ACTIVE SURFACE RECONSTRUCTION FROM OPTICAL FLOW

Marcel Mitran

Department of Electrical and Computer Engineering
McGill University, Montreal

August 2001

A Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering

© MARCEL MITRAN, MMI

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-79083-5

Canada

Abstract

This thesis describes the design and implementation of an active surface reconstruction algorithm for two-frame image sequences. The objective is to build a system that uses a passive sensor and an active viewer to accumulate information for disambiguating the depth sampling process involved in surface reconstruction. The viewer is considered to be restricted to a short baseline. Several ideas from the fields of optical flow, stereovision, and shape from motion will be drawn from and modified in the context of an active vision system.

The thesis begins by examining the optical flow estimation problem. Several algorithms are compared under the novel heading of maximal estimation theory. Each algorithm is decomposed into three parts: pixel-estimation, sub-pixel estimation and confidence measurement. The components are compared separately. A flow algorithm is then obtained by combining different components.

A Bayesian framework is adopted to provide a simple approach for propagating information in a bottom-up fashion in the system. This will also be used for combining information both temporally and spatially in the context of a Kalman filtering scheme.

The last part of this thesis examines how an active component can be integrated into the system to provide quicker convergence to the depth estimate. This approach is based on statistical grouping of image gradient features.

Synthetic and real experimental results are generated in each section. These results support ideas presented in the thesis, and suggest a basis for further research.

Résumé

Cette thèse décrit la conception d'un système actif de reconstruction de surface basé sur des paires d'images. L'objectif est de construire un système qui utilise un senseur passif en tandem avec un observateur actif pour réduire l'ambiguité sur les mesures de profondeurs pour la reconstruction de surface. Différents éléments des domaines du *flot* optique, de la vision en stéréo et de la reconstruction de surface seront abordés et modifiés dans le contexte du système actif désiré.

Cette thèse débute en examinant le problème du *flot* optique. Plusieurs algorithmes de ce domaine sont abordés et comparés. Le critère de comparaison utilisé est basé sur la théorie de l'estimation, ce qui constitue une approach nouvelle dans le domaine du flot optique. Chaque algorithme est décomposé en trois parties: la correspondance de pixels, la correspondance de sous-pixels et la mesure de confiance. Ces composantes sont examinées séparément. Un algorithme est obtenu en combinant plusieurs composantes différentes.

Une stratégie bayésienne est adoptée pour simplifier la propagation d'information dans le système. Ceci profite, en plus, au processus de fusion dans le domaine spatial et temporel. Tout cela sera regroupé dans le contexte d'un filter de Kalman.

La dernière étape de cette thèse discutera d'une stratégie active pour accélérer la convergence du système d'accumulation. Cette approche sera basée sur le regroupement de caractéristiques statistiques de l'image.

Des résultats expérimentaux sont présentés pour des données synthétiques et réelles dans chaque section. Ceux-ci supportent les idées présentées dans cette thèse.

Acknowledgments

I must begin by acknowledging my colleagues in the Artificial Perception Lab at McGill University. Philippe Simard, Haïg Djambazian, Stephen Benoit, Isabelle Bégin, Tal Arbel, Jean-Sébastien Valois and Fatima Drissi-Smaïli, have always made time to discuss the issues and help when I lost my way. I must also express my sincere appreciation to my thesis supervisor, Frank P. Ferrie, for his financial backing and encouragement, especially at those moments when things were at their worst.

I wish to extend a special thank-you to my brother, Patrick Mitran, who is one of the smartest people I know. Although his interest lie in a domain that is unrelated to computer vision, his insight into my work has always surpassed usefulness. I also thank my parents, Ovidiu and Birgitt Mitran, as well as my loving companion, Melissa Chee, for their undying encouragement and advice. Without their support, I certainly could not have *weather'd every rack.*

Finally, I wish to thank my three *amigos*: Jerrick Dangaran, Marie-Andrée Tessier and Jocelyn Lauzon. Their friendship is a precious gift. Joce, there will always be a Bud in my fridge for you. Guys... *O CAPTAIN! my Captain! The port is near, the bells I hear...*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The field of computer vision continues to play an important role in the development of autonomous robotic agents. Autonomous navigating robots should ideally be able to move through an unknown environment unaided. This involves path planning, obstacle detection, and scene recognition/interpretation [18]. Such tasks are all dependent on the robot's ability to quickly build sufficiently complete models of its environment. Thus, surface reconstruction remains an important motivator in the field of computer-vision.

The human visual system provides consistent proof that 2-D image sensing is sufficient for interacting with a 3-D world. Evolution has provided biological vision systems with a large set of tools for interacting with a 3-D world. Stereoscopic vision provides detailed representation for nearer objects (one meter away in humans) [26]. In most cases however, when moving through the world, objects are outside the stereoscopic range. Human experience in every day life demonstrates that, even under such conditions, it is possible to successfully perform many everyday tasks such as trajectory planning, obstacle detection and figure/ground separation. A similar task in computer vision involves recovering 3-D structure from a set of 2-D images. This problem requires the temporal accumulation of information through a monocular observer. The relationship between subsequent still images in a video stream provides a wealth of information in the form of spatio-temporal change. The temporal integration of such

velocity fields is essential for solving shape-from-motion [6, 16, 13, 42, 46, 54, 70], time-of-collision [18], object tracking [51], object-recognition [4], and figure/ground separation problems.

At first glance the problem of 3-D reconstruction from motion images seems trivial as it is intuitively sound to suggest that changes in intensity on an image plane are somewhat coupled with the projection of the apparent motion of the 3-D space surrounding the plane. It is however incorrect to say that such projections are unique and complete. The loss of a dimension, quantization of intensity, discrete sampling of infinitesimal spatial data and sensor noise make the problem of recovering 3-D structure from a set of 2-D intensity images ill-posed [10].

This begs the question, how does the 2-D human visual system successfully interact with the 3-D world with such consistency? Many suggest that the answer lies in considering the human observer not as a passive viewer, but rather as an active observer [2, 4, 7, 14, 68]. By interacting with the environment, a human can quickly and robustly achieve sufficiently stable representations of the world for navigation.

Although the human observer is active, it is wrong to assume complete freedom of motion exists in all six degrees of freedom under most conditions [18, 25, 45, 51, 59, 70]. For example, an individual driving a car is limited to very small lateral motions and a dominant forward motion. As such, a large baseline (motion parallel to the image plane) is not available to such an observer. Yet, people manage to navigate quite well without a wide baseline at their disposal.

This thesis examines *weakly active* surface reconstruction in the case of an autonomous monocular viewer. The term *weakly active* implies a severely constrained

configuration space for the viewer. Most active vision algorithms assume full motion control is available to a viewer. This is not often the case for a holonomically[1] constrained autonomous explorer, which must first see its world before moving through it. Thus a more realistic active motion model is considered, which constrains the viewer to small displacements between observations.

1. Overview of the Problem

Surface reconstruction can be defined as the process of inferring a mesh of interconnected points representing a three-dimensional surface. The surface is often assumed to be rigid and fixed. Points can be acquired using many types of sensors (e.g., range-finder, stereo-head). Computer vision systems generally wish to use image sensors to infer the state of the world. As such, computer vision systems ideally would like to be able to reconstruct objects or environments from a sequence of pictures. This measurement problem is inherently ill-posed [10] as projected image intensity fails to provide an invertible encoding of surface characteristics under most conditions. The conditioning of the system can be described by dividing it into parts. In general, image-based surface reconstruction system can be broken up into three elements (**Figure 1.1**):

i) Image correspondence,

ii) Depth estimation from triangulation or back-projection, and

iii) Depth integration.

[1] The physical construction of the robot and/or obstacles in the environment may prohibit certain configurations [18].

Each of these three elements comprises an important part of computer vision literature and, as such, can incur significant complexity in the system.

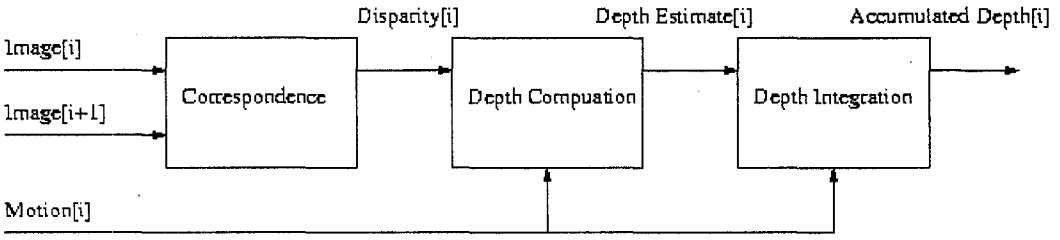


Figure 1.1 – Depth Accumulation System. Three structural blocks are represented: correspondence of intensity, depth estimation from correspondence and associated motion, and accumulation of depth estimates for surface reconstruction.

It is well accepted that the greatest difficulty encountered by image-based vision systems is the correspondence problem, which for small motions is referred to as the optical flow estimation problem. This involves measuring the motion of a projected point on the image plane. As mentioned earlier, recovering motion from a pair of intensity images is, for many reasons, ill-posed. It should also be noted that as the motion between images increases, correspondence becomes more difficult as image features will generally warp or fall outside the image-plane, and image intensity will change.

The depth estimation process, under perspective, involves transforming disparity[2] measurements taken from correspondence, and the associated viewer motion parameters into depth measurements. This task is ill-posed in some cases (e.g. pixels about the field-of-expansion when moving forward, or pure rotation about the viewer's origin), and

[2] Disparity measures the difference in retinal position between corresponding points in two stereo images.

increasingly ill-conditioned as motions between views become smaller and the angle of disparity approaches zero.

The depth integration process combines many depth measurements to reduce the effects of noise and increase the size of the data set. Depth measurements can be accumulated over time or joined in a batch process. The inherent difficulty in this process is in establishing correspondence between the many depth estimates. Much in the same spirit as intensity registration, as the motion between the two depth measurements increases, the correspondence becomes increasingly difficult and ill-posed. As the surface representation is necessarily discrete, interpolation must be used to merge points that fall in between points on the previous surface. As the depth samples become sparser, the interpolation process becomes increasingly ill-conditioned.

In general, when considering the above elements, two forms of surface reconstruction emerge. The first is feature based surface reconstruction [5, 9, 27, 28, 43, 49, 56]. This approach chooses features in the image that are stable for large motions. Thus, a sparse set of very confident depth estimates is obtained. Difficulties occur when trying to interpolate full surface representation. Often planarity assumptions must be used, or some underlying knowledge of the surface must be known *a priori*. This approach is in general not realistic for an autonomous robot that cannot afford to perform large motions without attending to the scene and thus running the risk of a critical collision.

The second approach to depth estimation is the iconic depth estimator [30, 46, 60, 70] in which all pixels contribute a depth estimate. This approach is more suitable for a navigating robot as it lends itself to small motions between viewpoints. Thus, the image

and depth correspondence problems are locally constrained and facilitated, and a dense disparity field is obtained. The difficulty in this approach is that the depth measurement process is very noisy. However, as the depth integration process is simplified, noisy measurements can be filtered out given sufficient depth estimates.

The two previous mentioned approaches to surface reconstruction illustrate a clear and well-known dichotomy in baseline stereo: smaller motions aid with the correspondence problem, while greater baseline motion provide more stable depth estimates but sparser data. Multi-baseline stereo [2, 37, 44, 53, 69] has sought to merge these two problems by tracking many points over many small motions. Once a sufficiently large baseline has been achieved, depth values are computed with less ambiguity. As mentioned earlier, this approach is not practical for an autonomous explorer that does not necessarily have a wide baseline at its disposition. However, the multi-baseline approach does provide inspiration for the approach suggested in this thesis.

2. The Approach

Given that a wide baseline is not available as in the multi-baseline approach, the system proposed in this thesis will use repeated sampling to disambiguate the measurements. If the measurement noise can be modeled as zero-mean, it is reasonable to assume that sufficient temporal integration can be used to make up for the shortcomings of a diminished baseline.

Maximal estimation theory provides a framework for the efficient accumulation of information in the context of a noisy measurement process. This methodology uses qualitative data to identify how much information is entering or is already in a system. Thus, measurements must have associated confidence values. These confidence values are propagated in a bottom up fashion through the system. If these confidence values are equivalent to inverse variance on the error of the associated data, it can be shown that the system will provide a minimal solution (in the least-squares sense).

The thesis will begin by examining the correspondence process in the context of maximal estimation theory. Different optical flow methods in the literature will be considered. Particular attention will be paid to the confidence measures identified for different optical flow algorithms. A novel approach to comparing these confidence values in the context of data accumulation will be introduced. A hybrid optical flow estimator will be constructed from the different flow-estimation algorithms considered.

The second part of the thesis draws heavily from previous work by Szeliski [60] and Matthies *et al.* [46] who discuss a Bayesian formulation for weighted accumulation of information using a maximal estimation framework [48]. Szeliski has shown that Bayesian modeling can be used for low-level vision systems. As such, measurements can successfully be represented as a mean and variance pair. Work by Matthies *et al.* shows how such a Bayesian model can be used to temporally accumulate information using a Kalman filter framework. This thesis will also improve on Matthies *et al.*'s framework by extending the maximal estimation framework to spatial data propagation. Mathur and Ferrie [47] provide the theory behind this improvement.

The last part of this thesis will introduce a simple feedback mechanism for active view selection. Work by Whaite and Ferrie [68], and Arbel and Ferrie [4] provide inspiration for the suggested approach. Whaite and Ferrie discuss active fitting of super-ellipsoids to range data. Arbel and Ferrie develop a similar application for active object recognition using optical flow. In this thesis, it will be shown that it is possible to predict the most informative motions based on the intensity in the prior image. As such, the temporal integration process for surface reconstruction will be shown to converge more consistently than a passive motion sequence.

3. Outline of Thesis

The rest of the thesis is organized as follows. Due to the importance of the correspondence problem and the particular need for a qualified measurement process for the application presented here, Chapter 2 will be dedicated to reviewing the problems and associated literature on optical flow, and molding a correspondence algorithm that suits the current purpose. Chapter 3 will review reconstruction geometry for a weak perspective camera model as well as the Kalman filter framework presented by Matthies *et al.* An improvement to the interpolation and regularization process will be introduced into the framework as well. Chapter 4 will discuss the addition of the epipolar constraint as well as the suggested feedback mechanism in detail. It will also provide results for a complete implementation of each previously described element of the system.

As each component of the system draws from different areas of computer vision, the literature will be cited and results will be provided when necessary. Chapter 5 will

conclude with general review of the thesis content, and some general observations with respect to the results.

4. Contributions

The contributions of this thesis consist of the following:

- A novel method for comparing the confidence measures of optical flow estimation is introduced. This approach provides a formal methodology and has the benefit of lending itself directly to maximal-estimation theory.
- An improved confidence measure for optical flow is provided. This method is a more general representation of previous approaches and is shown experimentally to provide more consistent results for a standard optical flow test set.
- The introduction of a maximal-estimation interpolator for iconic spatial depth accumulation. This is shown, from an information theoretic point of view, to be more appropriate for providing spatial support than current interpolation methods. Experimental results are provided to support the suggested interpolator.
- The development of an active procedure for accumulating depth. It is shown that the gradient of the intensity in the image can successfully be used to drive the viewer's motion. This effectively increases the convergence rate of the depth estimator significantly (3 to 4 times) over that of a passive viewer.

- Design and implementation of a fully functional active surface reconstruction system using a gantry robot and an off-the-shelf NTSC camera setup.

CHAPTER 2

Correspondence and Optical Flow

An important first step in developing the desired active surface reconstruction system is to consider the correspondence problem. As such, this chapter will be dedicated to first discussing where ambiguity arises in optical flow estimation and how the measurement process can be used to qualify it. The chapter will review key elements in the optical flow literature.

1. Previous Work in Optical Flow Estimation

Optical flow is defined by Horn [32] as: the problem of estimating the apparent motion of a brightness pattern. Barron *et al.* [8] consider optical flow as the process of: computing the approximation to the 2-D motion field – a projection of the 3-D velocities of surface points onto the imaging surface – from spatio-temporal patterns. These two definitions are not identical, as the first treats optical flow as being independent of scene structure, while the latter considers optical flow to be a consequence of scene structure and noise. It is suggested later in this section that these two definitions can be reconciled under certain conditions for an active viewer.

Horn formalizes a constraint for relating the projected 3-D motion in the scene and the image flow. This is referred to as the image flow constraint. Under specific conditions, it provides a relationship between the projected 3-D motion and the observed

change in intensity. This constraint can be interpreted as the assumption that a point in the 3-D surface, when projected onto the 2-D image plain, maintains a constant intensity over time. This assumption generally holds for small motions. Mathematically it is formulated as

$$I(x, y, t) = I(x - \Delta x \cdot t, y - \Delta y \cdot t, 0), \tag{2.1}$$

where $I(x, y, t)$ is the intensity distribution of the image of a pixel at a point (x, y) at a time t and $\vec{f} = (\Delta x, \Delta y)$ is the desired optical flow vector.

The image flow constraint provides limited conditioning to the problem for recovering the 2-D velocity field from a sequence of intensity images. The problem of projected motion recovery from images is ill-posed because local intensity alone fails to completely encode motion information. Three different levels of ambiguity may occur.

The first and most severe ambiguity involves reflective and transparent surfaces. Horn's mirror ball problem is a good example of this. Consider a mirrored sphere. As the sphere rotates about its center, no change in intensity is observed, yet the sphere does possess a 3-D motion field and texture is present in the reflected image. This is the most extreme case of ambiguity. Even biological visual systems fail under such conditions. In general, it is assumed that the surface is matt and that its texture is glued to it.

The second and less extreme example of how intensity fails to encode projected 3-D motion involves regions of constant intensity. In such regions, motion cannot be detected and an infinite number of solutions exist. Such cases rely on propagation of information from surrounding estimates to interpolate a measurement. Thus regularization is necessary to guess a solution for such areas. Ju. *et al.* provide an example of this in their Skin and Bones paper [36]. According to Bajcsy [7],

interpolation and regularization should be avoided whenever possible, as they impose a biased estimate that often resembles an educated guess.

The last form of ambiguity in flow estimation occurs when the intensity is constant in a given direction $\vec{u}$. This is referred to as the aperture problem [32]. In such cases the solution is only partially available as it is only possible to provide the component of the solution that is perpendicular to $\vec{u}$, called the normal velocity estimate. Under such conditions, an active viewer can effectively provide sufficient constraints to remove the ambiguity. As the scene is assumed static and rigid and the motion of the viewer is controlled up to a given certainty, it is possible to constrain the direction of the flow measurement using geometric properties of the projection model [21]. In fact, if the direction of the optical flow vector is parallel to the image gradient, the normal velocity and the flow velocity become equivalent, as suggested by Verri and Poggio [66]. Thus, the two definitions of optical flow mentioned earlier can be reconciled.

These different levels of ambiguity provide an important result in the development of a motion strategy for the viewer. The viewer should maximize the confidence in measurement by detecting when and what type of ambiguity occurs. The first case cannot be detected and is removed by assumption. The second and the third can in fact be detected. Different flow algorithms provide different approaches for resolving this problem.

Two comprehensive papers on the subject of optical flow performance exist. Work by Liu *et al.* [39] has studied the efficiency/accuracy tradeoff of different algorithms. The authors produce curves of accuracy versus efficiency for comparing different optical flow algorithms. A curve is constructed for each algorithm by changing

its parameters. Barron *et al.* have produced a paper that compares nine classic flow algorithms on the basis of accuracy and density. They provide a clear test set of image sequences that can be used for quantitative and qualitative comparison of the different algorithms. Most importantly, they discuss the different confidence measures used by different flow algorithms.

Barron *et al.* [8] classify flow algorithms into four groups: differential techniques, energy-based methods, phase-based techniques and region-based matching. Differential [41, 50, 52, 65], energy-based [29, 40] and phase-based [22] techniques can all be classified under the heading of gradient methods. These all perform discrete temporal filtering and require strong temporal support to work well. The energy and gradient-based methods, generally require families of velocity tuned filters to work well, which generally renders them much slower than differential or region matching methods. Thus, gradient methods are unlikely candidates for the autonomous explorer implementation. Still there are aspects concerning the confidence measurements of these algorithms that can be of use. This becomes apparent from the differential approaches. Thus, differential techniques will be considered briefly below, after which region matching will be discussed.

1.1 Differential Methods

Differential techniques are characterized by gradient search performed on first and second order spatial derivatives and temporal derivatives extracted from the image sequence. From the Taylor expansion of the flow constraint equation, the gradient constraint equation is obtained,

$$\nabla I(x,y,t)\cdot(\Delta x,\Delta y,0)+I_t(x,y,t)=0\,. \tag{2.2}$$

Horn and Schunk [33] combine a global smoothness term with the gradient constraint equation to obtain a functional for estimating optical flow. Their choice of smoothness term minimizes the absolute gradient of the velocity using

$$\int\int\left[\left(\nabla I\cdot\vec{f}+I_t\right)^2+\lambda^2\left(\left\|\nabla_2^2\Delta x\right\|+\left\|\nabla_2^2\Delta y\right\|\right)\right]dxdy\,. \tag{2.3}$$

This functional can be reduced to a pair of recursive equations that must be solved iteratively. It provides no confidence measure.

Lucas and Kanade [41] also construct a flow estimation technique based on first-order derivatives of the image flow constraint. In contrast to Horn and Schunk's approach of post-smoothing regularization, they choose to pre-smooth the data. This is represented mathematically as,

$$\mathbf{min}\sum_{\vec{x}\in\Omega}W^2(\vec{x})\left[\nabla I(\vec{x},t)\cdot\vec{f}+I_t(\vec{x},t)\right]^2\,, \tag{2.4}$$

where $W(\vec{x})$ is a window that gives more influence to constraints near the center of the neighborhood Ω. A closed form least-squares fit is then used to provide an optical flow estimate,

$$\vec{v}=[A^TW^2A]^{-1}A^TW\vec{b}\,, \tag{2.5}$$

where

$$\begin{aligned} A&=\left[\nabla I(\vec{x}_1),\ldots,\nabla I(\vec{x}_n)\right]^T,\\ W&=diag\left[W(\vec{x}_1),\ldots,W(\vec{x}_n)\right],\\ \vec{b}&=-\left[I_t(\vec{x}_1),\ldots,I_t(\vec{x}_n)\right]^T. \end{aligned} \tag{2.6}$$

This approach can be described as a weighted minimization of normal constraints where the weights are the magnitude of the spatial gradient of the image.

Barron et *al.* report in their survey [8] that Lucas and Kanade's algorithm provides the second most accurate results. Liu et *al.* [39] evaluate Lucas and Kanade as providing the third best efficiency-accuracy curve. Thus, there is noted interest in this approach. However, there is an important disadvantage to this algorithm. As discrete temporal differentiation is necessary, strong temporal support is required. The Barron *et al.* implementation required 15 frames of temporal support, and a 7-frame delay. This may be unacceptable for a real-time autonomous explorer. Aside from the noted phase delay, an autonomous observer should not be required to provide long continuous sequences in its path planning process [70].

There have been efforts to reduce the required temporal support. Fleet and Langley [23] attempt a more efficient implementation of Lucas and Kanade's work using infinite impulse response (IIR) temporal pre-filtering and temporal recursive estimation for regularization. They reduced the temporal support to three frames while improving computational efficiency. Unfortunately, the IIR filter mechanism comes at a price of decreased precision. Also, this approach does not lend itself well for non-smooth motions. If, for example, the motion of the viewer were to change directions, the filters would have to be reset and returned to stability.

One important contribution of these first order approaches is the suggested confidence measure, which is independent of the temporal component of the flow constraint. The spatial component of the first order gradient constraint is often referred to as the Normal matrix,

$$A^T W^2 A = \begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_x I_y & \sum W^2 I_y^2 \end{bmatrix}. \tag{2.7}$$

The eigenvalues of the Normal matrix have important significance when considering the conditioning of the flow estimation problem. Barron *et al.* suggest that when the smallest eigenvalue, λ_2, is less than 1.0, the aperture problem prevails. Fleet and Langley provide additional support to this statement.

There has been some interest in the second order image flow constraint for estimating optical flow [50, 65]. The second order constraint equation takes the following form,

$$H(I) \cdot \vec{f} + \vec{I}_t = \vec{0}, \tag{2.8}$$

where,

$$H(I) = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix}, \tag{2.9}$$

$$\vec{I}_t = \begin{bmatrix} I_{tx} \\ I_{ty} \end{bmatrix}.$$

Much in the same manner as Horn and Schunk, and Lucas and Kanade, different forms of regularization and minimization are used to solve the system of equations (2.8). Barron *et al.* confirm that a second order system requires increased constraint on the estimation process, as the higher order implies increased instability. Thus sparser and less accurate results are obtained.

The spatial component, $H(I)$, of the system of equations (2.8), often referred to as the Hessian, can be used in this case to determine the conditioning of the system. Thus, two confidence measures are suggested. Uras *et al.* [65] consider the smallest condition

number of the Hessian, $\kappa(H)$. Barron *et al.* suggest using the determinant of the Hessian, $\mathbf{det}(H)$. The later is shown to provide better results by Barron *et al.*

It is important to note that all the suggested confidence measures used by the differential methods do not require computing the actual flow. They only require knowledge of the current image intensity. Thus, these confidence measures provide an ideal mechanism for predicting ambiguity in the correspondence process for any flow algorithm. This will prove to be very useful for developing a closed loop active viewer control paradigm.

1.2 Region Matching Methods

Region matching is set apart from gradient methods as it forms the temporal filters for features extracted from the previous image in the sequence. Tiles from the previous image are matched with the next image using some metric. The best match provides the most likely displacement. This is equivalent to searching a spatially shifted and temporally differentiated space, where spatial shifts are in unit pixel distances.

This approach is better suited for the autonomous explorer application as it provides robustness with respect to temporal differentiation. It is generally quicker since it constructs a highly quantized solution space. The main disadvantage of region matching is that it only provides coarse depth unless extra interpolants or sub-pixel estimators are added.

The distance measure used by more classical algorithms such as Anandan [3], and Singh and Allen's [58] is referred to as the sum-of-square differences (SSD). It is formulated as

$$SSD_{1,2}(x,y,dx,dy) = \sum_{j=-n}^{n}\sum_{i=-n}^{n} W(i,j)\left[I_1(x+i,y+j) - I_2(x+dx+i,y+dy+j)\right]^2 . \tag{2.10}$$

where I_1 and I_2 are an image pair, W is a 2-D window function, and (dx,dy) denotes the suggested displacement vector.

Anandan constructs a multi-scale method based on the Burt Laplacian pyramid [12]. A coarse-to-fine strategy is adopted such that larger displacements are first determined from less resolved versions of the images and then improved with more accurate higher resolution versions of the image. This strategy is well suited for cases where the range of pixel motions is large.

Confidence measures, $c_{\max}$ and $c_{\min}$, which are based on the principle curvatures, $C_{\min}$ and $C_{\max}$, of the SSD surface, are used to steer the smoothing process. These are represented mathematically as

$$c_{\max} = \frac{C_{\max}}{k_1 + k_2 SSD_{\min} + k_3 C_{\max}}, \tag{2.11}$$

$$c_{\min} = \frac{C_{\min}}{k_1 + k_2 SSD_{\min} + k_3 C_{\min}}. \tag{2.12}$$

where k_1, k_2 and k_3 are normalization constants. The smoothness constraint is based on the directions, $\vec{e}_{\min}$ and $\vec{e}_{\max}$, of the principle axes of the SSD surface, the estimated displacements $\vec{d} = (dx,dy)$, and the sought best-fit velocity estimate $\vec{f} = (\Delta x, \Delta y)$. Anandan also includes Horn and Schunk's [33] formulation of the smoothness constraint. Mathematically,

$$\iint\left(\Delta x_x^2+\Delta x_y^2+\Delta y_x^2+\Delta y_y^2\right)+c_{max}\left(\vec{f}\cdot\vec{e}_{max}-\vec{d}\cdot\vec{e}_{max}\right)^2+c_{min}\left(\vec{f}\cdot\vec{e}_{min}-\vec{d}\cdot\vec{e}_{min}\right)^2 dxdy\,, \tag{2.13}$$

where the x,y subscripts represent partial derivatives along x,y respectively.

Anandan's sub-pixel approach is equivalent to fitting a parabolic surface to the SSD distribution. The 1-D parametrization is

$$SSD(x)=ax^2+bx+c\,. \tag{2.14}$$

The sub-pixel flow is obtained by solving for the minimum of this surface.

Singh and Allen provide another approach to region matching based on SSD correlation. They use a three-frame approach to the region matching method to average out temporal error in the SSD. For a frame 0, they form an SSD distribution with respect to frame –1 and frame +1 as such

$$SSD_0=SSD_{0,1}\left(\vec{x},\vec{d}\right)+SSD_{0,-1}\left(\vec{x},-\vec{d}\right). \tag{2.15}$$

A two-frame method could also be implemented.

From SSD_0, Singh and Allen build a probability distribution

$$R_c(\vec{d})=e^{-k\,SSD_0}\,, \tag{2.16}$$

where k is a normalization constant. The sub-pixel flow estimates $\vec{f}_c=(\Delta y_c,\Delta x_c)$ are then obtained by considering the mean of the distribution with respect to $\vec{d}=(dx,dy)$,

$$\Delta x_c=\frac{\sum R_c\left(\vec{d}\right)dx}{\sum R_c\left(\vec{d}\right)}\,, \tag{2.17}$$

$$\Delta y_c=\frac{\sum R_c\left(\vec{d}\right)dy}{\sum R_c\left(\vec{d}\right)}\,.$$

Singh and Allen employ a Laplacian pyramid strategy similar to that of Anandan. This provides a more symmetric distribution about displacement estimates in the SSD. A covariance matrix is then constructed from these estimates as,

$$C_c = \frac{1}{\sum R_c(\vec{d})}\begin{bmatrix} \sum R_c(\vec{d})(dx-\Delta x_c)^2 & \sum R_c(\vec{d})(dx-\Delta x_c)(dy-\Delta y_c) \\ \sum R_c(\vec{d})(dx-\Delta x_c)(dy-\Delta y_c) & \sum R_c(\vec{d})(dy-\Delta y_c)^2 \end{bmatrix}. \tag{2.18}$$

Singh suggests that the eigenvalues of the inverse of C_c provide a measure of confidence for $\vec{f}_c$.

For a given flow field $\vec{f}_i = (\Delta x_i, \Delta y_i)$, the least-squares estimate in a $(2w+1)\times(2w+1)$ neighborhood about $\vec{f}_n = (\Delta x_n, \Delta y_n)$ can be obtained from

$$\Delta x_n = \frac{\sum R_n(\vec{f}_i)\Delta x_i}{\sum R_n(\vec{f}_i)}, \tag{2.19}$$

$$\Delta y_n = \frac{\sum R_n(\vec{f}_i)\Delta y_i}{\sum R_n(\vec{f}_i)}.$$

A covariance matrix C_n can then be generated in the same manner as (2.18) from (2.17). Flow regularization is then obtained by minimizing the sum of the Mahalanobis distances between the estimated flow field $\vec{f}$ and the two distributions $\vec{f}_c$ and $\vec{f}_n$,

$$\iint (\vec{f}-\vec{f}_n)C_n^{-1}(\vec{f}-\vec{f}_n)+(\vec{f}-\vec{f}_c)C_c^{-1}(\vec{f}-\vec{f}_c)dxdy. \tag{2.20}$$

The eigenvalues of the covariance matrix $\left[C_c^{-1}+C_n^{-1}\right]^{-1}$ serve as confidence measures for the regularization process.

An interesting region matching flow algorithm is Camus' quantized flow [13]. Liu et *al.* [39] report that Camus' algorithm provides one of the two best accuracy-efficiency ratio curves. Camus notes the simple relationship between velocity, distance and time,

$$velocity = \frac{\Delta d}{\Delta t}. \tag{2.21}$$

Classical region matching algorithms set Δt to 1. The range of Δd is defined by the extent of the correlation search area. Camus proposes that the search be extended in time, s frames deep, and reduced in space. For example, **Figure 2.1** shows a search two frames deep ($S = 2$). The winning displacement is (2,2) in *Image[2]*. Thus, the velocity vector is (0,1/2).

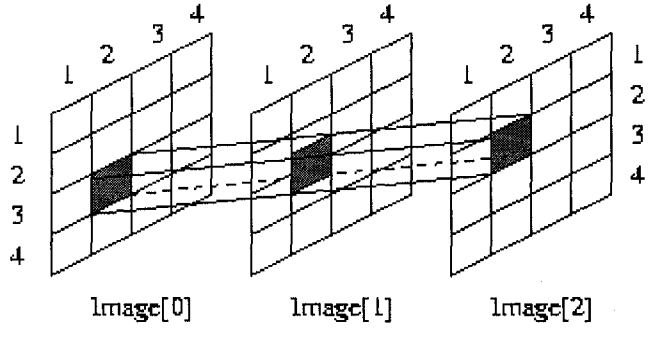


Figure 2.1 – Motion of pixel (2,3) in *Image[0]* to pixel (2,2) in *Image[2]*, an optical flow of (0,1/2) pixels per frame.

The advantage of this approach is that performing a search over time instead of over space is linear in nature rather than quadratic. Another efficient element of this algorithm is its suitability for integer arithmetic by suggesting additional optimizations for the correlation process under this framework. Camus proposes a box filter for $W(i,j)$ in (2.10) and memory management methods that make this approach extremely efficient

and robust. The price paid for this speed is that the algorithm only provides a quantized flow field containing $S(2n+1)^2$ different possible velocities.

2. Constructing a Correspondence Mechanism

It should be noted that the objective here is not to design a perfect optical flow estimator, as this is an impossible task. Instead, a set of criteria is established to provide guidelines for selecting components of the different flow algorithms in the construction of a correspondence mechanism and the motion selection mechanism. After considering the different optical flow algorithms in the literature, three elements must be chosen for constructing a suitable optical flow algorithm: a pixel motion estimator, a sub-pixel motion estimator and a confidence measurement process. Smoothing of the flow field is neglected to thus provide the accumulation process with unbiased data.

Under the current framework, the following criteria should be kept in mind when designing a flow estimator:

i) The flow algorithm should require minimal temporal support. The correspondence algorithm should be robust even when motion sequences are not smooth. This, ideally, implies a two-frame correspondence problem.

ii) It is important to be able to associate a confidence value with the measurement. As long as the relative confidence in flow measurements between images can be well estimated, the actual quality of the flow

measurement is left as a parameter to be decided on for the particular application.

iii) The algorithm should lend itself well to real-time navigation system. In the extreme case, if an algorithm is too slow, the assumption of a rigid environment may not necessarily hold.

Finally, the optical flow estimator results presented in this section are qualified using Barron *et al.*'s standard angular error metric. This metric is described as follows. Let a motion vector, $\vec{f}_2 = (\Delta x, \Delta y)^T$, be represented as a 3-D directional vector,

$$\vec{f}_3 \equiv \frac{(\Delta x, \Delta y, 1)^T}{\sqrt{\Delta x^2 + \Delta y^2 + 1}}. \tag{2.22}$$

For flow estimate, $\vec{f}_{2e}$, and corresponding ground truth, $\vec{f}_{2c}$, the angular error is defined as

$$\Psi_E = \cos^{-1}\left(\vec{f}_{3e} \cdot \vec{f}_{3c}\right). \tag{2.23}$$

This metric is chosen, as it appears to be the standard used by optical flow evaluation literature.

2.1 Pixel Correspondence

The first key point in selecting an optical flow algorithm is condition i), which suggests that shorter temporal support is desired. An important issue with differential algorithms is their stability under such conditions. To investigate this stability, the Lucas

and Kanade algorithm was modified to use 2-pt and 3-pt [3] numerical temporal differentiation with equivalent temporal smoothing. Fleet and Langley's IIR filter approach was also implemented. These were tested on the translating tree sequence **Figure 2.2** [8]. This sequence is chosen as it contains both sharp and smooth intensity features while providing a uni-modal flow field.

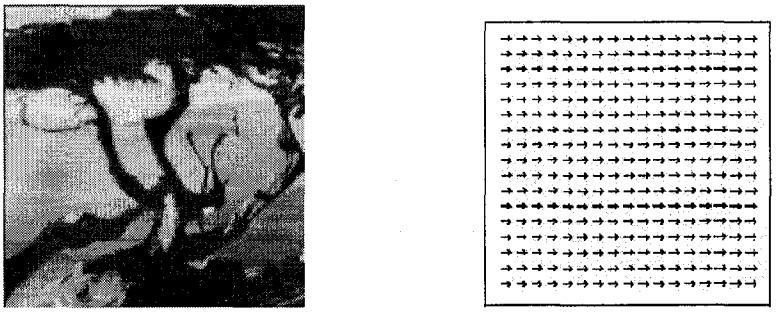


Figure 2.2 – Frame 4 from the translating tree sequence, and the ground truth optical flow field.

The differential approach is compared to a region-matching algorithm. Small pixel motions may be assumed as the viewer's motion assumed small. Thus, the multi-scale pyramid implementation used by Anandan and Singh may be neglected. Camus' optimized region matching algorithm is used. This algorithm is extremely efficient while providing flexible and robust temporal support. The implementation presented here uses two frames of temporal support. Spatial support was set to match that of the differential algorithms (9 pixels).

Results for Lucas and Kanade's, and Fleet and Langley's algorithms are compared to the traditional 5-pt temporal differentiation and Camus' region-matching algorithm in **Table 2.1**. Camus' algorithm provides the best result for this sequence.

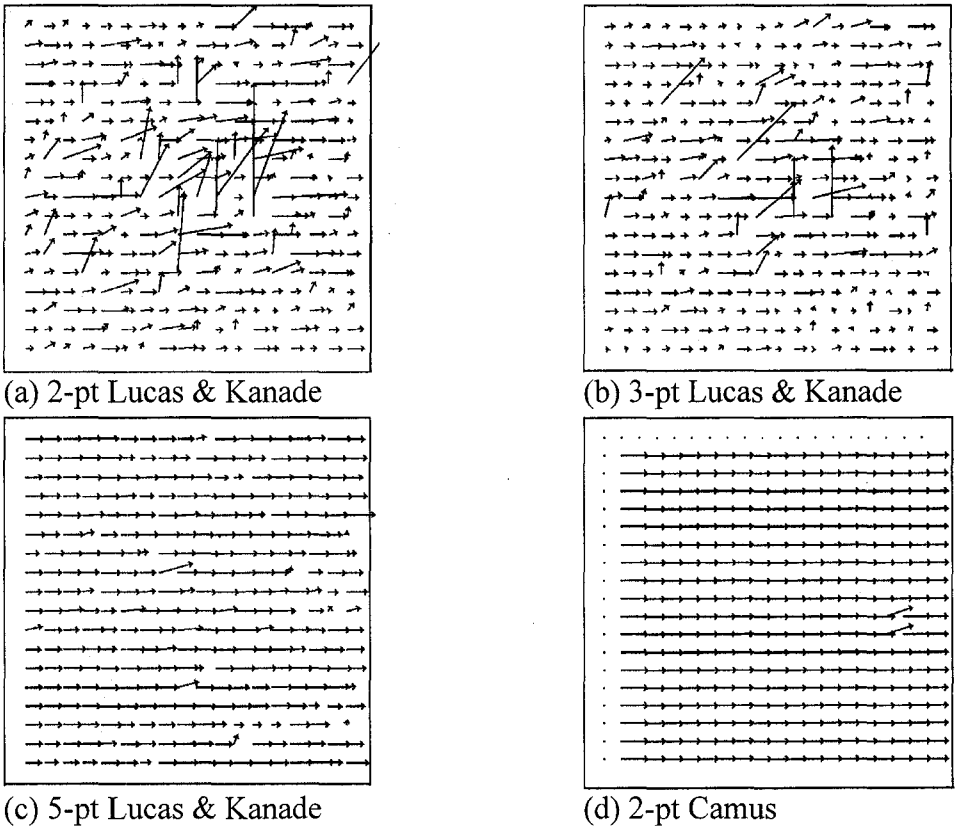
[3] Two-point and three-point differential operators were implemented as (1,-1) and (1/2,0,-1/2).

Algorithm	Angular Error (Deg)
2-pt Lucas & Kanade	22.6343
3-pt Lucas & Kanade	20.9849
3-pt Fleet and Langley	18.5682
5-pt Lucas & Kanade ($\lambda = 0.0$)	3.7262
5-pt Lucas & Kanade ($\lambda = 1.0$)	1.9250
2-pt Camus	1.4649

Table 2.1 – Flow error computation for translating tree sequence.

When comparing **Figure 2.3**(a) and **Figure 2.3**(b) to **Figure 2.3**(c) and **Figure 2.3**(d), it becomes apparent that, for the translating tree sequence, that the differential method is unreliable for two- and three-image sequences. It is noted that the flow estimates are most unstable along strong intensity changes. Normally, such features are band-limited through temporal smoothing. From a signal processing point of view, this smoothing seems counter-intuitive, as features with higher frequency signatures should provide more information for the correspondence process. Region matching algorithms avoid the temporal smoothing process, and as such, provide a broadband approach for matching signals with higher-band frequencies.

The disadvantage of Camus' region matching algorithm is that it doesn't provide a true sub-pixel estimate or a measure of confidence. Thus, it becomes necessary to invent or borrow these components from other algorithms. Sub-pixel estimation approaches and confidence measures are discussed in the next two sections.



(a) 2-pt Lucas & Kanade (b) 3-pt Lucas & Kanade

(c) 5-pt Lucas & Kanade (d) 2-pt Camus

Figure 2.3 – Optical Flow fields for Lucas & Kanade, and Camus for translating tree sequence.

2.2 Sub-pixel Correspondence

Four approaches to sub-pixel estimation are described. Three approaches are first considered: a bilinear interpolation method, Singh's method and Anandan's method. A fourth method, that combines Anandan and the bilinear approaches, is derived. **Table 2.2** provides results for each of these approaches for several standard synthetic image sequences for which the respective ground-truth is known. Camus' pixel estimation

approach, as well as the percentage of pixel motion estimates that are within a single pixel range of the ground truth optical flow, are provided to help identify how much room for improvement is available to the sub-pixel estimator.

	Angular Error for Image Sequence (Deg)				
Sub-Pixel Type	Sine-B	Sine-C	Trans. Tree	Diver. Tree	Yosemite (no sky)
Camus (% in pixel range)	5.21324 (100%)	0 (100%)	1.46495 (99.5%)	16.2144 (98.9%)	13.0404 (90.9%)
Bilinear	2.80501	0	1.25434	7.41948	7.92528
Singh	4.86211	0	2.66364	16.5368	13.1318
Anandan	1.17886	2.3822	2.57259	6.95023	6.46842
Bilinear & Anandan	0.0124328	0.525342	1.22560	5.68054	7.21033

Table 2.2 – Sub-pixel estimation results.

The first sub-pixel flow estimator uses a bilinear interpolator to up-sample the local patches of image around the SSD minimum by a factor of four. A new refined sub-SSD is obtained at this point. Results for this approach demonstrate that this method provides a robust estimate that is, however, still coarse. This is noted in sequences where the distribution of phase of the flow-field is well spread such as the Diverging Tree and Yosemite sequences. Singh's approach provides poor results overall. It relies on statistical estimation of the sub-pixel displacement. Thus large patches of the image must be used to obtain statistical correctness. This is an undesirable property as the larger the image patch becomes, the less local information is represented. Barron *et al.*'s comment concerning a bias for sub-pixel values that approach zero is noted as well, as Singh's algorithm performs well for the Sine-C which has no sub-pixel component, and the translating tree which only has a sub-pixel component in its x- component. Anandan's method was computationally efficient and provided better results than the previous three

mentioned. Conversely to Singh's method, it is overambitious in cases where no sub-pixel displacement exists. This is observed for the Sine-C sequence. A fourth flow estimation algorithm is implemented that takes advantage of the robustness of the linear method to constrain Anandan's estimation. Thus, a bilinear sub-SSD is first computed; Anandan's quadric surface fit is then applied to the sub-SSD surface to refine the estimate. This method provides the best sub-pixel flow estimate overall and is adopted. It is similar to Matthies *et al.*'s approach of using a cubic interpolator on top of Anandan's sub-pixel flow estimator to provide a sub-sub-pixel flow estimation.

2.3 Measure of Confidence

The last step in building a correspondence estimator involves developing a confidence measure. This section provides a novel and formal mechanism for comparing different optical flow confidence measures in the context of maximal estimation theory. From this, a more general and improved confidence measure is suggested and demonstrated experimentally.

Three measures of confidence are proposed: the differential method, Singh's approach, and a modified Anandan technique. Barron *et al.* evaluate these confidence measures by applying thresholds to reject undesirable flow estimates and then re-evaluating the sparser flow-field. The evaluation approach taken here is to determine how well the confidence values represent the information contained in the estimate. As such, flow estimates are never rejected. Instead, emphasis is placed on using the confidence values as weights to maximally merge information. Thus, a weighted least-

squares approach is used to determine how well error is attenuated by the confidence measures. For a linear measurement model,

$$\bar{z} = H\bar{x} + \gamma, \tag{2.24}$$

where H is a linear operator relating the state vector $\bar{x}$ to a measurement $\bar{z}$, and γ is White Gaussian measurement noise with an associated covariance matrix,

$$C_m = \begin{bmatrix} \sigma_{z1}^2 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & \sigma_{zn-1}^2 \end{bmatrix}. \tag{2.25}$$

In the case of an ideal measurement process, the inverse confidences should be equivalent to the variances on the flow measurements. In the flow-estimation framework, each of a confidence pair (c_x, c_y) is provided for the x,y components of the flow estimate,

$$C_m = \begin{bmatrix} \frac{1}{c_x} & 0 \\ 0 & \frac{1}{c_y} \end{bmatrix}. \tag{2.26}$$

This framework requires modifying the original confidence formulations. For the differential approach, the diagonal components of the Normal matrix (2.7), I_x^2 and I_y^2, are used. For Singh's approach the diagonals of (2.18) are used. Similar to Anandan's approach, Matthies *et al.* choose to use the curvature along the x,y directions of the SSD surface about its minimum as a confidence measurement. This is reasonable as sharpness of the parabolic SSD provides a direct metric of dissimilarity between the minimum and its neighbors. The attribute is fully described by the second order derivative of *SSD(x)* which is parameter a of (2.14),

$$(c_x, c_y) = \left(\frac{1}{a_x}, \frac{1}{a_y} \right), \tag{2.27}$$

where (a_x, a_y) represent the curvatures along the *x,y* directions of the SSD. Anandan's original confidence measure can be expressed as,

$$(c_x, c_y) = \left(\frac{SSD_{\min}}{a_x}, \frac{SSD_{\min}}{a_y} \right). \tag{2.28}$$

A more general and novel form for Anandan's method is introduced here,

$$(c_x, c_y) = \left(\frac{SSD_{\min}^K}{a_x}, \frac{SSD_{\min}^K}{a_y} \right), \tag{2.29}$$

where K determines the weight of the SSD_{min} values.

The maximal weighted least-squares estimator of (2.24) is [48],

$$\hat{x} = \left(H^T C_m^{-1} H \right)^{-1} H^T C_m^{-1} (\bar{z} - \gamma), \tag{2.30}$$

and the associated state variance, $\sigma_{\hat{x}}^2$, is obtained from the following expression

$$\sigma_{\hat{x}}^2 = \left(H^T C_m^{-1} H \right)^{-1}. \tag{2.31}$$

For perfect confidence values, γ should contribute minimally to $\hat{x}$. As the error in the flow measurement is known for synthetic sequences, it is possible to determine how the confidence measure predicts the error in the measurement. For simplicity, H, is chosen to be

$$H = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{2.32}$$

and, γ, is chosen as

$$\gamma = \left(\Delta x_e^2, \Delta y_e^2 \right), \tag{2.33}$$

where $(\Delta x_e, \Delta y_e)$ is the flow error for a pixel (i,j). The weighted error is then

$$e_{i,j} = \frac{c_y \Delta x_e^2 + c_x \Delta y_e^2}{c_x + c_y}. \tag{2.34}$$

For an *MxN* flow field, the full weighted error, based on the state variance, $\sigma_{\hat{x}}^2$, is provided by the following expression,

$$E_w = \frac{\sum_{i,j} \frac{e_{i,j}}{\sigma_{x_{i,j}}^2}}{\sum_{i,j} \frac{1}{\sigma_{\hat{x}_{i,j}}^2}}. \tag{2.35}$$

When $c_x = c_y$, no information about how to combine the measurements is available. As such, $e_{i,j}$ becomes the squared average of $(\Delta x_e, \Delta y_e)$. Thus, the average error over the flow-field is provided by the following expression,

$$E_A = \frac{\frac{1}{2}\sum_{i,j}\left(\Delta x_e^2 + \Delta y_e^2\right)}{MxN}. \tag{2.36}$$

To evaluate how effective the confidence values are for merging measurements over the full field, the following ratio of E_W and E_A is computed as

$$E_{Gain} = \frac{E_A - E_w}{E_A}. \tag{2.37}$$

Table 2.3 shows a compilation of results for synthetic sequences and a pair of real zero-flow sequences. Two values are provided in each cell. The top value is the effective confidence ratio E_{Gain}, the second in parentheses indicates the percentage of flow estimates in the image that had

$$e_{i,j} < \frac{\Delta x_e^2 + \Delta y_e^2}{2}. \tag{2.38}$$

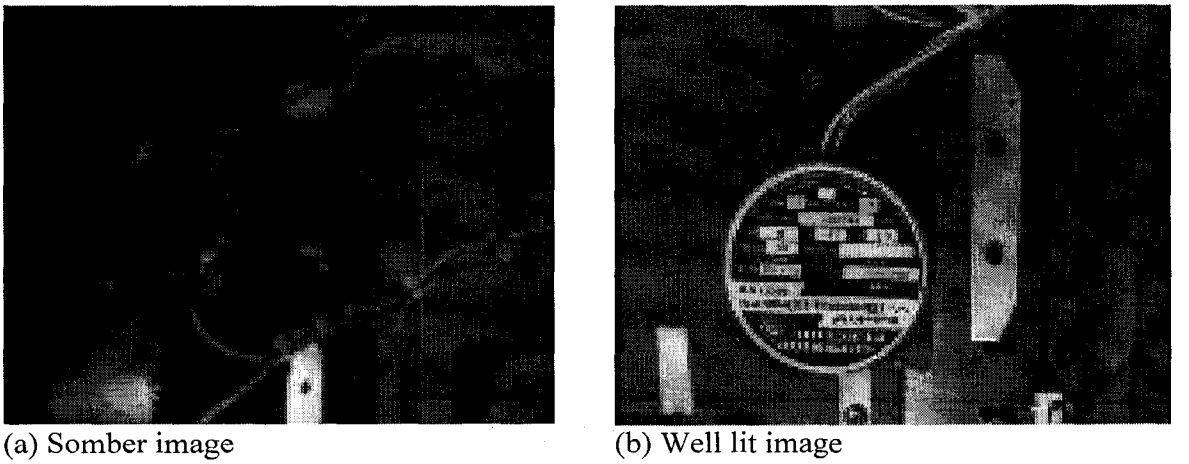
As such, comparing these two values provides a measure of how effective the state variance, $\sigma_{\hat{x}}^2$, is for weighing the errors of the *MxN* field.

	Effective Confidence Ratio (%) (Improved Estimate Ratio (%))					
Confidence Type	Sine-B	Trans. Tree	Diver. Tree	Yosemite (no sky)	Still Camera (Well lit)	Still Camera (Somber)
Differential	3.05444% (53.09%)	10.3638% (59.70%)	-1.2124% (43.71%)	72.2748% (58.01%)	67.0919% (8.69%)	78.3251% (55.93%)
Singh	0.1851% (77.37%)	61.7852% (66.85%)	-0.4616% (40.69%)	11.6308% (57.85%)	61.9962% (9.01%)	33.9959% (56.45%)
Matthies *et al.* (*K*=0)	5.8855% (60.73%)	25.1118% (65.11%)	-4.1464% (46.01%)	69.8526% (61.68%)	67.2141% (8.76%)	81.9122% (57.57%)
Anandan (*K*=1)	6.3547% (60.73%)	42.5485% (65.11%)	-1.7531% (46.01%)	64.4385% (61.68%)	66.0995% (8.76%)	74.9688% (58.52%)
Mod. Anandan (*K*=2)	6.6879% (60.73%)	45.0589% (65.11%)	25.4441 (46.01%)	51.6768% (61.68%)	64.8500% (8.76%)	45.7176% (58.52%)
Mod. Anandan (*K*=3)	6.8727% (60.73%)	43.0914% (65.11%)	64.4974% (46.01%)	47.4398% (61.68%)	63.4580% (8.76%)	35.1616% (58.52%)
Mod. Anandan (*K*=4)	6.8991% (60.73%)	41.9948% (65.11%)	77.9971% (46.01%)	47.4786% (61.68%)	61.9814% (8.76%)	35.0015% (58.52%)

Table 2.3 – A study of how different optical flow confidence values model noise covariance in the context of a maximal estimation framework. Two results are provided for each table entry. Comparing these results establishes how well the confidence values reduce error when combining the x,y components of a single measurement (2.38), as well as how good confidence values are for merging many measurements (2.37).

The results presented in **Table 2.3** show that the most consistent estimator is the generalized Anandan method for *K=2, 3* and *4*. It is consistently positive for all sequences. It out-performs Singh's method five times out of six. It performs similarly or slightly better than the differential confidence measure five times out of six. Also, the *K=2, 3* and *4* results provide more consistent results than the *K=0, 1* implementations, which return negative values for the translating and diverging tree sequences.

To determine how stable these confidence estimators are under noisy conditions, two real image sequences were tested. These involved taking two pairs of still pictures under well and poorly lit conditions (**Figure 2.4**). The ground truth flow is assumed to be a zero vector flow field. Comparing these two sets of results indicates how sensitive the confidence measurement is to noise. It is noted that the differential method is relatively insensitive to noise. This is due to the spatial smoothing process involved in computing I_x and I_y. Both Singh and Anandan's methods are sensitive to noise. Singh's method performs much more poorly under the somber conditions. It is noted that as K gets bigger, Anandan's approach becomes more sensitive, as noise is accumulated in the SSD distribution, and the SSD_{min} value gets amplified. For this reason $K=2$ is chosen over $K=3$ *or* 4.



(a) Somber image (b) Well lit image

Figure 2.4 – Real images for zero flow confidence tests.

CHAPTER 3

Small Motion Surface Reconstruction

The problem of shape-from-motion is defined in two parts: the recovery of viewer motion parameters, followed by 3-D shape reconstruction from a group of images [1, 31, 54, 63]. This part of the thesis is mainly concerned with the second part of this problem -- recovering structure given that the motion has already been determined up to a given certainty. For a two-image approach, the literature sometimes groups this problem with stereovision. An important discrepancy with stereovision problem resides in the fact that the motion is not known deterministically in the current problem. As such, it is probably more appropriate to call this problem monocular motion-stereo. Independently of the nomenclature, this problem clearly does draw strongly from the area of traditional stereovision.

In this chapter different concepts of stereovision and shape-from-motion will be examined. The projective geometry for estimating depth from a pair of images will be reviewed, methods for accumulating depth estimates will be examined and the Kalman filter framework of Matthies *et al.* will be described. The latter will be modified to include maximal-estimation for spatial support in the depth interpolation process. This is a novel improvement to the Kalman framework, and as such, results will be provided to support this claim.

1. Perspective Projection Stereo

The notation used in this thesis refers to points in 3D space using capital letters. Points in the camera's image plane are denoted using lower case characters. Bold characters indicate homogeneous motion operators. A subscript i is used to denote different iterations of the camera motion being considered.

A pinhole camera of focal length f is assumed and a viewer-based coordinate system is adopted (**Figure 3.1**). The origin is at the focal point of the camera. The image plane is at $Z = f$. The Z-axis runs along the optical axis, and the X- and Y-axes are parallel to the x- and y-axis of the image plane respectively. This is a common right-handed projection model for most shape-from-motion and stereovision problems.

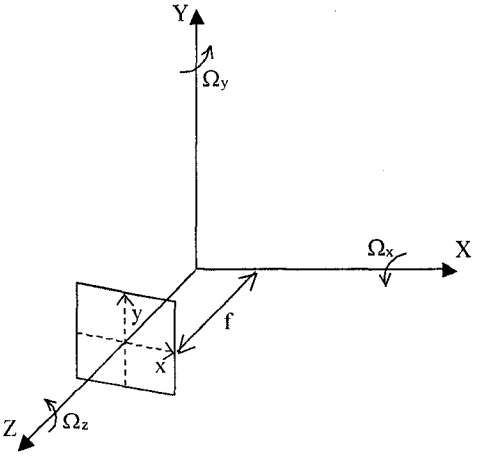


Figure 3.1 – Pinhole camera of focal length f with a viewer-based coordinate system where, $\mathbf{\Omega_i} = (\Omega_x\ \Omega_y\ \Omega_z)$, about an axis passing through the origin, and a translation, $\mathbf{T_i} = (T_x\ T_y\ T_z)$.

Equation (3.1) provides mathematical representations for the pinhole camera,

$$\begin{bmatrix} x_i \\ y_i \\ f \\ f/Z_i \end{bmatrix} = \begin{bmatrix} f/Z_i & 0 & 0 & 0 \\ 0 & f/Z_i & 0 & 0 \\ 0 & 0 & f/Z_i & 0 \\ 0 & 0 & 0 & f/Z_i \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}. \tag{3.1}$$

The ego-motion of the camera is decomposed into a rotation about an axis passing through the origin, and a translation. Any three-dimensional motion can be represented as such [20]. This is denoted as $(T_i \circ \Omega_i)$, where ° indicates a composite operator. Given a point in three-dimensional space, (X_i, Y_i, Z_i), and a camera motion, $(T_x, T_y, T_z)_i \circ (\Omega_x, \Omega_y, \Omega_z)_i$, the new location of the point, $(X_{i+1}, Y_{i+1}, Z_{i+1})$, is given by (3.2). This homogenous operator assumes a small rotation approximation, $\sin\theta \approx \theta$,

$$\begin{bmatrix} X_{i+1} \\ Y_{i+1} \\ Z_{i+1} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & -\Omega_z & \Omega_y & T_x \\ \Omega_z & 1 & -\Omega_x & T_y \\ -\Omega_y & \Omega_x & 1 & T_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}. \tag{3.2}$$

From expressions (3.1) and (3.2) an expression for the projected motion, $(\Delta x_i, \Delta y_i)$, of a point on the image plane can be derived [54]. For clarity, the iteration subscript, *i*, is dropped from this point on. Thus,

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \frac{1}{(Z - T_z)} H\mathbf{T} + R\boldsymbol{\Omega}, \tag{3.3}$$

where,

$$H = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix}, \tag{3.4}$$

and

$$R = \begin{bmatrix} xy & -(1+x^2) & y \\ (1+x^2) & -xy & -x \end{bmatrix}. \tag{3.5}$$

Matthies *et al.* derive a similar expression,

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \frac{1}{Z} H\mathbf{T} + R\boldsymbol{\Omega}. \tag{3.6}$$

They use a differential formulation that assumes infinitesimal image sampling. This assumption is correct as long as any motion in the Z-direction is much smaller than the depth of the point (i.e. $T_z << Z$). If this condition is met, expression (3.3) and (3.6) can be considered equivalent. Considering that a small motion constraint is imposed on the autonomous viewer system described here, this assumption is deemed valid, and (3.6) is adopted instead of (3.3).

2. Multi-Image Depth Accumulation

There are two general approaches to multi-image shape-from-motion. The first involves simultaneously considering all data collected for computing a minimum solution. This is referred to as the batch method [31, 54, 63]. This approach is elegant, and generally very robust, but it does not lend itself well to an active autonomous scenario in which real-time interaction is required. The second approach is recursive and often takes the form of a Kalman filter. This approach and related elements in the literature are described in the following section.

2.1 The Kalman Filter

The Kalman filter is a weighted sum mechanism derived from maximum-estimation theory. It is often used in robotics and shape-from-motion application for temporal integration of data [6, 9, 11, 16, 30, 35, 38, 46, 60, 70]. The different

components of a Kalman filter are: measurement model, state transition model and update phase [18].

The measurement model relates a current measurement to an estimated current belief about the state, $\vec{x}(k)$, of the world. The forward measurement process, where Λ represents the linear relationship between the expected measurements given a current state, and $\vec{w}(k)$ is the noise function, is described as such,

$$\vec{z}(k) = \Lambda\vec{x}(k) + \vec{w}(k). \tag{3.7}$$

This relationship is generally not invertible. If $\vec{w}(k)$ is zero-mean and Gaussian with covariance matrix C_w, least-squares minimum solution can be estimated [48],

$$\hat{x}(k) = \left(\Lambda^T C_w^{-1}\Lambda\right)^{-1}\Lambda^T C_w^{-1}\vec{z}(k), \tag{3.8}$$

as well as its associated covariance matrix,

$$P_k = \left(\Lambda^T C_w^{-1}\Lambda\right)^{-1}. \tag{3.9}$$

The state transition model provides a prediction of the new state given a current action, $\vec{u}(k)$, is applied to the system. For linear systems, a state transition matrix, Γ, models physical characteristics of the state-space and takes the form of a predictive element (interpolator or extrapolator). This is represented as

$$\hat{x}(k+1 \mid k) = \Gamma\vec{u}(k) + \vec{\gamma}(k), \tag{3.10}$$

where $\vec{\gamma}(k)$ is zero-mean Gaussian noise with covariance matrix C_γ introduced into the system during state transitions. The associated projected covariance matrix is obtained from

$$P_{k+1|k} = \Gamma P_k \Gamma^T + C_\gamma. \tag{3.11}$$

The update phase is used for integrating the current measurement with the predicted state to provide a new current state estimate

$$\hat{x}(k+1) = \hat{x}(k+1 \mid k) + K_{k+1}\left(\bar{z}(k+1) - \Lambda\hat{x}(k+1 \mid k)\right), \tag{3.12}$$

with associated covariance matrix

$$\hat{P}_{k+1} = (I - K_{k+1}\Lambda)P_{k+1|k} . \tag{3.13}$$

The Kalman gain, K_{k+1}, is used to weigh the importance of the new state estimate, obtained from the measurement, with respect to the predicted estimate, obtained from the state-transition model. Mathematically,

$$K_{k+1} = P_{k+1|k}\Lambda^T\left(\Lambda P_{k+1|k}\Lambda^T + C_w\right)^{-1} . \tag{3.14}$$

The Kalman filter framework has the important property of providing a provable least-squares solution to a state-space estimation problem. An important component of this solution includes confidence measure in the form of a covariance matrix. The following criteria are required of a state-space model for it to be optimal in the context of a Kalman filter:

- Zero mean system noise.
- Independent noise.
- A linear model for evolution over time.
- A linear relationship between the system state and the measurement made.

It is often impossible to ensure that all the above requirements are met. This does not, however, imply that the Kalman filter cannot be used. Rather, it suggests that the estimator may not be optimal. Despite non-optimal conditions, there is much work in the

areas of computer vision and robotics that supports the claim that the Kalman filter is a valid tool for accumulating noisy measurements [18, 48, 60]. Most often the Kalman filter is used for tracking and motion estimation problems. There are however applications in which this mechanism is used for depth estimation.

Beardsey *et al.* [9] demonstrate how an Iterated Extended Kalman Filter (IEKF) can be used for feature-based depth accumulation. This work shows how corners can be used to reconstruct planar scenes in an autonomous navigation scenario. The extended Kalman filter is used to approximate a non-linear projective measurement model by a first order Taylor expansion. Several iterations (the authors suggest three) are used to improve the linearized approximation.

Kumar *et al.* [38] also develop a feature based Kalman framework. They use a two-step approach in which shallow structure is first estimated, and in the second step model refinement and extension are applied. The first step is obtained from a pseudo-batch approach, while the second involves a Kalman mechanism.

Azarbayekjani *et al.* [6] use an Extended Kalman Filter (EKF) framework to extract 3-D motion parameters and point-wise depth for rigid objects in a scene. They follow Tomasi and Kanade's [63] approach for extracting image features. Motion is described with respect to the camera frame, while depth is described with respect to the object's coordinate frame.

Although, the Kalman framework is generally used for feature tracking, it is also used for iconic depth estimation. In such a framework it is standard practice to treat each depth element as being independent, despite the fact that these local estimates are not usually independent. This is necessary to make the solution computationally tractable

[30, 35, 46, 70]. As such, the image becomes an array of scalar and independent Kalman filters to which spatial support is applied separately. Several different approaches to iconic depth accumulation exist in the literature.

A particularly interesting approach is Heel's work [30]. The author shows how to integrate the image flow constraint (2.1) into the Kalman framework using the following direct depth measurement (3.15), thus avoiding the computation of the optical flow,

$$Z = \frac{-\bar{s} \cdot \mathbf{T}}{E_t + \bar{v} \cdot \boldsymbol{\Omega}}, \tag{3.15}$$

where,

$$\bar{s} = \begin{bmatrix} -I_x \\ -I_y \\ xI_x + yI_y \end{bmatrix}, \quad \bar{v} = \begin{bmatrix} xyI_x + (1+y^2)I_y \\ -xyI_y - (1+x^2)I_x \\ yI_x - xI_y \end{bmatrix}. \tag{3.16}$$

Hung and Ho [35] build on Heel's approach. They use the image gradient in the predictive phase of the filter. They also integrate a local smoothness constraint into their framework. Just as the gradient optical flow methods, this approach requires temporal derivation. Thus, sufficient temporal support and smoothness must be supplied for proper temporal numerical differentiation.

Xiong and Shafer [70] implement an iconic depth estimator using an Extended Kalman Filter. They concentrate on augmenting the depth uncertainty with motion information. They suggest mathematical techniques such as Sherman–Morrison-Woodbury matrix inversion and a weighted principal component analysis framework for making the approach computationally tractable. They use the current depth estimate to bootstrap the next motion estimate.

The Kalman framework used in this paper is very closely related to that of Matthies *et al.* The small motion assumption makes the depth-estimation formulation expressed in (3.6) valid. This approach is advantageous as it provides a linear measurement model using the inverse-depth (or disparity) instead of the true depth as the state variable. The use of a correlation-based optical flow estimator provides temporal robustness for non-smooth image sequences and a scalable, efficient computational framework.

The measurement model is described first. As mentioned in the introduction, the correspondence problem and triangulation problem are ill-posed and ill-conditioned, respectively. A qualified estimator of the projected motion should reflect and weight these issues. In the same manner as (3.8), Matthies *et al.* provide a least-squares solution to this problem. They begin by rectifying the image by removing the rotational component of the optical flow (which is independent of the depth),

$$\begin{bmatrix} \Delta x_r \\ \Delta y_r \end{bmatrix} = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} - R\boldsymbol{\Omega} . \tag{3.17}$$

The least-squares disparity estimate can then be obtained from the rectified optical flow as

$$\hat{d}_m = \left((HT)^T C_f^{-1} (HT)\right)^{-1} (HT)^T C_f^{-1} \begin{bmatrix} \Delta x_r \\ \Delta y_r \end{bmatrix}, \tag{3.18}$$

where C_f represents the covariance of the measurement error of the optical flow vector,

$$C_f = \begin{bmatrix} \sigma_{\Delta x}^2 & 0 \\ 0 & \sigma_{\Delta y}^2 \end{bmatrix}. \tag{3.19}$$

The corresponding variance, σ_d^2, of the disparity estimate is obtained from

$$\sigma_d^2 = \left((HT)^T C_f^{-1}(HT)\right)^{-1} = \frac{\sigma_{fx}^2 \sigma_{fy}^2}{\sigma_{\Delta y}^2 (xT_z - fT_x)^2 + \sigma_{\Delta x}^2 (yT_z - fT_y)^2}. \quad (3.20)$$

For baseline stereovision ($T_z \equiv 0$) expression (3.20) illustrates the advantage of a larger baseline motion mentioned in the Introduction of this thesis. A larger baseline between views provides more robust triangulation of a 3-D point. This becomes apparent from (3.20) as the magnitude of the baseline, $\|(T_x, T_y)\|$, increases, σ_d^2 decreases. As the autonomous system here has limited control over the magnitudes of T_x and T_y parameters, it is only natural to instead take advantage of an active strategy to maximize $\sigma_{\Delta x}^2$ and $\sigma_{\Delta y}^2$. This will be the approach used for selecting the motion strategy described in the next chapter.

For looming motions, where $T_z \neq 0$, a vanishing point is observed where the vector connecting the two centers-of-projection of the two images traverses the projection plane. This point of contraction/expansion is referred to as the focus-of-expansion (FOE) and is computed from (3.20) as

$$(x_{FOE}, y_{FOE}) = \left(\frac{fT_x}{T_z}, \frac{fT_y}{T_z}\right). \quad (3.21)$$

The depth estimate becomes increasingly ill-conditioned around the FOE, while it is completely ill-posed at the FOE.

Matthies *et al.* introduce spatial support into their system using interpolation and regularization stages [62]. They suggest that the state transition model can be treated equivalently to a polygonal mesh. Thus, the iconic depths are transformed as if they were a polygonal mesh under homogenous transformation. Heel [30], and Xiong and Shafer [70] use a similar approach to interpolation. A bilinear interpolation scheme, similar to

Gouraud shading is used to predict the new variance values of the predicted depth estimates. This can be expressed as such: given a triplet of connected disparity estimates on the surface, d_0, d_1 and d_2, the new disparity value, d_i, is computed as

$$d_i = w_{0i}d_0 + w_{1i}d_1 + w_{2i}d_2, \tag{3.22}$$

where w_0, w_1 and w_2 represent the weighted distances to the interpolated disparity, d_i, for each point, d_0, d_1 and d_2, respectively. The associated variance for the inverse depth is computed by pre- and post-multiplying the Jacobian of (3.22) onto the covariance matrix constructed from σ_{d0}^2, σ_{d1}^2 and σ_{d2}^2. This effectively results in

$$\sigma_{di}^2 = w_0^2\sigma_{d0}^2 + w_1^2\sigma_{d1}^2 + w_2^2\sigma_{d2}^2. \tag{3.23}$$

The authors then suggest that a pure blend may be used to interpolate the new confidence values,

$$\sigma_{di}^2 = w_0\sigma_{d0}^2 + w_1\sigma_{d1}^2 + w_2\sigma_{d2}^2. \tag{3.24}$$

The interpolation method of Matthies *et al.* and Heel leads to an increase in uncertainty when interpolating. Information theory suggests the opposite; on average, conditional entropy of a random variable should not increase as more measurements are combined into an estimate [15]. The entropy of a random variable measures its uncertainty. As such, uncertainty should not increase. Thus,

$$\sigma_{di}^2 \leq \mathbf{min}\left(\sigma_{d0}^2, \sigma_{d1}^2, \sigma_{d2}^2\right). \tag{3.25}$$

The upper bound for expressions (3.24) and (3.23) is

$$\sigma_{di}^2 \leq \mathbf{max}\left(\sigma_{d0}^2, \sigma_{d1}^2, \sigma_{d2}^2\right), \tag{3.26}$$

which implies that the approach used by Matthies *et al.* and Heel does not conform to basic information theory.

In the implementation presented here, regularization is dropped and the maximum-estimation approach is extended to the prediction procedure. It seems that, as the spatial and temporal estimation processes have already been decoupled, and that confidence information is available from the temporal estimator, a maximal estimation approach to the spatial interpolation of the surface is the natural extension to the current framework. Work by Mathur and Ferrie explains how to do this for local curvature models such a Bezier frames [47]. The approach taken here will involve a simpler local surface model -- the triangle. As such, the depth interpolator is described as follows

$$d_i = \left(\frac{w_{0i}}{\sigma_{d0}} d_0 + \frac{w_{1i}}{\sigma_{d1}} d_1 + \frac{w_{2i}}{\sigma_{d2}} d_2 \right) \Bigg/ \left(\frac{w_{0i}}{\sigma_{d0}} + \frac{w_{1i}}{\sigma_{d1}} + \frac{w_{2i}}{\sigma_{d2}} \right). \tag{3.27}$$

The variance associated to the new disparity value is

$$\sigma_{di}^2 = \sigma_{d0}^2 \sigma_{d1}^2 \sigma_{d2}^2 / \left(w_0^2 \sigma_{d1}^2 \sigma_{d2}^2 + w_1^2 \sigma_{d0}^2 \sigma_{d2}^2 + w_2^2 \sigma_{d0}^2 \sigma_{d1}^2 \right). \tag{3.28}$$

This approach conforms to the rules of information theory. It performs well provided that the linear interpolation model is correct. Computer graphics theory has shown that for dense depth fields this is an acceptable assumption [24].

The last step in the Kalman framework is the update phase. The Kalman gain is computed as

$$K_{i+1} = \frac{P_{k+1|k}}{P_{k+1|k} + \sigma_d^2}, \tag{3.29}$$

where p_k represents the current depth estimate covariance. The new measurement is integrated into the current disparity estimate as such

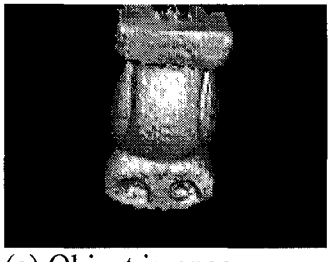
$$\hat{d}_{i+1} = d_{i+1|i} + K_{i+1} \left(\hat{d}_m - d_{i+1|i} \right) \tag{3.30}$$

and the updated confidence is obtained as

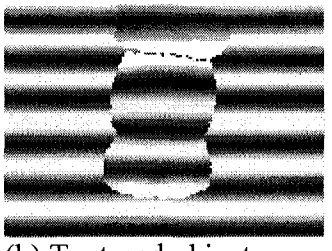
$$P_{k+1} = \frac{P_{k+1|k}}{P_{k+1|k} + \sigma_d^2} \,. \tag{3.31}$$

3. Results

This section presents results for the depth accumulation procedure described above. Synthetic data is used to demonstrate that the maximum estimation approach to spatial interpolation outperforms the previous approach used by Matthies *et al.* The synthetic environment is constructed from the rendering of a range image (an owl). The object is placed 3 units away from the viewer. A plane is placed perpendicular to the camera's viewing axis 6 units away (**Figure 3.2**). A horizontal texture is applied to reduce the aperture problem along the vertical direction. The viewer performs fifteen up-and-down iterations of **T** = (0,0.044,0) and **T** = (0,-0.044,0).



(a) Object in space



(b) Textured object

(c) Ground truth depth

Figure 3.2 – Sample images of synthetic experimental setup.

The error in estimated depth is measured as the root-mean-square (RMS) of the difference between the estimate and the ground truth over the full *MxN* depth image,

$$Err = \sqrt{\frac{1}{MxN}\sum\left(d_{gt}(i,j) - d_{Est}(i,j)\right)^2} \tag{3.32}$$

Although this metric is not ideal for providing a global description of the reconstruction, it is certainly reasonable for comparing convergence rates for the depth estimation process.

Figure 3.3 provides convergence plots for the three different interpolation methods suggested. The square weighted sum provides the worst results. The maximal estimation method provides the best results by converging to a lower RMS-Error. **Figure 3.4**, **Figure 3.5** and **Figure 3.6** depict confidence, depth and error maps associated to the fifteenth iteration of the estimation process. Confidence is represented as $I_c(i,j) = -\log(p_k(i,j))$. Thus brighter intensity indicates higher confidence. For the error maps, $I_{Err}(i,j) = (d_{gt}(i,j) - d_{est}(i,j))^2$, and as such darker intensity imply less error.

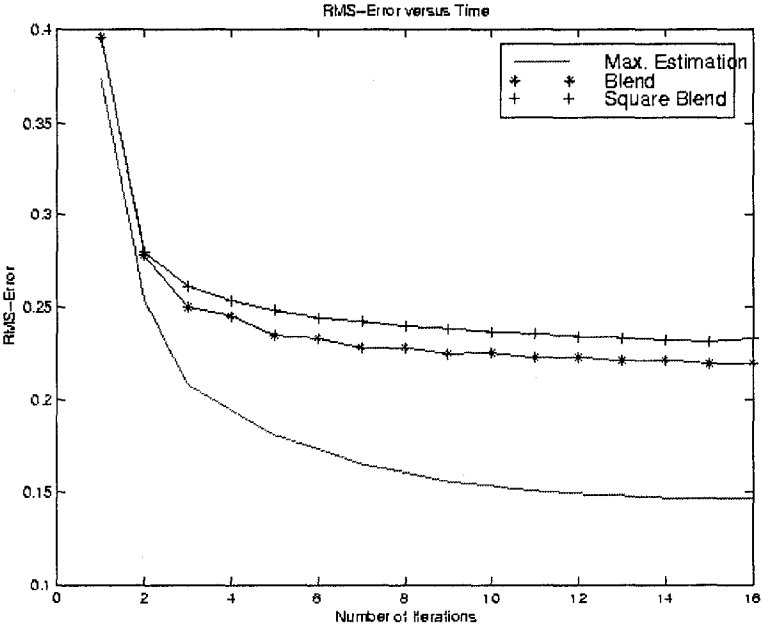
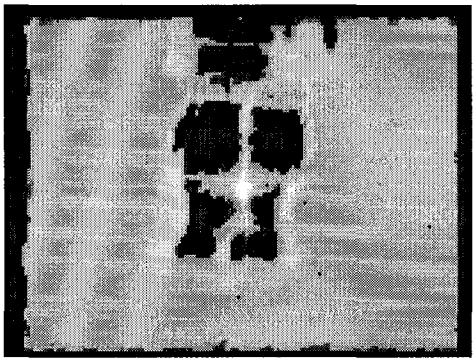


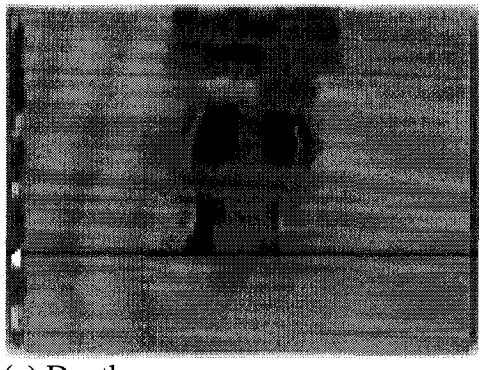
Figure 3.3 – RMS-Error as a function of time for squared interpolation method, for blend method, and for maximal-estimation method.

Figure 3.4 and **Figure 3.5** indicate that Matthies *et al*'s approach relies heavily on regularization to guess the surface where the measurement has low confidence. When the regularization process is removed, less confident estimates have the upper hand and propagate. This results in large holes in the confidence maps. Even if the regularization process were included, no framework for generating new confidence values is provided for the interpolated depths elements. As well, this method fails to take advantage of confidence measures already available. The maximal estimation approach provides a complete, compact and robust method for simultaneously interpolation and propagating information to areas of low confidence, as can be seen when examining **Figure 3.6** in contrast to results presented in **Figure 3.4** and **Figure 3.5**[4].



(a) Confidence map

(b) Error map



(c) Depth map

[4] Figures were all scaled linearly. Confidence values ranged between 0 and 1e30. Depth and error values ranged between 0 and 30.

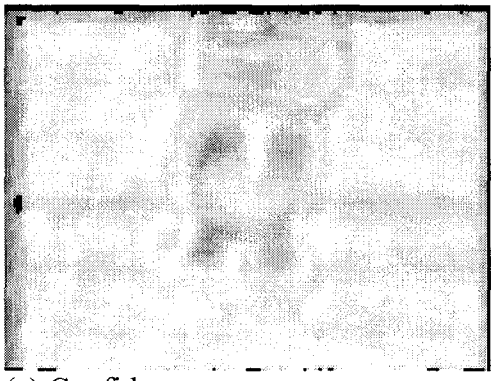
Figure 3.4 – Confidence, depth and error maps after 15 iterations of SQUARED interpolation method.

(a) Confidence map

(b) Error map

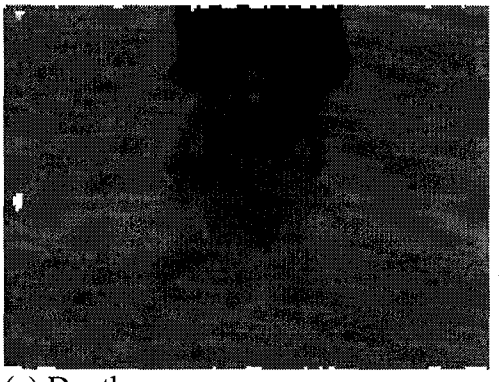
(c) Depth map

Figure 3.5 – Confidence, depth and error maps after 15 iterations of BLEND interpolation method.



(a) Confidence map

(b) Error map



(c) Depth map

Figure 3.6 – Confidence, depth and error maps after 15 iterations of MAXIMAL-ESTIMATION interpolation method.

CHAPTER 4

Active Reconstruction

1. Active Vision

From earlier discussion on correspondence and depth estimation, it should be clear that both multi-image feature and iconic approaches to surface reconstruction can fail to recover depth for passive motion, even if the images are textured and the viewer motion is not ambiguous. This occurs when the viewer's motion fails to take advantage of image features that require a selective direction for correspondence. **Figure 4.1** shows an example of a horizontally textured scene and its associated depth maps after 10 horizontal and vertical motions. **Figure 4.2** shows the RMS-Error of the measured depth for several iterations of horizontal and vertical motions. This result suggests that a passive viewer moving horizontally will fail to recover depth for this scene. Thus, it is suggested in this thesis that an active control strategy should be adopted to attempt to maximally extract information for image features.

(a) Horizontally textured scene (b) Depth map after 10 horizontal motions (c) Depth map after 10 vertical motions

Figure 4.1 – Vertical and horizontal motions for reconstructing a horizontally textured scene.

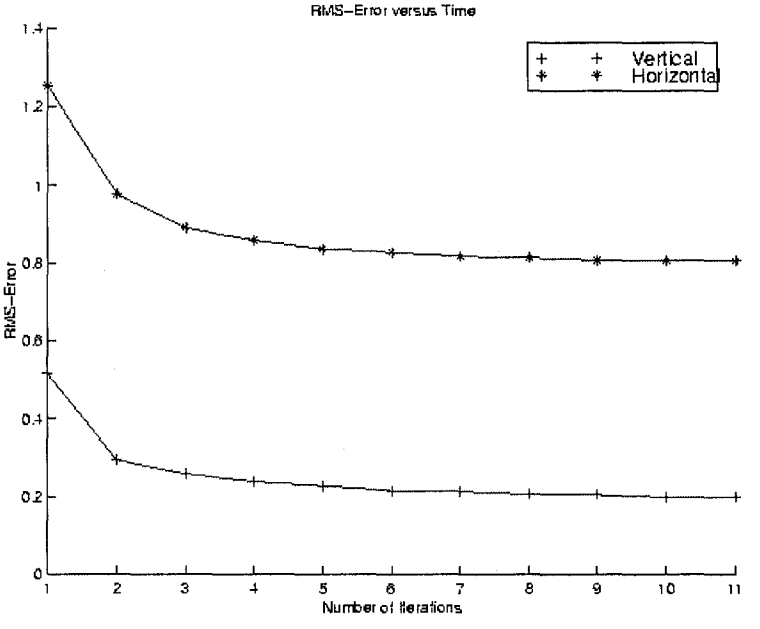


Figure 4.2 – RMS-Error plot of horizontal and vertical motions for the horizontally textured scene.

Active vision generally provides two approaches for reducing the ambiguity that results in the measurement process: active sensing, and passive sensing from an active viewer. The first approach involves controlling the lighting conditions in the scene such that ambiguity in the image formation process is removed. Active photometry and laser-range finders are examples of such systems [14, 64, 68]. These methods provide good

results under the assumption that the lighting over the sampled surface is controllable and that the reflective properties of the surface are sufficient for correct image formation. Such methods, clearly, also assume that the projection of a laser or colored light will not impede the measurement or cause detriment to the environment. These conditions do not necessarily lend themselves well to exploration of an unknown environment, as surface characteristics are generally unknown *a priori.*

The second approach involves a passive sensor and an active viewer. As such, the camera geometry is used to constrain the depth estimation process. For this thesis, the camera is moved to provide multiple samples of the surface from different points of view. The relationship between the viewpoints is actively controlled to profit from geometrical properties of the sensory apparatus [2]. The active viewer introduces a known relationship (up to a given certainty) between the different views, thus providing some constraints for inverting the correspondence and depth measurement sub-problems. This is often referred to as the epipolar constraint, and is discussed in section 2. Second, the active viewer selects the *next-step* that is predicted to best reduce ambiguity or increase confidence in next measurement. **Figure 1.1** is updated with a new process block for selecting the next motion of the viewer to obtain **Figure 4.3**. This strategy is developed in the section 3.

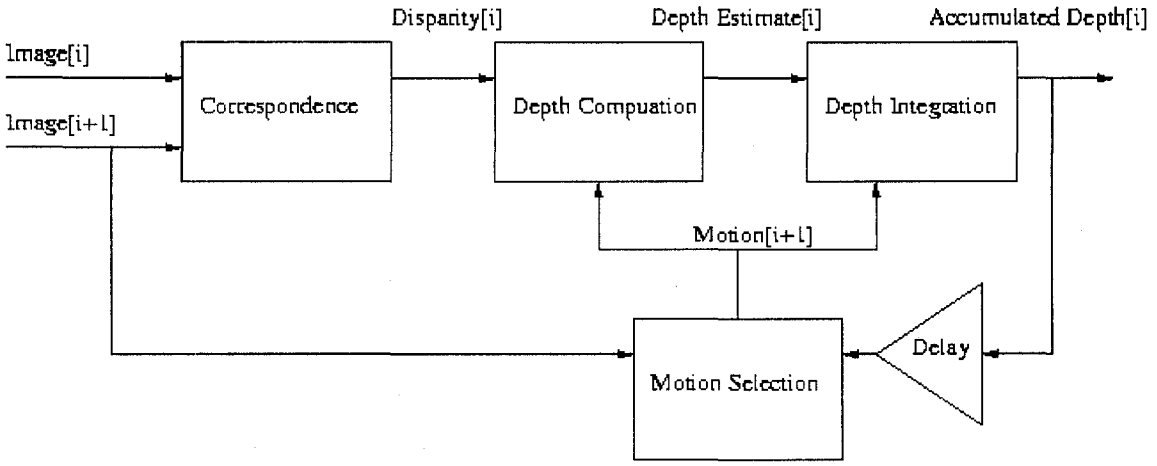


Figure 4.3 – Block diagram of active depth accumulation system. A new component is added that uses depth confidence values and image features to select a new motion.

2. Epipolar Constraint

An important element in the geometry of stereovision is the epipolar constraint [21]. Epipolar geometry constrains the angular components of an optical flow field. When the motion parameters between two views are fully known, the search space in the correspondence problem can be restricted to a line.

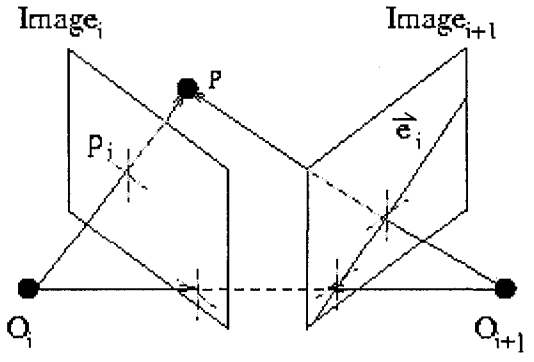


Figure 4.4 – Epipolar geometry for two image frames with respect to a point *P*.

As depicted in **Figure 4.4**, a ray can be constructed by connecting the centers of projection, O_i and a 3-D point, P. This ray projects to a point p_i in the first image. The ray, $\overline{O_i P}$, projects to a line, $\vec{e}_i$, in the second image. This suggests that all 3-D points that project to p_i in the first image must project onto the line $\vec{e}_i$ in the second image. The essential matrix represents this structure mathematically [21, 64],

$$E = \begin{bmatrix} 1 & -\Omega_z & \Omega_y \\ \Omega_z & 1 & -\Omega_x \\ -\Omega_y & \Omega_x & 1 \end{bmatrix} \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \tag{4.1}$$

Thus, it is sufficient to search along this line to match the pixels in the two images. This can be interpreted as an angular constraint on the flow measurement. The epipolar line in the second frame, given that the image point in the prior frame is (x_p, y_p, f), is computed as follows,

$$\begin{bmatrix} m_x \\ m_y \\ b \end{bmatrix} = E \begin{bmatrix} x_p \\ y_p \\ f \end{bmatrix}, \tag{4.2}$$

given the following form for the equation of a line,

$$y = \frac{m_y}{m_x} x + b. \tag{4.3}$$

The offset term b is ignored as it only contains information about the position of the epipolar line in the new image. This information is available under the trivial condition where a null motion is applied and x_p does not move. When expanding (4.1) and (4.2), the terms m_x and m_y are described as

$$m_x = -(\Omega_z T_z + \Omega_y T_y) x_p + (\Omega_y T_x - T_z) y_p + (T_y + \Omega_z T_x) f, \tag{4.4}$$

$$m_y = \left(T_z + \Omega_x T_y\right)x_p - \left(\Omega_z T_z + \Omega_x T_x\right)y_p + \left(T_y \Omega_z - T_x\right)f \,. \tag{4.5}$$

Epipolar geometry can also be used when the motion parameters are not known deterministically. More precisely, an additional constraint that takes into account the expected motion and respective covariance can be imbedded into the correspondence procedure by adding a bias to each cell of the *SSD(i,j)* surface, where *i,j* are the coordinated pixel distances from the center point of the SSD distribution. The bias is derived from the square of the Mahanalobis distance, $M_{ep}(i,j)^2$, of the *i,j*th cell from the expected epipolar line. Thus a robust constraint is used to sway the minimum of the SSD surface in the direction of the epipolar line, and SSD_{min} becomes

$$SSD_{\mathbf{min}} = \mathbf{min}\left(SSD(i,j) + e^{M_{ep}(i,j)^2} \right). \tag{4.6}$$

Uncertainty in the motion parameters and image point parameters can be projected into the epipolar parameter space, to obtain (4.7), by pre and post-multiplying the diagonal covariance matrix by the Jacobian of expressions (4.4) and (4.5),

$$\sigma_{mx}^2 = \left[\frac{\partial(m_x)}{\partial(\mathbf{T},\boldsymbol{\Omega},x_p,y_p)}\right]^T \begin{bmatrix} C_{\mathbf{T}} & 0 & 0 & 0 \\ 0 & C_{\boldsymbol{\Omega}} & 0 & 0 \\ 0 & 0 & \sigma_{xp}^2 & 0 \\ 0 & 0 & 0 & \sigma_{yp}^2 \end{bmatrix} \left[\frac{\partial(m_x)}{\partial(\mathbf{T},\boldsymbol{\Omega},x_p,y_p)}\right]. \tag{4.7}$$

Given that a normalized directional vector at each pixel in the SSD distribution is

$$\hat{f}(i,j) = \frac{(i,j)}{\sqrt{i^2 + j^2}}, \tag{4.8}$$

the squared Mahanalobis distance between the unit epipolar vector, $\hat{m} = (\hat{m}_x, \hat{m}_y)$, and the *i,j*th cell of the SSD is

$$Mep(i,j)^2 = \left(\hat{f}(i,j) - \hat{m}\right) \begin{bmatrix} \frac{1}{\sigma_{mx}^2} & \\ & \frac{1}{\sigma_{my}^2} \end{bmatrix} \left(\hat{f}(i,j) - \hat{m}\right), \tag{4.9}$$

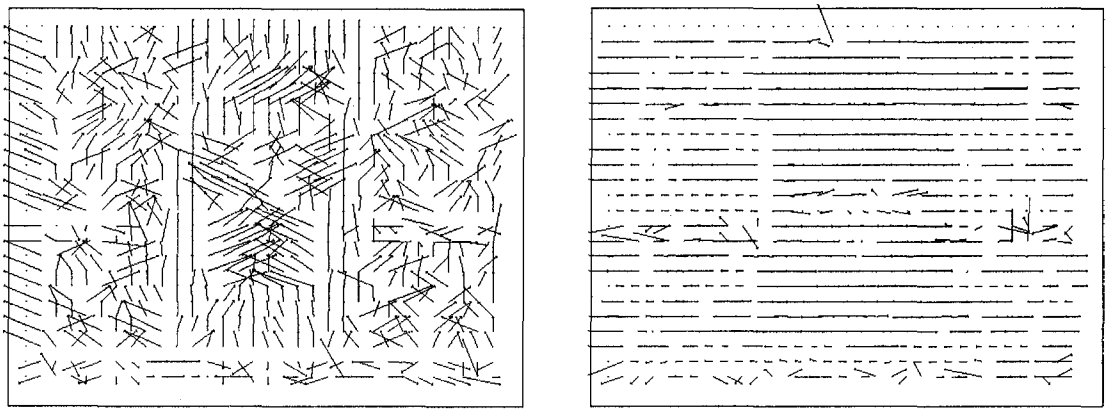
which, after some manipulation, becomes

$$M_{ep}(i,j)^2 = \frac{1}{\sigma_m^2}\left(\left(\hat{f}_x(i,j) - \hat{m}_x\right)^2 + \left(\hat{f}_y(i,j) - \hat{m}_y\right)^2\right), \tag{4.10}$$

where from the symmetry of expressions (4.4) and (4.5),

$$\sigma_m^2 = \sigma_{mx}^2 = \sigma_{my}^2 . \tag{4.11}$$

The bias is added when searching the SSD for its minimum. It is removed again when computing the actual confidence in the flow measurement.



(a) Sample field for vertical motion (b) Sample field for horizontal motion

Figure 4.5 – Flow fields for vertical and horizontal motions in horizontally textured environment.

Figure 4.5 shows sample flow fields for the horizontal and vertical motions when epipolar geometry is ignored. When moving vertically, the y- components of the flow-field are constrained, while the x- components are unstable. In this case, one could say that the magnitude of the flow-field is constrained, while the phase is not. When moving horizontally to the image gradient, the opposite condition occurs -- the image constrains

the angle of the flow field, while leaving the magnitude unconstrained. It is thus necessary to either reduce the ambiguity in the magnitude or the angle of the flow vectors respectively in **Figure 4.5**(a) and **Figure 4.5**(b). Epipolar geometry offers a solution for removing ambiguity in the flow angle. **Figure 4.6** shows the vertical motion when the epipolar constraint is applied to the correspondence process. The epipolar constraint effectively reduces instability of the flow-field's x-components.

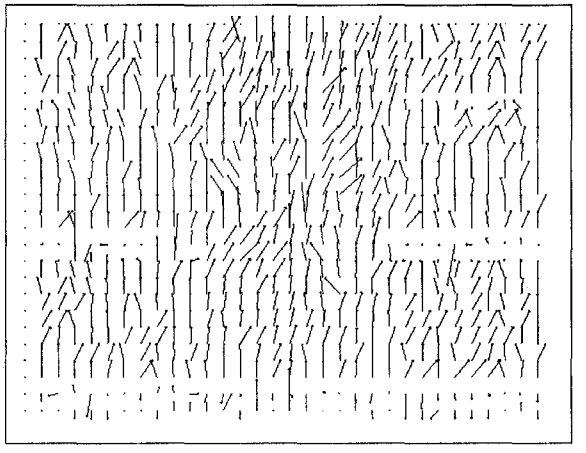


Figure 4.6 – Flow field for vertical motion with epipolar constraint applied.

3. Defining a Motion Space

The first step in building an active viewer is to define the space of all motions from which the next viewpoint must be selected. The union of three different constraints must be considered:

- the holonomic constraints on the viewer,
- computational constraints for correspondence, triangulation and interpolation; and
- a minimal spanning search space from which the motion is selected.

The full configuration space for a viewer can be defined for the six degrees of freedom, $\mathbf{M} = (T_x, T_y, T_z, \Omega_x, \Omega_y, \Omega_z)$. In an unconstrained motion space the vector elements live in unrestricted intervals,

$$\begin{aligned} &T_x,\ T_y \text{ and } T_z \in (-\infty, \infty) \\ &\Omega_x, \Omega_y \text{ and } \Omega_z \in (-\pi, \pi]. \end{aligned} \tag{4.12}$$

The *weakly active* viewer is forcibly restricted to a subset of this space. For the holonomically handicapped application described here, these intervals become

$$\begin{aligned} &T_x,\ T_y, T_z \in (-SMALL, SMALL), \\ &\Omega_x, \Omega_y \text{ and } \Omega_z \in (-\pi, \pi]. \end{aligned} \tag{4.13}$$

The definition of *SMALL* is dependent on the physical parameters of the system at hand. It is, however, desirable to make *SMALL* as large as possible as suggested by expression (3.20). For this reason it will be assumed that the total baseline motion is fixed at K_{Bmax}. The range of the rotations assumes a pan-tilt setup.

Additional constraints must be applied to the T_z, Ω_x, Ω_y and Ω_z due to computational assumptions made in the previous chapter. As described in (3.6) $T_z << Z$. Thus,

$$T_z \in (-SMALL, SMALL) \tag{4.14}$$

and, to satisfy the small rotation constraint of (3.2),

$$\Omega_x, \Omega_y \text{ and } \Omega_z \in \left[-\frac{\pi}{36}, \frac{\pi}{36}\right]. \tag{4.15}$$

Although the magnitudes are small, the motion space remains a six degree space and thus difficult to search quickly. It turns out that the motion space from which the active viewer will select its next step can be further constrained. Justification for further

reducing the configuration space is provided by [19, 46, 61], who all demonstrate that forward motion provides very little information for depth estimation as the ill-posedness of the vanishing point dominates. Thus, T_z will be zero when performing depth accumulation. Similarly, rotations about the focal point (**Figure 3.1**) provide no insight or additional information about the depth of the image, and are thus omitted. It should be noted that this does not imply that the viewer cannot move forward or rotate, as the interpolation process can predict views for small forward and rotational motions. It is just assumed here that the information introduced by such a motion is negligible. As such, these motions are not included in the motion space. This leaves short baseline motions with a fixed magnitude as the motion space available to the active viewer,

$$\sqrt{T_x^2 + T_y^2} = K_{B\max} \ . \tag{4.16}$$

This can be reduced to a single angular parameter,

$$\Theta_{\mathbf{T}} \equiv \tan^{-1}\left(\frac{T_y}{T_x}\right). \tag{4.17}$$

This is a severely restricted 1-D motion space (hence the term *weakly active*), which has the computational advantage of providing a quick solution to the motion selection problem, yet offering a near maximal basis for matching gradient distributions in natural scenes. As a final note concerning the motion of the viewer, each motion $\mathbf{T}_i$ is followed by a motion $\mathbf{T}_{i+1}$ such that,

$$\mathbf{T}_{i+1} = -\mathbf{T}_i = \left(-T_x, -T_y, 0\right). \tag{4.18}$$

This ensures that the scene remains relatively centered in the image.

4. Choosing the Next Motion

This section examines how the viewer can be actively controlled to optimally extract depth information from the intensity projection of a scene. It begins by reviewing key elements in the active vision literature. It then introduces a strategy for selecting a *next-step* based on statistical grouping of image gradient features.

Aloimonos and Bandyopadhyay did early work on the active vision paradigm in their landmark paper [2]. In this work they examine the advantages of an active observer for several shape-from-X problems. In particular, they discuss a multi-baseline approach for recovering Lambertian surfaces to which an adaptive image coordinate system, based on epipolar geometry and isophotes, is applied. They demonstrate that controlled view selection can be used to reduce the ambiguity involved in the correspondence process, and yields a stable and robust framework for shape recovery.

Bajcsy [7] provides a more general methodology for active perception. She defines active vision as an intelligent data acquisition strategy for which measurement parameters, which reflect ambiguity in the scene, are used as a feedback mechanism to the acquisition process. The author discourages regularization by suggesting that the *computational effort should not be spent on processing and artificially improving imperfect data, but rather on accepting imperfect, noisy data as matter of fact and incorporating it into the overall processing stage.*

Whaite and Ferrie [68] suggest a formal mathematical framework in which ambiguity is equated to uncertainty. The measurement model is defined as a general linear system,

$$d = G(x)\, m\,, \tag{4.19}$$

where d is the observation, G is the forward sensor model for a given sensor configuration x (including location), and m is the set of model parameters, which can also be considered as the state vector. A least-squares solution is suggested for inverting (). The associated uncertainty to this solution is,

$$C_m = C_w\left(G(x)^T G(x)\right)^{-1}, \tag{4.20}$$

where C_m is the model covariance that results from projecting the measurement noise into model space. The active viewer chooses a sensor configuration x, such that

$$H = G(x)^T G(x), \tag{4.21}$$

maximally reduces the uncertainty, C_m. The authors apply their theory in the context of autonomous model fitting application. The measurement consists of laser-range data, and the model space is defined as the set of super-ellipsoids.

Arbel and Ferrie [4] develop similar work for selecting the most informative view for autonomous object recognition. Training involves the acquisition of a cross section of short arc optical flow measurements on a tessellated view sphere surrounding the object. The state space is the set of confidence values associated to each object. A Bayesian inference is used to compute the confidence values associated to each object for each viewpoint. Principal components analysis (PCA) is used to build a compact parametric space representing the flow sphere of each object. A set of entropy maps, which provides a measure of distinctiveness for each object given a viewpoint and pose, is also constructed during the training phase. The recognition process starts by placing the viewer at a random pose on the view sphere of an unknown object. The optical flow cross-section at this viewpoint is projected into the PCA space and a confidence value is returned for each object. A Bayesian chaining process is used to accumulate evidence for

each hypothesis object. The winning hypothesis is defined as the object that has accumulates the greatest certainty. Given the hypothesized object and its associated entropy map, the viewer moves to what it estimates is the most informative new viewpoint.

In the context of the image-based surface reconstruction problem, Huang and Aloimonos [34] have developed an approach for obtaining relative (purposive) depth using the normal components of the optical flow. They suggest that when the local intensity has high gradient, the normal optical flow approximates the component of the projected motion field parallel to the image gradient. This agrees with the earlier mentioned analysis of Verri and Poggio [66]. This work fails to provide an accumulation strategy or respective confidence values, and does not suggest a strategy for actively choosing the viewer's motion. It only provides depth estimates where the optical flow happens to be parallel to the image gradient. This results in a sparse depth image and fails to ensure that the full potential of image features is used.

Sandini and Tistarelli [57] also propose a depth estimation system based on normal flow for computing scene depth. They use a DOG operator to extract edges in the image. They perform correspondence on the edges until a sufficient baseline is achieved and then triangulate. As is the case for Huang and Aloimonos, no feedback is applied in the system and the measurements are not qualified. Also, depth measurements are only available along distinguishable edges, and as such are sparse.

The approach presented in this thesis draws inspiration from all the above-mentioned active systems. It attempts to improve the convergence rate of the system by using the image gradient to estimate the most informative camera motion angle, Θ_T.

4.1 Predicting the Most Informative Motion

Arbel and Ferrie, and Whaite and Ferrie's definition of what is the best motion is borrowed here. The objective of the active viewer is to select the most informative motion over the whole *MxN* array of Kalman filters. In the context of the Kalman filter framework, the most informative view is the one that maximally reduces $P_k(i,j)$. This is equivalent to minimizing (3.20) by maximally reducing the values of $\sigma_{\Delta x}^2$ and $\sigma_{\Delta y}^2$. The strategy adopted to obtain this behaviour is described as follows:

i) where the gradient information is unidirectional, the viewer should be directed to move parallel to the image gradient, thus providing the best measurement and maximal information;

ii) in the opposite case where the aperture problem is negligible, the choice of the motion is less important, as, ideally, any motion should provide an equivalent increase in information; and

iii) when no intensity information is available, the point should be ignored, as it provides no contribution to the solution and is completely dependent on the interpolation process.

These characteristics are fully encompassed by the eigenvalues, λ_1 and λ_2, of the Normal matrix (2.7), where $\lambda_1 > \lambda_2$. **Table 4.1** provides an intuitive association of eigenvalues to the three conditions mentioned above:

λ_1	λ_2	Condition Number
LARGE	LARGE	(ii)
LARGE	SMALL	(i)
SMALL	SMALL	(iii)

Table 4.1 – Interpretation of Normal matrix eigenvalues.

Extending this idea to a full *MxN* array (image) of depth estimates, $d(i,j)$, involves developing some statistical tools. The approach taken here strongly resembles that of the Hough transform. Thus, a weighted histogram approach is adopted. The histogram represents a voting function in which each patch votes according to its gradient angle. The gradient angle with the most votes is adopted as the best motion angle, Θ_T. The weight, $w(i,j)$, of each depth element's vote is set according to the system variance of the respective Kalman filter $P_k(i,j)$ and the predicted conditioning of the patch according to $N(i,j)$. The weighting function should have the following characteristics:

- be strong for large system uncertainty when the aperture problem prevails,
- be weak for large system uncertainty where there is no aperture effect, and
- be weak when the system is very certain of its estimate.

The suggested expression for the weighting function is

$$w(i,j) = P_k(i,j)\frac{\lambda_1(i,j)}{\left(1+\lambda_1(i,j)\right)\lambda_2(i,j)}. \tag{4.22}$$

The choice of the inverted λ_2 term in (4.22) is based on the observation made by Barron *et al.* [8] and, Fleet and Langley [23] that the normal matrix predicts the aperture

problem for the condition where $\lambda_2 < 1.0$. The λ_1 ratio is used to neglect votes of elements where no gradient information is available.

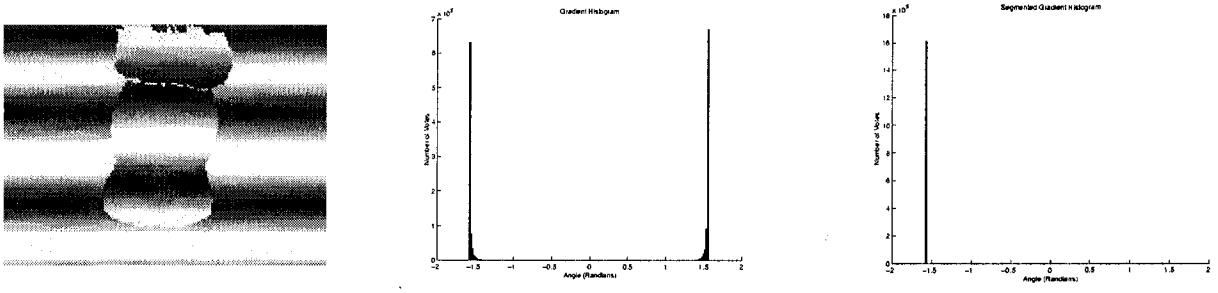
It is accepted that most natural and indoor scenes contain some form of structured gradient information. The traditional feature-based stereovision approaches have fit parametric models (e.g. lines, corners, polygons) to image pairs, thus taking advantage of spatial relationships between these somewhat invariant features to provide robust correspondence. Difficulties arise in the fitting process, which can be time consuming and ill-conditioned. Additional difficulties arise in matching these high-level features which may be numerous, small and difficult to detect.

In the context of the gradient-based weighted histogram, the spatial structure of the features is ignored. However, there still remains a strong relationship between the gradient-structures in the image and the histogram's distribution. Generally, different features with common intensity orientations will result in a peak. As there may be several dominant orientations in the image, several such peaks may occur. To distinguish these features, some form of clustering is necessary for segmenting the gradient distribution histogram. A slightly modified version of Puziacha *et al.'s* [55] unsupervised histogram clustering algorithm is used to group the votes into histogram clusters. The original implementation of Puziacha *et al.'s* clustering algorithm was for image segmentation. The algorithm uses annealed maximum a-posteriori estimation in a Bayesian framework to compute an optimal clustering solution. The authors report that this algorithm performs better and more efficiently than standard K-mean and proximity-based clustering approaches. Slight modification was made as the context of this thesis. As the algorithm is used for angular values, which live on a $(-\pi/2 \cdots \pi/2)$ interval, the

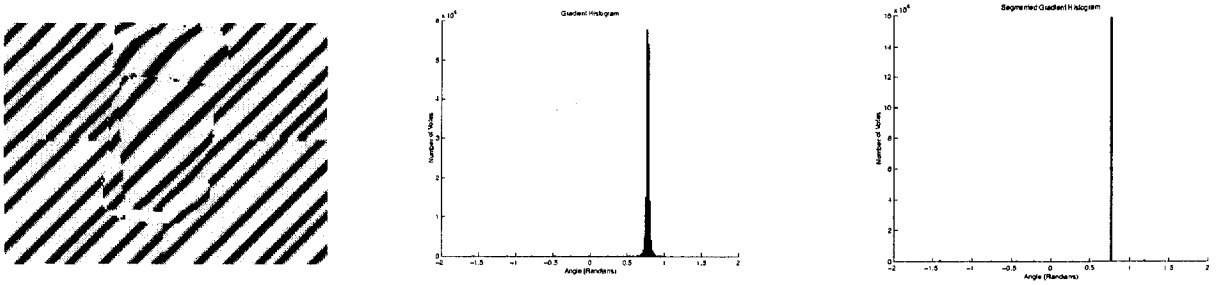
ends of the histogram were joined to provide correct clustering of angles near $-\pi/2$ and $\pi/2$.

The direction of the camera motion, Θ_T, is based on the mean value of the cluster with the most votes. This ensures that a maximum number of optical flow estimates are agreeable to the expected direction of the flow field, and a maximum amount of depth information is thus extracted from the flow-field. The histogram is recomputed after each motion pair, $\mathbf{T}_i$ and $\mathbf{T}_{i+1}$. Thus, an attention-like mechanism is obtained for driving the viewer's motion and closing the *next-step* control loop.

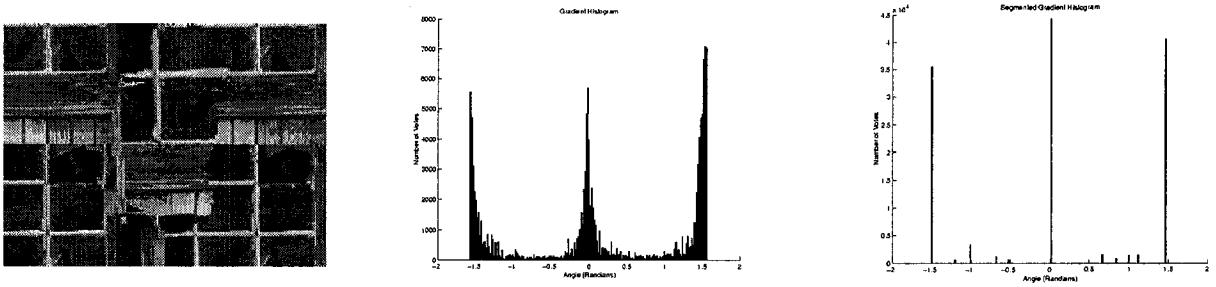
Figure 4.7 shows the histogram distribution and segmentation for several different textures. The first two are synthetic horizontal and diagonal textures. The last three are natural images of a window, a desert and the surface of the planet Mars. Each of these textures is mapped onto the synthetic owl scene. The segmented histograms show that the natural images do indeed contain gradient structure.



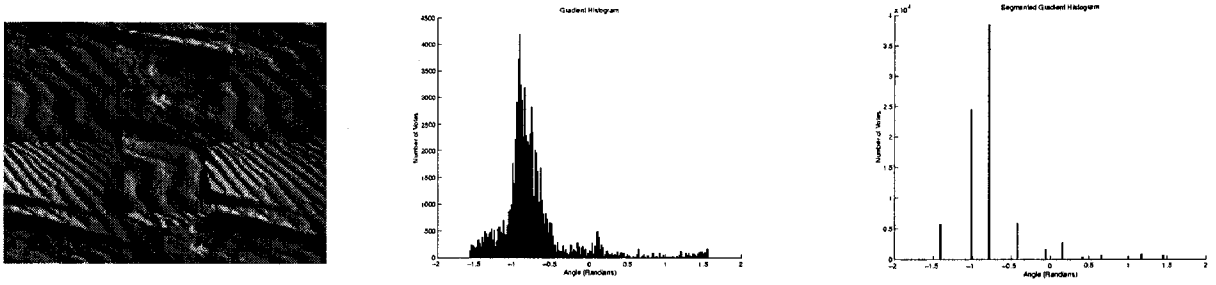
(a) Horizontal texture



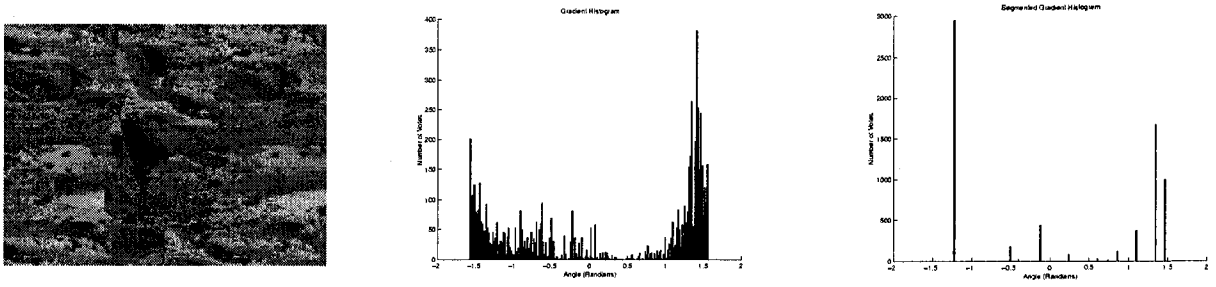
(b) Diagonal texture



(c) Window image



(d) Desert scene



(e) Mars image

Figure 4.7 – Real and synthetic textures with associated gradient and segmented gradient histograms.

The histograms indicate that a dominant gradient direction is present in each texture. The synthetic textures confirm the obvious dominant gradient directions of -90 degrees and 45 degrees. The window image results in a triplet of dominant directions in the horizontal and vertical directions. The desert texture provides a dominant gradient around 42 degrees. Most interesting is the Mars texture, which has no visually dominant gradient direction. The histogram segmentation indicates a dominant gradient direction of –72 degrees. Thus, the observer, in each case, chooses its first motion angle, Θ_{T0}, as: 90 degrees, 45 degrees, 0 degrees, 42 degrees and –72 degrees, respectively for each of the textured scenes in **Figure 4.7**. After each motion, the histogram is updated with the new confidence values and a new motion angle is selected.

5. Generating a Passive Trajectory

To provide a measure of effectiveness of the active algorithm over a passive approach, it is necessary to first define a paradigm for the passive viewer. The passive viewer is approximated as a series of successive random angular motion pairs, for which depth values are accumulated, where no angular motion is repeated,

$$\Theta_{Ti} = Random[0..\pi], \quad \Theta_{Ti} \neq \Theta_{Tj} \ \forall j \in (0 \ldots i-1) \tag{4.23}$$

This approximation to the passive viewer is however somewhat inexact. In general, some form directed bias is observed for the passive viewer. Thus, a random angular motion sequence does not truly represent the passive motion sequence. When considering this, it is important to note that the random angular motion has the advantage of conditioning the noise in the measurement process to the desired zero-mean. Thus, the

random angular motion provides better depth estimation than a true passive observer. Still, it is used in the next section to draw some understanding as to how well the active algorithm works.

6. Results

For each of texture in **Figure 4.7** a series of thirty different passive motion sequences were tested. The mean of the RMS-Error and respective standard deviation are provided for each step of the group of passive sequences. These are compared to the RMS-Error for the active motion sequence. Each sequence was constructed from five successive motion pairs. The results are presented in **Figure 4.8**, **Figure 4.9**, **Figure 4.10**, **Figure 4.11** and **Figure 4.12** below.

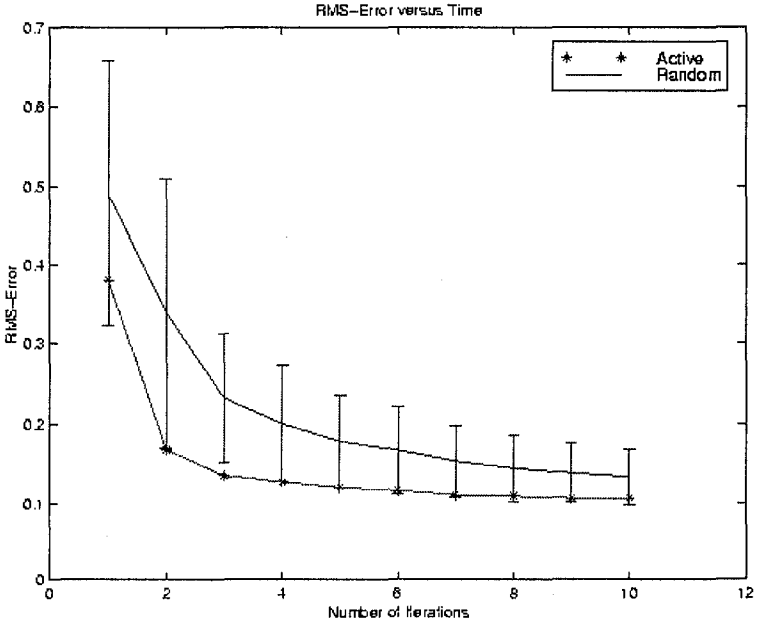


Figure 4.8 – RMS-Error plot for vertical texture.

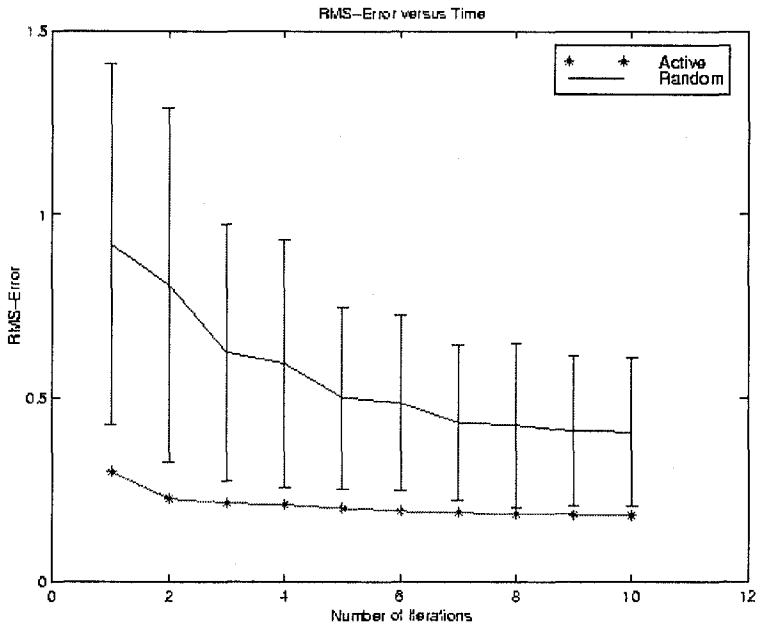


Figure 4.9 – RMS-Error plot for diagonal texture.

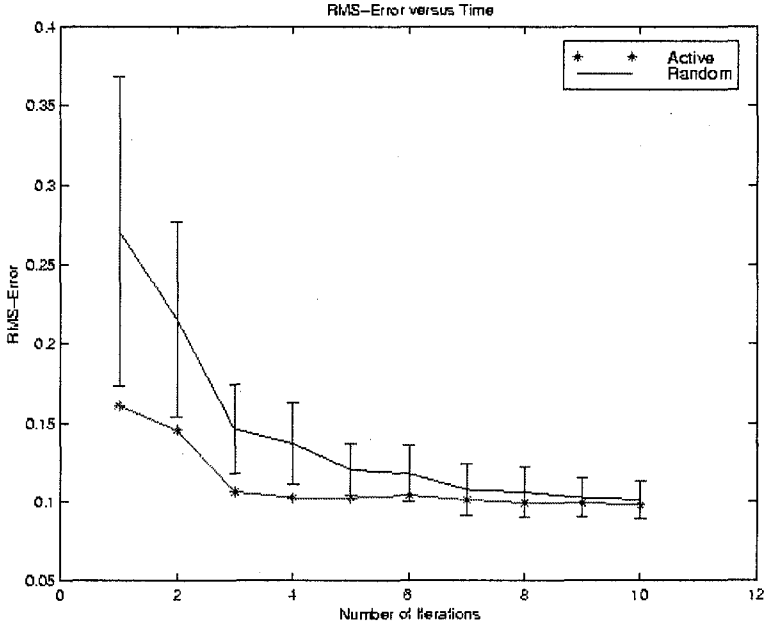


Figure 4.10 – RMS-Error plot for window image.

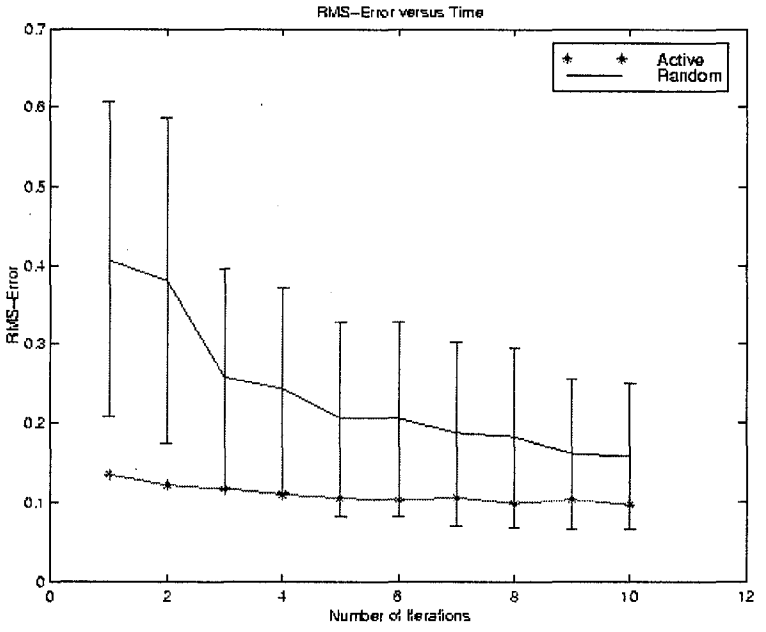


Figure 4.11 – RMS-Error plot for desert image.

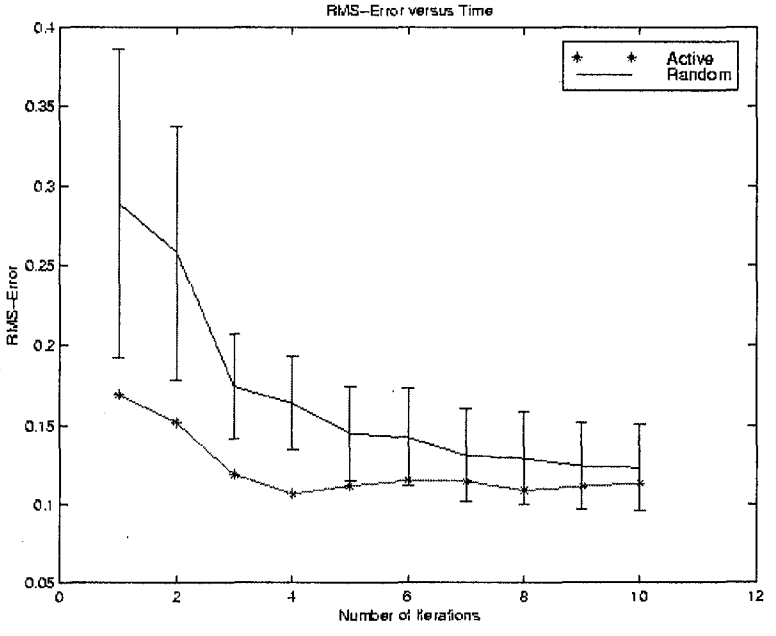
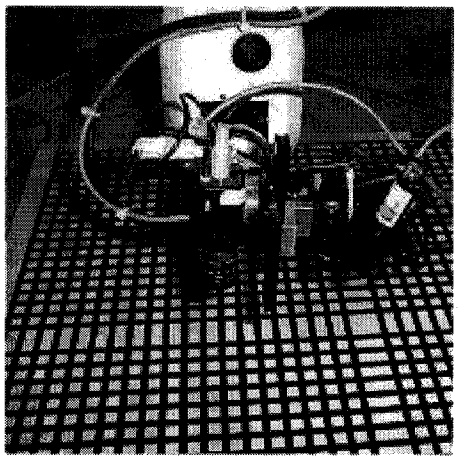


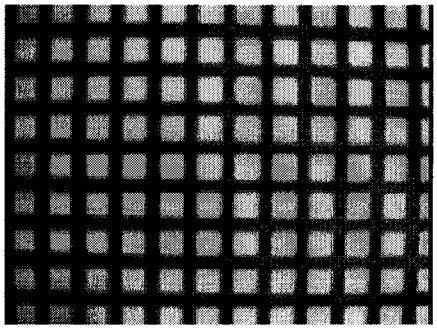
Figure 4.12 – RMS-Error plot for Mars image.

These results show that the active algorithm is indeed capable of taking advantage of texture features of synthetic and natural scenes for improving the convergence rate of the depth estimation process. In all cases, the active method falls below the mean RMS-Error of the passive observer. The active error is generally around a standard deviation better than the passive viewer RMS-Error for the earlier iterations in the sequence. The noise conditioning caused by the random element of the passive viewer is observed in the later iterations, where the active viewer is usually well within the standard deviation range. In general, the active method can be said to converge between 3 and 4 times faster than the passive viewer.

As a last addition to this thesis a pair of real surfaces were scanned using the active reconstruction system. A Panasonic GP-KS152 camera with a focal length of approximately 7mm was mounted on the end effecter of a gantry robot. The robot provided a pose that was repeatable up to 1mm. The camera was place 32cm away from a calibration grid (**Figure 4.13**), and permitted to explore two surfaces: a flat calibration grid and a step edge 10cm tall (**Figure 4.14**). The baseline was fixed at 4mm. Figures show the experimental setup and the camera's view during the experiment. The resulting surface reconstructions after 10 iterations are provided in Figure.

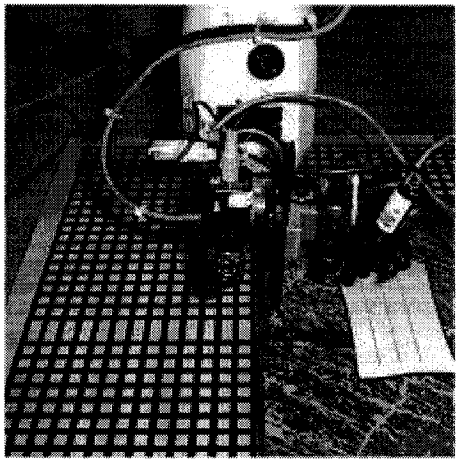


(a) Experimental setup for calibration grid.

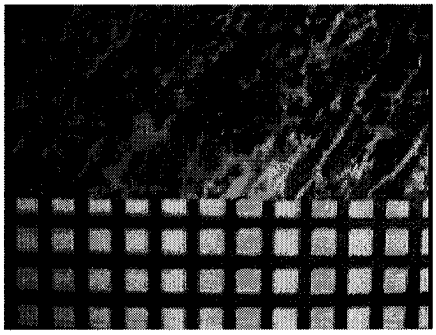


(b) Camera view of calibration grid experiment.

Figure 4.13 – Calibration grid experiment.



(a) Experimental setup for step edge.



(b) Camera view of step edge experiment.

Figure 4.14 – Step edge experiment.

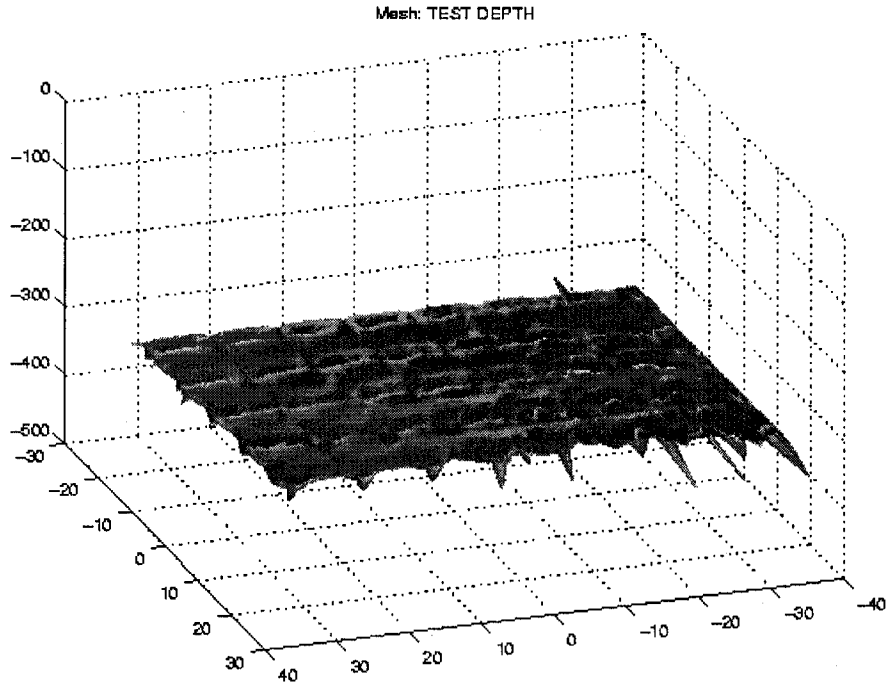


Figure 4.15 – Reconstruction of calibration grid.

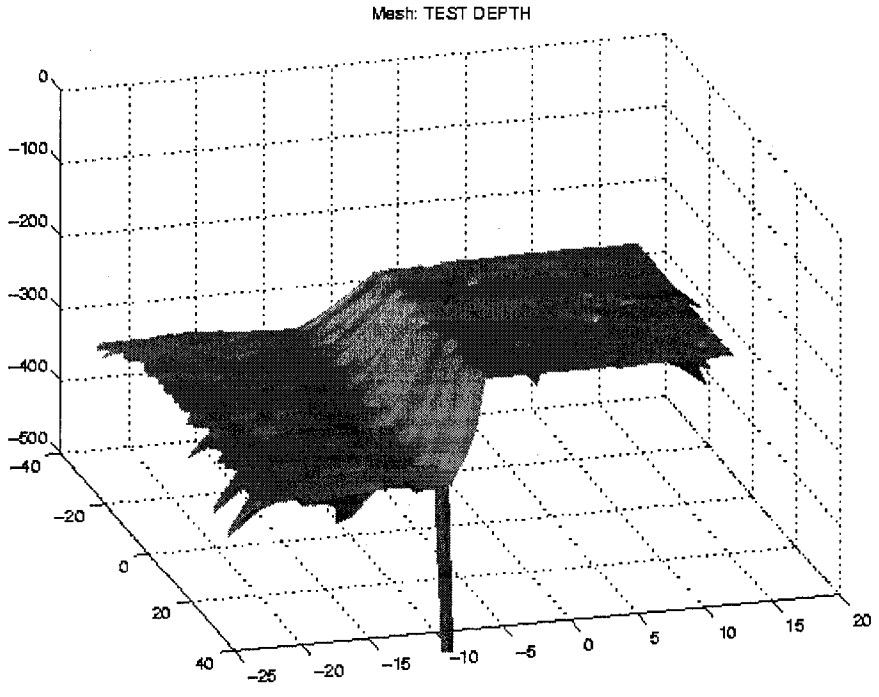


Figure 4.16 – Reconstruction of step edge.

Results from these experiments (**Figure 4.15** and **Figure 4.16**) show that the system is capable of estimating depths with an error of approximately 5mm. The surfaces are surprisingly smooth given that no explicit regularization was performed, and that the scene was only partially textured. The estimated step edge was nearly perpendicular and the height of the edge was successfully recovered.

CHAPTER 5

Conclusion

This thesis has described the design and implementation of an active surface reconstruction algorithm. The system was designed in the context of an autonomous explorer and does not assume continuous image sampling is available. As such, it was constrained to two frames of temporal support and a short baseline.

Surface reconstruction is known to be ill posed for several reasons. Under a small-motion assumption, the reconstruction can be simplified, leaving correspondence as the main source of ambiguity in the system. The problem is thus formulated in a maximal-estimation theory framework. Using this formulation, it is possible to recast previous work that uses a multi-baseline strategy and/or invariant image feature selection. New insight is provided by suggesting that it is not necessarily sufficient to select a wide enough baseline or invariant features. It is shown that, to ensure maximal information is extracted from the image sequence, the epipolar angles of the flow field and the directional predisposition of image features must be considered. This thesis shows that an adaptive active strategy can be used to improve the conditioning of the problem.

The thesis begins by examining the optical flow estimation problem. Different optical flow algorithms are examined at three levels: pixel correspondence, sub-pixel estimation and confidence measures. For the last criterion, a formalized framework for evaluating confidence measures in the anticipated maximal estimation context is

introduced. Results from this section of the thesis suggest that Camus' optical flow algorithm is best suited for two-frame flow estimation due to its important computational advantages and its robustness for short temporal support. Sub-pixel estimation was found to be most effective when combining a bilinear interpolation method with Anandan's sub-pixel method. A generalized confidence measure was suggested and shown to be more consistent for attenuating error for the standard flow estimation test set and an additional pair of real image sequences. This generalized confidence measure was shown to include previously suggested confidence measures by Anandan and Matthies *et al.*

It is expected that the suggested formal approach for determining the effectiveness of confidence measures is an important contribution to the fields of optical flow estimation and dense depth estimation. This methodology certainly offers a clear path for future work in the development of optical flow confidence measures as well as possible improvements to other confidence measures that were not considered in the context of this thesis.

The next important theme discussed in this thesis involved the accumulation of depth information. The Kalman filter was described and various elements of the literature were reviewed. It was shown that the current polygonal mesh models for interpolating the surface measurements are inconsistent with information theory. As such, it was suggested that maximal estimation approach used for temporal accumulation should indeed be extended to the spatial interpolation step of the Kalman Filter. Experimental results were provided to support this. It was shown that when regularization was removed from the current approaches, the algorithms actually regressed their surface estimation, thus creating large gaps in the surface estimates. The

maximal-estimation approach successfully filled the gaps by implicitly interpolating/extrapolating the depth values.

Although the current implementation of this approach works sufficiently well, it is felt that there is still room for improvement. Future work in this area should attempt to develop a relationship between the size of the correlation windows and the current confidence of the surface estimate. Thus more confident depth estimates should reduce image patch sizes to avoid smoothing out edges, while areas of low confidence should increase the patch sizes to include greater spatial support.

The last part of this thesis demonstrated how a generalized statistical model for local image gradient features could be used for improving the estimation process. As such, a statistical histogram-clustering algorithm was modified, and shown to successfully provide correct gaze guidance to the viewer. Several synthetic and real textures were tested experimentally. The active strategy was compared to a pseudo-passive viewer that was composed of thirty random motion sequences. Results show that the active approach was in general a full standard deviation bellow the mean RMS-Error of the passive walks for the first five iterations of the estimation process. More generally, the active strategy improves the convergence rate of the accumulation process by a factor of 3 to 4. This effectively demonstrates that the directional predisposition of the image features does in fact have an important impact on insuring that information in the image is maximally extracted.

It should also be noted that the system should be tested under conditions where segmentation of image gradient results in a more evenly distributed histogram. Under such conditions, the active strategy would probably fail to provide any advantage. It can

even be anticipated that it might be better to use a motion sequence for which the motion angles are evenly distributed, thus conditioning the noise as zero-mean.

Thus, this thesis has provided a consistent approach to demonstrating that statistical grouping of local gradient direction can indeed be used for directing the motion of a viewer. This effectively does significantly improve the depth estimation process. There is still much work to be done at all levels of the system described here, but it is felt that the contents of this thesis provide a clear basis for future work.

REFERENCES

[1] Aguiar, P.M.Q. and Moura J.M.F., *A Fast Algorithm for Rigid Structure from Image Sequences*, International Conference on Image Processing, Vol. 3, pp. 125-129, 1999.

[2] Aloimonos, Y. and Bandyopadyay, A., *Active Vision*, First International Conferenceon Computer Vision, June 1987.

[3] Anandan, P., *A Computational Framework and an Algorithm for Measurement of Visual Motion*, International Journal of Computer Vision, Vol. 2, pp. 283-310, 1989.

[4] Arbel, T. and F.P., Ferrie, *Entropy-based Gaze Planning*, IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, June 1999.

[5] Ayache, N. and Faverjon, B., *Efficient Registration of Stereo Images by Matching Graph Descriptions of Edge Segments*, International Journal of Computer Vision, pp. 107-131, 1987.

[6] Azarbayejani, A., Horowitz, B. and Pentland, A., *Recursive Estimation of Structure and Motion using Relative Orientation Constraints*, IEEE Conference on Computer Vision and Pattern Recognition, 1993.

[7] Bajcsy, R., *Active Perception*, Proceedings of the IEEE, Vol. 76, No 8, August 1988.

[8] Barron, J.L., Fleet, D.J. and Beauchemin, S.S., *Performance of Optical Flow Techniques*, International Journal of Computer Vision, Vol. 12:1, pp. 43-77, 1994.

[9] Beardsley, P.A., Zisserman, A., and Murray, D.W., *Sequential Updating of Projective and Affine Structure from Motion*, International Journal of Computer Vision, vol. 23, no. 3, pages 235-259, 1997.

[10] Bertero, M., Poggio, T.A., and Torre, V., *Ill-Posed Problems in Early Vision*, Proceedings of the IEEE, Vol. 76, No. 8, pp. 869-889, August 1988.

[11] Broida T.J. and Chellappa R., *Estimation of Object Motion Parameters from Noisy Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, pp. 90-99, January 1986.

[12] Burt, P.J., *Fast Filter Transforms for Image Processing*, Computer Graphics and Image Processing, Vol. 16, pp. 20-51, 1981.

[13] Camus, T., *Real-Time Quantized Optical Flow*, IEEE Conference on Computer Architecture for Machine Perception, Como, Italy, pp. 126-131, 1995.

[14] Clark, J.J. and Yuille, A.L., *Shape from Shading via Fusion of Specular and Lambertian Image Components*, Havard Robotics Laboratory, Technical Report no. 89-13.

[15] Cover, T.M. and Thomas, J. A., *Elements of Information Theory*, John-Wiley and Sons, 1991.

[16] Cui, N., Weng, J. and Cohen, P., *Extended Structure and Motion Analysis from Monocular Image Sequences,* International Conference on Computer Vision, 1990.

[17] De Micheli, E., Torre, V. and Uras, S., *The Accuracy of the Computation of Optical Flow and of the Recovery of Motion Parameters*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 5, May 1993.

[18] Dudek, G. and Jenkins, M., *Computational Principles of Mobile Robotics*, Cambridge University Press, Cambridge, 2000.

[19] Dutta, R. and Snyder, M.A., *Robustness of Correspondence-Based Structure from Motion*, IEEE Workshop on Visual Motion, pp. 81 -86, 1991

[20] Fang, J.-Q., and Huang, T.S., *Solving Three Dimensional Small-Rotation Motion Equation*, IEEE Conference on Computer Vision and Pattern Recognition, pp. 253-258, 1983.

[21] Faugeras, O. *Three-Dimensional Computer Vision*, MIT Press, Boston, Massachusetts, 1993.

[22] Fleet, D.J. and Jepson, A.D., *Computation of Component Image Velocity from Local Phase Information*, International Journal of Computer Vision, 5:1, pp. 77-104, 1990.

[23] Fleet, D.J. and Langley, K., *Recursive Filters for Optical Flow*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 1, pp.61-67, Jan 1995.

[24] Foley, J.D., van Damn, A., Feiner, S.K., and Huges, J.F., *Computer Graphics: Practices and Principles 2nd Edition*, Addison Wesley, 1990.

[25] Fradkin, M., Roux M., Maître, H., and Leloğlu, U.M., *Surface Reconstruction from Aerial Images in Dense Urban Areas*, IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 262-267, 1999.

[26] Goldstein, E.B., *Sensation and Perception*, Wadsworth Publishing Company, Belmont, California, 1993.

[27] Grimson, W.E.L., *Computational Experiments with Feature-Based Stereo Algorithms*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 7., no 1. pp. 17-34, Jan 1985.

[28] Haralick, R.M., *Digital Step Edges from Zero Crossing of Second Directional Derivatives*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, No. 1, pp. 58-68, Jan 1984.

[29] Heeger, D.J., *Optical Flow using Spatiotemporal Filters*, International Journal of Computer Vision, Vol. 1, pp. 279-302, 1988.

[30] Heel, J., *Temporal Surface Reconstruction*, IEEE Conference on Computer Vision and Pattern Recognition, pp 607-612, 1991.

[31] Ho, P.-K. and Chung, R., *Stereo-Motion with Stereo and Motion in Complement*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.2, pp. 215-220, Feb 2000.

[32] Horn, B.K.P, *Robot Vision*, The MIT Press, Cambridge, Massachusetts, 1986.

[33] Horn, B.K.P. and Schunk, B.G., *Determining Optical Flow*, Artificial Intelligence, Vol. 17, pp. 185-201, 1981.

[34] Huang, L. and Aloimonos, Y., *Relative Depth Motion using Normal Flow: An Active and Purposive Solution*, Proceedings of the IEEE Workshop on Visual Motion, 1991.

[35] Hung, Y.S. and Ho, H.T., *A Kalman Filter Approach to Direct Depth Estimation Incorporating Surface Structure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No.6, pp. 570-575, June 1999.

[36] Ju, S.X., Black, M.J, and Jepson A.D., *Skin and Bones: Multi-layer, Locally Affine, Optical Flow and Regularization with Transparency*, IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, June, 1996.

[37] Kang, S.B. and Szeliski, R., *3-D Scene Data Recovery using Omnidirectional Multibaseline Stereo*, Cambridge Research Lab, Technical Report, Oct 1995.

[38] Kumar, R., Sawhney, H.S. and Hanson, A.R., 3D *Model Acquisition from Monocular Image Sequences*, IEEE Conference on Computer Vision and Pattern Recognition, pp. 209 –215, 1992.

[39] Liu, H., Hong, T.H., Herman, M., Camus, T. and Chellappa, R., *Accuracy vs Efficiency Trade-offs in Optical Flow Algorithms*, Computer Vision and Image Understanding, Vol. 72, No. 3, pp. 271-286, 1998.

[40] Liu, H., Hong, T.H, Herman, M., and Chellappa, R., *A Generalized Motion Model for Estimating Optical Flow Using 3-D Hermite Polynomials*, Proceedings of the IEEE International Conference on Pattern Recognition, Jerusalem, Israel, pp. 360-366, 1994

[41] Lucas, B. and Kanade, T., *An Iterative Image Registration Technique with Applications in Stereo Vision*, Proceedings of the DARPA Image Understanding Workshop, pp. 121-130, 1981.

[42] Mandelbaum, R., Salgian, G, and Sawhney, *Correlation-Based Estimation of Ego-Motion and Structure from Motion and Stereo*, International Conference on Computer Vision, 1998.

[43] Marr, D. and Poggio, T., *A Computational Theory of Human Stereo Vision*, Proceedings of the Royal Society of London, Vol. B204, pp. 301-328, 1979.

[44] Maru, N., Nishikawa, A., Miyazaki, F. and Arimoto, S., *Active Detection of Binocular Disparities*, IEEE/RSJ International Workshop on Intelligent Robots and Systems, pp. 263 -268 vol.1, 1991.

[45] Masutani, Y., Mikawa, M., Maru, N. and Miyazaki, F., *Visual Servoing for non-Holonomic Mobile Robots*, Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems, Vol. 2, pp. 1133 -1140, 1994.

[46] Matthies, L., Kanade, T. and Szeliski, R., *Kalman Filter-based Algorithms for Estimating Depth from Image Sequences*, International Journal of Computer Vision, 3, 209-236, 1989.

[47] Mathur, S., and Ferrie, F.P., *Edge Localization in Surface Reconstruction Using Optimal Estimation Theory,* Computer Vision and Pattern Recognition, pp. 833-838, 1997

[48] Maybeck, P.S., *Stochastic Models, Estimation, and Control Volume I*, Academic Press, New York, 1979.

[49] Medioni, G. and Nevatia, R., *Segment-Based Stereo Matching*, Computer Vision, Graphics, and Image Processing, Vol. 31, pp. 2-18, 1985.

[50] Nagel, H.H., *Displacement Vectors Derived from Second-Order Intensity Variations in Image Sequences*, Computer Graphics and Image Processing, Vol. 21, pp-85-117, 1983.

[51] Negahdaripour, S., Yu, C.H, and Shokrollahi A.H., *Recovering Shape and Motion From Undersea Images*, IEEE Journal of Oceanic Engineering, Vol. 15, No. 3, pp 189-198, July 1990.

[52] Nesi, P., Del Bimbo, A., and Ben-Tzvi, D., *A Robust Algorithm for Optical Flow Estimation*, Computer Vision and Image Understanding, Vol. 62, No.1, pp 59-68, July 1995.

[53] Okutomi Kanade, *A Multiple-Baseline Stereo*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp. 355-363, April 1993.

[54] Oliensis, J., *A Multi-Frame Structure-from-Motion Algorithm under Perspective Projection*, International Journal of Computer Vision, Vol. 34, pp. 163-192, 1999.

[55] Puzicha, J., Hofman, T., and Buhmann J., *Histogram Clustering for Unsupervised Image Segementation*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 602-608, 1999.

[56] Rosenfeld, A., Hummel, R.A., and Zucker, S.W., *Scene Labeling by Relaxation Operation*, IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-6, pp. 420-423, June 1976.

[57] Sandini, G. and Tistarelli, M., *Active Tracking Strategy for Monocular Depth Inference Over Multiple Frames*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 1, Jan 1990.

[58] Singh, A. and Allen, P., *Image-Flow Computation: An Estimation-Theoretic Framework and a Unified Perspective*, Computer Vision, Graphics and Image Processing, Vol. 56, pp. 152-177, Sept 1992.

[59] Smith, P.N., Sridhar, B. and Hussien, B., *Vision-based Range Estimation using Helicopter Flight Data*, IEEE Conference on Computer Vision and Pattern Recognition, pp. 202-208, 1992.

[60] Szeliski, R., *Bayesian Modeling of Uncertainty in Low-Level Vision*, International Journal of Computer Vision, 5:3, pp. 271-301, 1990.

[61] Szeliski, R. and Kang, B., *Shape Ambiguities in Structure from Motion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5):506-512, May 1997.

[62] Terzopoulos, D., *Regularization of Inverse Visual Problems Involving Discontinuities*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 4, July 1986.

[63] Tomasi, C. and Kanade, T., *Shape and Motion from Image Streams under Orthography: A Factorization Method*, International Journal of Computer Vision, Vol. 9, No. 12, pp. 137-153, Nov 1992.

[64] Trucco E. and Verri A., *Introductory Techniques for 3-D Computer Vision*, Prentice-Hall, New Jersey, 1998.

[65] Uras, S., Girosi, F., Verri, A. and Torre, V., *A Computational Approach to Motion Perception*, Biological Cybernetics, 60: 79-97.

[66] Verri, A. and Poggio, T.A., *Motion Field and Optical Flow: Qualitative Properties*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, No. 5, May. 1989.

[67] Weber, J. and Malik J., *Rigid Body Segmentation and Shape Description from Dense Optical Flow Under Weak Perspective*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, pp. 139-143, February, 1997.

[68] Whaite, P. and Ferrie, F.P., *Autonomous Exploration Driven by Uncertainty*, IEEE Conference on Computer Vision and Pattern Recognition, June 1994.

[69] Williamson, T. and Thorpe, C., *A Specialized Multi-baseline Stereo Technique for Obstacle Detection*, IEEE Conference on Computer Vision and Pattern Recognition, June, 1998.

[70] Xiong, Y. and Shafer, S., *Dense Structure from a Dense Optical Flow Sequence*, Computer Vision and Image Understanding, Vol. 69, No. 2, pp. 222-245, 1998.