Disparity Estimation and Direct Passive Navigation using Stereo Gabor Filters

R. Neil Braithwaite
College of Engineering
University of California at Riverside
Riverside, CA 92521, USA
neilb@constitution.ucr.edu

Michael P. Beddoes
Department of Electrical Engineering
University of British Columbia
Vancouver, B.C., Canada
mikeb@ee.ubc.ca

Abstract

This paper discusses the use of stereo Gabor filters for the extraction of scene structure and ego-motion. Gabor filters are used to estimate stereo disparity and normal image velocity. The stereo and temporal correspondence problems are reduced in severity by using multi-scale prediction and phase-based measurements. Scene structure is estimated from the disparity. Ego-motion is estimated by combining image measurements using direct passive navigation. Experimental results demonstrate the utility of the approach.

1 Introduction

Stereo vision can be used as a primary sensing system for the interpretation of scene structure and ego-motion. The advantages of using a camera as a primary sensor are as follows: high resolution, commercial availability, low power consumption, and passive operation. One of the disadvantages associated with camera-based systems, particularly stereo vision, is the feature correspondence problem. This paper discusses the use of multi-scale Gabor filtering to reduce the serverity of the correspondence problem, and to estimate stereo disparity and the normal component of image velocity. Disparity and normal image velocity are subsequently combined using direct passive navigation to estimate the motion of the stereo cameras.

The correspondence problem is a hard perennial problem related to the matching of features or intensity patterns in image pairs that belong to the same scene entity. The correspondence problem appears in two forms for stereo motion sequences. The *stereo correspondence problem* is associated with the matching of features across stereo image pairs. The *temporal correspondence problem* is associated with matching features in successive images of an image sequence. There are three main challenges to overcoming the stereo and temporal correspondence problems: viewpoint/scale distortion, multiple candidate matches, and missing parts.

The Gabor filter approach presented in this paper can be described as coarse matching of features from bandpass filtered images followed by a refinement stage based on phase correlation. Gabor filters are useful because they are robust to image deformations caused by viewpoint differences, and their phase response can estimate image shifts to sub-pixel accuracy. When implemented as a multi-scale filter set, Gabor filters can simplify feature correspondence, reducing the likelihood of multiple candidate matches. The magnitude response can be used to test candidate matches thereby avoiding problems associated with missing parts.

The scope of this paper is limited to static scenes. However, the algorithm presented in this paper is the basis of a larger system [2]. Extensions of this algorithm include the integration of sensor translation over long stereo image sequences [3], the compensation of transient sensor rotations [3], the estimation of object trajectories from a moving platform [3], and the prediction of collisions of moving objects with a moving sensor.

The outline of this paper is as follows. Section 2 reviews the Gabor filter and defines local attributes that are used for selecting features and measuring image displacements. Section 3 discusses the use of these local attributes to measure disparity and normal image velocity, and to estimate sensor motion. Section 4 covers techniques for overcoming the feature correspondence problems. Results for real image sequences are shown in section 5.

The remainder of the introduction gives a review of technical prerequisites.

Disparity Measurements – Once matching projections, or corresponding image features, are found in the left and right images, the apparent shift, or disparity, can be measured. In this paper, it is assumed that the camera setup and calibration produces an ideal epipolar geometry which constrains the direction of the disparity to follow the horizontal axes of the two images. As a result, disparity is reduced to a scalar term. This ideal epipolar geometry requires that the stereo cameras are setup such that their optical axes are parallel, their image coordinate axes are aligned, and any lens distortions are compensated.

Depth Estimation – Depth estimation uses the stereo disparity and knowledge of the camera setup. If, in addition to the ideal epipolar geometry, it is assumed that the camera focal lengths are matched, then the three-dimensional position is calculated using

$$[x\ y\ z] = \frac{b}{d_{\hat{x}}}[\frac{(\hat{x}_L + \hat{x}_R)}{2}\ \hat{y}\ z_f], \tag{1}$$

where (x, y, z) represents the horizontal, vertical, and depth axes; $(\hat{x}_L, \hat{y})$ and $(\hat{x}_R, \hat{y})$ are the image coordinates for the left and right images, respectively; b is the baseline separation between cameras; z_f is the camera focal length; and $d_{\hat{x}} = (\hat{x}_L - \hat{x}_R)$ is the stereo disparity along the horizontal image axis.

Camera Motion and Image Velocity – A point $P(x, y, z)$ in the field-of-view of a camera is projected onto the image plane at coordinates $(\hat{x}, \hat{y})$:

$$[\hat{x}\ \hat{y}] = \frac{z_f}{z}[x_c\ y], \tag{2}$$

where

$$x_c = \begin{cases} x + \frac{b}{2} \text{ for the left camera} \\ x - \frac{b}{2} \text{ for the right camera.} \end{cases} \tag{3}$$

If the cameras move, the image velocity corresponding to a stationary point $P(x, y, z)$ is given by

$$[V_{\hat{x}}\ V_{\hat{y}}]^T = \mathbf{A}(z^{-1})\mathbf{B}(z)\bar{\mathbf{\Theta}}, \tag{4}$$

where $V_{\hat{x}}$ and $V_{\hat{y}}$ are the image velocities in the $\hat{x}$ and $\hat{y}$ directions, respectively; $\bar{\mathbf{\Theta}} = [T_x\ T_y\ T_z\ \Omega_x\ \Omega_y\ \Omega_z]^T$, referred to as the "inter-frame sensor motion," is the translational (T) and rotational (Ω) velocities of the stereo cameras with respect to the ground reference;

$$\mathbf{A}(z^{-1}) = z^{-1}\begin{bmatrix} z_f & 0 & -\hat{x} \\ 0 & z_f & -\hat{y} \end{bmatrix}; \tag{5}$$

and

$$\mathbf{B}(z) = \begin{bmatrix} -1 & 0 & 0 & 0 & -z & y \\ 0 & -1 & 0 & z & 0 & -x \\ 0 & 0 & -1 & -y & x & 0 \end{bmatrix}. \tag{6}$$

Since image velocity is usually measured using local regions of the image, the "aperture problem" [8] will arise. As a consequence, only the component of the image velocity normal to an image contour can be measured. This "normal image velocity," denoted by V_n, is given by

$$V_n = \bar{\mathbf{n}}^T [V_{\hat{x}}\ V_{\hat{y}}]^T, \tag{7}$$

where $\bar{\mathbf{n}} = [\cos\phi_n\ \sin\phi_n]^T$, and ϕ_n is the direction that is normal to the contour relative to the $\hat{x}$-axis. The superscript T indicates transpose.

Direct Passive Navigation – In this work, the inter-frame sensor motion is estimated using a "direct passive navigation" approach [11]. If depth information is available, the sensor motion can be estimated from the set of normal image velocities:

$$\bar{\mathbf{\Theta}} = \mathbf{Q}_{int}^{-1}\bar{\mathbf{p}}, \tag{8}$$

where $\mathbf{Q}_{int} = \sum_i c_i \bar{\mathbf{J}}_i \bar{\mathbf{J}}_i^T$, $\bar{\mathbf{p}} = \sum_i c_i \bar{\mathbf{J}}_i V_n(i)$, $\bar{\mathbf{J}} = \bar{\mathbf{n}}^T \mathbf{A}(z^{-1})\mathbf{B}(z)$, and c_i is a weighting term based on the certainty of $V_n(i)$. $\mathbf{Q}_{int}^{-1}$ is the error covariance matrix for the inter-frame sensor motion estimate, $\bar{\mathbf{\Theta}}$. An expression for the weighting term, c_i, appears in section 3.3.

2 Image Processing

In this work, the Gabor wavelet decomposition is used as a preprocessing stage for both disparity and normal image velocity estimation. The wavelet filter set decomposes the image into a set of bandpass (Gabor) channels with log-polar frequency separation. The Gabor channels are characterized by two responses: (a) a magnitude response that measures localized signal energy, and (b) a phase response that encodes the relational structure of an intensity pattern with respect to its spatial neighborhood. The magnitude response is useful for identifying significant features. The spatial phase gradient of the neighborhood about a feature and the local phase difference with its corresponding feature are used to estimate image displacements to sub-pixel accuracy [3] [5].

The quadrature Gabor filter kernels in a wavelet representation are given by

$$G_+(\hat{x}, \hat{y}, \tilde{\omega}, \tilde{\phi}) = g(\hat{x}, \hat{y})\cos[\tilde{\omega}(\hat{x}\cos\tilde{\phi} + \hat{y}\sin\tilde{\phi})], \tag{9}$$

$$G_-(\hat{x},\hat{y},\tilde{\omega},\tilde{\phi}) = g(\hat{x},\hat{y})\sin[\tilde{\omega}(\hat{x}\cos\tilde{\phi}+\hat{y}\sin\tilde{\phi})], \quad (10)$$

$$g(\hat{x},\hat{y}) = \exp(-\frac{\lambda^2\tilde{\omega}^2(\hat{x}^2+\hat{y}^2)}{4\pi}), \quad (11)$$

where $\tilde{\omega}$ and $\tilde{\phi}$ are the modulation (center) frequency and orientation, respectively, of the Gabor filter kernel (and the Gabor channel); and λ is the ratio of the channel bandwidth and the modulation frequency. The spatial output of each filtered image is typically sub-sampled, producing a spatial sampling lattice whose density is appropriate for the channel bandwidth. The spatially sampled output is referred to as a "Gabor coefficient", and is given by

$$a_+(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) = \iint I(\hat{x},\hat{y})G_+(\tilde{x}-\hat{x},\tilde{y}-\hat{y},\tilde{\omega},\tilde{\phi})\,d\hat{x}\,d\hat{y}, \quad (12)$$

$$a_-(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) = \iint I(\hat{x},\hat{y})G_-(\tilde{x}-\hat{x},\tilde{y}-\hat{y},\tilde{\omega},\tilde{\phi})\,d\hat{x}\,d\hat{y}, \quad (13)$$

where $I(\hat{x},\hat{y})$ is the input image and $(\tilde{x},\tilde{y})$ is the spatial coordinates of the lattice point.

Quadrature Gabor coefficients define a magnitude m and a phase θ:

$$m(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) = [a_+^2(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) + a_-^2(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})]^{0.5}, \quad (14)$$

$$\theta(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) = \arctan\left[\frac{a_-(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})}{a_+(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})}\right]. \quad (15)$$

The spatial phase gradient, denoted by $[\omega_{\hat{x}}\ \omega_{\hat{y}}]$, is used to estimate the mean frequency ω_n and orientation ϕ_n of the signal energy within a bandpass channel:

$$\phi_n(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) = \arctan\left[\frac{\omega_{\hat{y}}(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})}{\omega_{\hat{x}}(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})}\right], \quad (16)$$

$$\omega_n(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) = [\omega_{\hat{x}}^2(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi}) + \omega_{\hat{y}}^2(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})]^{0.5}. \quad (17)$$

In general, the mean frequency and orientation (ω_n,ϕ_n) are close to, but not the same as, the center frequency and orientation $(\tilde{\omega},\tilde{\phi})$, respectively, of the bandpass channel. The finite bandwidth of the channel, which allows this frequency/orientation difference, allows for robust matching of corresponding intensity patterns in successive or stereo images even if the patterns are distorted due to viewpoint changes.

The normal image displacement between successive images, when small, can be measured directly using phase information:

$$V_n(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})\ \Delta t = \frac{\Delta\theta_t(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})}{\omega_n(\tilde{x},\tilde{y},\tilde{\omega},\tilde{\phi})} \text{ when } |V_n\Delta t| \approx 0, \quad (18)$$

where Δt is the temporal sampling interval, and $\Delta\theta_t$ is the phase difference over this interval. For a large image displacement, a coarse alignment of features, or equivalently, a temporal correspondence of lattice points in successive images, is required. Thus, the normal image displacement is given by

$$V_n\Delta t = \Delta\hat{x}_{off}\ \cos\phi_n + \Delta\hat{y}_{off}\ \sin\phi_n + \frac{\Delta\theta_t}{\omega_n}, \quad (19)$$

where $\Delta\hat{x}_{off}$ and $\Delta\hat{y}_{off}$ are the lattice offsets between successive images. Similarly, the disparity is estimated using the phase difference between corresponding lattice points in stereo images:

$$d_{\hat{x}}(\tilde{x}_L,\tilde{y},\tilde{\omega},\tilde{\phi}) = (\tilde{x}_R-\tilde{x}_L) - \frac{\Delta\theta_{L,R}(\tilde{x}_R,\tilde{x}_L,\tilde{y},\tilde{\omega},\tilde{\phi})}{\omega_{\hat{x},ave}(\tilde{x}_R,\tilde{x}_L,\tilde{y},\tilde{\omega},\tilde{\phi})}, \quad (20)$$

where $\Delta\theta_{R,L}$ is the phase difference; and $\omega_{\hat{x},ave}$ is the average frequency of a pattern measured along the epipolar line, as viewed by the left and right Gabor filtered images.

Expected Errors – It is important to model the measurement uncertainty so that the expected errors in stereo disparity and normal image velocity, as well as the error covariance of camera motion, can be estimated. The expected frequency errors along, and orthogonal to, the channel modulation $(\tilde{\phi})$ are measured using phase variances $(\delta\omega_{\hat{x}},\ \delta\omega_{\hat{y}})$. The error in the local phase difference, either $\Delta\theta_t$ or $\Delta\theta_{R,L}$, can be estimated if it is assumed that the error in the Gabor coefficients (a_+,a_-) 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

$$[\delta(\Delta\theta)]^2 \approx \left[\frac{1}{m_0^2}+\frac{1}{m_1^2}\right]\sigma_G^2, \quad (21)$$

where m_0 and m_1 are the local magnitudes at temporal or stereo corresponding lattice points. Note that the expected phase shift error is inversely dependent on the signal-to-noise ratio of the filtered image.

Interesting Features – Certain points within an image will be easier to match than others. These salient image intensity patterns, which are both distinct and robust to image noise, are referred to as "features". These features are high peaks or plateaus in the magnitude response that exhibit local phase stability (see [2] or [5]). Each feature has a low expected error in terms of $\delta\omega_{\hat{x}}$, $\delta\omega_{\hat{y}}$, and $\delta(\Delta\theta)$.

Lattice Spacing – The output of the Gabor channel is spatially subsampled with respect to the original image. In this paper, the spatial sampling lattice is a Cartesian grid that is rotated by $\tilde{\phi}$ so that $\hat{x}$ is aligned with the channel modulation. The spatial sampling intervals, denoted by $\Delta\hat{x}_s$ and $\Delta\hat{y}_s$, are selected as a function of the channel modulation $\tilde{\omega}_k$:

$$\Delta\hat{x}_s = \Delta\hat{y}_s = \frac{\pi}{\tilde{\omega}_k}. \tag{22}$$

3 Using Local Attributes

The magnitude and phase responses from the Gabor filters can be used to estimate disparity, normal image velocity, and sensor motion along with their respective expected errors. Each of these topics are discussed in this section.

3.1 Disparity Estimation

Limitations of the strictly phase-based approach to disparity estimation result in multiple candidate feature matches. This subsection discusses criteria for rejecting incorrect feature matches. The estimation of the expected error in the disparity measurement is also discussed.

Consider an epipolar channel ($\tilde{\phi}_k = 0$) whose modulation frequency is $\tilde{\omega}_k$. The local spatial frequency of a detected feature, measured along the epipolar axis, is denoted by $\omega_{\hat{x}}$. If corresponding lattice points are identified in the left and right images, the relative disparity measured using the phase difference is given by

$$d_{rel} = \frac{\Delta\theta_{L,R}}{\omega_{\hat{x}}}. \tag{23}$$

Since the local phase difference is modulo 2π the measurable disparity is restricted to

$$-\frac{\pi}{\omega_{\hat{x}}} < d_{rel} \leq \frac{\pi}{\omega_{\hat{x}}}. \tag{24}$$

This interval, referred to as the "disparity interval," is dependent on the local frequency of the filtered image. Since the nature of feature selection (high magnitude, stable phase) ensures that $\omega_{\hat{x}} \approx \tilde{\omega}_k$, the disparity interval described by (24) will be very small for high frequency channels.

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

$$E_{offset} = n_o \Delta\hat{x}_s, \tag{25}$$

where n_o is a non-negative integer.

Multiple disparity intervals produce a set of possible disparities: the correct disparity, and many aliased disparities. Aliased disparity estimates must be identified and rejected. There a number of constraints which, if violated, can be used to reject unlikely feature pairings: the depth is positive; the local magnitudes are similar at corresponding points; and the best match should have a small phase shift. Even after applying these constraints, more than one potential match may still exist. The use of a priori prediction to select the correct disparity interval is discussed in section 4.

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

$$\delta d_{\hat{x}} = (E_{offset} - \hat{E}) + \delta d_{rel}, \tag{26}$$

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. Errors in the relative disparity are due primarily 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

$$(\delta d_{rel})^2 \approx (d_{rel}^2)\left(\frac{\delta\omega_{\hat{x}}}{\omega_{\hat{x}}}\right)^2 + \left(\frac{\sigma_G}{\omega_{\hat{x}}}\right)^2 \left[\frac{1}{m_L^2} + \frac{1}{m_R^2}\right]. \tag{27}$$

Equation (27) is used in section 3.3 as the expected disparity error: $(\delta d_{\hat{x}})^2 = (\delta d_{rel})^2$.

3.2 Normal Image Velocity

Multiple candidate matches can also occur when measuring image displacements between features in successive images. Like the disparity, the similar magnitude and the small phase requirements are used to test candidate temporal matches.

The normal image velocity is measured using (19). Assuming that the correct lattice offset ($\Delta\hat{x}_{off}$, $\Delta\hat{y}_{off}$) is selected, the expected error in the normal image velocity and direction are given by

$$(\delta V_n)^2 \approx \left(\frac{\omega_t}{\omega_n}\right)^2\left(\frac{\delta\omega_n}{\omega_n}\right)^2 + \left(\frac{\sigma_G}{\omega_n}\right)^2 \left[\frac{1}{m_1^2} + \frac{1}{m_0^2}\right], \tag{28}$$

$$\delta\phi_n \approx \frac{\delta\omega_{\hat{y}}}{\omega_n}, \tag{29}$$

respectively, where $\delta\omega_n \approx \delta\omega_{\hat{x}}$; $\delta\omega_{\hat{x}}$ and $\delta\omega_{\hat{y}}$ are the expected frequency errors along, and orthogonal to, the direction of channel modulation.

3.3 Direct Passive Navigation

Direct passive navigation (equation (8)) is a weighted least square parameter estimation technique that fits a set of normal image velocity measurements to a rigid-body motion model. The rigid-body model requires the scene structure, which is obtained from disparity measurements. The weighting term associated with a feature i is selected based on the uncertainty in the measured V_n and modeling errors caused by uncertainties in ϕ_n and $d_{\hat{x}}$: that is,

$$c_i^{-1} = (\delta V_n)^2 + \left[\frac{\delta V_n}{\delta \phi_n}\right]^2 (\delta \phi_n)^2 + \left[\frac{\delta V_n}{\delta d_{\hat{x}}}\right]^2 (\delta d_{\hat{x}})^2 + e_o, \tag{30}$$

where e_o is a constant offset error used to limit the effect of inaccuracies in the error estimation, particularly when the expected error is underestimated.

The above weight calculation assumes that each source of error is uncorrelated. For the most part, this assumption is valid except when the spatial lattice is oversampled and the feature density is high. The weight must be attenuated in image regions where the density of features is higher than the minimum complete sampling density (see [2]).

4 Feature Correspondence

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 scene entity. The implementation of a correspondence method, however, is difficult. In the previous section, it was shown how local attributes can be compared to test candidate correspondences. This local test will reject some feature pairings, but there is still the possibility of multiple candidate matches. In this section, a priori information is used to limit the search space such that a single best match or no match is found.

In this work, stereo correspondences are established using a priori information from lower frequency Gabor channels (multi-scale prediction), as well as using a heuristic ordering constraint. *Multi-scale prediction* uses scene structure measured in lower frequency channels to predict the local disparity and to limit the correspondence search space. At each scale, a priori estimates of the disparity, $d_{\hat{x}}$, and the expected squared error, $(\delta d_{\hat{x}})^2$, are interpolated from the lower frequency channels. For each detected feature, the a priori disparity estimate is used to select two candidate epipolar offsets. The magnitude and phase tests (section 3.1) are used to determine which, if either, offset produces a matching feature. If matching features are found, the phase-based (direct) measurements of the disparity (equation (20)) and the associated expected error (equation (27)) replace their respective a priori estimates.

After completing the direct measurements, a "diffusion" stage is applied to the disparity map which spreads the influence of new measurements to neighboring regions of large uncertainty, thereby limiting the disparity gradient where the measurement density is low. In this work, the diffused disparity is estimated from a weighted sum of neighboring lattice points:

$$\bar{d}_{\hat{x}}(0,0) = \frac{\sum_{ij} a_{i,j} d_{\hat{x}}(i,j)}{\sum_{ij} a_{i,j}}, \tag{31}$$

where $-1 < i, j < 1$,

$$a_{i,j}^{-1} = \begin{cases} E[(\delta d_{\hat{x}})^2] & \text{if } i = j = 0 \\ E[(\delta d_{\hat{x}} + e_{drift})^2] & \text{otherwise,} \end{cases} \tag{32}$$

and e_{drift} is the drift penalty. The updated error is set to the minimum error in the neighborhood: that is, $\bar{a}_{0,0} = \min_{ij}[a_{i,j}]$.

The above diffusion is inhibited at stereo features so that direct measurements are unaffected. The drift penalty, e_{drift}, limits the radius of influence of new measurements so that lattice points far from direct measurements retain their a priori disparity estimates. Once the diffusion stage is complete, the disparity and error values are then interpolated and projected to the next higher frequency channel. The prediction stage is then repeated at this new scale. A cross-scale penalty is added to the a priori error estimate to reduce the influence of the lower frequency data in the subsequent diffusion stage.

The density of stereo matches is increased further by enforcing a *heuristic ordering constraint.* If a viewed surface is sufficiently smooth, corresponding features along an epipolar line will appear in the same order in the left and right images [8]. The application of the heuristic ordering constraint is shown in figure 1. The existing correspondences, established by the scale-based matching algorithm, 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 1). This limited space is

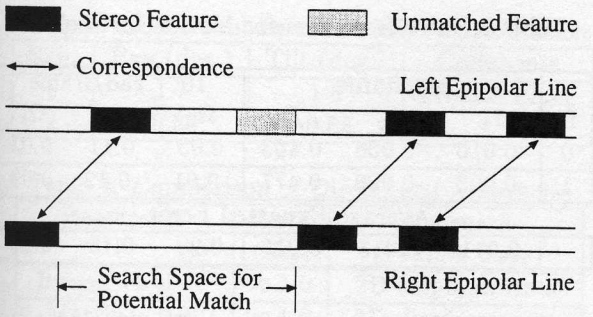

Figure 1: Heuristic Ordering Constraint

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, as in [10]. The utility of the heuristic ordering constraint is its ability to generate correspondences in image regions with large disparity gradients. Such correspondences are sometimes missed by multi-scale prediction. Once the heuristic ordering constraint establishes a correspondence, a temporal-based matching scheme will propagate the correspondence forward in time.

The temporal correspondence requires the coarse prediction of the two-dimensional image velocity (*cf.* $\Delta\hat{x}_{off}$ and $\Delta\hat{y}_{off}$ in equation (19)), not just the normal component. The two-dimensional displacement can be predicted using (4) along with estimates of the sensor motion $\bar{\Theta}$ and the depth z. In this work, the set of candidate matches consists of the four lattice offsets nearest to the prediction $\mathbf{AB}\bar{\Theta}\,\Delta t$. The best lattice offset is determined by applying magnitude and phase tests to the candidate features.

5 Results

In this section, two experiments demonstrate the robustness of the stereo Gabor filter-based disparity algorithm and direct passive navigation algorithm to various scene structures. The objectives of these experiments are to determine to accuracy of the image displacement measurements (disparity and normal image velocity estimation) as well as the accuracy of the estimated direction of sensor translation. Comparisons with other published results appear at the end of this section.

Experiment 1: In this experiment, stereo cameras move towards a stationary poster of a tiger's face. The image sequence contains three stereo pairs, one of which is shown in figure 2. The poster

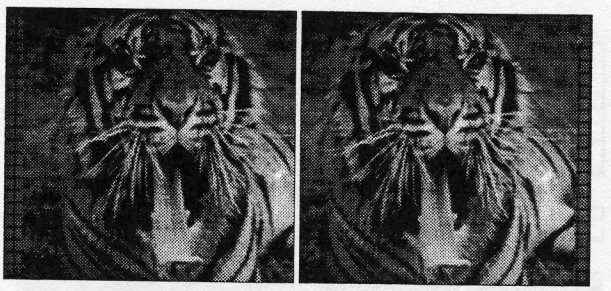

Figure 2: Experiment 1, Stereo Images of Poster, *(left)* Left Image, *(right)* Right Image.

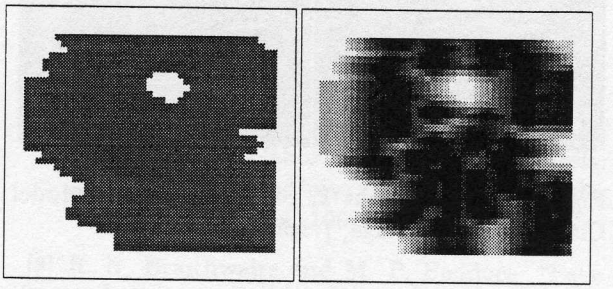

Figure 3: Experiment 1, *(left)* Interpolated Disparity. The minimum (white) and maximum (black) responses are 40 pixels and 60 pixels, respectively. *(right)* Uncertainty. Light regions have large uncertainties. Dark regions denote direct disparity measurements. A region that is white in both the disparity and uncertainty maps indicate a disparity estimate whose uncertainty is too large to be meaningful.

is a planar surface whose normal is parallel to the z-axis; that is, the scene structure has one depth.

The interpolated disparity and the associated uncertainty for the middle of three epipolar channels is shown in figure 3. 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.

Direct passive navigation provides an estimate of the inter-frame sensor motion parameters (T_x, T_y, T_z, Ω_x, Ω_y, Ω_z), and the expected errors (estimated from the error covariance matrix, $\mathbf{Q}_{int}^{-1}$). The two inter-frame sensor motions and the average expected error appear in table 1. 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 the expected errors. The inter-frame sensor motions are also consistent with the axial motion; the estimated direction of translation ($\frac{T_x}{T_z}$, $\frac{T_y}{T_z}$) and the rotation (Ω_x, Ω_y, Ω_z)

Table 1: Inter-frame Sensor Motion for Exp. 1

	cm/frame			10^{-3} rad/frame		
fr	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0	0.045	0.042	0.719	0.86	-1.23	0.04
1	-0.002	0.004	0.720	0.06	-0.14	-0.05
Average Expected Error (±)						
	0.134	0.195	0.016	4.15	2.86	0.36

Figure 4: Experiment 2, Stereo Images of Model City, *(left)* Left Image, *(right)* Right Image.

are approximately zero (within the expected errors). The average error in the direction of translation along the x- and y-axes (pan and tilt) are 0.030 radians (1.7 degrees) and 0.031 radians (1.8 degrees), respectively. These results are surprisingly accurate considering the inherent translation-rotation ambiguity that exists when viewing frontal planes [1].

Experiment 2: In this experiment, stereo cameras move towards a stationary model of a city. A stereo pair from the image sequence is shown in figure 4. The scene structure has a variety of depths, including some large depth gradients. The images also contain specular reflections.

The interpolated disparity and its uncertainty for the middle and highest frequency channels are shown in figure 5. In this example, the lowest frequency channel (not shown) provides only a vague

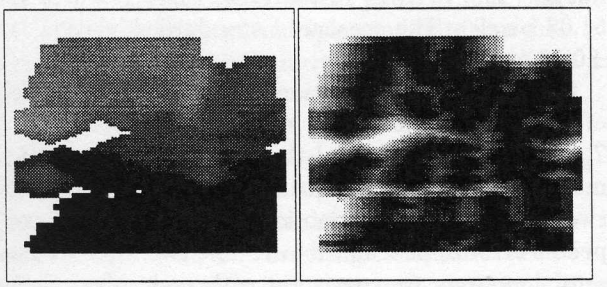

Figure 5: Exp. 2, *(left)* Interpolated Disparity. The minimum (white) and maximum (black) responses are 10 pixels and 45 pixels, respectively. *(right)* Uncertainty. Light regions have large uncertainties. Dark regions denote direct disparity measurements.

Table 2: Inter-frame Sensor Motion for Exp. 2

	cm/frame			10^{-3} rad/frame		
fr	T_x	T_y	T_z	Ω_x	Ω_y	Ω_z
0	-0.010	-0.006	0.463	0.03	0.11	0.03
1	-0.003	-0.009	0.477	-0.01	-0.12	0.09
Average Expected Error (±)						
	0.011	0.013	0.016	0.20	0.15	0.29

description of the scene structure. The scene details become discernible in the higher frequency channels. There is a total of 274 stereo feature pairs across the three epipolar channels.

The inter-frame sensor motions and the average expected error appear in table 2. The inter-frame sensor motions are consistent with the axial motion (within the expected errors). The average errors in the pan and tilt directions of translation are $(-0.014, -0.015)$ radians, or $(-0.80, -0.88)$ degrees.

Comparisons: The sliding stereo algorithm of Matthies et al, described in [9], was tested on the same tiger poster as that used in experiment 1. Matthies' algorithm produced an RMS disparity error of ± 0.12 pixels, which is slightly better than the RMS error of ± 0.16 pixels in experiment 1. The improved performance of Matthies' algorithm may be due to its use of eleven images along the epipolar line, instead of two images. In this sliding stereo approach, the disparity between image pairs is approximately 1 pixel. In contrast, the disparity in experiment 1 is 51.07 pixels.

An alternative performance measure is the RMS error in depth as a percentage of the actual value. Using this measure, the results of experiment 1 are very respectable: the RMS error in depth is 0.3 percent of the average value (46.30 ± 0.14 cm). Matthies' sliding stereo algorithm, after processing eleven images, produces a RMS error that is 0.5 percent of the average depth. Thus, the depth results of experiment 1 are better than those reported in [9] and better than a laser range-finder with 8 bits of precision.

Weng et al [12] judge the accuracy of the image velocity using the RMS difference between the measured image velocity field and the field predicted by the inter-frame sensor motion. For the real image sequence found in [12], the RMS error is 0.84 pixels. It is claimed in [12] that an RMS error less than one pixel is "satisfactory." For experiments 1 and 2, the RMS differences between the normal image velocity field and the field predicted by the inter-frame sensor motion are ± 0.10 and ± 0.09 pixels, respectively, which are significantly less than the one pixel

Table 3: Directional Errors in Sensor Translation

Exp.	Pan (deg.)	Tilt (deg.)	Comments
1	1.7	1.8	Frontal Plane
2	-0.80	-0.88	

Table 4: Comparison of Directional Accuracies

Researcher	Error (*deg*)	Approach (*Feature Type*)
Burger et al [4]	≈ 1.0	2D displacements
Matthies [10]	< 1.0	3D displacements
Hayashi et al [7]	4.3	normal image vel.
Braithwaite [3]	0.4	normal image disp.

"satisfactory" level.

The directional errors in the sensor translation are summarized in table 3. The dependence of directional accuracy on scene structure can be seen by comparing the results of experiments 1 and 2. The directional error along either the pan or tilt axes is less than 2 degrees for experiment 1 and less than 1 degree for experiment 2. The improved accuracy in experiment 2 is due primarily to the variation in depth. The dependence on the speed of translation can be seen by noting that the algorithm described in this paper has demonstrated a directional accuracy of ± 0.4 degrees for a fast moving stereo image sequence (see [2], [3]). The published results of other researchers (see table 4) suggest that a directional error of 1.0 degree is "very good."

In each experiment, the deviation of the estimated direction of translation from axial motion is within the expected error. Since the well-conditioned estimates have an error of less than 1 degree, and the other estimates are within the expected error of the actual motion, it can be concluded that this algorithm produces very good estimates of the direction of sensor translation.

6 Conclusion

An algorithm based on stereo Gabor filters has been presented which estimates stereo disparity, normal image velocity, and sensor motion. The image measurements obtained using Gabor filter 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 direction of sensor translation is accurately estimated. The results demonstrate the robustness of the algorithm to difficult conditions: images containing shadows and specular reflections; and scenes with large depth gradients.

Acknowledgements

The authors would like to thank Dr. Larry Matthies for providing the stereo image sequences used in experiments 1 and 2.

References

[1] G. Adiv, "Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 11, no. 5, pp. 477-489, 1989.

[2] R. N. Braithwaite, "Stereo-based obstacle detection using Gabor filters," Ph.D. thesis, Dept. Elec. Eng., Univ. of British Columbia, Vancouver, Canada, 1992.

[3] R. N. Braithwaite and M. P. Beddoes, "Estimating camera and object translation from a moving platform," in *proc., IEEE Symp. Intelligent Vehicles, Tokyo, Japan, 1993.*

[4] W. Burger and B. Bhanu, "Estimating 3-D egomotion from perspective image sequences," *PAMI*, vol. 12, no. 11, pp. 1040-1058, 1990.

[5] D. Fleet, A. Jepson, and M. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understand.*, vol. 53, no. 2, pp. 198-210, 1991.

[6] D. Gabor, "Theory of communication," *J. Inst. Elec. Eng.*, vol. 93, pp. 429-457, 1946.

[7] B. Hayashi, S. Negahdaripour, "Direct motion stereo: recovery of observer motion and scene structure," in *proc., ICCV*, 1990, pp. 446-450.

[8] B. K. P. Horn, *Robot Vision*. MIT Press, 1986.

[9] L. Matthies, R. Szeliski, and T. Kanade, "Kalman filter-based algorithms for estimating depth from image sequences," *Int. J. of Computer Vision*, vol. 3, pp. 209-236, 1989.

[10] L. H. Matthies, *Dynamic Stereo Vision*. PhD thesis, CMU-CS-89-195, 1989.

[11] S. Negahdaripour and B. K. P. Horn, "Direct passive navigation," *IEEE Trans. Pattern An. Machine Intell.*, vol. 9, pp. 168-176, 1987.

[12] J. Weng, T. Huang, and N. Ahuja, "Motion and structure from two perspective views: algorithms, error analysis, and error estimation," *PAMI* , vol. 11, no. 5, pp. 451-476, 1989.