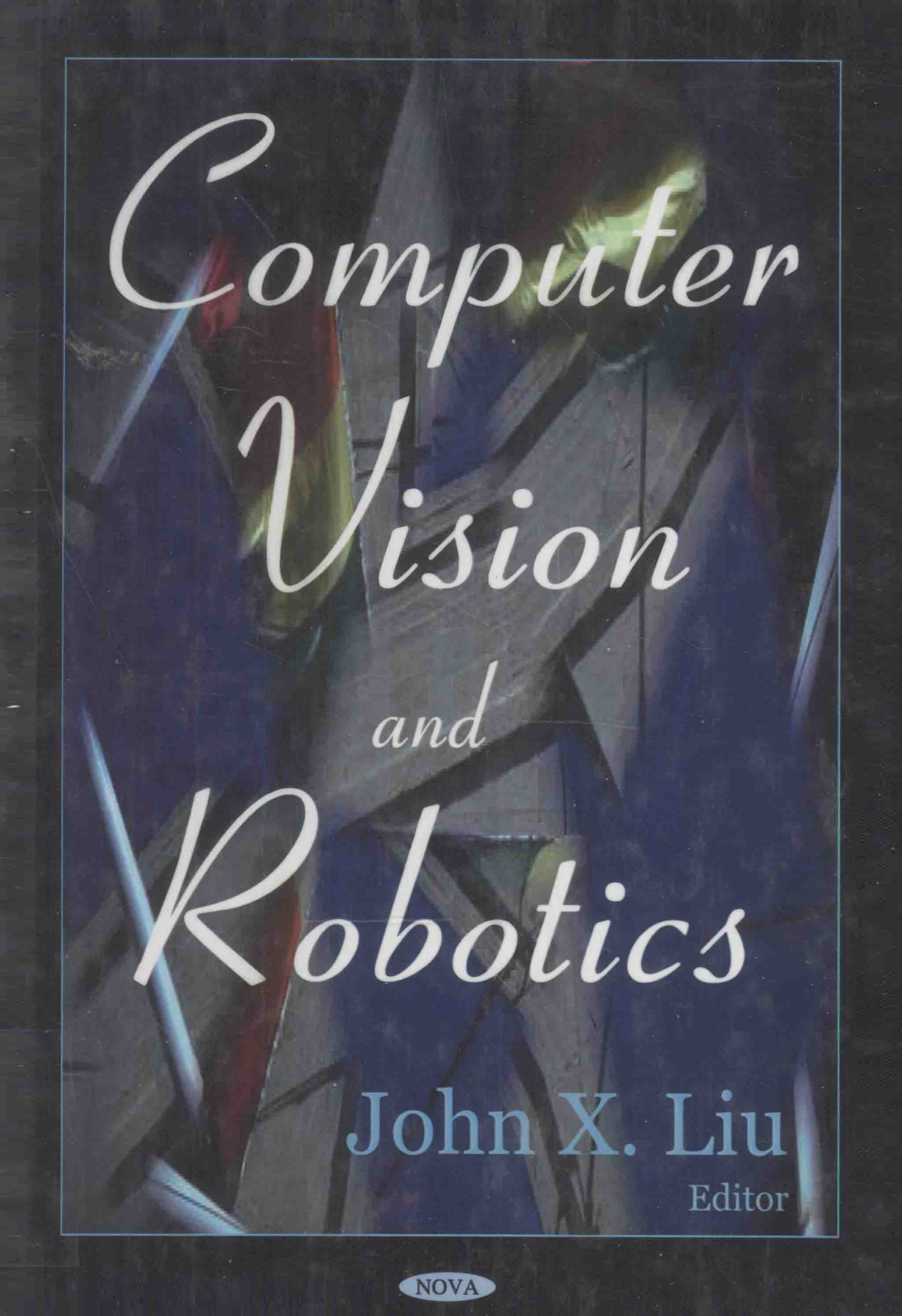
Computer
Vision
and
Robotics
John X. Liu
Editor
NOVA

# COMPUTER VISION AND ROBOTICS

JOHN K. LIU
EDITOR



**Nova Science Publishers, Inc.**
*New York*

**LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA**
Computer vision and robotics / John X. Liu (editor).
p. cm.
Includes bibliographical references and index.
ISBN 1-59454-357-7
1. Robotics. 2. Computer vision. I. Liu, John X.
TJ211.C621854 2005
629.8'92637--dc22 2005011341

*Published by Nova Science Publishers, Inc. + New York*

# COMPUTER VISION AND ROBOTICS

# PREFACE

This new book deals with control and learning in robotic systems and computers.

In Chapter 1, the authors discuss about the stereo vision and visual motion are two vision cues that allow three-dimensional (3D) information of a scene to be recovered from multiple images. When a mobile platform with two fixed camera heads is available to capture stereo pair of image streams, both cues are applicable. Yet, the cues have complementary advantages: while feature correspondence is simpler in visual motion, stereo vision offers more accurate 3D reconstruction. This paper presents an approach of integrating the two cues, that retains their advantages and removes their disadvantages. It is shown that by adopting the affine camera model for the projection model of the video cameras, the two sets of motion correspondences (on the two cameras) are actually related to the stereo correspondences (across the cameras) by a matrix rank property. The rank property is important, as it allows the inference from the more readily available motion correspondences to stereo correspondences that give more accurate 3D reconstruction. In addition, the inference process could be achieved in a time only linear with respect to the total size of the image data. With the inferred stereo correspondence, both the 3D structure of the scene as well as the motion of the mobile platform could be recovered. It is also shown that with the use of all stereo pairs of image data, not only could reconstruction accuracy be boosted, even errors in the initial motion correspondences could be detected. Experiments on real image data show that 3D reconstruction is accurate even with relatively short motion of the mobile platform.

Chapter 2 presents a new method for progressive transmission of 3D images that has four components: (1) decomposition of the image into regions using Singular Value Decomposition (SVD), (2) a reconstruction algorithm for progressive rendering that uses matrix polynomial interpolation along with approximations which are derived from SVD, (3) exploitation of a matrix norm for analyzing goodness of approximation, and (4) an optimal adaptive strategy for selecting "the next region to transmit".

SVD of matrices is used in some areas of image processing, such as restoration, but not usually in transmission. For an image (matrix) of size $m \times n$, its SVD produces an $m \times m$ matrix, an $n \times n$ matrix, and a vector of size min $\{m,n\}$. That is, the SVD generates more than double the amount of original data. Despite this fact, however, a design of an appropriate adaptive transmission strategy within this four-component procedure provides an algorithm, for lossy progressive transmission, with excellent rendering and computational performance at low percentages of data transmission.

Range image registration is a fundamental problem in range image analysis, as outlined in Chapter 3. The main task for range image registration is to establish point correspondences between overlapping range images to be registered. Since the relationship between point correspondences can be represented using motion parameters, in this chapter, we thus review the main image registration techniques from five aspects: motion representation, motion estimation, image registration using motion properties, image registration using both motion and structural properties, and the two-way constraint. In contrast with existing image registration review techniques, our review starts from the investigation of the relationship among point correspondences, motion parameters, and rigid constraints. Consequently, the review not only deepens our understanding about the relationship among motion parameters, rigid constraints, and point correspondences, but also possibly identifies the potential culprit for false matches in the process of image registration and points out future research direction to range image registration.

Given that errors in the estimates for the intrinsic and extrinsic camera parameters are inevitable, it is important to understand the behaviour of the resultant distortion in depth recovered under different motion-scene configurations. The main goal of the study in Chapter 4 is to look for a generic motion type that can render depth recovery more robust and reliable. To this end, lateral and forward motions are compared both under calibrated and uncalibrated scenarios. For lateral motion, it find that although Euclidean reconstruction is difficult, ordinal depth information is obtainable; while for forward motion, depth information (even partial one) is difficult to recover. It obtains the same conclusion in the uncalibrated case when the intrinsic camera parameters are fixed. However, when these parameters are not fixed, then lateral motion allows only a local recovery of depth order. In general, the depth distortion transformation is a Cremona transformation, and becomes a simple projective one in the case of lateral motion. It applied the above analysis to the scenario of recovering curvature of a quadric surface under lateral motion and showed that the shape estimates are recovered with varying degrees of uncertainty depending on the motion-scene configuration. Specifically, the reconstructed second order shape tends to be more distorted in the direction parallel to the translational motion than that in the orthogonal direction. It present the result of a psychophysical experiment, which confirms that in human vision, curvature estimates tend to be more erroneous and variable along the direction of lateral motion, than along its orthogonal direction.

Chapter 5 presents a stereo panoramic depth imaging system, which builds depth panoramas from multiperspective panoramas while using only one standard camera.

The basic system is mosaic-based, which means that we use a single standard rotating camera and assemble the captured images in a multiperspective panoramic image. Due to a setoff of the camera's optical center from the rotational center of the system, we are able to capture the motion parallax effect, which enables the stereo reconstruction.

The system has been comprehensively analysed. The analyses include the study of influence of different system parameters on the reconstruction accuracy, constraining the search space on the epipolar line, meaning of error in estimation of corresponding point, definition of the maximal reliable depth value, contribution of the vertical reconstruction and influence of using different cameras. They are substantiated with a number of experiments, including experiments addressing the baseline, the repeatability of results in different rooms, by using different cameras, influence of lens distortion presence on the reconstruction

accuracy and evaluation of different models for estimation of system parameters. The analyses and the experiments revealed a number of interesting properties of the system.

According to the basic system accuracy we definitely can use the system for autonomous robot localization and navigation tasks.

As explained in chapter 6, the estimation of 3-D motion and structure is one of the most important function-alities of an intelligent vision system. In spite of the best efforts of a generation of computer vision researchers, we still do not have a practical and robust system for accurately estimating motion and structure from a sequence of moving imagery under all motion-scene configurations. The authour's put forth in this study a geometrically motivated 3-D motion and structure error analysis which is capable of shedding light on global effect such as inherent ambiguities. This is in contrast with the usual statistical kinds of error analyses which can only deal with local effect such as noise perturbations, and in which much of the results regarding global ambiguities are empirical in nature. The error expression that we derive allows us to predict the exact conditions likely to cause ambiguities and how these ambiguities vary with motion types such as lateral or forward motion. Such an investigation may alert us to the occurrence of ambiguities under different conditions and be more careful in picking the solution. Our formulation, though geometrically motivated, was also put to use in modeling the effect of noise and in revealing the strong influence of feature distribution. Given the erroneous 3-D motion estimates caused by the inherent ambiguities, it is also important to understand the impact such motion errors have on the structure reconstruction. In this study, various robustness issues related to the different types of second order shape recovered from motion cue are addressed. Experiments on both synthetic and real image sequences were conducted to verify the various theoretical predictions.

This study would be most beneficial for an intelligent vision system that needs to have an estimate of the robustness of the 3-D motion and structure information recovered from the world. Such information would allow the system to carry out its tasks more effectively and to seek more information if necessary.

Chapter 7 introduces multiple-view geometry for algebraic curves, with applications in both static and dynamic scenes. More precisely, it shows when and how the epipolar geometry can be recovered from algebraic curves. For that purpose, it introduce a generalization of Kruppa's equations, which express the epipolar constraint for algebraic curves. For planar curves, it shows that the homography through the plane of the curve in space can be computed. It investigates the question of three-dimensional reconstruction of an algebraic curve from two or more views. In the case of two views, it shows that for a generic situation, there are two solutions for the reconstruction, which allows extracting the right solution, provided the degree of the curve is greater or equal to 3. When more than two views are available, it shows that the reconstruction can be done by linear computations, using either the dual curve or the variety of intersecting lines. In both cases, no curve fitting is necessary in the image space.

For dynamic scenes, it is addressed the question of recovering the trajectory of a moving point, also called trajectory triangulation, from moving, non-synchronized cameras. Two cases are considered. First it address the case where the moving point itself is tracked in the images. Secondly, it focus on the case where the tangents to the motion are detected in the images. Both cases yield linear computations, using the dual curve or the variety of intersecting lines.

Eventually, it presents several experiments on both synthetic and real data, which demonstrate that our results can be used in practical situations.

In Chapter 8, a new scheme of vision based navigation was proposed for flying vehicles. In this navigation scheme, the main navigation tool is a camera, plus an altimeter. The feasibility of this navigation scheme was carefully studied both from theory and numerical analysis. Unlike most of vision based navigation approaches in which feature trajectories were utilised to compute 3D-platform motion, it was used the image geometrical transformation parameters between consecutive frames to infer 3D displacement of camera. Due to this change, the navigation process can be conducted even if there is no salient features that can be extracted from in the image sequence, for example, in the case of flying over the sea. As a result, the long-range navigation becomes possible by use EO sensor. Moreover, the way of improvement navigation accuracy was also addressed. The experiment results demonstrated that the navigation accuracy of this system is compatible to GPS (Global Positioning system), much higher than all kinds of INS (Inertial Navigation System) in terms of position estimation. It is a good alternative choice when the GPS signal is not available

In Chapter 9 an evaluation metric for calculate the behavior of a video tracking system is proposed. This metric is used for adjusting several parameters of the tracking system in order to improve the performance. The optimization procedure is based on evolutionary computation techniques. The system has been tested in an airport domain where several cameras are deployed for surveillance purposes.

# CONTENTS

In: Computer Vision and Robotics
Editor: John X. Liu, pp. 1-27

ISBN 1-59454-357-7


*Chapter 1*

# STRUCTURE AND MOTION RECOVERY VIA STEREO-MOTION

***R. Chung*** and P.K. Ho***
Computer Vision Laboratory, Department of ACAE
The Chinese University of Hong Kong, Shatin, Hong Kong

**Abstract**

Stereo vision and visual motion are two vision cues that allow three-dimensional (3D) information of a scene to be recovered from multiple images. When a mobile platform with two fixed camera heads is available to capture stereo pair of image streams, both cues are applicable. Yet, the cues have complementary advantages: while feature correspondence is simpler in visual motion, stereo vision offers more accurate 3D reconstruction. This paper presents an approach of integrating the two cues, that retains their advantages and removes their disadvantages. It is shown that by adopting the affine camera model for the projection model of the video cameras, the two sets of motion correspondences (on the two cameras) are actually related to the stereo correspondences (across the cameras) by a matrix rank property. The rank property is important, as it allows the inference from the more readily available motion correspondences to stereo correspondences that give more accurate 3D reconstruction. In addition, the inference process could be achieved in a time only linear with respect to the total size of the image data. With the inferred stereo correspondence, both the 3D structure of the scene as well as the motion of the mobile platform could be recovered. It is also shown that with the use of all stereo pairs of image data, not only could reconstruction accuracy be boosted, even errors in the initial motion correspondences could be detected. Experiments on real image data show that 3D reconstruction is accurate even with relatively short motion of the mobile platform.

*E-mail address: rchung@acae.cuhk.edu.hk

# 1 Introduction

The capability of recovering 3D structure of a scene from visual data is important for applications like autonomous navigation and robotic manipulation. If more than one image of the scene are available the estimation problem is potentially easier because of the more information available about the imaged scene. There exists two major vision cues that employ such a multi-ocular approach. One is visual motion, in which 3D structure is recovered from an image sequence that is 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 to solve 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 feature correspondences.

The motion cue has the advantage that the correspondence problem is relatively easy to solve, because successive images are alike. However, it generally requires a long image sequence, up to hundreds of frames (for instance in [19]), for accurate 3D reconstruction. The reason is, 3D determination from multi-ocular vision is based upon intersecting the respective images' corresponding projection rays. To reduce the effect of disturbances like image noise etc. to the reconstruction, the physical separation between the spatial positions of the images, i.e., the baseline, must be wide enough.

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 generally 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 two viewpoints could reduce the originally 2D search to 1D search along the so-called epipolar lines [12].

With the above observations, we outlined in [8] a framework of combining the two vision cues, in which the affine projection model is used for the cameras. In contrast with previous work on stereo-motion like [21, 14, 25, 24], the framework emphasizes not on how to exploit the redundancy in the image data to boost the accuracy in 3D reconstruction, but on how to couple the two vision cues in a complementary way, so that their advantages are retained and their disadvantages removed. The framework relates motion correspondences to stereo correspondences, and allows inference from the former to the latter. Accurate 3D reconstruction was demonstrated, even with relatively short platform motion.

However, several points are yet to explore. First, only one stereo pair were used for 3D reconstruction. This is not entirely reasonable, as any stereo pair is as good as the other in the image data for recovering 3D information. Second, the platform motion was not recovered. This paper presents how to compute 3D structure and motion from all stereo pairs of images. It is demonstrated that not only could more accurate reconstruction results be obtained by using all image data, even false initial motion correspondences could be detected, and thus establishment of wrong stereo correspondences could be avoided by comparing the results from different stereo pairs.

## 2 Previous Work

Much has been done on stereo vision; good surveys can be found in [4, 10]. Yet due to the difficulty of its correspndence problem, it hasn't been widely used in industry and society.

Visual motion has also been well-studied; classical references are listed in [13, 22]. The correspondence problem is much simpler than that in stereo vision, as consecutive images are alike, which means a feature point could not move too far between consecutive images. Very good 3D reconstruction results have been obtained, for example in [19]. One drawback is that a long image sequence is required so as to have a wide enough triangulation for accurate 3D determination. Such a drawback is not unimportant, as the longer distance the camera needs to travel, the more probable are the needed assumptions (e.g., a stationary scene) violated.

Below a few works on motion analysis that are closely related to this work are outlined.

In an elegant work, Tomasi and Kanade [19] proposed a method for reconstructing 3D from an orthographically projected image sequence. It factorizes the image measurements of object points into shape and motion matrices through singular value decomposition (SVD). Later, Poelman and Kanade [17] extended the factorization method to the case of paraperspective projection, which produces more accurate results than the original method. 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 [15] presented a sequential approach for the factorization method. The sequential approach is much faster than the original one, but a long image sequence is still required.

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. A few studies [21, 14, 25, 1, 24, 16, 7] have looked into this so-called *stereo-motion* cue.

However, the focus of the above stereo-motion work was on the exploitation of the input data's redundancy 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. The emphasis of this work, in contrast, 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.

# 3 Stereo and Motion in Complement

## 3.1 The Motion Model

In [19] Tomasi and Kanade proposed an elegant discrete model for the motion cue. Below the model, with some variations to pave the way for further development, is described.

Suppose F image frames observing P points in space are available. Assume an affine camera. The image position $\mathbf{p}_{fp} = (u_{fp}, v_{fp})^T$ of point p ($p = 1,2,\ldots,P$) in image frame f ($f = 1,2,\ldots,F$), is related to its 3D position $\mathbf{P}_p = (x_p, y_p, z_p)^T$ (with reference to the last image frame: frame F), by

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

where $J_f$ is the affine projection matrix (a $2 \times 4$ matrix), and $(\mathbf{R}_f, \mathbf{t}_f)$ are the rotational and translational relationships between image frame f and the last image frame F. By combining the image positions of all P object points in F image frames, we have

$$\underbrace{\begin{bmatrix} \vdots \\ \cdots\ \mathbf{p}_{fp}\ \cdots \\ \vdots \end{bmatrix}}_{\mathbf{W}} = \underbrace{\begin{bmatrix} \ddots & \bigcirc \\ & J_f \\ \bigcirc & \ddots \end{bmatrix}}_{\mathbf{J}} \underbrace{\begin{bmatrix} \vdots \\ \mathbf{M}_f \\ \vdots \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} \cdots & \mathbf{P}_p & \cdots \\ & 1 & \end{bmatrix}}_{\mathbf{S}}$$

Here **W**, **J**, **M**, **S** represent the image measurements, the image projection process, the camera motion, and the scene or the object structure respectively. Each row in **W** contains the u-coordinates or v-coordinates of image points the same image frame, while each column contains the observations over the same object point. Since **W** can be factorized into matrices involving dimension four, **W** is of rank at most four (it is exactly four under general motion and general 3D structure).

## 3.2 The Stereo-Motion Model

The above motion model has been applied successfully to recover 3D structure [19]. 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 motion the 3D structure can still be estimated accurately, since widely separated views are always in the image data.

The motion model could be extended to the stereo-motion problem in the following way. Suppose a rigid stereo setup consisting of two cameras: Cameras 1 and 2, are available to capture image data as the whole setup moves in space relative to a scene. As shown in Figure 1, let $(\bar{\mathbf{R}}, \bar{\mathbf{t}})$ be the rotational and translational relationships between the stereo cameras (which are invariant with the motion of the stereo setup), in the sense that the 3D coordinates of any point with respect to the two camera coordinates frames (of Cameras 1 and 2 respectively), $\mathbf{P}$ and $\mathbf{P}'$, are related by

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

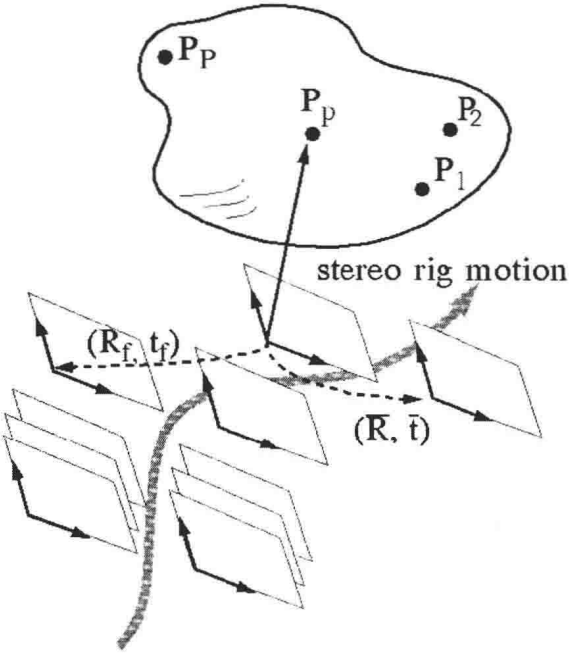


Figure 1: 3D Structure recovery from stereo-motion.

On applying Tomasi-Kanade's motion model to the two cameras separately, we have two image measurement matrices for feature points in the two cameras respectively:

$$\begin{aligned} \mathbf{W} &= \mathbf{JMS} \\ \mathbf{W}' &= \mathbf{J}'\mathbf{M}'\mathbf{S}' \end{aligned}$$

Here $\mathbf{W}', \mathbf{J}', \mathbf{S}'$ are matrices analogous to $\mathbf{W}, \mathbf{J}, \mathbf{S}$, but $\mathbf{W}', \mathbf{J}', \mathbf{S}'$ are with respect to the second 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 $\mathbf{S}$. If columns of $\mathbf{W}'$ and $\mathbf{S}'$ are so listed and $\mathbf{W}'$ is stacked beneath $\mathbf{W}$, a new image measurement matrix is obtained for the stereo-motion data:

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

$$= \underbrace{\left[\begin{array}{c|c}\mathbf{J} & \bigcirc\\ \hline \bigcirc & \mathbf{J}'\end{array}\right]}_{\mathcal{J}} \underbrace{\left[\begin{array}{c}\mathbf{I}_{4\mathsf{F}}\\ \hline \widetilde{\mathbf{M}}\end{array}\right]}_{\widetilde{\mathcal{M}}} \mathbf{M})\mathbf{S} \qquad (1)$$

where $\widetilde{\mathbf{M}}$ is a $4\mathsf{F} \times 4\mathsf{F}$ matrix representing the stereo camera geometry :

$$\widetilde{\mathbf{M}} = \begin{bmatrix} \ddots & & \bigcirc \\ & \bar{\mathbf{M}} & \\ \bigcirc & & \ddots \end{bmatrix}$$

The matrices $\mathcal{W}$ and $\mathcal{J}$ are analogues of $\mathbf{W}$ and $\mathbf{J}$ (which are for single-camera motion) in stereo-motion. The matrix $\mathcal{W}$ (size: $4\mathsf{F} \times \mathsf{P}$) represents the image measurements from the stereo cameras with the stereo correspondences correctly established, and $\mathcal{J}$ (size: $4\mathsf{F} \times 8\mathsf{F}$) represents the image projection parameters of the stereo cameras. $\widetilde{\mathcal{M}}$, a term not present in the original motion model, is a $8\mathsf{F} \times 4\mathsf{F}$ matrix representing the geometry of the stereo camera setup. Notice that although here we have a stereo pair of cameras not one camera, the 3D structure term $\mathbf{S}$, like the counterpart in the motion model, is with reference to the camera coordinate frame of Camera 1 over the last image frame (i.e., Image $\mathsf{F}$).

Since the factorization in Equation (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 and in general four (under general 3D structure and motion). Such a property is unlikely to be satisfied accidentally, as $\mathcal{W}$ is $4\mathsf{F} \times \mathsf{P}$ large; it is however satisfied when $\mathcal{W}$ is constructed under fully correct stereo matching. As will be discussed in the next section, the property allows stereo correspondences to be inferred from motion correspondences which are easier to obtain.

### 3.3 Inferring Stereo Correspondences from Motion Correspondences

Our stereo-motion system proceeds in the following way. A rigid stereo rig of cameras is constructed, and it undergoes a motion during which $\mathsf{F}$ pairs of images are taken from the cameras. Distinct feature points are then extracted independently from the two image sequences, and tracked in the two sequences separately.

We assume most of the estimated motion correspondences are correct since the image frames are dense and thus adjacent images are very much alike (however, we do allow mistakes in the motion correspondences, which are to be addressed in Section 3.6). With

such motion correspondences the image measurement matrices $\mathbf{W}^*$ and $\mathbf{W}'^*$ for the two image sequences 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 necessarily properly ordered, i.e., stereo correspondences are not 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.

Our idea is to transfer the motion correspondences, which are easier to obtain, to stereo correspondences, which allows more accurate 3D reconstruction. Establishing stereo correspondences across the two images sequences is equivalent to matching columns of $\mathbf{W}^*$ with columns of $\mathbf{W}'^*$, so as to have matched pairs of columns to form the matrix $\mathcal{W}$ in Equation (1).

As $\mathcal{W}$ is 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 four basis vectors of $\mathcal{W}$ are available as $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4$, and let $\mathbf{B}$ be $[\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4]$. Since $\mathbf{W}$ and $\mathbf{W}'$ are sub-matrices of $\mathcal{W}$ and of rank 4, the upper and lower sub-matrices of $\mathbf{B}$ — $\mathbf{B}_W$ and $\mathbf{B}_{W'}$ (size: $2\mathrm{F} \times 4$) — are also matrices consisting of basis vectors for $\mathbf{W}$ and $\mathbf{W}'$ respectively.

Take any column in $\mathcal{W}$, which has $\mathbf{h}_W$ as its upper sub-column and $\mathbf{h}_{W'}$ as its lower sub-column. $\mathbf{h}_W$ represents a column of $\mathbf{W}$ that corresponds to the motion correspondence of a feature point in Image Sequence 1, and $\mathbf{h}_{W'}$ represents a column of $\mathbf{W}'$ that corresponds to the motion correspondence of the same feature point in Image Sequence 2. If there is a way to predict $\mathbf{h}_{W'}$ for every $\mathbf{h}_W$, the problem of inferring stereo correspondences is essentially solved.

It turns out if Basis $\mathbf{B}$ of $\mathcal{W}$ is available (and thus Basis $\mathbf{B}_W$ of $\mathbf{W}$ and Basis $\mathbf{B}_{W'}$ of $\mathbf{W}'$ as well), $\mathbf{h}_{W'}$ of $\mathbf{W}'$ could indeed be predicted for every $\mathbf{h}_W$ of $\mathbf{W}$ as:

$$\mathbf{h}_{W'} = \mathbf{B}_{W'} (\mathbf{B}_W^{\mathrm{T}} \mathbf{B}_W)^{-1} \mathbf{B}_W^{\mathrm{T}} \mathbf{h}_W \tag{2}$$

Derivation of the above formula is simply based upon the fact that the set of linear combination coefficients that generate $[\mathbf{h}_W^{\mathrm{T}}, \mathbf{h}_{W'}^{\mathrm{T}}]^{\mathrm{T}}$ from Basis $\mathbf{B}$ also generate $\mathbf{h}_W$ from Basis $\mathbf{B}_W$, and $\mathbf{h}_{W'}$ from Basis $\mathbf{B}_{W'}$ as well.

In other words, 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 epipolar constraint of stereo cameras. If more than 4 matches are available, a more accurate basis can be determine by the SVD technique [19].

Thus, the stereo-motion framework could use Equation (2) for inferring stereo correspondences from motion correspondences. With noise, the estimated column $\mathbf{h}_{W'}$ may not be exactly that in $\mathbf{W}'^*$, but should be quite close to it. A column is then selected from $\mathbf{W}'^*$ that has least-squares-error with it. This way stereo correspondences can be fully established, and $\mathbf{W}^*$ and $\mathbf{W}'^*$ can be organized to form $\mathbf{W}$ and $\mathbf{W}'$ and also $\mathcal{W}$. An input-output description of the inference mechanism is summarized in Figures 2.

Notice that once $\mathbf{B}$ is known, for each feature point whose motion correspondence in