STEREO-BASED OBSTACLE DETECTION USING GABOR FILTERS

By

Richard Neil Braithwaite

B. Sc. (Electrical Engineering) University of Calgary

M.A.Sc. (Electrical Engineering) University of British Columbia

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

ELECTRICAL ENGINEERING

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 1992

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of ELECTRICAL ENGINEERING

The University of British Columbia
Vancouver, Canada

Date JULY 3, 1992

DE-6 (2/88)

Abstract

This work presents a new obstacle detection algorithm that uses Gabor filters. The task performed by this algorithm is the detection of moving and stationary obstacles from an autonomous vehicle undergoing predominantly rectilinear motion. Image measurements from stereo cameras are used to extract three-dimensional properties of viewed objects and of the vehicle. Properties such as depth and motion are used to predict if (and when) the object will collide with the vehicle.

Three inherently difficult problems associated with the estimation of depth and motion from stereo images are solved. (1) Stereo and temporal correspondence problems are solved using predictive matching criteria. (2) Segmentation of the image measurements into groups belonging to stationary and moving objects is achieved using error estimation and the "Mahalanobis distance." (3) Compensation for transient rotations produced by a shaking camera is achieved by internally representing the inter-frame (short-term) camera rotations in a rigid-body dynamical model. These three solutions possess a circular dependency, forming a "cycle of perception." A "seeding" process is developed to correctly initialize the cycle.

An additional complication is the translation-rotation ambiguity that sometimes exists when sensor motion is estimated from an image velocity field. Eigenvalue decomposition is used to detect such ambiguity. Temporal averaging using Kalman filters reduces the effect of motion ambiguities.

The obstacle detection algorithm operates correctly in a variety of difficult conditions such as: stereo images with different brightness; image sequences with large image velocities; transient sensor rotations; and concurrent object and sensor motion. Under

these difficult conditions, the obstacle detection algorithm presented in this thesis is able to identify moving objects, and distinguish between obstacles that will collide with the vehicle and objects that will pass safely by the vehicle.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Autonomous vehicles are designed to operate without human intervention. They can operate in hostile environments making them ideal for tasks like space and undersea exploration, as well as inspection of contaminated areas. A common task for an autonomous vehicle is to move safely from its current location to another. An important requirement for safe motion is the reliable detection of unexpected obstacles. This thesis introduces a new obstacle detection algorithm that can be used in the navigational control system of an autonomous vehicle. The use of collision parameters—the point-of-collision and the time-to-collision—is a departure from standard methods.

Section 1.1 describes the utility of obstacle detection and differentiates the detection of obstacles, which impede the vehicle motion, from the detection of objects, which can be used as landmarks for passive navigation. Section 1.2 describes the modules within a navigational control system and how they interact. Section 1.3 introduces the point-of-collision and the time-to-collision as important concepts of obstacle detection.

Section 1.4 defines the scope of the work and parameters used in the thesis. The section discusses the choice of sensors, the choice of prediction models, and the choice of error models. Section 1.5 describes existing camera-based implementations that relate to autonomous navigation. It also introduces a new obstacle detection algorithm that uses Gabor filters and discusses its contribution to the field. Section 1.6 concludes the introduction with an outline of the thesis.

1.1 Obstacle Detection

Obstacle detection can be useful in three different contexts. It can be used to assist a human operator, it can be used to increase the autonomy of a vehicle, or it can be used as part of the navigational control system of an autonomous vehicle. As an assistant to a human operator, the obstacle detection module acts as an alarm, alerting the operator if an object impedes the planned path. This is useful if the operator has more than one task to perform or if the operator is fatigued. Increased autonomy is useful for remote operations. Remotely operated vehicles, such as Martian land rovers, experience large transmission delays between commands. Direct control of the vehicle is not possible. Obstacle detection, combined with a simple obstacle avoidance scheme, can protect the remote controlled vehicle from unanticipated hazards. As part of the navigation control system of an autonomous vehicle, the obstacle detection module must provide the computer pilot with sufficient information to avoid obstacles. The information includes the "point-of-collision" and the "time-to-collision" for each viewed object. The point-of-collision specifies either the location (on the vehicle) at which the object will collide with the vehicle or how close the object will be as it passes by the vehicle. The time-to-collision specifies how much time will elapse before the object collides with (or passes by) the vehicle.

Obstacle detection is often grouped, and sometimes confused with, two related tasks: object detection, and obstacle avoidance. To avoid confusion, it is important to distinguish between objects and obstacles. An "object" is any physical feature within the sensor's field of view. An "obstacle" is an object that obstructs the vehicle's desired motion trajectory [1]. From these definitions it is apparent that object detection is based

[1] In a practical implementation, the definition of an obstacle is expanded to include objects that obstruct small deviations from the desired trajectory. The expanded definition allows the determination of feasible escape routes when the desired trajectory is obstructed.

on measurement: sensor information is used to localize the object's three-dimensional position and velocity. Obstacle detection is based on prediction. Using the position and velocity of the object relative to the sensor, the obstacle detection module predicts the point-of-collision and the time-to-collision [2]. Obstacle avoidance is a control problem. The obstacle avoidance module generates control signals to maneuver around any detected obstacles.

1.2 Navigation Control for an Autonomous Vehicle

This section describes the modules within a navigation control system, displayed in figure 1.1. The goal is assigned by the human task master. The goal is usually time-invariant; it only changes when the task is completed or aborted. A typical goal would be to move safely to a new location. The next levels contain a hierarchy of maps and path planners. The maps model the position of known objects in the ego-vehicle's environment. The path planners navigate around them.

There is always a degree of uncertainty in the absolute position of the ego-vehicle. Any uncertainty in the vehicle's position will affect the accuracy of the relative position of known obstacles. In addition, the position of moving obstacles is continually changing. Both the obstacle detection and obstacle avoidance modules, whose basis is sensor data, are required to safely navigate around these unexpected obstacles.

The purpose of the obstacle detection module is to predict imminent collisions of the ego-vehicle with an obstacle. A detailed description of the obstacle detection module appears in chapter 4. The obstacle avoidance module selects evasive maneuvers when a collision is imminent. It is an emergency module that is used when there is insufficient time to update the local map and plan a new path. In such cases the control signals from

[2]The object is an obstacle if the point-of-collision is on the ego-vehicle.

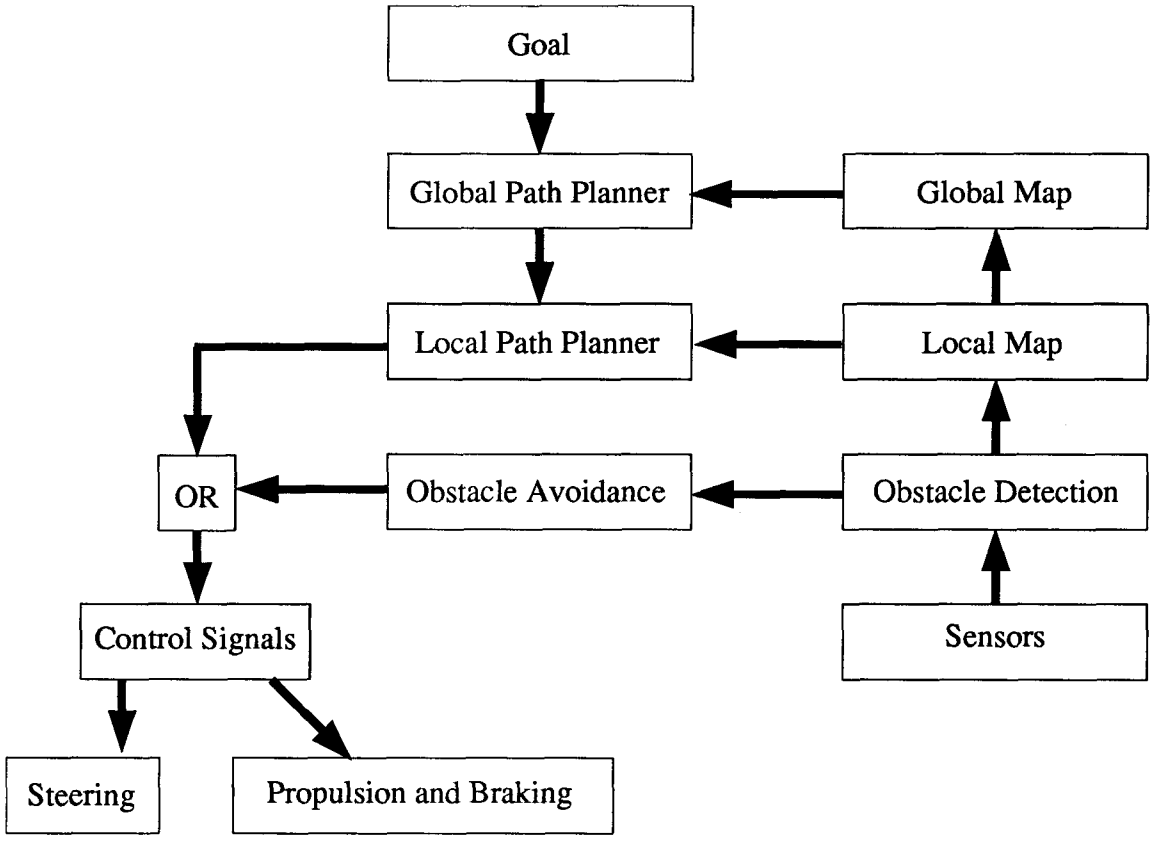


Figure 1.1: Modules of an Autonomous Navigation System

the obstacle avoidance module override the speed and direction signals from the local path planner.

The modules in figure 1.1 interact with each other. Higher order objectives are filtered down through the path planners to the actuator (steering, braking) level. Sensor information rises to higher levels as obstacles are detected and maps are updated. The frequency of these interactions is related to the module's position in the control hierarchy. The sensor information and the actuator control signals, at the bottom of the control hierarchy, are updated frequently; the goal, at the top of the hierarchy, might be time-invariant.

Complete vehicle autonomy is not always practical. In many cases a human operator is still required. As the degree of autonomy is increased, less operator supervision is required; the frequency of interaction between man and machine is reduced. To increase the degree of autonomy of a vehicle, the design of higher level modules must be preceded by the proper selection of sensors and the robust design of the obstacle detection module.

1.3 Collision Parameters

In this section, important obstacle detection parameters—the time-to-collision and the point-of-collision—are discussed. Models of the sensor and object motion are required to predict these collision parameters. Three possible kinetic models are proposed.

The path travelled by an object relative to the sensor group is referred to as the "observer frame trajectory" of the object. This trajectory is given by

$$x(t) = x_0 + \int_{t_0}^{t} \dot{x}(t)\, dt, \tag{1.1}$$

$$y(t) = y_0 + \int_{t_0}^{t} \dot{y}(t)\, dt, \tag{1.2}$$

$$z(t) = z_0 + \int_{t_0}^{t} \dot{z}(t)\, dt, \tag{1.3}$$

where t_0 is the current time, $(x_0,\ y_0,\ z_0)$ is the current position of an object, and $(\dot{x},\ \dot{y},\ \dot{z})$ is the object velocity relative to the observer coordinate frame. The point-of-collision is the intersection of the object's trajectory and the plane defined by $z(t) = 0$; that is

$$x_{col} = x_0 + \int_{t_0}^{t_0+t_{col}} \dot{x}(t)\ dt, \tag{1.4}$$

$$y_{col} = y_0 + \int_{t_0}^{t_0+t_{col}} \dot{y}(t)\ dt, \tag{1.5}$$

$$0 = z_0 + \int_{t_0}^{t_0+t_{col}} \dot{z}(t)\ dt, \tag{1.6}$$

where $(x_{col},\ y_{col})$ is the point-of-collision, and t_{col} is the time-to-collision. If the point-of-collision is less than the extent (height, width) of the ego-vehicle, a collision will occur. An obstacle detection algorithm that uses the point-of-collision and time-to-collision can be considered a very intelligent (predictive) proximity sensor.

The use of the point-of-collision and the time-to-collision in obstacle detection represents a departure [3] from traditional methods. Most implementations are designed for operation in a stationary environment [6] [12] [40] [41] [42] [43] [48] [49]. In such cases, the observer frame trajectory is simply defined by the position of the object relative to the ego-vehicle, and the velocity of the ego-vehicle. As a result, most research has been concentrated on localizing the three-dimensional position of objects (particularly the depth), and maintaining the positional representation as the ego-vehicle moves [6].

The time-to-collision is a measure of the urgency associated with a colliding obstacle. In a stationary environment, either depth (z_0) or time-to-collision can be used to measure urgency. Close obstacles present a more immediate danger than distant obstacles. Distant obstacles can be temporally ignored while the obstacle avoidance module maneuvers the

[3]The use of point-of-collision and time-to-collision is a departure from traditional computational vision approaches. The terms have been used in psychology for quite some time. Gibson [25], in a 1938 paper, described an automobile driver's perception of obstacles in terms of clearance lines and point-of-potential-collision, which are similar to the point-of-collision. Lee [37] described the (perceived) proximity of an obstacle using time-to-collision.

vehicle around close obstacles. In a stationary environment, the time-to-collision increases if the ego-vehicle slows down; a halted vehicle has no chance of collision. In a dynamic environment, however, stopping does not ensure safety. A moving obstacle can collide with a halted vehicle. In such an environment, depth is no longer the best measure of urgency; time-to-collision becomes more important.

The point-of-collision and the time-to-collision are predicted from the past sensor and object motions. Prediction requires models of the sensor and object motion. Three possible motion models include pure translation, predominantly rectilinear, and general translation and rotation. Pure translation is the simplest model. It assumes there is no sensor rotation. The advantage of this model is that the collision parameters can be estimated from the current position and velocity of the object relative to the ego-vehicle. The predominantly rectilinear model assumes that the sensor is undergoing both translation and small rotation. The advantage of this model is that the solution can be obtained using a simple least squares approach if the scene structure, that is the depth, is known. In the general motion case, the sensor rotation is unconstrained. The solution to this nonlinear problem is obtained using iterative schemes that require good initial guesses. The choice of models may be limited by the physical properties of the sensor motion and the environment in which the ego-vehicle operates. If a choice is available, it is a good idea to choose the simplest model that adequately represents the motion.

In this work, the predominantly rectilinear model is used for the inter-frame (short-term) sensor motion. The extended (long-term) sensor motion and the object motion are modelled as pure translation. The implementation of these models can be found in chapter 4.

1.4 Task and Scope

This section defines the task to be completed and the scope of the thesis. The operating environment of the ego-vehicle is defined, the sensors are selected, and the models of sensor and object motion are chosen. Sources of additional knowledge that can improve performance are identified.

The task in this work is to develop an obstacle detection algorithm for an autonomous vehicle undergoing rectilinear motion in a dynamic environment. By dynamic environment, it is assumed that in addition to a moving ego-vehicle, there are also moving objects. Both the ego-vehicle and objects are undergoing translational motion, but experience disturbances that cause small transient rotations. It is also assumed that the ground surface is stationary and rigid, and that the moving objects are rigid.

The work presented in this thesis concentrates on object detection and obstacle detection. Object detection requires a high resolution imaging sensor to localize 3D position. In this work, a stereo pair of CCD cameras are used. A stereo camera system is selected for the following reasons: it is passive, it has low power consumption, and it is available. Passive sensors are very important in military applications. Active sensors tend to reveal the presence of the ego-vehicle. A vehicle equipped with passive sensors has a greater probability of surviving in a battle zone. Power consumption is an important constraint for space vehicles. Active sensors, such as laser range finders, consume too much power to be useful on a Martian rover [41]. Finally, the CCD cameras are commercially available.

Obstacle detection requires models of sensor and object motion to predict the collision parameters. Since the cameras used in this work have a narrow field of view [4], the inter-frame image measurements are very sensitive to sensor rotation. The predominantly rectilinear model is appropriate for the inter-frame sensor motion because it internally

[4]Most of the image sequences in chapter 5 are obtained using cameras with a field of view spanning 30 degrees.

represents the transient sensor rotations. In contrast, small transient rotations of a viewed object have little effect on the inter-frame image measurements. As a result, the inter-frame object motion is modelled as pure translation.

The extended sensor and object motions are modelled as pure translation. The extended motion is an estimate of the sensor/object translation integrated over the image sequence. The extended sensor and object motions are combined to obtain the relative translation of each object to the sensor. The relative translation, $(\dot{x}, \dot{y}, \dot{z})$, is used in the following equations to estimate the collision parameters:

$$x_{col} = x_0 + \dot{x}\, t_{col}, \tag{1.7}$$

$$y_{col} = y_0 + \dot{y}\, t_{col}, \tag{1.8}$$

$$t_{col} = -\frac{z_0}{\dot{z}}. \tag{1.9}$$

The use of stereo cameras as the sensor creates problems that must be addressed. Image measurements are very sensitive to camera rotations. As previously mentioned, the solution is to use an inter-frame sensor motion model that internally represents the transient rotations. Another problem, which is related to the co-existence of stationary and moving objects in the scene, is the segmentation of the image measurements into groups belonging to stationary objects and moving objects. The solution to the segmentation problem is discussed in section 4.3. Solutions to the stereo and temporal correspondence problems are discussed in chapter 3 and section 4.6.1.

The final significant problem is the inherent translation-rotation ambiguity that sometimes exists when sensor motion is estimated using image measurements [3]. This ambiguity can be resolved using additional knowledge sources, such as motion constraints or auxiliary sensors. The (camera) sensor and object motion models can incorporate constraints such as planar motion. The planar constraint has two forms: a planar surface

with a known surface normal, and a planar surface with an unknown surface normal. In this work, such constraints are incorporated using penalty terms (see chapter 4). The ego-vehicle can also be equipped with auxiliary sensors that measure inter-frame motion parameters directly: such as a speedometer, or (changes in) compass heading. In general, motion constraints and auxiliary sensors are not necessary, but the information can improve the inter-frame (camera) sensor motion estimate for certain scene structures (such as a frontal plane, see chapter 5, experiment 1). The effect of the translation-rotation ambiguity on the extended sensor motion is reduced by temporal integration.

1.5 Previous Works and Author's Contributions

There are numerous camera-based implementations relating to navigation of an autonomous vehicle. They typically belong to one of three classes: restrictive environment; three-dimensional positional integration; and image displacement methods. This section discusses the strengths and weaknesses of the "better" implementations from each class. This section also introduces the new Gabor filter-based obstacle detection algorithm and discusses its principal contributions.

The restrictive environment implementation uses knowledge of its operating environment to navigate safely. The restrictive environment has a simple structure that can be modelled using a small number of scene parameters. Road-following autonomous vehicles, such as [48] [49], are examples. German researchers, Dickmanns et al [16] [17] [18] [19] [36] [44] [53], have produced the most impressive [5] of the road-following implementations.

The strength of Dickmanns' design stems from the use of the Kalman filter, which consists of the following: state variables that comprise scene and motion parameters; dynamical (process) models that enforces world knowledge about the physical laws of

[5] It has been tested at speeds upto 100 km/hr.

motion and the conventions used in road design; and a measurement model that enforces knowledge of the perspective transformation of state variables into image features (road edges). Dickmanns' implementation of the Kalman filter forms a "cycle of perception:" the dynamical models predict future state variables from the current state and future control inputs; the measurement model and state variables are used to predict the position of road edge in the camera image; and the detected road edges are used to update the state variables. The prediction of the position of the road edge in the image defines the "context of perception [4]," reducing the search space for a road edge and improving the robustness road detection.

The weaknesses in the Kalman filter approach are related to the accuracy with which the state variables represent the actual process, and the initialization of the cycle of perception. Dickmanns' state variables and dynamical models do not account for transient rotations. It is noted in [18] that the quality of the state estimation is affected by the pitching motion of the vehicle (one component of transient rotation). The initialization of the cycle of perception is very important. If the initial search for the road edge produces incorrect estimates of the state variables, then the cycle of perception will search the wrong regions of future images for new road edges. Groupings of image features and knowledge about road structure are used by Dickmanns to ensure a correct start-up state [18].

Dickmanns et al [19] have developed an obstacle detection algorithm for their road-following vehicle. The image projection of the road lane is searched for edges belonging to other vehicles. Detected edges are used in a Kalman filter to estimate the range (depth, z) and the range rate ($\dot{z}$) of the ego-vehicle relative to the other vehicle. The use of range and range rate together provides the same information as the time-to-collision. Knowledge of road convention (in this case, lane markings) eliminates the need for an equivalent to the point-of-collision; it is only necessary to determine if the lane is blocked.

The drawback of this approach is that Dickmanns' obstacle detection algorithm can not anticipate collisions when the obstacle is moving across the road (as often occurs at intersections).

There are a number of implementations that use three-dimensional position information to form local maps (in place of parameterized scene models) and to determine the sensor motion [6] [40] [43]. The position measurements of distinctive physical features are integrated over time. For optimal integration, a positional error estimate, in the form of a 3 by 3 error covariance matrix, must be maintained for each physical feature. These physical features are used as landmarks for estimating sensor motion. The sensor motion model enforces the following constraints: the environment is rigid and stationary.

The main weakness of [6] and [40] is that they are not designed for scenes with moving objects. Another weakness, with respect to obstacle detection, is the over-head associated with modelling positional uncertainty. For the case of obstacle detection, motion information is more important than positional information [6]. In [40], the inter-frame sensor motion is estimated in two stages: image measurements are transformed into three-dimensional position estimates for a set of features; then, the sensor motion is estimated from the differences in three-dimensional position of features in successive images. The positional error covariance matrices ensures that no information is lost in this intermediate step. It is possible to by-pass the intermediate positional representation by measuring positional changes directly from the image. In such an approach, the positional error covariance matrices are not necessary for the estimation of sensor motion.

Image displacement methods measure the positional change of brightness pattern in the image over time [5] [27]. A map of the local two-dimensional image displacements is referred to as an "optical flow field." Adiv [2] uses the optical flow field to segment

[6] It can be seen from (1.7) and (1.8) that an error in the in-plane velocities, $\dot{x}$ and $\dot{y}$, will alter the point-of-collision more than positional errors in x_0 and y_0, when the time-to-collision is large.

an image sequence and to estimate motion. The strength of Adiv's method is that it internally represents sensor rotation.

The weakness of Adiv's method is the (improper) use of world knowledge and his choice of inputs. Adiv segments the image sequence by grouping image measurements that are consistent with rigid-body motion of a planar surface patch. As a result, the segmentation is based on structure, as well as motion. An extra knowledge source is needed to identify which groupings belong to stationary objects. Another drawback of this method is that it relies on the optical flow field, which is difficult and computationally expensive to calculate [2] [3] [31].

The calculations associated with the optical flow field can be avoided by using the component of image velocity that can be directly (and locally) measured: that is, the "normal image velocity." Nelson and Aloimonos [46] use flow-field divergence in the component (normal) direction to avoid obstacles. Collisions are avoided by steering away from image regions with large flow-field divergence. The strength of this method is that it does not require an optical flow field, and it is not affected by the translation-rotation ambiguity. The weakness of the flow-field divergence method is that it is a primitive form of obstacle detection: there is no prediction of the time-to-collision or the point-of-collision. A second weakness is that the large divergence criterion can produce false positive responses; in addition to a close object that is approaching the sensor, a large divergence can be caused by a depth gradient or a motion boundary.

Jenkin and Jepson [34] use stereo camera-based trajectory detectors to identify obstacles. Stereo Gabor filters, tuned to a particular disparity and pair of normal image velocities [7], are used to estimate the "out-of-plane" motion, $\dot{z}$, and the "in-plane" motion, $\dot{x}$ and $\dot{y}$. The disparity and the out-of-plane motion provide sufficient information

[7]Disparity is the image displacement between stereo viewpoints. It is used to calculate depth. The normal image velocity is the component of image displacement over time in the direction normal to the image contour. Both terms are described in chapter 2.

to calculate the time-to-collision. The weakness of the trajectory detector is that it is unable to predict the point-of-collision because it can not distinguish between motion induced by in-plane translation and non-axial sensor rotation.

The inability to distinguish between translation and rotation is due to the fact that the trajectory detector is based on local image measurements [31]. Global characteristics of a set of image measurements can be used to determine the sensor motion. One computationally efficient technique is "direct passive navigation" [31] [33] [45]. Subject to certain assumptions (listed below), a set of local image gradient measurements can produce an estimate of the inter-frame sensor motion. The gradient of any image measurement can be used, but image intensity is the most common [5] [33]. The strengths of the direct passive navigation approach is that it avoids the calculation of the optical flow field (resulting in an order of magnitude speed-up [30]), and it internally represents both sensor translation and rotation in a rigid-body motion model.

The weaknesses of the direct passive navigation approach stems from two assumptions. It is assumed that all the image gradient measurements belonging to stationary objects. In a dynamic environment, the image sequence has to segmented into image measurements belong to stationary and moving objects. The second assumption is that the depth (or at least the scene structure) is known at each image measurement. Stereo cameras can be used to extract depth. However, to avoid the "difficulties" and the "computational burden" [31] associated with the measurement of depth from stereo, either the structure of the scene must be constrained (as in [45]) or the motion must be constrained (as in [31]). Such constraints have utility only if the operating environment enforces the constraints.

The above-mentioned camera-based navigation approaches provide an insight into useful obstacle detection design. The use of world models that enforce natural constraints in the operating environment can reduce the computational requirements and improve

the robustness of state estimation (as demonstrated by Dickmanns et al [18]). There is a tendency, however, to choose constraints that make the algorithm tractable (for example, [2] [31] [45]), instead of evaluating the environmental constraints. To ensure that all chosen constraints are acceptable, it is very important to test constraint-based algorithms using real or "realistic" data. The obstacle detection algorithm presented in this thesis (chapter 4) is thoroughly tested using realistic data (chapter 5).

The new obstacle detection algorithm presented in this thesis possesses many of the strengths mentioned above: it uses a cycle of perception to measure normal image velocity and sensor/object motion; and it internally represents sensor rotation to compensate for motion transients. In addition, the new algorithm has the ability to predict collisions with moving objects in an unrestricted environment. It can even predict a collision for the case of an object crossing the forward path of the ego-vehicle (a problematic case for Dickmanns).

Constraints on the scene structure limit the generality of an implementation. In this thesis, the depth is estimated using disparity from stereo cameras. The disparity is measured using Gabor filters. Since the preprocessing of images using Gabor filters is common to both the disparity and normal image velocity modules, the estimation of depth represents only a modest additional computational burden to the proposed obstacle detection algorithm.

The new Gabor filter-based obstacle detection algorithm combines the direct passive navigation and the trajectory detector approaches. In order to use these two approaches simultaneously, the image sequence must be segmented into regions belonging to stationary objects and moving objects. The rigidity constraint is used to detect the moving objects, whose image measurements are inconsistent with those induced by sensor motion only. Image measurements belonging to stationary objects are used as landmarks in the direct passive navigation approach to estimate sensor motion. After compensating

for sensor motion, the motion of each moving object is estimated using the trajectory detector approach.

The new obstacle detection algorithm can be described briefly. The phase and magnitude responses from Gabor filtered images are used to extract disparity and normal image velocity. Those image measurements belonging to stationary objects are combined to obtain an estimate of the inter-frame sensor motion (both translation and rotation). The inter-frame translation of each moving object is estimated using the excess normal image velocity; that is, the measured normal image velocity minus the image velocity induced by the sensor motion. The inter-frame sensor and object translations are integrated over the image sequence using Kalman filters. The difference between the object and sensor translations is the observer frame trajectory of the object. The object's trajectory, along with its current position, is used to predict the collision parameters.

The work presented in this thesis makes the following principal contributions:

- it implements direct passive navigation using phase-differences instead of intensity derivatives;
- it develops error models for the Gabor-based image measurements that are propagated into collision parameter uncertainty;
- it develops a "seeding" technique that is necessary to initialize the segmentation process;
- it stabilizes the image sequence from transients caused by camera shake.

Image displacements measured using phase-differences are more stable than intensity derivatives [21] [23]: to changes in image contrast and brightness; and to geometrical deformations caused by changes in sensor position, orientation, or viewpoint. The expected error for an image measurement is needed to perform segmentation of stationary objects

and moving objects. The expected error determines if the difference between the measured and predicted (sensor motion-induced) normal image velocity is significant. The seeding process is necessary to obtain the initial estimate of the sensor motion. The stabilization of the image sequence is obtained by dropping the inter-frame rotation terms when the model of sensor motion changes from predominantly rectilinear (for inter-frame sensor motion) to pure translation (for the Kalman filter). All of these contributions improve the robustness of the algorithm.

1.6 Outline of the Thesis

Chapter 2 reviews the technical prerequisites for performing obstacle detection using stereo cameras. The effect of object and sensor motion on the stereo image sequence is discussed. Chapter 2 also contains algorithms for determining the position and velocity of obstacles, and the velocity of the autonomous vehicle. These algorithms use local image displacements obtained from the stereo image sequence, such as stereo disparity and normal image velocity.

Chapter 3 discusses the Gabor representation and describes how it is used to measure the disparity and the normal image velocity. The description includes the selection of image features, the testing of feature stability with respect to viewpoint or motion-induced image deformations, and the testing of stereo and temporal correspondences. The phase-based refinements of the disparity and normal image velocity estimates are discussed. Expressions for the expected error in disparity and the expected error in normal image velocity are derived.

Chapter 4 discusses the obstacle detection algorithm and its various modules. The discussion includes: estimation of the inter-frame sensor motion; segmentation of the image sequence into stationary and moving object features; temporal integration of the

sensor and object translation; and estimation of the collision parameters. Implementation details, such as the seeding process and the generation of stereo and temporal correspondences, are addressed.

Chapter 5 contains the results of applying the obstacle detection algorithm to real stereo image sequences. Chapter 6 is the summary.

Chapter 2

Technical Prerequisites

The purpose of this chapter is to review the technical prerequisites for understanding stereo camera-based object and sensor motion estimation. The review begins with background information on image formation (section 2.1). In section 2.2, various coordinate systems, used to represent position and velocity, are defined. Sections 2.3 and 2.4 describe the image velocity as functions of object and sensor motion. Later sections discuss the inverse problem: estimating the motion of objects and the sensor from the image velocity field. The inverse problem involves estimating the depth from stereo images (section 2.5), estimating the localized time-to-collision from the stereo image velocity field (section 2.6), and estimating the sensor motion (section 2.7). Section 2.8 summaries the key concepts presented in this chapter.

2.1 Image Formation

This section briefly reviews image formation for a pinhole camera. Certain geometric and radiometric phenomena that can cause problems for image measurements are described.

A typical camera can be modelled as a pinhole camera. The projection geometry of a pinhole camera is shown in figure 2.2. In figure 2.2, a virtual image plane has been placed in front of the camera lens at $z_s = z_f$ for illustrative convenience. The actual image plane is behind the lens at $z_s = -z_f$. A camera is a passive sensor that measures the reflectance of a scene. The brightness pattern at the image plane is formed by rays of light traveling in straight lines from the surface of an object, through the pinhole, onto

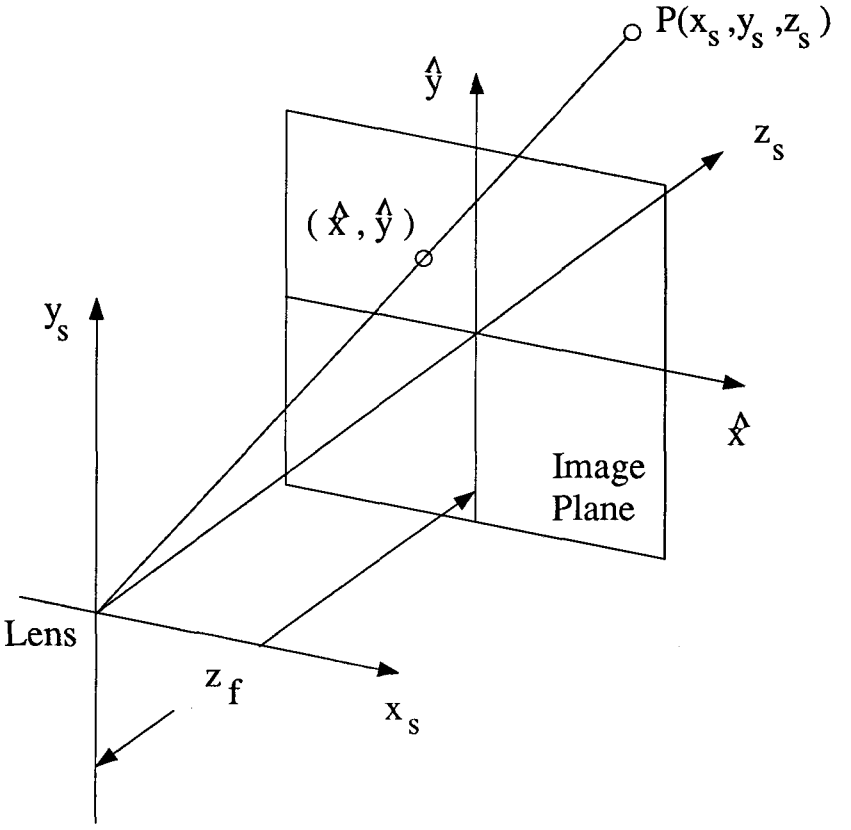


Figure 2.2: Projection Geometry of a Pinhole Camera

the image plane. The resulting two-dimensional image is a perspective projection of the three-dimensional scene.

Since the image formation process measures the reflectance along various rays passing through the pin-hole (or lens), an image is a function of both geometry and radiometry. Subsequent sections in this chapter deal exclusively with the geometric aspects of imaging. Errors will occur as a result of this geometric-only model of the camera. Certain radiometric and geometric phenomena will produce image brightness patterns that are

unstable with respect to changes in viewpoint. The unstable radiometric phenomena include specular reflections and shadows. Geometric instabilities can occur due to chance alignment of foreground and background features at surface discontinuities. Points on the occluding contour of curved objects are also unstable; the image outline of a curved object is dependent on the viewpoint as well as the surface shape.

The existence of these unstable image features can lead to errors in the local estimation of depth and motion. Thus, any camera-based method for estimating depth and motion should perform consistency tests on the measurement set to identify, and reject, unstable image features. The obstacle detection algorithm described in chapter 4 fulfills this requirement.

2.2 Coordinate Systems

The measurement of position and velocity must be preceded by the assignment of a coordinate system. In this work, five coordinate systems are used: two-dimensional image coordinates, three-dimensional scene coordinates, three-dimensional observer coordinates, three-dimensional vehicle coordinates, and three-dimensional world coordinates. The first two coordinate systems correspond to the sensor module. The remaining three coordinate systems—the observer, vehicle, and world coordinates—correspond to the obstacle avoidance, the local map, and the global map modules, respectively. Although a global map is not used in this work, the three-dimensional world coordinates are necessary for measuring the vehicle motion (both translational and rotational) and its subsequent effect on the image sequence.

2.2.1 Sensor Coordinates

There are two sensor coordinate systems: scene coordinates and image coordinates. The scene coordinate system represents the position of an object relative to the camera lens; the image coordinate system represents the position of the object's image projection relative to the center (origin) of the image plane. The scene coordinate system uses three-dimensional coordinates denoted by x_s, y_s, and z_s (see figure 2.2). The z_s-axis, also referred to as the "optical axis," is defined as the viewing direction. The other scene coordinates, x_s and y_s, represent the position along the horizontal and vertical axes, respectively. The image coordinates are given by $\hat{x}$ and $\hat{y}$. The origin of the image is the point at which the optical axis intersects the image. The image and scene origins are offset by the focal length z_f along the optical axis z_s.

The position of object can be represented using image coordinates or scene coordinates. The transformation of a point $P(x_s, y_s, z_s)$ from scene coordinates to image coordinates is given by

$$\hat{x} = x_s \frac{z_f}{z_s} \tag{2.10}$$

$$\hat{y} = y_s \frac{z_f}{z_s}. \tag{2.11}$$

A mixed coordinate system, given by $(\hat{x}, \hat{y}, z_s)$, defines the depth z_s at each image pixel.

2.2.2 Observer Coordinate System

The observer coordinate system represents the three-dimensional position of an object relative to a sensor group. A sensor group is configured such that each camera has a similar, but slightly different, view of a scene; that is, the angular separation between pairs of optical axes is small. If the observer coordinate axes are denoted by x, y, and z, the z-axis is chosen such that it is parallel or nearly parallel to each optical axis. The x-axis is defined to be tangential to the sensor group platform; the y-axis is normal to the

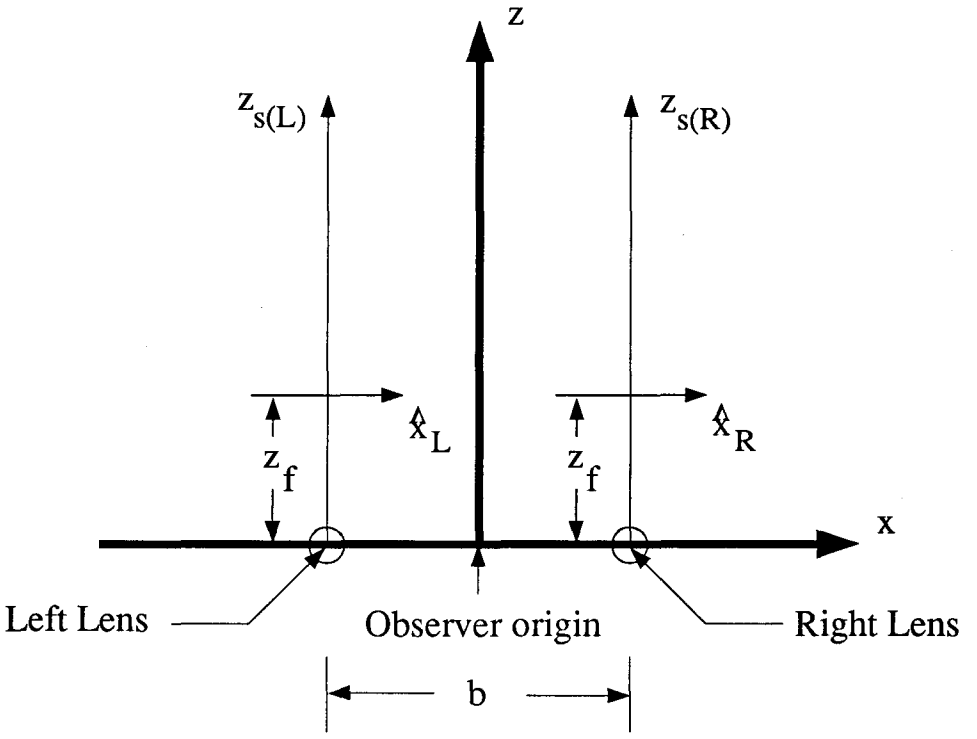


Figure 2.3: Stereo Camera Setup

platform. For the case of stereo cameras, the x-axis is parallel to the baseline separation of the camera.

A stereo camera setup is shown in figure 2.3. Consider the left camera. The scene coordinate origin for the left camera is offset (by $-\frac{b}{2}$) from the observer coordinate origin. In addition, there may be an angular offset between the observer coordinate frame and the left scene coordinate frame. The transformation from the observer coordinate frame to the left scene coordinate frame is given by

$$\begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} = R_{os(L)} \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \frac{b}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}, \tag{2.12}$$

where (small angle approximation)

$$R_{os(L)} = \begin{bmatrix} 1 & -\alpha_L & \beta_L \\ \alpha_L & 1 & -\gamma_L \\ -\beta_L & \gamma_L & 1 \end{bmatrix}^T , \tag{2.13}$$

and α_L, β_L, γ_L are the z, y, and x angular offset of the left scene coordinate frame relative to the observer coordinate frame. The rotational difference about the x, y, and z axes are referred to as tilt, pan, and roll, respectively. The angular offsets for the right camera are denoted by α_R, β_R, γ_R. The rotation matrix that comprises the angular offset for the right camera is denoted by $R_{os(R)}$

2.2.3 Vehicle Coordinate System

The vehicle coordinate system represents the three-dimensional position of an object relative to the vehicle. The vehicle coordinate axes are denoted by x_v, y_v, and z_v. The z_v-axis is defined as the vehicle heading. The x_v-axis is tangential to the floor of the vehicle; the y_v-axis is normal to the floor.

The relationship between the vehicle coordinates and the image coordinates is worth noting. If the scene is stationary and the vehicle and scene origins coincide, the intersection of the z_v-axis with the image plane is the "focus of expansion." The significance of the focus of expansion is discussed in section 2.4.

2.2.4 World Coordinate System

The world coordinate system is needed to estimate the motion of the vehicle. This three-dimensional Cartesian coordinate system is fixed relative to the ground surface. The world coordinate axes are denoted by x_w, y_w, and z_w.

2.3 Image Velocity and Scene Motion

This section defines the image velocity in terms of scene motion. Using the projection geometry of the camera, spatial shifts in image brightness patterns are predicted from changes in the position of viewed objects. The "aperture problem" associated with the apparent shift of a local brightness pattern is reviewed. The effect of sensor and object motion on the normal image velocity is discussed. Two important transformation vectors, used in chapter 4 to predict normal image velocity from sensor motion and object motion, respectively, are defined.

The motion of the sensor, and objects in its field of view, affects the brightness patterns in an image sequence. If the point $P(x_s, y_s, z_s)$ moves relative to the camera, the corresponding brightness pattern in the image sequence also moves (as shown in figure 2.4). This differential motion of the brightness pattern is referred to as the "image velocity." The $\hat{x}$ and $\hat{y}$ components of image velocity are denoted by $V_{\hat{x}}$ and $V_{\hat{y}}$, respectively. The differential motion of the point $P(x_s, y_s, z_s)$ is represented by the vector $[\dot{x}_s\ \dot{y}_s\ \dot{z}_s]^T$. The transformation from the relative motion of the point $P(x_s, y_s, z_s)$ to the image velocity is given by the matrix $A(z_s^{-1})$:

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = A(z_s^{-1}) \begin{bmatrix} \dot{x}_s \\ \dot{y}_s \\ \dot{z}_s \end{bmatrix}, \tag{2.14}$$

$$A = z_s^{-1} \begin{bmatrix} z_f & 0 & -\hat{x} \\ 0 & z_f & -\hat{y} \end{bmatrix}. \tag{2.15}$$

It is difficult to measure locally both the $\hat{x}$ and $\hat{y}$ components of image velocity. Consider a line contour viewed through a circular aperture, as shown in figure 2.5. If the contour moves, only motions normal to the line will produce shifts in the local brightness pattern (as viewed through a stationary aperture); motions along the line have no effect

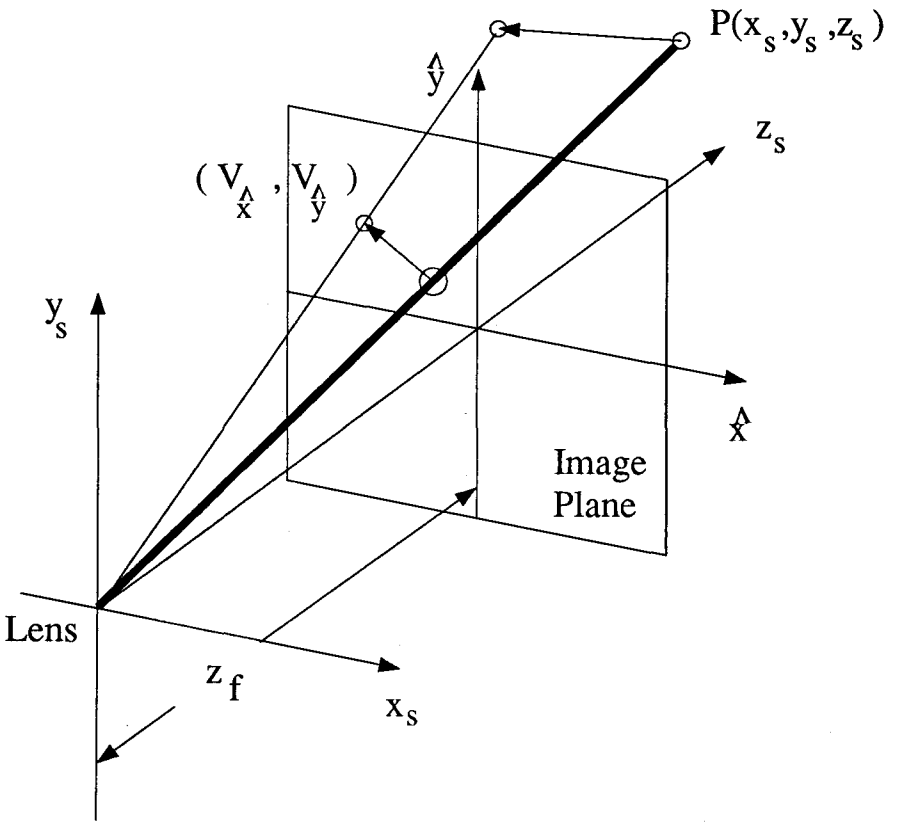


Figure 2.4: Image Projection of Motion

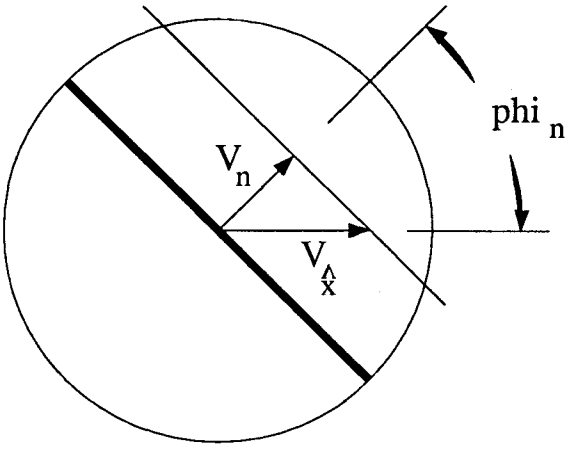


Figure 2.5: Aperture Problem

on the local image. This phenomena is referred to as the "aperture problem" [38] [54]. Thus, if the line moves, only the component of image velocity normal to the line can be measured. This component is referred to as the "normal image velocity" and is denoted by V_n. The normal direction of the line is measured relative to the $\hat{x}$-axis, and is denoted by ϕ_n. The transformation from the image velocity to the normal image velocity is given by the vector $\bar{n}$:

$$V_n = \bar{n}^T \begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix}, \tag{2.16}$$

where

$$\bar{n}^T = [\cos\phi_n \ \sin\phi_n]. \tag{2.17}$$

Image velocity is caused by motion of objects relative to the observer (sensor group). The motion of the sensor group, can be described by six parameters: three translational velocities, represented by the vector $\bar{T}$,

$$\bar{T} = [T_x \ T_y \ T_z]^T; \tag{2.18}$$

and three rotational velocities, represented by the vector $\bar{\Omega}$,

$$\bar{\Omega} = [\Omega_x \ \Omega_y \ \Omega_z]^T. \tag{2.19}$$

These velocities represent the instantaneous motion of the observer coordinate frame relative to the world coordinate frame. If the world coordinate frame is defined as the initial position of the observer coordinate frame, then T_x, T_y, T_z are the translational velocities of the observer origin along the x_w, y_w, z_w axes, and Ω_x, Ω_y, Ω_z are the rotational velocities about the x_w, y_w, z_w axes. The six parameters of instantaneous sensor motion [1] are represented by the vector $\bar{\theta}$:

$$\bar{\theta} = \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix}. \tag{2.20}$$

The motion of an object is described using three translational parameters localized about a point $P(x, y, z)$. The three translational velocities, measured with respect to the ground surface (world coordinate frame), are represented by

$$\bar{T}_{obj} = [\dot{x}_{obj} \ \dot{y}_{obj} \ \dot{z}_{obj}]^T. \tag{2.21}$$

The velocity of a point $P(x, y, z)$, with respect to the observer coordinate frame, is given by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = B(z)\bar{\theta} + \bar{T}_{obj}, \tag{2.22}$$

$$B(z) = \begin{bmatrix} -1 & 0 & 0 & 0 & -z & y \\ 0 & -1 & 0 & z & 0 & -x \\ 0 & 0 & -1 & -y & x & 0 \end{bmatrix}, \tag{2.23}$$

[1] This is more accurately described as the instantaneous observer motion or the instantaneous sensor group motion.

The matrix $B(z)$ transforms the instantaneous sensor motion into three translational parameters localized about the point $P(x, y, z)$.

If the previously mentioned matrices are combined, the normal image velocity, localized about the image coordinates $(\hat{x}, \hat{y})$, is given by

$$V_n(\hat{x}, \hat{y}) = \bar{J}^T \bar{\theta} + \bar{J}_{obj}^T \bar{T}_{obj}, \tag{2.24}$$

where $\bar{J}^T = \bar{n}^T A(z_s^{-1}) R_{os} B(z)$. and $\bar{J}_{obj}^T = \bar{n}^T A(z_s^{-1}) R_{os}$. The matrix R_{os} converts the observer velocity into the sensor coordinate frame. The transformation vectors $\bar{J}$ and $\bar{J}_{obj}$ convert the sensor motion and the object motion, respectively, into the normal image velocity. These transformation vectors are used in chapter 4.

From (2.22) and (2.24), it can be seen that the sensor and object motions have different effects on the image sequence. The sensor motion changes the position, with respect to scene coordinates, of every object within the scene. It will induce global (coherent) shifts in the brightness patterns throughout the image. For the case of an object, motion induces shifts that are localized to a small region of the image.

2.4 Image Velocity Field and Sensor Motion

The "image velocity field" is the set of image velocity vectors defined at each pixel in the two-dimensional image. This section discusses the effects of sensor motion, both translation and rotation, on the image velocity field. As mentioned in the previous section, the sensor motion has a global effect on the image velocity field.

If a point $P(x, y, z)$ is stationary with respect to the ground surface (world coordinate frame), it is possible to estimate the velocity of its image projection from the instantaneous sensor motion. The transformation of the instantaneous sensor motion into normal image velocity is given by

$$V_n(\hat{x}, \hat{y}) = \bar{J}^T \bar{\theta}. \tag{2.25}$$

In this section, the image velocity is being examined. The transformation of the instantaneous sensor motion into image velocity is given by

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = A(z_s^{-1}) R_{os} B(z) \bar{\theta}. \tag{2.26}$$

The transformation described by (2.26) uses two difference coordinate systems: observer coordinates in matrix $B(z)$; and mixed image/scene coordinates in matrix $A(z_s^{-1})$. It is useful to convert the observer coordinates of $B(z)$ into the mixed image/scene coordinates. The matrix $B(z)$ can be written as

$$B(z) = [B_T \; B_\Omega], \tag{2.27}$$

where $B_T = -I$ (I is the identity matrix) and

$$B_\Omega = \begin{bmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{bmatrix}. \tag{2.28}$$

The submatrix B_T is not a function of position, so it does not need to be modified. If the scene and observer origins are offset by $\frac{b}{2}$ along the x-axis, the matrix product $R_{os} B_\Omega(z)$ can be written as

$$R_{os} B_\Omega = \frac{z_s}{z_f} \begin{bmatrix} 0 & -z_f & \hat{y} \\ z_f & 0 & -\hat{x} \\ -\hat{y} & \hat{x} & 0 \end{bmatrix} R_{os} - \frac{b}{2} \begin{bmatrix} 0 & \beta & \alpha \\ -\beta & 0 & -1 \\ -\alpha & 1 & 0 \end{bmatrix} R_{os}. \tag{2.29}$$

Substituting (2.27) and (2.29) into (2.26), we get

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = \frac{z_f}{z_s} [-C(\hat{x}, \hat{y}) R_{os} \bar{T} - \frac{b}{2} E(\hat{x}, \hat{y}) R_{os} \bar{\Omega}] + D(\hat{x}, \hat{y}) R_{os} \bar{\Omega}, \tag{2.30}$$

where

$$C(\hat{x},\hat{y}) = z_f^{-1}\begin{bmatrix} z_f & 0 & -\hat{x} \\ 0 & z_f & -\hat{y} \end{bmatrix}, \tag{2.31}$$

$$D(\hat{x},\hat{y}) = C(\hat{x},\hat{y})\begin{bmatrix} 0 & -z_f & \hat{y} \\ z_f & 0 & -\hat{x} \\ -\hat{y} & \hat{x} & 0 \end{bmatrix}, \tag{2.32}$$

and

$$E(\hat{x},\hat{y}) = C(\hat{x},\hat{y})\begin{bmatrix} 0 & \beta & \alpha \\ -\beta & 0 & -1 \\ -\alpha & 1 & 0 \end{bmatrix}. \tag{2.33}$$

It can be seen that for a given image coordinate $(\hat{x}, \hat{y})$, the transformation from sensor motion to image velocity has only one unknown: the depth z_s.

Equation (2.30) contains two terms: the first term is normalized by the depth, z_s; and the second term is constant. The first term is the component of image velocity due to sensor translation; the second term is the component of image velocity due to sensor rotation. Both of these terms are measured using scene coordinates. Note that when there is an offset between the origins of the observer and scene coordinate frames (in this case $\frac{b}{2}$ along the x-axis), the observer rotation $\bar{\Omega}$ produces both translation and rotation in the scene coordinate system.

It is interesting to examine the effects of each sensor motion parameter on the image velocity field. In the following subsections, the effects of scene translation and rotation on image velocity are examined.

2.4.1 Translation

This subsection investigates how the image velocity field is affected by sensor translation (measured with respect to the scene coordinate frame). The focus of expansion is defined,

and its significance is discussed.

The image velocity due to scene translation is given by

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = -\frac{z_f}{z_s} C(\hat{x}, \hat{y}) \bar{T}_s. \tag{2.34}$$

where

$$\bar{T}_s = R_{os}\bar{T} + \frac{b}{2}\begin{bmatrix} 0 & \beta & \alpha \\ -\beta & 0 & -1 \\ -\alpha & 1 & 0 \end{bmatrix} R_{os}\bar{\Omega}. \tag{2.35}$$

When there is no rotation, $\bar{T}_s = [\dot{x}_s \ \dot{y}_s \ \dot{z}_s]^T$

The effects of the sensor translation on the image velocity field can be seen by examining (2.34). The x and y translations appear in the image as $\hat{x}$ and $\hat{y}$ image velocities: the direction of motion is the same, but the speed is reduced by a scale factor $-\frac{z_f}{z_s}$. The z translational motion causes an expansion (or contraction) of the velocity field about the origin of the image. The direction of the image velocity vectors radiate from (point towards) the origin. The speed increases with the radial distance from the image origin and decreases with the depth of the object.

If a camera is moving towards an object, all image velocity vectors will diverge from a point referred to as the "focus of expansion." Under the pure translation assumption, the focus of expansion is given by

$$\hat{x}_{foe} = z_f \frac{\dot{x}_s}{\dot{z}_s}, \tag{2.36}$$

$$\hat{y}_{foe} = z_f \frac{\dot{y}_s}{\dot{z}_s}. \tag{2.37}$$

The focus of expansion is not dependent on the depth of an object.

Equations (2.36) and (2.37) show that the focus of expansion is determined by the relative translational velocity of an object in the scene coordinate system. In general,

each moving object will have a different focus of expansion; only objects with common velocities will have a common focus of expansion. A camera translating through static environment is an example of a scene with one common focus of expansion. The focus of expansion for stationary objects is determined by the sensor translation; this special case is referred to as the "sensor's focus of expansion."

A number of researchers use the sensor's focus of expansion to extract properties of the sensor translation from the image velocity field. The direction of the image velocity at various image locations is used to estimate the sensor's focus of expansion, which is subsequently used to determine the direction of camera translation (see section 2.7). This estimate is obtained without depth information. Although the sensor's focus of expansion is useful for constraining the direction of translation when the depth is unknown, the focus of expansion becomes a singularity point if depth estimation from a known sensor motion is attempted (see section 2.5).

2.4.2 Rotation

This subsection investigates how the image velocity field is affected by sensor rotation. It is shown that rotations about the x_s- and y_s-axes produces an approximately constant offset in the image velocity field.

The image velocity due to sensor rotation, measured using scene coordinates, is given by

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = D(\hat{x},\hat{y})\bar{\Omega}_s, \tag{2.38}$$

where $\bar{\Omega}_s = R_{os}\bar{\Omega}$. The image velocity due to rotation about the x_s-axis (tilting the sensor) is given by

$$V_{\hat{x}} = \frac{\hat{x}\hat{y}}{z_f}\Omega_{x,s}, \tag{2.39}$$

$$V_{\hat{y}} = (z_f + \frac{\hat{y}^2}{z_f})\Omega_{x,s}. \tag{2.40}$$

Both the speed and direction of the $\Omega_{x,s}$-induced image velocity changes with the image coordinates. If the field of view of the camera is small ($z_f >> \hat{x}_{max}, \hat{y}_{max}$), the image velocity can be approximated by $(V_{\hat{x}}, V_{\hat{y}}) \approx (0, z_f)\Omega_{x,s}$. When this approximation is valid, the rotation about the x_s-axis contributes a constant flow throughout the image along the $\hat{y}$-axis. This constant depth-independent image velocity can be easily mistaken for a translation in the y_s direction.

The image velocity due to rotation about the y_s-axis is given by

$$V_{\hat{x}} = -(z_f + \frac{\hat{x}^2}{z_f})\Omega_{y,s}, \tag{2.41}$$

$$V_{\hat{y}} = -(\frac{\hat{x}\hat{y}}{z_f})\Omega_{y,s}. \tag{2.42}$$

If the field of view is small the image velocity can be approximated by $(V_{\hat{x}}, V\hat{y}) \approx (-z_f, 0)\Omega_{y,s}$. This constant image velocity can be mistaken for a translation in the x_s direction.

The image velocity due to rotation about the z_s-axis is given by

$$V_{\hat{x}} = \hat{y}\Omega_{z,s}, \tag{2.43}$$

$$V_{\hat{y}} = -\hat{x}\Omega_{z,s}. \tag{2.44}$$

The component of image velocity produced by $\Omega_{z,s}$ is orthogonal to that produced by T_z.

All of the above scene rotation-induced image velocities are independent of depth. Thus, rotation of a camera will not provide depth information.

2.4.3 Discussion

In the previous subsections, the effect of scene motion on the image velocity field has been discussed. It was shown that rotations about the x_s- and y_s-axes produce similar flow

patterns to translations in the y_s and x_s directions, respectively. The primary difference between the two flow patterns is that the translation induced pattern is dependent on depth. If the variation of depth within the scene is small, a given image velocity field can be produced (within a small error) by a set of translation-rotation combinations [3]. In such cases, the problem of estimating the three-dimensional sensor motion from the image velocity field is poorly conditioned (see section 4.6.3).

2.5 Estimating Depth

This section describes how stereo cameras are used to estimate depth. It is assumed that the optical axes of the stereo cameras are approximately parallel. A method of compensating for small convergence/divergence and differential tilt angles between the pair of cameras is discussed. This compensation transforms the configuration into an equivalent parallel stereo configuration. The remainder of the section addresses characteristics of the parallel stereo configuration. These characteristics include: the accuracy of the depth estimate, deformation of image features due to changes in viewpoints, and the sensitivity of depth estimates to the normal direction of an image feature.

Under certain conditions, it is possible to estimate the depth z_s using (2.30) if the sensor motion is known. The conditions are: that the viewed object must be stationary with respect to the ground surface (world coordinate frame); and its image velocity due to scene translation must be non-zero. The second condition is violated at the sensor's focus of expansion. In autonomous vehicle operations, the sensor is translating primarily along the z_s-axis, placing the focus of expansion near the origin of the image. Any depth measurements near the focus of expansion will be undefined or very sensitive to measurement error. Thus, forward translation along the z-axis is not a good motion direction for measuring depth [39] [40].

For the purpose of measuring depth, translational motion parallel to the image plane is preferred. Sensor motion along the x_s- or y_s-axis places the focus of expansion far from the image origin. Rather than continuously altering the course of the vehicle, tacking like a sailboat travelling upwind, the motion parallel to the image plane is simulated using matched stereo cameras that are offset along the x_s-axis. Stereo cameras can also measure the depth of moving objects if the cameras are synchronized using shutters. If the stereo images are obtained at the same time instant, all objects in the scene will appear stationary; that is, the effects of object motion disappear. Thus, synchronized stereo cameras, offset along the x_s-axis, fulfill the two above-mentioned conditions. In this section, extracting depth using stereo cameras is investigated.

Since the stereo vision system contains two imaging sensors, a three-dimensional observer coordinate system must be defined. Assume that the scene coordinate origins, for the left and right cameras, are separated by b. The baseline connecting the two scene origins is defined as the x-axis of the observer coordinate frame. The origin of the observer coordinate frame is defined as the midpoint between the scene origins. The z-axis is chosen to be nearly parallel to both optical axes. A typical stereo configuration is shown in figure 2.3.

A stereo vision system represents a three-dimensional scene by two image projections that differ in viewpoint. The apparent shift of an object due to changes in viewpoint is referred to as "disparity." Disparity information provided by stereo cameras is used to estimate the depth of an obstacle relative to the observer. Consider a point $P(x, y, z)$, whose position is defined using observer coordinates. The position of the point $P(x, y, z)$

using the mixed image/scene coordinates of the left camera, $P(\hat{x}_L, \hat{y}_L, z_{s(L)})$, is given by

$$\frac{z_{s(L)}}{z_f}\begin{bmatrix}\hat{x}_L\\ \hat{y}_L\\ z_f\end{bmatrix} = R_{os(L)}\left\{\begin{bmatrix}x\\ y\\ z\end{bmatrix} + \frac{b}{2}\begin{bmatrix}1\\ 0\\ 0\end{bmatrix}\right\}. \tag{2.45}$$

Similarly, the position using the mixed image/scene coordinates of the right camera, $P(\hat{x}_R, \hat{y}_R, z_{s(R)})$, is given by

$$\frac{z_{s(R)}}{z_f}\begin{bmatrix}\hat{x}_R\\ \hat{y}_R\\ z_f\end{bmatrix} = R_{os(R)}\left\{\begin{bmatrix}x\\ y\\ z\end{bmatrix} - \frac{b}{2}\begin{bmatrix}1\\ 0\\ 0\end{bmatrix}\right\}. \tag{2.46}$$

If it assumed that $R_{os(L)}$ and $R_{os(R)}$ are orthonormal and that the focal length z_f is the same for each camera, then

$$z_{s(L)}R_{os(L)}^T\begin{bmatrix}\hat{x}_L\\ \hat{y}_L\\ z_f\end{bmatrix} - z_{s(R)}R_{os(R)}^T\begin{bmatrix}\hat{x}_R\\ \hat{y}_R\\ z_f\end{bmatrix} = z_f b\begin{bmatrix}1\\ 0\\ 0\end{bmatrix}. \tag{2.47}$$

The two scene depths, $z_{s(L)}$ and $z_{s(R)}$, are obtained by solving the simultaneous equations.

In this work, since the stereo cameras are being used to simulate motion along the x-axis, the stereo cameras are configured in a nearly parallel setup; that is, the orientation of the scene coordinate frame for each camera is nearly the same as the orientation of the observer coordinate frame. In such a case, the small angle approximation of $R_{os(L)}$ and $R_{os(R)}$ can be used. Further simplifications are possible. The left and right coordinate rolls, α_L and α_R, represents the rotation of the respective image planes about their optical axis. This rotation can be set to zero by resampling each image such that the $\hat{x}$- and $\hat{y}$-axes in the stereo images are parallel. Thus, it can be assumed without loss of generality that α_L and α_R are both zero.

Using the small angle approximation and the zero roll assumption, (2.47) produces the following three equations:

$$z_{s(L)}(\hat{x}_L + \beta_L z_f) - z_{s(R)}(\hat{x}_R + \beta_R z_f) = z_f b, \tag{2.48}$$

$$z_{s(L)}(\hat{y}_L - \gamma_L z_f) - z_{s(R)}(\hat{y}_R - \gamma_R z_f) = 0, \tag{2.49}$$

$$z_{s(L)}(-\beta_L \hat{x}_L + \gamma_L \hat{y}_L + z_f) - z_{s(R)}(-\beta_R \hat{x}_R + \gamma_R \hat{y}_R + z_f) = 0. \tag{2.50}$$

When the tilt (β) and yaw (γ) angles are small and the field of view of each camera is small ($z_f >> \hat{x}_{max}$ or $\hat{y}_{max}$), it can be seen using (2.50) that $z_{s(L)} \approx z_{s(R)} \approx z$. Thus, (2.48) and (2.49) can be written as

$$\hat{x}_L - \hat{x}_R + (\beta_L - \beta_R) z_f = \frac{z_f}{z} b, \tag{2.51}$$

$$\hat{y}_L - \hat{y}_R = (\gamma_L - \gamma_R) z_f. \tag{2.52}$$

The disparity of a point $P(x, y, z)$ projected onto the left and right images in the $\hat{x}$ and $\hat{y}$ directions are respectively given by

$$d_{\hat{x}} = \hat{x}_L - \hat{x}_R, \tag{2.53}$$

$$d_{\hat{y}} = \hat{y}_L - \hat{y}_R. \tag{2.54}$$

The differential tilt and yaw between the left and right cameras are respectively given by

$$\Delta\beta = \beta_L - \beta_R, \tag{2.55}$$

$$\Delta\gamma = \gamma_L - \gamma_R. \tag{2.56}$$

The differential yaw angle is also referred to as the "vergence angle." Using the small angle approximation and the zero roll assumption, the depth is given by

$$z = \frac{z_f b}{d_x + \Delta\beta\, z_f}. \tag{2.57}$$

It can be seen that the depth of an object is dependent on the $\hat{x}$ disparity and the vergence angle, not on the absolute image coordinates or the absolute yaw angles. The $\hat{y}$ component of disparity is independent of depth:

$$d_{\hat{y}} = \Delta\gamma \; z_f. \tag{2.58}$$

For a given differential tilt (assuming small angle, zero roll), the disparity $d_{\hat{y}}$ is constant throughout the stereo image pair.

The parallel stereo configuration ($\Delta\beta = 0$ and $\Delta\gamma = 0$) is examined in the remainder of this section to provide a better understanding of the characteristic of a stereo image pair. The parallel stereo configuration is shown in figure 2.6. The position of the lens for the left and right cameras are denoted by O_L and O_R, respectively. Consider a point in the observer coordinate space: $P(x, y, z)$. The left and right lens positions and the point $P(x, y, z)$ form a triangle in the three-dimensional space. These three points also define a plane, referred to as the "epipolar plane" [5] [7] [8]. The projection of the point $P(x, y, z)$ onto the left and right images must lie on the epipolar plane. The intersection of the epipolar plane and the image plane defines the "epipolar line."

The epipolar plane can be defined by the two lens positions and a point in the left (or right) image plane. Once the projection of a point is identified in the left (right) image, the matching projection will be located on the right (left) image's epipolar line. Thus, the direction of the disparity is constrained by the camera configuration. This is referred to as the "epipolar constraint." For the parallel configuration shown in figure 2.6, the epipolar line is parallel to the x-axis.

Once matching projections (corresponding image features) are found in the left and right images, the position of the point $P(x, y, z)$ can be estimated using triangulation.

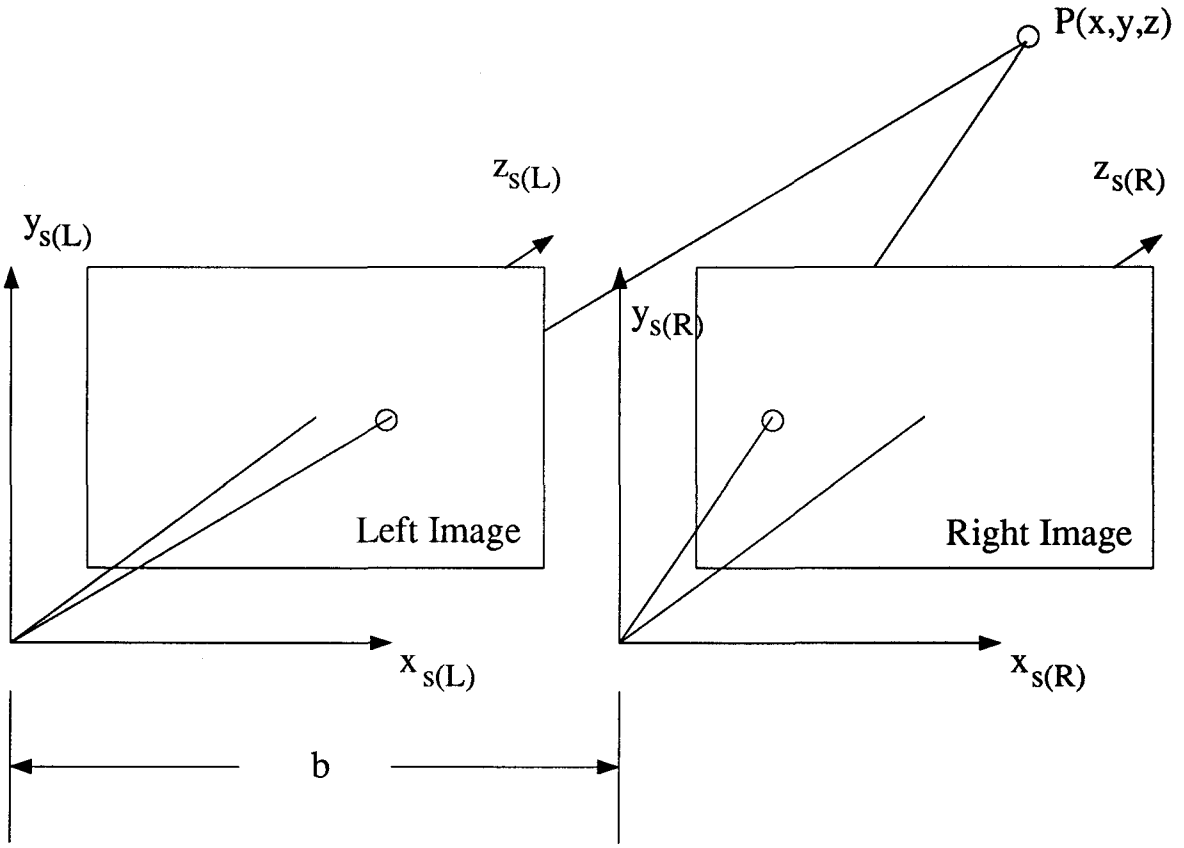


Figure 2.6: Point in Three-dimensional Space

The three-dimensional position, with respect to the observer coordinate frame, is calculated using [2]

$$x = \frac{b}{2}\frac{(\hat{x}_L + \hat{x}_R)}{d_{\hat{x}}}, \tag{2.59}$$

$$y = \frac{b}{2}\frac{(\hat{y}_L + \hat{y}_R)}{d_{\hat{x}}}, \tag{2.60}$$

$$z = \frac{z_f b}{d_{\hat{x}}}. \tag{2.61}$$

The accuracy of the three-dimensional position is dependent on the depth of the point, baseline separation and the focal length of the cameras, and the resolution of the disparity.

The resolution of the disparity affects the accuracy of depth estimates. The fractional error in the depth estimate is given by

$$\frac{\Delta z}{z} = -\frac{\Delta d_{\hat{x}}}{d_{\hat{x}} + \Delta d_{\hat{x}}}, \tag{2.62}$$

where $\Delta d_{\hat{x}}$ is the error in the $\hat{x}$ disparity estimate. If the error in the disparity estimate is fixed (say one pixel), the fractional error in depth is small for large disparities. The disparity increases as an obstacle approaches the cameras. Thus, the parallel stereo camera configuration estimates the depth of close objects with better accuracy than distance objects. Since the obstacle avoidance module usually assigns a high urgency (as defined in section 1.2) to close objects, the improved close depth accuracy is an asset. The disparity, and the accuracy of the depth estimate, can be increased by increasing the baseline separation or the focal length.

There are a number facts that must be considered when choosing a camera focal length or baseline separation. The field of view is affected by the baseline separation and the focal length. The field of view for a single camera is defined by the size of the image plane and the focal length. For a fixed size image plane, the monocular field of

[2]These equations are valid for the parallel stereo configuration. For nearly parallel configurations, the disparity $d_{\hat{x}}$ must be replaced by $d_{\hat{x}} + \Delta\beta\, z_f$.

view contracts as the focal length increases. The stereo field of view is the intersection of the left and right monocular fields of view. Increasing the baseline separation or the focal length contracts the stereo field of view.

Increasing the baseline separation makes the stereo images less similar. Due to differences in viewpoint, an object is projected differently onto the two images. As a result, establishing stereo correspondences becomes more difficult. Consider as an example a flat surface marked with periodic vertical stripes (see figure 2.7). If the projection of the surface normal onto the x-z plane is not parallel to the z-axis (and the optical axes of the cameras), the frequency of the pattern in the left and right images will be different. The difference in frequency at the two image projections of $P(x, y, z)$ is given by

$$\delta f = f_{ave} b \frac{(\frac{\delta z}{\delta x})}{z - x(\frac{\delta z}{\delta x})}, \tag{2.63}$$

where $(\frac{\delta z}{\delta x})$ is the slope of the surface in the x-z plane relative to the x-axis, and f_{ave} is the average frequency of the pattern as viewed by the left and right cameras. The fractional difference in frequency $(\frac{\delta f}{f_{ave}})$ increases with the baseline separation and the slope of the surface. Note that the difference in frequency becomes more pronounced as the obstacle approaches the cameras. Since the difference is larger for close obstacles, obstacle detection is hindered by this effect [3].

Similar to the image velocity measurements, the measurement of depth using stereo is subject to the aperture problem. The $\hat{x}$ component of disparity can be best measured at vertical image features ($\phi_n \approx 0$). If the normal component of disparity is denoted by d_n, then the $\hat{x}$ component of disparity is given by

$$d_{\hat{x}} = \frac{d_n}{\cos \phi_n}. \tag{2.64}$$

[3]The problems associated with the frequency differences are less significant for the Gabor filter-based method, presented in chapter 3, than pixel correlation methods, such as [43].

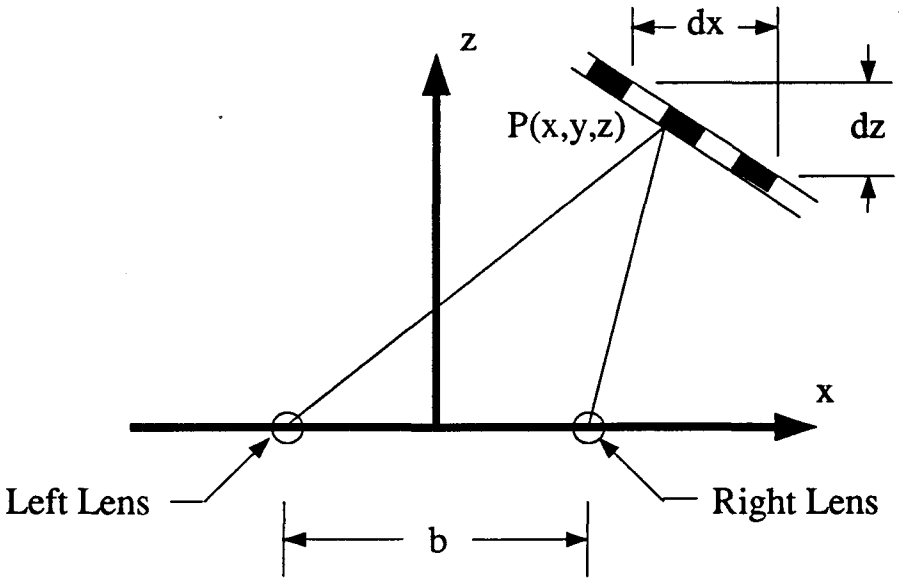


Figure 2.7: Viewing an Inclined Surface

It can be seen from (2.64) that the depth of horizontal image features ($\phi_n = \frac{\pi}{2}$) can not be measured using parallel stereo cameras whose baseline separation is along the x-axis.

The previous paragraphs have addressed three problems that are dependent on: the choice of focal length and baseline separation; and the normal direction of the image feature. Assume that the focal length is fixed. Increasing the baseline separation increases the accuracy of the depth estimate but makes stereo correspondence more difficult. This tradeoff can be eliminated if extra cameras are placed along the baseline. If both the vehicle and the scene are stationary, this collinear camera configuration can be simulated by moving the camera along a sliding mount [39]. The sensitivity of depth estimation to the normal direction of an image feature can be reduced by adding extra cameras that form noncollinear baselines [6]. Consider the case where a third camera is positioned such that one pair of cameras has a vertical baseline separation and another pair of cameras has a horizontal baseline separation. Since the epipolar lines in the horizontal

and vertical camera pairs are orthogonal, the depth of both horizontal and vertical image features can be estimated. In this work, only the binocular stereo camera setup is used.

2.6 Stereo Image Velocity

In the previous section, it was shown that stereo images can be used to estimate depth. In this section, the stereo image velocity field is used to produce local estimates of the time to collision and the velocity $\dot{z}$. The stereo camera setup is as follows: the baseline separation is given by b, and the orientation of the cameras relative to the observer coordinate frame is given by $(0, \beta_L, \gamma_L)$ and $(0, \beta_R, \gamma_R)$. It is assumed that the angular offsets are small enough that the small angle approximation introduced in section 2.5 is valid, and that the epipolar line is approximately parallel to the $\hat{x}$-axis. The time to collision is estimated using the difference of image velocity vectors at corresponding points in the left and right images.

The image velocity due to sensor motion is given by (2.26). For small angles, the left image velocity can be approximated as

$$\begin{bmatrix} V_{\hat{x},L} \\ V_{\hat{y},L} \end{bmatrix} = z^{-1} \begin{bmatrix} z_f & 0 & -(\hat{x}_L + z_f\beta_L) \\ 0 & z_f & -(\hat{y}_L - z_f\gamma_L) \end{bmatrix} [B(z)\bar{\theta} + \bar{T}_{obj}]. \tag{2.65}$$

A similar expression exists for the right image. The matrix $B(z)$ is defined in terms of observer coordinates. The observer coordinates can be expressed in terms of the mixed image/scene coordinates for the left image:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \approx \frac{z}{z_f} \begin{bmatrix} \hat{x}_L + z_f\beta_L \\ \hat{y}_L - z_f\gamma_L \\ z_f \end{bmatrix} - \frac{b}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{2.66}$$

The expression for the right image is given by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \approx \frac{z}{z_f} \begin{bmatrix} \hat{x}_R + z_f \beta_R \\ \hat{y}_R - z_f \gamma_R \\ z_f \end{bmatrix} + \frac{b}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{2.67}$$

Substituting (2.66) and (2.67) into $B(z)$, the difference between the image velocity along the $\hat{x}$-axis at corresponding image points is given by [51]

$$V_{\hat{x},L} - V_{\hat{x},R} = (d_{\hat{x}} + \Delta\beta\ z_f)[t_{col}^{-1} + \frac{\hat{y}_{st}}{z_f}\Omega_x - \frac{\hat{x}_{st}}{z_f}\Omega_y], \tag{2.68}$$

where

$$t_{col} = \frac{z}{T_z - \dot{z}_{obj}} = -\frac{z}{\dot{z}}, \tag{2.69}$$

$$\hat{y}_{st} = \frac{\hat{y}_L + \hat{y}_R - z_f(\gamma_L + \gamma_R)}{2}, \tag{2.70}$$

$$\hat{x}_{st} = \frac{\hat{x}_L + \hat{x}_R + z_f(\beta_L + \beta_R)}{2}. \tag{2.71}$$

It is often the case that the field of view of the camera is small and $R_{os(L)} = R_{os(R)}^T$; that is $z_f >> \hat{x}_{max}, \hat{y}_{max}$, $\beta_L = -\beta_R$ and $\gamma_L = -\gamma_R$. In such a case,

$$t_{col} \approx \frac{d_{\hat{x}} + \Delta\beta\ z_f}{V_{\hat{x},L} - V_{\hat{x},R}}. \tag{2.72}$$

If the depth is known, the object translation along the z-axis relative to the sensor is given by

$$\dot{z} = -\frac{z}{t_{col}}. \tag{2.73}$$

The localized $\dot{z}$ can be used to segment the image into projections of moving objects and stationary objects. Since $\bar{T}_{obj} = 0$ for a stationary object, a necessary condition for a stereo image feature to belong to a stationary object is $\dot{z} \approx -T_z$. Thus, any stereo image feature with a $\dot{z}$ significantly different than $-T_z$ is processed as belonging to a moving object [4].

[4]The sensor velocity T_z can be estimated from the extended sensor motion (section 4.4), if it is available. T_z can also be approximated by the vehicle speed if the vehicle is travelling along the z-axis.

Another local measurement, namely depth, can be used to verify this image velocity-based estimate of $\dot{z}$. The change in depth at corresponding points in the temporal sequence can also be used to estimate $\dot{z}$. If there is a significant difference between the image velocity-based and the depth-based estimates, then either an unstable point has been chosen or a correspondence error (stereo or temporal) has occurred. Such an estimate can be rejected.

The above mentioned segmentation scheme is one step used to segment the image. A more detailed description of the segmentation (seeding) process appears in chapter 4.

2.7 Estimating Three-dimensional Motion

The previous sections described how the image velocity field is generated by the three-dimensional motion of an observer in a rigid environment. This section reviews various techniques for solving the inverse problem: estimating the three-dimensional motion from an image velocity field.

The image velocity can be calculated directly from the three-dimensional motion parameter, as shown in (2.74):

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = -\frac{z_f}{z_s} C(\hat{x},\hat{y})\bar{T} + D(\hat{x},\hat{y})\bar{\Omega}. \tag{2.74}$$

The inverse problem can be solved if there is a degree of coherence in the image velocity field. If rigidity assumptions are valid, and all objects in the scene are moving at a common velocity [5], the three-dimensional motion parameters can be estimated by minimizing the difference between the measured image velocity field and the field that would be induced by the three-dimensional motion. If an image velocity field (containing

[5] A group of stationary objects fulfill both the rigidity and common velocity assumptions.

N flow vectors) is available, the least square solution is obtained by minimizing

$$r_{iv} = \sum_{i=1}^{N} \bar{e}_i^T \bar{e}_i \tag{2.75}$$

where

$$\bar{e}_i = \begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix}_i + \frac{z_f}{z_s} C(\hat{x}, \hat{y}) \bar{T} - D(\hat{x}, \hat{y}) \bar{\Omega}. \tag{2.76}$$

If only normal image velocity measurements are available, the least square solution is obtained by minimizing

$$r_{niv} = \sum_{i=1}^{N} e_i^2, \tag{2.77}$$

where

$$e_i = V_n(i) - \bar{J}^T \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix}. \tag{2.78}$$

The summations in the above equations include selected points in the image velocity field (or set of normal image velocities) that are believed to belong to stationary objects. The selection is done by pre-segmenting the velocity field or by assuming that all objects are stationary. The accuracy of any assumptions is indicated by the size of the residual, r_{iv} or r_{niv}: a large residual indicates errors.

The image velocity is a non-linear function of depth, translation, and rotation. Most methods for determining the three-dimensional motion attempt to create a linear problem by assuming (or measuring) the translation, or by assuming (or measuring) the depth. Methods based on assumed values use the residual to determine the "goodness of fit" to the measured image velocity field.

Many image velocity-based methods use only one camera to estimate sensor motion. Monocular solutions, however, can only determine the translation and the depth up to a scale factor. From (2.74), it can be seen that an image velocity field produced by a given translation and depth is identical to the image velocity field produced by doubling both

the translation and the depth. As a result, monocular implementations attempt to find the direction of translation or the structure of the scene (the depth of an object relative to its neighbours).

A number of researchers use the Hough transform to determine the direction of translation [2] [12] [28]. The space of possible directions of translation is sampled. The residual for each candidate direction is calculated by the implicit (using complementary subspaces) selection of a sensor rotation and scene structure that best fits the measured image velocity field. The direction of translation (or set of directions) producing the smallest residual (or near smallest set) is accepted as the actual direction. Once the direction of translation is specified, the the rotation is determined by examining the portion of the image velocity field that is orthogonal to the translation-induced field. The direction of each image velocity vector is required to obtain the orthogonal component. Thus, the image velocity field must be estimated; the normal image velocity measurements are not sufficient.

The alternative approach is to assume or measure the depth of the scene. If the depth is known, the three-dimensional motion can be solved using a least square solution that is based on the normal image velocity measurements [31] [33]. If only the parametric form of the scene structure is known, it is possible to iterate to a solution. Parametric structures include planar and quadratic surfaces. For the case of planar scenes, the surface normal is iteratively updated until a good fit with the set of normal image velocities is obtained [45]. For quadratic surfaces, a set of surface parameters are selected as starting points for the iterative process [33]. All of the known depth methods use a set of normal image velocities instead of the image velocity field. Approaches that use the normal image velocity are referred to as "direct methods," because the image velocity field is not estimated as an intermediate stage.

It can be seen from the above discussion that there is a tradeoff: either the depth of the

scene must be measured, or the image velocity field must be estimated. In this work, the known depth case of the direct methods is used. The stereo cameras are used to measure the depth of selected features. The stereo correspondence problem exists, and it is often cited as a motivation for choosing the monocular approaches. Although correspondence errors are possible, the effect of such errors can be reduced using error estimation and outlier testing. The known depth method and the normal image velocity-based outlier test (Mahalanobis distance) are described in chapter 4.

2.8 Summary

This chapter has reviewed various aspects of stereo camera-based motion estimation. The two most important aspects are: the geometric properties of image formation for the case of a moving object and a moving sensor; and how to exploit these properties to estimate the object and sensor motion. It was shown that the sensor motion induces global changes to the image velocity field and that object motion induces localized changes to the image velocity field. Localized parameters, such as the depth and the time-to-collision, are used in the segmentation process to identify moving objects. After segmenting the image into moving and stationary objects, image measurements (normal image velocity and disparity) belonging to stationary objects are combined to obtain an estimate of the sensor motion.

The stereo image sequence provides sufficient information to perform obstacle detection. The stereo images provide the depth that is required for direct sensor motion estimation. With the depth information, there in no scale ambiguity in the sensor translation estimate that is typically associated with monocular camera systems. Even if the sensor motion is known, stereo cameras provide better depth estimates than the monocular counterpart. The forward motion of a monocular camera system produces a

singularity for depth estimation at the focus of expansion. In addition, monocular camera systems are not suited for concurrent estimation of depth and object motion. Concurrent estimation is not a problem for stereo configurations. Finally, the stereo image velocity fields provide local quantities that are used to identify moving and stationary objects; they provide the initial segmentation of the normal image velocity set. Thus, the stereo image sequence makes estimation of three-dimensional sensor motion simple and reliable, as well as providing the local information needed to estimate the object motion.

Chapter 3

Measuring Normal Image Velocity and Disparity

In this chapter, techniques for measuring normal image velocity and disparity from a stereo image sequence are discussed. The basis of these techniques is the Gabor filter whose magnitude and phase responses are used to extract interesting features, estimate disparity, and estimate normal image velocity. Before proceeding with the Gabor-based methods, a general overview of measuring image properties is presented.

3.1 General Overview

This section reviews the general requirements for measuring normal image velocity and disparity. The characteristics of image features that are suitable for measuring normal image velocity or disparity are discussed. An approach for establishing the correspondence of image features over time and viewpoint is presented. A combined feature matching-phase gradient approach to measure normal image velocity and disparity is recommended.

3.1.1 Interesting Image Features

An image region comprises a group of pixel intensities whose spatial distribution forms a brightness pattern. This subsection discusses different types of brightness patterns: omni-directional features, uni-directional features, and textureless regions. The omni-directional and uni-directional features are suitable references for measuring image displacement; the textureless regions provide no motion information. These brightness patterns are characterized by the magnitude and distribution of spectral energy within

the image region.

A camera image is a projection of a physical scene. The physical scene contains changes in surface properties (such as depth, texture, reflectance) that are projected as brightness patterns onto the image. Both the brightness pattern and its physical source (changes in surface properties) are referred to as a "feature." If there is a need to discriminate between the two types of features, the brightness pattern will be referred to as an "image feature;" the change in surface property will be referred to as a "physical feature."

Interesting features are selected from the stereo image sequence to measure the position and velocity of objects. For the purpose of obstacle detection, an interesting feature is a distinct region on an object that is easily found in images. The interesting feature should be detectable in a set of images that differ slightly in time or viewpoint; that is, the feature must be stable with respect to image deformations caused by sensor/object motion or by the stereo camera separation. These features become references that can be used to measure inter-frame displacements over time or viewpoint.

A region in an image must have some variation in pixel intensity in at least one direction to be distinct enough to be a feature. A good feature has significant intensity variations in orthogonal directions. Directional intensity variations can be measured using information from the spectral (Fourier) domain. By applying a two-dimensional Fourier transform to a local region of an image, the frequency and orientation of the dominant modes of intensity variation are made explicit. In this work, the dominant modes within a local region are made explicit using a Gabor representation of an image.

A region must have multiple modes, containing at least two (preferably orthogonal) orientations, to unambiguously resolve local motion. These regions (corners, crosses, T-junctions) are referred to as "omni-directional features." Regions containing a single line, edge, or grating can not fully resolve local motion (aperture problem). These types of

regions have a single dominant spectral orientation, and are referred to as "uni-directional features." Uni-directional features are defined in the frequency domain in terms of the dominant spectral orientation. Note that the dominant spectral orientation of a uni-directional feature is the same as the normal direction ϕ_n. Regions without intensity variations can not be used as references. These textureless regions do not contain enough spectral energy to locally resolve motion.

In this work, a set of oriented bandpass filters, known as Gabor filters, is used to highlight uni-directional features (see section 3.2.2). A given image brightness pattern may appear in the output of a number of bandpass (Gabor) channels. In this work, multiple channel responses, produced by a wide bandwidth brightness pattern, are treated as separate uni-directional features.

3.1.2 Measuring Feature Displacement

This subsection discusses the measurement of feature displacement. The requirements of reliable feature correspondence are addressed. The gradient constraint equation, used to measure small displacements, is introduced. A combined feature matching-phase gradient approach is proposed.

For a continuous image sequence, normal image velocity is the differential motion of a brightness pattern. In this work, the image sequences are discrete; thus, inter-frame displacements are measured instead of image velocities. Both disparity and normal image velocity can be described as inter-frame displacements. Disparity is the displacement of a feature over viewpoint; normal image velocity is estimated from the displacement of a feature over time.

To measure inter-frame displacement, a correspondence between features in each image must be formed. Once an interesting feature is identified in a given image, the matching feature must be found within the other image. A search of the set of features

will produce many potential matches. A method or set of criteria is needed to reduce the number of potential matches to one or zero. No match can be made if the feature is viewed by one image only.

A method of reducing the number of potential matches is to provide a "rich" description of the feature. In this work, local attributes are assigned to a feature. Certain local attributes are stable with respect to changes in viewpoint or sensor/object motion: they include local magnitude and normalized moment of inertia (see section 3.2.2). Candidate matches with significantly different local attributes are rejected.

A priori information can be used to reduce the number of potential matches. A priori information can be obtained from other Gabor channels, from past measurements, or from spatial constraints (see section 4.6.1). This information is used to predict the image position of the corresponding feature. A limited search around the predicted position is performed. The feature with the most similar local attributes is considered the corresponding feature.

Although both normal image velocity and disparity measurements are based on inter-frame displacements; the two cases differ. The size of the inter-frame displacement for a normal image velocity measurement is typically small, but the direction is unknown. For disparity measurements, the reverse is true: the inter-frame displacement is typically large, but the direction of displacement is known (due to the epipolar constraint). The remainder of this section will address these two cases.

When the inter-frame displacement of a feature is small, gradient-based techniques can be used. The gradient-based approach uses the partial derivatives of local image measurements to constrain image velocity field [5]:

$$e_g = \frac{\delta c}{\delta \hat{x}} V_{\hat{x}} + \frac{\delta c}{\delta \hat{y}} V_{\hat{y}} + \frac{\delta c}{\delta t} = 0, \tag{3.79}$$

where c is a candidate function based on local image measurements. Possible candidate

functions include intensity, contrast, entropy, average, and directional derivatives [5]. In this work, the local phase of a bandpass (Gabor) filtered image is used as the candidate function.

The partial derivatives of phase are equivalent to the local image frequencies in each spatial/temporal direction:

$$\omega_{\hat{x}} = \frac{\delta c}{\delta \hat{x}}, \tag{3.80}$$

$$\omega_{\hat{y}} = \frac{\delta c}{\delta \hat{y}}, \tag{3.81}$$

$$\omega_t = \frac{\delta c}{\delta t}. \tag{3.82}$$

The spatial phase derivatives, $\omega_{\hat{x}}$ and $\omega_{\hat{y}}$, can be written in terms of the normal image direction $\bar{n}$:

$$[\omega_{\hat{x}}\ \omega_{\hat{y}}] = \omega_n \bar{n}^T, \tag{3.83}$$

where

$$\omega_n = [\omega_{\hat{x}}^2 + \omega_{\hat{y}}^2]^{0.5}. \tag{3.84}$$

From (3.79) and (3.83), it can be seen that the normal image velocity, speed and direction, is given by [22] [23]

$$V_n = -\frac{\omega_t}{\omega_n}, \tag{3.85}$$

$$\cos\phi_n = \frac{\omega_{\hat{x}}}{\omega_n}, \tag{3.86}$$

$$\sin\phi_n = \frac{\omega_{\hat{y}}}{\omega_n}. \tag{3.87}$$

The basis of the gradient method is the constraint equation (3.79). Equation (3.79) is derived from the assumption that the candidate function c at the image point $(\hat{x},\hat{y})$ and time t has the same value after it moves to the image point $(\hat{x}+\Delta\hat{x},\ \hat{y}+\Delta\hat{y})$ at time $t+\Delta t$; that is

$$g(\hat{x}+\Delta\hat{x}, \hat{y}+\Delta\hat{y}, t+\Delta t) = g(\hat{x},\hat{y},t). \tag{3.88}$$

If the left hand side of (3.88) is approximated by a first-order Taylor expansion about $(\hat{x}, \hat{y}, t)$, a variation of (3.79) is obtained:

$$\frac{\delta c}{\delta \hat{x}}\Delta\hat{x} + \frac{\delta c}{\delta \hat{y}}\Delta\hat{y} + \frac{\delta c}{\delta t}\Delta t \approx 0. \tag{3.89}$$

The accuracy of (3.89) depends on the size of the higher order terms in the Taylor expansion. The higher order terms are dependent on the characteristics of the candidate function to motion-induced deformations, such as image translation, image dilation, and image rotation. In this work, the phase is chosen as the candidate function because it is stable with respect to image deformations associated with sensor/object motion [23].

The phase gradient approach is designed to measure small motions. The maximum measurable image displacement, $V_n\ \Delta t$, is limited by phase wraparound. Since the phase is modulo 2π, the phase difference is restricted:

$$-\pi < \Delta\theta \leq \pi. \tag{3.90}$$

Thus, the measurable image displacement is limited to

$$|V_n\ \Delta t| < \frac{|\Delta\theta_{max}|}{\omega_n} = \frac{\pi}{\omega_n}. \tag{3.91}$$

Features undergoing larger image motions will be aliased. Even if the phase is unwrapped, the motion between successive images may be larger than the spatial extent of the filter. To accommodate large motions, the spatial portion of the filter must be moved with the feature. If the spatial portion is moved along a predicted trajectory, the phased-based image displacement would measure the prediction error. In this work, large (but coarse) motion is obtained using a feature matching approach; the residual error is estimated using a gradient-based approach. The magnitude response of a bandpass (Gabor) filtered image is used for feature matching, and the phase response is used in the gradient approach. This approach is referred to as a "combination method" because it uses both feature matching and gradient-based techniques.

The combination method is ideal for measuring disparity in stereo images. The coarse estimate of disparity produces an offset along the epipolar line that effectively aligns the left and right images. This "epipolar offset," denoted by E_{offset}, tunes the stereo cameras to a preferred depth [1]:

$$z_o = \frac{z_f b}{E_{offset}}. \tag{3.92}$$

The residual error, measured using the phase gradient approach, is referred to as the "relative disparity." The relative disparity, denoted by d_{rel}, is obtained using the following constraint:

$$d_{rel}\omega_{\hat{x}} + (\theta_R - \theta_L) = 0, \tag{3.93}$$

where θ_R and θ_L are the local phase at corresponding points in the right and left images, respectively. The measured disparity is given by

$$d_{\hat{x}} = E_{offset} + d_{rel}. \tag{3.94}$$

3.2 Using the Gabor Representation to Process Images

This section introduces the Gabor representation. It shows how the local magnitude information is used to select interesting image features, and how local phase differences are used to estimate local spatial frequency. The local magnitude and phase are also used to implement a combined feature matching-phase gradient approach for extracting both disparity and normal image velocity.

3.2.1 Gabor Representation

This subsection discusses aspects of the Gabor representation: the Gabor function, the log-polar set of Gabor filters, and the spatial sampling lattices. It also discusses how the

[1] The epipolar offset only alters the relative position of the origin for the left and right images; it does not affect the the baseline separation or the viewing direction of the cameras.

local magnitude and phase are measured.

The Gabor representation is a joint position-frequency representation of an image. There are two commonly used versions: the Gabor expansion [9] [14] [15] and the Gabor filter [1] [20] [27] [50]. The expansion decomposes an image into a weighted sum of Gabor functions; the filter method convolves an image with a set of Gabor functions. The Gabor filter method is used in this work.

Gabor Function

The following paragraphs describe the Gabor function: the kernel used in a Gabor filter. The spatial and spectral characteristics of the Gabor function are discussed. The uncertainty constraint, which illustrates the trade-off with respect to spatial and spectral localization, is reviewed.

The basis of the representation is the Gabor function [13] [24]. In two-dimensions, it is a frequency-modulated elliptical Gaussian window given by

$$G_i = g(\acute{x}, \acute{y}) \cos[u\hat{x} + v\hat{y} + p], \tag{3.95}$$

where

$$g(\acute{x}, \acute{y}) = \exp -\pi[(\frac{\acute{x}}{\sigma})^2 + (\frac{\acute{y}}{\alpha\sigma})^2], \tag{3.96}$$

$$(\acute{x}, \acute{y}) = (\hat{x}\cos\phi_G - \hat{y}\sin\phi_G, \hat{x}\sin\phi_G + \hat{y}\cos\phi_G). \tag{3.97}$$

The variables u and v represent the frequency of the modulating wave in the $\hat{x}$ and $\hat{y}$ directions, respectively. The phase of the modulating wave relative to the spatial centroid of the Gaussian window is denoted by p. $(\acute{x}, \acute{y})$ represents a position in a rotated coordinate system; ϕ_G is angle of rotation for the elliptical Gaussian window relative to the $\hat{x}$-axis. The final two variables, α and σ, are the aspect ratio and the scale of the elliptical Gaussian window. In this work, the frequency variables u,v are defined in polar

form:

$$\tilde{\omega} = (u^2 + v^2)^{0.5}, \tag{3.98}$$

$$\tilde{\phi} = \arctan[\frac{v}{u}], \tag{3.99}$$

where $\tilde{\omega}$ and $\tilde{\phi}$ are the frequency and orientation, respectively, of the modulating wave. In addition, the rotation of the elliptical Gaussian window is constrained such that one of the principle axes has the same orientation of the modulating wave ($\phi_G = \tilde{\phi}$).

Consider the case of a Gabor function whose spectral orientation is along the $\hat{x}$-axis ($\tilde{\phi} = 0$). The Gabor function and its Fourier transform are shown in figure 3.8. The window of the Gabor function has a Gaussian shape in both domains; however, only one contour is shown in figure 3.8 for simplicity. Note that the Gabor function has a finite frequency and orientation bandwidth. As a result, the Gabor filter can only extract information within the passband centered about $\tilde{\omega}_k$ and $\tilde{\phi}_l$. The passband for quadrature Gabor filters is referred to as the "Gabor channel," and it is described by its center frequency, $\tilde{\omega}$ and $\tilde{\phi}$.

With respect to energy, the Gabor function is localized in both the spatial domain and frequency domain [24] [13]. The effective frequency bandwidth $\Delta\tilde{\omega}$ and orientation bandwidth $\Delta\tilde{\phi}$ are shown in figure 3.8. These bandwidths are given by

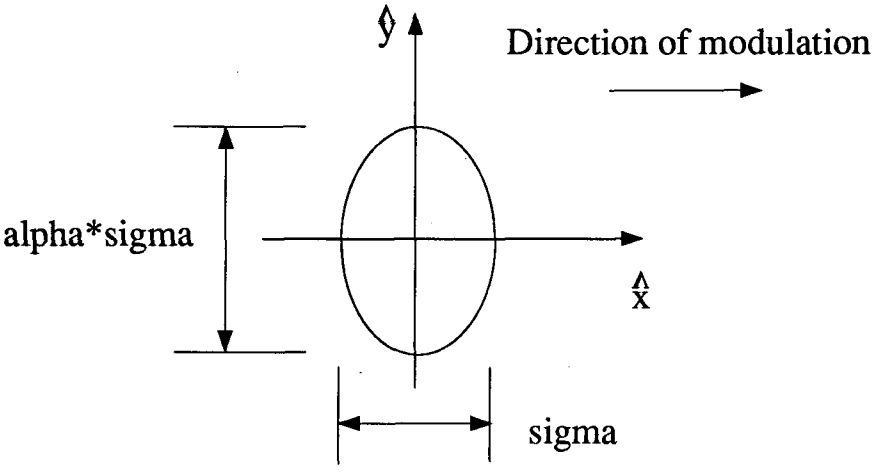
$$\Delta\tilde{\omega} = \frac{2\pi}{\sigma}, \tag{3.100}$$

$$\Delta\tilde{\phi} = 2\arctan[\frac{\pi}{\alpha\sigma\tilde{\omega}}] \approx \frac{2\pi}{\alpha\sigma\tilde{\omega}}. \tag{3.101}$$

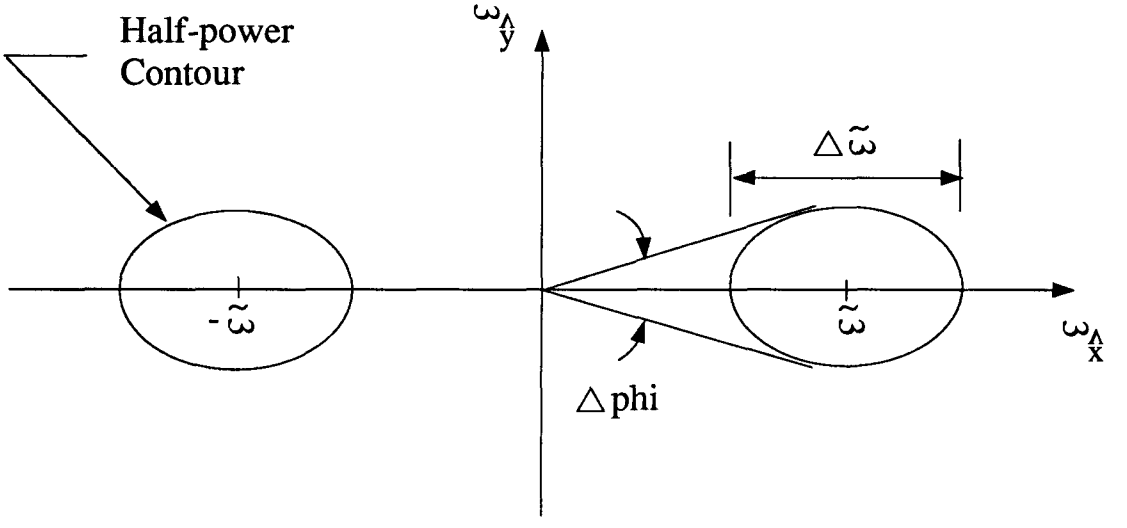
The effective spatial extent in the $\hat{x}$ and $\hat{y}$ directions are given by

$$\Delta\hat{x} = \sigma, \tag{3.102}$$

$$\Delta\hat{y} = \alpha\sigma. \tag{3.103}$$



(a) Gabor Function



(b) Fourier Transform

Figure 3.8: Gabor Function and its Fourier Transform

From (3.100), (3.101), (3.102), and (3.103), it can be seen that the bandwidth and spatial extent for $\tilde{\phi} = 0$ are subject to

$$\Delta\hat{x}\Delta\tilde{\omega} = 2\pi, \tag{3.104}$$

and

$$\Delta\hat{y}\,(\tilde{\omega}\Delta\tilde{\phi}) = 2\pi. \tag{3.105}$$

Equations (3.104) and (3.105) are the uncertainty constraints for the Gaussian window. The uncertainty constraints shows that arbitrary resolution can not be achieved in both the spatial and frequency domains simultaneously. Choosing a scale (σ) and/or aspect ratio (α) that provides a narrow bandwidth will result in a wide spatial extent. It is possible to adjust the frequency bandwidth and the orientation bandwidth independently because (3.104) and (3.105) are not coupled.

The above uncertainty constraints are valid for $\tilde{\phi} = 0$. For Gabor channels with other spectral orientations, the uncertainty constraint is defined in terms of rotated image coordinates: $\Delta\acute{x}$ and $\Delta\acute{y}$ are substituted for $\Delta\hat{x}$ and $\Delta\hat{y}$, respectively.

Log-polar Representation

The following paragraphs describe the sampling of the joint frequency-position space for the log-polar representation. The frequency domain is sampled by selecting the frequency and orientations, as well as the bandwidths, of the Gabor filters. By using the minimum number of Gabor channels that maintain completeness, the selection of a filter set is reduced to the selection of two parameters. Within each channel, a spatial sampling lattice is chosen. The minimally complete spatial sampling interval is defined.

A set of Gabor filters is defined by the selection of phases (p), orientations ($\tilde{\phi}_l$), modulation frequencies ($\tilde{\omega}_k$), scales (σ_k), and aspect ratio (α). The selection of these parameters is partially constrained because, in the log-polar representation, all Gabor

functions are rotation and/or dilations of each other [14]. Rotation is obtained by using the rotated spatial coordinates ($\acute{x}$,$\acute{y}$). This constrains the direction of a given spatial sampling lattice to be the same as the orientation of the Gabor function. Dilation requires that the scale of the Gaussian window be increased by the same factor as the modulation frequency is reduced; that is

$$\sigma_k = \frac{1}{\rho}\sigma_{k-1}, \tag{3.106}$$

$$\tilde{\omega}_k = \rho\tilde{\omega}_{k-1}. \tag{3.107}$$

Note that the frequency-scale product is constant; it is not affected by rotation or dilation. The inverse of this constant product is referred to as the "bandwidth-frequency ratio," and is given by

$$\lambda = \frac{2\pi}{\sigma_k\tilde{\omega}_k} = \frac{\Delta\tilde{\omega}_k}{\tilde{\omega}_k}. \tag{3.108}$$

Since λ is constant for the log-polar representation, defining the channel frequency also defines the channel bandwidth.

Since a Gabor channel has a finite frequency and orientation bandwidth, it is necessary to use a set of Gabor filters to "completely" extract the image information. Completeness means that the output obtained from the Gabor filters provides sufficient information to uniquely reconstruct the image. The smallest set of Gabor filters that preserves completeness is referred to as a "minimally complete" set.

If it is assumed that the filter set is minimally complete, the filter parameters can be systematically chosen. Two phases are sufficient for a minimally complete set. Each channel comprises a quadrature pair of Gabor filters. In this work, the phase for the quadrature pair is defined as $p = \pm\frac{\pi}{4}$.

In a minimally complete set, the angular difference between adjacent orientations is equal to the orientation bandwidth; that is

$$\tilde{\phi}_l - \tilde{\phi}_{l-1} = \Delta\tilde{\phi}. \tag{3.109}$$

The orientation has a π wrap-around making the orientations $\tilde{\phi}$ and $\tilde{\phi} + \pi$ dependent. Thus, the orientation bandwidth in the minimally complete set must be chosen such that

$$\Delta\tilde{\phi} = \frac{\pi}{n_{\tilde{\phi}}}, \tag{3.110}$$

where $n_{\tilde{\phi}}$ is the (integer) number of orientations in the filter set. In the log-polar representation, the orientation bandwidth is the same for all Gabor functions; the orientation bandwidth is not affected by rotation or dilation.

In the minimally complete set, the difference between adjacent modulation frequencies is given by

$$\tilde{\omega}_k - \tilde{\omega}_{k-1} = \frac{\Delta\tilde{\omega}_k + \Delta\tilde{\omega}_{k-1}}{2}. \tag{3.111}$$

Combining (3.100), (3.101), (3.107), (3.110), and (3.111), it can be seen that the frequency multiplication factor ρ is given by

$$\rho = \frac{2\pi\alpha + n_\phi}{2\pi\alpha - n_\phi}. \tag{3.112}$$

Once ρ is defined, the set of modulation frequencies and scales are defined by the selection of a base frequency $\tilde{\omega}_0$ and λ.

From the previous equations, it can be seen that there are four important constants used in forming a set of Gabor filters: $n_{\tilde{\phi}}$, α, ρ, and λ. The minimally complete frequency spacing of the Gabor channels is defined by the selection of $n_{\tilde{\phi}}$ and one of α, ρ, or λ.

A Gabor filter will oversample the spatial domain; an output is produced at each pixel location of the image. If minimal completeness is enforced in the spatial domain, the sampling interval in the $\acute{x}$ and the $\acute{y}$ (rotated axes) directions are determined by the respective spatial extents; that is

$$\acute{x}_n - \acute{x}_{n-1} = \Delta\acute{x}_k = \sigma_k, \tag{3.113}$$

$$\acute{y}_m - \acute{y}_{m-1} = \Delta\acute{y}_k = \alpha\sigma_k. \tag{3.114}$$

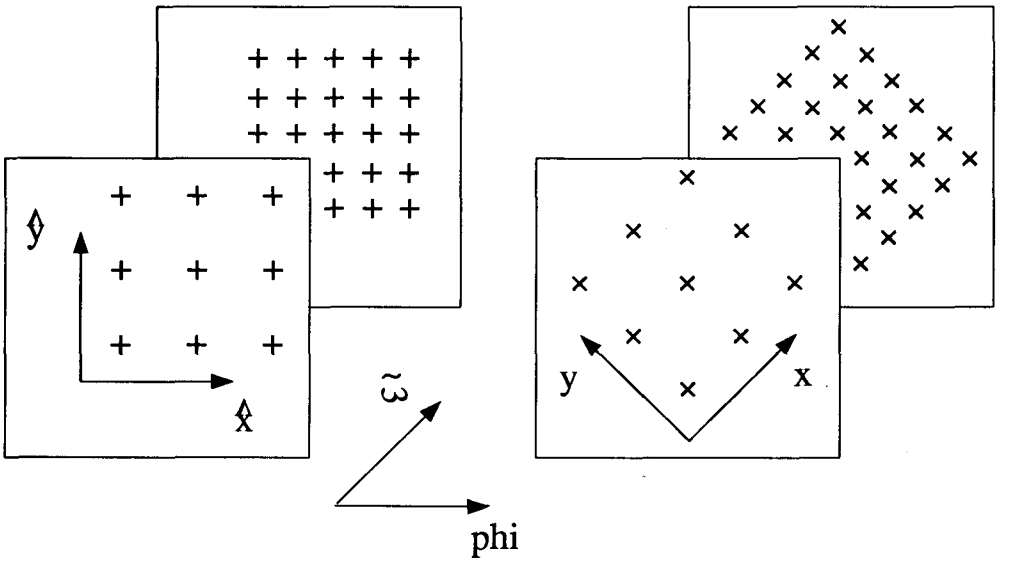


Figure 3.9: Spatial Sampling Lattices

The set of spatial sample points $(\acute{x}_n,\acute{y}_m)$ is referred to as the "spatial sampling lattice" for the channel $\tilde{\omega}_k$, $\tilde{\phi}_l$.

Each Gabor channel has its own spatial sampling lattice. The shape of the lattices vary with the frequency and orientation of the channel. It can be seen in figure 3.9 that higher frequency channels have higher spatial resolutions. It can also be seen that each lattice is rotated to match the orientation of the Gabor channel. The spatial sampling lattice is discussed further in section 3.3.

Local Magnitude and Phase

This paragraph discusses how the local magnitude and phase are measured. A Gabor filter produces a coefficient $a()$ at each lattice point (x_n,y_m):

$$a(\hat{x}_n, \hat{y}_m; \tilde{\omega}_k, \tilde{\phi}_l, p) = \int\int I(\hat{x}, \hat{y}) G_i(\hat{x} - \hat{x}_n, \hat{y} - \hat{y}_m; \tilde{\omega}_k, \tilde{\phi}_l, p) d\hat{x} d\hat{y}. \tag{3.115}$$

The coefficients from quadrature Gabor filters are used to calculate a magnitude and phase which is localized about $\acute{x},\acute{y}$ in the spatial domain and $\tilde{\omega}_k$,$\tilde{\phi}_l$ in the frequency domain. The local magnitude for the channel $\tilde{\omega}_k$, $\tilde{\phi}_l$ is given by

$$m(\acute{x},\acute{y};\tilde{\omega}_k,\tilde{\phi}_l) = [a(\acute{x},\acute{y};\tilde{\omega}_k,\tilde{\phi}_l,p)^2 + a(\acute{x},\acute{y};\tilde{\omega}_k,\tilde{\phi}_l,p+\frac{\pi}{2})^2]^{0.5}. \tag{3.116}$$

The local phase is given by

$$\theta(\acute{x},\acute{y};\tilde{\omega}_k,\tilde{\phi}_l) = \arctan[\frac{a(\acute{x},\acute{y};\tilde{\omega}_k,\tilde{\phi}_l,p+\frac{\pi}{2})}{a(\acute{x},\acute{y};\tilde{\omega}_k,\tilde{\phi}_l,p)}]. \tag{3.117}$$

The output of each Gabor channel forms a magnitude and a phase map. The set of maps provide a multiscale, multiresolution representation of the original image.

The magnitude and phase are used in subsequent subsections to extract image features (section 3.2.2), to estimate local image frequency (section 3.2.3), and to measure inter-frame displacements (sections 3.2.4 and 3.2.5).

3.2.2 Selecting Interesting Features

This subsection discusses how the local magnitude is used to extract uni-directional and omni-directional features. Three thresholds are applied to a magnitude map. The omni-directional features are identified using the normalized moment of inertia.

The Gabor representation makes interesting features explicit. The magnitude response within a Gabor channel represents the significance of the feature. A good feature will produce a large local magnitude.

In order to extract uni- and omni-directional features, large magnitudes must be identified in each channel. In this work, three thresholds are applied to the magnitude response to detect significant values. The first threshold is based on the absolute magnitude. It defines the minimum magnitude to be considered a feature. This threshold can be selected as a fraction of the maximum magnitude or as a multiple of the noise level.

The second threshold is based on the local magnitude at a given lattice point relative to its spatial neighbours. It rejects the hypothesis that a lattice point contains a significant image measurement (a feature) if one of its spatial neighbours has a large magnitude. The third threshold compares the local magnitude response in neighbouring channels. It is based on the relative magnitude of a potential feature viewed through channels with neighbouring orientations, but common frequency. A large magnitude in channel $\tilde{\phi}_1$ at spatial coordinates $(\hat{x}, \hat{y})$ inhibits the instantiation of a feature in channels $\tilde{\phi}_0$ and $\tilde{\phi}_2$ at $(\hat{x}, \hat{y})$.

Omni-directional references can be found by combining the magnitude responses from all orientations [2]. If, at a point $(\hat{x},\hat{y})$, the magnitude responses display significant spectral energy in different (preferably orthogonal) orientations, then the local region is an omni-directional feature. The minimum normalized moment of inertia is an estimate of the variance in the orientation of the local region's spectral energy. An omni-directional feature has a high normalized moment of inertia.

The normalized moment of inertia for the orientation ϕ_T is given by [9]

$$I(\hat{x}, \hat{y}; \tilde{\omega}_k) = \sum_l \frac{m(\hat{x}, \hat{y}; \tilde{\omega}_k, \tilde{\phi}_l)}{C_I} \sin^2[\tilde{\phi}_l - \phi_T], \tag{3.118}$$

where

$$C_I(\hat{x}, \hat{y}; \tilde{\omega}_k) = \sum_l m(\hat{x}, \hat{y}; \tilde{\omega}_k, \tilde{\phi}_l). \tag{3.119}$$

The minimum moment of inertia is obtained by substituting the following orientation into (3.118):

$$\phi_T = 0.5 \arctan \frac{\sum_l m(\hat{x}, \hat{y}; \tilde{\omega}_k, \tilde{\phi}_l) \sin 2\tilde{\phi}_l}{\sum_l m(\hat{x}, \hat{y}; \tilde{\omega}_k, \tilde{\phi}_l) \cos 2\tilde{\phi}_l}. \tag{3.120}$$

The minimum normalized moment of inertia is limited to an interval given by

$$0 \leq I(\hat{x}, \hat{y}; \tilde{\omega}_k) \leq 0.5. \tag{3.121}$$

[2] Interpolation is required because the spatial lattices from different Gabor channels are not aligned.

A uni-directional reference has a minimum normalized moment of inertia near zero; for an omni-directional reference, $I(\hat{x}, \hat{y}; \tilde{\omega}_k)$ is near one-half.

3.2.3 Measuring Phase Differences

This subsection discusses how the phase difference between lattice points is measured. The phase difference is used to estimate the local frequency and the expected frequency error. Criteria for identifying phase measurements that are likely to be unstable with respect to image deformations induced by sensor/object motion are established. The sensitivity of the phase difference measurement to image noise is discussed.

The local spatial frequency and small inter-frame displacements are estimated using phase differences. The local phase difference is given by

$$\Delta\theta = \theta_1 - \theta_0 = \arctan \frac{a_0(p)a_1(p+\frac{\pi}{2}) - a_1(p)a_0(p+\frac{\pi}{2})}{a_1(p)a_0(p) + a_1(p+\frac{\pi}{2})a_0(p+\frac{\pi}{2})}, \tag{3.122}$$

where the $a_1(p)$ and $a_0(p)$ are coefficients at lattice points 1 and 0, respectively, obtained from a Gabor filter with phase p. For the case of local frequency estimation, the two lattice points belong to the same channel, and the same image. For inter-frame displacements, the lattice points belong to to the same channel, but different images. Discussion of the inter-frame displacements can be found in later subsections on Gabor-based disparity (section 3.2.4) and normal image velocity (section 3.2.5) estimation.

The local spatial frequencies, $\omega_{\hat{x}}$ and $\omega_{\hat{y}}$, are estimated using the average phase shift between the lattice point $(\acute{x}_n, \acute{y}_m)$ and its spatial neighbours along the $\acute{x}$ and $\acute{y}$-axes, respectively; that is

$$\omega_{\acute{x}} = 0.5[\frac{\Delta\theta_{n,n-1}}{\acute{x}_n - \acute{x}_{n-1}} + \frac{\Delta\theta_{n+1,n}}{\acute{x}_{n+1} - \acute{x}_n}], \tag{3.123}$$

$$\omega_{\acute{y}} = 0.5[\frac{\Delta\theta_{m,m-1}}{\acute{y}_m - \acute{y}_{m-1}} + \frac{\Delta\theta_{m+1,m}}{\acute{y}_{m+1} - \acute{y}_m}], \tag{3.124}$$

where

$$\Delta\theta_{n,n-1} = \theta(\acute{x}_n, \acute{y}_m) - \theta(\acute{x}_{n-1}, \acute{y}_m). \tag{3.125}$$

The difference between the two phase shifts is used to measure the expected frequency error in the $\acute{x}$ and $\acute{y}$ directions:

$$|\Delta\omega_{\acute{x}}| = |\frac{\Delta\theta_{n,n-1}}{\acute{x}_n - \acute{x}_{n-1}} - \frac{\Delta\theta_{n+1,n}}{\acute{x}_{n+1} - \acute{x}_n}|, \tag{3.126}$$

$$|\Delta\omega_{\acute{y}}| = |\frac{\Delta\theta_{m,m-1}}{\acute{y}_m - \acute{y}_{m-1}} - \frac{\Delta\theta_{m+1,m}}{\acute{y}_{m+1} - \acute{y}_m}|. \tag{3.127}$$

The expected frequency errors, $\Delta\omega_{\acute{x}}$ and $\Delta\omega_{\acute{y}}$, are used to estimate the expected error in disparity and normal image velocity.

The expected frequency error is one measure of phase stability with respect to small image deformations. The phase is generally stable throughout the image. There are some image regions, referred to as "phase singularity neighbourhoods [21]," where the phase is not stable. When the expected frequency error is large, the gradient-based estimate of the inter-frame displacement will be unreliable; the assumption that the higher-order phase terms in the Taylor expansion are negligible is invalid.

The difference between the local frequency and the channel frequency is also used to detect phase singularity neighbourhoods. Although the local frequency ω_n is generally different than the channel frequency $\tilde{\omega}_k$, the difference should not be much larger than half the bandwidth of the filter. Consider the case of an image filtered using a Gabor function modulated along the $\hat{x}$-axis, that is, $\tilde{\phi}_l = 0$. The ratio of the $\hat{x}$ frequency difference and the effective bandwidth of the Gabor filter is given by [21]

$$\tau_{\omega(\hat{x})} = \frac{|\omega_{\hat{x}} - \tilde{\omega}_k|}{\Delta\tilde{\omega}_k}. \tag{3.128}$$

The ratio of the orthogonal frequency difference and the effective bandwidth (in the $\hat{y}$ direction) is given by

$$\tau_{\omega(\hat{y})} = \frac{|\omega_{\hat{y}}|\alpha}{\Delta\tilde{\omega}_k}. \tag{3.129}$$

In this work, an image region is removed from the active feature list if either

$$\tau_{\omega(\hat{y})} > 0.5, \tag{3.130}$$

or

$$\tau_{\omega(\hat{y})} > 0.5. \tag{3.131}$$

Note that the above constraints are valid for $\tilde{\phi}_l = 0$; rotated coordinates are used in (3.128) and (3.129) when $\tilde{\phi}_l \neq 0$.

The measurement of the local frequency imposes a new requirement on the spatial sampling interval. The local frequency is estimated using the phase change between lattice points. The spatial lattice must be oversampled (with respect to the minimally complete spatial sampling interval) to ensure that all frequencies within the channel passband can be measured. To avoid aliasing, the phase change between the reference lattice point and its neighbours must be less than $\pm\pi$. The demodulated phase change [3] between adjacent lattice points in the direction of modulation is given by

$$\Delta\theta_{demod} = (\omega_{\acute{x}} - \tilde{\omega}_k)\Delta\acute{x}_s, \tag{3.133}$$

where $\Delta\acute{x}_s$ is the sampling interval along the $\acute{x}$-axis. To ensure that $|\Delta\theta_{demod}|$ is less than a chosen threshold, θ_o, for all

$$|\omega_{\acute{x}} - \tilde{\omega}_k| < \frac{\Delta\tilde{\omega}_k}{2}, \tag{3.134}$$

the lattice spacing $\Delta\acute{x}_s$ must be constrained:

$$\Delta\acute{x}_s < \frac{2\theta_o}{\Delta\tilde{\omega}_k} = \frac{2\theta_o}{\lambda\tilde{\omega}_k}. \tag{3.135}$$

The phase change between adjacent lattice points that are orthogonal to the modulation is given by

$$\Delta\theta = \omega_{\acute{y}}\,\Delta\acute{y}_s, \tag{3.136}$$

[3]The phase response is demodulated before measuring $\Delta\theta$. The conversion from the demodulated phase derivative to local image frequency along the $\acute{x}$-axis is given by

$$\omega_{\acute{x}} \approx \tilde{\omega}_k + \frac{\Delta\theta_{demod}}{\Delta\acute{x}}. \tag{3.132}$$

where $\Delta\acute{y}_s$ is the sampling interval along the $\acute{y}$-axis. The constraint on the lattice spacing $\Delta\acute{y}_s$ that ensures that $|\Delta\theta| < \theta_o$, for all

$$|\omega_{\acute{y}}| < \frac{\Delta\tilde{\omega}_k}{2\alpha}, \tag{3.137}$$

is given by

$$\Delta\acute{y}_s < \frac{2\theta_o\alpha}{\Delta\tilde{\omega}_k} = \frac{2\theta_o}{\tilde{\omega}_k(\Delta\tilde{\phi})}. \tag{3.138}$$

A good choice for the threshold phase is $\theta_o = \frac{\pi}{2}$.

The relative magnitude between adjacent lattice points can be used to determine if the frequency measurement is aliased or is being influenced by neighbouring image features. Consider an edge whose normal is parallel to the $\acute{x}$-axis. Assume that it passes through the reference lattice point. If there are no other significant features in the neighbourhood, the magnitude at the adjacent point in the modulation direction will be attenuated:

$$\tau_{m(\acute{x})}(edge) = \frac{m_{1,0}}{m_{0,0}} = \exp[-a_k(\Delta\acute{x}_s)^2], \tag{3.139}$$

where

$$a_k = \frac{\lambda^2\tilde{\omega}_k^2}{4\pi}, \tag{3.140}$$

and $m_{i,j}$ is the magnitude at a lattice point that is offset from the reference point by $(i\Delta\acute{x}_s, j\Delta\acute{y}_s)$. The relative magnitude $\tau_{m(\acute{x})}$ can be less if the edge does not pass through the reference point. Consider an edge that passes to the left of the reference point by $\frac{\Delta\acute{x}_s}{2}$. The relative magnitude between the reference and the right adjacent point is

$$\tau_{m(\acute{x})}(edge, wc) = \frac{\exp[-a_k(1.5\Delta\acute{x}_s)^2]}{\exp[-a_k(0.5\Delta\acute{x}_s)^2]} = \exp[-2a_k(\Delta\acute{x}_s)^2], \tag{3.141}$$

where $(edge, wc)$ denotes the worst case attenuation for an edge.

A similar attenuation exists for adjacent lattice points in the orthogonal direction. Consider an edge whose normal is tilted relative to the modulation direction of the

Gabor channel. The largest tilt within the channel bandwidth is $\frac{\Delta\phi}{2}$. For such a case, the attenuation is given by

$$\tau_{m(\acute{y})}(edge, wc) = \frac{m_{0,1}}{m_{0,0}} = \exp[-\frac{\lambda^2\tilde{\omega}_k^2}{4\pi\alpha^2}(\frac{\lambda\,\Delta\acute{y}}{2\alpha})^2]. \tag{3.142}$$

The above-mentioned attenuations are valid for the special case of a edge. Most uni-directional features, such as sine wave gratings, will experience less attenuation. A magnitude test that identifies local frequency estimates that are aliased or are influenced by neighbouring image features can be formed. If the relative magnitude (for a uni-directional feature) is defined as

$$\tau_{m(\acute{x})}(uni) = \min[\frac{m_{1,0}}{m_{0,0}}, \frac{m_{0,0}}{m_{1,0}}], \tag{3.143}$$

$$\tau_{m(\acute{y})}(uni) = \min[\frac{m_{0,1}}{m_{0,0}}, \frac{m_{0,0}}{m_{0,1}}]. \tag{3.144}$$

then the test for valid phase change measurements are given by

$$\tau_{m(\acute{x})}(edge, wc) < \tau_{m(\acute{x})}(uni) \leq 1.0, \tag{3.145}$$

$$\tau_{m(\acute{y})}(edge, wc) < \tau_{m(\acute{y})}(uni) \leq 1.0. \tag{3.146}$$

Adjacent points whose relative magnitude does not satisfy the above constraints are rejected.

The local frequency is estimated using the average of two phase changes. If one of the phase changes is rejected, the local frequency can still be estimated using the remaining measurement. The expected error for the local frequency is set equal to half the channel bandwidth in this case. If both phase changes are rejected, the feature is consider unstable and is removed from the active feature list.

The expected frequency error is used in sections 3.2.4 and 3.2.5 to estimate the expected disparity error and the expected normal image velocity error, respectively. The

expected error in the phase shift for the inter-frame displacement is also used. An expression for the expected inter-frame phase error, obtained using sensitivity analysis, appears below.

The sensitivity of (3.122) to errors in the Gabor coefficients is given by the following derivatives:

$$S_0(p) = \frac{\delta\Delta\theta}{\delta a_0(p)} = \frac{a_0(p+\frac{\pi}{2})}{[m_0]^2}, \tag{3.147}$$

$$S_1(p) = \frac{\delta\Delta\theta}{\delta a_1(p)} = -\frac{a_1(p+\frac{\pi}{2})}{[m_1]^2}, \tag{3.148}$$

where

$$m = [a^2(p) + a^2(p+\frac{\pi}{2})]^{0.5}. \tag{3.149}$$

The error in the local phase difference ($\delta\Delta\theta$) due to errors in the Gabor coefficients (δa) is given by

$$[\delta(\Delta\theta)]^2 = S_0^2(p)\delta a_0^2(p)+S_0^2(p+\frac{\pi}{2})\delta a_0^2(p+\frac{\pi}{2})+S_1^2(p)\delta a_1^2(p)+S_1^2(p+\frac{\pi}{2})\delta a_1^2(p+\frac{\pi}{2}). \tag{3.150}$$

If it is assumed that the error in the Gabor coefficient is due to in-channel noise whose power is denoted by σ_G^2, the error in the local phase difference is approximately given by

$$\delta(\Delta\theta) \approx \frac{2\sigma_G}{m_0 + m_1}. \tag{3.151}$$

Thus, given a model of the image noise within the Gabor channel, the expected inter-frame phase error can be estimated. Note that the expected inter-frame phase error is inversely dependent on the signal-to-noise ratio of the filtered image.

3.2.4 Disparity

This subsection discusses how disparity is measured using Gabor filters. An oversampled lattice is proposed to extend to measurable range of depths. Criteria for rejecting incorrect feature matches are established. The expected error in the disparity measurement is estimated.

The use of stereo disparity to calculate depth was reviewed in chapter 2. It was shown that the disparity for parallel stereo cameras is constrained to the epipolar line. In this work, only the Gabor channels whose orientation is along the epipolar line are used for estimating disparity. These channels are referred to as the "epipolar channels."

Disparity is measured by comparing maps from the same channel, but from the left and right images. Once corresponding lattice points are identified in the left and right images, the relative disparity is measured using the phase difference:

$$d_{rel} = \frac{\Delta\theta_{L,R}}{\omega_{\hat{x}}}, \tag{3.152}$$

where $\Delta\theta_{L,R}$ is the phase difference between the left and right images. Because of the epipolar constraint, only the $\hat{x}$ component of the local frequency is estimated; $\omega_{\hat{y}}$ is not required.

The range of measurable depths for (3.152) is limited. Since the local phase difference is modulo 2π the measurable disparity is restricted to

$$-\frac{\pi}{\omega_{\hat{x}}} < d_{rel} \leq \frac{\pi}{\omega_{\hat{x}}}. \tag{3.153}$$

This interval, referred to as the "disparity interval," is dependent on the local frequency of the filtered image. Since $\omega_{\hat{x}} \approx \tilde{\omega}_k$ (see (3.128)), the disparity interval will be very small for high frequency channels.

The range of measurable depths can be extended by selecting a set of epipolar offsets with overlapping disparity intervals. The epipolar offsets can be chosen as multiples of the spatial sampling interval ($\Delta\hat{x}_s$); that is

$$E_{offset} = n_o \Delta\hat{x}_s, \tag{3.154}$$

where n_o is a non-negative integer. In such a case, a set of epipolar offsets can be achieved by matching a lattice point in the right image with incremental lattice points in the left

image. To ensure that the disparity intervals overlap and to avoid phase wraparound, the spatial domain is oversampled along the epipolar line. If adjacent spatial samples are separated by a phase shift of π (with respect to the channel frequency, $\tilde{\omega}_k$), the sampling interval is given by

$$\Delta\hat{x}_s = \frac{\pi}{\tilde{\omega}_k}. \tag{3.155}$$

Multiple disparity intervals produce a set of possible depths: the correct depth, and many aliased depths. Aliased depth estimates must be identified and rejected. There a number of constraints and measures that can be used to reject unlikely depth estimates: the disparity is positive (for parallel stereo cameras); the best match should have a small phase shift; and the local magnitude and normalized moment of inertia are similar at corresponding points.

The cameras can only view objects that are in front of them. As a result, the disparity for parallel stereo cameras is always positive. Candidate matches with negative disparities are identified as aliased and are rejected.

The phase difference can be used to reject bad matches. If the spatial domain is oversampled to satisfy (3.155), corresponding lattice points will usually have a phase shift that is less than $\frac{\pi}{2}$. The largest phase shift, accounting for the finite bandwidth of the Gabor channel, is restricted to

$$|\Delta\theta| \leq \frac{\pi}{2}(1 + \frac{\lambda}{2}). \tag{3.156}$$

Any potential match with a phase shift exceeding (3.156) can be rejected. If there are two neighbouring disparity intervals that satisfy (3.156) and the other matching criteria, the interval with the lower phase shift is the better match.

Differences in the local magnitude between stereo images can be used to detect aliasing. Consider as an example a scene that consists of a dark spot on a light coloured wall. The spot become a reference feature when the scene is viewed by stereo cameras. If the

centroid of the spot coincides with a lattice point in the right image, the spot will be displaced from the corresponding lattice point in the left image by the relative disparity d_{rel}. The spot produces the maximum magnitude response in the right image, and an attenuated response in the left image. The attenuation in the local magnitude [4] is given by

$$r_{R,L} = min[\frac{m_R(\hat{x},\hat{y})}{m_L(\hat{x},\hat{y})}, \frac{m_L(\hat{x},\hat{y})}{m_R(\hat{x},\hat{y})}] = \exp[-\pi(\frac{d_{rel}}{\sigma_k})^2]. \tag{3.157}$$

The disparity interval containing the correct depth produces a large $r_{R,L}$, near unity. If $r_{R,L}$ is small, then the relative disparity is too large and the depth estimate is identified as aliased.

If the relative disparity is small enough such that aliasing does not occur, (3.157) can be rewritten in terms of the local phase difference:

$$r_{R,L} = \exp[-\pi\lambda^2(\frac{\Delta\theta}{2\pi})^2]. \tag{3.158}$$

Since the onset of aliasing occurs at $|\Delta\theta| = \pi$,

$$r_{R,L} < r_{alias} = \exp[-\pi(\frac{\lambda}{2})^2] \tag{3.159}$$

indicates an aliased measurement [5].

Differences in the normalized moment of inertia can also be used to detect aliasing. The normalized moment of inertia is more difficult to use as a matching criterion than local magnitude. In this work, heuristic thresholds are used to identify uni-directional features and omni-directional features. A uni-directional feature is given by $I < 0.3$; an omni-directional feature is given by $I > 0.4$. Any potential pairing that attempts to match a uni-directional with an omni-directional features is rejected. In this work, the normalized moment of inertia is used only in the E_{offset} histogram described in chapter 4.

[4] It is assumed that the gains of the stereo cameras are matched.

[5] A tighter bound can be obtained by using $|\Delta\theta_{max}| = \frac{\pi}{2}(1 + \frac{\lambda}{2})$ instead of π.

The above criteria are used to reject bad matches. Even after applying these constraints, more than one potential match may still exist. Information from other sources is used to select the correct disparity interval. A priori information sources include disparity estimates from the lower frequency channels and from past images. The use of a priori prediction for feature correspondence is discussed in chapter 4. Note that the E_{offset} histogram mentioned in the previous paragraph is used when no a priori information is available.

It is desirable to model the accuracy of the disparity measurement. The error in a disparity measurement is given by

$$\delta d_{\hat{x}} = (E_{offset} - \hat{E}) + \delta d_{rel}, \tag{3.160}$$

where $\hat{E}$ is the selected epipolar offset, and δd_{rel} is the error in the relative disparity. The disparity estimate has two primary sources of error: an incorrect epipolar offset or an inaccurate estimate of the relative disparity. An incorrect epipolar offset, or equivalently a correspondence error, is usually large and difficult to model. If the image feature is moving, the correspondence error will be identified in later stages of processing, during the velocity-based outlier test (Mahalanobis distance in chapter 4). Errors in the relative disparity are primarily due to inaccurate estimation of the local frequency $\omega_{\hat{x}}$ and due to noise σ_G^2. The model of the relative disparity error is given by

$$(\delta d_{rel})^2 \approx (d_{rel}^2)\,(\frac{\delta\omega_{\hat{x}}}{\omega_{\hat{x}}})^2 + (\frac{\sigma_G}{\omega_{\hat{x}}})^2[0.5(m_R + m_L)]^{-2}, \tag{3.161}$$

where m_L and m_R are the local magnitude of the left and right images. Equation (3.161) is used in chapter 4 as the expected disparity error.

3.2.5 Normal Image Velocity

This subsection discusses the Gabor-based estimation of normal image velocity. The matching criteria established for disparity measurements are reformulated for normal

image velocity. The expected error in the normal image velocity measurement is estimated.

Measuring normal image velocity is similar to measuring the disparity except that the direction of image motion is not known. As a consequence, both the $\acute{x}$ and $\acute{y}$ components of local frequency must be estimated. Normal image velocity measurements are obtained from each channel. The Gabor channels whose orientation is not along the epipolar line are referred to as "oblique channels." A special case is the channel whose orientation is orthogonal to the epipolar channel. This channel is referred to as the "orthogonal channel."

Normal image velocity is measured by comparing maps from the same channel, but from successive images in an image sequence. To measure normal image velocity, we need the normal direction (ϕ_n), the local frequency along the normal (ω_n), and the temporal frequency (ω_t). The normal direction, with respect to the epipolar line, is defined as

$$\phi_n = \arctan \frac{\omega_{\hat{y}}}{\omega_{\hat{x}}} = \tilde{\phi}_l + \arctan \frac{\omega_{\acute{y}}}{\omega_{\acute{x}}}. \tag{3.162}$$

The local frequency along the normal direction is given by

$$\omega_n = [\omega_{\acute{x}}^2 + \omega_{\acute{y}}^2]^{0.5}. \tag{3.163}$$

The local temporal frequency, ω_t, is measured using the phase shift of corresponding points from two successive images:

$$\omega_t = \frac{\Delta\theta(t_i)}{\Delta t}, \tag{3.164}$$

where

$$\Delta t = t_i - t_{i-1}. \tag{3.165}$$

If there is no lattice offset between the corresponding lattice points, the normal image velocity is obtained by substituting ω_t and ω_n into (3.85). If there is a lattice offset, the

normal image velocity is given by

$$V_n = \frac{\acute{x}_{offset}}{\Delta t}\cos\phi_n + \frac{\acute{y}_{offset}}{\Delta t}\sin\phi_n - \frac{\omega_t}{\omega_n}, \tag{3.166}$$

where $\acute{x}_{offset}$ and $\acute{y}_{offset}$ are the lattice offsets along the $\acute{x}$- and $\acute{y}$-axes, respectively.

As in the case of disparity, the spatial domain must be oversampled. In this work, a lattice sampling interval of

$$\Delta\acute{x}_s = \frac{\pi}{\tilde{\omega}_k} \tag{3.167}$$

is used. It will ensure that some $\acute{x}_{offset}$ exists such that $\Delta\theta(t_i)$ does not experience phase wraparound. The $\acute{y}$ direction is also oversampled:

$$\Delta\acute{y}_s = \frac{\pi\alpha}{\tilde{\omega}_k}. \tag{3.168}$$

The lattice offset between corresponding lattice points is obtained using a hypothesis-test approach. The lattice offset is predicted from the current estimate of sensor motion. This estimate is based on information from lower frequency channels. At the lowest frequency channel, the lattice shift is assumed to be zero. This assumption is acceptable because the phase-based estimator of image displacement covers a large spatial region for low frequency channels. Once a lattice offset prediction is made, the four nearest lattice points (in the corresponding image) are tested as potential matches. The matching criteria is a subset of the disparity test: the local magnitude must be similar at corresponding lattice points; and the best match has a small phase shift.

The accuracy of the normal image velocity and the normal direction can be estimated. The constraints (3.128) and (3.129) ensure that the local frequencies within a Gabor channel are approximately given by

$$\omega_{\acute{x}} \approx \tilde{\omega}_k, \tag{3.169}$$

$$\omega_{\acute{y}} \approx 0. \tag{3.170}$$

In such a case, errors in the normal direction and the normal frequency, due to errors in the local frequency (but not correspondence errors), are approximated by

$$\delta\phi_n \approx \frac{\delta\omega_{\dot{y}}}{\tilde{\omega}_k}, \tag{3.171}$$

$$\delta\omega_n \approx \delta\omega_{\dot{x}}. \tag{3.172}$$

The expected error in the normal image velocity, which include the effects of noise, is given by

$$(\delta V_n)^2 \approx (\frac{\omega_t}{\omega_n})^2(\frac{\delta\omega_n}{\omega_n})^2 + (\frac{\sigma_G}{\omega_n})^2[0.5(m_1 + m_0)]^2, \tag{3.173}$$

where m_1 and m_0 are the local magnitude of the successive images.

3.3 Notes on the Sampling Lattice

This section discusses further the spatial sampling lattice. The utility of a multiscale, multiresolution representation for measuring and predicting image displacements is discussed. Sampling schemes that reduce the size of higher frequency lattices are examined. The effect of spatial oversampling on the estimation of the inter-frame sensor motion is discussed.

For the case of the log-polar representation, multiscale refers to a set of Gabor channels with different channel frequencies and different channel bandwidths; multiresolution refers to a set of lattice with varying sampling densities. Multiscale, multiresolution representations are useful for estimating disparity and normal image velocity. Both disparity and normal image velocity is easily estimated using lower frequency channels. In low frequency channels, phase-based measurements of image displacement cover a large interval. In addition, low frequency channels contain a sparse number of lattice points which simplifies correspondence. The drawback of low frequency measurements is that the expected error tends to be large.

Lower frequency measurements can be used as a priori predictions of the more accurate higher frequency measurements. Lower frequency disparity measurements provide a local constraint on the possible disparities in higher frequency channels. The lower frequency normal image velocity measurements are used to produce an estimate of the sensor motion; the sensor motion is subsequently used to predict the position of corresponding features in higher frequency channels. The prediction of disparity and normal image velocity is detailed in chapter 4.

The size of the spatial sampling lattices is large for high frequency channels. Both the size of the lattice and the sampling density increases with the square of the channel frequency. One may wish to restrict the size of the spatial sampling lattices of higher frequency channels in order to reduce the memory and computational requirements. These lattice restrictions have utility only if they exclude image measurements that are unimportant or redundant.

In autonomous vehicle operations, a stationary object near the sensor's focus of expansion will (eventually) obstruct the vehicle. To ensure reliable detection of stationary obstacles, lattice points near the sensor's focus of expansion must be retained. For the case of a forward translating sensor, the focus of expansion is near the image origin. Any lattice restrictions should retain the important samples near the image origin, and exclude periphery points. If the number of lattice points is fixed, the size of the region covered by the lattice is reduced as the channel frequency is increased. The restricted sampling lattice is shown in figure 3.10.

One limitation of this restricted sampling lattice is that it considers only the detection of stationary obstacles, and not the requirements for disparity measurements. In the restricted lattice shown in figure 3.10, the maximum epipolar offset reduces as the channel frequency increases. As a result, many stereo correspondences will be lost. Without a disparity estimate, the associated normal image velocity measurement is useless for

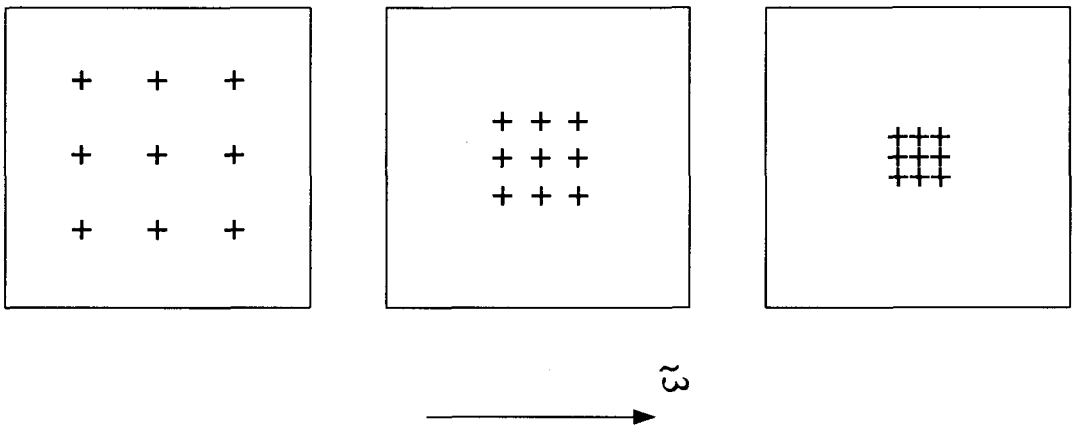


Figure 3.10: Restricted Sampling Lattice

estimating sensor motion.

An alternative approach is to use a band sampling lattice for the epipolar channel, and the restricted sampling lattice for the oblique channels. The band sampling lattice extends across the image in the direction of the channel orientation. Band sampling restricts the lattice in one direction only: the direction orthogonal to the channel orientation. For the epipolar channel, band sampling produces a high resolution band parallel to the $\hat{x}$-axis, as shown in figure 3.11. A virtue of band sampling is that the number of points along the epipolar line is increased. As a result, larger disparities can be measured. A second advantage of band sampling is that the number of stereo velocity measurements is increased. This will result in better detection of moving objects. The final advantage of band sampling is that peripheral measurements can help distinguish between rotation and translation when the sensor is viewing a scene with a constant depth. With the additional peripheral epipolar measurements, the T_x and Ω_y sensor motion parameters are more accurately estimated.

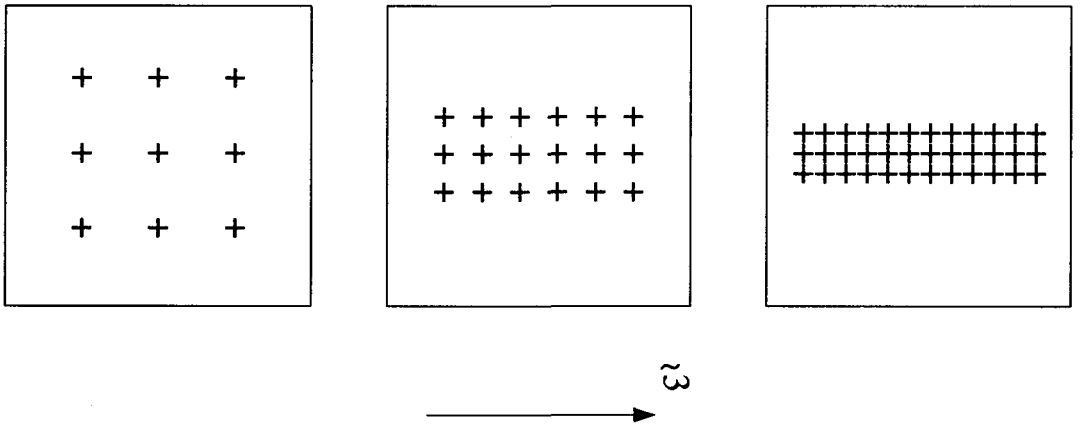


Figure 3.11: Band Sampling Lattice

The final advantage would seen to be a reason to use band sampling for oblique channels. Although more sample points would improve the accuracy of T_y and Ω_x, the accuracy may not be required. In a typical autonomous vehicle application, the vehicle is travelling on a planar surface and the optical axis of the sensor is approximately parallel to the ground plane. Knowledge of planar sensor motion can be used to improve the accuracy of T_y and Ω_x, as shown in chapter 4.

The spatial sampling lattice is oversampled to avoid aliasing in the measurement of local frequency, and to produce overlapping disparity (and normal image velocity) intervals. The oversampling does not increase the amount of information available from a Gabor channel; the larger number of measurements represent the same amount of information. As a result, when the disparity and normal image velocity measurements are combined to estimate the inter-frame sensor motion, the weight assigned to each measurement must be discounted. A brief discussion of the discount factor follows. A more detailed discussion can be found in appendix A.

In chapter 4, a weighted least squares technique is used to estimate the inter-frame sensor motion. If all the image measurements are independent, the assigned weight is the inverse of the expected squared error of the measurement. If an active feature has no active neighbours, the measurement can be considered independent. In such a case, no discount is required. If an active feature is surrounded by other active features, the measurement is dependent on, and highly correlated with, the neighbouring measurements. A larger discount factor is required to properly model the amount of new information produced by each correlated measurement. The discount factor is the inverse of the sum of the overlaps between the reference feature and all its active neighbours. If the neighbourhood of interest is restricted the eight closest points in the spatial lattice (the adjacent vertical and horizontal lattice points and the adjacent diagonal lattice points), the discount factor, $\beta_{discount}$, will be in the following range (see appendix A):

$$\frac{1}{1+2a+2b+4ab} \leq \beta_{discount} \leq 1.0, \tag{3.174}$$

where

$$a = \exp[-\frac{\pi}{2}(\lambda\tilde{\omega}_k\Delta\acute{x}_s)^2], \tag{3.175}$$

and

$$b = \exp[-\frac{\pi}{2}(\frac{\lambda}{\alpha}\tilde{\omega}_k\Delta\acute{y}_s)^2]. \tag{3.176}$$

The overlap between adjacent horizontal, vertical, and diagonal points are denoted by a, b, and ab, respectively.

3.4 Discussion and Summary

It should be apparent that normal image velocity and disparity measurements are not made at every pixel in the image. The image is encoded into a representation that comprises a set of spatially subsampled bandpass channels. Distinctive image features are

identified from the magnitude maps of each channel. Even though the spatial position of features may coincide, the maps from the set of Gabor channels are processed independently. The application of the thresholds to the magnitude maps result in a significant reduction in data without losing important information. The resulting set of features may have a sparse and nonuniform distribution. Data is also reduced using constraints on the local frequency. Features that are likely to be unstable with respect to sensor motion-induced image deformations are identified and rejected.

The remaining features in each channel are used to measure the disparity and the normal image velocity. At this stage, some cross channel coherence is assumed. Information from lower frequency channels is exploited to aid feature correspondence.

Along with the disparity and normal image velocity measurements, the Gabor-based approach estimates the expected error in each measurement. The expected error or other measure of uncertainty is important when information is combined. In chapter 4, all features with valid normal image velocity and disparity measurements are combined to estimate sensor motion. A weighted least square estimate of sensor motion reduces the influence of uncertain measurements. In addition, knowledge of the expected measurement error allows the calculation of the error covariance matrix for the inter-frame sensor motion. The expected measurement error and the error covariance matrix are used in the Mahalanobis distance test to identify and reject features that belong to nonstationary objects.

The propagation of the expected measurement error can go beyond the inter-frame sensor motion stage. In this work, the expected measurement errors eventually become part of the error covariance matrix for the extended sensor motion, the error covariance matrix for each object motion, and the expected error in the collision parameters for each moving object. It will be shown in chapter 4 that the various error covariance matrices are needed to integrate auxiliary sensor information and to exploit external knowledge

such as planar motion. The expected error in the collision parameters would be used by the obstacle avoidance module.

Chapter 4

Obstacle Detection using a Stereo Image Sequence

This chapter describes how the collision parameters are estimated from a stereo sequence of images. An overview of the obstacle detection algorithm is presented in section 4.1. The mathematical descriptions of the algorithm's important submodules—inter-frame sensor motion, Mahalanobis distance, and Kalman filter—are included in later sections. Subtle details that are necessary to implement the algorithm are investigated. A comparison of the various submodules of this algorithm with the work of other researchers is provided.

4.1 Overview of the Obstacle Detection Algorithm

This section provides a brief overview of the obstacle detection algorithm. Information is transformed from pixels in a stereo image sequence into collision parameters for moving objects.

The obstacle detection algorithm is shown in figure 4.12. The stereo image sequence is processed using the Gabor filter technique described in chapter 3. A set of interesting image features are selected (section 3.2.2). The disparity, the normal image velocity, and the associated expected errors are measured at corresponding features (sections 3.2.4 and 3.2.5). The disparity is converted into a depth estimate using (2.57). From the depth and the image coordinates, the transformation matrices $A(z^{-1})$ and $B(z)$ are calculated for each feature. The normal image velocity measurement provides both the magnitude V_n, and direction ϕ_n. The normal direction vector $\bar{n}$, along with $A(z^{-1})$ and $B(z)$, form

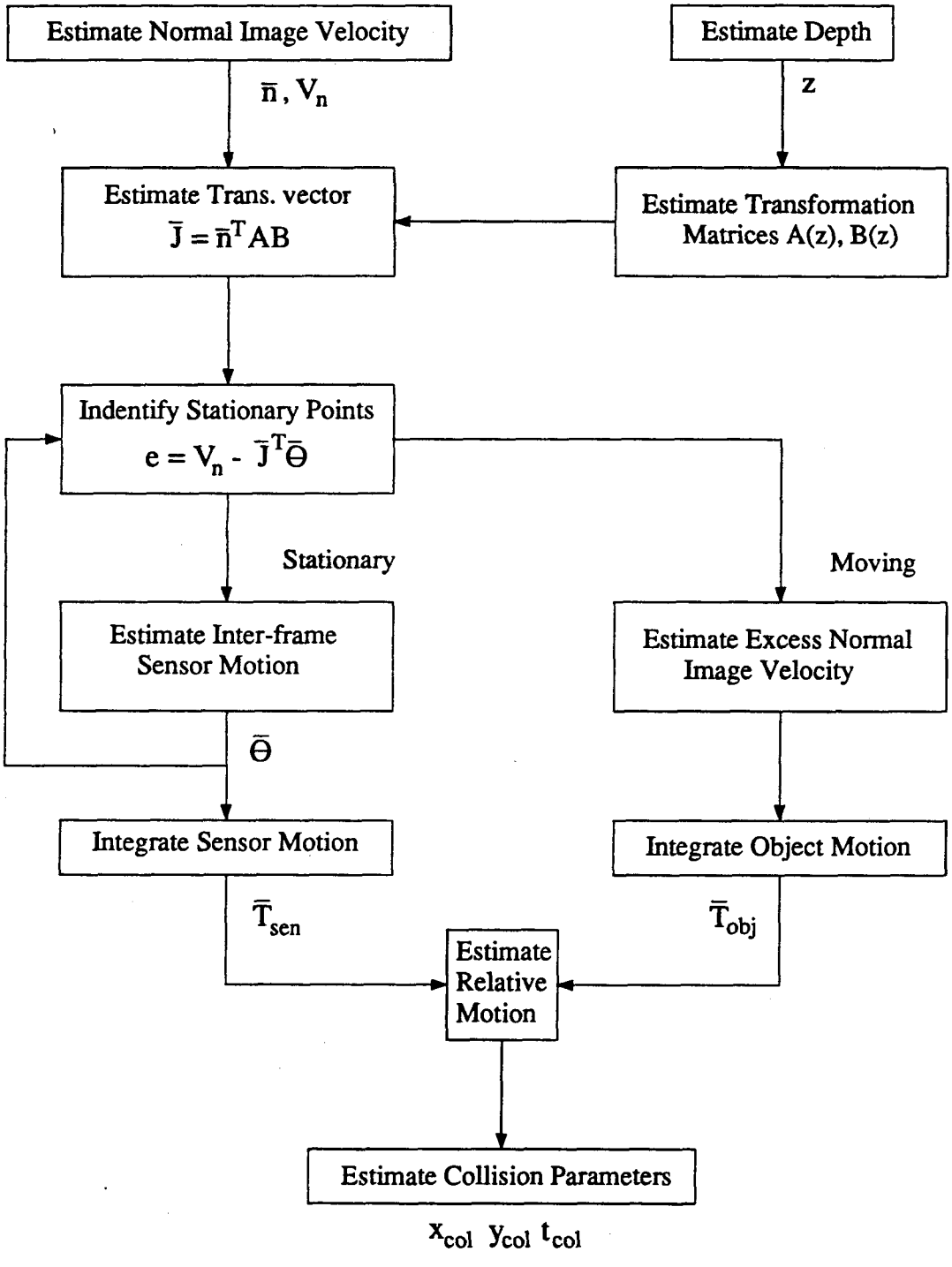


Figure 4.12: Obstacle Detection Algorithm

the transformation vector $\bar{J}$.

The vector $\bar{J}$ transforms the inter-frame sensor motion into an estimate (prediction) of the normal image velocity. This prediction is valid for stationary objects only because it does not account for object motion. The difference between the measured and predicted normal image velocity is the prediction error. If the prediction error is much larger than the expected error, the hypothesis that the feature belongs to a stationary object is rejected. This hypothesis is tested using the Mahalanobis distance (section 4.3).

The features that pass the Mahalanobis distance test are used to refine the estimate of inter-frame sensor motion. The inter-frame sensor motion is estimated using a weighted least square approach (section 4.2). The expected error in the normal image velocity is used to weight the sensor motion estimation.

Any normal velocity measurement inconsistent with its sensor motion-based prediction is processed as a moving object. The excess normal image velocity—the measured less the sensor motion-induced normal image velocity—is used to estimate the translational motion of the object relative to the ground surface.

The inter-frame motion estimates are integrated over the image sequence. The translational portion of the inter-frame sensor motion is integrated using a Kalman filter (section 4.4). Similar Kalman filters are used to integrate the translational motion of each moving object. The difference between these integrated motions is used to estimate the collision parameters for each moving object (section 4.5).

The above paragraphs are a simplified summary of the obstacle detection algorithm. Note that an estimate of the inter-frame sensor motion is required to test the stationary object hypothesis. During the start of the inter-frame estimation stage, no sensor motion estimate is available. The startup or bootstrapping stage uses a seeding process described in section 4.6 to resolve this problem.

4.2 Inter-frame Sensor Motion

This section provides a description of the inter-frame sensor motion estimation. It shows how a set of normal image velocity measurements is combined to estimate the inter-frame sensor motion. A by-product of the estimation process in the error covariance matrix for the inter-frame sensor motion. An expression for the expected error is derived.

If the image features belonging to stationary objects are identified, and the depth z is known, the inter-frame sensor motion can be estimated from a set of normal image velocities using weighted least squares:

$$\bar{\theta} = Q_{int}^{-1}\bar{p}, \tag{4.177}$$

where $Q_{int} = \sum_i w_i \bar{J}_i \bar{J}_i^T$, $\bar{p} = \sum_i w_i \bar{J}_i V_n(i)$, and w_i is a weighting term. The weighting term w is defined as the inverse of the expected squared error in V_n, discounted by $\beta_{discount}$ to account for feature overlap; that is

$$w = \frac{\beta_{discount}}{E[(\delta V_n)^2]}, \tag{4.178}$$

where $E[\,]$ denotes expected value. The matrix Q_{int} is referred to as the "Hessian" matrix; $\bar{p}$ is referred to as the "measurement vector." Note that the inverse of the Hessian matrix is the error covariance matrix for the inter-frame sensor motion.

The expected squared error of the normal image velocity has two components: measurement error and estimation error. The expected squared error in the measured normal image velocity is given by

$$E[(\Delta V_{n,meas})^2] = (\frac{\omega_t}{\omega_n})^2(\frac{\delta\omega_n}{\omega_n})^2 + \frac{\sigma_G^2}{(m_{ave}\,\omega_n)^2}. \tag{4.179}$$

The expected squared error in the estimated normal image velocity is given by [1]

$$E[(\Delta V_{n,est})^2] = (\frac{\delta\bar{J}}{\delta\phi_n}\bar{\theta})^2\, E[(\delta\phi_n)^2] + (\frac{\delta\bar{J}}{\delta\hat{x}}\bar{\theta})^2\, E[(\delta\hat{x})^2] + (\frac{\delta\bar{J}}{\delta\hat{y}}\bar{\theta})^2\, E[(\delta\hat{y})^2]$$

[1] Equation (4.180) assumes that the error in each of ϕ_n, $\hat{x}$, $\hat{y}$, and $d_{\hat{x}}$ are uncorrelated.

$$+(\frac{\delta \bar{J}}{\delta d_{\hat{x}}}\bar{\theta})^2 \; E[(\delta d_{\hat{x}})^2]. \tag{4.180}$$

If the sensor motion is known, the sensitivity derivatives of the normal image velocity estimate can be obtained from the derivative of $\bar{J}$ with respect to the parameters ϕ_n, $\hat{x}$, $\hat{y}$, and $d_{\hat{x}}$. Unfortunately, the sensor motion estimate may change as more measurements are incorporated.

To reduce the effect of the changing sensor motion estimate, the sensitivity derivative can be approximated using stable quantities. The most stable quantities are obtained from the Gabor sampling lattice: $\hat{x}$, $\hat{y}$, $\tilde{\omega}_k$, and $\tilde{\phi}_l$. Other stable quantities are estimated locally from the stereo image sequence: $d_{\hat{x}}$ (or z), V_n, ϕ_n and t_{col}. A moderately stable quantity is the image velocity:

$$\begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix} = AR_{os}B\,\bar{\theta}. \tag{4.181}$$

The image velocity is usually insensitive to changes in the inter-frame sensor motion estimate because errors in $\bar{\theta}$ are primarily due to the difficulty in distinguishing between rotation about the x-axis (y-axis) and translation along the y-axis (x-axis). Even if there is a confusion between sensor translation and rotation, the errors tend to balance, leaving a good estimate of the image velocity.

The approximate sensitivity derivatives of the estimated normal image velocity are listed below. The sensitivity to errors in the normal direction is given by

$$\frac{\delta V_n}{\delta \phi_n} = [-\sin\phi_n \;\; \cos\phi_n] \begin{bmatrix} V_{\hat{x}} \\ V_{\hat{y}} \end{bmatrix}. \tag{4.182}$$

The expected squared error in V_n due to errors in the image coordinates is approximated by

$$(\frac{\delta \bar{J}}{\delta \hat{x}}\bar{\theta})^2 \; E[(\delta\hat{x})^2] + (\frac{\delta \bar{J}}{\delta \hat{y}}\bar{\theta})^2 \; E[(\delta\hat{y})^2] \approx (\frac{\delta \acute{x}}{t_{col}})^2. \tag{4.183}$$

Only image coordinate errors in the normal direction affect the normal image velocity. The error along the rotated axis $\acute{x}$ is used as an approximation to the normal coordinate error:

$$\delta\acute{x} = 0.5\Delta\acute{x}_s = \frac{\pi}{2\tilde{\omega}_k}. \tag{4.184}$$

This error is equal to half the resolution of the Gabor spatial sampling lattice. The sensitivity of the estimated normal image velocity with respect to disparity errors is difficult to calculate using only stable quantities. The portion of normal image velocity caused by sensor translation is sensitive to errors in disparity; the rotation-induced normal image velocity is almost independent of disparity. To ensure repeatability, it is assumed that the normal image velocity is caused by sensor translation only. Under the pure translation assumption, the sensitivity due to disparity errors is given by

$$\frac{\delta V_n}{\delta d_{\hat{x}}} = \frac{V_n}{d_{\hat{x}}}. \tag{4.185}$$

Using the above approximations, the expected squared error in the normal image velocity is

$$E[(\delta V_n)^2] = E[(\delta V_{n,meas})^2] + E[(\delta V_{n,est})^2] + e_v^2, \tag{4.186}$$

where e_v is a constant term used to compensate for approximation errors. Note that the estimate of $E[(\delta V_n)^2]$ ignores the correlation [2] between $E[(\delta V_{n,meas})^2]$ and $E[(\delta V_{n,est}^2]$. As a result, the expected squared error is slightly over-estimated.

4.3 Mahalanobis Distance

In order to use (4.177) it is necessary to exclude feature belonging to moving objects. The "Mahalanobis distance" [6] can be used to test the hypothesis that a given normal image velocity measurement belongs to a stationary object. The Mahalanobis distance

[2] The measured normal image velocity and disparity share a common image. The noise in that image appears in both $E[(\delta V_{n,meas})^2]$ and $E[(\delta d_{\hat{x}})^2]$.

is an image velocity-based hypothesis tester that compares the prediction error with the expected error. In this application, the squared Mahalanobis distance is given by

$$d^2_{mah} = \frac{e^2}{E[e^2]}, \tag{4.187}$$

where $e = V_n - \bar{J}^T\bar{\theta}$. The expected squared error, $E[e^2]$, contains two parts: the expected squared error due to measurement noise, and the expected squared error due to motion parameter uncertainty:

$$E[e^2] = E[(\delta V_n)^2] + \bar{J}^T Q^{-1}_{int} \bar{J}. \tag{4.188}$$

A threshold is applied to the Mahalanobis distance to identify measurements that are inconsistent with the estimated sensor motion $\bar{\theta}$ and the stationary object assumption [27].

Note that the Mahalanobis distance requires Q^{-1}_{int} and $\bar{\theta}$. In order to insure that the inverse of Q_{int} exists and the current estimate of $\bar{\theta}$ is accurate, it is necessary to "seed" Q_{int} and $\bar{p}$ using measurements belonging to known stationary objects. The seeding process is examined in section 4.6.

4.4 Kalman Filtering

This section presents two Kalman filter implementations: one for the extended sensor motion; the other for the extended object motion. A batch solution to the extended sensor motion is provided to facilitate understanding of the transformation from the predominantly rectilinear model to the pure translation model of motion. This transformation effectively stabilizes the stereo image sequence. The recursive implementation of the Kalman filter follows. As part of the Kalman filter, the state transition and measurement equations for the sensor and object cases are defined.

Equation (4.177) estimates the inter-frame sensor motion; that is, the three-dimensional sensor motion over a short time interval. Because of the correlation between the x (y)

translation and y (x) rotation, the Hessian can be ill-conditioned. This ill-conditioning occurs when the features used in (4.177) are poorly distributed in three-dimensional space. Even if the feature are well distributed throughout the image, insufficient variation in depth relative to the average depth will lead to two small eigenvalues in the Hessian matrix (see section 4.6.3). Temporal consistency can be used to improve the three-dimensional velocity estimates.

The predominantly rectilinear model of sensor motion assumes that the vehicle translation is invariant over time. The sensor translation, however, will change as the observer coordinate frame is rotated. If, for the moment, the effect of observer coordinate rotation is ignored, the three-dimensional motion over longer image sequences can be estimated by minimizing

$$r_{seq} = \sum_i r_{niv}(i), \tag{4.189}$$

subject to

$$T_x(i) = T_x, \tag{4.190}$$

$$T_y(i) = T_y, \tag{4.191}$$

$$T_z(i) = T_z, \tag{4.192}$$

where i denotes the inter-frame motion over the time interval t_i to t_{i+1}. For the purpose of solving this constrained minimization, consider the normal image velocity of a feature being tracked over the image sequence. If the feature belongs to a stationary object, the transformation from sensor motion to normal image velocity is given by

$$\begin{bmatrix} V_n(0) \\ V_n(1) \\ \vdots \\ V_n(n) \end{bmatrix} = \begin{bmatrix} \bar{J}_T^T(0) & \bar{J}_\Omega^T(0) & 0 & \cdots & 0 \\ \bar{J}_T^T(1) & 0 & \bar{J}_\Omega^T(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \bar{J}_T^T(n) & 0 & 0 & \cdots & \bar{J}_\Omega^T(n) \end{bmatrix} \begin{bmatrix} \bar{T} \\ \bar{\Omega}(0) \\ \bar{\Omega}(1) \\ \vdots \\ \bar{\Omega}(n) \end{bmatrix}, \tag{4.193}$$

where the number in brackets denotes the inter-frame interval. A set of normal image velocities over the entire image sequence are combined to estimate the sensor motion.

A weighted least square estimate of the sensor motion is formulated using the set of inter-frame Hessian matrices and measurement vectors obtain from different time intervals. Using the block form of the inter-frame measurement vector and Hessian matrix, $\bar{p}(i) = [\bar{p}_a(i)\ \bar{p}_b(i)]^T$, and

$$Q_{int}(i) = \begin{bmatrix} Q_a(i) & Q_b(i) \\ Q_b^T(i) & Q_c(i) \end{bmatrix} \tag{4.194}$$

respectively, the least square solution is given by

$$\begin{bmatrix} \bar{T} \\ \bar{\Omega}(0) \\ \bar{\Omega}(1) \\ \vdots \\ \bar{\Omega}(n) \end{bmatrix} = \begin{bmatrix} \sum_i Q_a(i) & Q_b(0) & Q_b(1) & \cdots & Q_b(n) \\ Q_b^T(0) & Q_c(0) & 0 & \cdots & 0 \\ Q_b^T(1) & 0 & Q_c(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ Q_b^T(n) & 0 & 0 & \cdots & Q_c(n) \end{bmatrix}^{-1} \begin{bmatrix} \sum_i \bar{p}_a(i) \\ \bar{p}_b(0) \\ \bar{p}_b(1) \\ \vdots \\ \bar{p}_b(n) \end{bmatrix}. \tag{4.195}$$

The least square estimate of the extended sensor translation is given by (see appendix B for details)

$$\bar{T} = \Delta^{-1} \sum_i [\bar{p}_a(i) - Q_b(i) Q_c^{-1}(i) \bar{p}_b(i)] \tag{4.196}$$

where

$$\Delta = \sum_i [Q_a(i) - Q_b(i) Q_c^{-1}(i) Q_b^T(i)]. \tag{4.197}$$

Note that the extended sensor translation is calculated without explicitly determining the set of inter-frame rotations. This implicit calculation is achieved by decoupling the effects of rotation from the inter-frame matrices Q_{int} and $\bar{p}$. The new decoupled matrices, containing only inter-frame translational information, are given by

$$Q_T = Q_a - Q_b Q_c^{-1} Q_b^T, \tag{4.198}$$

$$\bar{p}_T = \bar{p}_a - Q_b Q_c^{-1} \bar{p}_b, \tag{4.199}$$

Equation (4.196) can be written as a temporal integration of the decoupled inter-frame matrices:

$$\bar{T} = [\sum_i Q_T(i)]^{-1} [\sum_i \bar{p}_T(i)]. \tag{4.200}$$

Equation (4.200) considers the entire image sequence as a batch. A recursive formulation is most useful for autonomous vehicle applications.

In this work the Kalman filter is used to integrate the sensor motion over the entire image sequence. The Kalman filter is a recursive parameter estimator that requires a model of the underlying process and a model of the measurements of the process. In this application, the parameters (state variables) to be estimated are the extended sensor translation along the x, y, and z-axes. The process is the sensor motion and the measurements are the normal image velocities at various points and at various times. The model of sensor motion includes the effects of the observer rotation.

To understand how the observer rotation is included in the Kalman filter, it is necessary to review the inter-frame sensor motion. The inter-frame sensor motion is a discrete approximation of the instantaneous sensor motion. In the instantaneous sensor motion, all translation and rotation components occur simultaneously. The inter-frame sensor motion must be described sequentially. In this work, it is assumed that the translation occurs before the rotation. Figure 4.13 illustrates the transformation of the old observer coordinate frame at time t_i into the new observer frame at time t_{i+1}. The new observer frame is obtained by translating, then rotating, the old observer frame using the inter-frame sensor motion:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_{t(i+1)} = R^T(t_i) \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{t(i)} + \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} \Delta t \right\}, \tag{4.201}$$

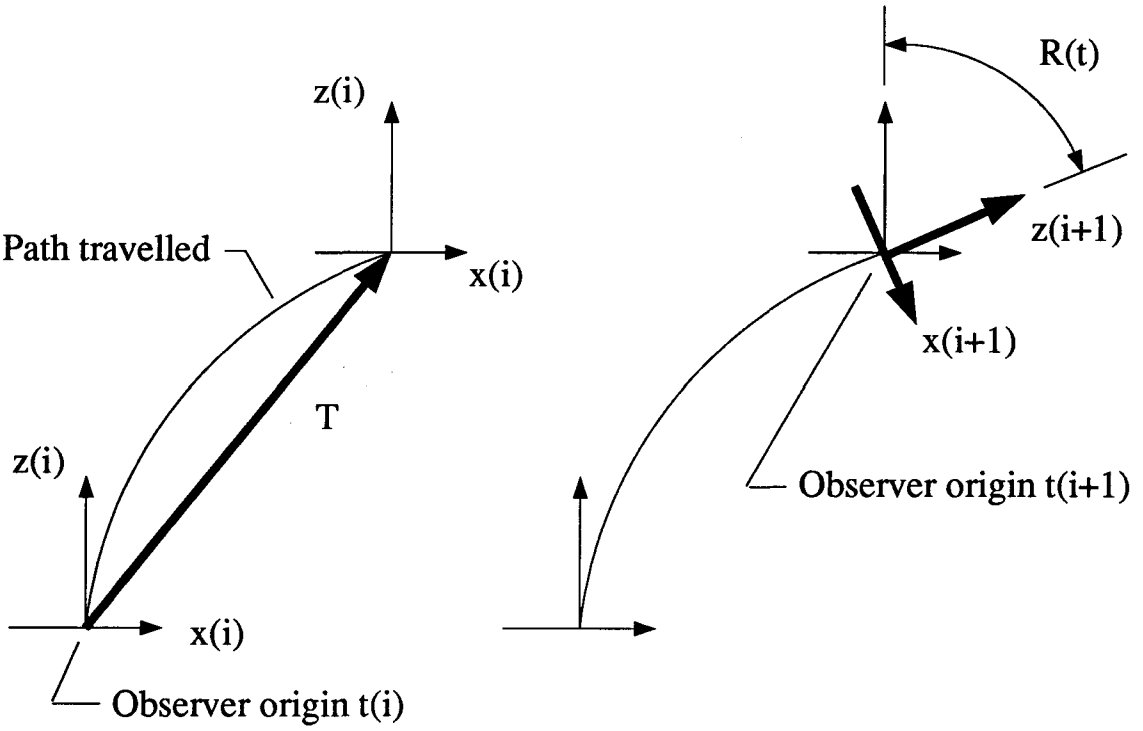


Figure 4.13: Model of Inter-frame Sensor Motion

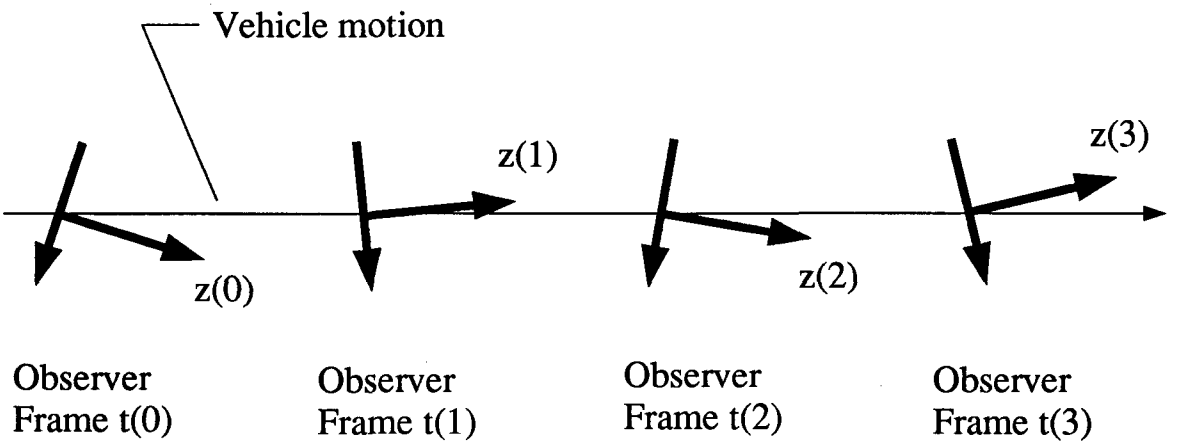


Figure 4.14: Model of the Extended Sensor Motion

where Δt is the time between successive images $(t_{i+1} - t_i)$. The rotation matrix R, using the small angle approximation, is given by

$$R = \begin{bmatrix} 1 & -\Omega_z \Delta t & \Omega_y \Delta t \\ \Omega_z \Delta t & 1 & -\Omega_x \Delta t \\ -\Omega_y \Delta t & \Omega_x \Delta t & 1 \end{bmatrix}. \tag{4.202}$$

The model of the extended sensor motion (the process model) is shown in figure 4.14. The vehicle (or equivalently, the origin of the observer coordinate frame) is translating at a constant velocity. The stereo cameras are undergoing transient rotations; the orientation of the observer coordinate frame is changing. In this work, the extended sensor translation is represented using the current observer coordinate frame. In order to predict the translation at t_{i+1}, the R matrix must be included in the process model to account for any changes in the orientation of the observer coordinate frame during the inter-frame transition from t_i to t_{i+1}. The extended sensor motion is describe in recursive form as

$$\bar{T}_{sen}(t_{i+1}) = R^T(t_i)\, \bar{T}_{sen}(t_i) + \bar{n}_p, \tag{4.203}$$

where $\bar{n}_p$ is the process noise vector. The state vector $\bar{T}_{sen}$ represents the estimate of the extended sensor translation, not the inter-frame translation.

The measurement model is given by

$$V_n = \bar{J}^T \bar{\theta} + n_m, \tag{4.204}$$

where n_m is the measurement noise. Note that the state vector for the process model ($\bar{T}_{sen}$) and the measurement model ($\bar{\theta}$) are different. It is necessary to convert the inter-frame sensor motion into a more compatible format. The inter-frame sensor motion combines a large number of normal image velocity measurements obtained at a given time instant. A model that combines all normal image velocity measurements obtained at time instant t_i can be described as

$$\bar{p} = Q_{int}\, \bar{\theta}. \tag{4.205}$$

The measurement noise is incorporated into the Hessian matrix Q_{int}. The measurement model (4.205) still needs modification. In order to use the inter-frame motion information in the Kalman filter, the effects of rotation must be decoupled from Q_{int} and $\bar{p}$. Using the decoupled matrices, Q_T and $\bar{p}_T$, the new measurement model becomes

$$\bar{p_T} = Q_T \bar{T}_{sen}. \tag{4.206}$$

The decoupling of the rotational parameters effectively stabilizes the image sequence, allowing the use of a pure translation model for extended sensor motion.

The following equations are a modified version of the "alternative Kalman filter" presented in [11]:

$$Q_{sen}(t_1) = Q_{sen}(t_1/t_0) + Q_T, \tag{4.207}$$

$$\bar{T}_{sen}(t_1) = \bar{T}_{sen}(t_1/t_0) + Q_{sen}^{-1}(t_1)\,[\bar{p}_T - Q_T \bar{T}_{sen}(t_1/t_0)], \tag{4.208}$$

$$\bar{T}_{sen}(t_2/t_1) = R^T(t_1)\,\bar{T}_{sen}(t_1), \tag{4.209}$$

$$Q_{sen}(t_2/t_1) = [R^T\, Q_{sen}^{-1}(t_1)R]^{-1} = R^T(t_1)Q_{sen}(t_1)R(t_1), \tag{4.210}$$

where Q_{sen} is the Hessian matrix for the extended sensor motion. The notation (t_1/t_0) indicates a prediction for time t_1 based on the estimate at time t_0. The second equality in (4.210) assumes that the rotation matrix R is orthonormal. This matrix is not orthonormal, but for small rotational angles the error in making such an assumption is small. Note that the inverse of the Hessian matrix, Q_{sen}^{-1}, is the error covariance matrix of the extended sensor motion.

The Kalman filter equations, (4.207), (4.208), (4.209), and (4.210), describe two distinct operations: updating Hessian and state variables with new data, and predicting the next Hessian and state variables from past information. In (4.207), the decoupled inter-frame Hessian Q_T is added to the predicted Hessian $Q_{sen}(t_1/t_0)$. The addition of new information reduces the error covariance of the translation estimates; new information improves the estimate of the extended sensor translation. Equation (4.208) updates the estimate of the extended sensor translation. The measurement error,

$$\bar{p}_T - Q_T\bar{T}_{sen}(t_1/t_0), \tag{4.211}$$

weighted by the error covariance, adjusts the predicted translation. The predicted translation is recursively defined by (4.209). The rotation matrix R transforms the translation estimates into next observer coordinate frame representation. Equation (4.210) is used to predict the next Hessian. In its current form, (4.210) only transforms the Hessian to the next observer frame representation. It is only valid when the process noise vector is zero over the entire sequence; it does not allow for deviations from the modelled vehicle motion trajectory. The process noise must be modelled if the stereo image sequence is long. If the covariance structure of the motion disturbance is not known, a forgetting factor (λ_{forget}) can be incorporated to reduce the influence of older data:

$$Q_{sen}(t_2/t_1) = \lambda_{forget}R^T(t_1)Q_{sen}(t_1)R(t_1) \tag{4.212}$$

where $0 < \lambda_{forget} \leq 1$. When the motion disturbance has a known error covariance structure, N_{sen}, (4.210) can be modified as follows:

$$Q_{sen}(t_2/t_1) = [R^T\, Q_{sen}^{-1}(t_1)R + N_{sen}]^{-1}. \tag{4.213}$$

In section 4.6.6, it is shown how rotational terms and cross terms from the inter-frame error covariance matrix can be used in N_{sen}.

The translational velocity of each object can be integrated over time using a Kalman filter. By subtracting the estimated normal image velocity due to sensor motion, the excess normal image velocity can be integrated. The translational velocity estimated using the excess normal image velocity is relative to the world coordinate frame. The process model for the object motion is given by

$$\bar{T}_{obj}(t_{i+1}) = R^T(t_i)\, \bar{T}_{obj}(t_i). \tag{4.214}$$

The inter-frame sensor rotation is included in the process model to account for changes in the orientation of the observer coordinate frame. The measurement model for the object motion is given by

$$V_{n,excess} = V_n - \bar{J}^T\bar{\theta} = \bar{J}_{obj}^T\bar{T}_{obj}, \tag{4.215}$$

where $\bar{J}_{obj}^T = \bar{n}^T A R_{os}$.

The Kalman filter equations for the object motion are as follows:

$$Q_{obj}(t_1) = Q_{obj}(t_1/t_0) + w\bar{J}_{obj}\bar{J}_{obj}^T, \tag{4.216}$$

$$\bar{K}(t_1) = w\, Q_{obj}^{-1}(t_1)\bar{J}_{obj}, \tag{4.217}$$

$$\bar{T}_{obj}(t_1) = \bar{T}_{obj}(t_1/t_0) + \bar{K}(t_1)\, [V_{n,excess} - \bar{J}_{obj}^T\bar{T}_{obj}(t_1/t_0)], \tag{4.218}$$

$$\bar{T}_{obj}(t_2/t_1) = R^T(t_1)\, \bar{T}_{obj}(t_1), \tag{4.219}$$

$$Q_{obj}(t_2/t_1) = R^T(t_1)Q_{obj}(t_1)R(t_1), \tag{4.220}$$

where Q_{obj} is the Hessian matrix for the extended object motion, and $w^{-1} = E[(\delta V_{n,excess})^2]$. The expected squared error in the excess normal image velocity is given by

$$E[(\delta V^2_{n,excess}] = E[(\delta V_n)^2] + \bar{J}^T Q^{-1}_{int} \bar{J}. \tag{4.221}$$

If the moving objects are processed after the inter-frame sensor motion, $E[(\delta V^2_{n,excess}] \approx E[(\delta V_n)^2]$. The vector $\bar{K}$ is referred to as the "Kalman gain." A high $\bar{K}$ indicates the influx of important new data.

As in the extended sensor motion case, the Hessian prediction equation, (4.220), must be modified to account for process noise if the stereo sequence is long (see section 4.6.6). If the unmodified version of (4.220) is used, the Kalman gain will approach zero as time elapses.

4.5 Estimating Collision Parameters

This section uses the extended object and sensor translation to estimate the collision parameters. The collision parameters, the point-of-collision and the time-to-collision, are important because they indicate if and when an object will collide into the sensor. It is also important, particularly for obstacle avoidance, to determine the accuracy of the collision parameters. The object and sensor motion error covariance matrices are used to estimate the expected error for the point of collision and the time to collision.

The relative velocity of each object can be estimated using the integrated object motion and the integrated sensor motion:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \bar{T}_{obj} - \bar{T}_{sen}. \tag{4.222}$$

The collision parameters for each object are given by

$$t_{col} = \frac{z}{\dot{z}}, \tag{4.223}$$

$$x_{col} = x + \dot{x}\, t_{col}, \tag{4.224}$$

$$y_{col} = y + \dot{y}\, t_{col}. \tag{4.225}$$

Uncertainty in the position and velocity of the object relative to the observer produces uncertainty in the collision parameters:

$$\begin{bmatrix} \delta x_{col} \\ \delta y_{col} \\ \delta t_{col} \end{bmatrix} = H_{col} \begin{bmatrix} \delta \dot{x} \\ \delta \dot{y} \\ \delta \dot{z} \\ \delta \hat{x} \\ \delta \hat{y} \\ \delta d_{\hat{x}} \end{bmatrix}, \tag{4.226}$$

where

$$H_{col} = [H_{col,v}\ H_{col,p}] = \begin{bmatrix} t_{col} & 0 & -\hat{x}_{foe} t_{col} & \frac{z}{z_f} & 0 & \frac{x_{col}}{d_{\hat{x}}} \\ 0 & t_{col} & -\hat{y}_{foe} t_{col} & 0 & \frac{z}{z_f} & \frac{y_{col}}{d_{\hat{x}}} \\ 0 & 0 & -\frac{t_{col}}{\dot{z}} & 0 & 0 & \frac{t_{col}}{d_{\hat{x}}} \end{bmatrix}. \tag{4.227}$$

The collision parameter uncertainty expression (4.226) has been linearized using a first-order Taylor series expansion about the actual collision parameters, $(x_{col}, y_{col}, t_{col})$. In this work, the estimated collision parameters are used instead of the actual.

In order to determine the expected error in the collision parameters, it is necessary to determine the error covariance of the relative position and velocity of the object. The error covariance for the position of a feature is

$$Q_p^{-1} = \begin{bmatrix} (\delta \hat{x})^2 & 0 & 0 \\ 0 & (\delta \hat{y})^2 & 0 \\ 0 & 0 & (\delta d_{\hat{x}})^2 \end{bmatrix}. \tag{4.228}$$

The error covariance of the relative translation for a moving object is given by

$$Q_{rel}^{-1} = Q_{obj}^{-1} + Q_{sen}^{-1}. \tag{4.229}$$

There is a cross-correlation between the position and velocity estimates, which is given by

$$Q_{pv} = \left[\begin{array}{cc} \bar{v}_p & \bar{v}_p \quad (\frac{w}{ds^2} \bar{J}_{obj} \bar{J}_{obj}^T \bar{T}_{obj}) \end{array} \right], \tag{4.230}$$

where $\bar{v}_p = [0 \ 0 \ z^{-2}]^T$. The effect of Q_{pv} is insignificant when the extended object translation is estimated over a sequence of image and/or using many features (grouping of moving object features is discussed in section 4.6.4). In such cases, Q_{pv} can be approximated as 0_{3x3}.

The expected error for the collision parameters of features belonging to stationary objects are calculated in a similar manner. The position error covariance is the same as (4.228). The error covariance for the relative translation is set equal to the error covariance for the extended sensor motion; that is,

$$Q_{rel}^{-1} = Q_{sen}^{-1}. \tag{4.231}$$

Since the sensor motion is estimated using a large number of features, the cross-correlation between the position and velocity is considered insignificant ($Q_{pv} \approx 0_{3x3}$).

The error covariance matrix for the collision parameters, for both the moving and stationary objects, is given by

$$Q_{col}^{-1} = H_{col} \left[\begin{array}{cc} Q_{rel} & Q_{pv}^T \\ Q_{pv} & Q_p \end{array} \right]^{-1} H_{col}^T. \tag{4.232}$$

When the cross-correlation between position and velocity is small, the error covariance is approximated by

$$Q_{col}^{-1} = H_{col,v} Q_{rel}^{-1} H_{col,v}^T + H_{col,p} Q_p^{-1} H_{col,p}^T. \tag{4.233}$$

Note that, for a constant relative translation ($\dot{x}$, $\dot{y}$, $\dot{z}$), the following terms are time-invariant: the object's focus of expansion ($\hat{x}_{foe}$, $\hat{y}_{foe}$) and the object's point-of-collision (x_{col}, y_{col}). If the relative translation and the error covariance matrices Q_{rel}^{-1} and Q_p^{-1} are constant, the first term in (4.233) is proportional to the square of t_{col}; the second term is proportional to the square of z. Thus, the accuracy of the collision parameters will improve as the object approaches the observer.

4.6 Implementation Details

This section explains implementation details that are necessary to run the proposed algorithm. It examines establishing feature correspondences, seeding the Hessian matrix to avoid startup problems, and grouping features that belong to a common moving object. It discusses how to improve inter-frame sensor motion estimates by exploiting constraints such as planar motion. Two extensions to the Kalman filter equations are proposed: the incorporation of rotational uncertainty and pilot commands into the process model.

4.6.1 Feature Correspondence

This subsection discusses how feature correspondence, stereo and temporal, are established. Stereo correspondence exploits a priori information from lower frequency Gabor channels, past measurements, and heuristic spatial constraints. Temporal correspondence is guided by the estimate of the inter-frame sensor motion and object motion.

The concept of feature correspondence is simple: given a feature in one image, find the feature in a companion image that corresponds to the same physical feature within the scene. The implementation of a correspondence method, however, is difficult. In the previous chapter, it was shown how local attributes can be compared to test the plausibility of two lattice points, each from different images, belonging to the same object

feature. This local test will reject some feature pairings, but there is still the possibility of multiple candidate matches. In this subsection, a priori information is used to limit the search space such that a single best match or no match is found. Once a feature is found in a given image, the a priori prediction of the position of the corresponding feature is used to define the search space in the companion image. The search space for a disparity measurement consists of the two lattice points along the epipolar line that are closest to the predicted position. The search space for a normal image velocity measurement consists of the four closest lattice points surrounding the predicted position. Note that the search space is defined in terms of lattice points within a Gabor channel. The search space for the lower frequency Gabor channels spans a larger number of pixels than the higher frequency channels. As a result, greater prediction accuracy is required for high frequency channels.

The multi-scale approach to feature correspondence is useful for estimating disparity. Coarse information from low frequency channels are used to predict the disparity in higher frequency channels. Direct predictions are not always possible. The application of the three thresholds (see section 3.2.2) to the magnitude map usually produces sparsely distributed features, and subsequently, sparsely distributed disparity measurements. The disparity estimates are interpolated to fill-in the disparity map for each epipolar channel. Interpolated disparity estimates from lower frequency epipolar channels are used to select the appropriate epipolar offset in the higher frequency epipolar channels. Since disparity measurements are obtained from the epipolar channel, interpolated data is also used to estimate disparity of features found in oblique channels of the same frequency.

Accompanying the disparity estimates are the expected disparity errors. The disparity errors are used to weight the interpolation process. The interpolation of disparity estimates consists of two stages: a filling-in stage, and a four-point interpolation stage. In the first stage, every lattice point in the epipolar channel is assigned a disparity value

and an error value. Direct measurements of disparity and error are left unchanged. Lattice points near direct measurements are assigned a similar disparity value but a larger error. A drift penalty assigns progressively larger errors as the distance from a direct measurement increases. In this work, the disparity is estimated from a weighted sum of neighbouring lattice points:

$$\bar{d}_{\hat{x}}(0,0) = \frac{1}{a_{sum}}[a_{0,0}d_{\hat{x}}(0,0) + a_{0,1}d_{\hat{x}}(0,1) + a_{1,0}d_{\hat{x}}(1,0) + a_{0,-1}d_{\hat{x}}(0,-1) + a_{-1,0}d_{\hat{x}}(-1,0)], \tag{4.234}$$

where

$$a_{sum} = a_{0,0} + a_{0,1} + a_{1,0} + a_{0,-1} + a_{-1,0}, \tag{4.235}$$

$$a_{i,j}^{-1} = \begin{cases} E[(\delta d_{\hat{x}})^2] & \text{if } i = j = 0 \\ E[(\delta d_{\hat{x}} + e_{drift})^2] & \text{otherwise,} \end{cases} \tag{4.236}$$

and e_{drift} is the drift penalty. The interpolation error is set to the minimum error in the neighbourhood:

$$\bar{a}_{0,0} = \min[a_{0,0}\ a_{0,1}\ a_{1,0}\ a_{0,-1}\ a_{-1,0}]. \tag{4.237}$$

Once the filling-in stage is complete, each lattice point in the epipolar channel will have a disparity and error estimate.

For oblique channels, the disparity and error are interpolated from the epipolar channel. A weighted average of the four nearest points provide the off-epipolar disparity estimates. The error is calculated in a similar fashion. This interpolation scheme is also used to project initial (predicted) values of disparity and error into higher frequency epipolar channels. The initial estimates are used to select the appropriate epipolar offset and to aid the subsequent filling-in stage. A cross-scale penalty is added to the error estimate to reduce the influence of the lower frequency data.

It is necessary to estimate the disparity for the lowest frequency channel without an a priori estimate. A histogram of the epipolar offset of all possible candidate matches

is formed. To reduce the number of potential matches, an additional test using the normalized moment of inertia is performed (see section 3.2.2). The dominant mode and the larger of its two neighbours are selected as the candidate epipolar offsets for the entire channel. Since the lowest frequency lattice contains a small number of points, the computational complexity is small. In addition, since the spatial extent of low frequency Gabor functions is large, the range of measurable depths is quite large.

The multi-scale approach is not perfect; no match is made when both candidate epipolar offsets fail the local attribute test. Past information is also used to establish matches. Past depth measurements from the channel of interest are projected to the current time instant. The image velocities due to sensor and object motion predict the changes in the lattice positions of corresponding points. In most cases, the temporal estimate of the epipolar offset will confirm the scale-based prediction. There will be instants when the temporal estimates will produce extra matches. These additional feature correspondences will increase the density of the stereo matches and improve the certainty of the interpolated disparity map.

The density of stereo matches can be increased by enforcing a heuristic ordering constraint. If a surface is sufficiently smooth, corresponding features along an epipolar line will appear in the same order in the left and right images [32]. The application of the heuristic ordering constraint is shown in figure 4.15. The existing correspondences, established by the scale-based and temporal-based matching algorithms, act as boundaries for other yet unseen matches. If an unmatched feature is detected in the left image, and it is bounded by two stereo features, the currently unmatched corresponding feature in the right image must be bounded by the two corresponding stereo features (see figure 4.15). This limited space is searched for a potential match. If one unambiguous match is found, the correspondence is established. In this work, multiple candidate matches are ignored. Multiple candidate matches could be resolved using dynamic programming

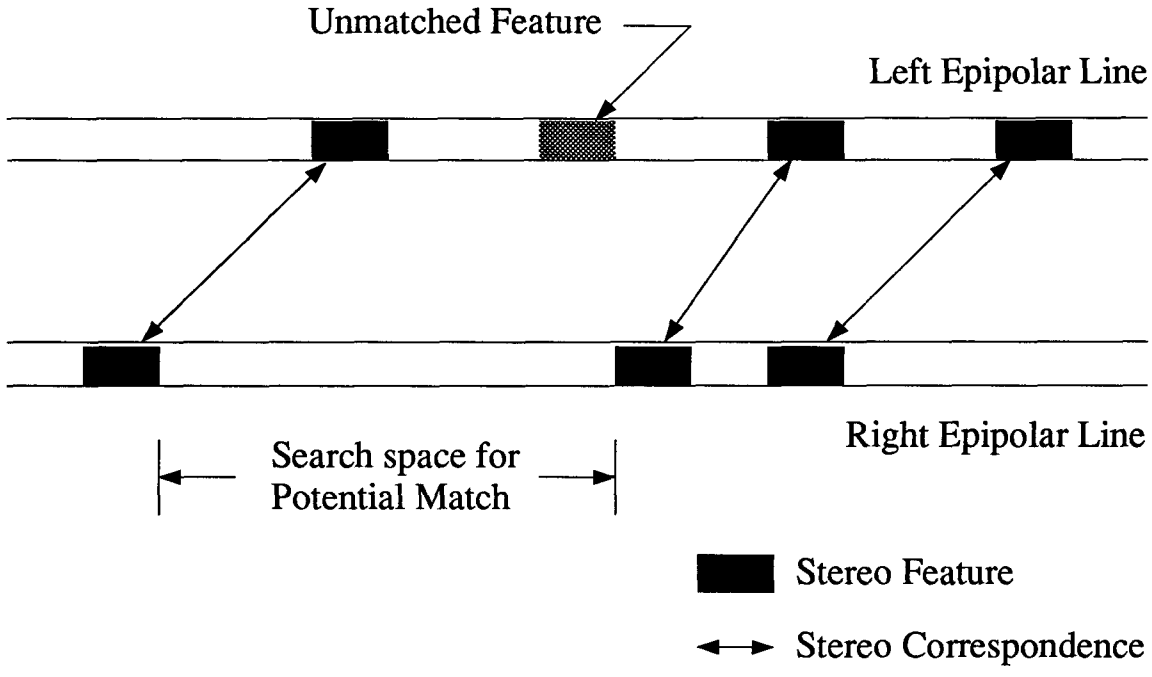


Figure 4.15: Heuristic Ordering Constraint

[40]. Once the heuristic ordering constraint instantiates a correspondence, the temporal constraint will propagate the correspondence over time.

A multi-scale approach, embedded within a cycle of perception, is useful for creating temporal correspondences and for estimating normal image velocities. Candidate temporal correspondences are generated using the estimates of inter-frame sensor motion and extended object motion, as shown in figure 4.16. The inter-frame sensor motion, estimated from the lower frequency channels [3], is used to predict the temporal shift in the lattice position of a feature belonging to a stationary object. The object motion, estimated using past information from the channel of interest (see subsection 4.6.4), is used along with the sensor motion to predict the lattice shift for moving object features.

[3]At the lowest frequency channel, the predicted image velocity is zero.

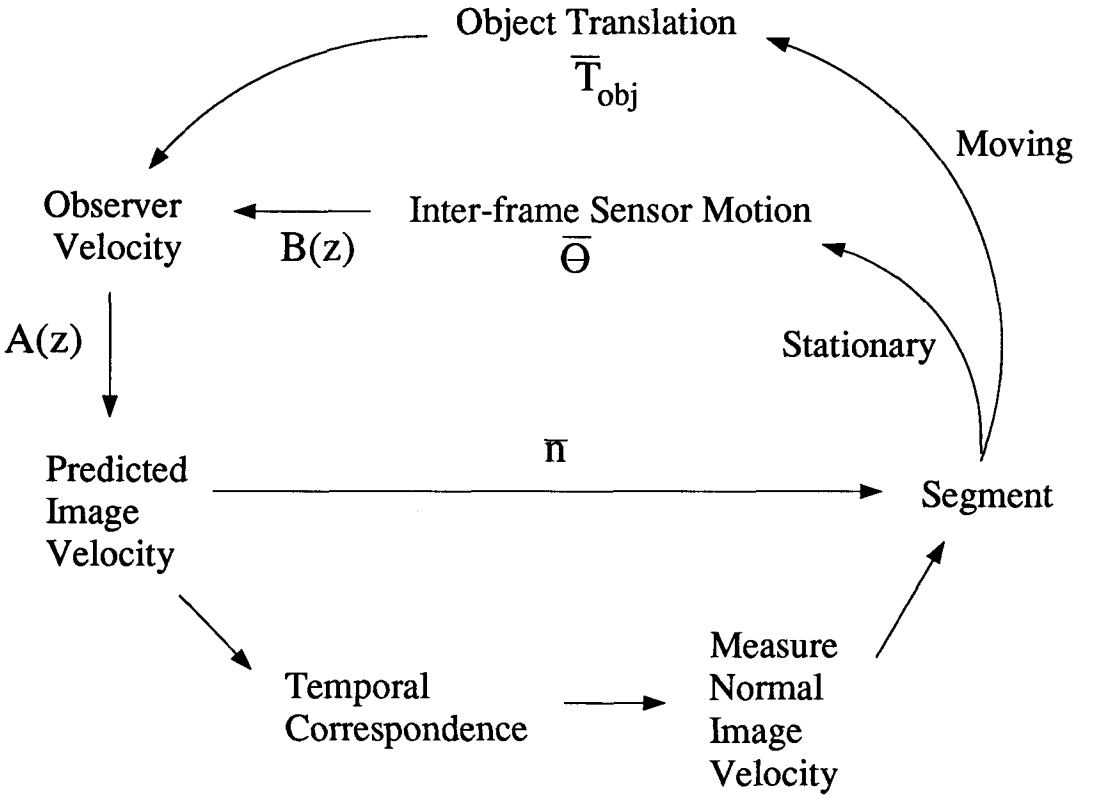


Figure 4.16: Cycle of Perception

In the epipolar channel, both stereo and temporal feature correspondences can be made. This four point correspondence produces a stereo motion detector, such as described in section 2.6. The stereo motion detector, also referred to as a "trajectory detector," is tuned to a preferred three-dimensional observer trajectory and has a finite velocity bandwidth.

As mentioned previously, candidate temporal correspondences are predicted from the sensor and object motion. For an epipolar channel, the candidate correspondences (the lattice offsets that are closest to the predicted image displacement) are generated using [4]

$$\begin{bmatrix} V_{\hat{x},R} \\ V_{\hat{x},L} \\ V_{\hat{y}} \end{bmatrix} = H_v[B(z)\bar{\theta} + \bar{T}_{obj}], \tag{4.238}$$

where

$$H_v = \frac{1}{z}\begin{bmatrix} z_f & 0 & -\hat{x}_R \\ z_f & 0 & -\hat{x}_L \\ 0 & z_f & -\hat{y} \end{bmatrix}. \tag{4.239}$$

The candidate correspondences are tested and the best match (if any) is accepted. The lattice offset along the $\hat{y}$-axis, $\hat{y}_{offset}$, is the same for both the left and right images. The lattice offsets along the $\hat{x}$-axis, $\hat{x}_{L,offset}$ in the left image and $\hat{x}_{R,offset}$ in the right image, can be different. The lattice offsets form a trajectory detector that is tuned to a preferred three-dimensional velocity $(\dot{x}_o, \dot{y}_o, \dot{z}_o)$:

$$\begin{bmatrix} \dot{x}_o \\ \dot{y}_o \\ \dot{z}_o \end{bmatrix} = H_v^{-1}\begin{bmatrix} \hat{x}_{R,offset} \\ \hat{x}_{L,offset} \\ \hat{y}_{offset} \end{bmatrix}\frac{1}{\Delta t}, \tag{4.240}$$

[4]For the case of parallel stereo cameras, the image velocity $V_{\hat{y}}$ is the same for both the right and left images.

where

$$H_v^{-1} = \frac{z}{z_f}\frac{1}{E_{offset}}\begin{bmatrix} \hat{x}_L & -\hat{x}_R & 0 \\ \hat{y} & -\hat{y} & -E_{offset} \\ z_f & -z_f & 0 \end{bmatrix}, \tag{4.241}$$

and $E_{offset} = \hat{x}_L - \hat{x}_R$. The term E_{offset} is the preferred disparity of the trajectory detector. Due to lattice quantization, the preferred velocity will not, in general, be the same as the actual velocity $B(z)\bar{\theta} + \bar{T}_{obj}$. The actual velocity will be within the velocity bandwidth of the trajectory detector.

The three-dimensional velocity bandwidth of the trajectory detector is described using a covariance matrix. If the largest change in image velocity without further offsetting the lattices is $\pm\frac{\Delta\hat{x}_s}{2\Delta t}$ or $\pm\frac{\Delta\hat{y}_s}{2\Delta t}$, the covariance matrix for the trajectory detector is defined as

$$C_{v3d} = \begin{bmatrix} E[(\Delta\dot{x})^2] & E[\Delta\dot{x}\Delta\dot{y}] & E[\Delta\dot{x}\Delta\dot{z}] \\ E[\Delta\dot{y}\Delta\dot{x}] & E[(\Delta\dot{y})^2] & E[\Delta\dot{y}\Delta\dot{z}] \\ E[\Delta\dot{z}\Delta\dot{x}] & E[\Delta\dot{z}\Delta\dot{y}] & E[(\Delta\dot{z})^2] \end{bmatrix} = H_v^{-1}\begin{bmatrix} (\frac{\Delta\hat{x}_s}{2\Delta t})^2 & 0 & 0 \\ 0 & (\frac{\Delta\hat{x}_s}{2\Delta t})^2 & 0 \\ 0 & 0 & (\frac{\Delta\hat{y}_s}{2\Delta t})^2 \end{bmatrix} H_v^{-T}. \tag{4.242}$$

If $\Delta\hat{y}_s = \alpha\Delta\hat{x}_s$, the covariance matrix can be written as

$$C_{v3d} = (\frac{z}{z_f})^2(\frac{\Delta x_s}{2E_{offset}})^2\begin{bmatrix} \hat{x}_L^2 + \hat{x}_R^2 & \hat{y}(\hat{x}_L + \hat{x}_R) & z_f(\hat{x}_L + \hat{x}_R) \\ \hat{y}(\hat{x}_L + \hat{x}_R) & 2\hat{y}^2 + \alpha^2 E_{offset}^2 & 2\hat{y}z_f \\ z_f(\hat{x}_L + \hat{x}_R) & 2\hat{y}z_f & 2z_f^2 \end{bmatrix}. \tag{4.243}$$

It can be seen that the velocity bandwidth along the coordinate axes are different:

$$\Delta\dot{x} \geq 2^{-0.5}\frac{z}{z_f}(\frac{\Delta\hat{x}_s}{2\Delta t}), \tag{4.244}$$

$$\Delta\dot{y} \geq \frac{z}{z_f}(\frac{\Delta\hat{y}_s}{2\Delta t}), \tag{4.245}$$

$$\Delta\dot{z} = 2^{0.5}\frac{z}{E_{offset}}(\frac{\Delta\hat{x}_s}{2\Delta t}). \tag{4.246}$$

The velocity bandwidth along the z-axis is usually the largest [5] of the three coordinate bandwidths.

Estimates of the object motion may be known from past images. In such cases, (4.238) can be used to predict candidate temporal correspondences. Unfortunately, the object motion $\bar{T}_{obj}$ is not always known. In such cases, candidate correspondences are generated using a default set of probable or important object translations. In this work, the default set contains the translation of stationary objects, moving objects that are currently being tracked, and objects on collision trajectories.

$\bar{T}_{obj} = \bar{0}$ is a probable object translation because it is assumed that most of the objects in the scene are stationary. A candidate correspondence predictor that assumes $\bar{T}_{obj} = \bar{0}$ is said to be "tuned to stationary objects," and is referred to as a "stationary object correspondence predictor." Such a predictor will also detect moving objects within its velocity bandwidth. Once a moving object is detected and its three-dimensional motion is estimated (see section 4.6.4), the information is used in subsequent images to predict temporal correspondences. It is important to retain the estimate of $\bar{T}_{obj}$ because the velocity bandwidth will contract as the object approaches the observer.

Once a moving object has been detected, it is useful to search for other features belonging to the same object. Such action is useful for finding moving object features in higher frequency channels where the velocity bandwidths are small. Since we are searching for features belonging to a specific object, the depth of the candidate correspondence must be close to the depth of the object. The matching depth requirement reduces the amount of computations and reduces the probability of chance (incorrect) matches. This is important when the scene contains many moving objects.

The most important objects are obstacles: objects with collision trajectories. There is

[5] The largest bandwidth is along the eigenvector of C_{v3d} that is approximately equal to $\bar{v}_{v3d} = [(\hat{x}_L + \hat{x}_R)\ 2\hat{y}\ 2z_f]$.

no guarantee that a candidate correspondence predictor tuned to stationary objects will detect an obstacle. A correspondence predictor that is tuned to objects with collision trajectories is required. The following paragraphs review collision prediction for the monocular case, then extend it for the stereo case.

For a monocular observer, a collision can be predicted from the projected image coordinates of the object, denoted by $(\hat{x}_{obj}, \hat{y}_{obj})$, and the object's focus of expansion, $(\hat{x}_{foe}, \hat{y}_{foe})$. Recalling section 2.4.1, the object's focus of expansion is defined as

$$k_t \begin{bmatrix} \hat{x}_{foe} \\ \hat{y}_{foe} \\ z_f \end{bmatrix} = \bar{T}_{obj} + B_T \bar{\theta} = \bar{T}_{obj} - \bar{T}, \tag{4.247}$$

where

$$k_t = \frac{\dot{z}_{obj} - T_z}{z_f} = -\frac{z}{z_f t_{col}}. \tag{4.248}$$

A collision will occur if $(\hat{x}_{obj}, \hat{y}_{obj}) = (\hat{x}_{foe}, \hat{y}_{foe})$. For the stereo camera setup [6], a collision will occur if the object's focus of expansion lies between the projections of the object in the left and right images; that is, if $\hat{x}_{R,obj} \leq \hat{x}_{foe} \leq \hat{x}_{L,obj}$ and $\hat{y}_{obj} = \hat{y}_{foe}$.

In order to determine which $\bar{T}_{obj}$ should be used to detect obstacles, we need to consider the stereo image velocity:

$$\begin{bmatrix} V_{\hat{x},R} \\ V_{\hat{x},L} \\ V_{\hat{y}} \end{bmatrix} = k_t H_v \begin{bmatrix} \hat{x}_{foe} \\ \hat{y}_{foe} \\ z_f \end{bmatrix} + H_v \, B_\Omega \bar{\Omega}. \tag{4.249}$$

If the obstacle will collide into the right camera, $(\hat{x}_{R,obj}, \hat{y}_{obj}) = (\hat{x}_{foe}, \hat{y}_{foe})$, the stereo

[6] For parallel stereo cameras, the object's focus of expansion will be the same in both the left and right images.

image velocity is

$$\begin{bmatrix} V_{\hat{x},R} \\ V_{\hat{x},L} \\ V_{\hat{y}} \end{bmatrix}_{(\hat{x}_R=\hat{x}_{foe})} = \frac{E_{offset}}{t_{col}} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + H_v \, B_\Omega \bar{\Omega}. \tag{4.250}$$

If the obstacle is heading for the left camera, $(\hat{x}_{L,obj}, \hat{y}_{obj}) = (\hat{x}_{foe}, \hat{y}_{foe})$, the stereo image velocity is

$$\begin{bmatrix} V_{\hat{x},R} \\ V_{\hat{x},L} \\ V_{\hat{y}} \end{bmatrix}_{(\hat{x}_L=\hat{x}_{foe})} = \frac{E_{offset}}{t_{col}} \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} + H_v \, B_\Omega \bar{\Omega}. \tag{4.251}$$

To detect an obstacle that will pass between the cameras, we wish to select $\bar{T}_{obj}$ such that

$$\begin{bmatrix} V_{\hat{x},R} \\ V_{\hat{x},L} \\ V_{\hat{y}} \end{bmatrix}_{(\hat{x}_R=\hat{x}_{foe})} - \begin{bmatrix} V_{\hat{x},R} \\ V_{\hat{x},L} \\ V_{\hat{y}} \end{bmatrix}_{(\hat{x}_L=\hat{x}_{foe})} < \frac{1}{2\Delta t} \begin{bmatrix} \Delta\hat{x}_s \\ \Delta\hat{x}_s \\ \Delta\hat{y}_s \end{bmatrix}. \tag{4.252}$$

If

$$\frac{E_{offset}}{t_{col}} < \frac{\Delta\hat{x}_s}{2\Delta t}, \tag{4.253}$$

then (4.252) is satisfied by setting $\bar{T}_{obj} = \bar{T}$. In this work, the extended sensor translation, $\bar{T}_{sen}$, is used instead of the inter-frame sensor translation $\bar{T}$. Thus, if the time-to-collision is sufficiently large, such that (4.253) is satisfied, tuning the correspondence predictor to the sensor translation $\bar{T}_{sen}$ will generate the correct correspondences for objects on collision trajectories.

In summary, the above-mentioned four point correspondences produce three types of trajectory detectors. The first type of trajectory detector is tuned to the detect stationary objects. This type is used in section 4.6.2 to identify features that belong to stationary objects. The second type of trajectory detector is tuned to the velocity of a known object at a specific depth. This type of trajectory detector is used to identify features

that belong to a specific moving object. The final type of trajectory detector is tuned to the sensor translation. This trajectory detector is used to identify features belonging to obstacles (objects that will collide into the observer).

4.6.2 Seeding the Hessian Matrix

In order to use the Mahalanobis distance as a test for stationary objects, the Hessian matrix Q_{int} must have a full rank. Thus, the Mahalanobis distance can not be used during the startup stage of estimating inter-frame sensor motion. A multi-stage seeding process is used to increase the rank of the Q_{int}. In the first stage, a priori predictions of the sensor motion (usually from auxiliary sensors), along with the associated errors increase the rank of the Hessian. The remaining stages identify features belonging to known (or at least probable) stationary objects. These stationary object features are combined to produce an initial estimate of the inter-frame sensor motion before testing the normal image velocity of other (uncertain) measurements.

A priori predictions can be obtained from other sensors, such as a speedometer, or from default values. The extended sensor translation, if available, can be used as a prediction of the inter-frame sensor translation. To illustrate how predicted parameters can be incorporated into the inter-frame sensor motion estimate, it is useful to formulate a cost function that penalizes deviations from the measured data and deviations from the predicted parameters:

$$C_{seed} = C_{data} + C_{pred}. \tag{4.254}$$

The cost function for deviating from the measured data is given by

$$C_{data} = 0.5 \sum_i w_i (V_n(i) - \bar{J}_i^T \, \bar{\theta})^2. \tag{4.255}$$

The matrix form of (4.255) is

$$C_{data} = 0.5 \, \bar{\theta}^T Q \, \bar{\theta} - \bar{p}^T \bar{\theta} + c_v^2, \tag{4.256}$$

where c_v is a constant term. The cost function for deviating from the predicted values is

$$C_{pred} = 0.5\ (\bar{\theta} - \bar{\theta}_{pred})^T W_{pred} (\bar{\theta} - \bar{\theta}_{pred}), \tag{4.257}$$

where $\bar{\theta}_{pred}$ is the predicted sensor motion and W_{pred} is a weighting matrix. The inverse of the weighting matrix is the error covariance matrix for the predicted sensor motion. The least square estimate of $\bar{\theta}$ is obtained by solving

$$\bar{p} + W_{pred}\bar{\theta}_{pred} = [Q_{int} + W_{pred}]\ \bar{\theta}. \tag{4.258}$$

The next stage of seeding attempts to order the data such that the processing of uncertain data is delayed. Candidate seed features are selected and tested for global consistency. Inconsistent features are culled from the seed set. This stage relies on four assumptions: an estimate of the sensor translation is available, the depth has been measured, the rotation about the z-axis is small, and that most of the features in the scene belong to stationary objects. The effect of errors in the above assumptions are discussed at the end of this subsection.

Initial seeding candidates are obtained from the epipolar channel. Stereo image features that have similar $\dot{z}$ are grouped together. The velocity $\dot{z}$ is estimated locally using the difference in image velocity at corresponding points in the stereo images (see section 2.6). Any stereo image feature whose $\dot{z}$ is significantly different than $-T_{z,sen}$ is excluded from the seed set. The estimate of $T_{z,sen}$ is obtained from the extended sensor translation or from an auxiliary sensor such as a speedometer.

The features fulfilling the $\dot{z}$ requirement are tested for in-plane motion consistency, as shown in figure 4.17. The normal image velocity is measured at each seed feature. A companion set of predicted normal image velocities is formed using the extended sensor translation. The difference between the measured and predicted normal image velocity is due to sensor rotation. Since the features are obtained from the epipolar channel,

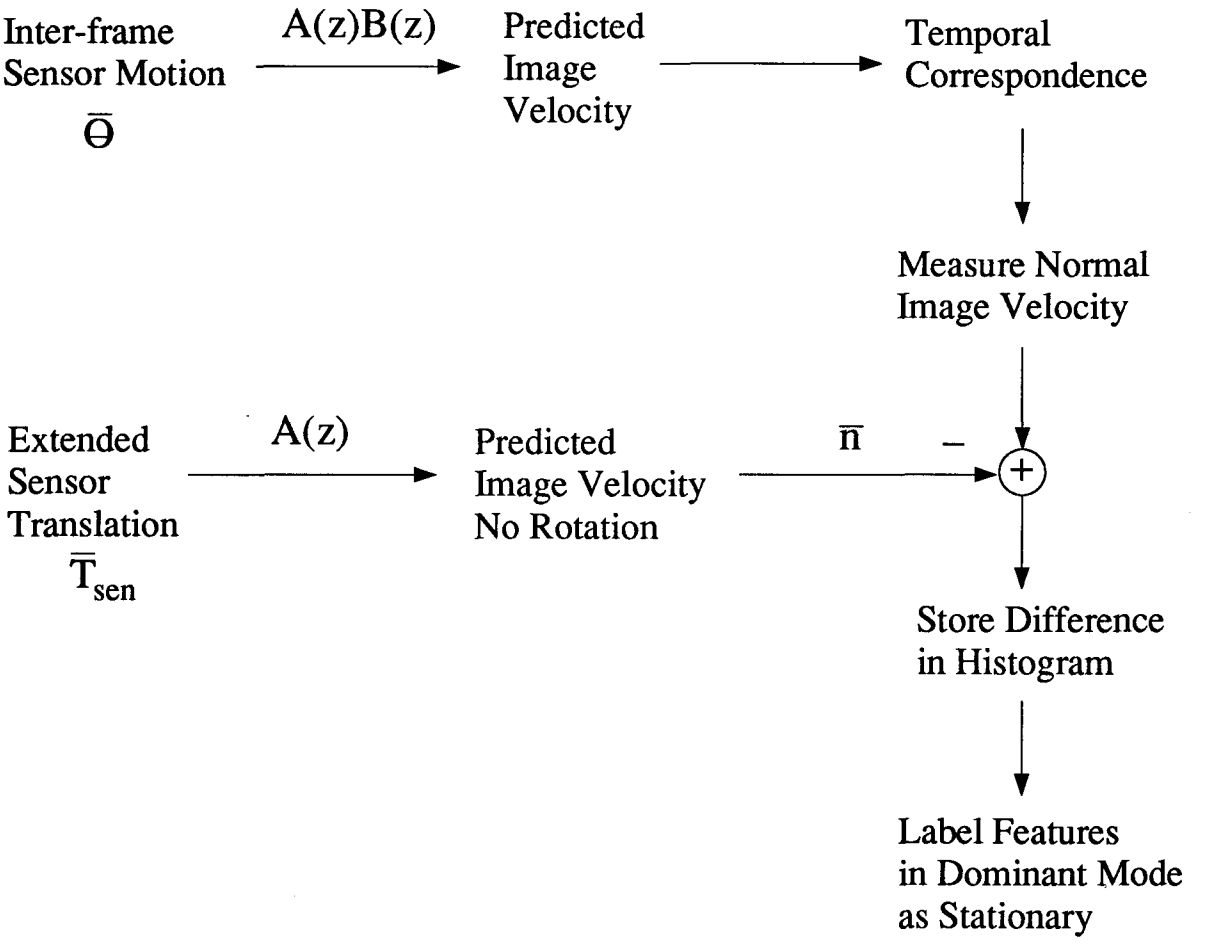


Figure 4.17: Identifying Stationary Object Features Using In-plane Motion Consistency

the difference is measured in one direction: along the $\hat{x}$-axis. If the rotation about the z-axis is zero, the difference between the measured and predicted set of normal image velocities should be a nearly constant offset caused by Ω_y (see section 2.4.2). The normal image velocity differences are stored in a histogram. If the first assumption (most of the features in the scene belong to stationary objects) is valid, the features associated with the histogram's dominant mode belong to stationary objects.

The previous paragraph describes how stationary object features from the epipolar

channel are identified. A similar histogram technique is used to identify stationary object features from the orthogonal channel. The features associated with the dominant mode are assumed to belong to stationary objects. The $\hat{x}$ and $\hat{y}$ image velocity offsets, estimated from the epipolar and orthogonal channels, respectively, are used to identify stationary object features within the remaining (oblique) channels.

The final stage of the seed process incorporates known stationary object features available from past segmentations. Most of these features will have been detected by the previous stage.

In summary, the estimation of the inter-frame sensor motion has four distinct steps. In the first step, information from auxiliary sensors is used to initialize the inter-frame Hessian matrix and the measurement vector. In the next step, a conservatively chosen set of seed measurements are used to update the Hessian matrix and the measurement vector. The Mahalanobis distance test is inhibited. In the third step, features previously identified as belonging to stationary objects are incorporated, subject to passing the Mahalanobis distance test. Finally, the remaining features passing the Mahalanobis distance test are incorporated into the inter-frame Hessian matrix and measurement vector. The four steps introduce a serial processing requirement to the estimation of inter-frame sensor motion. Parallel processing is possible within each step.

It is important to note that the seeding process (the first three steps) does not exclude features; culled features are retested during the final step. At the final step, the Mahalanobis distance is used to reject features that do not belong to stationary objects. The exceptional case is stereo features that are identified as belonging to moving objects. Such features are immediately removed from the seed set. The processing of these "moving object features" is discussed in in subsection 4.6.4.

Since the seed process does not exclude features, errors in seed assumptions are usually not serious. The sensitivity of the seeding histogram to inaccuracies in T_x and T_y is related

to the variation in depth within the scene. If the variation in depth is small, the errors will simply change the image velocity offset, but the same features will remain in the dominant mode. Large variations in depth coupled with errors in T_x and T_y tend to smear the histogram. In such a case, the dominant mode in the histogram will contain only features from the depth interval with the most features.

Inaccuracies in T_z and a non-zero Ω_z tend to degrade the histogram process. When the estimate of T_z is inaccurate, the dominant mode in the histogram will comprise measurements from a local part of the image containing a high density of features with small variations in depth. Sensor rotation about the z-axis produces a constant offset along each column of the Gabor lattice. The offset will be different in each column. The histogram process will pick features localized about a given column of the sampling lattice when Ω_z is not zero.

The critical assumption is that most of the features in the image belong to stationary objects. If most of the objects are moving at a common (non-zero) velocity, the histogram process will select the wrong mode. This is less likely to occur for stereo features (from the epipolar channel) because the local estimate of $\dot{z}$ is used is used to cull moving object features.

4.6.3 Exploiting Planar Motion

The poor distribution of features in the scene can produce image velocity fields that are not unique to a given sensor motion. This subsection discusses how non-uniqueness is identified. It also discusses how physical constraints such as planar motion can be exploited to improve the conditioning of the Hessian matrix. Two cases of planar motion are examined: the unknown plane and the known plane.

Strictly speaking, the set of normal image velocities will produce a unique least square estimate of sensor motion if Q_{int}^{-1} exists. However, if the conditioning of the Hessian matrix

is poor, the least square solution will not be stable to small changes in the measurement vector. Since measurement errors exists, an ill-conditioned solution may exhibit a large change in parameter estimates when a new measurement is incorporated.

The condition of the Hessian matrix is measured by the ratio of the largest and smallest eigenvalues; a large eigenvalue ratio indicates poor conditioning. Associated with each eigenvalue is an eigenvector. The eigenvectors are dependent on the distribution of features in the image and scene. Consider the case where the features are evenly distributed throughout the image. Four of the six eigenvectors will be approximately equal to [7]

$$\bar{v}_0 \approx [\frac{z_f}{z_{ave}} \; 0 \; 0 \; 0 \; -z_f \; 0], \tag{4.259}$$

$$\bar{v}_5 \approx [z_f \; 0 \; 0 \; 0 \; \frac{z_f}{z_{ave}} \; 0], \tag{4.260}$$

$$\bar{v}_1 \approx [0 \; \frac{z_f}{z_{ave}} \; 0 \; z_f \; 0 \; 0], \tag{4.261}$$

and

$$\bar{v}_4 \approx [0 \; z_f \; 0 \; -\frac{z_f}{z_{ave}} \; 0 \; 0], \tag{4.262}$$

where z_{ave} is the average depth of the scene [8]. The eigenvectors v_0 and v_1 represents the in-plane motion; the average shift in the image velocity field along the $\hat{x}$- and $\hat{y}$-axes, respectively. The eigenvectors v_5 and v_4 represents the variations in $V_{\hat{x}}$ and $V_{\hat{y}}$ due to variations in depth. In most cases, v_0 and v_1 will produce the largest eigenvalues; v_4 and v_5 will produce the smallest eigenvalues. When the eigenvalue associated with v_5 (v_4) is small, it is very difficult to distinguish between T_x and Ω_y (T_y and Ω_x) from the image velocity field.

[7] The following eigenvectors can be considered as an ideal case that is sufficiently accurate for illustrative purposes.

[8] This "average" depends on the weighted average of the features being processed as stationary objects. It is not explicitly calculated.

Some of the eigenvalues can be increased by enforcing physical constraints. In many cases, an autonomous vehicle is travelling on a planar surface, such as a floor. The planar surface constrains the sensor motion: sensor translation must be along the surface, and axis of sensor rotation must be normal to the surface. Thus, the translation vector is orthogonal to the rotation vector. In vector form, the constraint is given by

$$\bar{T}^T \bar{\Omega} = 0. \tag{4.263}$$

This constraint does not assume that the surface normal is known.

To illustrate how the planar motion constraint can be exploited, it is useful to define a cost function that penalizes deviations from the measured data and deviations from planar motion. The proposed cost function is given by

$$C_p = C_{data} + 0.5\, w_p e_p^2, \tag{4.264}$$

where w_p is a scalar weighting term and

$$e_p = (\bar{T}^T \bar{\Omega}). \tag{4.265}$$

The cost function can be written as

$$C_p = 0.5 \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix}^T \begin{bmatrix} Q_a & Q_b \\ Q_b^T & Q_c \end{bmatrix} \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix} - [\bar{p}_a\ \bar{p}_b] \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix} + c_v^2 + 0.5\, w_p (\bar{T}^T \bar{\Omega})^T (\bar{T}^T \bar{\Omega}). \tag{4.266}$$

The gradient with respect to inter-frame sensor translation and rotation are respectively given by

$$\nabla C_p(\bar{T}) = Q_a \bar{T} + Q_b \bar{\Omega} - \bar{p}_a + w_p e_p \bar{\Omega}, \tag{4.267}$$

and

$$\nabla C_p(\bar{\Omega}) = Q_b^T \bar{T} + Q_c \bar{\Omega} - \bar{p}_b + w_p e_p \bar{T}. \tag{4.268}$$

Setting both gradients to zero, we get

$$\begin{bmatrix} \bar{p}_a \\ \bar{p}_b \end{bmatrix} = \begin{bmatrix} Q_a & Q_b \\ Q_b^T & Q_c \end{bmatrix} \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix} + w_p e_p \begin{bmatrix} 0_{3x3} & I \\ I & 0_{3x3} \end{bmatrix} \begin{bmatrix} \bar{T} \\ \bar{\Omega} \end{bmatrix}, \tag{4.269}$$

where 0_{3x3} is a 3 by 3 matrix of zeroes and I is a 3 by 3 identity matrix. Equation (4.269) can be rewritten as

$$\bar{p} = \{Q_{int} + w_p e_p \begin{bmatrix} 0_{3x3} & I \\ I & 0_{3x3} \end{bmatrix}\} \bar{\theta}. \tag{4.270}$$

Assume that the surface normal, denoted by $\bar{n}_p = [n_x \ n_y \ n_z]^T$, is known. A cost function that penalizes deviation from this known planar motion can be made:

$$C_n = C_{data} + 0.5 \ w_T e_T^2 + w_\Omega e_\Omega^2, \tag{4.271}$$

where w_T and w_Ω are scalar weighting terms, $e_T = \bar{T} \cdot \bar{n}_p$ and $e_\Omega = \bar{\Omega} \times \bar{n}_p$. The least square solution is

$$\bar{\theta} = [Q_{int} + W_{plane}]^{-1} \bar{p}, \tag{4.272}$$

where

$$W_{plane} = \begin{bmatrix} W_a & 0_{3x3} \\ 0_{3x3} & W_c \end{bmatrix}, \tag{4.273}$$

$$W_a = w_T \begin{bmatrix} n_x^2 & n_x n_y & n_x n_z \\ n_x n_y & n_y^2 & n_y n_z \\ n_x n_z & n_y n_z & n_z^2 \end{bmatrix}, \tag{4.274}$$

and

$$W_c = w_\Omega \begin{bmatrix} n_y^2 + n_z^2 & -n_x n_y & -n_x n_z \\ -n_x n_y & n_x^2 + n_z^2 & -n_y n_z \\ -n_x n_z & -n_y n_z & n_x^2 + n_y^2 \end{bmatrix}. \tag{4.275}$$

It is useful to compare the two types of planar constraints. The known plane weighting matrix W_{plane} has a rank of three. It is applied directly to the Hessian matrix and should

be incorporated into the seed phase with the auxiliary sensor motion estimates. The unknown plane constraint has a rank of one. The constraint contains the error term e_p; thus, the unknown plane a non-linear problem that is dependent on the current estimate of sensor motion. As a result, an iterative process is required to find the solution. The unknown plane constraint is applied after all the data has been incorporated into the inter-frame sensor motion estimate. It has little value as a seeding tool.

The planar constraint reduces, but does not eliminate, the problem of poor conditioning. The unknown plane constraint only has a rank of one; there are typically two small eigenvalues in the Hessian matrix. In addition, there is no guarantee that the unknown plane constraint will affect either of the smallest eigenvalues. For the case of an observer travelling forward along a plane whose surface normal is given by $\bar{n}_p = [0\ 1\ 0]^T$, the unknown plane constraint will primarily affect the eigenvalues associated with Ω_z. The known plane constraint suffers from a similar limitation, but to a lesser extent. For the previously mentioned case, the known plane would constrain eigenvalues associated with T_y, Ω_x, and Ω_z. The eigenvalue associated with $\bar{v}_4$ would increase significantly. The eigenvalue associated with $\bar{v}_5$ would be unaltered. Thus, for the typical operation of a ground-based autonomous vehicle, the planar constraints provide only marginal improvements in the overall conditioning of the Hessian matrix. The planar constraints will improve certain individual motion parameter estimates.

4.6.4 Moving Objects

This subsection describes how features belonging to moving objects are processed. The grouping of features is discussed.

The stereo features that belong to moving objects are identified at two stages in the processing: early in the processing during the seed stage; and later, after the inter-frame sensor motion stage is complete. The selected stereo features are tested for $\dot{z}$

consistency; that is, to check if both the image velocity-based and depth-based estimates of $\dot{z}$ are similar (see section 2.6). A by-product of the $\dot{z}$ consistency test is a local estimate of the time-to-collision. A feature tracking process is activated if the feature has a short time-to-collision. The time-to-collision restriction is used to reduce the number of active (tracked) features. The segmentation of the image sequence is discussed in section 4.6.5.

Once the active features are identified, it is necessary to group features that belong to a common moving object into a common object class. The features within a given object class have the same translational velocity and approximately the same depth. For the purpose of testing the similarity of two object classes, a second Mahalanobis distance is defined:

$$\begin{aligned} d^2_{mah2} &= [\bar{T}_{obj(1)} - \bar{T}_{obj(2)}]^T [Q^{-1}_{obj(1)} + Q^{-1}_{obj(2)}]^{-1} [\bar{T}_{obj(1)} - \bar{T}_{obj(2)}] \\ &\quad + \frac{[d_{\dot{x}}(1) - d_{\dot{x}}(2)]^2}{E[(\Delta d_{\dot{x}}(1))^2] + E[(\Delta d_{\dot{x}}(2))^2]}. \end{aligned} \tag{4.276}$$

This Mahalanobis distance measures the difference in object velocity and in the disparity. The disparity term is particularly useful when an object class has a small number of features. A small class can have a large error covariance Q^{-1}_{obj}; thus, the expected motion error is too large to reject other object classes. If the maximum size of the viewed object is known, the difference in the average x and y position of two object classes can be included in the Mahalanobis distance.

In this work, the grouping process is sequential. Each stereo pair of features is assigned its own object class. The Mahalanobis distance is calculated for each pair of object classes. The two object classes that produce the smallest Mahalanobis distance (that does not exceed a given threshold) are merged. This procedure is repeated until all remaining pairs of object classes exceed the threshold.

The above-mentioned method groups stereo features together. Since all the stereo features are obtained from epipolar channels, the Hessian matrix Q_{obj} will be ill-conditioned.

Features from oblique channels are required. Candidate oblique features are selected from the set of features as belonging to non-stationary objects [9]. The disparity similarity is enforced to reduce the number of candidates for a given object class. If more than one candidate remains, flow-field divergence [46] is used to estimate the time-to-collision and subsequently, the velocity $\dot{z}_{obj}$. Once this initial culling process is complete, the remaining candidates are tested using the second Mahalanobis distance. Acceptable candidates are merged into the appropriate object class.

In some cases, there are no oblique candidate features. This can be a result of lack of oblique features, severe culling of features, or incorrectly identifying features as belonging to stationary objects. To allow the calculation of the error covariance Q_{obj}^{-1}, the y component of object motion, $\dot{y}_{obj}$, must be assumed. If the object and sensor are travelling on the same plane, the combined object/sensor motion has the same $\hat{y}_{foe}$ as the sensor motion. An alternative (worst case) approach is to assume a $\hat{y}_{foe}$ that will cause the object to collide with the sensor or pass-by at the height of the sensor. The object's "padded" Hessian matrix, that incorporates the assumed $\hat{y}_{foe}$, is given by

$$Q_{pad} = Q_{obj} + \frac{w_{pad}}{z_f^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & z_f^2 & (z_f\ \hat{y}_{foe}) \\ 0 & (z_f\ \hat{y}_{foe}) & \hat{y}_{foe}^2 \end{bmatrix}, \tag{4.277}$$

where Q_{pad} is the padded Hessian matrix and w_{pad} is the padding confidence term. The padding confidence is the inverse of the expected squared error in $\dot{y}_{obj}$ resulting from the assumption of $\hat{y}_{foe}$.

Not all features belonging to moving objects will be incorporated into an object class. It is desirable to identify these unassigned monocular image features for two reasons: to prevent the image features from accidentally being identified as belonging to

[9] Oblique features belonging to non-stationary objects are available after the inter-frame sensor motion stage has been completed.

stationary objects; and to identify regions of uncertainty in the scene. Certain moving objects can be identified, without depth information, by the normal image velocity. If the measured normal image velocity is incompatible with the positive depth/stationary object assumption, the viewed object must be moving. Consider, as an example, a forward translating camera with no rotation. Any feature whose normal image velocity points towards the sensor's focus of expansion must belong to a moving object. This directional test will identify moving objects that will pass in front of the sensor. This method for identifying moving objects is also valid for a rotating sensor. The effect of sensor rotation must be subtracted from the normal image velocity before applying the directional test. In summary, an image feature must belong to a moving object if [29]

$$(V_n - \bar{n}^T A R_{os} B_\Omega \, \bar{\Omega}) \, \bar{n}^T \bar{r}_{foe} < 0, \tag{4.278}$$

where $\bar{r}_{foe} = [(\hat{x} - \hat{x}_{foe}) \, (\hat{y} - \hat{y}_{foe})]^T$,

$$\hat{x}_{foe} = z_f \frac{T_x}{T_z}, \tag{4.279}$$

and

$$\hat{y}_{foe} = z_f \frac{T_y}{T_z}. \tag{4.280}$$

Equation (4.278) will only detect a moving object (that is passing in front of the camera) at certain parts of the object's trajectory. Consider a forward moving camera and the right moving object. If the projection of the moving object is at the left periphery of the image, the normal image velocity induced by the object motion opposes the normal image velocity induced by the forward translation of the sensor. Equation (4.278) is activated at the point in the object's trajectory where the object-induced normal image velocity exceeds the normal image velocity induced by the sensor translation. This will occur at some point because the normal image velocity induced by the sensor translation decreases to zero as the image feature approaches the sensor's focus of expansion. Once

the image feature crosses the sensor's focus of expansion, the two normal image velocities will have the same directions. At this point in the trajectory, (4.278) is no longer activated.

If a standoff depth is assigned, more types of moving objects can be detected. Very fast image features will be flagged as significant. These features belong either to moving objects at an arbitrary positive depth or to stationary objects within the standoff depth. Either case is consider interesting by the obstacle avoidance module. The fast moving/standoff depth test is given by

$$(V_n - \bar{n}^T A R_{os} B_\Omega \bar{\Omega})\, \bar{n}^T \bar{r}_{foe} > (\bar{n}^T \bar{r}_{foe})^2 \frac{T_z}{z_{stnd}}, \tag{4.281}$$

where z_{stnd} is the standoff depth. Note that the right side of the inequality is positive for forward translating sensors.

4.6.5 Interaction Between Modules

This section describes the interaction of the various modules in the obstacle detection algorithm. Particular attention is given to the segmentation of the image sequence into stationary objects and moving objects.

The modules of the obstacle detection algorithm are shown in figure 4.18. Significant features are selected from the Gabor-filtered images. The disparity and the normal image velocity, along with the expected errors, are measured. The stereo normal image velocity and the disparity produce a local estimate of $\dot{z}$.

The velocity $\dot{z}$ is used to segment the image sequence. If an estimate of the extended sensor translation is available, the object velocity $\dot{z}_{obj}$ can be estimated:

$$\dot{z}_{obj} = \dot{z} + T_{z,sen}. \tag{4.282}$$

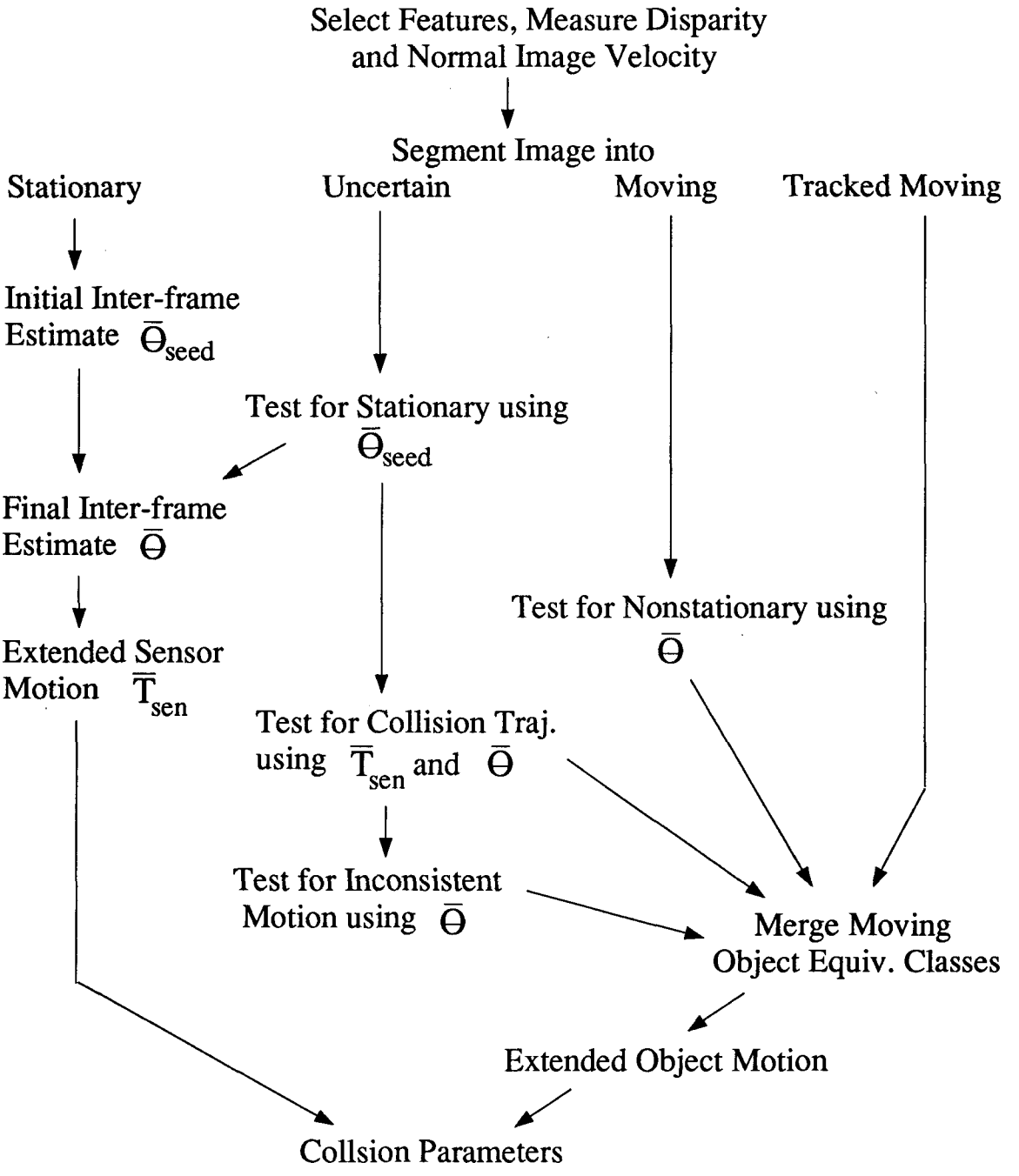


Figure 4.18: Modules of the Obstacle Detection Algorithm

A necessary condition for a stationary object is

$$|\dot{z}_{obj}| \approx 0. \tag{4.283}$$

A sufficient condition for a moving object is

$$|\dot{z}_{obj}| >> 0. \tag{4.284}$$

Two thresholds are applied to $\dot{z}_{obj}$, initially segmenting the stereo image features into stationary objects, moving objects, or uncertain. The stationary object threshold, $\dot{z}_{st}$, is non-zero (but small) to allow for errors in the local estimate of $\dot{z}$ and the sensor motion estimate $T_{z,sen}$. The moving object threshold, $\dot{z}_{mt}$, is chosen significantly larger than zero to avoid false detections. Stereo features whose $\dot{z}_{obj}$ is between the thresholds are labelled as uncertain.

The potential stationary object features are tested for in-plane motion consistency using the seeding histograms. Inconsistent measurements are labelled uncertain. The potential moving objects are tested for urgency. Any object with a large time-to-collision is labelled as uncertain because these measurements are sensitive to errors. Moving object that are being tracked from previous images, referred to as "tracked moving" in figure 4.18, by-pass both the $\dot{z}_{mt}$ and t_{col} tests.

The features identified by the seed stage as belonging to stationary objects are combined to obtain an initial estimate of the inter-frame sensor motion, denoted by $\bar{\theta}_{seed}$ in figure 4.18. The inter-frame Hessian associated with $\bar{\theta}_{seed}$ is usd in the Mahalanobis distance to test if uncertain features are consistent with the stationary object hypothesis. The image measurements of consistent features are used to update the inter-frame sensor motion estimate. Any inconsistent features remain labelled as uncertain.

After all the stationary object features have been identified and combined, the final inter-frame sensor motion estimate (denoted by $\bar{\theta}$ in figure 4.18) is integrated into the

extended sensor translation estimate (denoted by $\bar{T}_{sen}$). The Hessian matrix associated with $\bar{\theta}$ is used to test the stereo features belonging to moving objects. If both of the stereo features are consistent with the stationary object hypothesis, the features are removed from the moving object list.

Uncertain stereo features are checked for collision trajectories. The correspondence predictor, that is tuned to collision trajectories, uses both $\bar{T}_{sen}$ and $\bar{\theta}$. This step identifies moving obstacles whose velocity is orthogonal to the z-axis ($\dot{z}_{obj}$ is small). Next, the uncertain features are tested for inconsistent motion. Any normal image velocity measurement with an incompatible direction (that satisfies equation (4.278)) is identified. If matching stereo features exist, the stereo features are labelled as a moving object. This step identifies moving objects that will pass in front of the cameras.

The moving object features, both the new and tracked, are merged into equivalence classes. The extended object motion for each class is estimated. In the current implementation, only stereo features from the epipolar channel are used to estimate object motion. As a result, the $\hat{y}_{foe}$ of the object is assumed to be zero. The extended object and sensor motions are combined to estimate the observer frame trajectory. The collision parameters for each feature are predicted using the image coordinates, the depth, and the observer frame trajectory. Note that each feature belonging to a given object has a different point-of-collision.

4.6.6 Extensions to the Kalman Filter

This subsection investigates extensions to the Kalman filter equations presented in section 4.4. Uncertainty in the inter-frame rotation is modelled as process noise (motion disturbances) in the extended sensor and object translation equations. It is also shown how pilot commands, such as the steering angle and braking, can be incorporated into the process model to predict future states.

The process model of the extended sensor translation is given by

$$\bar{T}_{sen}(t_{i+1}/t_i) = R^T(t_i)\bar{T}_{sen}(t_i). \tag{4.285}$$

The error, the process noise, is given by

$$\delta\bar{T}_{sen}(t_{i+1}/t_i) = R^T(t_i)\ \delta\bar{T}_{sen}(t_i) - S\ \delta\bar{\Omega}, \tag{4.286}$$

where

$$S = \begin{bmatrix} 0 & -T_{z,sen} & T_{y,sen} \\ T_{z,sen} & 0 & -T_{x,sen} \\ -T_{y,sen} & T_{x,sen} & 0 \end{bmatrix} \Delta t. \tag{4.287}$$

In matrix form, the error can be written as

$$\delta\bar{T}_{sen}(t_{i+1}/t_i) = H_N^T \begin{bmatrix} \delta\bar{T}_{sen}(t_i) \\ \delta\bar{\Omega} \end{bmatrix}, \tag{4.288}$$

where $H_N^T = [R^T \ -S]$.

The error covariance matrix of this noise corrupted process is denoted by

$$P_{RT}(t_i) = \begin{bmatrix} Q_{sen}^{-1}(t_i) & E[\delta\bar{T}_{sen}\ \delta\bar{\Omega}^T] \\ E[\delta\bar{\Omega}\ \delta\bar{T}_{sen}] & E[\delta\bar{\Omega}\ \delta\bar{\Omega}^T] \end{bmatrix}. \tag{4.289}$$

The error covariance for the inter-frame sensor motion can be written in a similar form:

$$Q_{int}^{-1} = \begin{bmatrix} Q_T^{-1} & P_{T\Omega} \\ P_{\Omega T} & P_{\Omega\Omega} \end{bmatrix}. \tag{4.290}$$

The block elements of P_{RT} and Q_{int}^{-1} are related:

$$E[\delta\bar{T}_{sen}\ \delta\bar{\Omega}^T] = \eta P_{T\Omega}, \tag{4.291}$$

$$E[\delta\bar{\Omega}\ \delta\bar{T}_{sen}^T] = \eta P_{\Omega T}, \tag{4.292}$$

$$E[\delta\bar{\Omega}\ \delta\bar{\Omega}^T] = P_{\Omega\Omega}, \tag{4.293}$$

where

$$\eta = Q_{sen}^{-1} Q_T. \tag{4.294}$$

The matrix η represents the influence of the current inter-frame sensor motion estimate on the extended sensor translation. It will tend to decrease as more information is integrated.

The error covariance of the extended sensor translation at time t_{i+1} can be predicted from P_{RT} at time t_i:

$$Q_{sen}^{-1}(t_{i+1}/t_i) = H_N^T \, P_{RT}(t_i) H_N = [R^T \; -S] P_{RT}(t_i) \begin{bmatrix} R \\ -S^T \end{bmatrix}. \tag{4.295}$$

This equation can be rewritten as

$$Q_{sen}^{-1}(t_{i+1}/t_i) = R^T Q_{sen}^{-1}(t_i) R + N_{sen}(t_i), \tag{4.296}$$

where

$$N_{sen}(t_i) = -\eta [R^T P_{T\Omega} S^T + S P_{\Omega T} R] + S P_{\Omega\Omega} S^T. \tag{4.297}$$

The matrix N_{sen} is the error covariance of the motion disturbance described in section 4.4.

A similar model of process noise exists for the extended object translation. The process model for the object translation is given by

$$\bar{T}_{obj}(t_{i+1}/t_i) = R^T(t_i) \bar{T}_{obj}(t_i). \tag{4.298}$$

Since $\bar{T}_{obj}$ is based on the excess normal image velocity, the object translation and the inter-frame sensor rotation are uncorrelated; that is,

$$E[\bar{T}_{obj} \bar{\Omega}^T] = (E[\bar{\Omega} \; \bar{T}_{obj}^T])^T = 0_{3x3}. \tag{4.299}$$

Thus, the error covariance of the extended object translation is given by

$$Q_{obj}^{-1}(t_{i+1}/t_i) = R^T Q_{obj}^{-1}(t_i) R + S_{obj} P_{\Omega\Omega} S_{obj}^T, \tag{4.300}$$

where

$$S_{obj} = \begin{bmatrix} 0 & -T_{z,obj} & T_{y,obj} \\ T_{z,obj} & 0 & -T_{x,obj} \\ -T_{y,obj} & T_{x,obj} & 0 \end{bmatrix} \Delta t. \tag{4.301}$$

The rotational uncertainty increases the predicted error covariance $Q_{obj}^{-1}(t_{i+1}/t_i)$. As a result, the relative importance of future inter-frame object translations will increase. This result is also true for the extended sensor translation case.

It is possible to incorporate pilot commands into the process model for the extended sensor motion. In its current form, the model of vehicle motion is pure translation [10]. Certain forces applied to the vehicle, such as steering, propulsion, and braking, cause the vehicle to accelerate. The acceleration, by definition, will cause the velocity of the vehicle to change. A vehicle translation model, that includes these accelerations, is given by

$$\bar{T}_{veh}(t_{i+1}/t_i) = R_{veh}^T[\bar{T}_{veh}(t_i) + A_s], \tag{4.302}$$

where $\bar{T}_{veh}$ is the translation of the vehicle (relative to a world coordinate frame), R_{veh} is a rotation matrix representing the change in vehicle heading, and A_s is the change in vehicle speed. Note that $\bar{T}_{veh}(t_{i+1}/t_i)$ in (4.302) is represented using the same coordinate frame as $\bar{T}_{veh}(t_i)$. The vehicle translation at time t_i is (assuming no sideways slip)

$$\bar{T}_{veh} = \begin{bmatrix} 0 \\ 0 \\ s \end{bmatrix}, \tag{4.303}$$

where s is the speed of the vehicle.

Vehicle acceleration has two forms: changes in vehicle heading, and changes in vehicle speed. For a standard automobile, the steering angle is used to alter the vehicle heading.

[10]The matrix $R(t_i)$ changes the coordinate frame representing the sensor translation from the observer frame at time t_i to the observer frame at time t_{i+1}. The change in the orientation of the observer frame is assumed to be due to transients, such as camera shake, not due to changes in the vehicle heading.

The steering angle, γ_{str}, causes the vehicle to travel in a circle whose radius, r_{str}, is given by

$$r_{str} = \frac{L_{wb}}{\tan \gamma_{str}}, \tag{4.304}$$

where L_{wb} is the wheel base of the vehicle. The change in heading, $\Delta\theta_h$, is dependent on the steering angle and the distance the vehicle travels during Δt:

$$\Delta\theta_h = \frac{\tan \gamma_{str}}{L_{wb}} \int_{t_i}^{t_{i+1}} (s + at)\, dt, \tag{4.305}$$

where a is the linear acceleration. The vehicle speed is given by

$$s(t_{i+1}) = k_a s(t_i), \tag{4.306}$$

where

$$k_a = 1 + \frac{\int a\, dt}{s(t_i)}. \tag{4.307}$$

The new model of vehicle motion is

$$\bar{T}_{veh}(t_{i+1}/t_i) = k_a R_{veh} \bar{T}_{veh}(t_i), \tag{4.308}$$

where

$$R_{veh} = \begin{bmatrix} \cos\Delta\theta_h & 0 & \sin\Delta\theta_h \\ 0 & 1 & 0 \\ -\sin\Delta\theta_h & 0 & \cos\Delta\theta_h \end{bmatrix}. \tag{4.309}$$

The above equations describe the vehicle translation. The extended sensor translation is given by

$$\bar{T}_{sen}(t_i) = R_{vo}^T(t_i) \bar{T}_{veh}(t_i), \tag{4.310}$$

where R_{vo} is the difference in orientation between the observer and vehicle coordinate frames. When the roll angle between the coordinate frames is zero, and the pan and tilt

angles are small,

$$R_{vo}(t_i) = \frac{1}{z_f}\begin{bmatrix} z_f & 0 & -\hat{x}_{foe} \\ 0 & z_f & -\hat{y}_{foe} \\ \hat{x}_{foe} & \hat{y}_{foe} & z_f \end{bmatrix}. \tag{4.311}$$

The new model of the extended sensor translation is

$$\bar{T}_{sen}(t_{i+1}/t_i) = k_a R^T(t_i) R_{vo}^T R_{veh} R_{vo} \bar{T}_{sen}(t_i). \tag{4.312}$$

In this work, pilot commands are not incorporated into the process model for the extended sensor translation. Such a model would be useful for obstacle avoidance. Simple evasive maneuvers, described by the steering angle and braking/propulsion, could be tested to predict how long a given path can be followed before encountering an obstacle. Maneuvers that decrease the time-to-collision should be avoided.

This subsection has proposed extensions to the extended sensor and object Kalman filters. The uncertainty associated with the inter-frame rotation is incorporated into the Kalman filters as process noise. This rotation-induced process noise is included in the current implementation. The current implementation does not incorporate pilot commands. The ease in which an obstacle avoidance module can be added should be apparent.

4.7 Comparison

The implementation described in this chapter is a cascade of a number of different algorithms. This section provides a comparison of the submodules of this implementation with the works of other researchers.

The inter-frame sensor motion is similar to the direct passive navigation problem first studied by Negahdaripour and Horn [45], later by Ito and Aloimonos [33], and Horn and Weldon [31]. It is most similar to the known structure case found in both Ito and

Aloimonos, and Horn and Weldon. This implementation differs from the previously mentioned works in a number of ways. First, phase derivatives from bandpass (Gabor) filtered images are used in place of image intensity derivatives. The advantage of using phase is that, unlike image intensity, phase is stable with respect to changes in contrast and to geometrical deformations such as dilation and rotation [21]. The second difference is that error estimates are used to weight the normal image velocity measurements. Thus, a weighted least squared solution is obtained. Finally, this work is designed for a dynamic environment, the other researchers assume that the scene is stationary.

The depth of each feature must be measured to estimate the inter-frame sensor motion. The depth is measured using a set of epipolar Gabor filters. The depth module is a combination of the works of Sanger [47], Jenkin and Jepson [34], and Fleet, Jepson and Jenkin [21]. The interesting features obtained from their works include a multiscale approach for estimating disparity, and a matching criteria for reducing the likehood of false disparity estimates. I have extended their work by estimating the measurement error. This work also uses temporal consistency and a heuristic ordering constraint to increase the number of detected stereo features.

Gabor filters are also used to estimate the normal image velocity. This module follows the work of Fleet and Jepson [22] [23]. There are some differences, however. This implementation uses the inter-frame sensor and object motion to predict the normal image velocity. This extends the maximum inter-frame displacement. In addition, this implementation estimates the measurement error.

The Mahalanobis distance is used to identify features belonging to stationary objects. Heeger and Hager [29] applied the Mahalanobis distance to image velocity field; this implementation uses the normal image velocity. The Mahalanobis distance compares the estimated velocity error with the expected error. Both errors are available from the inter-frame sensor motion module. Thus, motion segmentation using the Mahalanobis

distance is a natural extension of the weighted least square approach to direct passive navigation when the structure is known.

Kalman filtering is used to estimate the object and sensor motion over the entire image sequence. The uniqueness of a given implementation of the Kalman filter is in the modelling of the motion and the error. In this implementation, a pure translation model is used. Dropping the inter-frame rotation parameters has the effect of stabilizing the image sequence. The inter-frame rotation terms are comparable to the extra terms used in an augmented Kalman filter [11].

In summary, the key difference between this implementation and the many works it draws from is error estimation. The error estimated from the Gabor filters is propagated through each stage of processing until it reaches the collision parameters. The error estimates are particularly important for motion segmentation. Error estimates are also used to fuse measurements of vehicle motion from auxiliary sensors. The error covariance matrices used in the inter-frame sensor motion estimation and the Kalman filter can be used to incorporate physical constraints, such as planar motion, into the motion estimates.

The rest of this section compares this implementation as a whole with that of Ayache and Faugeras [6]. Ayache and Faugeras use three-dimensional positional information to perform stereo (trinocular) camera-based navigation. The three-dimensional position of physical features (lines in a three-dimensional space) are tracked and integrated over time using a Kalman filter. The uncertainty in the three-dimensional position is stored in an error covariance matrix. As time elapses, and more information is accumulated, the error covariance for tracked three-dimensional feature decreases. The changes in the three-dimensional position of features over time are used to estimate the inter-frame sensor motion as well as the inter-frame error covariance matrix. The inter-frame motion is used to aid temporal correspondences by predicting the the future position of a physical

feature. The Mahalanobis distance is used to reject outliers (candidate correspondences that are inconsistent with the inter-frame motion). The Mahalanobis distance is also used to test groups of features for coherent properties (such a belonging to a common plane). Coherent features are merged.

There are many similarities between the implementation used by Ayache and Faugeras and the implementation presented in this work. Both implementations use Kalman filters to integrate information, use inter-frame motion to predict correspondences, use the Mahalanobis distance to reject outliers and to test features for coherent properties, and both merge coherent features to improve accuracy. There are notable differences. Ayache and Faugeras use the three-dimensional position as the state variables in the Kalman filter; I use the three-dimensional velocity as state variables. The obvious reason for this difference is that Ayache and Faugeras are building a three-dimensional description (map) of a static environment, where as I am estimating the collision parameters of obstacles. The position of physical features relative to the cameras is important for building a map of the environment where as the trajectory of the object relative to the sensor is important for obstacle detection. Merging features with coherent positions improves the accuracy of a map; merging features with coherent motion improves the accuracy of the estimated trajectory.

There are also differences in the data representations. In the implementation used by Ayache and Faugeras, feature information is transformed from an image representation to a three-dimensional representation immediately after the disparity is measured. The implementation presented in this work delays the transformation. Both depth and motion are measured in the image coordinate frame using stereo disparity and normal image velocity, respectively. The transformation from an image representation to a three-dimensional representation occurs during the estimation of the inter-frame sensor motion.

There is an advantage in using the image coordinate frame to measure both motion

and depth. In this work, changes in position are measured directly using phase differences; the positional information is only coarsely measured. The advantage of such an approach is that changes in position can often be measured more accurately than the individual positions. Consider as an example a sine wave grating. It is difficult to localize the position of such a feature, but it is easy to determine how much it has moved. Since the estimates of the inter-frame sensor and object motions are based on changes in position (normal image velocity and disparity), the implementation is less sensitive to positional errors. A much simpler error model can be used to represent positional uncertainty (only the disparity error in maintained). Error covariance matrices are maintained for the sensor and object motion, but not for positional information.

Chapter 5

Results

In this chapter, three data sets, which comprise eight stereo image sequences, are used to test the individual modules of the algorithm as well as to demonstrate the algorithm's robustness as a system to various scene structures, various lighting conditions, and various combinations of sensor and object motions. All image sequences contain scenes are captured using CCD cameras; there are no computer synthesized images.

Data set 1 is comprised of two stereo image sequences obtained from an optical bench. In both experiments 1 and 2, the cameras are undergoing pure translation in a stationary environment. The purpose of data set 1 is to test the accuracy of various modules in the obstacle detection algorithm.

Data set 2 is also obtained from an optical bench. Three controlled experiments are presented that are designed to imitate difficult, but typical, situations encountered by an autonomous vehicle. In experiment 3, the stereo cameras are panning the scene, as if the computer pilot is turning its "head" to better view interesting features (as in [18]). Experiments 4 and 5 are designed to imitate two vehicles (one of which is the ego-vehicle) approaching an intersection. In experiment 4, the two vehicles reach the intersection at the same time; in experiment 5, the other vehicle reaches the intersection before the ego-vehicle. The purpose of data set 2 is to test the accuracy of the algorithm in the presence of sensor rotation and the algorithm's ability to segment a moving object from the background.

Data set 3 presents three experiments which are performed under less controlled lighting conditions and less precise motion than in data sets 1 and 2. Experiment 6 contains an outdoor scene with shadows. Experiment 7 has a moving sensor that experiences transient rotations. Experiment 8 contains a scene with two independently moving objects. The purpose of data set 3 is to test the accuracy of the algorithm in realistic environments.

Before proceeding with the experimental results, section 5.1 will specify the system parameters used in the eight experiments, and section 5.2 will establish standards for judging the accuracy of the results.

5.1 System Parameters

This section specifies system parameters used to process the image sequences. The Gabor channels are defined by the selection of the filter set and spatial sampling lattices. Various thresholds, used to identify important features, to test the stability of features, and to test feature correspondences, are selected.

The set of Gabor filters used in the following experiments comprises 3 frequencies, 4 orientations, and 2 phases (a total of 24 filters). The three frequencies are $(0.040\pi, 0.092\pi, 0.210\pi)$ radians per pixel or $(0.020, 0.046, 0.105)$ cycles per pixel. The four orientations are $(0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4})$. The two phases are $(-\frac{\pi}{4}, \frac{\pi}{4})$. The three constants described in section 3.2.1—the aspect ratio of the Gabor function α, the ratio of adjacent frequencies ρ, and the bandwidth-frequency ratio λ—are given by

$$\alpha = 1.00, \tag{5.313}$$

$$\rho = 2.29, \tag{5.314}$$

$$\lambda = \frac{\pi}{4}. \tag{5.315}$$

The spatial lattice is oversampled, with respect to the minimally complete definition (equations (3.113) and (3.114)), by a factor of 2.55 in both the $\acute{x}$ and $\acute{y}$ directions. This sampling density exceeds the "local frequency estimation" sampling requirement (section 3.2.3) and matches the overlapping disparity interval requirement (section 3.2.4). A bandsampled lattice is used in the epipolar channels; a restricted sampling lattice is used in the oblique channels. The width of the bandsampled lattice is limited to 35 lattice points. The 35 lattice point limit is enforced in both the $\acute{x}$ and $\acute{y}$ directions for the restricted lattices (total number of lattice points is restricted to 1225). In the experiments presented in this chapter, the 35 lattice point limit affects only the spatial lattice in the highest frequency channel.

Important features have local magnitudes that exceed three thresholds: an absolute threshold, a relative orientation threshold, and a relative spatial threshold (see section 3.2.2). The absolute threshold is set between 0.1 and 0.2 of the maximum magnitude. The relative threshold for orientation neighbours is set to 0.95. The relative threshold for spatial neighbours is set to 1.0.

The size of the relative spatial threshold may be misleading. The spatial threshold is applied to a relative significance measure that is calculated using a peak detector method. The local magnitude at a given lattice point is compared to the attenuated magnitude of other nearby lattice points. The attenuation is obtained using a Gaussian window whose spatial support is twice the support width of the Gabor filter's kernel. The relative (spatial) significance is the ratio of the local magnitude and the largest of the Gaussian windowed (neighbouring) responses. A relative spatial threshold of 1.0 using this peak detector method is similar to a threshold of 0.8 when unattenuated adjacent spatial neighbours are compared.

The relative magnitude thresholds used to test stereo and temporal correspondences (see sections 3.2.4 and 3.2.5) are both set to 0.8. The thresholds for the relative magnitude

test for the stability of features (see section 3.2.3) are set to 0.5 for the $\hat{x}$ direction and 0.7 for the $\hat{y}$ direction. The thresholds for the local frequency test for stability (equations (3.128) and (3.129)) are set to 0.5 for both the $\hat{x}$ and $\hat{y}$ directions.

5.2 Standards for Comparisons

This section defines the standards for judging the utility of the obstacle detection algorithm and its various modules. Three standards will be used to compare measured data with actual values. The difference between the measured and actual values can be compared to the accuracy of methods used by other researchers, the expected error estimated by the algorithm, and the accuracy required to discriminate between obstacles and objects.

It is difficult to make useful comparisons with other researchers. As will be seen in the following eight experiments, camera-based results depend heavily on image, scene, and motion quantities. An additional problem (perhaps due to the previous observation) is that there are few published results for real image sequences. Despite these obvious difficulties, I will attempt to formulate standards for judging the accuracy of disparity and depth, normal image velocity, and the direction of sensor motion.

Matthies et al [39] published disparity and depth results for a flat tiger poster that is the same as the one used in experiment 1. The RMS error in the disparity is 0.12 pixels. The RMS error in depth is 0.5 percent of the actual depth. It is also useful to compare the accuracy to other sensors: a laser range finder with 256 levels [48] has a depth accuracy of 0.4 percent of the maximum depth [1]. I consider the results of [39] to be "very good."

Weng et al [52] judge the accuracy of the image velocity using the RMS difference

[1] In most cases, 0.4 percent of the maximum depth will be less accurate than 0.5 percent of the actual depth.

between the measured image velocity field and the field predicted by the inter-frame sensor motion; that is, the RMS error is given by

$$\Delta V_{RMS} = [\frac{r_{iv}}{N}]^{0.5}, \tag{5.316}$$

where r_{iv} is defined by (2.75). In this work, the RMS error is given by

$$\Delta V_{n,RMS} = [\frac{r_{niv}}{N}]^{0.5}, \tag{5.317}$$

where r_{niv} is defined by (2.77). In the real image sequence found in [52], the RMS error is 0.84 pixels. It is claimed in [52] that an RMS error less than one pixel is "satisfactory."

The accuracy of the direction of sensor translation is dependent on many conditions: the speed of translation, the distribution of features in the image, and the variety of depths in the scene. Large inter-frame translation tend to improve the directional accuracy; a sparse number of features, clustering of the features in a small portion of the image, or clustering of features in a small depth interval will produce ambiguous image velocity fields which tend to reduce the directional accuracy. Establishing a fixed "standard" for judging the directional accuracy ignores the (inherent) translation-rotation ambiguity [3] that can exist when sensor motion is estimated from the image velocity field. Acknowledging its limitations, I will attempt to determine a fixed standard by comparing the directional accuracy reported by other researchers. Table 5.1 contains published results for real image sequences produced by a camera (or cameras) undergoing predominantly axial translation ($T_z > T_x, T_y$). If the reference contains more than one example of axial translation, the best result is listed. Table 5.1 also lists any assumptions used to improve the results [2]. It appears that a directional error of less than 1.0 degree in each of the pan and tilt directions can be considered "very good."

The second standard for judging accuracy is the expected error. In this work, each measurement of disparity and normal image velocity is accompanied by an expected error

[2]The no rotation assumption is often used to avoid the translation-rotation ambiguity.

Table 5.1: Comparison of Directional Accuracies

Researcher	Directional Error	Assumptions
Matthies [40] [42]	< 1.0 degree	Stereo, 3D point matching, Planar motion (T_x, T_z, Ω_y)
Hayashi and Negahdaripour [26]	4.3 degrees	Correspondenceless stereo, Direct, No rotation
Heel and Negahdaripour [30]	0.7 degrees	Monocular, Direct, 10 image integration, No rotation, Frontal plane (one depth)
Adiv [2]	1.0 degree	Monocular, Requires optical flow

estimate. These expected errors are propagated to other modules as error covariance matrices. The expected error is useful for judging the accuracy of motion estimates and collision parameters because it accounts for any inherent ambiguities that arise due to poor feature distribution. If the difference between the measured and actual values (of a motion or collision parameter) is within the expected error, the measured value is said to be "consistent" with the actual value.

The third standard is the accuracy required to successfully complete the task at hand. In this work, the task is to estimate the collision parameters of objects with sufficient accuracy that the computer pilot can avoid any obstacles. The accuracy of the point-of-collision required for obstacle detection is related to the size of the object/obstacle, and the baseline separation of the stereo cameras (or the size of the ego-vehicle).

5.3 Data Set 1

Data set 1 tests the following modules: disparity, normal image velocity, inter-frame sensor motion, and the extended sensor motion. Two stereo image sequences are formed by viewing stationary environments with forward translating cameras. The first sequence has a flat scene structure with one depth; the second sequence has a variety of depths.

Both stereo image sequences were obtained from the Calibrated Imaging Lab at Carnegie Mellon University [3].

The image sequences are obtained from an optical bench, which ensures precise motion control. The stereo image sequence is produced using one camera; the baseline separation is obtained by moving the camera along the x-axis. This technique guarantees that the focal lengths used in the right and left image sequences are matched. The stereo baseline is 2.54cm (1.0 inch). The optical axes of the stereo cameras are parallel. The direction of sensor translation is along the z-axis (axial translation).

The nominal camera parameters are as follows: the focal length of camera is 16mm; the physical size of the CCD array is 6.6mm by 8.8mm; and the image size is 480 x 512 pixels. The effects of using nominal camera parameters instead of the actual values is discussed in appendix C and in the section summary.

5.3.1 Experiment 1: Tiger Poster

In this experiment, stereo cameras move towards a stationary poster. A stereo pair from the image sequence is shown in figure 5.19. The poster contains the face of a tiger. The poster is flat and its surface normal is parallel to the z-axis; that is, the scene structure has one depth. The image projection of the poster contains many uni-directional features that comprise a variety of normal directions.

Using the theory outlined in the previous chapters, we can make predictions regarding the performance of various modules of the algorithm for this particular image sequence. The disparity module should perform well because the scene structure is simple (constant depth). All of the stereo correspondences can be made using the E_{offset} histogram for the lowest frequency channel and the multiscale prediction for the higher frequency channels

[3]The sequences were supplied by Larry Matthies who is currently working at the Jet Propulsion Laboratory.

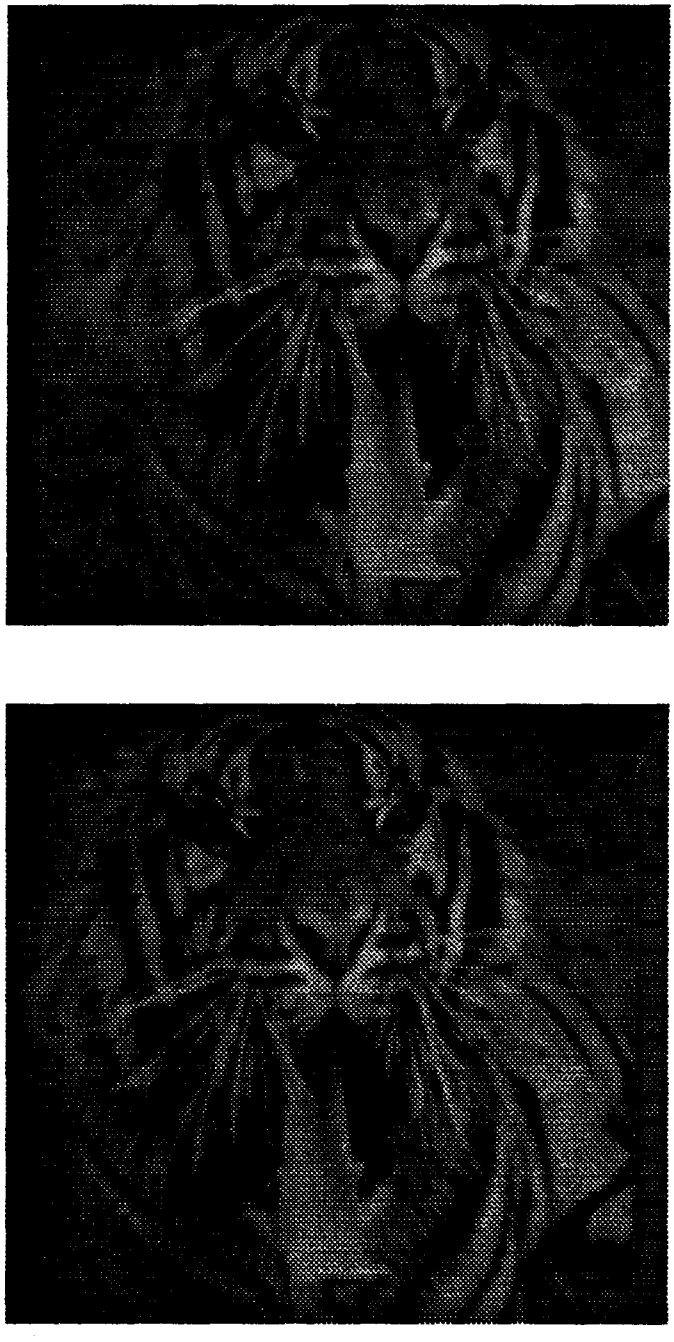


Figure 5.19: Experiment 1, Stereo Images *(upper)* Left Image, *(lower)* Right Image.

(see section 4.6.1); the heuristic ordering constraint will not produce additional matches. The normal image velocity module should also perform well because the axial motion is small. Generating the correct temporal correspondences will be simple because most of the matches will have no lattice offset.

The Hessian matrix used to estimate inter-frame sensor motion will be ill-conditioned due to the lack of variation in depth. As a result, the x component of motion may be incorrectly distributed between T_x and Ω_y. Similarly, the y component may be incorrectly distributed between T_y and Ω_x. Enforcing motion constraints should improve the inter-frame parameter estimates.

The objectives of this experiment are: to measure the structure of the scene and compare with the flat poster; to measure the normal image velocity and compare with the flow pattern predicted by the sensor motion; to measure and compare the two inter-frame sensor motions; to determine if the motion constraints improve the inter-frame sensor motion estimates; and to measure the extended sensor translation. Success of this experiment will verify the correct operation of the phase-based measurements of disparity and normal image velocity. It will also verify the correct operation of the primary stereo correspondence predictors: the E_{offset} histogram and the multiscale prediction.

The interpolated disparity and the associated uncertainty, for each of the three epipolar channels, are shown in figures 5.20, 5.21, and 5.22. There is a total of 284 stereo feature pairs across the three epipolar channels. The disparity is approximately constant throughout the image. The average disparity of these features is 51.07 pixels. The measured standard deviation is ± 0.16 pixels. It is comparable with the RMS error of ± 0.12 pixels reported in [39].

The top view and side view of the local map (observer coordinate frame) are shown in figure 5.23. The top view (the upper image) is the projection of stereo features from all epipolar channels onto the x-z plane. The side view (the lower image) is the projection

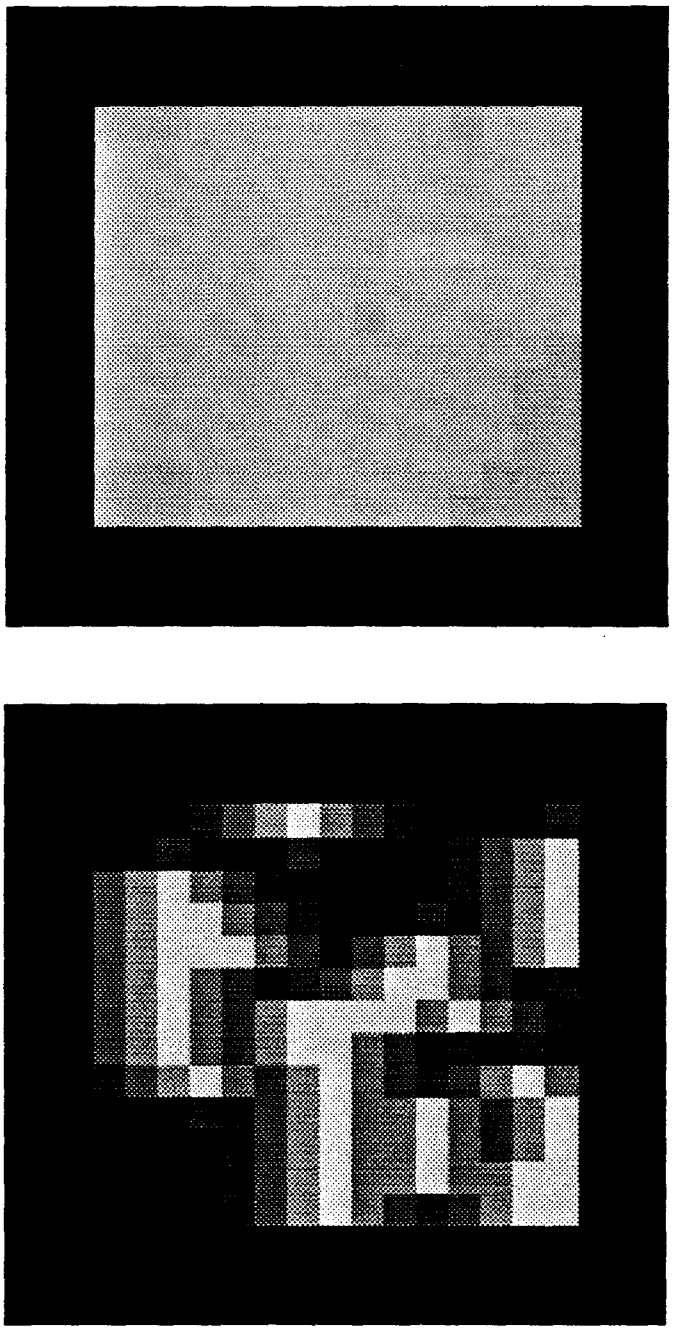


Figure 5.20: Experiment 1, $\tilde{\omega}_0 = 0.040\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 42 pixels and 52 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements.

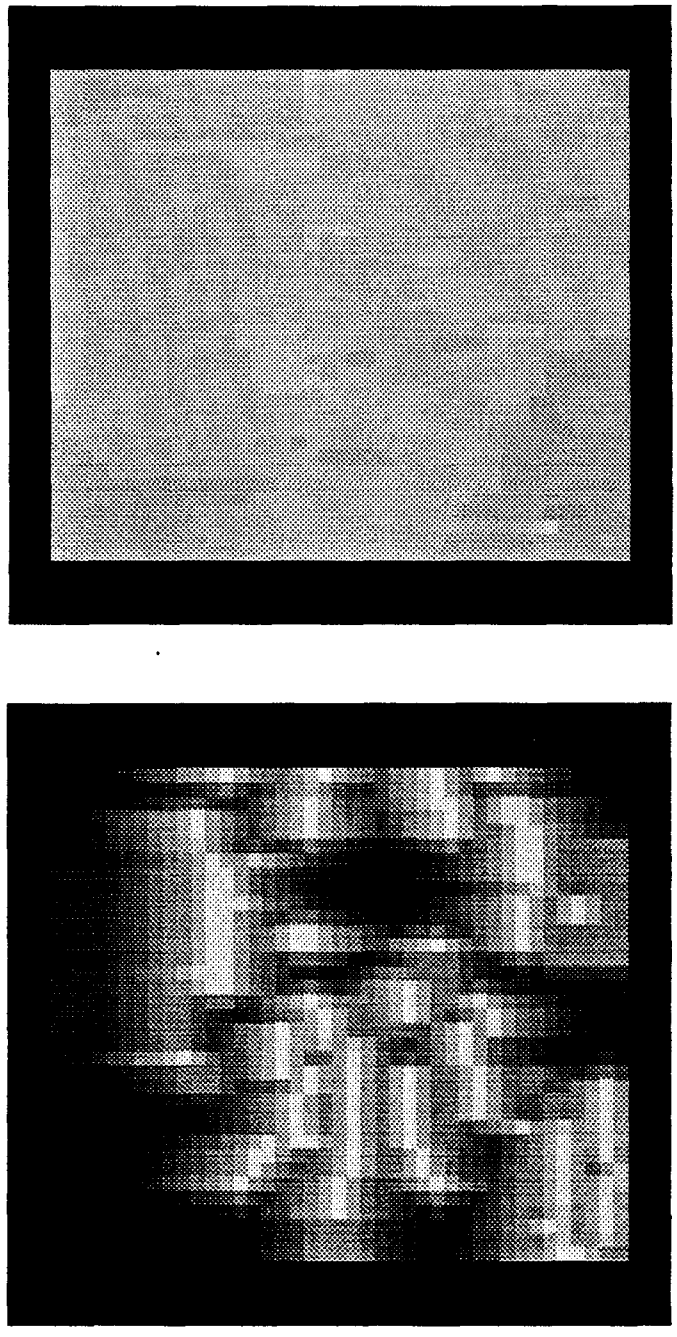


Figure 5.21: Experiment 1, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 42 pixels and 52 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements.

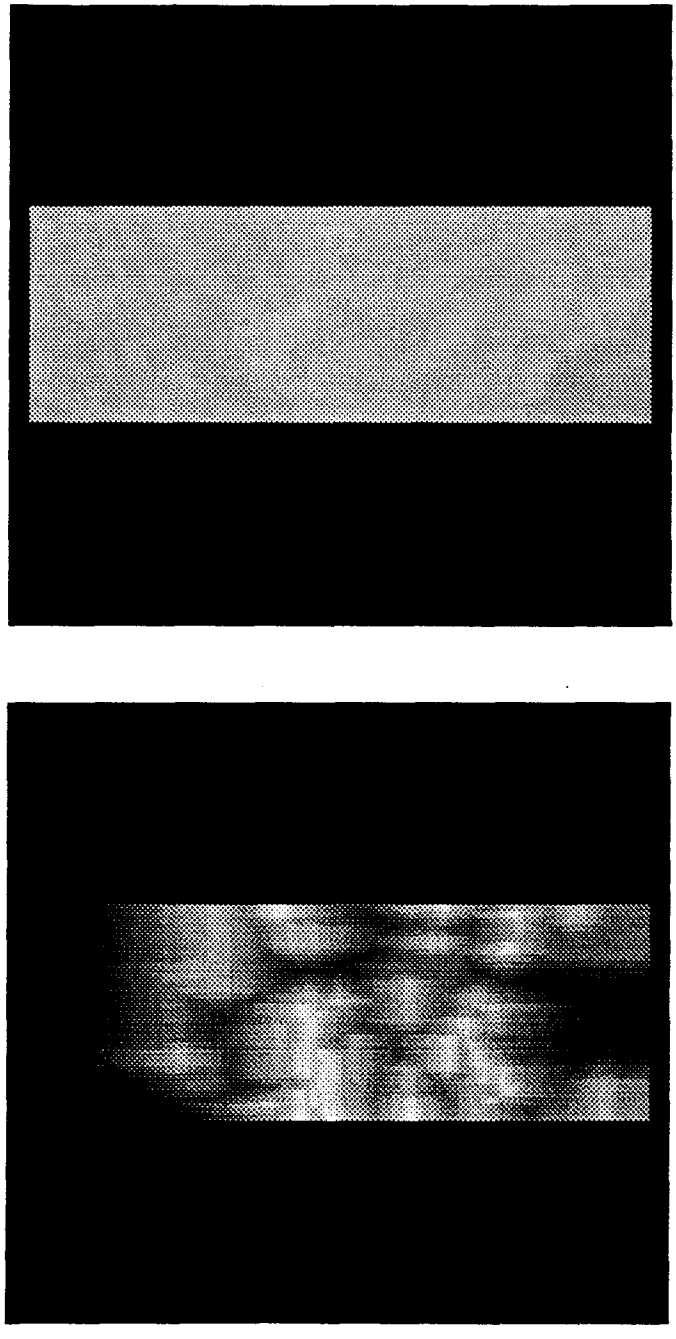


Figure 5.22: Experiment 1, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 42 pixels and 52 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements.

Table 5.2: Inter-frame Sensor Motion for Experiment 1

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.0452	0.0416	0.7194	0.856	-1.232	0.043
1-2	-0.0018	0.0036	0.7204	0.056	-0.144	-0.051

of the stereo features onto the y-z plane. In each view of the local map, the observer origin is positioned at the bottom-center tick.

The local map correctly illustrates the flat planar structure of the scene. The average depth is 46.30 cm and the standard deviation is ± 0.14 cm. Thus, the RMS error in depth is 0.3 percent of the average value, which is better than the standard described in section 5.2.

The normal image velocity measurements (also referred to as "component flow vectors") for the four orientations with the channel frequency of 0.092π radians per pixel are shown in figures 5.24 and 5.25. It can be seen that the normal direction of each component flow vector is within the orientation bandwidth of its respective channel. All four channels display flow patterns that are characteristic of a sensor undergoing axial motion: the component flow vectors point away from the image origin and the speed increases with the (normal) distance from the origin. The RMS error in the measured normal image velocity field, compared to the field predicted by the inter-frame sensor motion, is 0.10 pixels, which is better than the one pixel standard described in section 5.2.

The two inter-frame sensor motions and the expected errors appear in tables 5.2 and 5.3, respectively. The two inter-frame sensor motions are consistent with each other; that is, the parameter differences between the two inter-frame transitions are less than

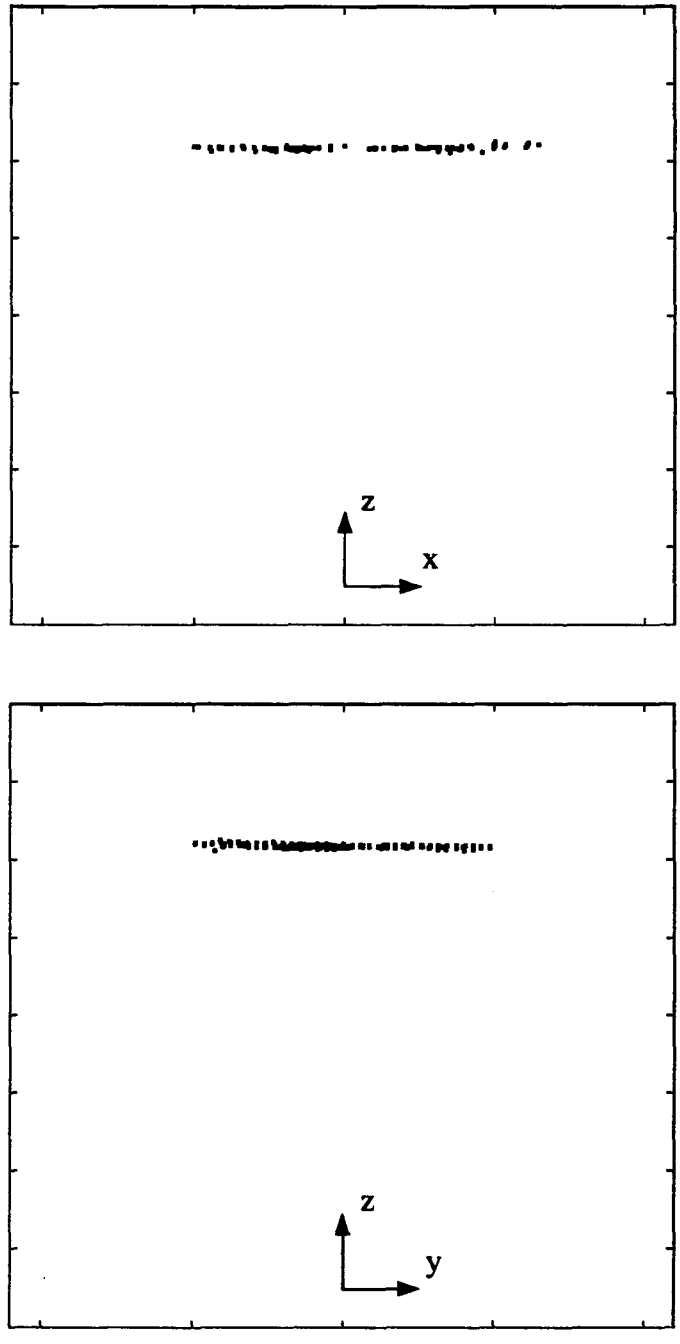


Figure 5.23: Experiment 1, Local Map. Stereo features are denoted by black squares. Distance between ticks is 7.5 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

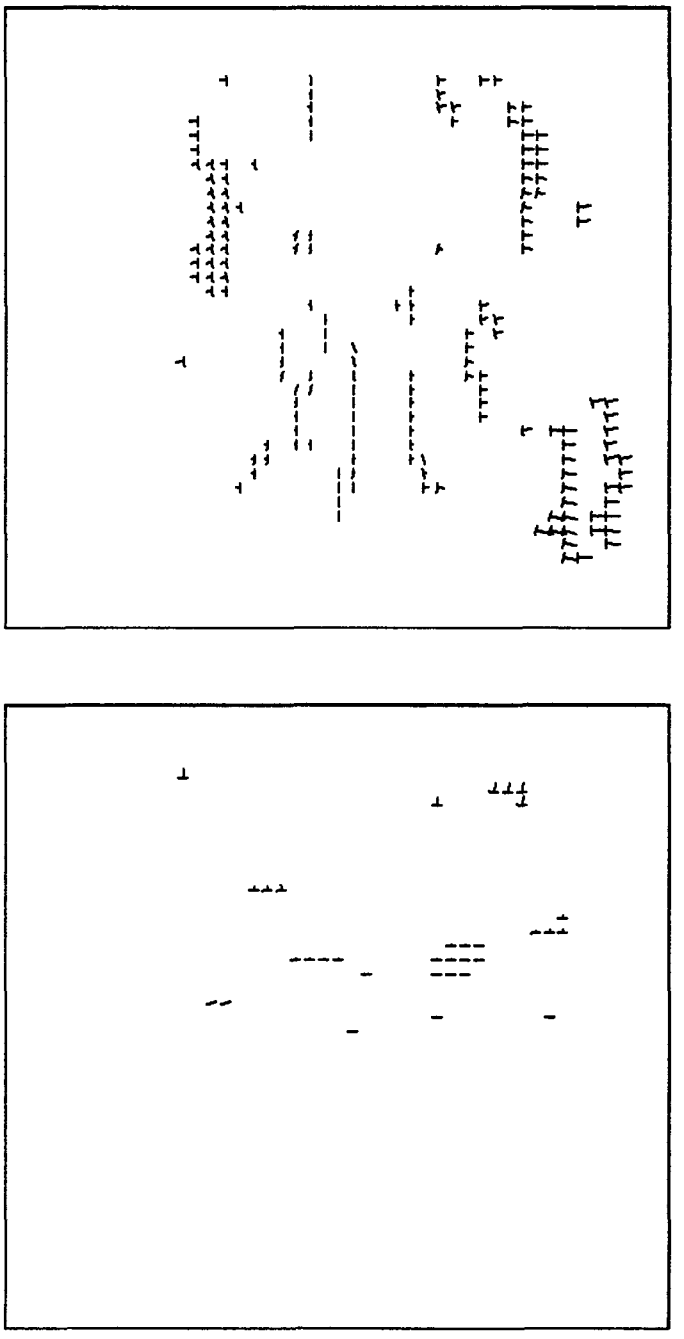


Figure 5.24: Experiment 1, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. The lengths of the vectors have been multiplied by 5.0. *(upper)* Epipolar Channel, $\tilde{\phi}_0 = 0$, *(lower)* Orthogonal Channel, $\tilde{\phi}_2 = \frac{\pi}{2}$ radians.

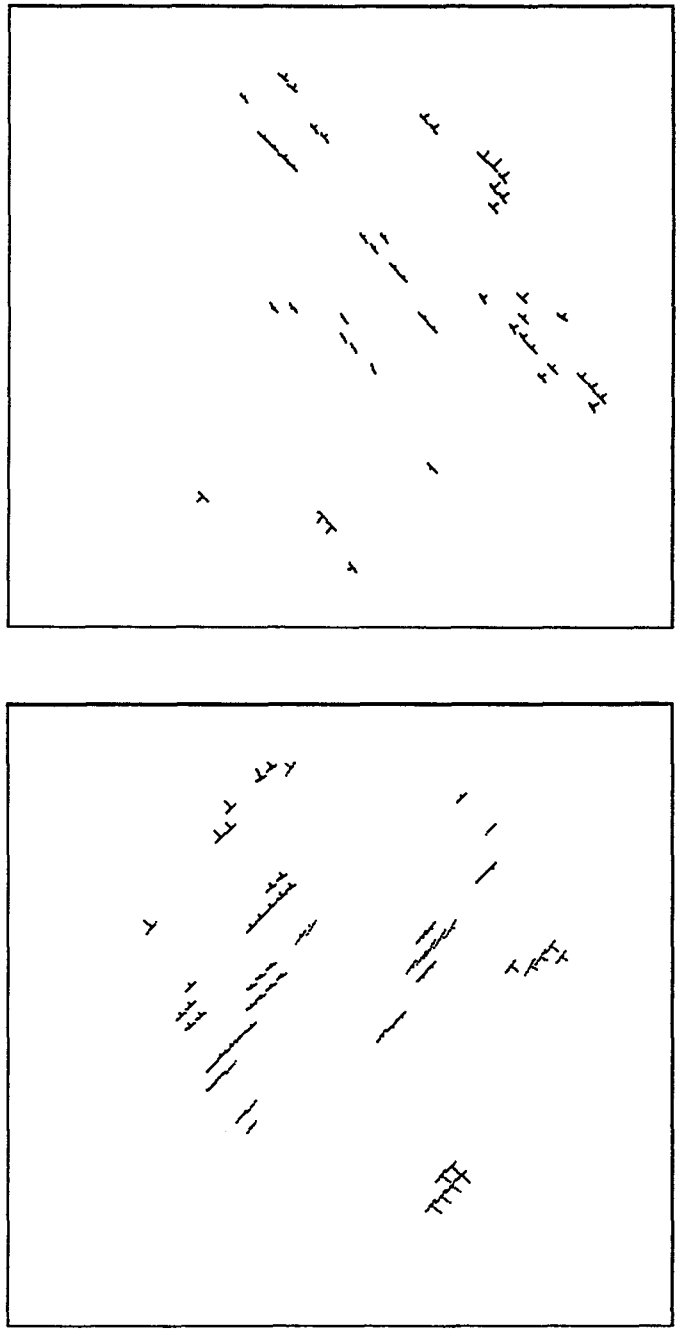


Figure 5.25: Experiment 1, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. The lengths of the vectors have been multiplied by 5.0. *(upper)* $\tilde{\phi}_1 = \frac{\pi}{4}$, *(lower)* $\tilde{\phi}_3 = \frac{3\pi}{4}$ radians.

Table 5.3: Expected Error in Inter-frame Sensor Motion for Experiment 1

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	± 0.1360	± 0.2002	± 0.0163	± 4.246	± 2.875	± 0.370
1-2	± 0.1329	± 0.1883	± 0.0144	± 4.052	± 2.848	± 0.349

the expected errors. The inter-frame sensor motions are also consistent with the axial motion; the direction of translation ($\frac{T_x}{T_z}$, $\frac{T_y}{T_z}$) and the rotation are approximately zero (within the expected errors).

An eigenvalue decomposition of Q_{int} shows that the inter-frame parameters T_x, T_y, Ω_x, and Ω_y are sensitive to measurement errors. The six eigenvalues and eigenvectors are as follows [4]:

$$\lambda_0 = 948781 \quad \bar{v}_0 = [\ 0.702 \quad 0.002 \quad -0.004 \quad -0.001 \quad 0.712 \quad 0.005\ ]^T,$$
$$\lambda_1 = 339083 \quad \bar{v}_1 = [\ -0.002 \quad 0.703 \quad -0.029 \quad -0.710 \quad -0.001 \quad 0.011\ ]^T,$$
$$\lambda_2 = 4482.5 \quad \bar{v}_2 = [\ -0.023 \quad -0.021 \quad -0.820 \quad 0.021 \quad 0.015 \quad 0.571\ ]^T,$$
$$\lambda_3 = 3727.3 \quad \bar{v}_3 = [\ -0.004 \quad -0.010 \quad -0.571 \quad 0.001 \quad 0.007 \quad -0.821\ ]^T,$$
$$\lambda_4 = 33.123 \quad \bar{v}_4 = [\ -0.674 \quad 0.229 \quad 0.024 \quad 0.227 \quad 0.664 \quad -0.010\ ]^T,$$
$$\lambda_5 = 13.362 \quad \bar{v}_5 = [\ 0.229 \quad 0.673 \quad 0.024 \quad 0.666 \quad -0.226 \quad -0.002\ ]^T.$$

It can be seen, by comparing eigenvalues λ_0 and λ_5, that the inter-frame motion estimate is poorly conditioned (the condition number is 71000). Any constraints that increase λ_4 and λ_5 will improve the estimates of T_x, T_y, Ω_x and Ω_y. Two constraints are considered: the known rotation constraint, which will increase both λ_4 and λ_5; and the motion along a known plane constraint, which will increase λ_5.

For the known rotation case, the rotation and the uncertainty are assumed to be

[4]The rotation terms in the eigenvectors have been normalized by the average scene depth, $z_{norm} = 46$ cm.

Table 5.4: Inter-frame Sensor Motion, Known Rotation, for Experiment 1

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	-0.0107	0.0016	0.7209	0.0069	-0.0522	0.0177
1-2	-0.0082	0.0010	0.7207	-0.0002	-0.0071	-0.0402

Table 5.5: Inter-frame Sensor Motion, Known Plane Constraint, for Experiment 1

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.0356	0.0010	0.7195	-0.0045	-1.029	0.0324
1-2	-0.0019	0.0005	0.7205	-0.0102	-0.1418	-0.0388

$(\Omega_x,\ \Omega_y,\ \Omega_z) = (0.000 \pm 1.000,\ 0.000 \pm 1.000,\ 0.000 \pm 1.000)\ 10^{-3}$ radians per frame. The inter-frame sensor motion for the known rotation constraint appears in table 5.4. The surface normal of the known plane is $n_p = [0\ 1\ 0]^T$. The weighting terms used in the known plane constraint are $\lambda_T = 472.6$ and $\lambda_\Omega = 10^6$, which produce the following auxiliary inter-frame estimates:

$$T_y = 0.000 \pm 0.046 \text{ cm per frame,} \tag{5.318}$$

$$\Omega_x = 0.000 \pm 1.000\ 10^{-3} \text{ radians per frame,} \tag{5.319}$$

$$\Omega_z = 0.000 \pm 1.000\ 10^{-3} \text{ radians per frame.} \tag{5.320}$$

The inter-frame sensor motion for the known plane constraint appears in table 5.5. The known rotation constraint improves the accuracy of T_x, T_y, Ω_x, and Ω_y. The known plane constraint improves the accuracy of T_y and Ω_x.

The extended sensor motion appears in table 5.6. The extended sensor motion is

Table 5.6: Extended Sensor Motion for Experiment 1

Frame	cm/frame			Pred. Error cm/fr		
	T_x	T_y	T_z	ΔT_x	ΔT_y	ΔT_z
0-1	0.0452	0.0416	0.7194	± 0.136	± 0.200	± 0.016
0-2	0.0211	0.0202	0.7198	± 0.093	± 0.136	± 0.011

consistent with the axial translation, and improves as more images are integrated. The final estimate of the direction of translation along the x- and y-axes (pan and tilt) are 0.029 radians (1.68 degrees) and 0.028 radians (1.61 degrees), respectively. The pan and tilt directional errors are slightly larger than the one degree standard established in section 5.2. These results are surprisingly accurate considering the inherent translation-rotation ambiguity that exists when viewing frontal planes.

To summarize experiment 1, the image measurements of disparity and normal image velocity are very good. The disparity measurements provide an accurate representation of the scene structure, verifying the correct operation of the E_{offset} histogram and the multiscale prediction. The normal image velocity measurements produce a flow pattern that is consistent with sensor motion. The two inter-frame sensor motions are consistent (within the expected error) with each other and with the axial motion. The known rotation and plane constraints improve the inter-frame sensor motion estimates, as predicted by the eigenvalue analysis of Q_{int}. The direction of the extended sensor translation estimate is within the expected error of axial motion.

5.3.2 Experiment 2: Model City

In this experiment, stereo cameras move towards a stationary model of a city. A stereo pair from the image sequence is shown in figure 5.26. The model city contains buildings,

cars, railroad tracks, and trees. The scene structure has a variety of depths, including some large depth gradients. The image projection of the model city contains many uni-directional features, primarily features with vertical and horizontal normal directions. The image contains some specular reflections from the railroad tracks.

Theoretical predictions can be made for this image sequence. The disparity module will be challenged by this scene structure. Since there are large depth gradients, the E_{offset} histogram and the multiscale prediction will miss some stereo features. The heuristic ordering constraint will make additional matches. The temporal constraint will propagate these matches into future stereo images.

The normal image velocity module should perform well because the axial motion is small. The inter-frame sensor motion estimate should be good: the conditioning of the inter-frame Hessian matrix will be much better than experiment 1 (the scene contains a variety of depths). As a result, the known rotation constraint will provide only a modest improvement in the estimate of inter-frame sensor motion.

The objectives of this experiment are: to measure the disparity and normal image velocity; to measure and compare the two inter-frame sensor motions; to determine if the known rotation constraint improves the inter-frame sensor motion estimates; and to measure the extended sensor translation. Success of this experiment will verify the correct operation of secondary stereo correspondence predictors: the temporal constraint and the heuristic ordering constraint.

The interpolated disparity and its uncertainty are shown in figures 5.27, 5.28, and 5.29. These figures illustrate the increasing resolution with channel frequency. The low frequency channel provides only a vague description of the scene structure. The scene details become discernible in the higher frequency channels.

The top and side views of the local map are shown in figure 5.30. There is a total of 298 stereo feature pairs across the three epipolar channels. The local map captures much

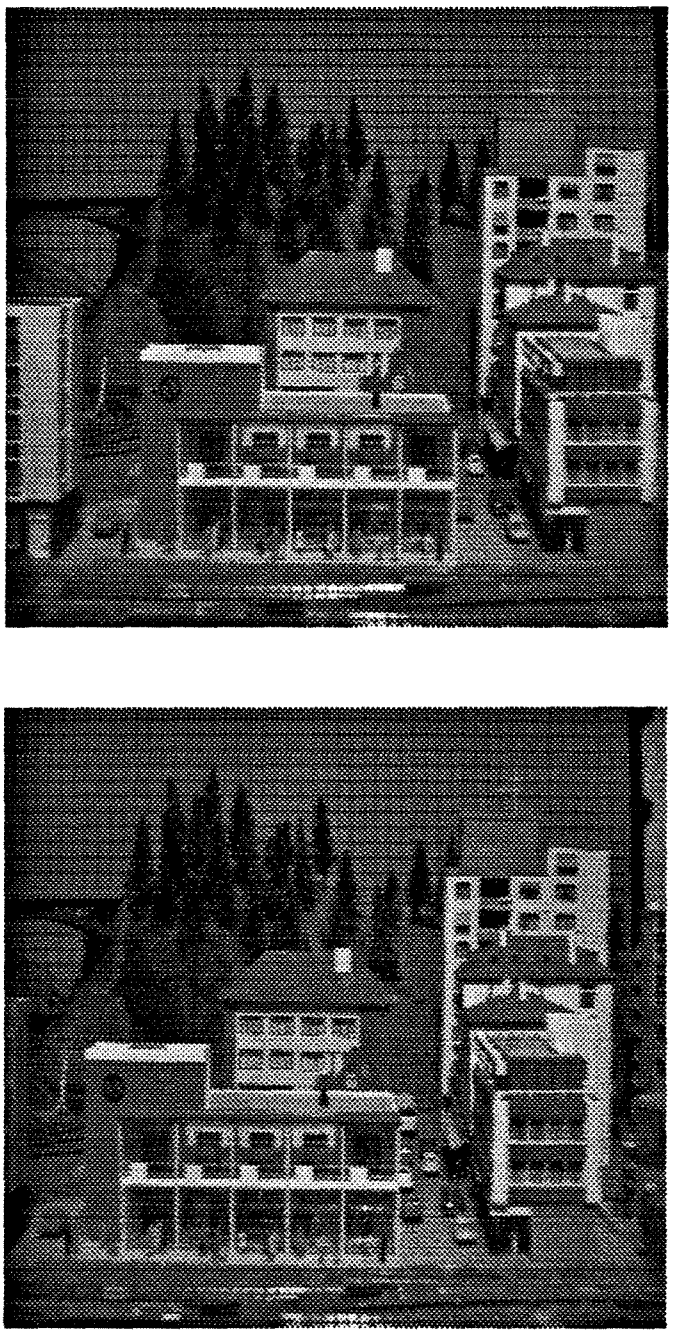


Figure 5.26: Experiment 2, Stereo Images *(upper)* Left Image, *(lower)* Right Image.

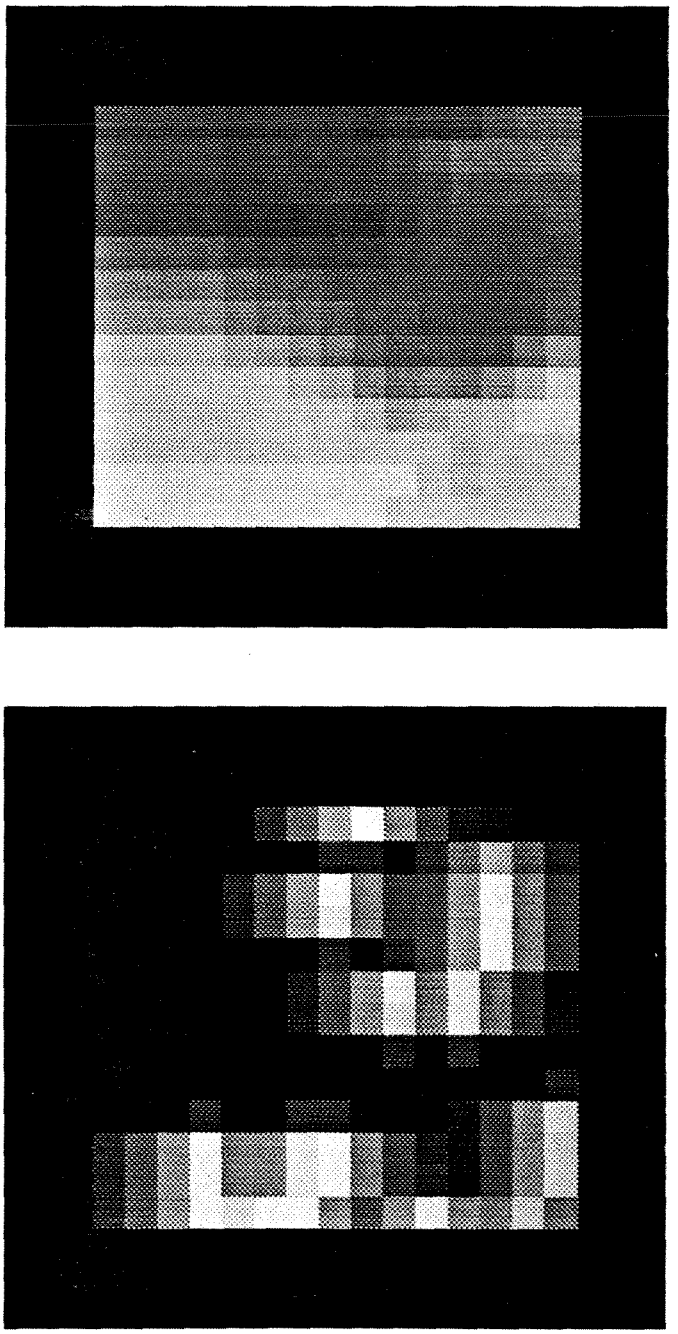


Figure 5.27: Experiment 2, $\tilde{\omega}_0 = 0.040\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 10 pixels and 45 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements.

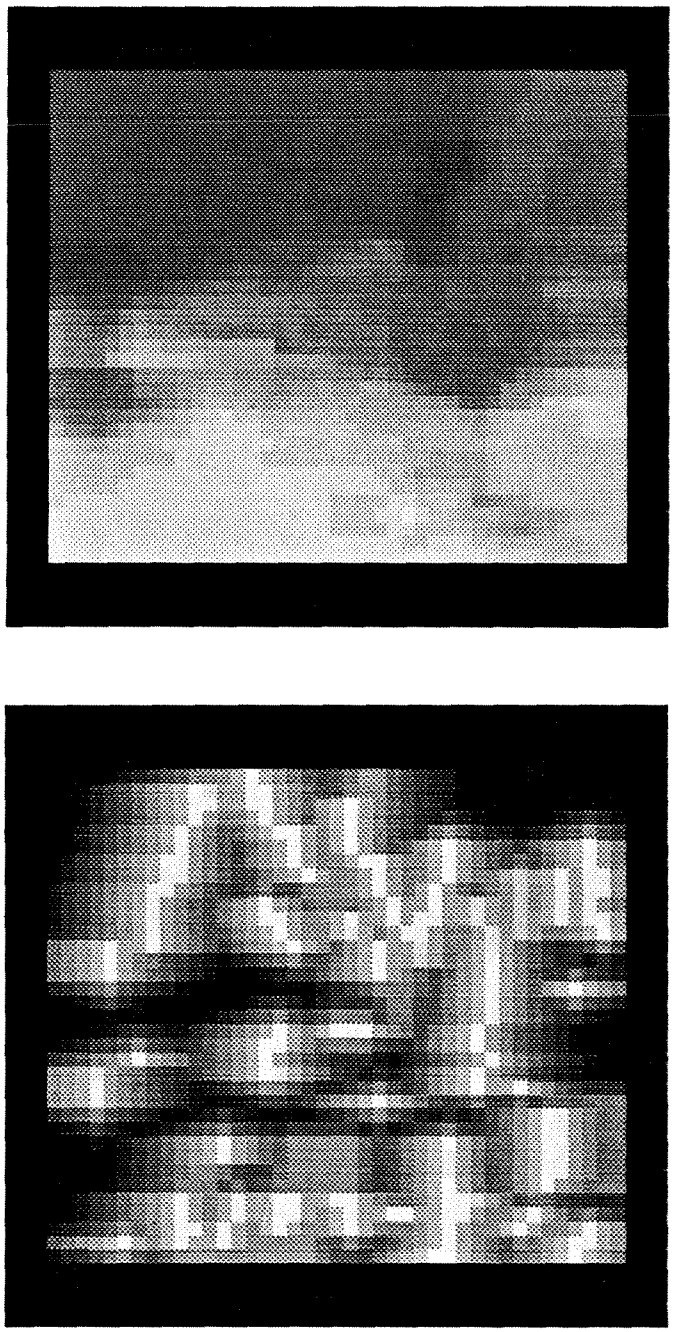


Figure 5.28: Experiment 2, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 10 pixels and 45 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements.

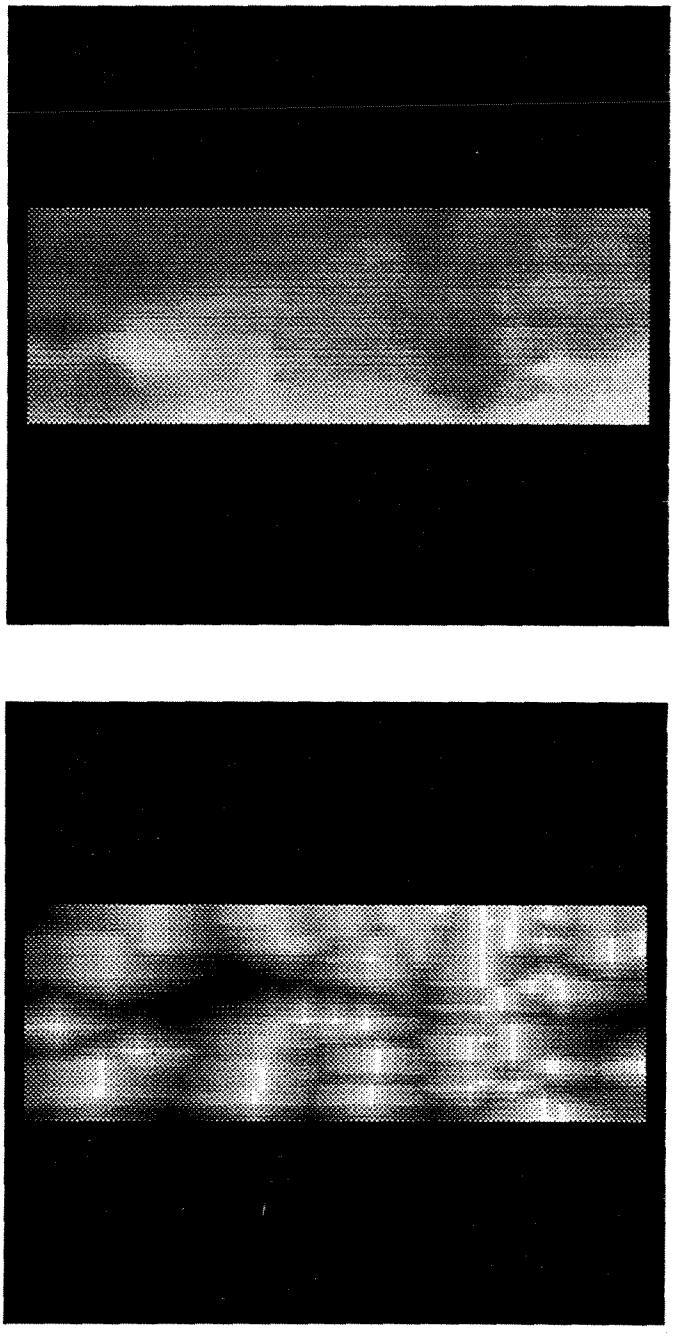


Figure 5.29: Experiment 2, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 10 pixels and 45 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements.

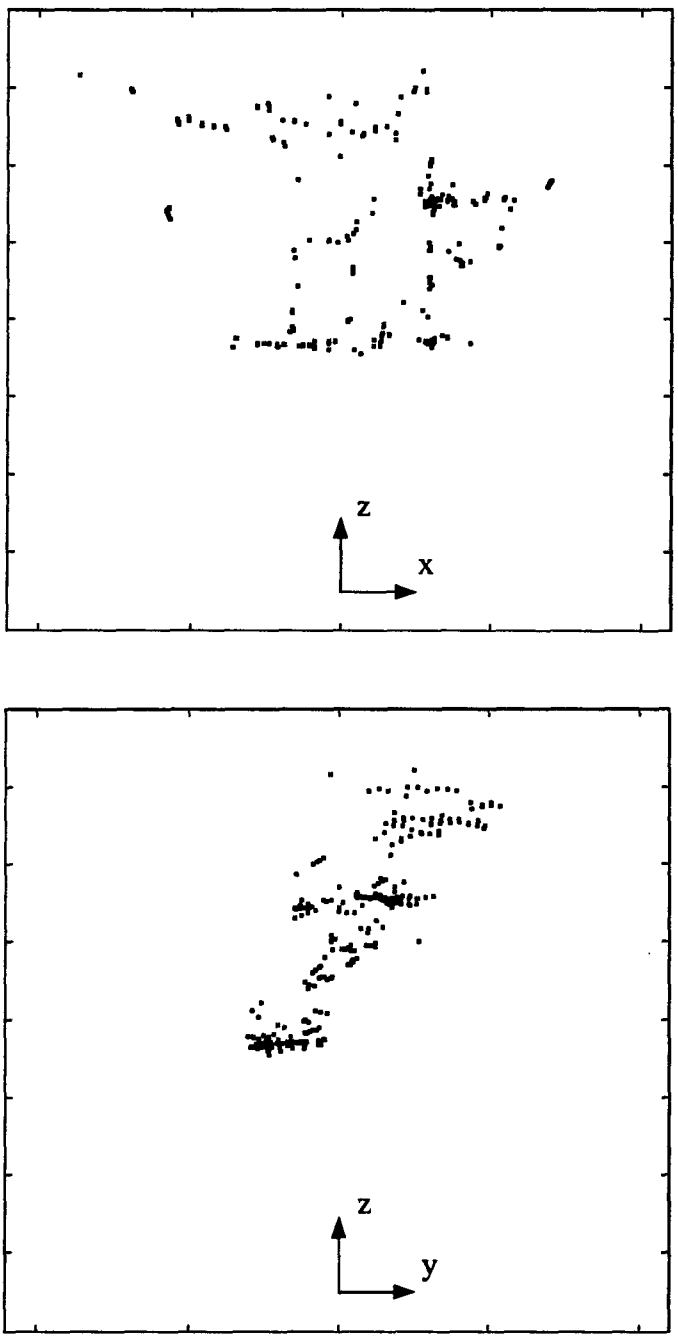


Figure 5.30: Experiment 2, Local Map. Stereo features are denoted by black squares. Distance between ticks is 15 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

Table 5.7: Inter-frame Sensor Motion for Experiment 2

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	-0.0101	-0.0055	0.4631	0.027	0.105	0.032
1-2	-0.0031	-0.0089	0.4772	-0.005	-0.117	0.094

Table 5.8: Expected Error in Inter-frame Sensor Motion for Experiment 2

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	± 0.0112	± 0.0135	± 0.0155	± 0.199	± 0.151	± 0.287
1-2	± 0.0111	± 0.0134	± 0.0158	± 0.197	± 0.150	± 0.290

of the scene structure. The flat surfaces, such as the front of the buildings, are correctly represented.

The normal image velocity measurements for the epipolar and orthogonal channels are shown in figure 5.31. The normal image velocity measurements display the outward flow associated with axial motion. The speed of the component flow vectors increase with the (normal) distance from the image origin and the disparity of the feature. The RMS error in the measured normal image velocity field, compared to the field predicted by the inter-frame sensor motion, is 0.09 pixels, which is better than the one pixel standard described in section 5.2.

The inter-frame sensor motions and the expected errors appear in tables 5.7 and 5.8, respectively. The inter-frame sensor motions are consistent with axial motion (within the expected errors). The two inter-frame translations are consistent with each other.

An eigenvalue decomposition shows that the inter-frame Hessian matrix is better conditioned in this experiment than in experment 1. The eigenvalues and eigenvectors

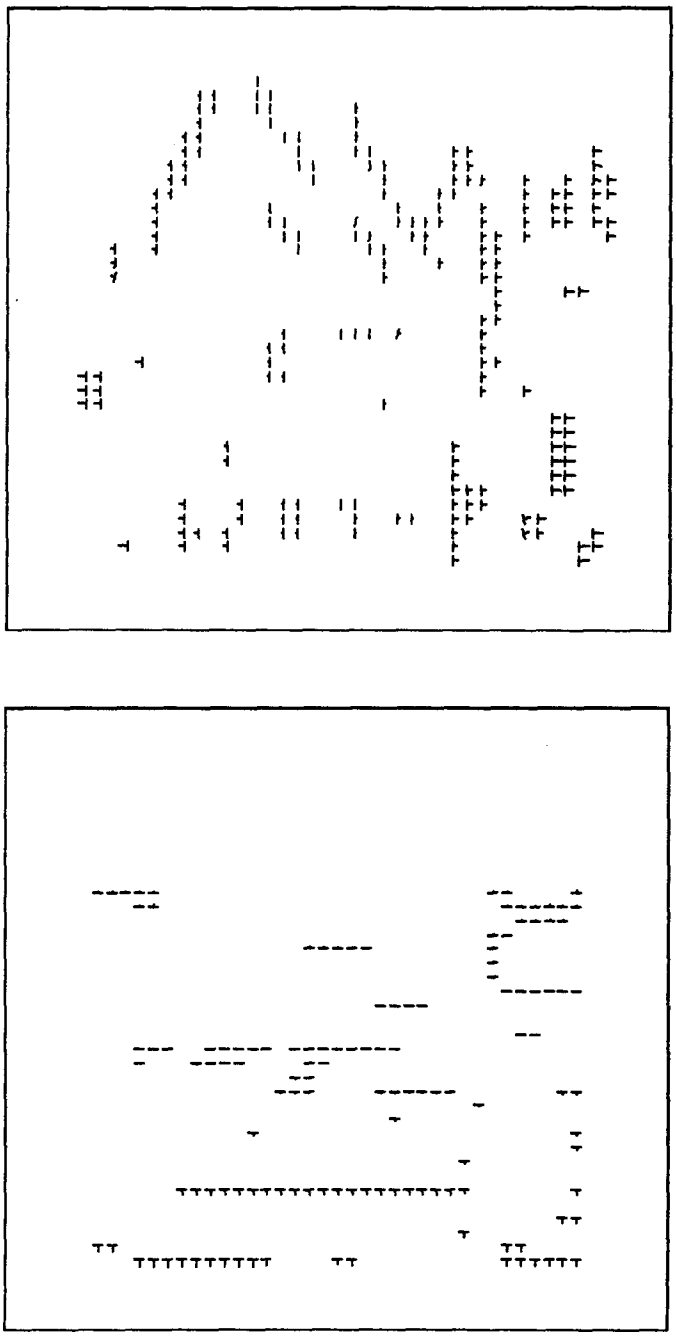


Figure 5.31: Experiment 2, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. The lengths of the vectors have been multiplied by 5.0. *(upper)* Epipolar Channel, $\tilde{\phi}_0 = 0$, *(lower)* Orthogonal Channel, $\tilde{\phi}_2 = \frac{\pi}{2}$ radians.

Table 5.9: Inter-frame Sensor Motion, Known rotation, for Experiment 2

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	-0.0095	-0.0057	0.4632	0.024	0.098	0.021
1-2	-0.0031	-0.0090	0.4772	-0.006	-0.115	0.083

are given by [5]

$$\lambda_0 = 544702 \quad \bar{v}_0 = [\ 0.723 \ -0.073 \ -0.030 \ 0.067 \ 0.683 \ 0.001\ ]^T,$$
$$\lambda_1 = 406505 \quad \bar{v}_1 = [\ -0.076 \ -0.758 \ -0.032 \ 0.644 \ -0.066 \ -0.009\ ]^T,$$
$$\lambda_2 = 7519.1 \quad \bar{v}_2 = [\ 0.583 \ -0.077 \ -0.096 \ -0.085 \ -0.622 \ 0.501\ ]^T,$$
$$\lambda_3 = 5646.7 \quad \bar{v}_3 = [\ 0.100 \ 0.288 \ 0.865 \ 0.386 \ -0.076 \ 0.064\ ]^T,$$
$$\lambda_4 = 2218.8 \quad \bar{v}_4 = [\ -0.164 \ -0.523 \ 0.448 \ -0.588 \ 0.195 \ 0.338\ ]^T,$$
$$\lambda_5 = 1486.3 \quad \bar{v}_5 = [\ -0.308 \ 0.241 \ -0.200 \ 0.280 \ 0.314 \ 0.794\ ]^T.$$

The condition number is 366, which is 194 times smaller than in experiment 1.

For comparison with experiment 1, the known rotation constraint is applied to the inter-frame sensor motion; the results appear in table 5.9. The rotation and uncertainty are assumed to be $(\Omega_x,\ \Omega_y,\ \Omega_z) = (0.000 \pm 1.000,\ 0.000 \pm 1.000,\ 0.000 \pm 1.000)\ 10^{-3}$ radians per frame. No significant improvement is obtained (for a rotational uncertainty of ± 0.001 radians per frame).

The extended sensor motion appears in table 5.10. The extended sensor translation is consistent with axial motion. The measured direction of translation is $(-0.014,\ -0.016)$ radians, or $(-0.79,\ -0.89)$ degrees, relative to the axial motion. The pan and tilt directional accuracies are better than the one degree standard established in section 5.2.

[5] The rotation terms in the eigenvectors have been normalized by the average scene depth, $z_{norm} = 78$ cm.

Table 5.10: Extended Sensor Motion for Experiment 2

	cm/frame			Pred. Error cm/fr		
Frame	T_x	T_y	T_z	ΔT_x	ΔT_y	ΔT_z
0-1	-0.0101	-0.0055	0.4631	± 0.0112	± 0.0135	± 0.0161
0-2	-0.0065	-0.0073	0.4702	± 0.0079	± 0.0095	± 0.0113

In summary, the disparity measurements are good, capturing the structure of the scene, and verifying the correct operation of the heuristic ordering constraint and the temporal constraint. The normal image velocity measurements produce a flow pattern consistent with the sensor motion. The variations in depth improve the estimate of inter-frame sensor motion over experiment 1. The motion constraint does not improve significantly the inter-frame sensor motion estimates, as predicted by the eigenvalue analysis. The direction of the extended sensor translation is within the expected error of axial motion.

5.3.3 Discussion and Summary

The actual motion of the cameras for experiments 1 and 2 are believed to be 0.762cm (0.3 inch) and 0.508cm (0.2 inch) per frame, respectively. The T_z component of the extended sensor translation for the two experiments are 0.720cm and 0.478cm, respectively. The percent error in the motion is approximately the same for both experiments: 5.8 for the tiger poster sequence, and 6.2 for the model city sequence. In both sequences, the error is limited to the speed of the sensor translation; the direction of translation is accurately measured.

The above-mentioned speed errors are caused by the camera parameter uncertainty. The nominal camera parameters are used instead of the actual values because calibration

data is not available. The speed errors are probably due to an incorrect focal length and to pixel scaling errors (see appendix C). Both errors affect the speed, not the direction, of translation [6].

A speed error can also be caused by an error in the baseline separation. Since the baseline separation is produced by precise lateral motion on the optical bench, the baseline error should be negligible.

To summarize this section, the image measurements are very good: stable features are selected, stereo and temporal correspondences are correctly made, disparity is measured to sub-pixel accuracy, and normal image velocity measurements are consistent with the component flow pattern predicted by the sensor motion. The local maps capture the scene structure. The inter-frame and extended sensor motions are consistent with axial motion. The known rotation and known plane constraints improve the inter-frame parameters when the Hessian is ill-conditioned. The condition number of the inter-frame Hessian matrix is a good measure of the extent to which the motion estimates are affected by the inherent translation-rotation ambiguity.

5.4 Data Set 2

The second data set contains three stereo image sequences which test the ability of the obstacle detection algorithm: to estimate motion in the presence of sensor rotation; to segment moving and stationary objects; and to predict collision parameters. In the first sequence, stereo cameras view a stationary scene while translating and rotating. In the second sequence, a translating object is on a collision trajectory with forward translating cameras. In the final sequence, the translating object has a trajectory that will pass safely in front of the forward translating cameras. The image sequences are obtained from an

[6] Focal length errors and symmetric pixel scaling errors are not serious problems for obstacle detection because they have no effect on the time-to-collision or the point-of-collision.

optical bench, insuring precise sensor and object motions. The stereo image sequences in data set 2 were obtained from the Laboratory for Computational Intelligence at the University of British Columbia [7].

Separate cameras are used to capture the right and left image sequences. Unfortunately, the stereo cameras could not be mounted with great precision. The left camera has a 0.01 radian roll, which is compensated by rotating (and re-sampling) the image about the optical axis. The left camera also has a slight upward tilt, requiring a -3 pixel offset along the $\hat{y}$-axis to approximate a parallel stereo setup (see section 2.5). A one pixel offset along the $\hat{x}$-axis compensates for a slight camera convergence. The stereo baseline is 5.2 ± 0.05cm. The nominal camera parameters are as follows: the focal length of each camera is 8.5mm; the physical size of the CCD array is 6.6mm by 8.8mm; and the image size is 480 x 512 pixels.

The sequences in data set 2 are affected by a problem with the image acquisition system. The acquisition system introduces random vertical offsets into the image sequence. Depending on when the frame grabbing process is initialized, relative to the camera synchronization pulse, the stereo images can either be registered correctly or offset vertically by two pixels. The random toggling between the zero and two pixel offsets is interpreted by the sensor motion module as transient rotations about the x-axis (pitch motion).

The same scene is used in all three experiments. A stereo image pair is shown in figure 5.32. The stereo images are viewing two plastics toys and a background poster. The toy at the right periphery of the image, the "eco-sub," is on a movable platform. The platform moves to the left in experiments 4 and 5. The toy on the left, the "toxic cannon," is stationary in each experiment. The toys and the rails of the movable platform produce specular reflections and shadows. Additional complications are that the brightness of the two images are slightly different, and the peripheral image features are compressed due

[7] Technical work was performed by Stewart Kingdon.

to lens distortion [8].

The interpolated disparity and its uncertainty are shown in figures 5.33, 5.34, and 5.35. The top and side views of the local map are shown in figure 5.36. There are 300 stereo feature pairs across the three epipolar channels. The background poster has a parabolic shape in figure 5.36, instead of planar. This parabolic shape is believed to be caused by lens distortion.

It is possible to compensate for the distortion in the depth estimate [9]. The compensated local map is shown in figure 5.37. Note that the surface normal of the plane is not parallel to the z-axis. This non-zero angle suggests that the focal lengths of the stereo cameras are mismatched.

In the following three experiments, the lens distorted (uncompensated) image sequences will be used.

5.4.1 Experiment 3: Camera Rotation

The purpose of this experiment is to test the robustness of the sensor motion module to sensor rotation. The stereo cameras are moving in a stationary environment: the sensor translation is along the z_w-axis; and the sensor rotation is about the y-axis.

Theoretical predictions regarding the performance of the various modules can be made for this image sequence. The normal image velocity module should perform well because all of the image features belong to stationary objects, and the axial motion is not large relative to the depth of the objects. The accuracy of the inter-frame sensor motion estimate will be degraded by the depth distortion. The parabolic depth distortion in this image sequence will cause T_z to be over-estimated (see appendix C).

[8] The arching of the top edge of the background poster board is believed to be due to radial lens distortion.

[9] The radial distortion constant k, found in appendix C, equation (C.376), is adjusted until the background poster is planar.

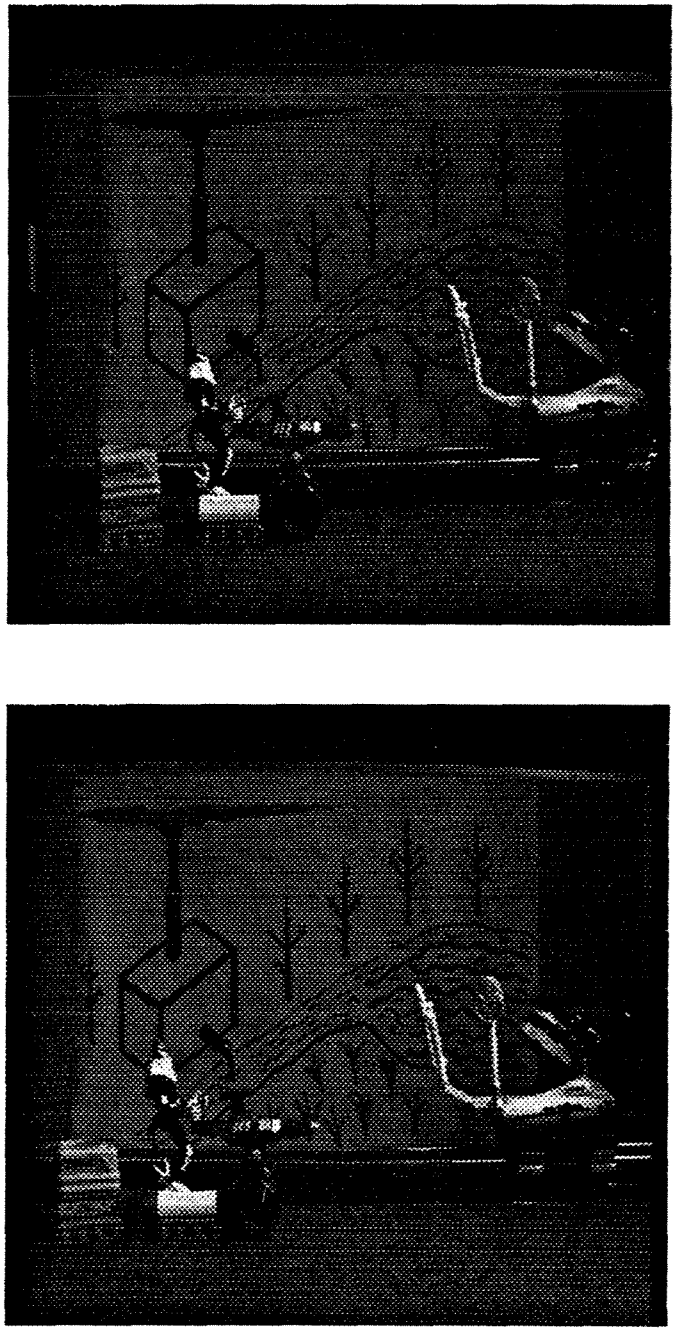


Figure 5.32: Experiment 3, Stereo Images *(upper)* Left Image, *(lower)* Right Image.

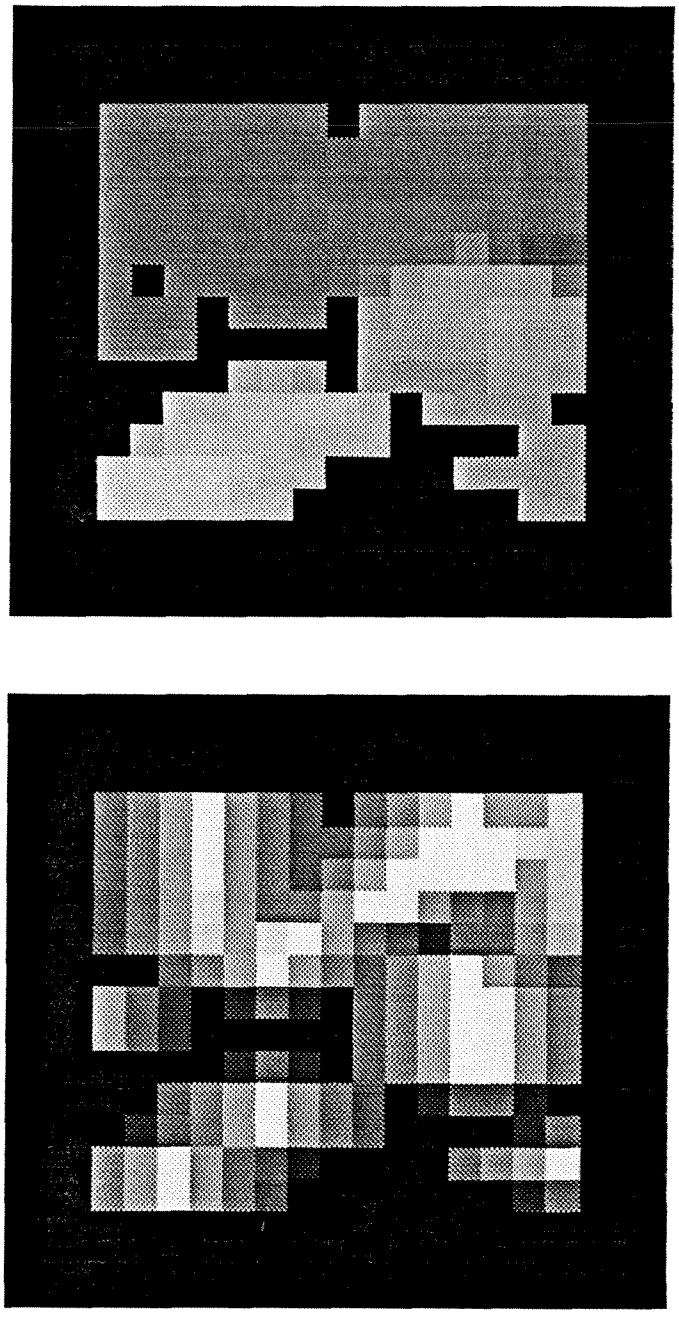


Figure 5.33: Experiment 3, $\tilde{\omega}_0 = 0.040\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 12 pixels and 30.3 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

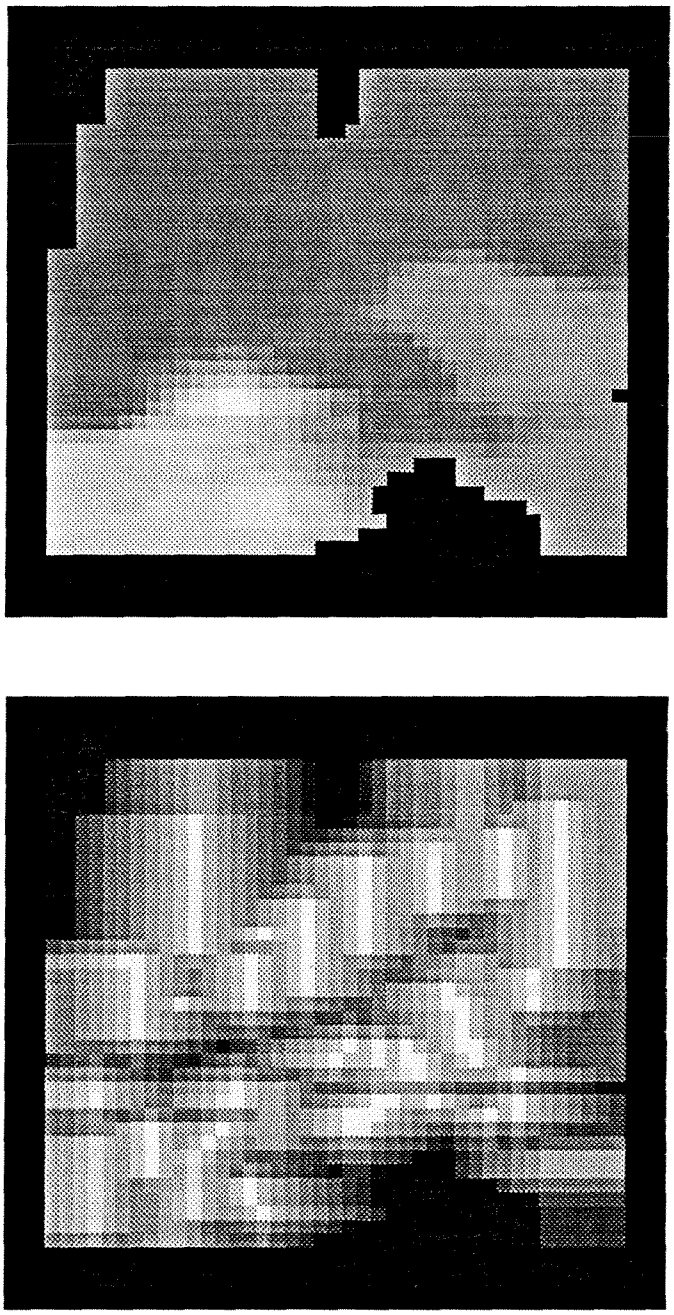


Figure 5.34: Experiment 3, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 12 pixels and 30.3 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

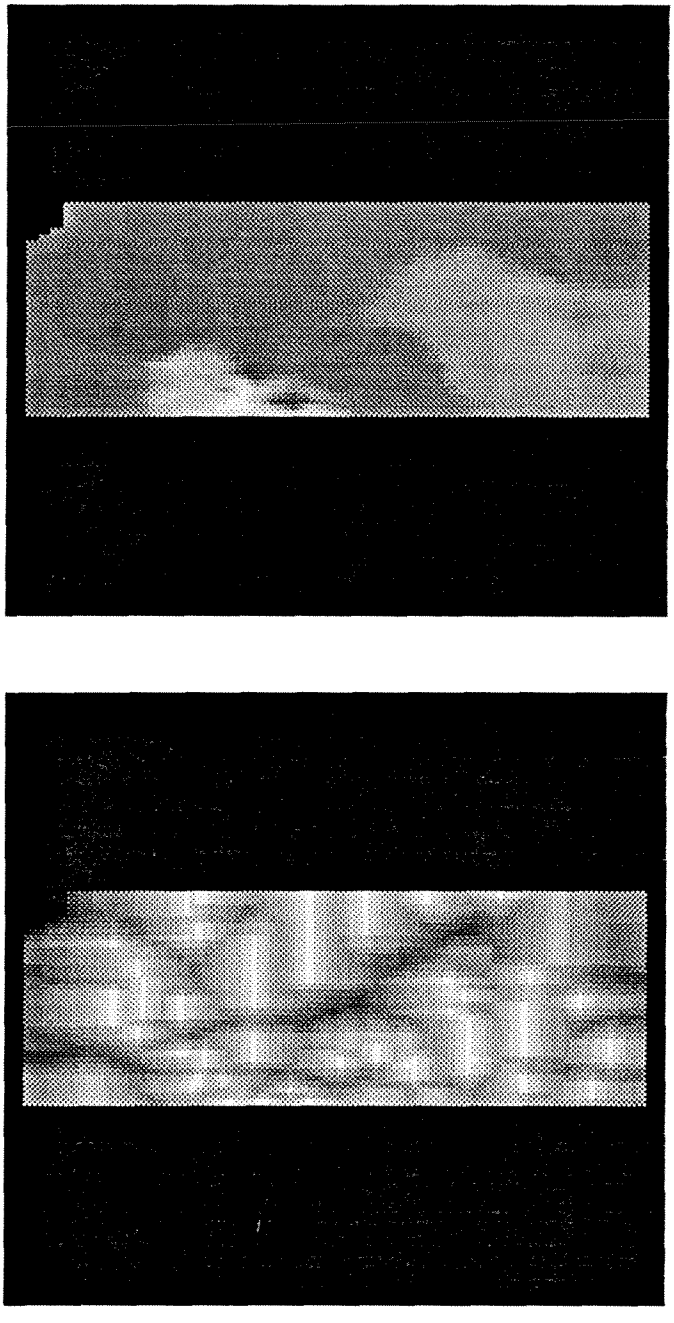


Figure 5.35: Experiment 3, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 12 pixels and 30.3 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

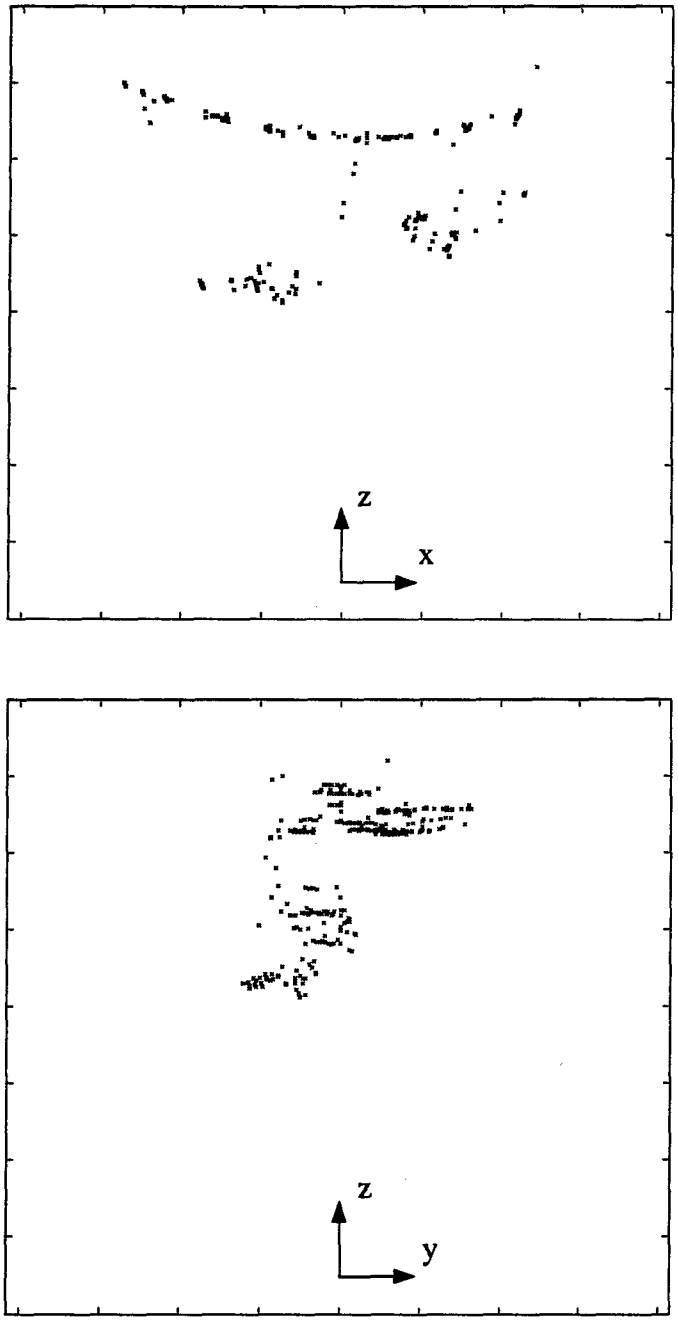


Figure 5.36: Experiment 3, Local Map. Stereo features are denoted by black "X"s. Distance between ticks is 20 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

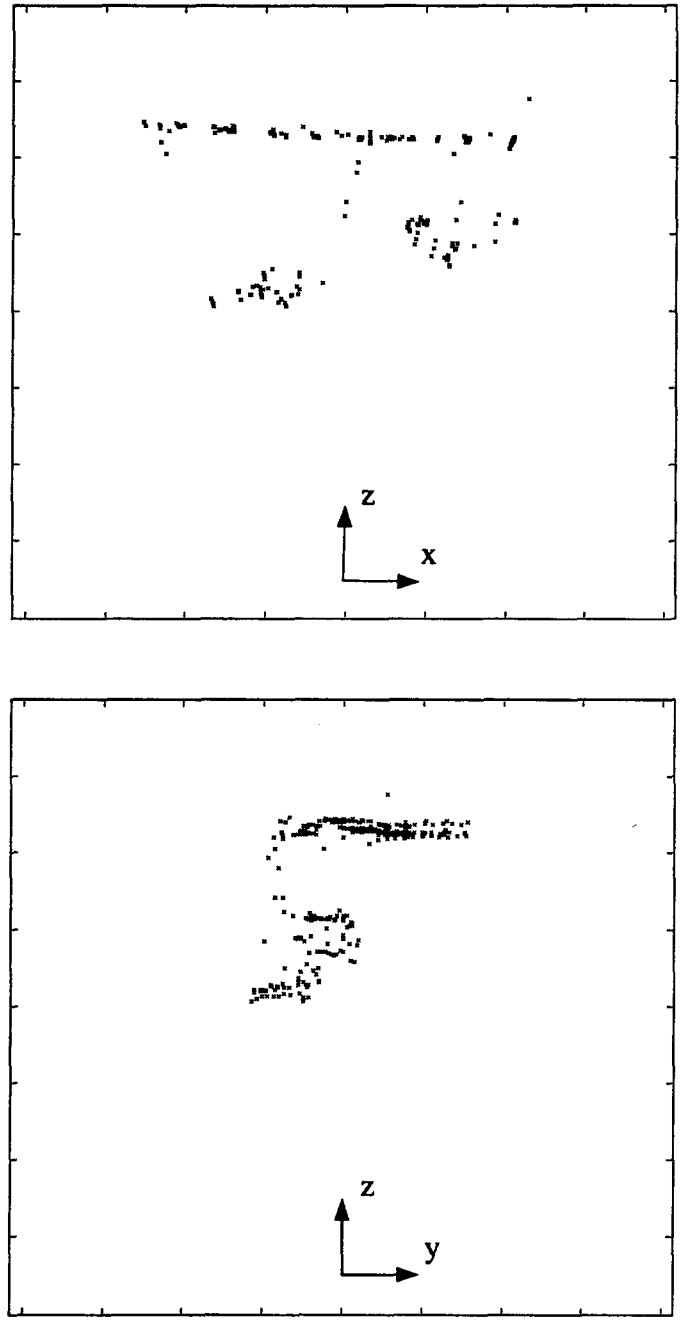


Figure 5.37: Experiment 3, Local Map with Distortion Compensated. Stereo features are denoted by black "X"s. Distance between ticks is 20 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

Table 5.11: Actual Inter-frame Sensor Motion for Experiment 3

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.0443	0.0000	2.5396	transient	8.727	0.000
1-2	0.0222	0.0000	2.5399	transient	8.727	0.000
2-3	0.0000	0.0000	2.5400	transient	8.727	0.000

The objectives of this experiment are: to measure the inter-frame sensor motions; and to measure the extended sensor translation. Success of this experiment will verify the ability of the inter-frame sensor motion stage to measure rotation, and the correct implementation of process model used in the Kalman filters.

The normal image velocity measurements for the epipolar and orthogonal channels are shown in figure 5.38. The epipolar channel displays the outward flow associated with axial motion and a -4 pixel offset associated with the rotation Ω_y. The orthogonal channel has an outward flow pattern. The RMS error in the measured normal image velocity field, compared to the field predicted by the inter-frame sensor motion, is 0.16 pixels, which is better than the one pixel standard described in section 5.2.

The actual inter-frame sensor motion over the image sequence appears in table 5.11. The precise movements of the cameras include a forward translation of 2.54cm (1.0 inch) and a rotation of 0.00873 radians (0.5 degrees) per frame. Other parameters are approximately known. The pan and tilt angles of the observer axes, relative to the direction of camera motion, are believed to be 0.0174 radians (-1.0 degree) and 0 radians, respectively, at the start of the sequence (time t_0). It is also believed that the center of rotation of the stereo cameras is at the origin of the observer coordinate frame. The apparent pitch velocity (Ω_x), which is caused by the random occurrence of zero or two pixel (vertical) offsets, will be one of three values: $-3.2x10^{-3}$, 0.0, or $3.2x10^{-3}$ radians

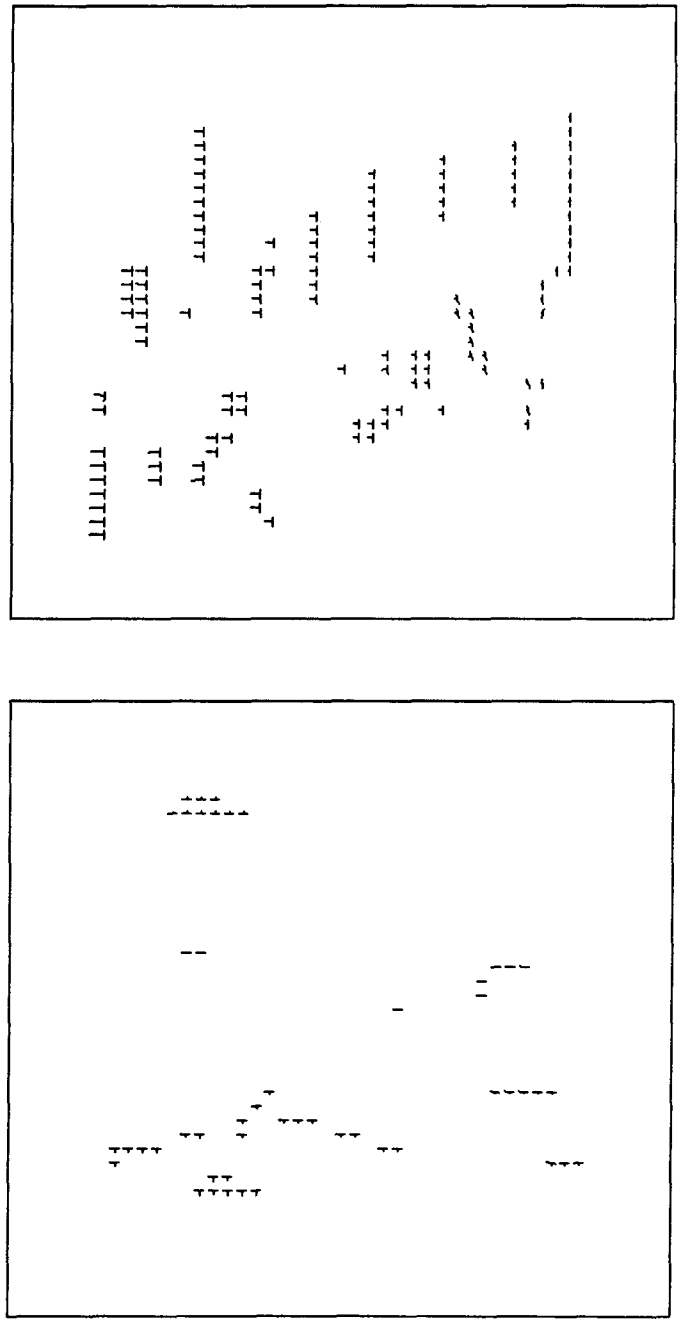


Figure 5.38: Experiment 3, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. *(upper)* Epipolar Channel, $\tilde{\phi}_0 = 0$, *(lower)* Orthogonal Channel, $\tilde{\phi}_2 = \frac{\pi}{2}$ radians.

Table 5.12: Inter-frame Sensor Motion for Experiment 3

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.0577	-0.0562	2.6986	-0.461	8.672	0.042
1-2	0.0178	-0.0329	2.6821	-3.536	8.864	0.161
2-3	0.0018	-0.0877	2.6834	-0.795	8.588	0.182

Table 5.13: Expected Error in Inter-frame Sensor Motion for Experiment 3

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	$\pm$ 0.0648	$\pm$ 0.1165	$\pm$ 0.0357	$\pm$ 1.089	$\pm$ 0.530	$\pm$ 0.621
1-2	$\pm$ 0.0594	$\pm$ 0.1263	$\pm$ 0.0337	$\pm$ 1.199	$\pm$ 0.495	$\pm$ 0.649
2-3	$\pm$ 0.0576	$\pm$ 0.1209	$\pm$ 0.0327	$\pm$ 1.200	$\pm$ 0.490	$\pm$ 0.604

per frame.

The estimated inter-frame sensor motion and expected errors appear in tables 5.12 and 5.13, respectively. The T_z is over-estimated (as expected), but it is consistent throughout the image sequence. The other translational parameters are within the expected error of the actual motion. The Ω_y rotation is within the expected error of the actual value. The largest difference between the measured and actual Ω_y is 0.00014 radians (0.008 degrees) per frame.

Table 5.14: Inter-frame Sensor Motion, Known Ω_z, for Experiment 3

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.0581	-0.0560	2.6987	-0.459	8.669	0.033
1-2	0.0192	-0.0314	2.6822	-3.521	8.852	0.124
2-3	0.0032	-0.0866	2.6835	-0.785	8.576	0.144

The eigenvalues and eigenvectors of the inter-frame Hessian matrix are given by [10]

$$\lambda_0 = 33406 \quad \bar{v}_0 = [\ 0.690 \quad 0.063 \quad -0.021 \quad -0.057 \quad 0.718 \quad -0.002\ ]^T,$$
$$\lambda_1 = 7485.1 \quad \bar{v}_1 = [\ 0.047 \quad -0.736 \quad -0.003 \quad 0.671 \quad 0.072 \quad -0.022\ ]^T,$$
$$\lambda_2 = 901.73 \quad \bar{v}_2 = [\ -0.106 \quad 0.015 \quad 0.983 \quad 0.012 \quad 0.130 \quad -0.068\ ]^T,$$
$$\lambda_3 = 289.44 \quad \bar{v}_3 = [\ 0.368 \quad 0.017 \quad 0.142 \quad 0.058 \quad -0.344 \quad 0.850\ ]^T,$$
$$\lambda_4 = 101.12 \quad \bar{v}_4 = [\ -0.612 \quad 0.025 \quad -0.108 \quad 0.025 \quad 0.586 \quad 0.518\ ]^T,$$
$$\lambda_5 = 33.552 \quad \bar{v}_5 = [\ -0.004 \quad -0.673 \quad 0.022 \quad -0.737 \quad 0.004 \quad 0.063\ ]^T.$$

The condition number is 996. It can be seen that the following parameters are sensitive to measurement errors: T_y and Ω_x; and to a lesser extent, T_x, Ω_y, and Ω_z. The random vertical offsets prevent the use of the known plane constraint (which would increase λ_5). The known Ω_z constraint is chosen in an attempt to improve the inter-frame sensor motion estimate.

The inter-frame sensor motion for the known Ω_z appears in table 5.14. For this experiment, the rotation is assumed to be $\Omega_z = 0.000 \pm 1.000\ 10^{-3}$ radians per frame. The known axial rotation does not alter significantly the inter-frame parameters.

The extended sensor motion appears in table 5.15. The extended sensor translation is consistent with the actual direction of translation throughout the sequence. In the

[10]The rotation terms in the eigenvectors have been normalized by the average scene depth, $z_{norm} = 117$ cm.

Table 5.15: Extended Sensor Motion for Experiment 3

Frame	cm/frame T_x	T_y	T_z	Pred. Error cm/fr ΔT_x	ΔT_y	ΔT_z
0-1	0.0577	-0.0562	2.6986	± 0.0648	± 0.1165	± 0.0357
0-2	0.0253	-0.0456	2.6900	± 0.0419	± 0.0853	± 0.0245
0-3	0.0011	-0.0656	2.6876	± 0.0307	± 0.0695	± 0.0196

Kalman filter's process model, the translation estimate is rotated after each inter-frame motion to account for the change in the orientation of the observer coordinate frame. Without this coordinate change, the direction estimate of the extended sensor motion would lag the actual motion. The final estimate for the direction of translation is (0.0004, −0.0244) radians, or (0.02, −1.40) degrees (actual direction is believed to be (0,0)). The pan directional error is less than the standard established in section 5.2; the tilt error is slightly larger than the one degree standard.

In summary, the inter-frame sensor motion parameters are good despite the sensor rotation and the lens distortion. The good distribution of features in the image suppressed many of the detrimental effects of lens distortion, such as biases in the direction of sensor translation (see appendix C). The direction of sensor translation is consistent with the actual motion, verifying the correct implementation of the process model for the sensor's Kalman filter.

5.4.2 Experiment 4: Moving Object on Collision Trajectory

The purpose of this experiment is to test the ability of the algorithm to segment the image sequence into stationary and moving objects, and to identify an object on a collision trajectory. In this experiment, the sensor translation is along the z_w-axis; the translation of the eco-sub is along the x_w-axis.

Theoretical predictions can be made for this image sequence. The normal image velocity module should perform well. Even though the eco-sub is moving, its normal image velocity is within the velocity bandwidth of a correspondence predictor tuned to stationary objects (see section 4.6.1). The accuracy of the inter-frame sensor motion estimate will be affected by the lens distortion and the feature clustering. Most of the stationary object features for the epipolar channel will be clustered in the left half of the image [11]; thus, T_z will be over-estimated, and T_x will be biased towards the left.

The objectives of this experiment are: to measure the inter-frame and extended sensor motions; to segment the image sequence; to measure the object translation; and to predict the collision parameters. Success of this experiment will verify the correct operation of the segmentation stage and the collision trajectory predictor.

The stereo image velocity measurements for the epipolar channel are shown in figure 5.39. The RMS error in the set of normal image velocity measurements belonging to stationary objects, compared to the set predicted by the inter-frame sensor motion, is 0.17 pixels. The normal image velocity measurements belonging to stationary objects exhibit an axial motion flow pattern. The normal image velocity measurements belonging to the eco-sub are small compared to the neighbouring axial flow vectors; the flow pattern of the eco-sub is typical of an object on a collision trajectory for the case of no sensor rotation.

The actual inter-frame motion includes a forward camera translation of 2.54cm (1.0 inch) and a leftward object (eco-sub) translation of 0.847cm (0.33 inch) per frame. Both the pan and tilt angles of the observer axes relative to the direction of camera motion are believed to be 0 radians throughout the image sequence.

The estimated inter-frame sensor motion and the expected errors appear in tables 5.16

[11] The inter-frame sensor motion is estimated using stationary object features only. Because it is moving, the eco-sub measurements are not included.

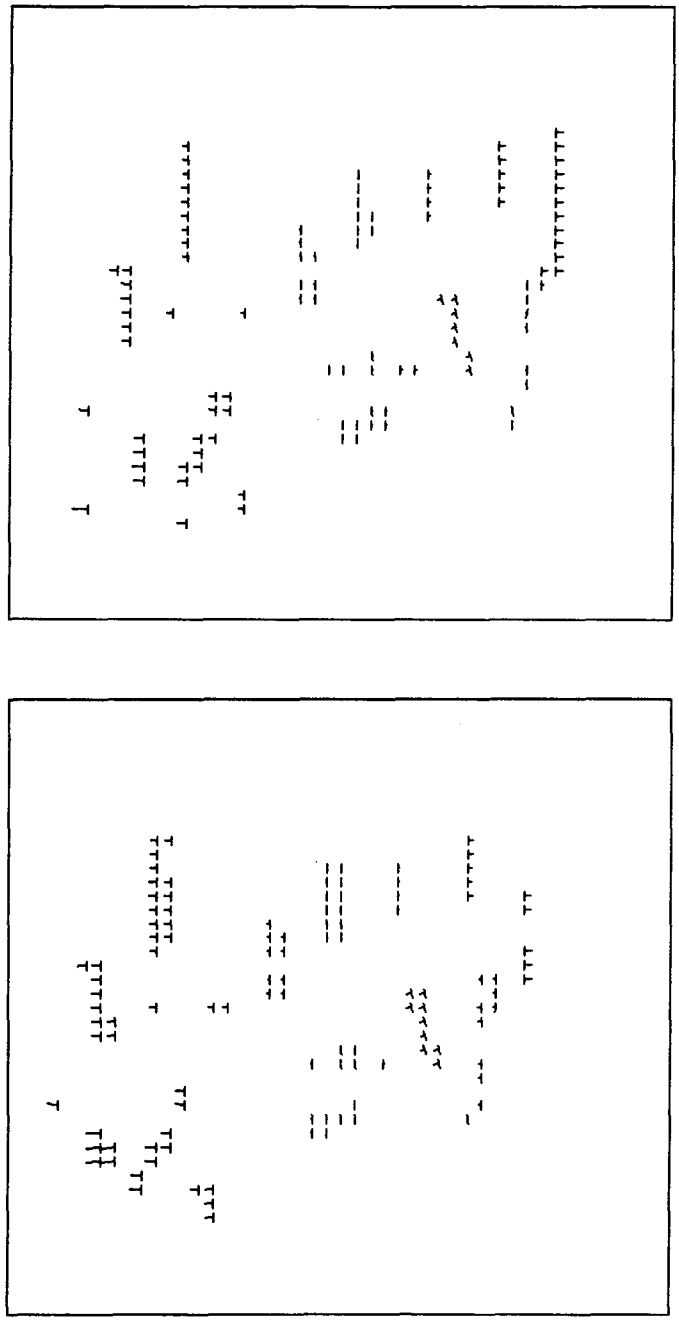


Figure 5.39: Experiment 4, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. The lengths of the vectors have been multiplied by 2.0. *(upper)* Left Epipolar Channel, *(lower)* Right Epipolar Channel.

Table 5.16: Inter-frame Sensor Motion for Experiment 4

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	-0.1545	-0.0436	2.6672	2.797	0.874	0.654
1-2	-0.1392	-0.0749	2.7166	-0.693	0.670	0.565
2-3	-0.1546	-0.0530	2.6816	-0.579	0.881	0.679

Table 5.17: Expected Error in Inter-frame Sensor Motion for Experiment 4

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	± 0.0620	± 0.1215	± 0.0536	± 1.132	± 0.497	± 0.664
1-2	± 0.0627	± 0.1116	± 0.0555	± 1.041	± 0.515	± 0.642
2-3	± 0.0608	± 0.1061	± 0.0506	± 1.026	± 0.506	± 0.641

and 5.17. The lens distortion and the clustering of stationary object features produces a bias in the x component of motion. The x component of sensor motion is given by (for $\Omega_z = 0$)

$$\dot{x} \approx -T_x - z\,\Omega_y. \tag{5.321}$$

The $\dot{x}$ bias is consistent over the sequence; the bias in $\dot{x}$ for a poster feature, whose depth is (127.5, 125.0, 122.5) cm at (t_0, t_1, t_2), is (0.0431, 0.0555, 0.0467) cm per frame. In addition to the $\dot{x}$ bias, the lens distortion-feature clustering alters the distribution of $\dot{x}$ between T_x and Ω_y. The lens distortion also results in an over-estimated T_z.

Table 5.18: Inter-frame Sensor Motion, Known Ω_z, for Experiment 4

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	-0.1492	-0.0453	2.6703	2.786	0.825	0.496
1-2	-0.1350	-0.0749	2.7193	-0.682	0.630	0.436
2-3	-0.1497	-0.0522	2.6841	-0.562	0.835	0.523

The eigenvalues and eigenvectors are given by [12]

$$\lambda_0 = 26357 \quad \bar{v}_0 = [\ 0.682 \quad 0.052 \quad 0.096 \quad -0.050 \quad 0.721 \quad -0.017\ ]^T,$$
$$\lambda_1 = 7394.1 \quad \bar{v}_1 = [\ 0.049 \quad -0.746 \quad -0.051 \quad 0.659 \quad 0.060 \quad 0.009\ ]^T,$$
$$\lambda_2 = 428.81 \quad \bar{v}_2 = [\ -0.020 \quad -0.014 \quad -0.937 \quad -0.094 \quad 0.131 \quad -0.308\ ]^T,$$
$$\lambda_3 = 236.34 \quad \bar{v}_3 = [\ 0.499 \quad -0.033 \quad -0.290 \quad -0.068 \quad -0.419 \quad 0.697\ ]^T,$$
$$\lambda_4 = 103.24 \quad \bar{v}_4 = [\ -0.531 \quad 0.014 \quad -0.126 \quad -0.013 \quad 0.553 \quad 0.647\ ]^T,$$
$$\lambda_5 = 29.918 \quad \bar{v}_5 = [\ 0.037 \quad 0.663 \quad -0.096 \quad 0.741 \quad -0.018 \quad 0.027\ ]^T.$$

The condition number is 881, which is similar to experiment 3.

The inter-frame sensor motion for the known Ω_z constraint appears in table 5.18. The rotation and uncertainty are assumed to be $\Omega_z = 0.000 \pm 1.000\ 10^{-3}$ radians per frame. As in experiment 3, the known axial rotation did not alter significantly the inter-frame parameters. Constraints on T_x, T_y, Ω_x and Ω_y would be more effective.

The extended sensor motion appears in table 5.19. The T_x bias, seen in the estimate of inter-frame sensor motion, also appears in the extended sensor translation. As a result, the directional error is larger than the one degree standard established in section 5.2. The measured direction of translation is (-0.057, -0.021) radians, or (-3.3, -1.2) degrees (actual direction is believed to be (0,0)).

The segmentation of the image sequence is in shown in figure 5.40. The segmentation

[12]The rotation terms in the eigenvectors have been normalized by $z_{norm} = 120$ cm.

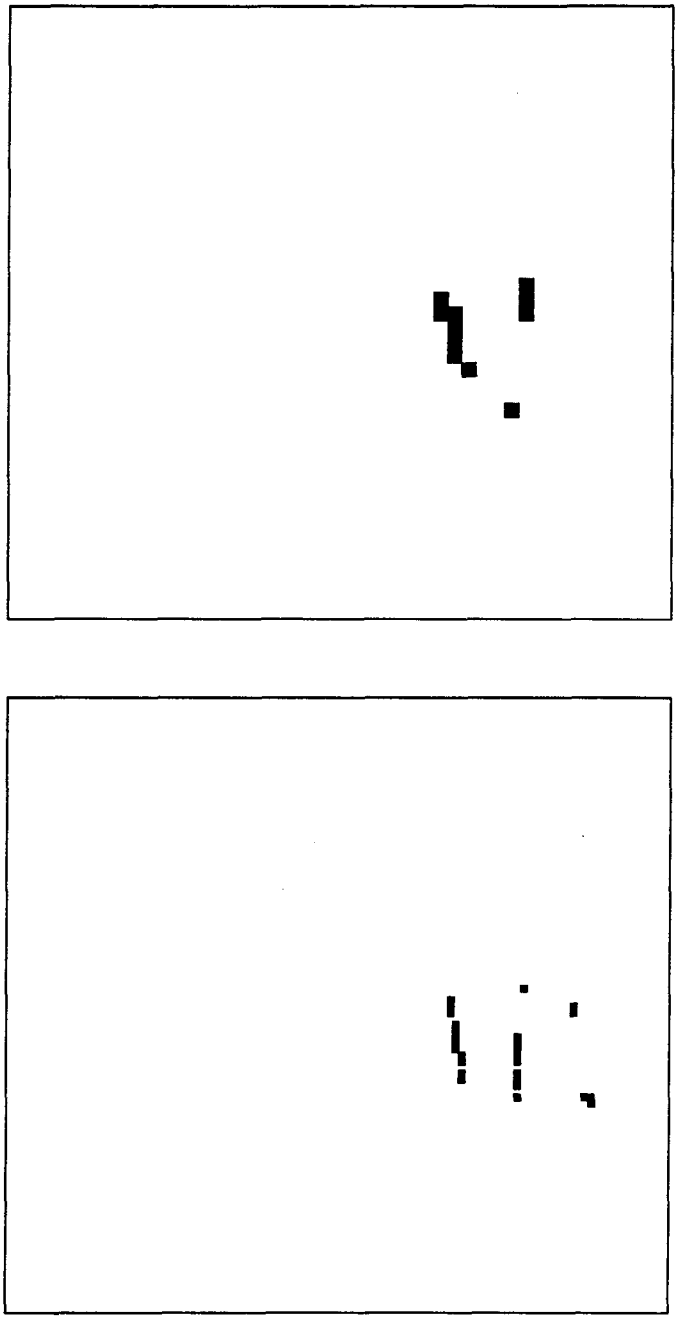


Figure 5.40: Experiment 4, Segmentation of Image Sequence. Stereo features identified as belonging to the eco-sub are denoted by black squares. *(upper)* Epipolar Channel, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel, *(lower)* Epipolar Channel, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel.

Table 5.19: Extended Sensor Motion for Experiment 4

Frame	cm/frame			Pred. Error cm/frame		
	T_x	T_y	T_z	ΔT_x	ΔT_y	ΔT_z
0-1	-0.1546	-0.0436	2.6672	± 0.0620	± 0.1215	± 0.0536
0-2	-0.1509	-0.0566	2.6906	± 0.0416	± 0.0820	± 0.0384
0-3	-0.1532	-0.0564	2.6874	± 0.0321	± 0.0646	± 0.0311

Table 5.20: Extended Object Motion for Experiment 4

Frame	cm/frame		Pred. Error cm/fr	
	T_x	T_z	ΔT_x	ΔT_z
0-1	-0.9458	-0.1680	± 0.0589	± 0.2339
0-2	-0.9550	-0.1239	± 0.0429	± 0.1620
0-3	-0.9494	-0.0768	± 0.0362	± 0.1339

of the eco-sub features from the stationary object features is successful: there are no false positive responses (no stationary features are identified as part of the eco-sub). Most of the features identified in figure 5.40 belong to the arms of the eco-sub.

The extended object motion appears in table 5.20. The $|\dot{x}_{obj}|$ is over-estimated. Since the object motion is estimated using the excess normal image velocity, the bias in the x component of sensor motion affects the object motion. If this bias (which is about 0.07 at the depth of the eco-sub) is removed, $\dot{x}_{obj}$ is approximately -0.87, which is 0.32 of the measured $T_{z,sen}$ (the ratio of the actual object and sensor motion is 0.33). There is a small bias in the $\dot{z}_{obj}$ that decreases as the object moves towards the origin. The cause of this bias is discussed in appendix C.

Each feature on a given object has a different point-of-collision. The arm of the eco-sub that is closet to the driver (the right-most arm in the image) is used as a reference

Table 5.21: Eco-sub Collision Parameters for Experiment 4

	Units: x_{col} cm, t_{col} frames					
	Estimate		Actual		Pred. Error	
Frame	x_{col}	t_{col}	x_{col}	t_{col}	Δx_{col}	Δt_{col}
0	-2.4	35.5	-9	40	± 2.4	± 3.1
1	-3.9	34.8	-9	39	± 1.6	± 2.1
2	-3.9	34.2	-9	38	± 1.3	± 1.8

point. The collision parameters (x_{col} and t_{col}) of the arm appear in table 5.21. Both the point-of-collision and the time-to-collision have been under-estimated. These errors are a result of the biases [13] in T_x and $\dot{z}_{obj}$. Despite the biases, the point-of-collision is accurate enough, relative to the spread of features (see figure 5.40), to identify the eco-sub as an obstacle [14]. The final time-to-collision under-estimates the actual value by about 10 percent.

In summary, the inter-frame and extended sensor motions are affected by the combination of lens distortion and clustering of stationary object features: T_z is over-estimated, and T_x is biased. The object translation is also biased and over-estimated. As a result, the time-to-collision is under-estimated by 10 percent. Despite the distortion, the segmentation of the eco-sub from the stationary background is successful. In addition, the estimated point-of-collision is sufficiently accurate (relative to the size of the eco-sub) to identify the eco-sub as an obstacle.

[13]The bias in T_x is partially compensated by the bias in the object velocity $\dot{x}_{obj}$. A residual of -0.08 cm per frame (-0.15 + 0.07) remains. This residual decreases the $|x_{col}|$ by about 3 cm. The bias in $\dot{z}_{obj}$ decreases $|x_{col}|$ further by about 1.5 cm.

[14]The spread of features is related to the body length of the eco-sub, which is 20 cm. A collision is predicted if x_{col} for the reference arm is between 0 and -20 cm.

Table 5.22: Inter-frame Sensor Motion for Experiment 5

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	-0.1477	-0.0706	2.6810	2.665	0.757	0.476
1-2	-0.1568	-0.0766	2.7241	-3.943	0.858	0.679
2-3	-0.1284	-0.1013	2.6971	-0.832	0.632	0.263

5.4.3 Experiment 5: Moving Object on Pass-by Trajectory

The purpose of this experiment is to test the ability of the algorithm to identify an object that will pass in front of the cameras. This experiment is similar to experiment 4 except that the speed of the eco-sub has been doubled.

The stereo image velocity measurements for the epipolar channel are shown in figure 5.41. The RMS error in the set of normal image velocity measurements belonging to stationary objects, compared to the set predicted by the inter-frame sensor motion, is 0.18 pixels. The normal image velocity measurements belonging to the eco-sub have a leftward direction in both the left and right images. This component flow pattern is characteristic of an object that will pass in front of the cameras for the case of no sensor rotation.

The actual inter-frame motion includes a forward camera translation of 2.54cm (1.0 inch) and a leftward object (eco-sub) translation of 1.69cm (0.67 inch) per frame. Both the pan and tilt of the observer axes are believed to be 0 radians throughout the image sequence.

The inter-frame sensor motion and expected errors appear in tables 5.22 and 5.23. As in experiment 4, $\dot{x}$ is biased and T_z is over-estimated. The $\dot{x}$ bias is consistent over the sequence; for a poster feature, whose depth is (127.5, 125.0, 122.5) cm at (t_0, t_1, t_2),

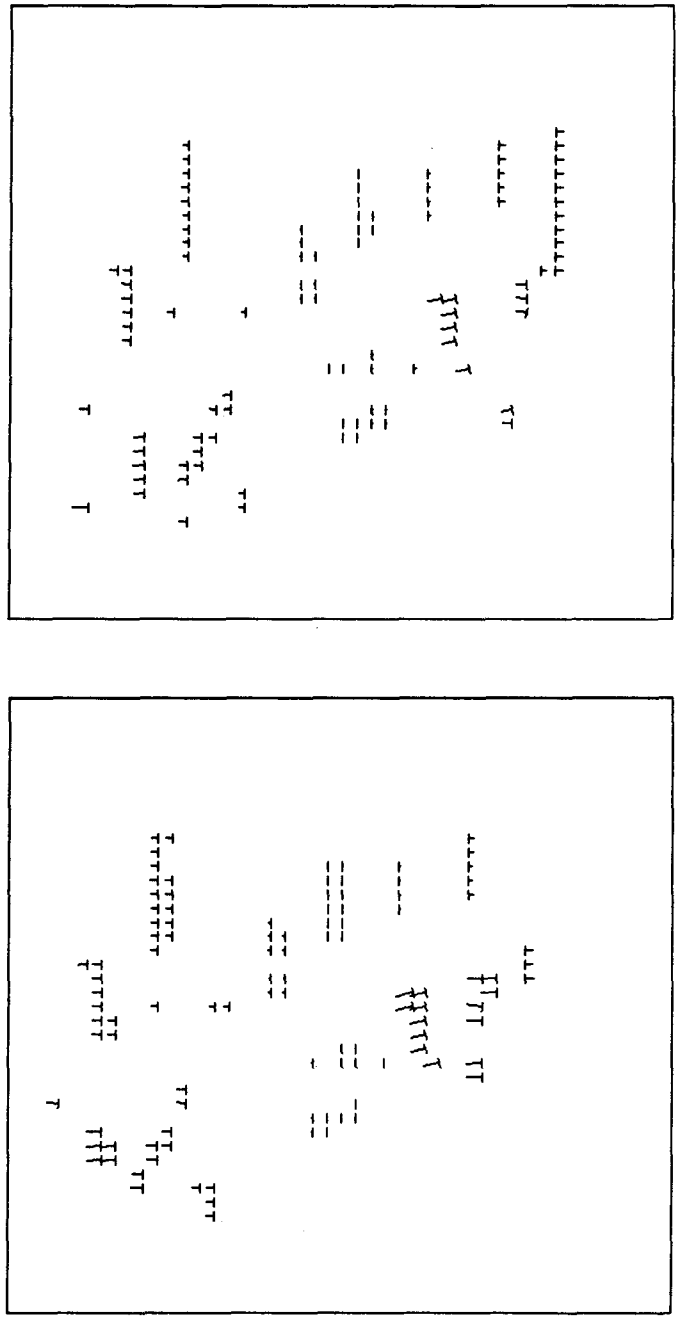


Figure 5.41: Experiment 5, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. The lengths of the vectors have been multiplied by 2.0. *(upper)* Left Epipolar Channel, *(lower)* Right Epipolar Channel.

Table 5.23: Expected Error in Inter-frame Sensor Motion for Experiment 5

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	± 0.0625	± 0.1227	± 0.0550	± 1.138	± 0.503	± 0.715
1-2	± 0.0609	± 0.1173	± 0.0567	± 1.097	± 0.501	± 0.680
2-3	± 0.0628	± 0.1129	± 0.0529	± 1.080	± 0.521	± 0.676

Table 5.24: Extended Sensor Motion for Experiment 5

	cm/frame			Pred. Error cm/fr		
Frame	T_x	T_y	T_z	ΔT_x	ΔT_y	ΔT_z
0-1	-0.1477	-0.0706	2.6810	± 0.0625	± 0.1227	± 0.0550
0-2	-0.1534	-0.0700	2.7018	± 0.0413	± 0.0844	± 0.0395
0-3	-0.1489	-0.0874	2.6994	± 0.0319	± 0.0674	± 0.0316

the $\dot{x}$ bias is (0.0512, 0.0496, 0.0510) cm per frame.

The extended sensor motion appears in table 5.24. The T_x bias seen in the estimate of inter-frame sensor motion also appears in the extended sensor translation. As a result, the directional error is larger than the one degree standard (see section 5.2). The measured direction of translation is (-0.055, -0.032) radians, or (-3.2, -1.9) degrees (actual direction is believed to be (0,0)).

The segmentation of the image sequence is in shown in figure 5.42. As in experiment 4, the segmentation of the eco-sub features from the stationary object features is successful: there are no false positive responses. Most of the features identified in figure 5.42 belong to the arms of the eco-sub.

The extended object motion appears in table 5.25. The object velocity is approximately double the speed of the eco-sub in experiment 4 (as expected). If $\dot{x}$ bias (which is

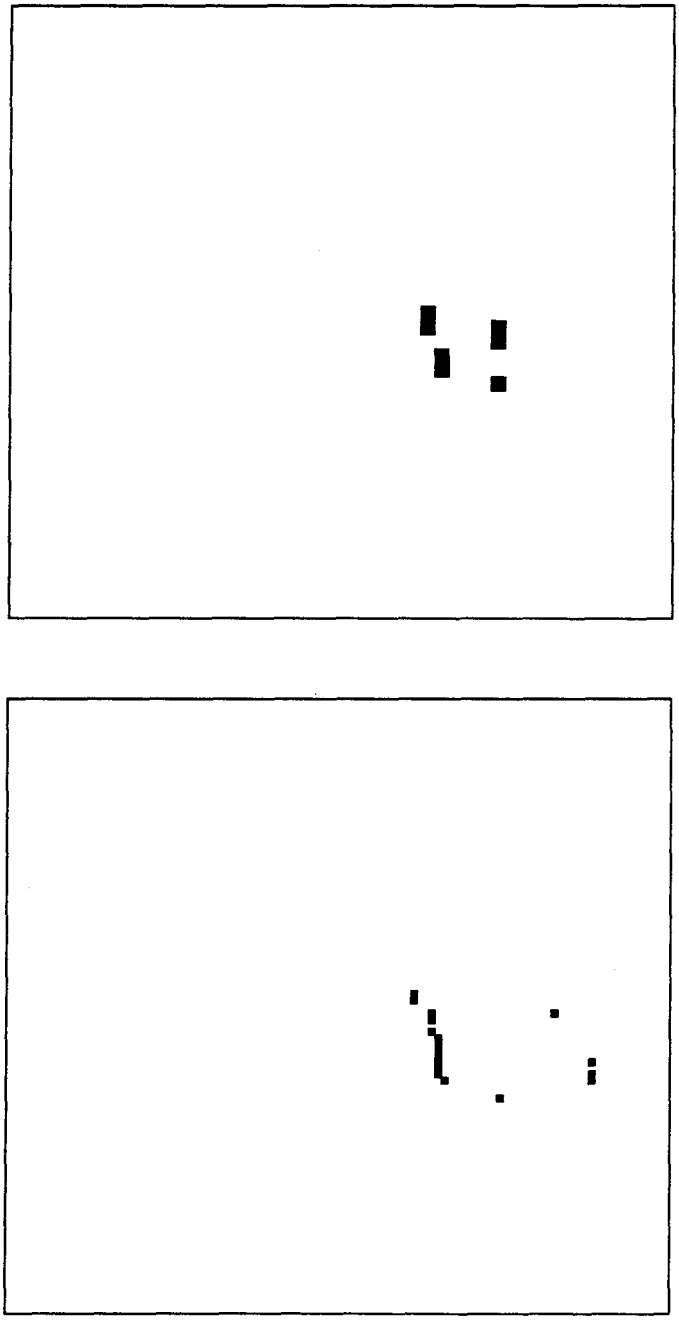


Figure 5.42: Experiment 5, Segmentation of Image Sequence. Stereo features identified as belonging to the eco-sub are denoted by black squares. *(upper)* Epipolar Channel, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel, *(lower)* Epipolar Channel, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel.

Table 5.25: Extended Object Motion for Experiment 5

Frame	cm/frame		Pred. Error cm/fr	
	T_x	T_z	ΔT_x	ΔT_z
0-1	-1.8977	-0.3223	± 0.0730	± 0.3438
0-2	-1.8568	-0.0332	± 0.0468	± 0.1841
0-3	-1.8787	-0.1144	± 0.0392	± 0.1570

Table 5.26: Eco-sub Collision Parameters for Experiment 5

Frame	Units: x_{col} cm, t_{col} frames					
	Estimate		Actual		Pred. Error	
	x_{col}	t_{col}	x_{col}	t_{col}	Δx_{col}	Δt_{col}
0	-44.3	34.7	-50	43	± 5.4	± 4.1
1	-48.1	37.2	-50	42	± 3.4	± 2.6
2	-47.8	34.9	-50	41	± 3.7	± 2.8

about 0.07 at the depth of the eco-sub) is removed, $\dot{x}_{obj}$ is approximately -1.81, which is 0.67 of the measured $T_{z,sen}$ (the ratio of the actual object and sensor motion is 0.67). As in experiment 4, there is a bias in $\dot{z}_{obj}$.

The forward arm of the eco-sub is used as a reference point. The collision parameters of the arm appears in table 5.26. As in experiment 4, both the point-of-collision and the time-to-collision have been under-estimated. The final time-to-collision under-estimates the actual value by 15 percent. The point-of-collision correctly identifies the eco-sub as an object that will pass safely in front of the cameras.

In summary, the inter-frame and extended sensor motions are affected by the combination of lens distortion and the clustering of stationary object features: T_z is over-estimated, and T_x is biased. The object translation is also biased and over-estimated.

As a result, the time-to-collision is under-estimated by 15 percent. Despite the distortion, the segmentation of the eco-sub from the stationary background is successful. The estimated point-of-collision is sufficiently accurate (relative to the size of the eco-sub) to identify the eco-sub as a pass-by object.

5.4.4 Summary

Lens distortion (or any other non-scalar error) has a detrimental effect on the estimated depth, motion, and collision parameters. The planar background appears parabolic in the local map. The T_z is over-estimated in each image sequence. In experiments 4 and 5, the combination of feature clustering and lens distortion causes biases in T_x. In addition, the magnitude of the collision parameters are under-estimated.

The algorithm performed fairly well in spite of the lens distortion. In experiment 3, the sensor rotation and the direction of sensor translation are within the expected errors. Because the image has a good distribution of stationary object features, motion biases did not appear. In experiments 4 and 5, the segmentation of the image sequence is successful; and the collision parameters are sufficiently accurate, relative to the size of the eco-sub, to determine if the eco-sub is an obstacle (as in experiment 4) or a pass-by object (as in experiment 5).

5.5 Data Set 3

In this final set of image sequences, we leave the precise environment of the optical bench. In the first of three experiments, an outdoor scene is analyzed. The second experiment measures the motion of tripod-mounted cameras that are moving through a stationary indoor environment. The last experiment estimates the collision parameters of two manually moved objects viewed from stationary cameras.

5.5.1 Experiment 6: Outdoor Scene

In this experiment, only one stereo pair is available [15]; it is an outdoor scene shown in figure 5.43. The ground surface is flat in the foreground, and hills are present in the background. The foreground contains stationary objects: rocks, dirt piles, and trees. Beyond the rocks and dirt piles, the shadows reduce the contrast of viewed objects.

The nominal camera parameters are as follows: the focal length of each camera is 16mm; the CCD array is 6.6mm by 8.8mm; and the image size is 240 x 256 pixels. The stereo baseline is 25cm. The cameras are convergent, requiring an 11 pixel offset along the $\hat{x}$-axis to approximate a parallel stereo setup (see section 2.5). This offset is estimated by manually measuring the disparity of features along the horizon [16].

Theoretical predictions can be made regarding the performance of the disparity module. The E_{offset} histogram and the multiscale prediction will detect large objects in the foreground, such as the big rock and the tree. The flat ground produces a large depth gradient in the image along the $\hat{y}$-axis, causing the disparity module to rely heavily on the heuristic ordering constraint. The horizon will be difficult to match because it is predominantly horizontal; the preferred spectral orientation is orthogonal to the epipolar channel. The shadow covering the background hides objects. As a result, the number of disparity measurements in background will be small.

The objective of this experiment is to measure the scene structure and to make a qualitative comparison with the presumed structure. Success of this experiment will verify the disparity module's robustness to outdoor conditions which include large depth gradients, and lighting phenomena such as shadows.

The interpolated disparity and its uncertainty are shown in figures 5.44 and 5.45. The black regions in the disparity map indicate that the uncertainty is too large for a

[15] This stereo image was supplied by Larry Matthies.

[16] It is assumed that the parallel stereo disparity of a distant horizon point is approximately zero.

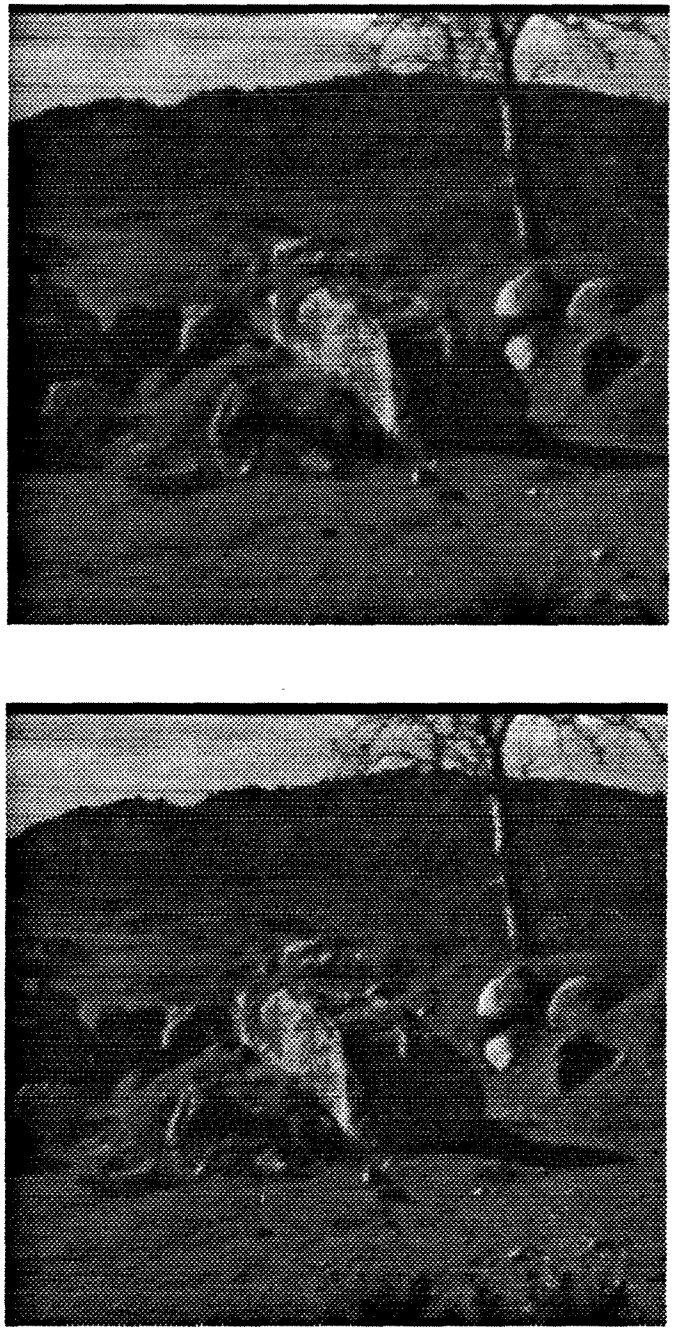


Figure 5.43: Experiment 6, Stereo Images *(upper)* Left Image, *(lower)* Right Image.

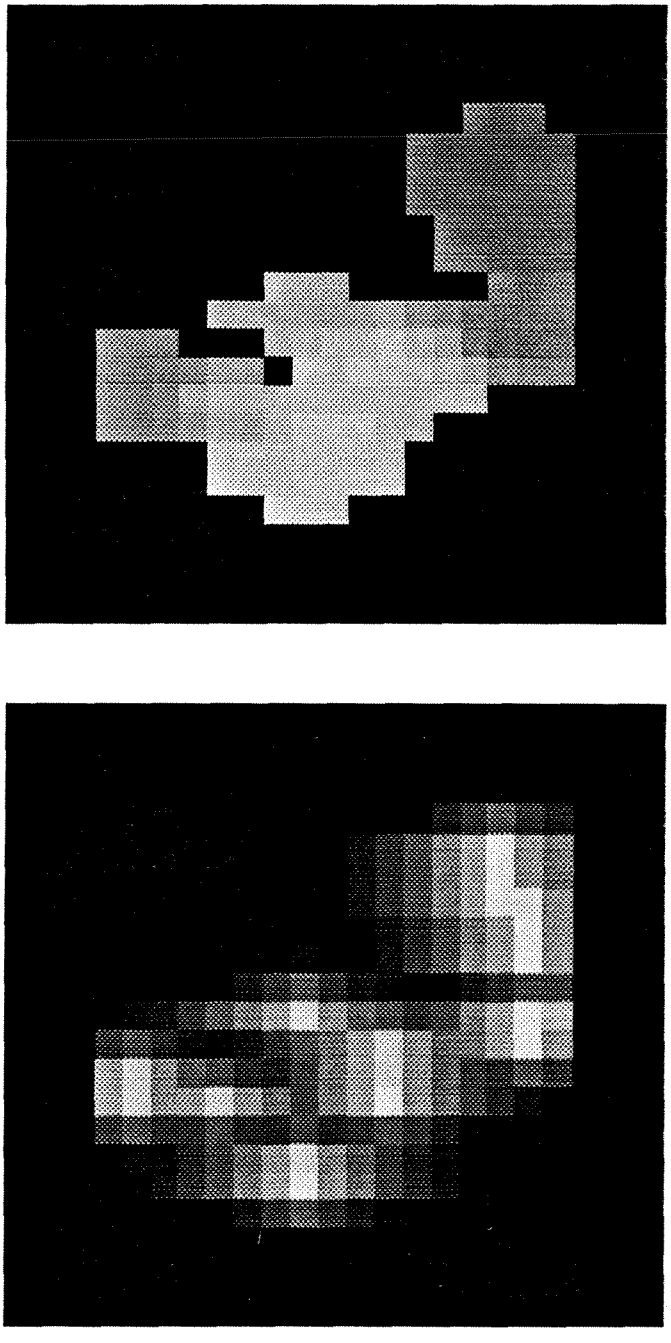


Figure 5.44: Experiment 6, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are -15 pixels and 35 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

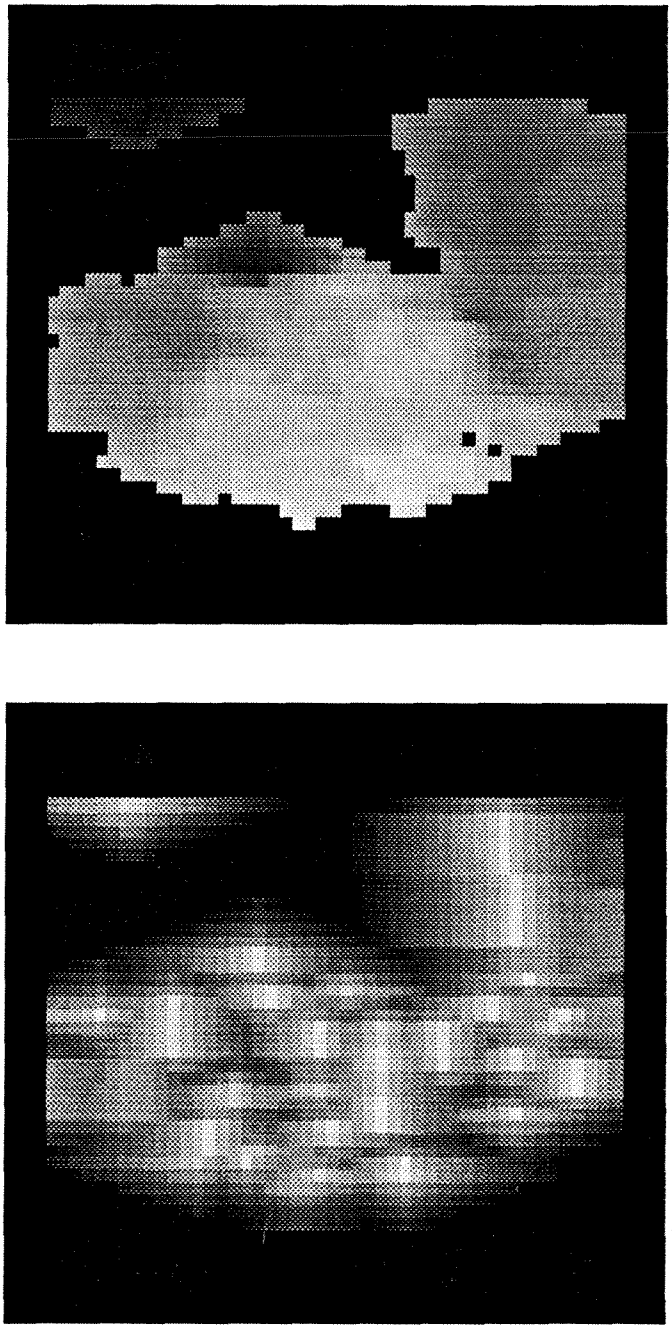


Figure 5.45: Experiment 6, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are -15 pixels and 35 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

meanful estimate of disparity. These regions include the shadowed background points.

Two local maps are shown in figures 5.46 and 5.47. Figure 5.46 displays most of the stereo features with the exception of four background features. Figure 5.47 is a more detailed map of the foreground features. The relative position of objects in the scene can be compared with the presumed structure. The presumed ordering of objects, from the most distant features to the foreground features, is as follows: the crest of the hill, the shadowed background, the tree, and the rock pile. The measured depths of these features have the correct order.

Experiment 6 has measured the disparity and depth for an outdoor scene. Although the disparity module did not detect many shadowed background features, it did not make any foolish matches. If an image sequence was available, the number of detected stereo features and the variation in depth would have been sufficient to accurately estimate the inter-frame sensor motion.

5.5.2 Experiment 7: Camera Motion with Transients

In this experiment, stereo cameras move in a stationary environment: namely, a graduate student office. A stereo pair from the image sequence is shown in figure 5.48. The office contains tables, chairs, bookshelves, boxes, beverage cans, and stacks of papers. The background is a uniform coloured wall with some posters. The image projection of the office contains uni-directional features from various orientations, but primarily features whose normal direction is either horizontal or vertical.

The image sequence is obtained by moving stereo cameras, which are mounted to a tripod, at approximately 10 cm per frame in a forward direction (believed to be the z-axis). Tripod flex introduces transient rotations into the sequence. The optical axes of the stereo cameras are believed to parallel. The stereo baseline is 11cm. The nominal camera parameters are as follows: the focal length of each camera is 16mm; the CCD

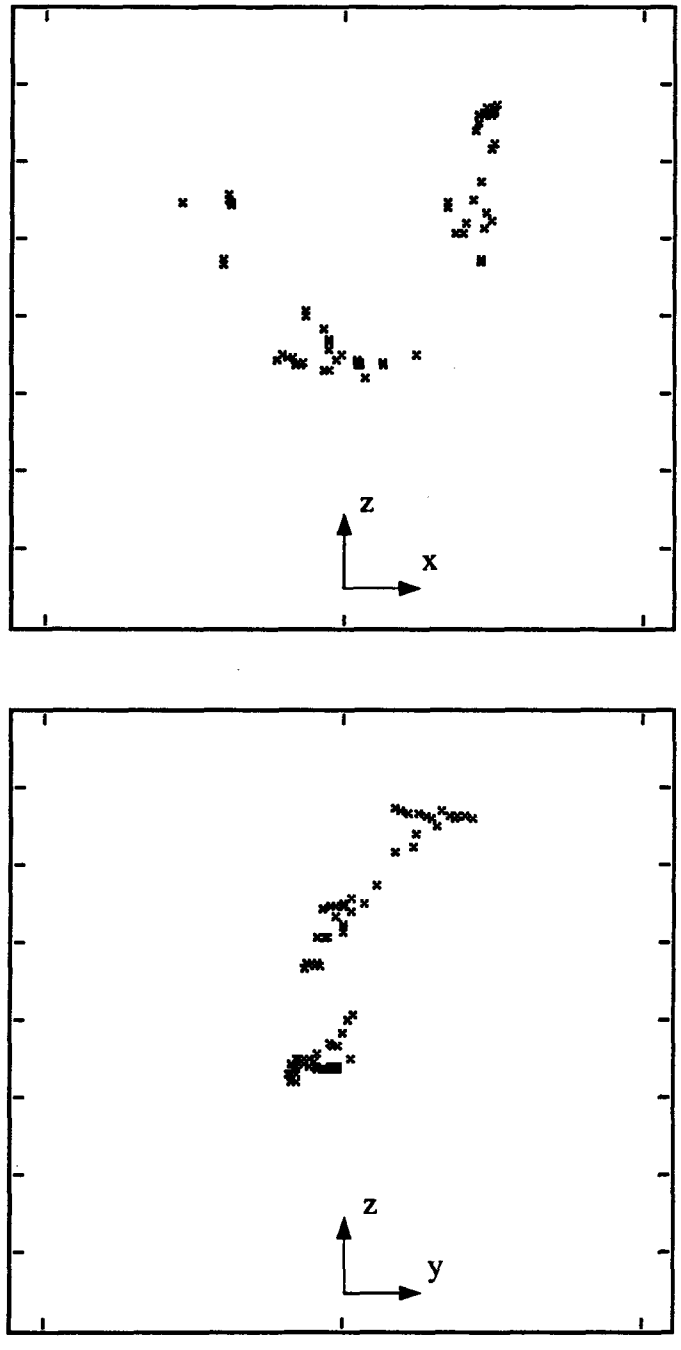


Figure 5.46: Experiment 6, Local Map. Stereo features are denoted by black "X"s. Distance between ticks is 250 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

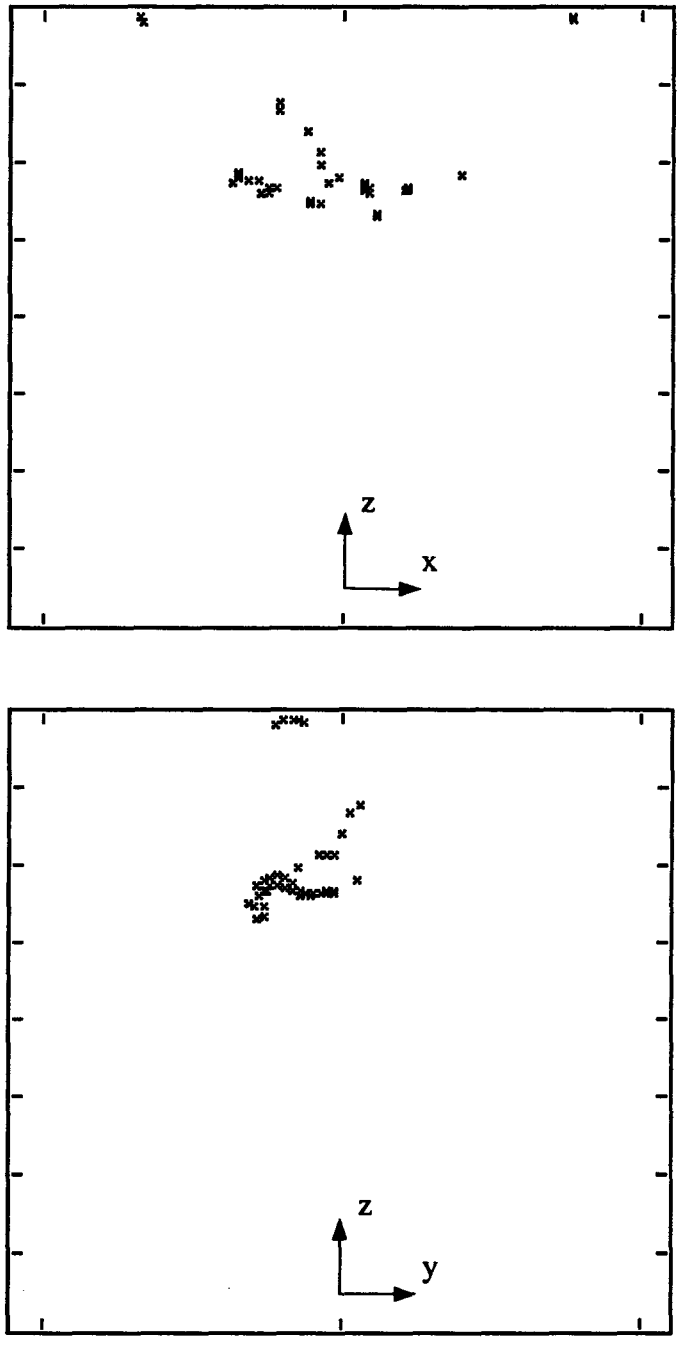


Figure 5.47: Experiment 6, Local Map of Foreground. Stereo features are denoted by black "X"s. Distance between ticks is 150 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

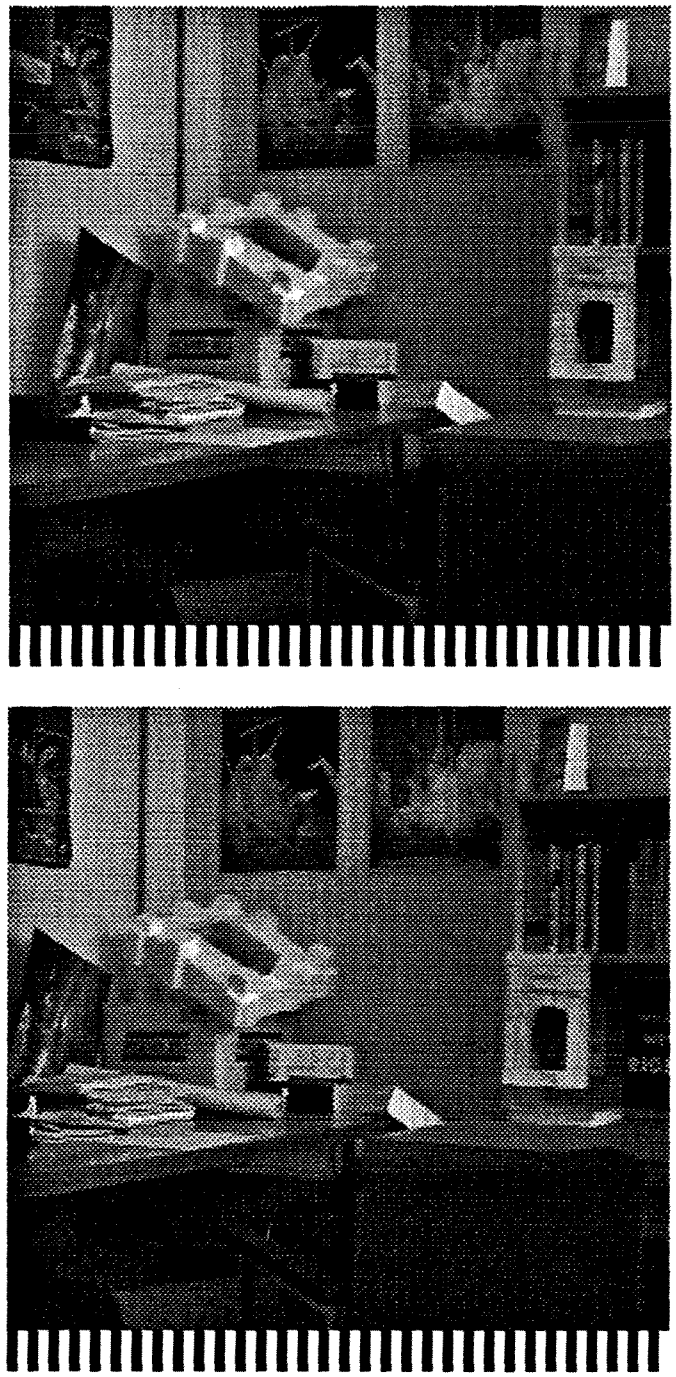


Figure 5.48: Experiment 7, Stereo Images *(upper)* Left Image, *(lower)* Right Image.

array is 6.6mm by 8.8mm; and the image size is 480 x 512 pixels.

The disparity module will work well because there are no large disparity gradients. Most of the stereo correspondences will be made by the E_{offset} histogram and the multi-scale prediction. The heuristic ordering constraint and temporal constraint will provide some additional matches. There will be large regions in the image with no epipolar features, and therefore, no direct disparity measurements. These include image regions with no spectral energy, such as the uniform coloured wall, and the shadowed region below the desk.

The normal image velocity module will be challenged by this sequence. Large axial translation and transient rotation result in non-zero lattice offsets between (temporal) corresponding points. The generation of correct correspondences in higher frequency channels will depend heavily on the accuracy in which the low frequency motion estimate can predict the image velocity field.

The inter-frame Hessian may be ill-conditioned because of the poor distribution of features in the image. Almost all of the features are found in the top part of the image. As a result, the axial rotation Ω_z will produce a similar image velocity field to those produced by T_x and Ω_y. The known Ω_z constraint may improve the estimate of inter-frame sensor motion.

The objectives of this experiment are: to measure the inter-frame sensor motion; to determine if the known Ω_z constraint improves the inter-frame motion estimate; and to measure the extended sensor motion. Success of this experiment will verify that the sensor motion module is insensitive to transient rotations induced by camera shake.

The interpolated disparity and its uncertainty are shown in figures 5.49, 5.50, and 5.51. The uncertainty maps display large regions of uncertainty: the shadowed area beneath the table, and the uniform coloured portions of the wall.

The top and side views of the local map are shown in figure 5.52. The local map

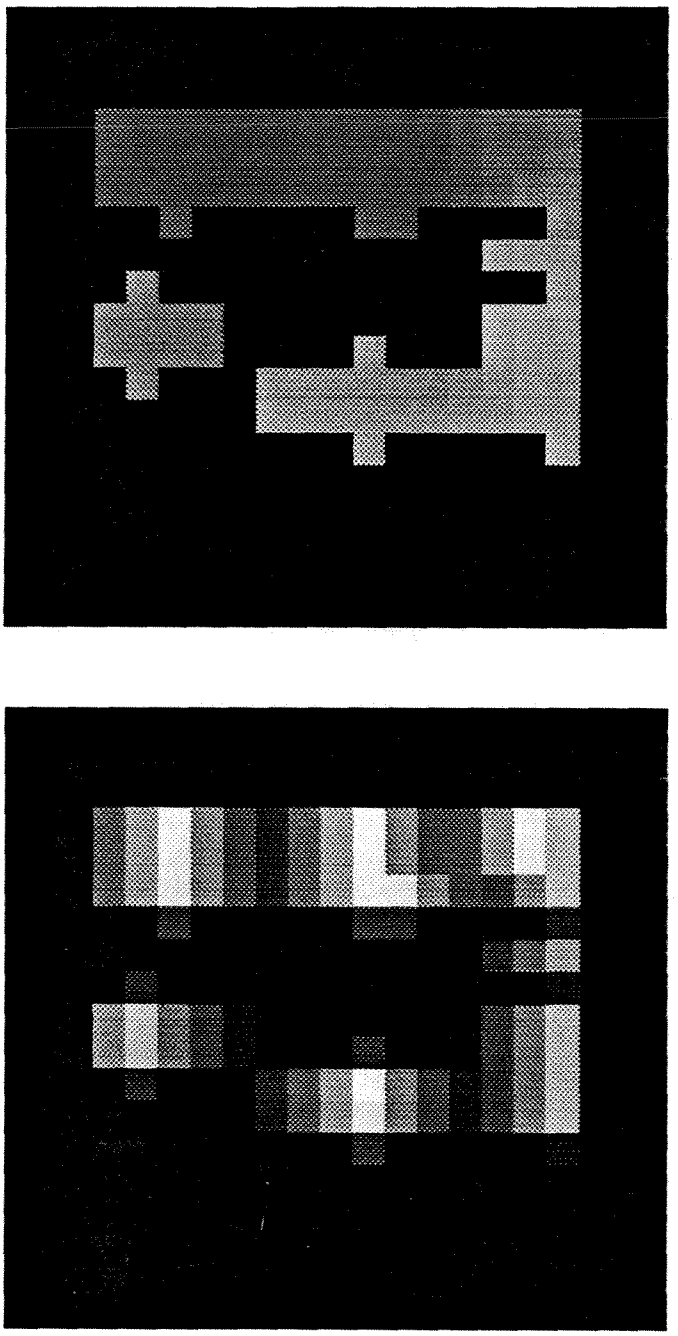


Figure 5.49: Experiment 7, $\tilde{\omega}_0 = 0.040\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 18 pixels and 39 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

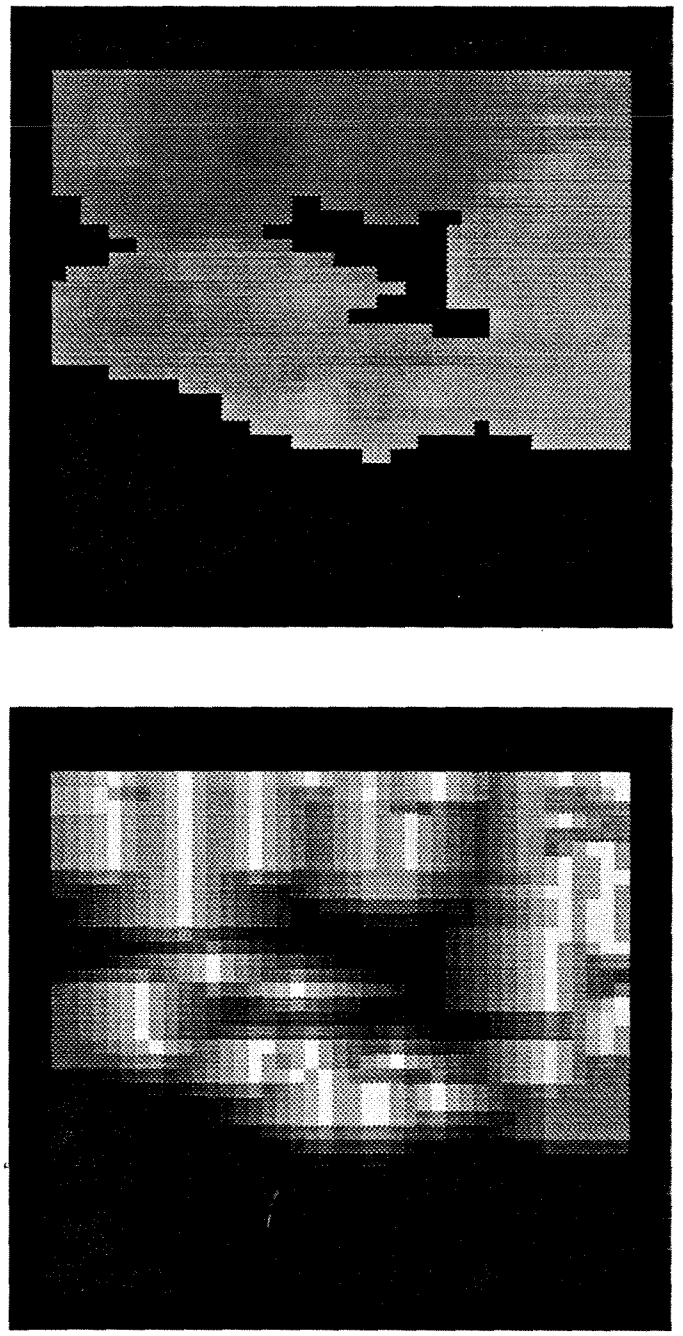


Figure 5.50: Experiment 7, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 18 pixels and 39 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

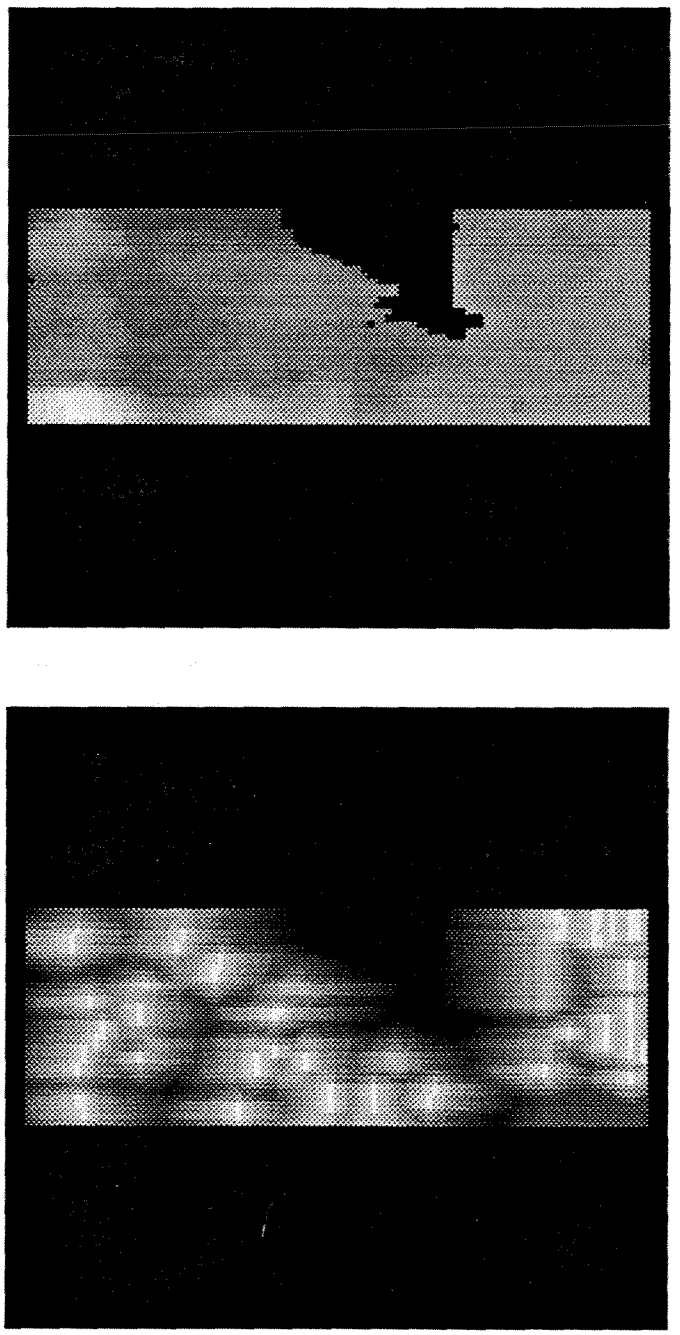


Figure 5.51: Experiment 7, *(upper)* Interpolated Disparity and *(lower)* Uncertainty for $\tilde{\omega}_2 = 0.210\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 18 pixels and 39 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

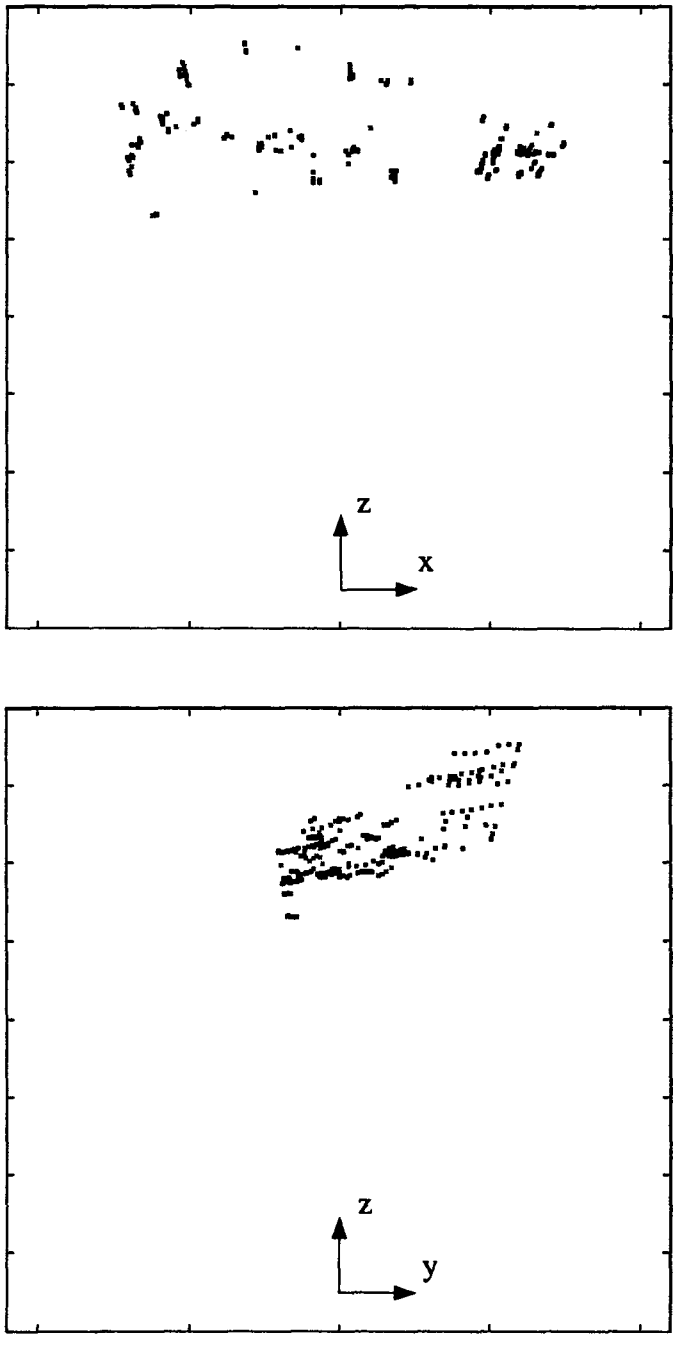


Figure 5.52: Experiment 7, Local Map. Stereo features are denoted by black squares. Distance between ticks is 50 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

Table 5.27: Inter-frame Sensor Motion for Experiment 7

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.1816	0.0794	9.513	3.649	10.265	0.078
1-2	-0.0068	0.0060	9.421	-0.166	-1.581	1.140
2-3	-0.2750	-0.0252	9.627	1.436	1.085	1.710

captures the scene structure: a corner of a room with assorted office materials. The distant features roughly define the wall. Note that the wall features, which are vertically aligned, appear slanted in the local map. This slant suggests that one of the stereo cameras is mounted with a non-zero roll angle.

The normal image velocity measurements for the epipolar and orthogonal channels, during the first inter-frame transition (from t_0 to t_1), are shown in figure 5.53. The normal image velocity measurements are consistent with an axial flow pattern offset by -10 pixels in the $\hat{x}$ direction. The nearly constant offset is induced by a transient sensor rotation about the y-axis. The RMS error in the measured normal image velocity field, compared to the field predicted by the inter-frame sensor motion, is 0.17 pixels, which is better than the one pixel standard described in section 5.2.

The inter-frame sensor motion and the expected errors appear in tables 5.27 and 5.28, respectively. The inter-frame sensor motion estimates are good despite transient rotations. The direction of translation is consistent with axial translation (within the expected errors). The inter-frame frame estimates of T_z are consistent with the final estimate of the extended sensor translation ($T_{z,sen} = 9.51$).

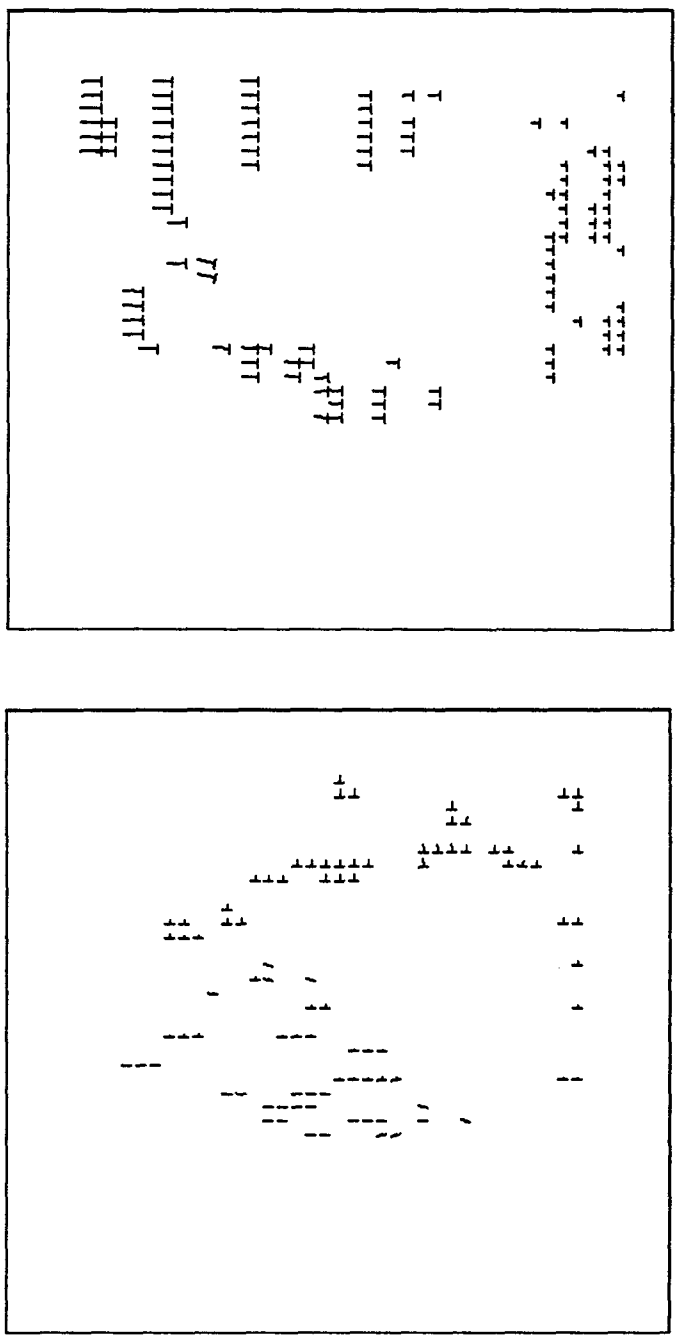


Figure 5.53: Experiment 7, Normal Image Velocity for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. *(upper)* Epipolar Channel, $\tilde{\phi}_0 = 0$, *(lower)* $\tilde{\phi}_2 = \frac{\pi}{2}$ radians.

Table 5.28: Expected Error in Inter-frame Sensor Motion for Experiment 7

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	± 0.3313	± 0.4109	± 0.0994	± 1.343	± 1.044	± 0.799
1-2	± 0.3038	± 0.3051	± 0.1019	± 1.019	± 0.985	± 0.712
2-3	± 0.3170	± 0.3419	± 0.1222	± 1.196	± 1.065	± 0.794

The eigenvalues and eigenvectors are given by [17]

$$\lambda_0 = 9894.2 \quad \bar{v}_0 = [\ 0.699 \ -0.005 \ -0.022 \ 0.006 \ 0.714 \ -0.016\ ]^T,$$
$$\lambda_1 = 5649.1 \quad \bar{v}_1 = [\ -0.006 \ -0.711 \ -0.014 \ 0.703 \ -0.005 \ 0.016\ ]^T,$$
$$\lambda_2 = 142.89 \quad \bar{v}_2 = [\ -0.082 \ 0.034 \ 0.988 \ 0.054 \ 0.111 \ 0.003\ ]^T,$$
$$\lambda_3 = 34.999 \quad \bar{v}_3 = [\ -0.234 \ -0.055 \ -0.036 \ -0.035 \ 0.207 \ -0.947\ ]^T,$$
$$\lambda_4 = 5.3994 \quad \bar{v}_4 = [\ -0.314 \ 0.626 \ -0.116 \ 0.628 \ 0.305 \ 0.089\ ]^T,$$
$$\lambda_5 = 4.8512 \quad \bar{v}_5 = [\ -0.593 \ -0.313 \ -0.087 \ -0.327 \ 0.585 \ 0.308\ ]^T.$$

The condition number is 2040. It can be seen, from the eigenvectors associated with λ_3, λ_4, and λ_5, that all the parameters except T_z are sensitive to measurement errors.

The known Ω_z constraint is applied to improve the inter-frame estimates [18]. It is assumed that the axial rotation Ω_z is zero with a standard deviation of $\pm 1.000\ 10^{-3}$ radians per frame. The inter-frame sensor motion for the known axial rotation can be found in table 5.29. The known Ω_z constraint appears to improve the parameter estimates for the inter-frame transition from t_2 to t_3.

The extended sensor motion appears in table 5.30. The temporal variations in the inter-frame translation are smoothed by the integral nature of the extended sensor motion

[17] The rotation terms in the eigenvectors have been normalized by the average scene depth, $z_{norm} = 303$ cm.

[18] Constraints on T_x, T_y, Ω_x, or Ω_y would also be useful. Unfortunately, the tripod flex makes most of the motion constraints invalid. The known Ω_z is itself a bold assumption.

Table 5.29: Inter-frame Sensor Motion, Known Axial Rotation, for Experiment 7

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.1924	0.0810	9.5149	3.655	10.230	-0.030
1-2	0.1457	0.0397	9.4481	-3.610	-2.086	0.496
2-3	-0.0184	0.0047	9.6837	1.568	0.212	0.654

Table 5.30: Extended Sensor Motion for Experiment 7

	cm/frame			Pred. Error cm/fr		
Frame	T_x	T_y	T_z	ΔT_x	ΔT_y	ΔT_z
0-1	0.1816	0.0794	9.5131	± 0.3313	± 0.4109	± 0.0994
0-2	0.0499	0.0405	9.4710	± 0.2179	± 0.2430	± 0.0726
0-3	-0.0653	0.0137	9.5142	± 0.1755	± 0.1949	± 0.0633

module. The extended sensor motion is consistent with the axial sensor translation. The (final) measured direction of translation is (-0.0069,0.0014) radians, or (-0.39,0.08) degrees, from the z-axis. The impressive directional accuracy (the error is less than half of the standard established in section 5.2) is partial due to the large axial translation.

In summary, the inter-frame sensor motion estimates displayed temporal variations, but they are within the expected errors. The known Ω_z constraint partially stabilized these variations. The extended sensor motion smoothed the variations, producing parameter estimates that are consistent with axial translation. Experiment 7 has shown that the sensor motion module is insensitive to transient sensor rotations.

5.5.3 Experiment 8: Multiple Moving Objects

In this experiment, two moving objects, in an otherwise stationary environment, are being viewed by stationary stereo cameras. Stereo pairs from the beginning (t_0) and end (t_4) of the image sequence are shown in figures 5.54 and 5.55, respectively. The two moving objects in the foreground are beverage cans on top of stools. The two moving objects can be easily distinguished because the attached cans are competing cola brands: the left object will be referred to as the "C-cola stool" and the right object will be referred to as the "P-cola stool." The stationary background consists of a large bookshelf filled with assorted manuals and equipment. A stationary chair appears in the foreground at the right periphery of the stereo images. The scene structure is very complex, containing large depth gradients and viewpoint sensitive (unstable) alignments of foreground and background features. The image projection of the background contains many uni-directional features that have horizontal or vertical normal directions.

The image sequence is produced by manually moving objects which are in the field of view of stationary cameras. The C-cola stool is heading towards the cameras along a collision trajectory at 10 cm per frame. The P-cola stool is heading towards the cameras at 20 cm per frame, but the stool will pass safely in front of the cameras.

The stereo cameras are divergent and are differently tilted, requiring a -12 pixel offset along the $\hat{x}$-axis and -10 pixel offset along the $\hat{y}$-axis to approximate a parallel stereo setup. The right camera has a 0.012 radian roll angle which is compensated by rotating the right image. The stereo baseline is 10.2 cm.

The nominal camera parameters are as follows: the focal length of each camera is 16mm; the CCD array is 6.6mm by 8.8mm; and the image size is 480 x 512.

The theory outlined in the previous chapters predict that this image sequence will be

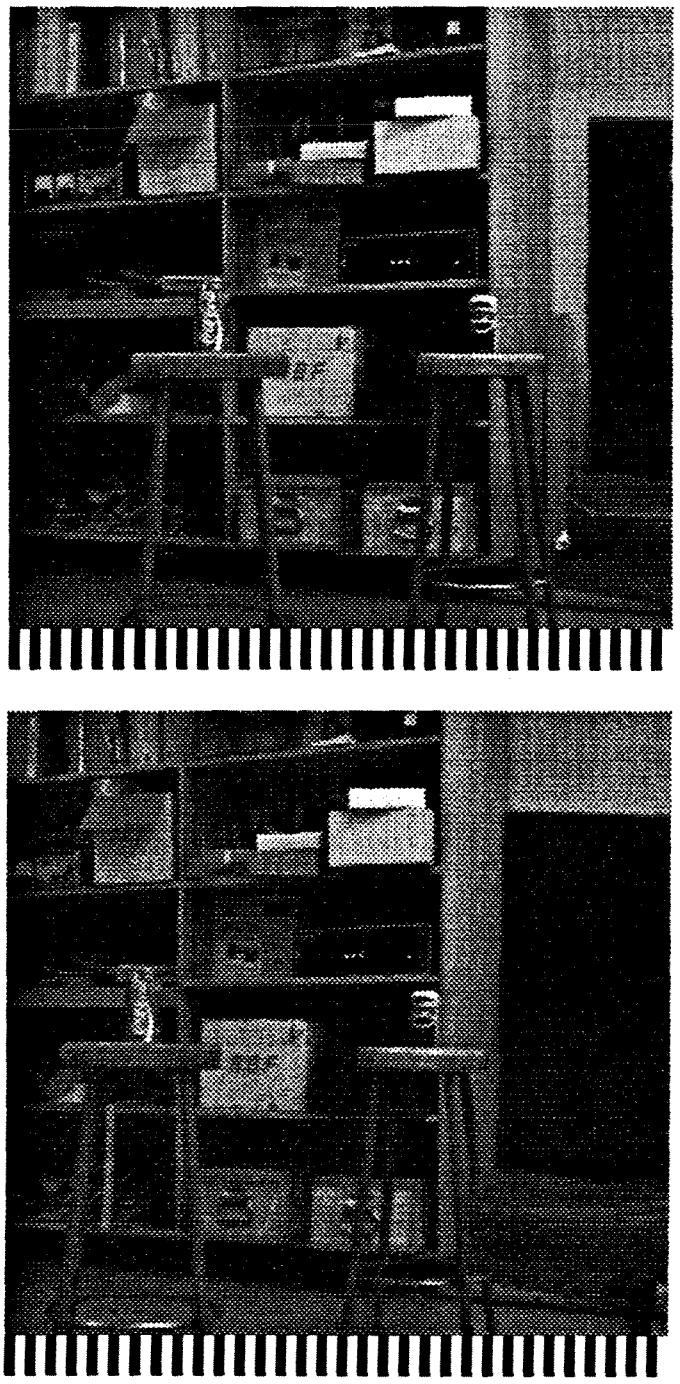


Figure 5.54: Experiment 8, Stereo Images at t_0, *(upper)* Left Image, *(lower)* Right Image.

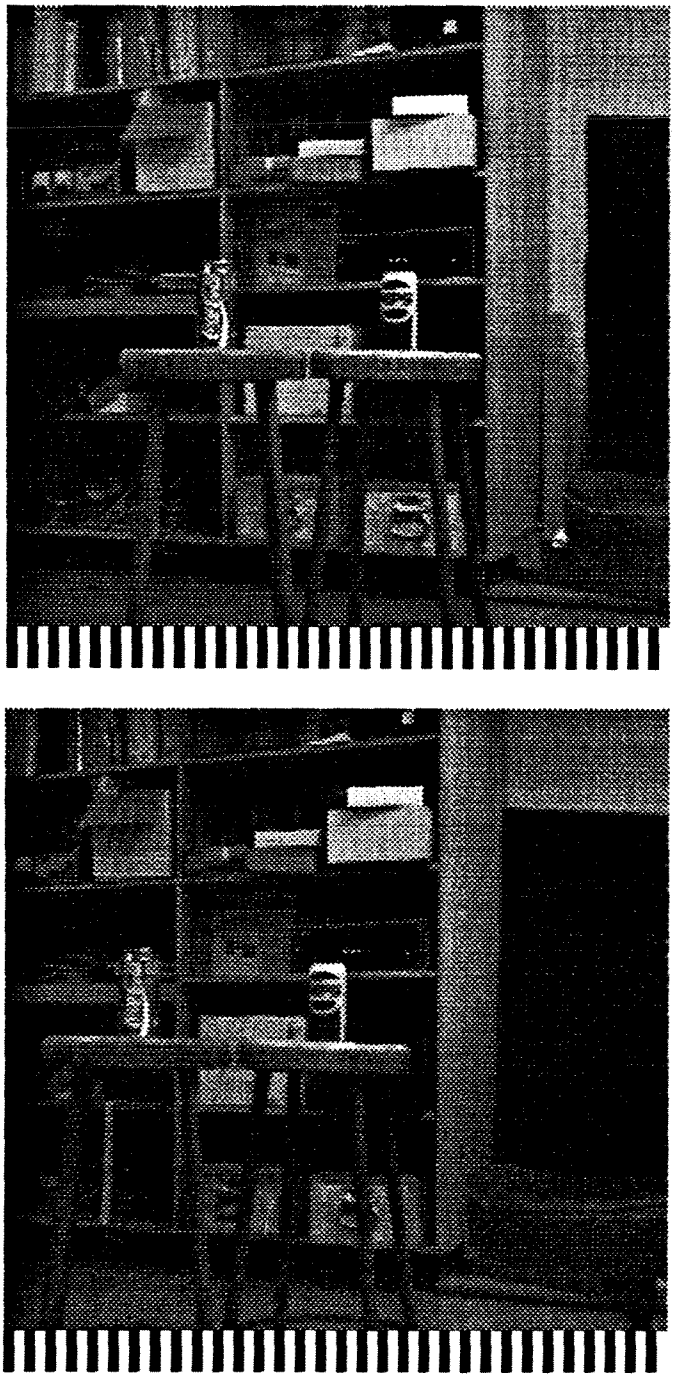


Figure 5.55: Experiment 8, Stereo Images at t_4, *(upper)* Left Image, *(lower)* Right Image.

extremely difficult to analyze. The image sequence contains viewpoint sensitive alignments of foreground and background features, which are unstable with respect to sensor motion as well as stereo camera separation. These unstable features appear in image regions near the stools.

Generating the correct stereo correspondences in image regions near the stools will be difficult: the depth gradients are very large; the presence of the stools changes the order in which features appear along an epipolar line in the left and right images, making the heuristic ordering constraint invalid; and there are unstable alignments of foreground and background features. The different alignment of the foreground and background in the left and right images can result in a low relative magnitude between stereo features, prompting a "no match" response from the correspondence tester. The "no match" response most likely to occur if the image projection of the aligned backgrounds (in left and right images) contain significant features. If the background is uniform over the spatial extent of the Gabor function, the correspondence testing and the disparity estimation of foreground features will not be affected. Note that the rejection of unstable measurements is desirable in moderation; excessive culling will make the stools "invisible" to the algorithm.

The normal image velocity module will perform in a similar manner. Temporal correspondences for features belonging to moving objects will be difficult to create. As the object moves, its position relative to the background changes, causing many candidate correspondences to be rejected in regions where the background not uniform locally. Generating temporal matches will be most difficult for the P-cola stool. The image velocity associated with the P-cola stool is very large; there is a good possibility that the velocity bandwidth of a correspondence predictor, tuned to stationary objects, will be exceeded. Generating the correct temporal match will be easier for the C-cola object because it has a collision trajectory. An object on a collision trajectory produces a near zero image

velocity when viewed by stationary cameras.

The objectives of this experiment are: to measure the inter-frame and extended sensor motions; to measure the object motions; and to predict the collision parameters. Success of this experiment will verify the ability of the obstacle detection algorithm to segment and track multiple objects, both collision and pass-by cases.

The interpolated disparity and its uncertainty are shown in figures 5.56, 5.57, and 5.58. Despite the difficult scene structure, the disparity module is able to produce many good disparity measurements.

The top and side views of the local map are shown in figure 5.59. The two forward clusters of features belong to the two stools. The other features belong to the bookshelf in the background. The planar structure of the bookshelf is correctly depicted in figure 5.59.

Figure 5.60 shows the stereo image velocity for the epipolar channel at time t_3. The actual normal image velocity of each background feature is zero. The RMS error in the set of normal image velocity measurements belonging to the stationary background, compared to the set predicted by the inter-frame sensor motion, is 0.14 pixels. The epipolar image velocity for the C-cola stool is small and has opposite directions in the left and right images. This stereo flow is typical for an object on a collision trajectory (when there is no sensor rotation). The stereo image velocity for the P-cola stool has the same direction in both images, as we would expect for an object that will pass in front of the cameras. The inter-frame displacement for the P-cola stool is very large: approximately 14 pixels during the first inter-frame transition. This displacement is near the limit of the velocity bandwidth [19] for a correspondence predictor, tuned to stationary objects, for the channel $\tilde{\omega}_1$.

[19]Depending on which four points are selected by the correspondence predictor, the maximum measurable displacement can vary. For the channel $\tilde{\omega}_1$, the range in the $\hat{x}$ direction varies from 5.5 to 16.5 pixels ($0.5\Delta\hat{x}_s$ to $1.5\Delta\hat{x}_s$).

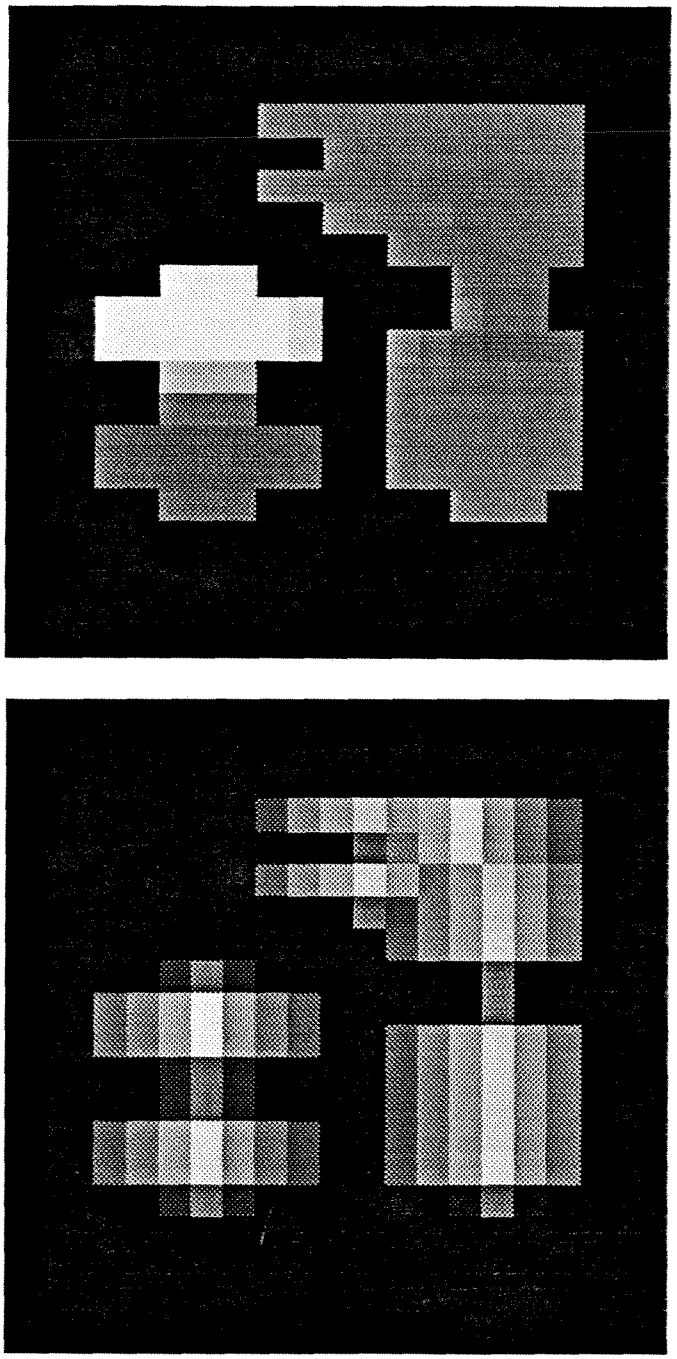


Figure 5.56: Experiment 8, $\tilde{\omega}_0 = 0.040\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 22 pixels and 57.3 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

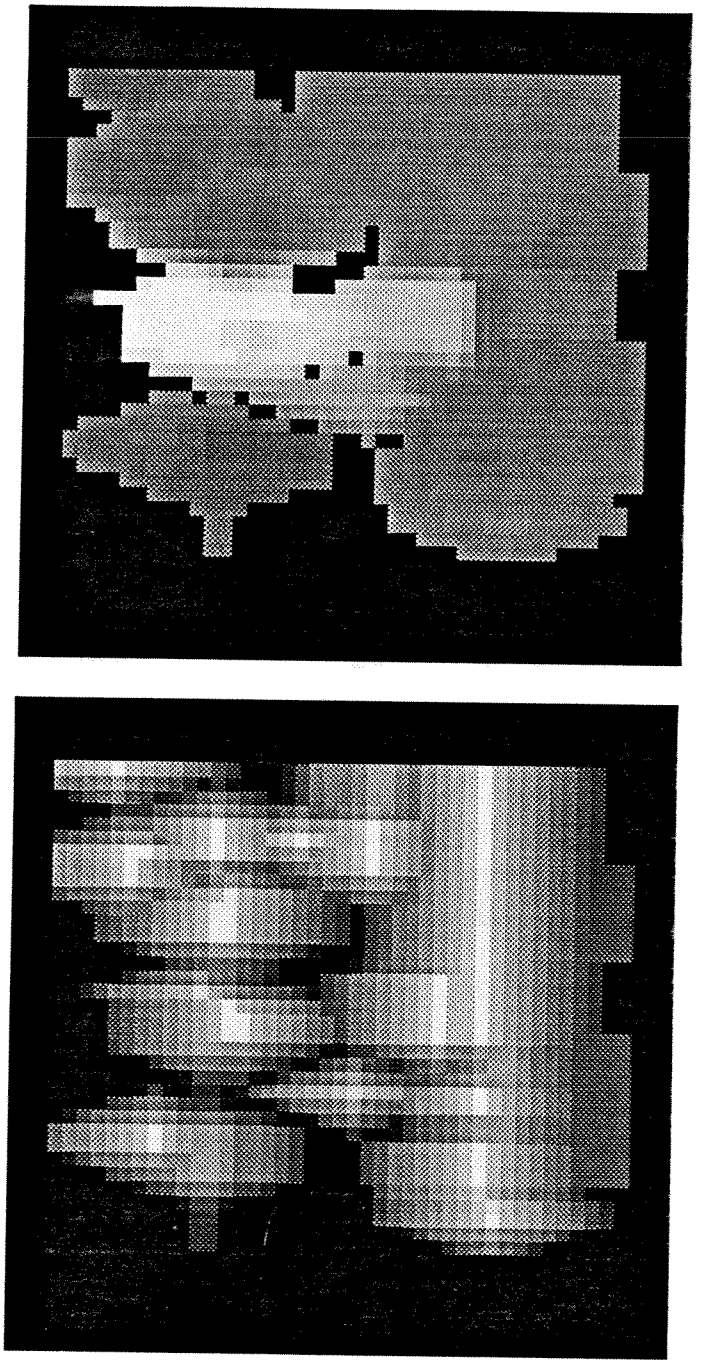


Figure 5.57: Experiment 8, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 22 pixels and 57.3 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

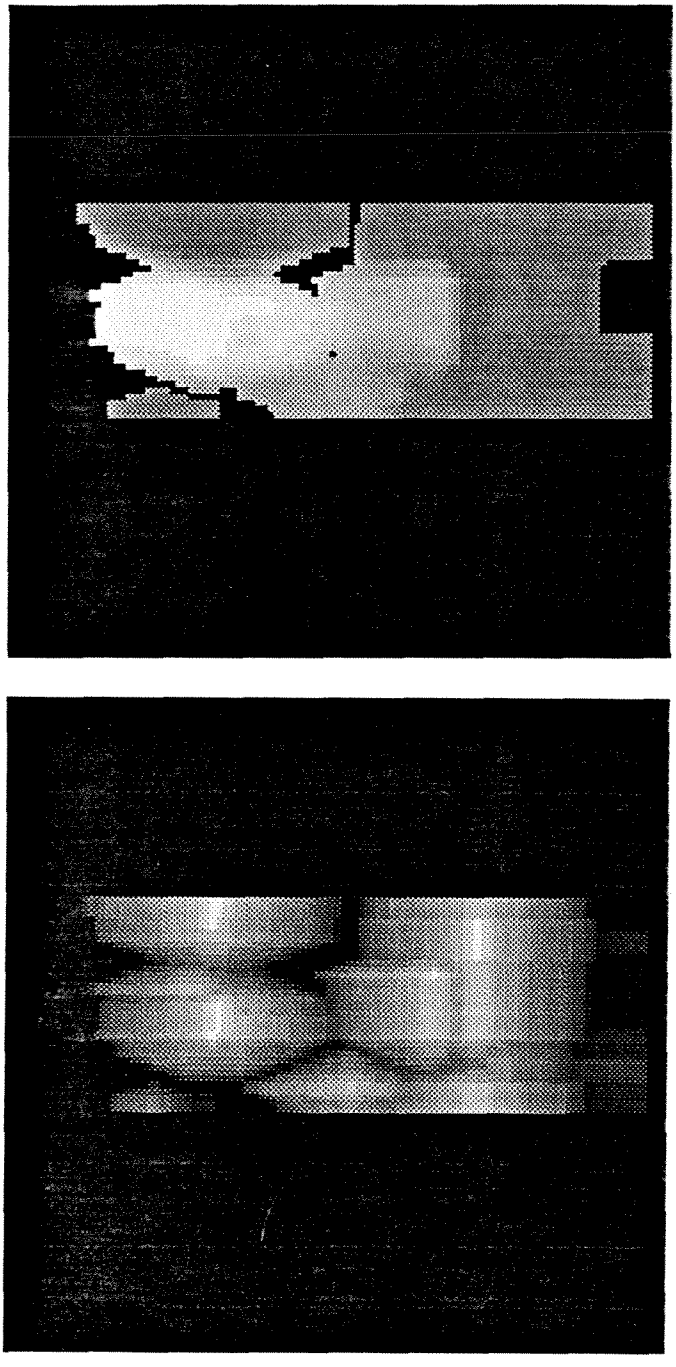


Figure 5.58: Experiment 8, $\tilde{\omega}_2 = 0.210\pi$ rad/pixel. *(upper)* Interpolated Disparity. The minimum (black) and maximum (white) responses are 22 pixels and 57.3 pixels, respectively. *(lower)* Uncertainty. Dark regions have large uncertainties. Light regions denote direct disparity measurements. A region that is black in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

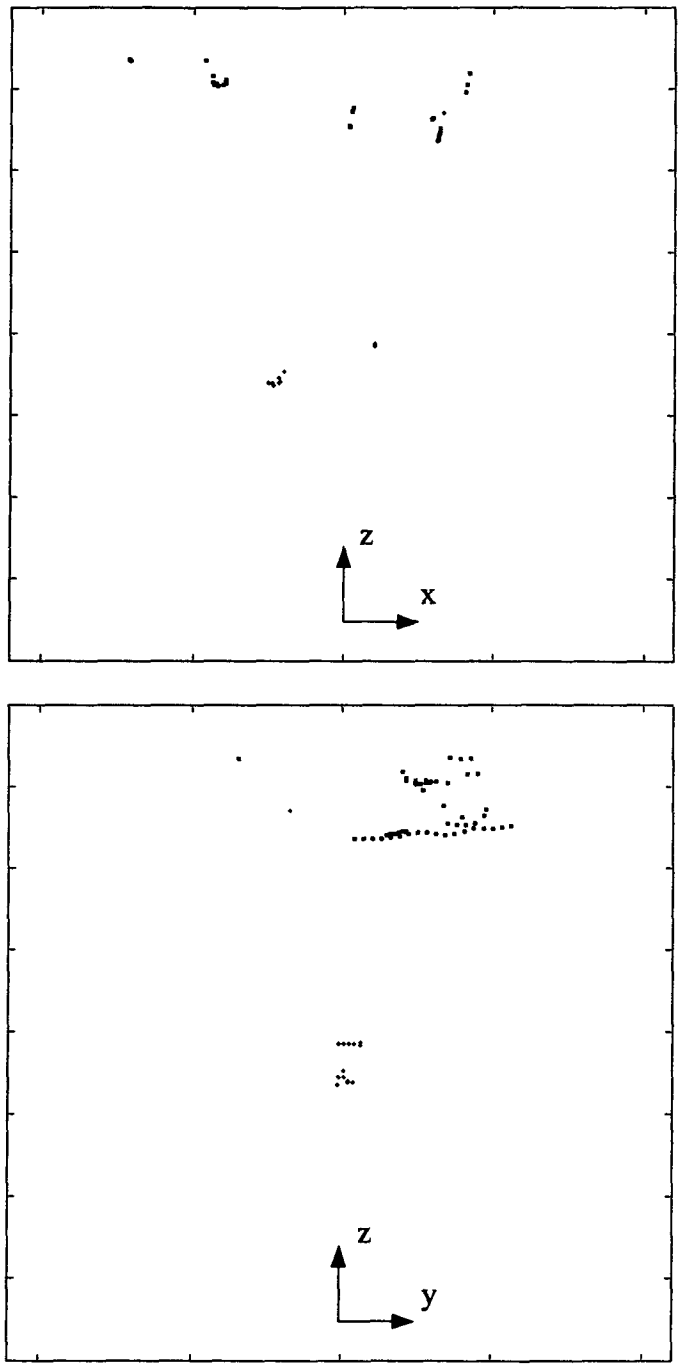


Figure 5.59: Experiment 8, Local Map. Stationary stereo features are denoted by black squares. Moving stereo features are denoted by black crosses (in foreground). Distance between ticks is 62.5 cm. *(upper)* Top View, x-z projection, *(lower)* Side View, y-z projection.

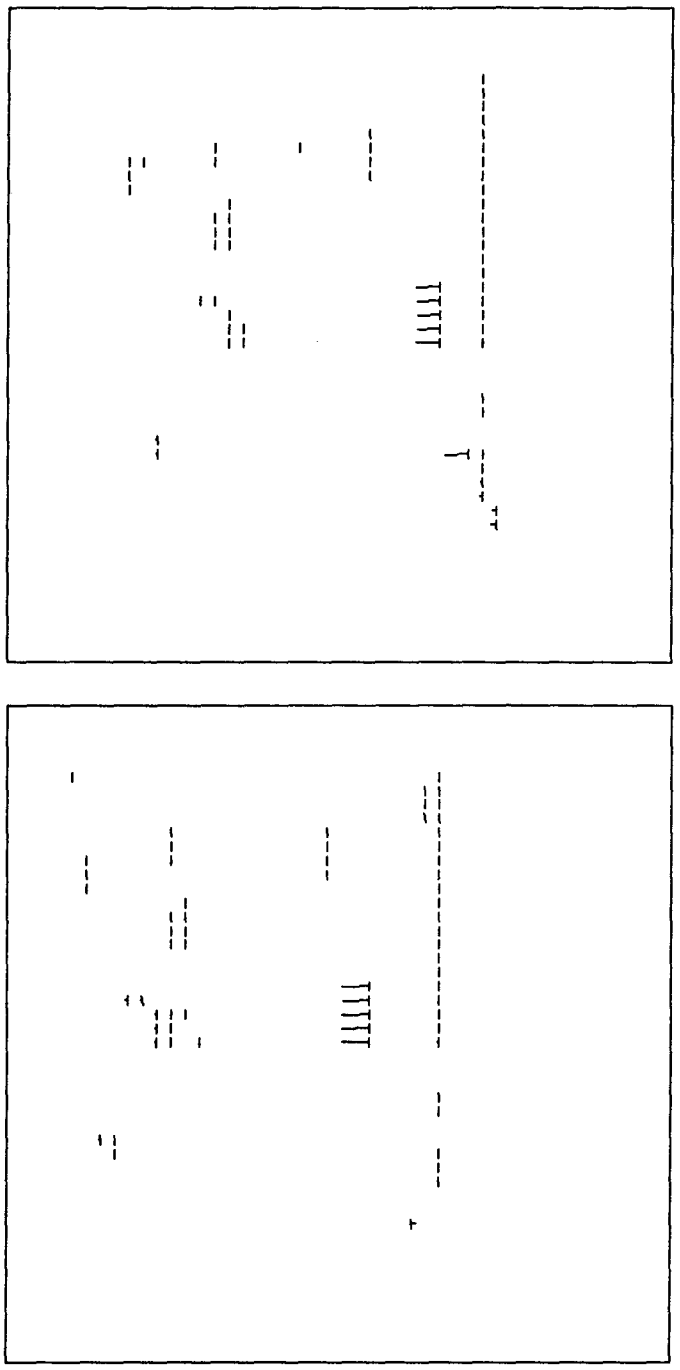


Figure 5.60: Experiment 8, Normal Image Velocity for Epipolar Channel, $\tilde{\phi}_0 = 0$, $\tilde{\omega}_1 = 0.092\pi$ rad/pixel. Component flow vectors are represented by a "T." The direction and length of the stem of the "T" denote the normal direction and the image displacement, respectively. *(upper)* Left View, *(lower)* Right View.

Table 5.31: Inter-frame Sensor Motion for Experiment 8

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	0.1082	-0.0008	-0.0417	-0.002	-0.224	0.344
1-2	-0.0124	-0.0685	0.0360	-0.198	0.086	0.576
2-3	-0.0473	-0.0996	-0.0085	-0.235	0.124	0.141
3-4	0.0718	-0.0368	-0.0552	-0.057	-0.118	0.560

Table 5.32: Expected Error in the Inter-frame Sensor Motion for Experiment 8

	cm/frame			10^{-3} rad/frame		
Frame	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0-1	± 0.1865	± 0.0676	± 0.1461	± 0.191	± 0.447	± 0.395
1-2	± 0.1359	± 0.0731	± 0.1451	± 0.206	± 0.336	± 0.529
2-3	± 0.1173	± 0.0705	± 0.1676	± 0.228	± 0.301	± 0.569
3-4	± 0.1216	± 0.0585	± 0.1759	± 0.201	± 0.306	± 0.511

The inter-frame sensor motions and the expected errors appear in tables 5.31 and 5.32, respectively. Since the cameras are stationary, each parameter should be zero. Most of the inter-frame parameters are within the expected errors.

The eigenvalues and eigenvectors are given by [20]

$$\lambda_0 = 6780.0 \quad \bar{v}_0 = [\ -0.065 \quad 0.739 \quad -0.036 \quad -0.666 \quad -0.068 \quad -0.022\ ]^T,$$

$$\lambda_1 = 5522.9 \quad \bar{v}_1 = [\ 0.676 \quad 0.069 \quad 0.010 \quad -0.062 \quad 0.729 \quad -0.046\ ]^T,$$

$$\lambda_2 = 124.24 \quad \bar{v}_2 = [\ -0.035 \quad 0.638 \quad 0.252 \quad 0.701 \quad 0.017 \quad -0.191\ ]^T,$$

$$\lambda_3 = 45.557 \quad \bar{v}_3 = [\ -0.128 \quad -0.129 \quad 0.948 \quad -0.198 \quad 0.110 \quad 0.129\ ]^T,$$

$$\lambda_4 = 39.569 \quad \bar{v}_4 = [\ -0.126 \quad -0.161 \quad 0.056 \quad -0.144 \quad 0.058 \quad -0.965\ ]^T,$$

$$\lambda_5 = 15.465 \quad \bar{v}_5 = [\ 0.711 \quad -0.019 \quad 0.181 \quad -0.028 \quad -0.669 \quad -0.115\ ]^T.$$

[20]The rotation terms in the eigenvectors have been normalized by z_{norm} = 400 cm.

Table 5.33: Extended Sensor Motion for Experiment 8

	cm/frame			Pred. Error cm/fr		
Frame	T_x	T_y	T_z	ΔT_x	ΔT_y	ΔT_z
0-1	0.1082	-0.0008	-0.0417	± 0.1865	± 0.0676	± 0.1461
0-2	0.0338	-0.0319	-0.0094	± 0.1093	± 0.0496	± 0.1022
0-3	-0.0059	-0.0558	-0.0182	± 0.0798	± 0.0403	± 0.0867
0-4	0.0169	-0.0503	-0.0231	± 0.0666	± 0.0330	± 0.0777

The condition number is 438. The inter-frame parameters associated with λ_3, λ_4, and λ_5 (T_x, T_z, Ω_x, and Ω_z) are sensitive to measurement errors.

The extended sensor motion appears in table 5.33. The extended sensor motion is approximately zero.

The segmentation of the image sequence at four time instants is shown in figures 5.61 and 5.62. The features on the right belong to the P-cola stool. During the inter-frame transition from t_0 to t_1, all correspondence predictors are tuned to stationary objects. As a result, only one stereo feature belonging to the P-cola stool is detected at time t_0 (the algorithm is fortunate to have detected this feature). After detecting this first feature, a new correspondence predictor, tuned to this moving object, is automatically generated. As a result, more features belonging to the P-cola object are detected in later images. Note that the tracked features for the P-cola stool change over time, as the alignment of the stool and the background changes.

The feature at the left in figures 5.61 and 5.62 belongs to the C-cola stool. Most of the tracked C-cola features are found in other epipolar channels ($\tilde{\omega}_0$ and $\tilde{\omega}_2$). Note that a correspondence predictor tuned to a stationary object is also tuned to a moving object with a collision trajectory, when the cameras are stationary. As a result, the detection of the C-cola stool does not cause a new correspondence predictor to be generated.

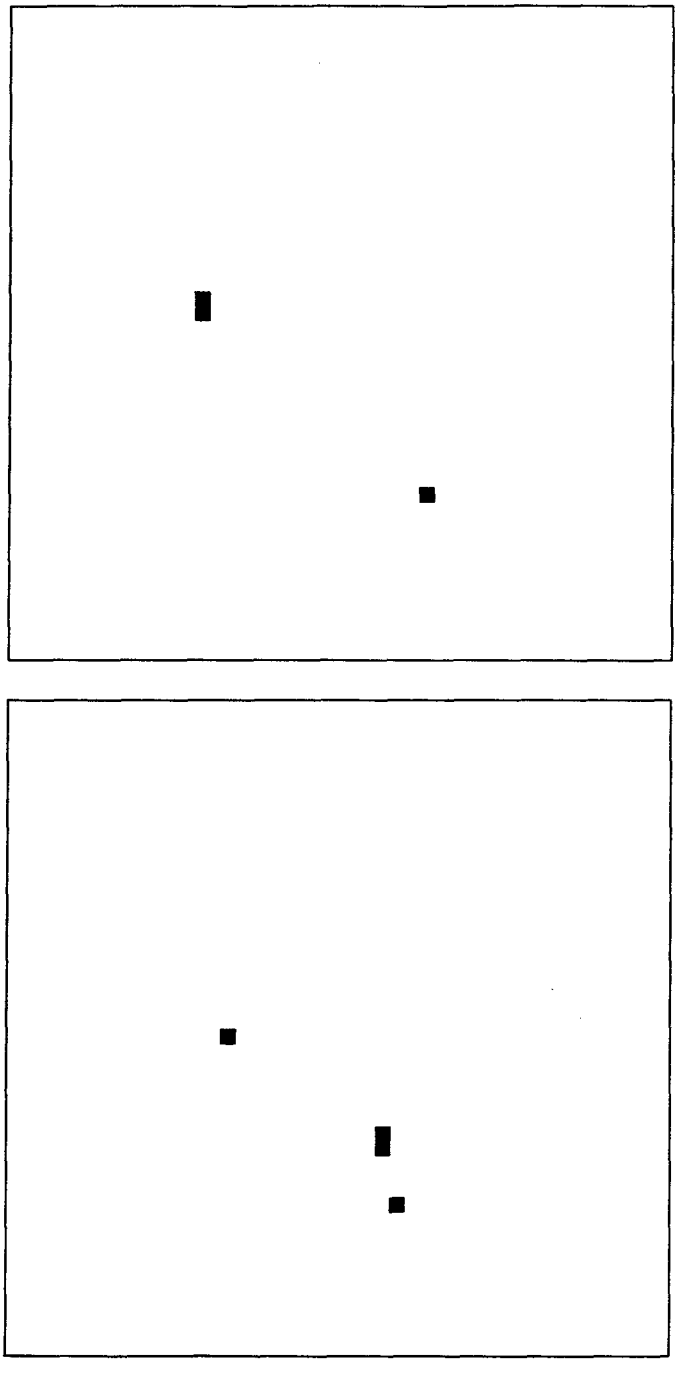


Figure 5.61: Experiment 8, Segmentation of Image Sequence for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel, Stereo features identified as belonging to the P-cola (C-cola) stool are denoted by black (gray) squares. *(upper)* Epipolar Channel, t_0, *(lower)* Epipolar Channel, t_1.

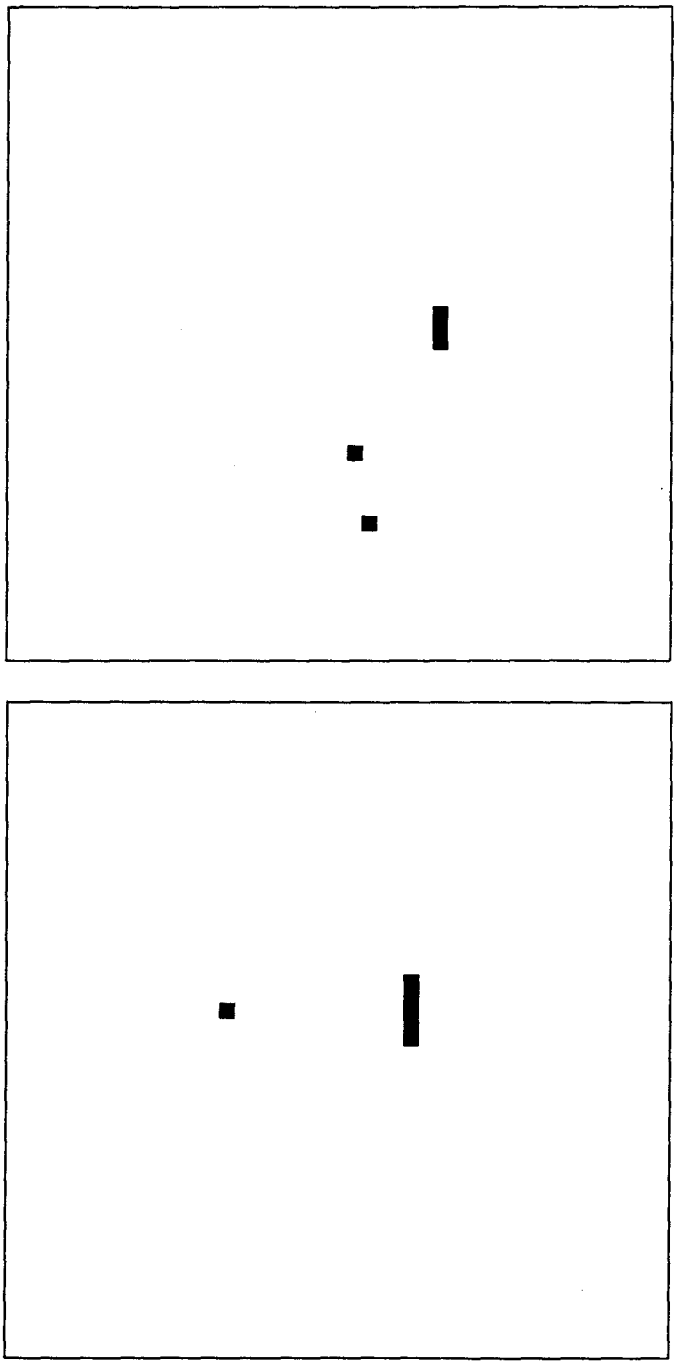


Figure 5.62: Experiment 8, Segmentation of Image Sequence for $\tilde{\omega}_1 = 0.092\pi$ rad/pixel, Stereo features identified as belonging to the P-cola (C-cola) stool are denoted by black (gray) squares. *(upper)* Epipolar Channel, t_2, *(lower)* Epipolar Channel, t_3.

Table 5.34: Extended P-cola Motion for Experiment 8

Frame	cm/frame $T_{x,obj}$	cm/frame $T_{z,obj}$	Pred. Error cm/fr $\Delta T_{x,obj}$	Pred. Error cm/fr $\Delta T_{z,obj}$
0-1	-5.912	-13.945	± 0.520	± 7.944
0-2	-6.320	-19.678	± 0.160	± 3.208
0-3	-6.198	-17.984	± 0.116	± 2.196
0-4	-5.905	-17.498	± 0.100	± 1.597

Table 5.35: Extended C-cola Motion for Experiment 8

Frame	cm/frame $T_{x,obj}$	cm/frame $T_{z,obj}$	Pred. Error cm/fr $\Delta T_{x,obj}$	Pred. Error cm/fr $\Delta T_{z,obj}$
0-1	1.481	-9.122	± 0.193	± 1.418
0-2	1.427	-8.985	± 0.134	± 1.003
0-3	1.373	-9.280	± 0.166	± 1.228
0-4	1.417	-9.841	± 0.131	± 0.969

The extended object motion for the P-cola and C-cola stools appear in tables 5.34 and 5.35. The actual P-cola velocity is approximately $(\dot{x}_{obj}, \dot{z}_{obj}) = (6.0, 19.1)$; the actual C-cola velocity is about $(\dot{x}_{obj}, \dot{z}_{obj}) = (1.4, 9.9)$. The C-cola estimates are closer to the actual value than the P-cola estimates. Most of the object translation parameters are within the expected errors.

Each feature on a given object has a different point-of-collision. The P-cola and C-cola collision parameters, at selected reference points, appear in tables 5.36 and 5.37. The reference point for the P-cola stool is at the left leg for t_0 and t_1, and at the right side of the P-cola can for t_2 and t_3 [21]. The reference point for the C-cola stool is at the white stripe of the C-cola can. Most of the estimated collision parameters are within the

[21]The change of the reference point alters the actual x_{col} in table 5.36.

Table 5.36: P-cola Collision Parameters for Experiment 8

	Units: x_{col} cm, t_{col} frames					
	Estimate		Actual		Pred. Error	
Frame	x_{col}	t_{col}	x_{col}	t_{col}	Δx_{col}	Δt_{col}
0	-119.0	22.2	-80	16	± 65.4	± 12.7
1	-88.4	15.0	-80	15	± 13.7	± 2.5
2	-74.4	14.3	-65	14	± 9.7	± 1.8
3	-74.1	13.8	-65	13	± 6.5	± 1.3

Table 5.37: C-cola Collision Parameters for Experiment 8

	Units: x_{col} cm, t_{col} frames					
	Estimate		Actual		Pred. Error	
Frame	x_{col}	t_{col}	x_{col}	t_{col}	Δx_{col}	Δt_{col}
0	5.0	26.5	3	25	± 5.2	± 4.2
1	6.0	25.8	3	24	± 3.1	± 2.9
2	4.5	23.9	3	23	± 2.2	± 3.2
3	3.2	21.6	3	22	± 1.6	± 2.2

expected tolerances. The collision parameters correctly identify the P-cola stool as an object that will pass safely in front of the cameras, and the C-cola stool as an obstacle. The final error in the time-to-collision is 6.2 percent and 1.9 percent of the actual values for the P-cola and C-cola stools, respectively.

In summary, experiment 8 has shown that the obstacle detection algorithm can estimate the collision parameters of multiple moving objects, simultaneously.

5.5.4 Summary

Data set 3 has tested the obstacle detection algorithm under realistic conditions. Experiment 6 demonstrates that the disparity module can operate in an outdoor environment which has large depth gradients and uneven lighting (shadows). Experiment 7 demonstrates the sensor motion module's robustness to transient rotations associated with camera shake. Experiment 8 demonstrates the obstacle detection algorithm's ability to process multiple moving objects.

5.6 Summary

The results presented in this chapter demonstrate the robustness of the obstacle detection algorithm to various difficult conditions: stereo images with different brightness and contrast (experiments 3, 4, 5); images containing shadows and specular reflections (experiments 2, 3, 4, 5, 6); unstable alignments of foreground and background features (experiment 8); image sequences with large image velocities (experiment 8); scenes with large depth gradients (experiment 6); constant and transient sensor rotations (experiments 3, 7); concurrent object and sensor motions (experiments 4, 5); and multiple moving objects (experiment 8). Experiments 1, 2, and 6 verified the correct implementation of the stereo candidate correspondence predictors. Experiment 8 demonstrated the

Table 5.38: RMS Error in the Normal Image Velocity

Experiment	RMS Error	Comments
1	0.10 pixels	
2	0.09 pixels	
3	0.16 pixels	Lens Distortion
4	0.17 pixels	Lens Distortion
5	0.18 pixels	Lens Distortion
7	0.17 pixels	
8	0.14 pixels	

automatic generation and the correct operation of temporal correspondence predictors tuned to moving objects. Experiment 4, 5, and 8 demonstrated the correct operation of the segmentation algorithm.

The accuracy of various modules within the obstacle detection algorithm have been measured. The RMS error of the disparity and the depth in experiment 1 is ± 0.16 pixels and 0.3 percent of the actual depth, respectively. Theses results are comparable with the published results of Matthies et al [39] (0.12 pixels and 0.5 percent of the actual depth for a similar scene). The RMS error in the normal image velocity field is summarized in table 5.38. All of the RMS errors are within the one pixel "satisfactory" level used by Weng et al [52].

The directional errors in the sensor translation are summarized in table 5.39. The directional accuracy tends to be better when the condition number of the inter-frame Hessian is low and the speed of sensor translation is large. The directional accuracy degrades in the presence of lens distortion and feature clustering. The better results presented in this chapter (experiments 2, 7) are below the one degree standard established in section 5.2; a standard which is based on the best reported results of other researchers.

The percentage error in the time-to-collision for moving objects is summarized in

Table 5.39: Directional Errors in Sensor Translation

Experiment	Directional Error		Comments
	Pan (deg.)	Tilt (deg.)	
1	1.68	1.61	Frontal Plane, ill-conditioned
2	-0.79	-0.89	
3	0.02	-1.4	Lens Distortion
4	-3.3	-1.2	Lens Dist., Feature Clustering
5	-3.2	-1.9	Lens Dist., Feature Clustering
7	-0.39	0.08	Large translation (10 cm/fr)

Table 5.40: Percentage Error in the Time-to-Collision

Experiment	Percent Error	Comments
4	10 percent	Lens Distortion
5	15 percent	Lens Distortion
8 (P-cola)	6.2 percent	
8 (C-cola)	1.9 percent	

table 5.40. The lens distortion causes the time-to-collision to be under-estimated in experiments 4 and 5. The accuracy of the time-to-collision for the P-cola and C-cola objects in experiment 8 are within the expected errors and should be sufficient for most obstacle detection/avoidance applications. The point-of-collision in each experiment is sufficiently accurate relative to the size of the object to discriminate between obstacles with collision trajectories (experiments 4, 8) and objects with pass-by trajectories (experiments 5, 8).

Chapter 6

Summary and Conclusion

In this chapter, the obstacle detection algorithm is summarized, possible extensions are discussed, and conclusions are made.

6.1 Summary

The obstacle detection algorithm transforms a stereo image sequence from pixels with time varying intensities to the collision parameters for viewed objects. The collision parameters, the point-of-collision and the time-to-collision, along with the expected errors can be used by a computer pilot to avoid obstacles.

The pixel-based stereo image sequence is first converted into a Gabor representation (section 3.2.1). The representation comprises a set of channels formed by filtering the image sequence with log-polar Gabor filters. Minimal completeness is used to form a laconic description of the filter set. Once a reference frequency, orientation, and phase have been chosen ($\tilde{\omega}_0$, $\tilde{\phi}_0$, and p_0), we need only to select the number of orientations $n_{\tilde{\phi}}$, and one of the following: the aspect ratio α, the ratio of adjacent frequencies ρ, or the bandwidth-frequency ratio λ.

The output of each channel is sampled, forming a two-dimensional sampling lattice. Two types of lattice are used: a band sampled lattice for epipolar channels, and a restricted sampling lattice for oblique channels. Constraints on the spatial sampling intervals that avoid aliasing in the measurement of spatial frequency, disparity, and normal image velocity are established.

The magnitude and phase responses in each channel is used to extract image features, estimate disparity, and estimate normal image velocity. To extract features, three thresholds are applied to the magnitude response: an absolute threshold, a relative spatial threshold, and a relative orientation threshold. The features with unstable phase responses are rejected using frequency and magnitude tests. The phase and magnitude responses are also used to estimate the expected error in the disparity and normal image velocity measurements.

The disparity is measured using a combination feature matching-phase gradient approach (section 3.2.4). The feature matching stage attempts to establish stereo correspondences between lattice points in the left and right images. The candidate lattice shifts are predicted using: an epipolar offset histogram, multiscale consistency, temporal consistency, and a heuristic ordering constraint. The candidate shifts are tested using a matching criteria that compares attributes such as the local magnitude and the phase shift. Once the correspondence has been established, the disparity estimate is refined using the phase gradient.

The normal image velocity is measured using a similar feature matching-phase gradient approach (section 3.2.5). The candidate correspondence prediction is achieved using estimates of the inter-frame sensor motion (obtained from lower frequency channels) and the object motion (obtained from past measurements). Three types of correspondence predictors are available. They are tuned to: stationary objects, objects with collision trajectories, and moving objects seen in past measurements. Correspondence predictors of the third kind (tuned to past moving objects) are generated automatically. Once the correspondence has been established, the normal image velocity is refined using the phase gradient.

Features in the image sequence are segmented into moving objects and stationary objects using: the local $\dot{z}$ velocity, seeding histograms, and two Mahalanobis distance

measures. The local $\dot{z}$ estimate, obtained from the stereo image velocity, is used to detect moving objects. The seeding histograms identify stationary object features by testing the set of image measurements for in-plane velocity consistency (section 4.6.2). Consistent measurements are combined to produce an initial estimate of the inter-frame sensor motion. The first Mahalanobis distance is used to test the hypothesis that a normal image velocity measurement belongs to a stationary object (section 4.3). The second Mahalanobis distance tests the hypothesis that two moving object classes belong to the same object (section 4.6.4).

The inter-frame sensor motion is estimated using image measurements (normal image velocity and disparity) from features belonging to stationary objects (section 4.2). The image measurements are weighted using the expected squared errors which are approximated using stable image quantities. The weighted least square estimation also produces an inter-frame Hessian matrix whose inverse is the error covariance. The inter-frame error covariance is used in the first Mahalanobis distance, in the eigenvalue analysis, and in the error covariance of the extended sensor motion.

The eigenvalues and eigenvectors of the inter-frame Hessian are useful for determining the stability of the motion parameter estimates, and for predicting which motion constraints will be most effective in improving the stability of the parameters. Three different constraints are used in this work: the known rotation, known plane, and the unknown plane. The known rotation and the known plane constraints improve the inter-frame sensor motion when the Hessian is ill-conditioned (see experiment 1).

A Kalman filter, which integrates inter-frame sensor translations, is used to estimate the translational parameters and error covariance for the extended sensor motion (section 4.4). The rotation terms are decoupled from the inter-frame Hessian matrix and measurement vector, which converts the sensor motion model from predominantly rectilinear to pure translation. The effect of the decoupling process is to stabilize the image

sequence from transient rotations.

Kalman filters are also used to estimate the extended translation for each moving object. The effect of inter-frame sensor motion is compensated by subtracting the predicted sensor motion-induced image velocity from the measured normal image velocity.

Collision parameters are estimated for each feature belonging to a moving object using the observer frame trajectory (section 4.5). The observer frame trajectory for each object is estimated using the difference between the extended object and sensor translations. The uncertainty in the collision parameters are calculated from the error covariance matrices for the extended object and sensor motions, and the feature's positional uncertainty.

The performance of the obstacle detection algorithm and its various modules—disparity, normal image velocity, depth, motion, segmentation, and collision parameters—are analyzed in chapter 5. Good results are obtained in difficult conditions: stereo images with different brightness and contrast (experiments 3, 4, 5); images containing shadows and specular reflections (experiments 2, 3, 4, 5, 6); unstable alignments of foreground and background features (experiment 8); image sequences with large image velocities (experiment 8); scenes with large depth gradients (experiment 6); constant and transient sensor rotations (experiments 3, 7); concurrent object and sensor motion (experiments 4, 5); and multiple moving objects (experiment 8).

The image measurements, which use the Gabor representation, are very good: the feature extraction stage selected stable image features, the disparity is measured to subpixel accuracy, the normal image velocity measurements had speeds and directions that are consistent with the sensor and object motions as well as the scene structure. The RMS error in the disparity for experiment 1 is 0.14 pixels, compared with 0.12 pixels reported in [39] for a similar scene. The RMS error in the normal image velocity measurements, compared with the component flow field induced by the inter-frame sensor motion, varies

from 0.09 pixels in experiment 2 to 0.18 pixels in experiment 5. Each RMS error is less than the 0.84 pixel error reported in [52].

The depth and motion accuracy is limited by the uncertainty associated with using nominal camera parameters. Deviations from the nominal camera parameters produce scale factor errors and distortion errors. The scale factor errors are not serious because they do not affect the structure of the scene (the relative depth of features), the direction of sensor/object translation, or the sensor rotation. Distortion errors alter the scene structure. If the features are clustered away from the image origin, distortion errors also introduce bias terms into the motion estimates, and alter the distribution of $\dot{x}$ ($\dot{y}$) between T_x and Ω_y (T_y and Ω_x).

The scene structure, the direction of translation, and the sensor rotation are accurately measured. In experiment 1, the RMS error in the measured depth, when fitted to a frontal plane structure, is ± 0.16 cm; the RMS error is 0.3 percent of the average depth (compared to 0.5 percent in [39]). The directional accuracy of the extended sensor translation for experiment 7 is within 0.4 degrees of the actual value (compared to 1.0 degree in [40]). The Ω_y rotation in experiment 3 is within ± 0.01 of a degree for an inter-frame rotation of 0.5 degrees.

The segmentation of moving objects and stationary objects is successful in each of the experiments containing moving objects. All the identified features belonged to moving objects. The accuracy of the object translation is good; that is, it is usually within the expected error. The accuracy tends to be better for obstacles with collision trajectories than pass-by objects.

The estimates of the collision parameters are good. The scale factor errors due to camera uncertainty have no effect on the time-to-collision; most scale factor errors do not affect the point-of-collision. The distortion errors in experiment 4 and 5 caused the magnitude of the collision parameters to be under-estimated. Despite the distortion

errors, the point-of-collision in each experiment is sufficiently accurate, relative to the size of the object, to determine if the object will collide with or pass-by the cameras.

6.2 Extensions

There are many possible extensions to the obstacle detection algorithm presented in this work. They include the use of a priori scene structure, and additional sensors.

In many operating environments, information regarding the scene structure is available, such as a scene model or the maximum (minimum) depth. A model of the scene, such a planar ground surface, can be used to produce default disparity estimates at each lattice point, replacing or enhancing, the epipolar offset histogram. Minimum and maximum disparities can be used as search bounds for the epipolar offset histogram and the heuristic ordering constraint.

The seeding histograms, which test in-plane velocity consistency, use the assumption that the axial rotation (Ω_z) is zero. Although failure of this assumption has little effect for stationary environments, it may prove more significant for scenes with many moving objects. The seeding histograms can be easily altered to use a known Ω_z. An auxiliary sensor that measures camera roll could provide this information. In certain cases, the camera roll can be estimated from the image sequence prior to the seeding stage. If the scene has a horizon, the changes in the image position and orientation of the horizon line can produce initial estimates of Ω_y and Ω_z. Horizon line is often easily identified; in experiment 6, it is a dark to light transition extending across the image.

A third camera that provides a vertical baseline separation would be useful for disparity estimation, seeding histograms, and object motion estimation. In the current implementation, disparity is estimated using features from the epipolar channel only. A vertical camera separation could provide direct disparity measurements for features

in the orthogonal channel [1]. The vertical cameras would also provide a stereo image velocity field for the orthogonal channel. The resulting estimate of $\dot{z}$ can be used to remove moving object features from the orthogonal channel's seeding histogram, increasing the reliability of the seeding process. The moving object features from the orthogonal channel can be combined into object classes, providing an accurate estimate of vertical component of object translation and an improved estimate of the axial component.

Pilot commands can be incorporated into the sensor's Kalman filter to predict changes in the collision parameters in response to an evasive maneuver. In a simpler Kalman filter implementation, pilot commands can be used to adjust the forgetting factor. When the vehicle changes heading or speed, the forgetting factor can be temporarily increased to flush old data. The Kalman filters for the object motion can also be expanded. If the maneuverability of the object is known, a process noise model can be formed. The uncertainty in the collision parameters would represent a probabilistic model accounting for possible perturbation in the object trajectory.

6.3 Conclusion

This work has demonstrated the utility of the Gabor representation as an image processing stage for stereo-based obstacle detection. The importance of error estimation and propagation have also been demonstrated.

The work presented in this thesis makes the following principal contributions:

- it develops sampling constraints and predictive matching criteria that detect and avoid aliasing in the measurements of local frequency, disparity, and normal image velocity;

[1] Orthogonal with respect to the horizontal stereo cameras.

- it provides robust detection of moving objects by automatically generating trajectory detectors tuned to moving objects seen in past images;
- it implements direct passive navigation using phase-differences instead of intensity derivatives;
- it develops image measurement error models using stable quantities, and propagates the image measurement errors into collision parameter uncertainty;
- it develops a "seeding" technique that is necessary to initialize the segmentation process;
- it stabilizes the image sequence from transients caused by camera shake.
- it uses eigenvalue/eigenvector analysis to determine if a motion estimate is stable and which motion constraint will most improve the stability of the sensor motion estimates.

All of these contributions produce a very robust obstacle detection algorithm.

Appendix A

Discount Factor

The "weight" used in the weighted least square estimate of the inter-frame sensor motion (see section 4.2) is given by

$$w = \frac{\beta_{discount}}{E[(\delta V_n)^2]}, \tag{A.322}$$

where $\beta_{discount}$ is a discount factor that compensates for the spatial oversampling. This appendix shows how the discount factor is determined.

Assume that an image is corrupted by white noise whose power is given by σ_n^2. When the image is filtered by the Gabor filter (Gaussian bandpass), the noise power is reduced and becomes correlated. The noise power within the Gabor channel is

$$\sigma_G^2 = \sigma_n^2 < G_i, G_i >= \frac{\sigma_n^2}{\sigma_s}, \tag{A.323}$$

where

$$< G_i, G_j >= \int\int G_i G_j \, d\acute{x} \, d\acute{y}. \tag{A.324}$$

The correlation matrix R_G is given by

$$R_G = \sigma_G^2 R_s, \tag{A.325}$$

where R_s is a matrix that contains the correlation (or overlap) between every pair of Gabor functions in the set of lattices. The matrix R_s usually contains the correlation between Gabor functions from different channels (referred to as "cross-channel correlation"). In this work, the cross-channel correlation is small because the set of Gabor filters is minimally complete (see section 3.2.1). The most significant correlation occurs

within a given channel because the spatial sampling lattice is oversampled (to calculate the phase gradient without aliasing; see section 3.2.3).

If we ignore the cross-channel correlation and consider only the correlation in the $\acute{x}$ direction, the correlation matrix R_s is given by

$$R_{s,\acute{x}} = \begin{bmatrix} & \vdots & \vdots & \vdots & \\ \cdots & < G_{-1}, G_{-1} > & < G_0, G_{-1} > & < G_1, G_{-1} > & \cdots \\ \cdots & < G_{-1}, G_0 > & < G_0, G_0 > & < G_1, G_0 > & \cdots \\ \cdots & < G_{-1}, G_1 > & < G_0, G_1 > & < G_1, G_1 > & \cdots \\ & \vdots & \vdots & \vdots & \end{bmatrix}. \tag{A.326}$$

If the Gabor functions are normalized such that $< G_i, G_i >= 1$, $R_{s,\acute{x}}$ can be rewritten as

$$R_{s,\acute{x}} = \begin{bmatrix} & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & a & a^4 & a^9 & \cdots \\ \cdots & a & 1 & a & a^4 & \cdots \\ \cdots & a^4 & a & 1 & a & \cdots \\ \cdots & a^9 & a^4 & a & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}, \tag{A.327}$$

where

$$a = \exp[-\frac{\pi}{2}(\lambda \tilde{\omega}_k \Delta \acute{x}_s)^2]. \tag{A.328}$$

It can be seen that the correlation decreases as the separation between pairs of Gabor functions increases; that is, the elements of $R_{s,\acute{x}}$ decrease from the diagonal. Thus, $R_{s,\acute{x}}$ is a local operator.

The matrix R_s used in (A.325) contains the correlation in both the $\acute{x}$ and $\acute{y}$. The

matrix R_s is written as a block matrix:

$$R_s = \begin{bmatrix} & \vdots & \vdots & \vdots & \\ \cdots & R_{s,\acute{x}} & bR_{s,\acute{x}} & b^4R_{s,\acute{x}} & \cdots \\ \cdots & bR_{s,\acute{x}} & R_{s,\acute{x}} & bR_{s,\acute{x}} & \cdots \\ \cdots & b^4R_{s,\acute{x}} & bR_{s,\acute{x}} & R_{s,\acute{x}} & \cdots \\ & \vdots & \vdots & \vdots & \end{bmatrix}, \tag{A.329}$$

where

$$b = \exp[-\frac{\pi}{2}(\frac{\lambda}{\alpha}\tilde{\omega}_k\Delta\acute{y}_s)^2]. \tag{A.330}$$

The blocks in R_s decrease from the diagonal as the separation along the $\acute{y}$-axis increases. Thus, R_s is a local operator.

The optimal weighting for a least square estimate of the inter-frame sensor motion is [40]

$$\bar{\theta} = (H^TWH)^{-1}H^TW\bar{V}_n, \tag{A.331}$$

where $\bar{V}_n$ is a vector containing all normal image velocity measurements $V_n(i)$, H is a matrix containing all the associated transformation vectors $\bar{J}_i$, and W is the optimal weighting matrix. $\bar{V}_n$ and H are given by

$$H = [\bar{J}_1\ \bar{J}_2\ \cdots\ \bar{J}_n], \tag{A.332}$$

$$\bar{V}_n = [V_n(1)\ V_n(2)\ \cdots\ V_n(n)]^T. \tag{A.333}$$

If the error due to $\delta\omega_n$ is ignored, inverse of W is given by

$$W^{-1} = M^{-1}R_GM^{-T}, \tag{A.334}$$

where

$$M = \begin{bmatrix} m_0 & 0 & 0 & \cdots & 0 \\ 0 & m_1 & 0 & \cdots & 0 \\ : & : & : & & : \\ 0 & 0 & 0 & \cdots & m_n \end{bmatrix}, \tag{A.335}$$

and m_i is the magnitude at lattice point i. The optimal weighting is impractical because the matrix R_s (as well as R_G and W^{-1}) has a very large size. In addition, the inverse of R_s is required to obtain W. Because of the oversampling of the spatial lattice, a unique inverse of R_s does not exist. It is desirable to obtain an approximate local solution that is conservative, but near optimal.

The optimal solution contains the matrix product $H^T W$. Consider measurements from the epipolar channel ($\tilde{\phi}_l = 0$). If all the features are vertical, H can written as the following six column vectors: [1]

$$H = [\bar{v}_1\ \bar{v}_2\ \cdots\ \bar{v}_6], \tag{A.336}$$

where

$$\bar{v}_1 = [-z_f\ -z_f\ \cdots\ -z_f]^T, \tag{A.337}$$

$$\bar{v}_2 = [0\ 0\ \cdots\ 0]^T, \tag{A.338}$$

$$\bar{v}_3 = [\hat{x}(1)\ \hat{x}(2)\ \cdots\ \hat{x}(n)]^T, \tag{A.339}$$

$$\bar{v}_4 = [\frac{\hat{x}(1)\hat{y}(1)}{z_f}\ \frac{\hat{x}(2)\hat{y}(2)}{z_f}\ \cdots\ \frac{\hat{x}(3)\hat{y}(n)}{z_f}]^T, \tag{A.340}$$

$$\bar{v}_5 = [(z_f + \frac{\hat{x}^2(1)}{z_f})\ (z_f + \frac{\hat{x}^2(2)}{z_f})\ \cdots\ (z_f + \frac{\hat{x}^2(n)}{z_f})]^T, \tag{A.341}$$

$$\bar{v}_6 = [\hat{y}(1)\ \hat{y}(2)\ \cdots\ \hat{y}(n)]^T. \tag{A.342}$$

Note that the matrices R_s and M are symmetric; thus

$$W^{-1} = R_s[MM]^{-1}\sigma_G^2. \tag{A.343}$$

[1] Similar column vector can be formed for the other channels.

The matrix product $H^T W$ can be written as

$$H^T W = [\bar{v}_1^T R_s^{-p}\ \bar{v}_2^T R_s^{-p}\ \cdots\ \bar{v}_6^T R_s^{-p}][MM]^{-1}, \tag{A.344}$$

where R_s^{-p} is a pseudo-inverse of R_s. The lower bound for $\bar{v}_i^T R_s^{-p}$ is given by

$$\bar{v}_i^T R_s^{-p} = R_s^{-p}\bar{v}_i > \frac{1}{\lambda_0}\bar{v}_i, \tag{A.345}$$

where λ_0 is the largest eigenvalue in R_s. A sub-optimal weighting matrix is given by

$$W_{sub} = \frac{1}{\lambda_0}\frac{1}{\sigma_G^2}[MM]. \tag{A.346}$$

Since the error due to $\delta\omega_n$ is ignored, the expected square error in $V_n(i)$ is

$$E[(\delta V_n)^2]_i = \frac{\sigma_G^2}{m_i^2}. \tag{A.347}$$

Using the sub-optimal weighting matrix, we get

$$H^T W_{sub}\bar{V}_n = \sum_i w_i \bar{J}_i V_n(i), \tag{A.348}$$

$$H^T W_{sub} H = \sum_i w_i \bar{J}_i \bar{J}_i^T, \tag{A.349}$$

where

$$w_i = \frac{1}{\lambda_0}E[(\delta V_n)^2]_i. \tag{A.350}$$

Thus, the discount factor is given by $\beta_{discount} = \lambda_0^{-1}$.

For the above case, all that remains is to determine the largest eigenvalue and to discuss the reduction in accurary introduced by using a sub-optimal weight. The eigenvector associated with λ_0 is $\bar{v}_e(0) = [1\ 1\ \cdots\ 1]^T$. The largest eigenvalue of R_s is the sum of all overlaps with the reference function:

$$\lambda_0 = \sum_i < G_i, G_0 > \approx 1 + 2a + 2b + 4ab. \tag{A.351}$$

Note that λ_0 is dependent on the Gabor filter, and the spatial sampling lattice spacing. Because it is not dependent on the signal information, it can be precomputed.

It is interesting to see how the eigenvector $\bar{v}_e$ compares with the actual vectors $\bar{v}_1 \cdots \bar{v}_6$. It can be seen that $\bar{v}_1$ and $\bar{v}_2$ are scalar multiples of the eigenvector when the normal image direction is the same for all sample points. In such a case, W_{sub} will produce the optimal estimates of T_x and T_y. Since R_s is a local operator, the condition stated previously can be made less restrictive: the normal image direction need only be constant within a small region about the reference point. W_{sub} will produce near optimal estimates of Ω_x and Ω_y if the normal direction is locally constant and if the field of view is small ($\frac{\hat{x}}{z_f}$ and $\frac{\hat{y}}{z_f} \approx 0$). The estimates of T_z and Ω_z will not be optimal. Since R_s is local, the estimates T_z and Ω_z should be reasonably close to optimal.

In this work, not every sample is used. Only significant features are incorporated into the weighted least square estimate of the inter-frame sensor motion. As a result, the matrix R_s must be altered so that it contains only the overlap of significant (active) features; all other elements of R_s must be set to zero. The lower bound obtained using λ_0 will be very conservative in a region containing a sparse number of features. Instead of using the largest eigenvalue λ_0, which is define globally, the approximate local bound for the spatial neighbourhood about the reference G_0 can be used. The local bound for G_i is approximated by

$$\lambda_{local}(i) \approx \sum_{active(j)} < G_j, G_i >, \tag{A.352}$$

where the subscript $active(j)$ indicates that the sum includes only the overlaps between active (significant) features. The local discount factor for the measurement $V_n(i)$ is

$$\beta_{discount} = \frac{1}{\lambda_{local}(i)}. \tag{A.353}$$

Appendix B

Least Square Estimate of Extended Sensor Translation

This appendix provides details of how the inter-frame sensor motion is used to update the estimate of the extended sensor translation. The least square solution of the extended sensor translation and the transient inter-frame rotations over an image sequence is given by

$$\begin{bmatrix} \bar{T}_{sen} \\ \bar{\Omega}(0) \\ \vdots \\ \bar{\Omega}(n) \end{bmatrix} = \begin{bmatrix} A_q & B_q \\ B_q^T & C_q \end{bmatrix}^{-1} \begin{bmatrix} D_p \\ E_p \end{bmatrix}, \tag{B.354}$$

where $A_q = \sum_i Q_a(i)$, $B_q = [Q_b(0) \cdots Q_b(n)]$,

$$C_q = \begin{bmatrix} Q_c(0) & 0 & \cdots & 0 \\ 0 & Q_c(1) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Q_c(n) \end{bmatrix}, \tag{B.355}$$

$D_p = \sum_i \bar{p}_a(i)$, and $E_p = [\bar{p}_b(0) \cdots \bar{p}_b(n)]^T$. Using the block matrix inversion, we get

$$\begin{bmatrix} A_q & B_q \\ B_q^T & C_q \end{bmatrix}^{-1} = \begin{bmatrix} \Delta^{-1} & -\Delta^{-1} B_q C_q^{-1} \\ -C_q^{-1} B_q^T \Delta^{-1} & C_q^{-1} + C_q^{-1} B_q^T \Delta^{-1} B_q C_q^{-1} \end{bmatrix}, \tag{B.356}$$

where

$$\Delta = A_q - B_q C_q^{-1} B_q^T. \tag{B.357}$$

The extended sensor translation is given by

$$\bar{T}_{sen} = \Delta^{-1}(D_p - B_q C_q^{-1} E_p). \tag{B.358}$$

Note that

$$B_q C_q^{-1} E_p = \sum_i Q_b(i) Q_c^{-1}(i) \bar{p}_b(i), \tag{B.359}$$

$$B_q C_q^{-1} B_q^T = \sum_i Q_b(i) Q_c^{-1}(i) Q_b^T(i), \tag{B.360}$$

and

$$\Delta = \sum_i [Q_a(i) - Q_b(i) Q_c^{-1}(i) Q_b^T(i)]. \tag{B.361}$$

Thus,

$$\bar{T}_{sen} = \Delta^{-1} \sum_i [\bar{p}_a(i) - Q_b(i) Q_c^{-1}(i) \bar{p}_b(i)]. \tag{B.362}$$

Appendix C

The Effect of Camera Uncertainty on Collision Parameters

This appendix discusses the effect of uncertainties in the camera parameters and the stereo setup on the estimates of depth, motion, and the collision parameters. The camera parameter uncertainties include an incorrect focal length, pixel scaling errors in the $\hat{x}$ and $\hat{y}$ directions, and lens distortion. Uncertainties in the stereo setup include mismatches of the focal lengths; baseline separation errors; and non-parallel camera configurations.

C.1 Incorrect Focal Length

An incorrect focal length causes errors in the estimated depth, and sensor/object translation. The error in depth, due to an incorrect focal length, is given by

$$\delta z = z\frac{\delta z_f}{z_f}. \tag{C.363}$$

The error in sensor translation is given by

$$\delta \bar{T} = -\bar{T}\frac{\delta z_f}{z_f}. \tag{C.364}$$

Object translation is affected in a similar manner. In each case, the error is a scale factor of the actual value. Note that the speed of the translation is altered by δz_f, but not the direction.

The collision parameters are unaffected by incorrect, matched focal lengths. The collision parameters can be expressed in terms of image measurements, independent of the z_f:

$$t_{col} = -\frac{z}{\dot{z}} = \frac{ds}{V_{\hat{x},L} - V_{\hat{x},R}}, \tag{C.365}$$

$$x_{col} = x + \dot{x}\, t_{col} = (\hat{x}_{av} - \hat{x}_{foe})\frac{b}{ds}, \tag{C.366}$$

where $\hat{x_{av}} = 0.5(\hat{x}_L + \hat{x}_R)$.

C.2 Pixel Scaling Errors

The horizontal and vertical pixel sampling intervals are the distances between adjacent pixels on the CCD array. The horizontal and vertical pixel scale factors, denoted by $s_{\hat{x}}$ and $s_{\hat{y}}$, respectively, convert the pixel coordinates into physical distances (in cm):

$$s_{\hat{x}} = \frac{x_{CCD}}{n_{\hat{x}}}, \tag{C.367}$$

$$s_{\hat{y}} = \frac{y_{CCD}}{n_{\hat{y}}}, \tag{C.368}$$

where (x_{CCD}, y_{CCD}) are the physical dimensions of the the CCD array (in cm), and ($n_{\hat{x}}$, $n_{\hat{y}}$) are the number of pixels along the $\hat{x}$- and $\hat{y}$-axes. The ratio of the vertical and horizontal pixel scale factors is given by

$$r_{pix} = \frac{s_{\hat{y}}}{s_{\hat{x}}}. \tag{C.369}$$

This section considers the effect two types of pixel scaling errors: symmetric pixel scaling errors, which have the correct r_{pix}; and asymmetric pixel scaling errors, which alter r_{pix}.

Asymmetric pixel scaling errors affect the estimated motion and collision parameters. Consider the case where nominal value of $s_{\hat{x}}$ is correct, but the nominal value of $s_{\hat{y}}$ is too large. Because $s_{\hat{x}}$ is correct, the depth estimates will not be affected. The large $s_{\hat{y}}$ will distort the image velocity field by over-estimating $V_{\hat{y}}$. The effect on the sensor motion parameters will depend on the distribution of features (both position and normal direction) in the image. In most cases, $|T_z|$ will be over-estimated; thus, the time-to-collision of an object being approached by forward translating sensors will be under-estimated.

A horizontal scaling error can be introduced by the frame grabbing process. The CCD camera converts the image into an analog output which is redigitized by an image acquisition board. A horizontal scaling error can occurred due to mismatches between the clock frequencies of the camera and acquisition board. The error in $s_{\hat{x}}$ will have a similar effect as an incorrect focal length: it will alter the depth and motion estimates. The asymmetry in the pixel scale error will alter the time-to-collision and further altering the motion estimate.

Symmetrical pixel scaling errors will have no effect on the time-to-collision and point-of-collision.

C.3 Lens Distortion

This section discusses the effect of radial lens distortion on the depth, motion, and collision parameter estimates. The radial lens distortion can be modelled as

$$\hat{x}_d = \frac{\hat{x}}{1 + k_d(\hat{x}^2 + \hat{y}^2)}, \tag{C.370}$$

$$\hat{y}_d = \frac{\hat{y}}{1 + k_d(\hat{x}^2 + \hat{y}^2)}, \tag{C.371}$$

where $(\hat{x}_d,\ \hat{y}_d)$ represents the image coordinate of the distorted image. In this section, only the changes in the $\hat{x}$ coordinate are considered. For convenience, the distortion model is modified:

$$\hat{x} = \hat{x}_d(1 + k\hat{x}_d^2). \tag{C.372}$$

The lens distortion will alter the measured depth. If the distortion is not compensated, the measured depth is given by

$$z_m = \frac{z_f b}{\hat{x}_{d,L} - \hat{x}_{d,R}}, \tag{C.373}$$

where $\hat{x}_{d,L}$ and $\hat{x}_{d,R}$ are the distorted image coordinates for the left and right images. The actual depth, z, (for parallel stereo cameras) is obtained by solving the following equation:

$$\hat{x}_{d,L}(1 - k\hat{x}_{d,L}^2) - \hat{x}_{d,R}(1 - k\hat{x}_{d,R}^2) = \frac{z_f b}{z}. \tag{C.374}$$

The relationship between the measured and actual depth is given by

$$z_m = k_z z, \tag{C.375}$$

where

$$k_z = 1 + k(\hat{x}_{d,L}^2 + \hat{x}_{d,L}\hat{x}_{d,R} + \hat{x}_{d,R}^2). \tag{C.376}$$

It can be seen that the peripheral depth measurements will over-estimate the actual depth.

The lens distortion will also affect the image velocity measurements. The velocity in the distorted image plane will be reduced compared to the undistorted image velocity:

$$V_{\hat{x},d} = k_v V_{\hat{x}}, \tag{C.377}$$

where

$$k_v = [1 + 3k\hat{x}_d^2]^{-1}. \tag{C.378}$$

The effect of lens distortion on the estimated sensor motion is difficult to model. The sensor motion is estimated by combining a set of image measurements. Consider the case of forward axial sensor motion, which produces an outward flow pattern from the image origin. The radial distortion will affect each image measurement: it will reduce the image velocity by k_v, it will reduce the $\hat{x}$ position by $(1 + k\hat{x}^2)$, and it will increase the depth by k_z. If the features are evenly distributed, T_z will be over-estimated and T_x will be approximately zero. If the features are clustered to the left or right of the image origin, biases in T_x will appear.

The effect of lens distortion on the time-to-collision varies depending the object's position and motion. Consider the case of data set 2 where the sensor is forward translating ($T_z > 0$) and the objects have no axial translation ($\dot{z}_{obj} = 0$). The time-to-collision will be under-estimated near the image origin, and over-estimated at the periphery.

In experiments 4 and 5, the object is moving along the x-axis. Because of the lens distortion, the estimated object motion ($\dot{x}_{m,obj}$, $\dot{z}_{m,obj}$) (obtain from the stereo image velocity) will have a directional bias:

$$\dot{x}_{obj} \approx \dot{x}_{m,obj} - \dot{z}_{m,obj} \frac{4k\hat{x}_d^3}{k_z^2 z_f}. \tag{C.379}$$

It can be seen that the directional bias decrease as the object approaches the image origin.

C.4 Mismatched Focal Lengths

This section investigates the effect of mismatches in the focal length for the left and right cameras on the measured depth and motion. For the case of parallel cameras, we have

$$\frac{z}{z_{f(L)}}\hat{x}_L - \frac{z}{z_{f(R)}}\hat{x}_R = b, \tag{C.380}$$

where $z_{f(L)}$ and $z_{f(R)}$ are the focal lengths of the left and right cameras, respectively. The relationship between the actual and measured depth is given by

$$z^{-1} = z_{meas}^{-1} - \delta z_{f(L,R)} \frac{1}{z_f b} \frac{\hat{x}_{av}}{z_f}, \tag{C.381}$$

where $\delta z_{f(L,R)} = z_{f(L)} - z_{f(R)}$, $z_f = 0.5[z_{f(L)} + z_{f(R)}]$, and $\hat{x}_{av} = 0.5[\hat{x}_L + \hat{x}_R]$. The mismatched focal lengths produces an apparent gradient along the $\hat{x}$-axis for the disparity measurements.

The measured direction of sensor translation will be altered by mismatches in the focal lengths. Consider the case of a forward translating sensor. If the features are evenly

distributed, the measured direction of translation will be biased towards the camera with the smaller focal length. This bias will affect the accuracy of the measured point-of-collision.

C.5 Baseline Separation Errors

The baseline separation is an important parameter for converting the disparity into depth. Errors in the measured baseline separation (b_m) will affect the estimate of depth, translational motion, and the point-of-collision.

The error in depth is

$$\delta z = z\frac{\delta b}{b}, \tag{C.382}$$

where $\delta b = b_m - b$. Consider the case where b_m is too large. The depth will be over-estimated by a factor $\frac{\delta b}{b}$. Since the image velocity measurements are not affected, the translation estimates will be over-estimated by the same factor. The time-to-collision will be unaffected. The point-of-collision will be over-estimated by $\frac{\delta b}{b}$.

C.6 Non-parallel Camera Configurations

This section discusses the effect of non-parallel camera configurations on the estimated depth, motion, and collision parameters. In chapter 2, a first-order compensation, (2.57), is described that transforms a convergent stereo configuration into a parallel approximation. The effect of errors in the compensation term $\Delta\beta$, as well as higher-order effects associated with (2.57), are discussed.

The measured depth, which incorporates the first-order compensation, is given by

$$z_m = \frac{z_f b}{ds + \Delta\beta_m\, z_f}. \tag{C.383}$$

If $\Delta\beta_m$ is less than the actual $\Delta\beta_m$, both the depth and translational motions will be under-estimated. Note that an incorrect $\Delta\beta_m$ has a larger effect on distant objects than close ones, causing a distortion along the z_m-axis. If a scene has a large depth gradient across the camera's field of view, the z_m distortion will alter the estimated direction of the sensor translation and the estimated point-of-collision.

The measured direction of object translation is not affected by an incorrect $\Delta\beta_m$ (the depth variation of an object is small). However, an incorrect $\Delta\beta_m$ will cause the measured speed of the object to change as it moves in the z direction. If $\Delta\beta_m$ is less than the actual value, then the measured speed will decrease as the object approaches the cameras. The measured speed of the object will become more accurate as the object approaches the camera. The time-to-collision will not be affected.

The first-order compensation is not exact. Even if the compensation parameters, $\Delta\beta$ and $\Delta\gamma$, correctly selected, the measured depth will be distorted. If $\beta_L = -\beta_R$ and $\gamma_L = -\gamma_R$, the relationship between the actual and measured depth (after the first-order compensation) is given by

$$z_{meas} = z + (k_{\hat{x}}\,\Delta\beta)x - (k_{\hat{x}}\,\Delta\gamma)y, \tag{C.384}$$

where

$$k_{\hat{x}} = \frac{\hat{x}_L + \hat{x}_R}{2(\hat{x}_L - \hat{x}_R + \Delta\beta\, z_f)}. \tag{C.385}$$

This relationship can be written in terms of image coordinates and depth:

$$z^{-1} = z_{meas}^{-1} + \frac{\hat{x}_{ave}^2}{z_f^2}\frac{\Delta\beta}{b} - \frac{\hat{x}_{ave}\hat{y}_{ave}}{z_f^2}\frac{\Delta\gamma}{b}, \tag{C.386}$$

where $\hat{x}_{ave} = 0.5(\hat{x}_L + \hat{x}_R)$ and $\hat{y}_{ave} = 0.5(\hat{y}_L + \hat{y}_R)$. It can be seen that camera convergence, $\Delta\beta > 0$, causes a parabolic distortion. A differential tilt, $\Delta\gamma \neq 0$, causes a distortion that makes vertical object appear slanted (in depth).

For the case of a forward translating sensor viewing well-distributed features, the parabolic distortion causes T_z to be over-estimated; the slanting introduces a T_y bias. In the presence of a parabolic distortion, the time-to-collision will be under-estimated near the image origin, and over-estimated near the periphery. The slanting distortion will affect the point-of-collision.

C.7 Summary

Two types of errors have been discussed in this appendix: scale factor errors and distortion errors. The scale factor errors include: an incorrect (matched) focal length; symmetric pixel scaling errors; and baseline separation errors. Each error alters the measured depth by a scale factor. None of the scale factor errors affect the time-to-collision. The incorrect focal length and the symmetric pixel scaling errors do not affect the point-of-collision. An incorrect baseline separation alters the point-of-collision by a scale factor (no sign changes). Thus, scale factor errors have little effect on obstacle detection.

The distortion errors include: asymmetric pixel scaling errors, radial lens distortion, mismatched focal lengths, and non-parallel camera configurations. With the exception of the asymmetric pixel scaling errors, the distortion errors affect the depth estimates. In addition to altering the estimated depth, the ordering of feature along the z-axis is often changed. As a result, the effect of the distortion errors on the estimated sensor motion and collision parameters is dependent on the distribution of features in the scene (distribution in image position, normal direction, and depth). The distortion errors, combined with the clustering of features away from the image origin, introduce biases in the estimated motion which affect both the estimated point-of-collision and time-to-collision. Thus, distortion errors have a negative effect on obstacle detection.

Bibliography

[1] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A*, vol. 2, pp. 284-299, 1985.

[2] G. Adiv, "Determining three-dimensional motion and structure from optical flow generated by several moving objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 7, no. 4, pp.384-401, 1985.

[3] G. Adiv, "Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 5, 1989.

[4] J. S. Albus, "Outline for a theory of intelligence," *IEEE Trans. Sys. Man Cybern.*, vol. 21, no. 3, 1991.

[5] J. K. Aggarwal and N. Nandhakumar, "On the Computation of Motion from sequences of images—A review," *Proc. IEEE*, vol. 76, no. 8, pp. 917-935, 1988.

[6] N. Ayache and O. D. Faugeras, "Maintaining representations of the environment of a mobile robot," *IEEE Trans. on Robotics and Automation*, vol. 5, no. 6, pp.804-819, 1989.

[7] H. H. Baker and R. C. Bolles, "Generalizing epipolar-plane analysis on the spatiotemporal surface," *Int. J. Comp. Vision*, vol. 3, pp. 33-49, 1989.

[8] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane analysis: an approach to determining structure from motion," *Int. J. Comp. Vision*, vol. 1, pp. 7-55, 1987.

[9] R. N. Braithwaite, "The use of the Gabor expansion in computer vision systems," Master thesis, Dep. Elec. Eng., Univ. of British Columbia, Vancouver, Canada, 1989.

[10] R. N. Braithwaite and M. P. Beddoes, "Obstacle detection for an autonomous vehicle," in *proceedings, Third Conf. on Military Robotic Applications, Medicine Hat, AB, Canada, 1991.*

[11] R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering.* New York, NY: Wiley and Sons, 1983.

[12] W. Burger and B. Bhanu, "Estimating 3-D egomotion from perspective image sequences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 11, pp. 1040-1058, 1990.

[13] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, vol. 2, no. 7, pp. 1160-1169, 1985.

[14] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 7, pp. 1169-1179, 1988.

[15] J. G. Daugman, "Entropy reduction and decorrelation in visual coding by oriented neural receptive fields," *IEEE Trans. Biomedical Eng.*, vol. 36, no. 1, pp. 107-114, 1989.

[16] E. D. Dickmanns and A. Zapp, "A curvature-based scheme for improving road vehicle guidance by computer vision," *Proc. SPIE Mobile Robots Conference*, vol. 727, pp. 161-168, 1986.

[17] E. D. Dickmanns, "4D-dynamic scene analysis with integral spatio-temporal models," *Proc. 4th International Symposium on Robotics Research*, Santa Cruz, Aug. 1987, pp. 311-318.

[18] E. D. Dickmanns and B. D. Mysliwetz, "Recursive 3-D road and relative ego-state recognition," *Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 2, pp. 199-213, 1992.

[19] E. D. Dickmanns, B. Mysliwetz, and T. Christians, "An integrated spatio-temporal approach to automatic visual guidance of autonomous vehicles," *IEEE Trans. Systems Man Cybern.*, vol. 20, no. 6, pp. 1273-1284, 1990.

[20] D. J. Fleet and A. D. Jepson, "Hierarchical construction of orientation and velocity selective filters," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 3, pp. 315-325, 1989.

[21] D. J. Fleet, A. D. Jepson, and M. R. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understanding*, vol. 53, no. 2, pp. 198-210, 1991.

[22] D. J. Fleet and A. D. Jepson, "Computation of normal image velocity from local phase information," in *proceedings, IEEE CVPR, San Diego, 1989*, pp. 379-386.

[23] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comp. Vision*, vol. 5, no. 1, pp. 77-104, 1990.

[24] D. Gabor, "Theory of communication," *J. Inst. Elec. Eng.*, vol. 93, pp. 429-457, 1946.

[25] J. J. Gibson, "A theoretical field-analysis of automobile-driving," in E. Reed and R. Jones (Eds.), *Reasons for realism.* Hillsdale, N.J.: Erlbaum, 1982.

[26] B. Y. Hayashi and S. Negahdaripour, "Direct motion stereo: recovery of observer motion and scene structure," in *proceedings, IEEE Third Int. Conf. Comp. Vision*, 1990, pp. 446-450.

[27] D. J. Heeger, "Optical flow from spatiotemporal filters," *Proc. First Int. Conf. Comp. Vision*, 1987, pp. 181-190.

[28] D. J. Heeger and A. Jepson, "Simple method for computing 3D motion and depth," in *proceedings, IEEE Third Int. Conf. of Comp. Vision, 1990*, pp. 96-100.

[29] D. J. Heeger and G. Hager, "Egomotion and the stabilized world," in *proceedings, Second Int. Conf. on Comp. Vision, Tampa, FL, 1988*, pp. 435-440.

[30] J. Heel and S. Negahdaripour, "Time-sequential structure and motion estimation without optical flow," in *proceedings, SPIE Sensing and Reconstruction of Three-Dimensional Objects and Scenes*, vol. 1260, pp. 50-61, 1990.

[31] B. K. P. Horn and E. J. Weldon Jr., "Direct methods for recovering motion," *Int. J. of Comp. Vision*, vol. 2, pp. 51-76, 1988.

[32] B. K. P. Horn, *Robot Vision.* MIT Press, 1986.

[33] E. Ito and J. Aloimonos, "Determining three dimensional transformation parameters from images: theory," *IEEE 1987 Int. Conf. on Robotics and Automation*, 1987, pp. 57-61.

[34] M. R. Jenkin and A. D. Jepson, "Response profiles of trajectory detectors," *IEEE Trans. Systems Man Cybern.*, vol. 19, no. 6, pp. 1617-1622, 1989.

[35] A. D. Jepson and M. R. Jenkin, "The fast computation of Disparity from phase differences," in *proceedings, IEEE CVPR, San Diego, 1989*, pp. 398-403.

[36] K. Kuhnert, "A vision system for real time road and object recognition for vehicle guidance," *Proc. SPIE Mobile Robots Conference*, vol. 727, pp. 267-272, 1986.

[37] D. N. Lee, "A theory of visual control of braking based on information about time-to-collision," *Perception*, vol. 5, pp. 437-459, 1976.

[38] D. Marr, *Vision.* San Francisco, CA: Freeman, 1982.

[39] L. Matthies, R. Szeliski, and T. Kanade, "Kalman filter-based algorithms for estimating depth from image sequences," CMU-CS-87-185, 1987.

[40] L. H. Matthies, *Dynamic Stereo Vision.* PhD thesis, CMU-CS-89-195, 1989.

[41] L. Matthies, "Stereo vision for planetary rovers: stochastic modeling to near real-time implementation," JPL D-8131, 1991.

[42] L. Matthies and S. A. Shafer, "Error modeling in stereo navigation," vol. 3, no. 3, pp. 239-248, 1987.

[43] H. P. Moravec, "The Stanford Cart and the CMU Rover," *Proc. IEEE*, vol. 71, no. 7, pp. 872-884, 1983.

[44] B. D. Mysliwetz and E. D. Dickmanns, "Distributed scene analysis for autonomous road vehicle guidance," *Proc. SPIE Mobile Robots II*, vol. 852, pp. 72-79, 1987.

[45] S. Negahdaripour and B. K. P. Horn, "Direct passive navigation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, 1987.

[46] R. C. Nelson and J. Aloimonos, "Obstacle avoidance using flow field divergence," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 10, pp. 1102-1106, 1989.

[47] T. D. Sanger, "Stereo disparity using Gabor filters," *Biol. Cybern.*, no. 59, pp. 405-418, 1988.

[48] C. Thorpe, M. H. Hebert, T. Kanade, and S. A. Shafer, "Vision and navigation for the Carnegie-Mellon Navlab," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, no. 3, pp. 362-373, 1988.

[49] M. A. Turk, D. G. Morgenthaler, K. D. Gremban, and M. Marra, "VITS—A vision system for autonomous vehicle navigation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, no. 3, pp. 342-361, 1988.

[50] A. B. Watson and A. J. Ahumada, Jr., "Model of human visual motion sensing," *J. Opt. Soc. Amer. A*, vol. 2, pp. 322-341, 1985.

[51] A. M. Waxman and J. H. Duncan, "Binocular image flows: steps towards stereo-motion fusion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, no. 6, pp. 715-729, 1986.

[52] J. Weng, T. S. Huang, and N. Ahuja, "Motion and structure from two perspective views: algorithms, error analysis, and error estimation," *IEEE Trans. Pattern Anal, Machine Intell.*, vol. 11, no. 5, pp. 451-476, 1989.

[53] H. Wunsche, "Detection and control of mobile robot motion by real-time computer vision," *Proc. SPIE Mobile Robots Conference*, vol. 727, pp. 100-109, 1986.

[54] M. Yamamoto, "A general aperture problem for direct estimation of 3-D motion parameters," *IEEE Trans. Pattern Anal. Machine Intell.* vol. 11, no.5, 1989.