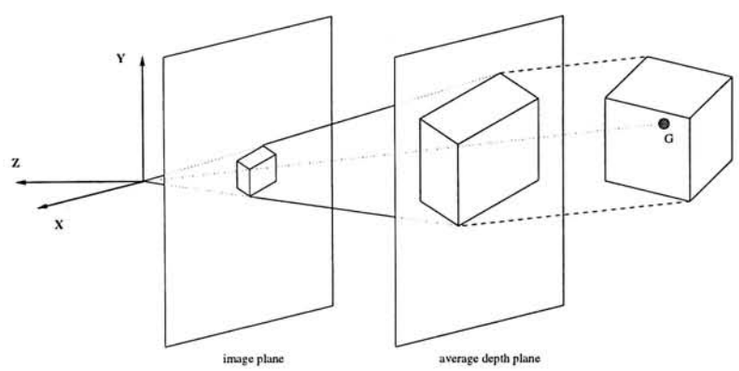


Figure A.4 Paraperspective projection model. The average depth plane is drawn in front for clarity.

Affine Projection Model

The projection matrices of all orthographic, weak perspective and paraperspective projection models can be written in form of a 2×4 matrix. The affine projection model is a generalization of all these projection models. Its projection matrix is a general 2×4 matrix. The equation for this model is

$$\begin{bmatrix} u \\ v \end{bmatrix} = J \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

where **J** is the affine projection matrix.

A.2 Epipolar Constraint in Stereo Vision

A binocular cameras system is shown in Figure A.5. Consider an image point p in the left image. The plane containing the two camera centers and point p is called the epipolar plane. The corresponding point in the scene must lie on the line passing through the point p and the camera center. The projection of all possible object point position on the right image is called the *epipolar line* (i.e., the intersection of epipolar plane and right image plane). The corresponding point of p of the right image is constrained to lie on the epipolar line [27].

Epipolar constraint is very useful in stereo matching, since it can reduce the 2D search space into a 1D space.

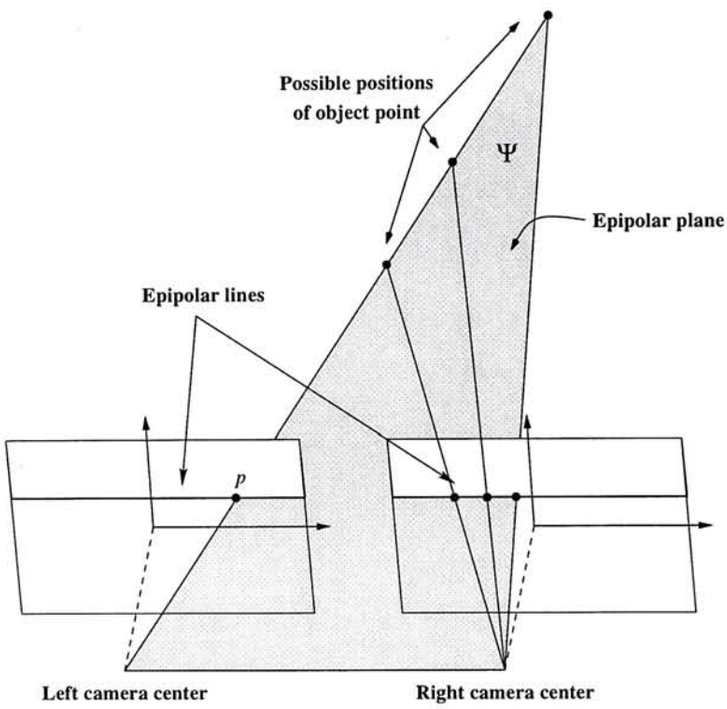


Figure A.5 Epipolar geometry of stereo camera system. Note that in general the epipolar lines may not be parallel to the x-axis.

Appendix B

Inferring Stereo Correspondences with Matrices of Rank < 4

In Section 3.3, an equation was derived for inferring stereo correspondences from motion correspondences (Equation 3.3). This equation was derived under the assumption that all measurement matrices are of rank 4. However, the rank of measurement matrices may be less than four in some cases. For example, a measurement matrix is rank 3 when all points in the scene lie on a plane. The reason is that the coordinates of any object point (X_p, Y_p, Z_p) is a linear combination of two vectors (basis of the 2D space), and so a structure matrix in homogeneous form would be rank 3. Thus the measurement matrix is also rank 3. Another example is when the camera motion is purely rotational and the object is close to the optical axis, the rank of measurement matrix would be close to 3, i.e. the fourth singular value of matrix is close to zero. In these cases, the correspondence-inference-mechanism may not work properly, since one measurement matrix may not contain enough information to infer the elements in the other matrix.

In our approach, $\mathbf{W}'$ is estimated from $\mathbf{W}$ through Equation 3.3. If rank($\mathbf{W}$) $\geq$ rank($\mathbf{W}'$), $\mathbf{W}'$ can still be estimated by the equation. However, the equation does not work if rank($\mathbf{W}$) is less than rank($\mathbf{W}'$). For example, if stereo rig rotates about the first camera's center, then the first camera's motion will be purely rotational and the second camera's one will be both translational and rotational. So the matrices $\mathbf{W}$ and $\mathbf{W}'$ will be rank 3 and 4 respectively. In this case, Equation 3.3 cannot correctly infer the stereo correspondences from

$\mathbf{W}$, because $\mathbf{W}$ does not contain sufficient information to estimate the value of $\mathbf{W}'$. However, this problem can be easily solved by swapping the role of $\mathbf{W}$ and $\mathbf{W}'$ in the inference mechanism, i.e. estimating $\mathbf{W}$ from matrix $\mathbf{W}'$. So, if the rank of $\mathbf{W}$ is smaller than that of $\mathbf{W}'$, then the inference mechanism should be applied using the following equation

$$\mathbf{W} = \mathbf{B}_W \left(\mathbf{B}_{W'}^{\mathsf{T}} \mathbf{B}_{W'}\right)^{-1} \mathbf{B}_{W'}^{\mathsf{T}} \mathbf{W'} \tag{B.1}$$

Appendix C

Generating Image Reprojection

Figure C.1 illustrates the procedure of generating an image reprojection. The coordinates of some object points are first computed. Then a 2D triangular mesh is generated from image points using the Delaunay triangulation method [20]. Each triangle in the mesh corresponds to a triangular patch in 3D. These triangular patches together form a continuous surface.

This surface is a linear approximation of the object's surface. It can closely approximate the object's surface if dense object points are available. The pattern on the surface is then rendered onto the reprojecting image. It can be done by mapping the pixel values in the original image onto the resulting surface.

The results of this method is satisfactory, provided that the object's surface is smooth and continuous. The performance of this rendering method can be improved if non-planar triangular patches are used.

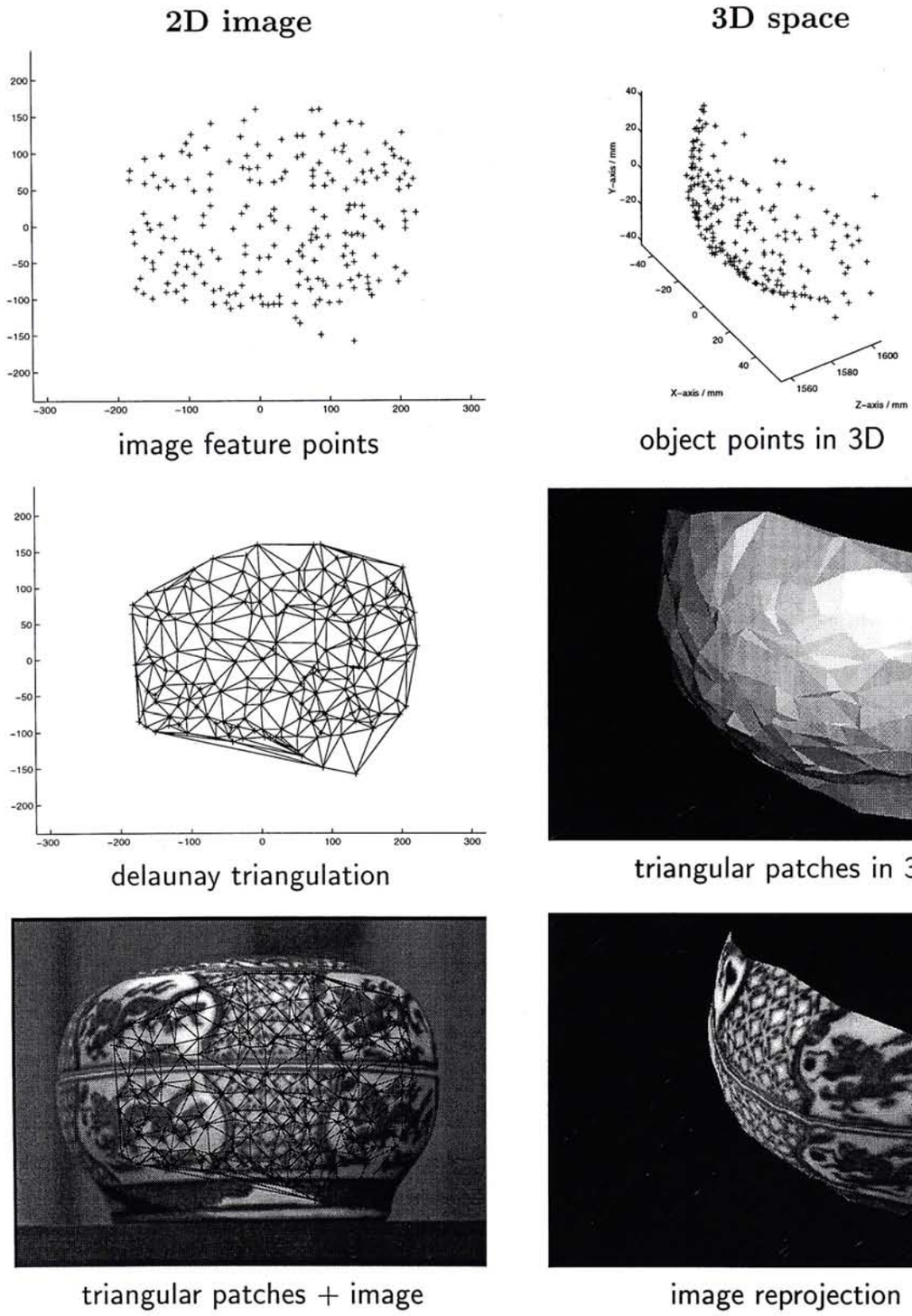


Figure C.1 Generating image reprojection

Appendix D

Singular Value Decomposition

By the singular value decomposition [5][23], any m by n matrix $\mathbf{A}$ can be factored into

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U}$ (m by m) and $\mathbf{V}$ (n by n) are orthogonal matrices and $\mathbf{D}$ is a m by n matrix. The columns of $\mathbf{U}$ are eigenvectors of $\mathbf{A}\mathbf{A}^T$, and the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$. The diagonal elements in $\mathbf{D}$ are the singular values of $\mathbf{A}$ (sorted in descending order), which are equal to the square roots of nonzero eigenvalues of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. The remaining elements in $\mathbf{D}$ are all equal to zero.

The columns of $\mathbf{U}$ and $\mathbf{V}$ give orthonormal bases for the fundamental subspaces. For example, if $\mathbf{A}$ is a rank r matrix, then the first r columns of $\mathbf{U}$ are its column space bases, and the first r columns of $\mathbf{V}$ are bases of its row space. The remaining columns in $\mathbf{U}$ and $\mathbf{V}$ are the left nullspace and nullspace of $\mathbf{A}$ respectively.

Let $a = min(m, n)$. The SVD equation can also be written as

$$\mathbf{A} = \sum_{i=1}^{a} \sigma_i u_i v_i^T \tag{D.1}$$

where u_i and v_i are the i-th column in $\mathbf{U}$ and $\mathbf{V}$ respectively, and σ_i is the i-th singular value in $\mathbf{D}$. If matrix $\mathbf{A}$ is of rank r, the $(r+1) - th$ to a-th singular values are all equal to zero. Then $\mathbf{A}$ is equal to the summation of the first r-th terms in Equation D.1.

Suppose $\mathbf{A}$ is a rank r matrix corrupted by noise. Then $\mathbf{A}$ may become full rank and all singular values are non-zero. The best estimate of matrix $\mathbf{A}$,

under least-square-error criterion, is equal to summation of the first r-th terms in Equation D.1.

Appendix E

Quaternion

A quaternion [2][9][19] is a quadruple of ordered real numbers, s, a, b, c, associated, respectively, with four units: the real number $+1$, and three other units $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, having cyclical permutation:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1 \qquad \begin{array}{lll} \mathbf{ij} = \mathbf{k} & \mathbf{jk} = \mathbf{i} & \mathbf{ki} = \mathbf{j} \\ \mathbf{ij} = -\mathbf{k} & \mathbf{jk} = -\mathbf{i} & \mathbf{ki} = -\mathbf{j} \end{array}$$

A quaternion Q is written as

$$Q = s + a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$$

The operation on quaternion is different from ordinary vector. Some basic properties of quaternion algebra are:

Scalar part of Q	:	s
Vector part of Q, v	:	$a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$
Conjugate of Q	:	$s - (a\mathbf{i} + b\mathbf{j} + c\mathbf{k})$
Norm of Q	:	$s^2 + a^2 + b^2 + c^2$
Reciprocal of Q	:	$(s - a\mathbf{i} + b\mathbf{j} + c\mathbf{k})/(s^2 + a^2 + b^2 + c^2)$
Addition of Q_1 and Q_2	:	$(s_1 - s_2) + (v_1 - v_2)$
Multiplication of Q_1 and Q_2	:	$s_1 s_2 - v_1 \cdot v_2 + s_2 \cdot v_1 + s_1 \cdot v_2 + v_1 \times v_2$

A unit quaternion vector can represent rotation in space. A rotation of angle α about an axis $\mathbf{n}$ can be written as

$$Q = \mathbf{Rot}(\mathbf{n}, \alpha) = \left[\cos\left(\frac{\alpha}{2}\right) + \sin\left(\frac{\alpha}{2}\right)\mathbf{n}\right]$$

A rotation about an arbitrary axis can also be represented by quaternion. For example, a rotation of 120^o about an axis equally inclined to the $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ axes is equivalent to a rotation of 90^o about $\mathbf{k}$ followed by a rotation of 90^o about $\mathbf{j}$. The corresponding quaternion representation is $\mathbf{Rot}(\mathbf{j}, 90)\mathbf{Rot}(\mathbf{k}, 90)$.

With quaternionic parameterization, a rotation matrix $\mathbf{R}$ can be written in terms of s, a, b, c

$$\mathbf{R} = \begin{bmatrix} s^2 + a^2 - b^2 - c^2 & 2(ab - sc) & 2(ac + sb) \\ 2(ab + sc) & s^2 - a^2 + b^2 - c^2 & 2(bc - sa) \\ 2(ac - sb) & 2(bc + sa) & s^2 - a^2 - b^2 + c^2 \end{bmatrix}$$

The orthonormal property of rotation matrix can be ensured as long as the norm of quaternion vector is equal to one.

Reference List

[1] P. Balasubramanyam and M. A. Snyder. The P-Field: A Computational Model for Binocular Motion Processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 115–120, Maui, Hawaii, June 1991.

[2] O. Bottema and B. Roth. *Theoretical Kinematics.* North-Holland Pub. Co., New York, 1979.

[3] R. Chung and S.-k. Wong. Stereo Calibration from Correspondences of OTV projections. *IEE Proceedings: Vision, Image and Signal Processing*, 142(5):289–296, October 1995.

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

[5] K. I. Diamantaras and S. Y. Kung. *Principal component neural networks: theory and applications.* Wiley, New York, 1996.

[6] O. Faugeras. Stratification of three-dimensional vision: projective, affine, and metric representations. *Journal of the Optical Society of America - A*, 12(3):465–484, March 1995.

[7] A. Ho and T. Pong. Cooperative Fusion of Stereo and Motion. *Pattern Recognition*, January 1996.

[8] P.K. Ho and R. Chung. Stereo-Motion that complements Stereo and Motion Analyses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 213–218, Puerto Rico, June 1997.

[9] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4(4):629–642, 1987.

[10] R. Jain, R. Kasturi, and B. G. Schunck. *Machine Vision.* McGRAW-HILL, Inc., New York, 1995.

[11] G. A. Jones. Constraint, Optimization, and Hierarchy: Reviewing Stereoscopic Correspondence of Complex Features. *Computer Vision and Image Understanding*, 65(1):57–78, January 1997.

[12] Simon K. Kearsley. An algorithm for the simultaneous superposition of a structural series. *Journal of Computational Chemistry*, 11(10):1187–1192, 1990.

[13] T. J. Keating, P. R. Wolf, and F. L. Scarpace. An Improved Method of Digital Image Correlation. *Photogrammetric Engineering and Remote Sensing*, 41(8):993–1002, August 1975.

[14] Stephen Maybank. *Theory of reconstruction from image motion.* Springer series in information sciences 28. Springer-Verlag, Berlin, 1993.

[15] A. Mitiche. A Computational Approach to the Fusion of Stereo and Kineopsis. In W. N. Martin and J. K. Aggarwal, editors, *Motion Understanding: Robot and Human Vision*, pages 81–95. Kluwer Academic Publishers, 1988.

[16] T. Morita and T. Kanade. A sequential factorization method for recovering shape and motion from image streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):858–867, 1997.

[17] M. Okutomi and T. Kanade. A Multiple-baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:353–363, 1993.

[18] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):137–154, 1997.

[19] Fu K. S., Gonzalez R. C., and C. S. G. Lee. *Robotics: control, sensing, vision and intelligence.* McGRAW-HILL, Inc., New York, 1987.

[20] J. R. Shewchuk. Triangle: Engineering a 2d quality mesh generator and delaunay triangulator. In *First Workshop on Applied Computational Geometry*, pages 124–133, Philadelphia, Pennsylvania, May 1996.

[21] J. Shi and C. Tomasi. Good Features to Track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, Seattle, Washington, June 1994.

[22] Y. C. Shiu and S. Ahmad. Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form ax=xb. *IEEE Transactions on Robotics and Automation*, 5(1):16–29, 1989.

[23] G. Strang. *Linear Algebra and Its Applications, 2nd Ed.* Academic Press, New York, 1980.

[24] C. Tomasi and T. Kanade. Shape and Motion from image streams under orthography: A Factorization Method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[25] A. M. Waxman and J. H. Duncan. Binocular Image Flows: Steps toward stereo-motion fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:715–729, November 1986.

[26] J. Weng, T. S. Huang, and N. Ahuja. *Motion and structure from image sequences.* Springer series in information sciences 29. Springer-Verlag, Berlin, 1993.

[27] Gang Xu and Zhengyou Zhang. *Epipolar Geometry in Stereo, Motion and Object Recognition : a unified approach.* Kluwer Academic Publishers, Boston, 1996.

[28] Z. Zhang and O. D. Faugeras. Three-Dimensional Motion Computation and Object Segmentation in a Long Sequence of Stereo Frames. *International Journal of Computer Vision*, 7(3):211–241, 1992.

[29] Z. Zhang, O. D. Faugeras, and N. Ayache. Analysis of a sequence of stereo scenes containing multiple moving objects using ridigity constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, Tampa, Florida, 1988.