Detecting Floor Anomalies

Michael R. M. Jenkin[1] and Allan Jepson[2,‡]

[1]Department of Computer Science, York University,
[2]Canadian Institute for Advanced Research, and
Department of Computer Science, University of Toronto

March 14, 1995

Abstract

When a robot moves about a 2D world such as a planar surface, it is important that obstacles to the robot's motions be detected. This classical problem of "obstacle detection" has proven to be difficult. Many researchers have formulated this problem as being the process of determining where a robot *cannot* move due to the presence of obstacles. An alternative approach presented here is to determine where an robot *can* go by identifying floor regions for which the planar floor assumption can be verified. A stereo vision system is developed for Floor Anomaly Detection (FAD), and its relationship to existing stereo obstacle detection algorithms is described.

1 Introduction

When an agent moves about its environment it is important that the agent is sure that the surface over which it is to move is safe. In robotic applications this safety is usually expressed in terms of the robot being able to detect obstacles to its motion. Here we are concerned with the development of a stereo vision system capable of identifying safe places to move on a floor plane. Previous work [3, 1, 12] on this problem has taken the goal to be to identify obstacles on the 2D plane over which the robot is to move. What is tacitly assumed in the literature is that the floor plane itself is safe and can be traversed (for an exception, see [2]). This is an important issue in many robotics environments as the floor may not be particularly safe for the robot to navigate.

As a concrete example, the ARK (Autonomous Robot for a Known Environment) Project involves the development of a sensor-based mobile robot that can autonomously navigate in a known, previously-mapped industrial environment. The environment presents many difficulties for safe navigation, including people and forklifts moving about; oil and water spills on the floor; floor drains (which can be uncovered); hoses, tools, and piping on the

‡Authors contributed equally and are listed in alphabetical order. This paper appeared in the Proceedings of the British Machine Vision Conference 1994, Univ. of York, Ed. E. Hancock, BMVA Press, pp. 731-740.

floor [7]. Thus although the floor of the ARK environment can be expected to be planar, local regions of the floor can be expected to contain structure which violates the planarity assumption. In addition, regions that violate the planarity assumption are not easily mapped and they cannot be completely avoided.

An ideal stereo vision approach would be to use the stereo information to accurately locate both the floor and any anomalies in 3D. This is still a formidable task and is beyond the reach of moderate cost image processing hardware for real-time applications. Instead, we consider locating regions in the image which provide evidence for being on the floor plane. To do this we use a novel method for fitting an exact model of stereo disparities arising from a 3D plane to the provided image data. This fitting method allows us to deal with deviations from a particular floor model, which often arise due to the robot tilting, but can also arise from structure in the floor, such as ramps, or from changes in the stereo system's configuration over time. Once the image correspondences for the floor have been determined, and the anomalies marked, we can then map the extracted information onto a robot-centered representation so that the robot can plan a safe motion.

2 Calibration of a Fixed Stereo Head

In order to map results from image coordinates to robot centered world coordinates, we first need to calibrate the stereo system. In this paper we consider the camera model proposed by Horn [6]. This model includes a general 3D camera position and focal length, along with several important perturbations from an ideal pinhole camera. In particular, Horn's model can represent defects due to a general 3D-misalignment of the sensor array with the optical axis. The most common remaining distortions are nonlinear radial distortions of the image, which can be significant for some lenses. Such radial distortions are not modeled here and we assume they are negligible.

In order to specify Horn's model, we define $\vec{X} = (X_1,\ X_2,\ X_3)^T$ to be a point in a 3-dimensional global coordinate system, and let $\vec{x} = (x_1, x_2)^T$ be its image. It is most convenient to write both $\vec{X}$ and $\vec{x}$ in terms of homogeneous coordinates. That is, a given four-vector $\vec{W}$ represents $\vec{X}$ if $X_i = W_i/W_4$ for $i = 1, 2, 3$. Similarly, a three vector $\vec{w}$ represents the image point $\vec{x}$ whenever $x_i = w_i/w_3$ for $i = 1, 2$. Now, following [6], we define the transformation from an arbitrary point $\vec{W}$ in a global coordinate system to an image point $\vec{w}$ as

$$\vec{w} = T\vec{W}, \tag{1}$$

where T is a 3x4 matrix of coefficients which specify the transformation.

It is straightforward to calibrate a camera using this model along with a known 3D calibration object. In particular, points on the calibration object are identified by hand in each image. Linear least squares is then used to compute the transformation matrix, T, in equation (1), separately for each camera. We denote the resulting matrices for the left and right cameras simply as L and R. These matrices cleanly capture all the required information about the fixed stereo system.

3 Image Warping for a Plane

We are interested in the exact form of image warp which can be used to map the images of points on the ground plane in one stereo image to the corresponding image points in the second view. Suppose the ground plane (or an arbitrary plane, for that matter) is described in normal-distance form as $\vec{n} \cdot \vec{X} = d$. Here $\vec{n} = (n_1, n_2, n_3)^T$ is a unit normal to the plane, and d measures the perpendicular distance of the plane from the origin of the 3D coordinate frame. Then, for pinhole camera models, Faugeras [3] has shown that the mapping from a point, $\vec{w}^{\,r}$, in the right image to the corresponding point, $\vec{w}^{\,l}$, in the left image is given by

$$\vec{w}^{\,l} = K(\vec{n}, d)\vec{w}^{\,r}, \tag{2}$$

where $K(\vec{n}, d)$ is a 3×3 matrix. We have shown [9] that the same form of equation suffices for the more general camera models described in the previous section, and we have derived a closed-form expression for K in terms of L, R, $\vec{n}$ and d.

It turns out that $K(\vec{n}, d)$ is a linear function of $(\vec{n}, d)$. That is, using the slightly unusual notation

$$K(\vec{n}, d) = \begin{pmatrix} k_1 & k_2 & k_3 \\ k_4 & k_5 & k_6 \\ k_7 & k_8 & k_9 \end{pmatrix}, \tag{3}$$

we find the vector $\vec{k} = (k_1, \ldots, k_9)^T$ is given by

$$\vec{k}(\vec{n}, d) = M(L, R) \begin{pmatrix} \vec{n} \\ -d \end{pmatrix}. \tag{4}$$

Here $M(L, R)$ is a 9×4 matrix which depends only on the left and right calibration matrices L and R. This makes it a relatively simple matter, using least squares, to convert a coefficient vector $\vec{k}$ to the corresponding parameters $\vec{n}$, d for the ground plane in 3D [9].

Note that we can rewrite equation (2) in terms of the more familiar image coordinates $\vec{x}^{\,l}$ and $\vec{x}^{\,r}$ as

$$\begin{aligned} x_1^l &= m_1(\vec{x}^{\,r}, \vec{k}) = (k_1 x_1^r + k_2 x_2^r + k_3)/(k_7 x_1^r + k_8 x_2^r + k_9), \\ x_2^l &= m_2(\vec{x}^{\,r}, \vec{k}) = (k_4 x_1^r + k_5 x_2^r + k_6)/(k_7 x_1^r + k_8 x_2^r + k_9), \end{aligned} \tag{5}$$

which is obtained by dividing through by the third coordinate of $\vec{w}^{\,l}$.

4 Disparity Measurement

Once the left and right views have been brought into a rough alignment, perhaps by using the warp specified by (5) along with a rough guess for the floor parameters $(\vec{n}, d)$, we need to be able to do two things. First, we must refine the coefficients $\vec{k}$ of the mapping to accommodate imperfections in the slope of the floor and/or small tilts of the robot. Secondly, once an accurate estimate of the image warp has been obtained, we need to estimate the probability that various image points correspond to the floor, rather than floor anomalies. Both of these steps require the measurement of the local relative shifts, that is, the *disparity* between the two images.

Many different disparity measurement techniques can be used (see [8] for a survey). Here we consider so called gradient-based methods which rely on a constancy assumption of the form

$$B_l(\vec{x} + \vec{d}(\vec{x}); \vec{k}) = B_r(\vec{x}), \tag{6}$$

where B_l and B_r are filtered and possibly warped versions of the original left and right images, respectively. The constancy assumption (6) provides a constraint on the disparity $\vec{d}(\vec{x})$ which, in gradient-based approaches, is approximated by linearizing about some initial guess $\vec{d}_0(\vec{x})$. This leads to the linear disparity constraint equation

$$(c_1(\vec{x}), c_2(\vec{x})) \cdot (\vec{d}(\vec{x}) - \vec{d}_0(\vec{x})) + c_3(\vec{x}) = 0, \tag{7}$$

where (c_1, c_2) involve the gradient of B_l at $\vec{x} + \vec{d}_0$, and c_3 involves the difference between B_l and B_r.

For our test cases reported below we have chosen a phase-based disparity scheme [4]. In particular, given a stereo pair we have the option of prewarping these images according to an initial guess, $\vec{k}_0$, for $\vec{k}$. This prewarping is done using bilinear interpolation on the grey levels. The resulting images are then bandpass filtered using complex 15×15 kernels based on the steerable quadrature pair G_2 and H_2 developed in [5]. This pair has been scaled so that the peak frequency occurs at wavelength $\lambda = 8$ pixels, and four spatial orientations separated by 45 degrees are used. The convolution responses are subsampled every second pixel both horizontally and vertically, and then quantized to 8 bits. Phase gradients are computed using the technique described in [4]. To reduce the number of noisy disparity constraints we discarded filter responses with too low an amplitude (i.e. below the response obtained from a grating having a half-amplitude of 3 grey levels), or too close to a phase singularity (i.e. we used $\tau = 1.5$ for the singularity neighbourhood detection in [4]). The phase constancy assumption between the left and right images then provided disparity constraints of the form (7), for each of the four filter orientations. The error in the linearization used in the derivation of this constraint equation was kept as small as possible by using discrete shifts in the subsampled grid, according to the current guess for the disparity at each pixel.

From the results of the previous section, we know the disparity field for the floor plane is

$$\vec{d}(\vec{x}) = \vec{m}(\vec{x}, \vec{k}) - \vec{m}(\vec{x}, \vec{k}_0), \tag{8}$$

where $\vec{k}_0$ are is related to any prewarping used. Substitution of equation (8) into the disparity constraint equation (7) provides

$$(c_1(\vec{x}), c_2(\vec{x})) \cdot \left[\vec{m}(\vec{x}, \vec{k}) - \vec{m}(\vec{x}, \vec{k}_0)\right] + c_3(\vec{x}) = 0. \tag{9}$$

We get one constraint of this form for each disparity constraint vector $\vec{c}(\vec{x})$ that we measure during the image matching stage. One might consider solving the resulting set of equations for $\vec{k}$ using nonlinear least squares. The trouble is that such a straight forward approach cannot be expected to work when there are significant outliers in these regression equations due to floor anomalies. A successful solution strategy must be robust to the presence of such outliers.

5 Mixture Model for Disparity

For a robust solution we follow the mixture model approach described in [11, 10]. The idea is to consider the disparity $\vec{d}(\vec{x})$ as arising from one of several simple distributions. In particular, the disparity may arise from a point on the floor, in which case we assume its value is distributed according to the simple Gaussian model

$$p_f(\vec{d'}|\vec{x},\vec{k}) = N(\vec{d'}|\vec{d}(\vec{x},\vec{k}),C). \tag{10}$$

Here $N(\vec{d'}|\vec{d},C)$ is given by

$$N(\vec{d'}|\vec{m},C) = \frac{1}{\sqrt{2\pi|C|}} \exp(-\frac{1}{2}(\vec{d'}-\vec{d})^T C^{-1}(\vec{d'}-\vec{d})). \tag{11}$$

with $\vec{d}(\vec{x},\vec{k})$ the mean and C the covariance of the distribution. Note that we have taken the mean to be the disparity provided by the warp parameters $\vec{k}$. Also, the covariance C represents both a tolerance for imperfections in the floor and errors in the disparity measurements themselves. In our experiments below we take

$$C = \sigma^2 I, \quad \sigma = \lambda/16. \tag{12}$$

For $\lambda = 8$, then, the standard deviation of the disparity is thus just a half a pixel in any direction.

In addition to points on the floor, the actual disparity may arise from anomalies close to the floor. The phase-based method can produce constraints within a disparity range of $\pm\lambda/2$ of the initial guess. Disparities outside of this range will produce false targets and outlier measurements (see below). The stereo configuration specifies that the true disparities must lie along particular epipolar lines, which can be easily computed from L, R and the current image position $\vec{x}$ [9]. To model this set of disparities we use the anisotropic Gaussian

$$p_e(\vec{d'}|\vec{x},\vec{k}) = N(\vec{d'}|\vec{d}(\vec{x},\vec{k}),C(\vec{x})). \tag{13}$$

Again the mean of the distribution is just our current disparity estimate, $\vec{d}(\vec{x},\vec{k})$. The covariance $C(\vec{x})$ is taken to have a principle axis directed along the epipolar line at $\vec{x}$, with the second axis perpendicular to the epipolar line. We take the standard deviation along the epipolar line to be $\lambda/2$, which is roughly the range of disparities that can be measured with the phase-based approach. In the perpendicular direction we use the standard deviation, $\sigma = \lambda/16$, which is the same as the one used for p_f above.

The two distributions p_e and p_f are used to define the likelihoods of measuring a constraint, $\vec{c}$, given that the disparity arises from that distributions. For example, the likelihood that $\vec{c}$ arises from the distribution p_f is

$$p_1(\vec{c}|\vec{x},\vec{k}) = a_0\frac{1}{\lambda} + a_1 \int_{\vec{c}\cdot(\vec{d'},1)=0} p_f(\vec{d'}|\vec{x},\vec{k})ds. \tag{14}$$

Here ds in the integral represents arclength along the specified line. We have also included a flat distribution in this expression for p_1 in order to account for outliers in the measurement process which arise when the phase constancy assumption fails. In all the experiments we set the proportion of measurement outliers, a_0, at 0.25. This allows for 25% of the constraints measured off of the floor to be outliers. The above expression for p_1 can be simplified to

$$p_1(\vec{c}|\vec{x},\vec{k}) = a_0\frac{1}{\lambda} + \frac{a_1}{\sqrt{2\pi}\sigma}\ \exp(-\frac{[\vec{c}\cdot(\vec{d}^T(\vec{x},\vec{k}),1)]^2}{2\sigma^2}), \tag{15}$$

which is just a mixture of a uniform distribution and a Gaussian distribution having standard deviation σ. A similar approach provides a closed form expression for $p_2(\vec{c}|\vec{x},\vec{k})$, which is the likelihood of observing constraint $\vec{c}$ from the distribution $\vec{p}_f$.

Finally, 'outlier' disparities beyond the range of the phase-based method can also occur, as well as bogus constraints from regions which are only visible in one of the images. The distribution of these outlier constraints is taken to be a uniform distribution

$$p_0(\vec{c}|\vec{d}) = \frac{1}{\lambda} \tag{16}$$

in the disparity range of $\pm\lambda/2$ about the current estimate $\vec{d}(\vec{x},\vec{k})$. This range is again the limit imposed by phase-based methods. Note that we are using four separate orientation channels and therefore we can model each outlier distribution as effectively one dimensional.

Given the three likelihood functions, p_0, p_1 and p_2, the mixture model for the distribution of a constraint vector $\vec{c}(\vec{x})$ at image location $\vec{x}$ is then taken to be

$$p(\vec{c}|\vec{x},\vec{m},\vec{k}) = m_0p_0(\vec{c}|\vec{x},\vec{k}) + m_1p_1(\vec{c}|\vec{x},\vec{k}) + m_2p_2(\vec{c}|\vec{x},\vec{k}). \tag{17}$$

We refer to p_n as the n^{th} *component probability* distribution of this mixture model. The corresponding mixture proportion, m_n, gives the prior probability that a constraint arises from this component of the distribution. Of course, the mixture proportions m_n, $n = 0, 1, 2$ must sum to one.

6 The EM-Algorithm

Given a set of disparity constraint vectors, $\{\vec{c}_j(\vec{x}_j)\}_{j=1}^J$, we seek parameter values $\vec{k}$ and mixture probabilities $\{m_0, m_1, m_2\}$ which provide a *maximum likelihood* fit to this data set. In particular, assuming that the observations are independently distributed, the log likelihood of generating this set of observations from a specific mixture model of the form (17) is

$$\log L(\vec{m},\vec{k}) = \sum_{j=1}^{J} \log p(\vec{c}_j|\vec{m},\vec{d}(\vec{x}_j,\vec{k})). \tag{18}$$

At a local extrema, it can be shown that the parameters $\vec{m}$ and $\vec{k}$ must satisfy

$$\begin{aligned} \sum_{j=1}^{J} q_{nj} &= \gamma m_n, \\ \sum_{j=1}^{J} q_{nj} \frac{\partial}{\partial \vec{k}} \log p_n(\vec{c}_j | \vec{d}(\vec{x}_j, \vec{k})) &= 0, \end{aligned} \tag{19}$$

for $n = 0, 1, 2$. Here γ arises as a Lagrange multiplier in imposing the constraint that the mixture probabilities sum to one. The quantity q_{nj} satisfies

$$q_{nj} = \frac{m_n p_n(\vec{c}_j | \vec{x}_j, \vec{k})}{\sum_{n=0}^{2} m_n p_n(\vec{c}_j | \vec{x}_j, \vec{k})}. \tag{20}$$

and is called an *ownership probability.* It is the probability that the j^{th} constraint belongs to the n^{th} component. These equations for a maximum likelihood fit have been derived by a number of authors; for further details see [13].

These equations suggest an iterative algorithm, known as the EM-algorithm [13], for obtaining a maximum likelihood fit for the parameters $\vec{m}$ and $\vec{k}$. Given an initial guess for these parameters, we first estimate the ownership probabilities q_{nj} for each constraint belonging to each component. This expectation step, or "E-step", simply involves evaluating (20). Next, the maximization step, or "M-step", maximizes L with these ownerships held fixed. The result is a simple iterative algorithm which is guaranteed to increase the log likelihood of its fit each iteration [13]. In the experiments discussed below we do not accurately locate a maximum value for $\vec{k}$ during the M-step; instead we use only one iteration of Newton's method applied to equation (19b).

It is important to understand precisely what is being modeled by the mixture model given in equation (17). Note that the mixture proportion m_n is independent of the image location, and thus represents a spatial average across the entire image of the occurrence of the n^{th} mode. Clearly, for a stereo pair which includes some floor anomalies (see Figure 1) the mixture proportions should depend on image location. In any region consisting of a floor anomaly we would expect m_0 and m_2 to account for most of the constraints, while in a region containing only the floor, we expect m_1 to dominate. This spatial variation is not being modeled here.

This fact that the mixture proportions represent averages over the entire image is important when we come to display ownership results at specific pixels. For the purposes of display it is convenient to show *ownership likelihoods* for the various components of the mixture model at each pixel. These likelihoods avoid the global averaging process discussed above, and clearly exhibit the information available at each pixel separately. The ownership likelihoods are computed simply by clamping the mixture proportions at uniform probability levels, that is, for a three component mixture $m_n = 1/3$ for each n, and then evaluating q_{nj} according to equation (20).

One final implementation detail, which is important for inaccurate initial guesses, is that it is often convenient to use a coarse to fine strategy. In particular, by decimating the original images by a factor of 2, 4, or 8, we obtain disparity measurements corresponding

Figure 1: A calibrated stereo image pair.

to wavelengths 16, 32, and 64, respectively. These longer wavelengths increase the range of disparities that a single constraint vector $\vec{c}(\vec{x}_j)$ can measure. In addition to using longer wavelengths, we have also found it useful to begin with the standard deviation, σ, of the disparity model at $\lambda/8$ and decrease it slowly to the value $\lambda/16$. This provides a form of graduated non-convexity which helps avoid local minima.

7 Experimental Results

The calibrated image pair shown in Figure 1 was obtained using a camera separation of 56cm, with the distance to the floor near the middle of the image about 180cm and the angle of the center ray with the floor about 45 degrees. The floor anomalies consisted of books, which have thicknesses of of 0cm (top row, a piece of paper), 5.0cm (second row), 0.5cm, 4.3cm, 1.0cm (third row, left to right), and 1.2cm, 1.7cm, 1.8cm (bottom row). The mixture model procedure described in the previous section was used to fit the rational warp parameters $\vec{k}$. The convergence behaviour was found to be excellent in that the initial guess could be inaccurate, and the method converged with 5 to and 10 iterations. The fitted warp parameters were observed to bring points on the floor into a nearly perfect correspondence. Similar results have been observed on many other image pairs.

The ownership likelihoods for the three components of our mixture model are shown in Figure 2. Note that in regions with little texture, no disparity constraints were obtained, and thus the likelihood is left at the intermediate level of 1/3 for all three components. Also, for regions in which the texture of the floor (or an object) is roughly horizontal, we see that the likelihoods for mixture components p_1 and p_2 are roughly equal. This is as expected, since gradients which are nearly perpendicular to the epipolar lines provide no information about depth, and are thus equally likely to arise from either process p_1 of p_2. In the remaining places on the floor, the likelihood maps show strong evidence for the floor model, while the objects out of the floor plane are identified as outliers. The book on the left end of the second row from the bottom, which is 0.5cm high, appears to be close to the resolution limit

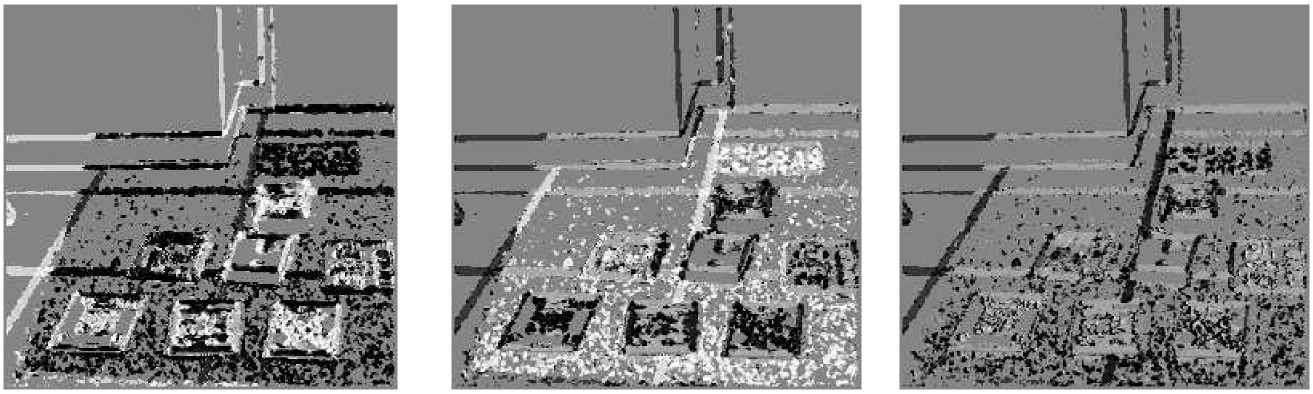

Figure 2: Log likelihood ownership maps for outliers (left), the floor model (middle), and for objects near the floor (right), consistent with the epipolar geometry (right), for the stereo pair in Figure 1. Here white corresponds to high likelihood, and the uniform grey background represents equal likelihood.

of our system in that its top surface and left edge are not clearly identified as an anomaly.

We have used the recovered warp parameters $\vec{k}$ to compute the postion of the floor plane to within the accuracy of our "ground truth" measurements. In fact, during a first pass at running the system we found that our recovered warp parameters were not well modeled by equation (4) for any values of $\vec{n}$ and d. It turned out that the problem was due to an inaccurate calibration of L and R, and a satisfactory solution was found after recalibrating the system. This raises the possibility that the calibration of the FAD system can be monitored on-line.

In summary, we have demonstrated a simple approach for fitting planar models to stereo data which is robust in the presence of a significant number (eg 50%) of outliers. The model involves fitting eight parameters for a rational image warp, which we have shown is exact for planar surfaces and for cameras free from radial distortion. Moreover the results of the fitting procedure can be used to verify which portions of the floor are safe for transit.

References

[1] N. Ayache. *Artificial vision for mobile robots.* MIT Press, Cambridge, MA, 1991.

[2] P. Burt, P. Anandan, K. Hanna, G. van der Wal, and R. Bassman. A front-end vision processor for vehicle navigation. In *IAS-3*, pages 653–662, Pittsburgh, PA, 1993.

[3] O. Faugeras. *Three-dimensional computer vision.* MIT Press, Cambridge, MA, 1993.

[4] D. Fleet, A. Jepson, and M. Jenkin. Phase-based disparity measurement. *CVGIP: IU, 3(2)*, pages 198–210, 1991.

[5] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE PAMI, 13(9)*, pages 891–906, 1991.

[6] B. Horn. *Robot Vision.* MIT Press, Cambridge, MA, 1986.

[7] M. Jenkin, N. Bains, J. Bruce, T. Campbell, B. Down, P. Jasiobedzki, A. Jepson, B. Majarais, E. Milios, B. Nickerson, J. Service, D. Terzopoulos, J. Tsotsos, and D. Wilkes. ARK: Autonomous mobile robot for an industrial environment. In *IEEE/RSJ IROS*, Munich, Germany, 1994. (in press).

[8] M. Jenkin, A. Jepson, and J. Tsotsos. Techniques for disparity measurement. *CVGIP: IU*, 53(1):14–30, 1991.

[9] M. R. M. Jenkin and A. Jepson. Detecting floor anomalies. *ARK Tech. Rep.* in preparation.

[10] A. Jepson and M. Black. Mixture models for optical flow computation. In *Proc. of the DIMACS Workshop on Partitioning Data Sets*, Providence, RI, 1994. AMS Pub.

[11] A. Jepson and M. J. Black. Mixture models for optical flow computation. In *Proc. Computer Vision and Pattern Recognition, CVPR-93*, pages 760–761, New York, June 1993.

[12] D. Kim and R. Nevatia. Indoor navigation without a specific map. In *IAS-3*, pages 268–277, Pittsburgh, PA, 1993.

[13] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker Inc., N.Y., 1988.

Acknowledgments

Funding for this work was provided, in part, by the ARK (Autonomous Robot for a Known environment) Project, which receives its funding from PRECARN Associates Inc., Industry Canada, the National Research Council of Canada, Technology Ontario, Ontario Hydro Technologies, and Atomic Energy of Canada Limited. The authors also gratefully acknowledge the financial support of NSERC Canada.