Active Surface Reconstruction Using the Gradient Strategy

Marcel Mitran and Frank P. Ferrie

Centre for Intelligent Machines, McGill University,
3480 University Street Room 410, Montreal, Canada, H3A 2A7
{mmitran, ferrie}@cim.mcgill.ca
http://www.cim.mcgill.ca/~apl

Abstract. This paper describes the design and implementation of an active surface reconstruction algorithm for two-frame image sequences using passive imaging. A novel strategy based on the statistical grouping of image gradient features is used. It is shown that the gradient of the intensity in an image can successfully be used to drive the direction of the viewer's motion. As such, an increased efficiency in the accumulation of information is demonstrated through a significant increase in the convergence rate of the depth estimator (3 to 4 times for the presented results) over traditional passive depth-from-motion. The viewer is considered to be restricted to a short baseline. A maximal-estimation framework is adopted to provide a simple approach for propagating information in a bottom-up fashion in the system. A Kalman filtering scheme is used for accumulating information temporally. The paper provides results for real-textured data to support the findings.

Keywords: Image-features, surface geometry, structure-from-motion, active vision and autonomous robot navigation.

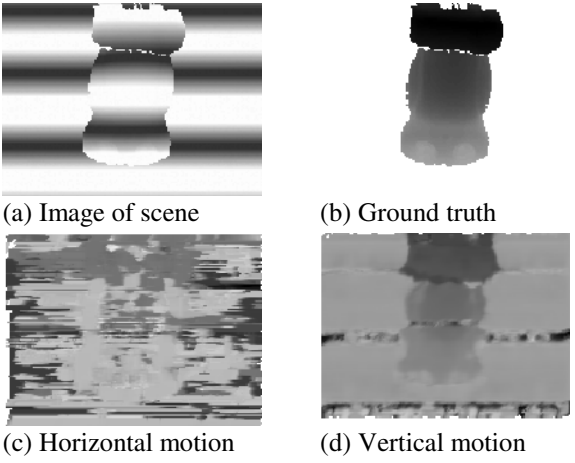
1 Introduction

Structure-from-motion can be defined as the process of inferring a mesh of interconnected points representing a three-dimensional surface from time-varying patterns of image change [3,16,20,25]. This problem is very difficult for several reasons [5,7,14]: the projected image intensity fails to provide an invertible encoding of surface characteristics under most conditions, the correspondence problem becomes increasingly difficult as the displacement between images increases, and the triangulation process becomes increasingly ill-conditioned as the motion between frames becomes small. As such, this problem has received considerable attention in the computer vision literature over the years, most often in the context of image sequences generated by the motion of a passive, mobile observer in a stationary environment [1,6,18]. The contribution of this paper is to examine improvements to structure-from-motion algorithms afforded by an active observer, i.e., one that can alter its trajectory on the basis of visual feedback using an active gaze-planning strategy [1,4,24]. Specifically, it introduces a

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2353, pp. 267–281, 2002.

method for inferring next best view based on an analysis of the gradient structure of the image and for determining the conditions under which it can be applied. This is referred to as the gradient strategy.

The motivation for this paper can best be explained with Fig. 1, which demonstrates the need for active trajectory generation. Fig. 1b shows a depth map of an owl obtained with a laser rangefinder. The depth map is painted with an artificial horizontal texture and rendered as the image shown in Fig. 1a. Using a conventional structure-from-motion algorithm [18], the depth maps shown in Fig. 1c and Fig. 1d are obtained for 10 horizontal and vertical displacements respectively. These resulting depth maps suggest that a passive viewer moving horizontally will fail to recover depth for this scene.



(a) Image of scene (b) Ground truth

(c) Horizontal motion (d) Vertical motion

Fig. 1. Horizontally textured scene.

Thus, an active vision system is introduced. The novelty of the gradient strategy, when compared to previous active systems, lies in suggesting that controlling the angle of the motion of the viewer is as important as controlling the magnitude alone. As such, the gradient strategy is based on the angular predisposition of scene features and is adopted to extract maximal information from the scene. Maximal information is represented by an increased convergence rate for the estimation process.

An additional constraint on the system is the restriction on the configuration space inherent to an autonomous navigator [9,13,19,25]. Most active vision algorithms assume full motion control is available to a viewer (e.g. wide baseline). This is not often the case for a holonomically 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.

The rest of this paper develops a system with these key points in mind. Section 2 reviews key elements in the literature related to monocular stereo and active surface reconstruction, Section 3 discusses different elements of the computer vision theory

involved in developing the active system, Section 4 provides experimental results, and Section 5 concludes.

2 Previous Work

Two general forms of passive surface reconstruction have emerged. The first is feature based surface reconstruction [3,6]. This approach fits 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 over large baselines. Difficulties arise in the fitting process, which can be time consuming and ill-conditioned. Additional complexity arises in matching these high-level features which may be numerous, small and difficult to detect.

The second approach is the iconic depth estimator [18,25] 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. Generally, due to the smaller baseline, depth measurements are noisy. However as depicted in Fig. 1, even when using a maximal estimation framework (such as the Kalman filter) in a textured environment for accumulating information, this process is not guaranteed to converge.

Multi-baseline stereo [1,16,20] has emerged as a solution to the well-known dichotomy in baseline stereo: smaller motions aid with the correspondence, while greater baseline motions provide more stable depth estimates but sparser data. Notable in the multi-baseline category, is Aloimonos and Bandyopadhyan's [1] work on the active vision paradigm. In particular, they discuss an adaptive image coordinate system, based on epipolar geometry and isophotes. They demonstrate that a controlled baseline magnitude can reduce ambiguity in the correspondence process. However, they offer no solution for textured environments and assume an unrestricted motion space.

Whaite and Ferrie [24] suggest a formal mathematical framework for active vision in which ambiguity is equated to uncertainty. The measurement model is defined as a general linear system,

$$d = G(x)\, m\,, \tag{1}$$

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 (1).

The associated uncertainty to this solution is,

$$C_m = C_w \left(G(x)^T G(x)\right)^{-1}\,, \tag{2}$$

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

$$H = G(x)^T G(x)\,, \tag{3}$$

maximally reduces the uncertainty, C_m. The authors apply their theory in the context of autonomous model fitting applications. The measurement consists of laser-range data, and the model space is defined as the set of super-ellipsoids. This methodology can, however, be generalized to the structure-from-motion paradigm.

Huang and Aloimonos [15] 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 component of the optical-flow vector parallel to image gradient best approximates the component of the projected motion field in the same direction. This agrees with the analysis of Verri and Poggio [23]. 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 [22] 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.

3 Theory

The gradient strategy attempts to improve the convergence rate for depth estimation by using the image gradient to predict the most informative camera baseline angle, Θ_T. It is assumed that the viewer's motion is controlled up to some given certainty, C_T and $C_{\cdot}$, for the translation and rotation parameters respectively (Fig. 2).

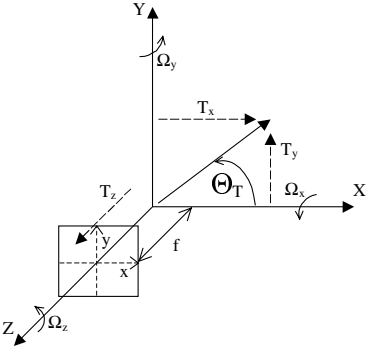


Fig. 2. Pinhole camera of focal length f with a viewer-based coordinate system where, $\mathbf{\Omega_i} = (\Omega_x\ \Omega_y\ \Omega_z)$, about an axis passing through the origin, and a translation, $\mathbf{T_i} = (T_x\ T_y\ T_z)$.

This section begins by describing a novel method for embedding the epipolar constraint in the correspondence process. An improvement to Matthies *et al.*'s [18] iconic

depth estimator is suggested. And finally, the anticipated gradient strategy for improving the depth accumulation process is discussed.

3.1 Epipolar Constraint

A novel method for embedding the epipolar constraint into the correspondence process is suggested here for the case where the motion parameters are only known up to a given certainty, C_T and C_Ω. The epipolar constraint is an important element in the geometry of stereovision. It is generally used to constrain the angular components of an optical flow field to a line when the motion parameters between two views are fully known.

Using a standard perspective projection and camera-centered frame (Fig. 2), the essential matrix, can be represented mathematically as [10],

$$E = \begin{bmatrix} 1 & -\Omega_z & \Omega_y \\ \Omega_z & 1 & -\Omega_x \\ -\Omega_y & \Omega_x & 1 \end{bmatrix} \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}. \tag{4}$$

This can be interpreted as an angular constraint on the flow measurement. The epipolar line in the new frame, given that the image point in the prior frame is, (x_p, y_p, f), is computed as follows,

$$\begin{bmatrix} m_x \\ m_y \\ b \end{bmatrix} = E \begin{bmatrix} x_p \\ y_p \\ f \end{bmatrix}, \tag{5}$$

given the following form for the equation of a line,

$$y = \frac{m_y}{m_x} x + b\,. \tag{6}$$

When expanding (5) and (6), the slope terms m_x and m_y are described as

$$\begin{aligned} m_x &= -\left(\Omega_z T_z + \Omega_y T_y\right)x_p + \left(\Omega_y T_x - T_z\right)y_p + \left(T_y + \Omega_z T_x\right)f\,, \\ m_y &= \left(T_z + \Omega_x T_y\right)x_p - \left(\Omega_z T_z + \Omega_x T_x\right)y_p + \left(T_y \Omega_z - T_x\right)f\,. \end{aligned} \tag{7}$$

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 embedded into the correspondence process. In this case, correspondence is performed using a region-matching algorithm similar to that of Anandan [2]. By adding a bias to each cell of the optical-flow's sum-of-square-differences,

$$SSD_{x,y}(dx,dy) = \sum_{j=-n}^{n} \sum_{i=-n}^{n} W(i,j)\left[I_1(x+i, y+j) - I_2(x+dx+i, y+dy+j)\right]^2, \tag{8}$$

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, the optical-flow estimate can be constrained. The bias is derived from the square of the Mahanalobis distance, $M_{ep}(dx,dy)^2$, of the (dx,dy)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

$$SSD_{\min} = \min\left(SSD(dx,dy) + e^{M_{ep}(dx,dy)^2}\right). \tag{9}$$

Uncertainty in the motion parameters, C_T and C_Ω, and image point parameters, σ_{xp}^2 and σ_{yp}^2, can be projected into the epipolar parameter space, to obtain,

$$\sigma_{mx}^2 = \left[\frac{\partial(m_x)}{\partial(\mathbf{T},\boldsymbol{\Omega},x_p,y_p)}\right]^T \begin{bmatrix} C_{\mathbf{T}} & 0 & 0 & 0 \\ 0 & C_{\boldsymbol{\Omega}} & 0 & 0 \\ 0 & 0 & \sigma_{xp}^2 & 0 \\ 0 & 0 & 0 & \sigma_{yp}^2 \end{bmatrix} \left[\frac{\partial(m_x)}{\partial(\mathbf{T},\boldsymbol{\Omega},x_p,y_p)}\right], \tag{10}$$

and similarly for σ_{my}^2.

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

$$\hat{f}(dx,dy) = \frac{(dx,dy)}{\sqrt{dx^2+dy^2}}, \tag{11}$$

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

$$M_{ep}(dx,dy)^2 = \left(\hat{f}(dx,dy) - \hat{m}\right)^T \begin{bmatrix} \frac{1}{\sigma_{mx}^2} & \\ & \frac{1}{\sigma_{my}^2} \end{bmatrix} \left(\hat{f}(dx,dy) - \hat{m}\right), \tag{12}$$

which, after some manipulation, becomes

$$M_{ep}(dx,dy)^2 = \frac{1}{\sigma_m^2}\left(\left(\hat{f}_x(dx,dy) - \hat{m}_x\right)^2 + \left(\hat{f}_y(dx,dy) - \hat{m}_y\right)^2\right), \tag{13}$$

where from the symmetry of expressions (7),

$$\sigma_m^2 = \sigma_{mx}^2 = \sigma_{my}^2. \tag{14}$$

The bias is used when searching the SSD for its minimum.

3.2 Kalman Framework for Iconic Depth Estimation

The depth-accumulation framework used for the system presented here is very similar to that of Matthies *et al.* [18]. The spatial interpolation method is modified to provide conformance with the rules of information theory. Thus, the maximal estimation framework is extended to the interpolation phase of the Kalman framework.

For baseline motion, Matthies *et al.*'s measurement model is described as follows:

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}=\frac{1}{Z}\begin{bmatrix}-f & \\ & -f\end{bmatrix}\begin{bmatrix}T_x\\ T_y\end{bmatrix}. \tag{15}$$

where the state vector is represented by inverse depth (or disparity).

The system uses a least-squares solution to the measurement problem. The least-squares disparity estimate is obtained from the optical flow as

$$\hat{d}_m=\left(H^T C_f^{-1} H\right)^{-1} H^T C_f^{-1}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}, \tag{16}$$

where

$$H=\begin{bmatrix}-f & \\ & -f\end{bmatrix}\begin{bmatrix}T_x\\ T_y\end{bmatrix}, \tag{17}$$

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

$$C_f=\begin{bmatrix}\sigma_{\Delta x}^2 & 0\\ 0 & \sigma_{\Delta y}^2\end{bmatrix}. \tag{18}$$

For baseline motions, the corresponding variance, σ_d^2, of the disparity estimate is obtained from

$$\sigma_d^2=\left(H^T C_f^{-1} H\right)^{-1}=\frac{\sigma_{\Delta x}^2\sigma_{\Delta y}^2}{\sigma_{\Delta y}^2\left(fT_x\right)^2+\sigma_{\Delta x}^2\left(fT_y\right)^2}. \tag{19}$$

Matthies *et al.* introduce spatial support into their system using interpolation and regularization stages. 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. 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

$$d_i=w_{0i}d_0+w_{1i}d_1+w_{2i}d_2, \tag{20}$$

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 (20) onto the covariance matrix constructed from σ_{d0}^2, σ_{d1}^2 and σ_{d2}^2. This effectively results in

$$\sigma_{di}^2=w_0^2\sigma_{d0}^2+w_1^2\sigma_{d1}^2+w_2^2\sigma_{d2}^2. \tag{21}$$

It is then suggested by Matthies *et al.* that a pure blend may be used to interpolate the new confidence values,

$$\sigma_{di}^2=w_0\sigma_{d0}^2+w_1\sigma_{d1}^2+w_2\sigma_{d2}^2. \tag{22}$$

The interpolation method of Matthies *et al.* leads to an increase in uncertainty when interpolating. Information theory suggests the opposite; on average, conditional en-

tropy of a random variable (which is a measure of its uncertainty) should not increase as more measurements are combined into an estimate [8]. Thus,

$$\sigma_{di}^2 \leq \mathbf{min}\left(\sigma_{d0}^2, \sigma_{d1}^2, \sigma_{d2}^2\right). \tag{23}$$

The upper bound for expressions (21) and (22) is

$$\sigma_{di}^2 \leq \mathbf{max}\left(\sigma_{d0}^2, \sigma_{d1}^2, \sigma_{d2}^2\right), \tag{24}$$

which implies that the approach used by Matthies *et al.* 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 [17] describes how to do this for local curvature models such as Darboux frames. The approach taken here will involve a simpler local surface model – the triangle. As such, the depth interpolator is described as

$$d_i = \left(\frac{w_{0i}}{\sigma_{d0}} d_0 + \frac{w_{1i}}{\sigma_{d1}} d_1 + \frac{w_{2i}}{\sigma_{d2}} d_2\right) \Bigg/ \left(\frac{w_{0i}}{\sigma_{d0}} + \frac{w_{1i}}{\sigma_{d1}} + \frac{w_{2i}}{\sigma_{d2}}\right). \tag{25}$$

The variance associated to the new disparity value is

$$\sigma_{di}^2 = \sigma_{d0}^2 \sigma_{d1}^2 \sigma_{d2}^2 \Big/ \left(w_0^2 \sigma_{d1}^2 \sigma_{d2}^2 + w_1^2 \sigma_{d0}^2 \sigma_{d2}^2 + w_2^2 \sigma_{d0}^2 \sigma_{d1}^2\right). \tag{26}$$

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 [12].

The last step in the Kalman framework is the update phase. The Kalman, K_{i+1}, gain is computed as

$$K_{i+1} = \frac{P_{k+1|k}}{P_{k+1|k} + \sigma_d^2}, \tag{27}$$

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

$$\hat{d}_{i+1} = d_{i+1|i} + K_{i+1}\left(\hat{d}_m - d_{i+1|i}\right) \tag{28}$$

and the updated confidence is obtained as

$$P_{k+1} = \frac{P_{k+1|k}}{P_{k+1|k} + \sigma_d^2}. \tag{29}$$

3.3 The Gradient Strategy

The gradient strategy provides a novel method for selecting the next best motion. It is strongly influenced by Whaite and Ferrie's work as it selects the most informative motion for which the Kalman state uncertainty, $P_k(i,j)$ is maximally reduced over the *MxN* image. It is noted that wide-baseline stereo achieves this by increasing the size of *Tx* and *Ty* in (19). The gradient strategy increases confidence in the estimate by minimizing $\sigma_{\Delta x}^2$ and $\sigma_{\Delta y}^2$ instead. This is achieved by selecting an appropriate baseline angle, Θ_T. The strategy is summarized 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 image matrix

$$\Pi(i,j) = \begin{bmatrix} \sum WI_x^2 & \sum WI_x I_y \\ \sum WI_x I_y & \sum WI_y^2 \end{bmatrix}, \tag{30}$$

where W is a Gaussian operator, I_x and I_y are partial derivative of the image in the x and y directions, and $\lambda_1 > \lambda_2$. Table 1 provides an intuitive association of eigenvalues to the three conditions mentioned above.

Table 1. Interpretation of Normal matrix eigenvalues.

λ_1	λ_2	Condition Num.
LARGE	LARGE	(ii)
LARGE	SMALL	(i)
SMALL	SMALL	(iii)

Extending this idea to a full *MxN* 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 $\Pi(i,j)$. 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

$$w(i,j) = P_k(i,j)\frac{\lambda_1(i,j)}{\left(1+\lambda_1(i,j)\right)\lambda_2(i,j)} \quad . \tag{31}$$

The choice of the inverted λ_2 term in (31) is based on the observation made by Barron *et al.* [5] and, Fleet and Langley [11] 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.

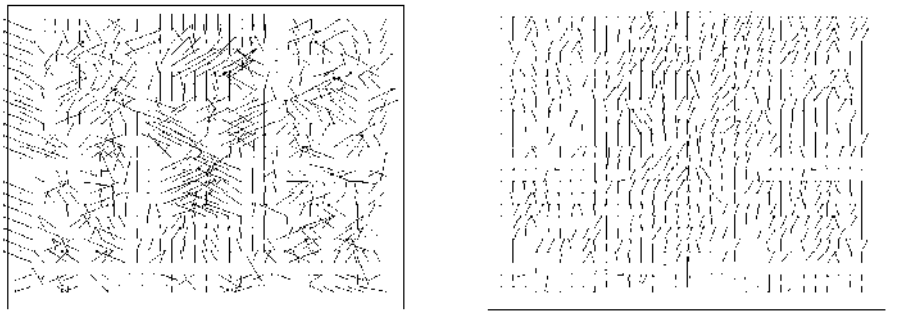
In the context of the gradient-based weighted histogram, the spatial structure of the features is ignored. However, there still remains a strong relationship between the gradient-structures in the image and the histogram's distribution. Generally, different features with common intensity orientations 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. Puzicha *et al.'s* [21] unsupervised histogram clustering algorithm is used to group the votes. The original implementation of their clustering algorithm was for image segmentation. The algorithm uses annealed maximum a-posteriori estimation in a Bayesian framework to compute an optimal clustering solution. Puzicha *et al.* report that this algorithm performs better and more efficiently than standard K-means and proximity-based clustering approaches.

The direction of the camera baseline, 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 with higher state uncertainty are agreeable to the expected direction of the flow field, and a maximum amount of depth information can thus be extracted. The histogram is recomputed after each motion pair, Θ_{Ti} and $\Theta_{T\,i+1}$, where $\Theta_{T\,i} = \Theta_{T\,i+1} + \pi$ (back and forth). As such, an attention-like mechanism is obtained for driving the viewer's motion and closing the *next-step* control loop.

4 Experimental Results

The active surface reconstruction system was tested for several real textured scenes. Results for different components of the system are provided below.

Fig. 3a shows a sample flow field for the vertical motion described in Fig. 1 when epipolar geometry is ignored. The y- components of the flow-field are constrained by the image features, while the x- components are unstable. As suggested earlier, epipolar geometry offers a solution for removing ambiguity in the flow angle. Fig. 3b 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.



(a) Without epipolar constraint (b) With epipolar constraint

Fig. 3. Flow fields for vertical and horizontal motions in horizontally textured environment.

Fig. 4 shows a depth map obtained for Matthies *et al*'s interpolation method, as well as, the maximal estimation method. When the regularization process is removed, less confident estimates have the upper hand and propagate. This results in large instabilities in the depth map (black holes). Even if the regularization process were included, no framework for generating new confidence values is provided for the interpolated depth 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 interpolating and propagating information to areas of low confidence. The resulting depth map in Fig. 4 is smooth and consistent with the ground truth (Fig. 1b).

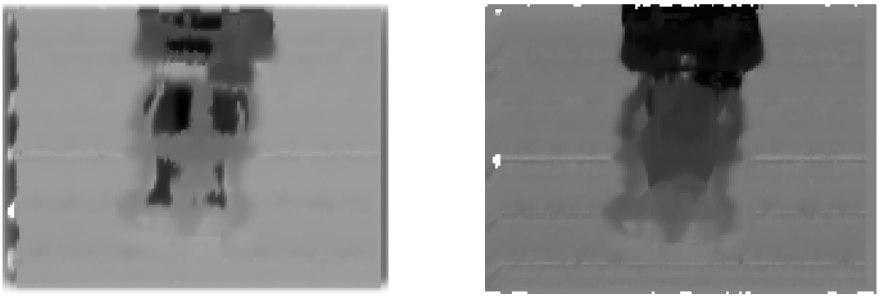


Fig. 4. Depth maps from interpolation methods. Left: Matthies *et al*. method when regularization is removed. Right: Maximal estimation method.

Fig. 5 shows the histogram segmentation for several different textures. The first is the synthetic horizontal texture. The other three are natural textures of a window, a desert and the surface of the planet Mars. Each of these textures is mapped onto the range-image of the owl. The segmented histograms show that the natural images do indeed contain gradient structure.

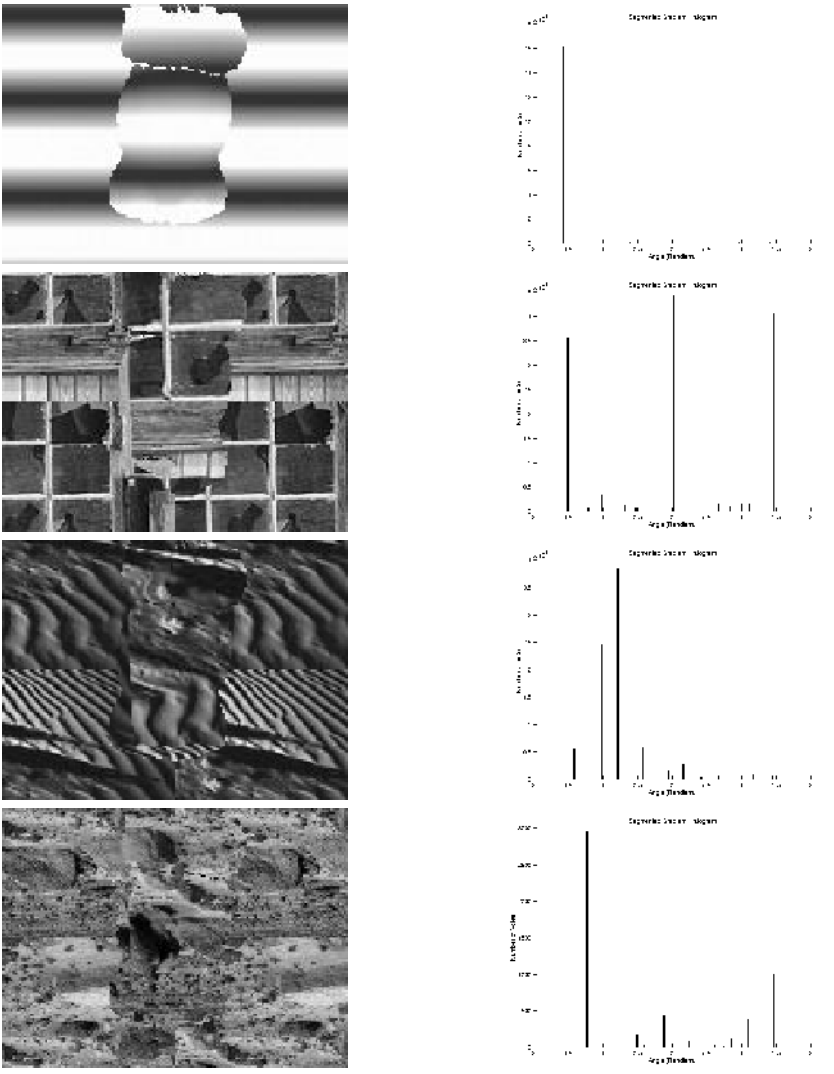


Fig. 5. Real and synthetic textures with associated segmented gradient histograms. Histograms indicate number of votes versus angle ($-\pi/2 \ldots \pi/2$). From top to bottom: Synthetic horizontal, window texture, desert texture, and Mars picture texture

The last element described in the results section considers the effectiveness of the active control strategy. To do this, 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,

$$\Theta_{\mathbf{T}i} = Random[0..\pi], \quad \Theta_{\mathbf{T}i} \neq \Theta_{\mathbf{T}j} \ \forall j \in (0 \ldots i-1). \tag{32}$$

This approximation to the passive viewer is however somewhat inexact. In general, some form of directed bias is observed for a true 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.

For each of the textures in Fig. 5 a series of thirty different passive motion sequences were tested. The mean of the RMS-Errors, which was computed as

$$Err = \sqrt{\frac{1}{MxN}\sum\left(d_{Ground\_Truth}(i,j) - d_{Estimate}(i,j)\right)^2} \tag{33}$$

over the *MxN* depth image, 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 Fig. 6 below.

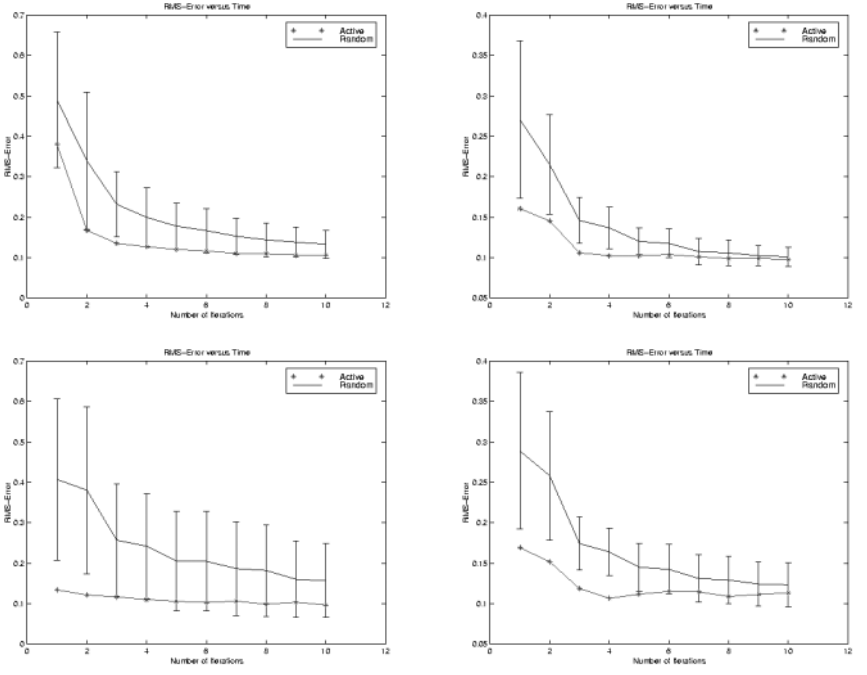


Fig. 6. RMS-Error versus iterations. Top-left: Horizontal texture. Top-right: Diagonal texture. Middle-left: Window texture. Middle-right: Desert texture. Bottom-left: Mars texture.

Results show that the active algorithm is indeed capable of taking advantage of texture features of synthetic and natural scenes to improve 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. Generally, the passive experiments converge on the last two iterations. The active methods converge between the first and fourth iteration. As such, the active method can be said to converge between 3 to 4 times faster than the randomly sequenced viewer.

5 Conclusion

This paper has described the design and implementation of an active surface reconstruction algorithm. The problem is presented in a maximal-estimation 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. It is shown that, to ensure that 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. Thus, the active gradient strategy is suggested.

Implementation issues relating to epipolar geometry and iconic depth accumulation were discussed. A novel method for embedding the epipolar constraint in the correspondence process was introduced, and an improvement to previously suggested depth interpolation was suggested. These components were successfully implemented as part of the active system.

Finally, it was demonstrated that a generalized statistical model for local image gradient features could be used to drive the angle of the camera motion, thus improving the estimation process. As such, a statistical histogram-clustering algorithm was shown to successfully provide correct gaze guidance to the viewer. Several synthetic and real textures were tested experimentally. The effectiveness of the gradient strategy, with respect to the pseudo-passive viewer, was gauged by comparing convergence plots for the two methods. Results show the active strategy improves the convergence rate of the accumulation process by a factor of 3 to 4 for the given test set.

References

1. Aloimonos, Y. and Bandyopadhyan, A.: Active Vision. First International Conference on Computer Vision, June 1987.
2. Anandan, P.: A Computational Framework and an Algorithm for Measurement of Visual Motion, International Journal of Computer Vision, Vol. 2, pp. 283-310, 1989.
3. 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.
4. Bajcsy, R.: Active Perception. Proceedings of the IEEE, Vol. 76, No 8, August 1988.

5. 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.
6. 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.
7. 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.
8. Cover, T.M. and Thomas, J. A.: Elements of Information Theory. John-Wiley and Sons, 1991.
9. Dudek, G. and Jenkins, M.: Computational Principles of Mobile Robotics. Cambridge University Press, Cambridge, 2000.
10. Faugeras, O.: Three-Dimensional Computer Vision, MIT Press, Boston, Mass. , 1993.
11. 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.
12. Foley, J.D., van Damn, A., Feiner, S.K., and Huges, J.F.: Computer Graphics: Practices and Principles 2nd Edition. Addison Wesley, 1990.
13. 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.
14. Horn, B.K.P: Robot Vision. The MIT Press, Cambridge, Massachusetts, 1986.
15. 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.
16. Kang, S.B. and Szeliski, R.: 3-D Scene Data Recovery using Omnidirectional Multibaseline Stereo. Cambridge Research Lab, Technical Report, Oct 1995.
17. Mathur, S., and Ferrie, F.P.: Edge Localization in Surface Reconstruction Using Optimal Estimation Theory. Computer Vision and Pattern Recognition, pp. 833-838, 1997
18. 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.
19. 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, 1990.
20. Okutomi M. and Kanade, T.: A Multiple-Baseline Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp. 355-363, April 1993.
21. 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.
22. 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, January 1990.
23. 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.
24. Whaite, P. and Ferrie, F.P.: Autonomous Exploration Driven by Uncertainty. IEEE Conference on Computer Vision and Pattern Recognition, June 1994.
25. 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.