

International Journal of Computer Vision 39(1), 5–23, 2000
© 2000 Kluwer Academic Publishers. Manufactured in The Netherlands.

Practical Structure and Motion from Stereo When Motion is Unconstrained

NICHOLAS MOLTON AND MICHAEL BRADY
Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK

Abstract. This paper describes a system which robustly estimates motion, and the 3D structure of a rigid environment, as a stereo vision platform moves through it. The system can cope with any camera motion, and any scene structure and is successful even in the presence of large jumps in camera position between the capture of successive image pairs, and when point matching is ambiguous. The system was developed to provide robust obstacle avoidance for a partially sighted person.

The process described attempts to maximise use of the abundant information present in a stereo sequence. Key features include the use of multiple stereo match hypotheses, efficient motion computation from three images, and the use of this motion to ensure reliable matching, and to eliminate multiple stereo matches. Points are reconstructed in 3D space and tracked in a static coordinate frame with a Kalman Filter.

This results in good 3D scene reconstructions. Structure which is impossible to match with certainty is absent, rather than being incorrectly reconstructed. As a result, the system is appropriate for obstacle detection. The results of processing some indoor and outdoor scenes, are given in the paper, and practical issues are highlighted throughout.

Keywords: structure from motion, stereo vision, unconstrained motion, map building, obstacle avoidance, walker-mounted

1. Introduction

Structure from motion is an important computer vision method, applicable to many complex sensing situations. While analysis of a single image sequence provides much scene information, a truly robust system, capable of recovering full 3D structure is possible with a stereo camera system. This paper describes an implementation of stereo structure from motion and discusses some of the issues arising.

1.1. Overview of Stereo Structure from Motion Methods

Most 3D computer vision systems analyse image streams either by optic flow or region matching. The power of optic flow methods is that they calculate structure entirely from image derivatives. A good example for stereo vision is given in the paper by Stein and Shashua (1998). Methods of this type are limited by the assumption of incremental camera movement, and will fail if anything in the scene moves faster than a few pixels per frame. In contrast, matching methods can cope with much larger movements per frame, but the process of matching is much more difficult. Since rapid camera motions are inevitable in our application, a matching method is used for this work.

Many structure from motion systems have been devised for monocular vision (for example, Harris, 1987; Bouguet and Perona, 1995), but relatively few for stereo vision. In a matching approach, what is matched varies between points of interest (features) to almost the entire image (area-correlation). Moravec implemented a successful feature based, nine-eye-stereo system for non real-time obstacle avoidance on a cart in the early 1980s (Moravec, 1983). Stereo image pairs were used to calculate 3D structure, and the relationship between the structure sets at different times used to estimate camera motion and to identify outliers. This system was enhanced by Matthies and Shafer (1987), who used a linear approximation to the uncertainty in stereo

reconstruction to deduce motion 5–10 times more accurately than could be done through least squares minimisation of distance between points in the two point sets. Other work on the use of feature-based stereo vision systems for obstacle avoidance includes the system by Li et al. (1995).

Young and Chellappa (1990) describe a method for identifying the motion of a camera which moves according to a motion model (they analyse a constant acceleration and constant precession model) from sets of matched 3D point positions. The parameters of the model are obtained using a Kalman Filter. Uniqueness of the solution is proved and the maximum achievable accuracy for a particular model, which also gives an indication of noise immunity, is calculated. Zhang and Faugeras (1992) describe a method for obtaining motion from two sets of 3D line positions, using models of uncertainty appropriate to information obtained by computer vision triangulation. They also use a Kalman Filter to estimate motion, as it enables hypothesised motion estimates to be updated continually during the motion verification process.

Area-correlation work includes the system by Matthies et al. (1994, 1996), which uses dense area correlation stereo to produce reconstructions in real-time for obstacle detection by slow moving vehicles. They describe a method for estimating the probability of detection and of false detection as a function of the vision system properties and vehicle speed, which is useful in the design of a stereo vision system (Matthies and Grandjean, 1994). A detectability analysis was also done, as part of the ASMONC project, by Se and Brady (1998).

Several real time area-correlation stereo systems are currently available. Ferrari et al. (1990) devised a simple non-calibrated, real-time stereo system for obstacle detection on a vehicle. Murray and Little (1998) use a trinocular system for building simple maps. They include a novel method for removing locally-supported mismatched image areas. Konolige (1998) describes a commercial real-time stereo vision system mounted on a single circuit board capable of dynamic camera recalibration.

Area-based and feature-based matching are ultimately quite similar. Area-based methods can be thought of as being corner feature-based with a very loose feature strength threshold (almost every pixel is a feature) defined by the interest operator used to cut-out homogeneous image regions. The systems developed by Matthies make the transition from feature-based to area-correlation without the underlying method being changed.

The density of area-based reconstruction methods offers a level of robustness through checks for continuity in the recovered structure. Robustness is limited by the fact that structure is not usually tracked through time (because there are too many features to process). However, since temporal matching is avoided, rigidity need not be assumed.

In a feature-based method density is typically sacrificed to accommodate the relatively computationally expensive operation of tracking features through time. This tracking process results in very robust and accurate structure measurements. Future processing power may allow the robustness benefits of feature tracking to be included in a dense system, highlighting the need for research into the exploitation of the feature tracking approach.

1.2. Application

The particular application within which this work was carried out was the ASMONC (Autonomous System for Mobility, Orientation, Navigation and Communication) project (Molton et al., 1999; Se, 1998; Molton, 1999). This project aimed to provide a mobility system for a partially sighted person. A key requirement is obstacle avoidance without the aid of maps or beacons. Structure and motion from stereo vision has the potential to achieve this. Since the vision system is carried by a person through unstructured environments, it must be able to cope with motions and scene structure which are essentially unconstrained, even in the presence of a walker model (Molton and Brady, accepted for publication).

The system is based around a backpack containing the system electronics. Greyscale cameras are attached to the end of rigid arms which extend from the backpack over each of the wearer's shoulders. This design enables the user to walk naturally, and to negotiate any hazard which a sighted person would be able to traverse. Between each camera and arm is a mounting bracket which allows the camera to be tilted towards the ground, and a small amount of camera roll and pan (for proper camera alignment). The cameras are interfaced to an image capture and processing board based on the TI C40 processor, which though it would now be considered slow, was current at the time of the project. Images can be captured as stereo-pairs (an image captured from the right and left cameras at the same

time) but only at 1 Hz. The method used for camera calibration is that described by Tsai (1987). Camera alignment was achieved by adjusting camera position until two suitably positioned marks on a wall in front of the backpack appear at the centre of the two images.

The structure and motion from stereo method described in this paper can be applied generally to any task requiring the structure of a largely static scene from stereo vision. The system is particularly suited to applications where camera motion is unconstrained, and will not fail if feature motion between frames is large.

1.3. Overview of Paper and Contribution

Most stereo vision systems tend to isolate image information captured at different times. The scene is typically reconstructed for each stereo pair before comparison of 3D positions in order, for example, to calculate motion. Section 2.1 argues that this may not be the most efficient approach to take.

The contribution of the work described here is a method with a structure designed to more fully exploit the information available in stereo sequences with the aims of maximum efficiency and robustness. Contributions are:

- The use of multiple initial stereo hypotheses, to retain many possibilities in the presence of matching uncertainty.
- An analysis which determines the most efficient approach to motion estimation to be calculation from three images before performing the reconstruction.
- A method of motion calculation from three sets of points in three images is described. This gives better results than calculating the motion using the fundamental matrix.
- Feature matching is only done in earnest when a reliable motion estimate is available which can be used to assess the consistency of feature matches.

Robustness is critical for an obstacle avoidance system, and this is where the system succeeds most: the method will not reconstruct strong texture which cannot be matched reliably, for example, the leaves on a tree.

The next section describes the camera motion through observation of image feature points. The method chosen for this has a substantial influence on the format of the structure from motion system as a whole, which is introduced in Section 3. Section 4 describes the practical calculation of motion within the structure from motion process, and Section 5 then describes in detail the methods used to calculate structure. Section 6 discusses uncertainty in the resulting point structure and the use of a Kalman Filter to give continuity to the point reconstruction process. Section 7 then describes results from using the system and some practical issues. The last section summarises and suggests enhancements which might be made to the system. Most of the work described here is discussed in greater detail in Molton (1999).

2. Motion Measurement by Stereo Vision

This section describes the calculation of motion from the positions of points in a stereo image sequence. Motion is calculated before structure, because structure can be found more reliably given camera motion. The section begins by deciding which images should be included in the feature matching process; after which motion calculation methods from the matches are described.

2.1. Images to be Matched for Motion from Stereo

When a new stereo image pair arrives, it is assumed that, in the steady state, matches between the left and right images of the previous stereo pair already exist. With reference to Fig. 1, the set of matches SM_{t-1} is known. Sufficient further matching must be carried out to allow the motion, M, to be calculated from the possible sets of matches: *LIM*, *RIM*, and SM_t in the most efficient manner. Two 'diagonal' match sets also exist, corresponding to, for example, matching between a left image and the following right image. These are ignored for the reason given below.

Alternative methods for finding motion from different sets of matches (match sets yet to be computed are

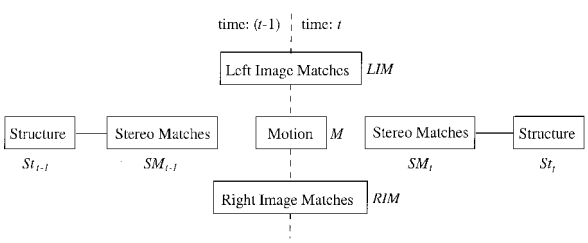


Figure 1. Image matching options.

shown underlined) are:

1. $SM_{t-1} + \underline{LIM} \rightarrow M$
2. $SM_{t-1} + \underline{RIM} \rightarrow M$
3. $\underline{LIM} + \underline{RIM} \rightarrow {}^{1}M$
4. $\underline{SM_t} + \underline{LIM} \rightarrow M$
5. $\underline{SM_t} + \underline{LIM} \rightarrow M$

Since (1) and (2) necessitate only one matching process, and matching is relatively computationally expensive, one of these (with no loss of generality, the first) is chosen for the motion calculation. Options under failure of this method are discussed in Molton (1999).

One of the diagonal match sets referred to earlier might also have been used as the addition required for motion. This would result in more difficult matching, and a more complex calculation than would method (1), and so was not attempted. If processing was rapid, more than one method might be used (perhaps methods (1) and (2)), and the results compared, for added robustness. This would be a straightforward extension to the method developed here. The limited bandwidth of the system developed precluded this.

2.2. Motion Recovery Methods

Six motion parameters describe camera movement between frames: three components of rotation about the camera centre; followed by three components of translation. Each point match (if chosen sensibly) provides one piece of information to help retrieve these, so at least six matches are needed, taken from SM_{t-1} and *LIM*.

One possible means of calculating the motion is through the fundamental matrix, which relates two consecutive left images. Theoretically, this would require five left image matches; but in practice seven are needed for the computation (Torr and Murray, 1997). One of the stereo matches (SM_{t-1}) is then required to complete the calculation of motion (resolving the speed-scale ambiguity). An alternative is to recover motion from a set of three points, matched in stereo at $t-1$ (3 SM_{t-1} matches) and also matched between the latest two left images (3 *LIM* matches). These matches enable the exact positions in space (relative to the cameras) of the three points at time $t-1$ to be recovered, and their position at time t to be limited to a line in space. This situation can be formulated as Eq. (1)

$$\begin{aligned}&(\lambda_1\mathbf{x}'_1 \quad \lambda_2\mathbf{x}'_2 \quad \lambda_3\mathbf{x}'_3)\\ &\quad = (\mathbf{R})(\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3) + (\mathbf{t} \quad \mathbf{t} \quad \mathbf{t}),\end{aligned} \tag{1}$$

in which the three-dimensional position vectors of the three points before the motion are $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, relative to the left camera. After the motion, the 3D position of each point is known to lie on a ray entering the left camera's optical centre, defined in relation to the left camera by, for example $\lambda_1\mathbf{x}'_1$, where the scalar λ_1 is unknown. $\mathbf{R}$ (3 dof) is the camera rotation matrix, and $\mathbf{t}$ the translation (3 dof) between $t-1$ and t. Since Eq. (1) represents nine linear equations and there are three degrees of freedom in $\mathbf{R}$, three in $\mathbf{t}$, and three scalar unknowns, λ_{1-3}, a non-linear solution is possible, at least in theory. The solution that we propose breaks the problem into two parts: (a) calculation of λ_{1-3}; and (b) the calculation of $\mathbf{R}$, and $\mathbf{t}$ (both parts are also non-linear). An intuitive geometric solution is used for each.

Calculation of the λ values exploits the fact that the 3D distances and angles in the triangle formed by the three points does not change between time instants. This results in λ_1 to λ_3 given as the solution of Eqs. (2) (computed numerically), and (3).

$$\begin{aligned}&\gamma_1\lambda_1^2 + \lambda_1\left(\gamma_2\sqrt{\lambda_1^2\phi_1 + \rho_1} + \gamma_3\sqrt{\lambda_1^2\phi_2 + \rho_2}\right)\\ &\quad + \gamma_4\sqrt{(\lambda_1^2\phi_1 + \rho_1)(\lambda_1^2\phi_2 + \rho_2)} + \gamma_5 = 0\end{aligned} \tag{2}$$

$$\begin{aligned}\lambda_2 &= \frac{\lambda_1(\mathbf{x}'_1 \cdot \mathbf{x}'_2)}{(\mathbf{x}'_2 \cdot \mathbf{x}'_2)} + \sqrt{\lambda_1^2\phi_2 + \rho_2}\\ \lambda_3 &= \frac{\lambda_1(\mathbf{x}'_1 \cdot \mathbf{x}'_3)}{(\mathbf{x}'_3 \cdot \mathbf{x}'_3)} + \sqrt{\lambda_1^2\phi_1 + \rho_1}\end{aligned} \tag{3}$$

$$\gamma_1 = \frac{2[(\mathbf{x}'_1 \cdot \mathbf{x}'_2)(\mathbf{x}'_2 \cdot \mathbf{x}'_3)(\mathbf{x}'_1 \cdot \mathbf{x}'_3) - (\mathbf{x}'_1 \cdot \mathbf{x}'_2)^2(\mathbf{x}'_3 \cdot \mathbf{x}'_3) - (\mathbf{x}'_1 \cdot \mathbf{x}'_3)^2(\mathbf{x}'_2 \cdot \mathbf{x}'_2) + (\mathbf{x}'_1 \cdot \mathbf{x}'_1)(\mathbf{x}'_2 \cdot \mathbf{x}'_2)(\mathbf{x}'_3 \cdot \mathbf{x}'_3)]}{(\mathbf{x}'_2 \cdot \mathbf{x}'_2)(\mathbf{x}'_3 \cdot \mathbf{x}'_3)}$$

$$\gamma_2 = \frac{2[(\mathbf{x}'_2 \cdot \mathbf{x}'_3)(\mathbf{x}'_1 \cdot \mathbf{x}'_2) - (\mathbf{x}'_1 \cdot \mathbf{x}'_3)(\mathbf{x}'_2 \cdot \mathbf{x}'_2)]}{(\mathbf{x}'_2 \cdot \mathbf{x}'_2)}$$

$$\gamma_3 = \frac{2[(\mathbf{x}'_2 \cdot \mathbf{x}'_3)(\mathbf{x}'_1 \cdot \mathbf{x}'_3) - (\mathbf{x}'_1 \cdot \mathbf{x}'_2)(\mathbf{x}'_3 \cdot \mathbf{x}'_3)]}{(\mathbf{x}'_3 \cdot \mathbf{x}'_3)}$$

$$\gamma_4 = 2(\mathbf{x}'_2 \cdot \mathbf{x}'_3)$$

$$\gamma_5 = 2[(\mathbf{x}_1 \cdot \mathbf{x}_2) + (\mathbf{x}_1 \cdot \mathbf{x}_3) - (\mathbf{x}_1 \cdot \mathbf{x}_1) - (\mathbf{x}_2 \cdot \mathbf{x}_3)]$$

$$\phi_1 = \frac{(\mathbf{x}'_1 \cdot \mathbf{x}'_3)^2 - (\mathbf{x}'_1 \cdot \mathbf{x}'_1)(\mathbf{x}'_3 \cdot \mathbf{x}'_3)}{(\mathbf{x}'_3 \cdot \mathbf{x}'_3)^2}$$

Medit Ex-1014, Page 4

$$\phi_2 = \frac{(\mathbf{x}'_1 \cdot \mathbf{x}'_2)^2 - (\mathbf{x}'_1 \cdot \mathbf{x}'_1)(\mathbf{x}'_2 \cdot \mathbf{x}'_2)}{(\mathbf{x}'_2 \cdot \mathbf{x}'_2)^2}$$

$$\rho_1 = \frac{(\mathbf{x}_1 \cdot \mathbf{x}_1) + (\mathbf{x}_3 \cdot \mathbf{x}_3) - 2(\mathbf{x}_1 \cdot \mathbf{x}_3)}{(\mathbf{x}_3 \cdot \mathbf{x}_3)}$$

$$\rho_2 = \frac{(\mathbf{x}_1 \cdot \mathbf{x}_1) + (\mathbf{x}_2 \cdot \mathbf{x}_2) - 2(\mathbf{x}_1 \cdot \mathbf{x}_2)}{(\mathbf{x}_2 \cdot \mathbf{x}_2)}$$

Up to four solutions for these positions are possible (Wolfe, 1991), but the one closest to the positions at time $t-1$ is chosen, as it is most likely to be the correct solution in the usual case of limited camera movement.

Once the positions of the three points at both time instants is fully known, the Euclidean transformation between the two sets of positions is determined. The method that we used forms a triangle from the three points at each time, then calculates the rotation required to make the triangles parallel, and the translation required to make the triangles coincident after the rotation is applied to the first.

The main weakness of the fundamental matrix method is that it fails under certain point configurations and motions. For success, the points must not be coplanar; however this is far from unlikely for ASMONC, as the cameras are pointing at the ground. Also, translatory motion must occur between every image pair capture. The three-point triangulation method that we developed works under *any* type of motion, or under zero motion, so long as the points are not collinear. This does not present a problem in practice.

We found that the three-point triangulation method gave slightly better results in simulated and real image data tests, with motion and structure suitable for non-degenerate computation of the fundamental matrix method (Molton, 1999). This may be due to the stereo matches, upon which the method is more reliant, corresponding to a wider camera baseline than the temporal matches.

Given the image matching strategy chosen, $SM_{t-1} + \underline{LIM} \rightarrow M$, there is a range over which the balance between the temporal (*LIM*) and stereo (SM_{t-1}) match significance can be altered. The three-point and fundamental matrix methods are at the extremes of the range of methods which could be used given these sets of matches. One might fail, but the other work, if one set of matches (*LIM* or SM_{t-1}) is of poor quality.

3. Structure and Motion from Stereo

This section presents an overview of the structure from motion process that we adopted, defining objectives and describing the corner feature extraction and matching processes used.

3.1. System Objectives

For our application, obstacle avoidance, the principle requirements are that indications of the presence of structure can be relied upon, and that the reconstruction is sufficiently dense. It follows that the point reconstruction should have the following properties:

- No grossly erroneous points should appear in the reconstruction[2]
- As many points as possible should be reconstructed

There is also a set of secondary objectives: to minimise computer power required (for real time operation, or to minimise the cost of the computer used); and to maximise accuracy. Accuracy is *not* a high priority for this application, since the ultimate aim is a reconstruction of the world suitable for obstacle avoidance, rather than for measurement. Moderately sized reconstruction errors (up to 10 cm) are acceptable.

3.2. Overview of the Structure and Motion from Stereo Method

Some vision systems limit the representation of the world to stereo disparity space (Harris and Pike, 1987; Se and Brady, 1998). Given the stereo disparity of points in a scene, it is possible to detect and track obstacles, without using disparity with camera calibration to calculate positions explicitly in Euclidean space. A full Euclidean reconstruction is used here, because the removal of camera platform motion from the motion of moving world obstacles, and the ability to send meaningful information to the person or robot carrying the platform is complicated under a system restricted to disparity space. However, this decision is not critical, as these tasks could be performed in a disparity space system, and reliance on accurate camera calibration avoided.

The approach adopted here is to calculate a global 3D world motion, which can, assuming scene rigidity, be used as a strong guide to stereo and temporal matching, making these relatively easy to achieve. Features which do not move rigidly with the scene are not included in the reconstruction. Because a correct stereo match is necessary to determine the structure of each point, particular emphasis is given to ensuring that

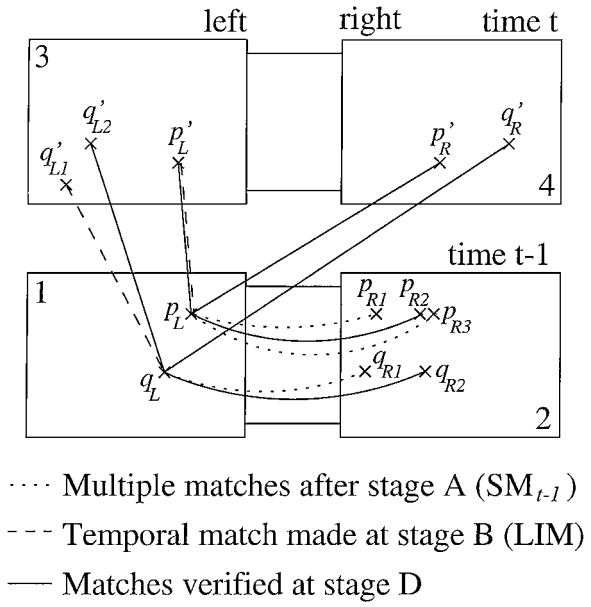


Figure 2. Point matching between four image pairs.

the matches are correct by allowing multiple hypotheses in the stereo matches until the motion estimate is available.

Structure and motion recovery is achieved in four stages, A–D below, performed between the capture of successive image pairs. Just stage A is performed on the first stereo image pair. Stages B, C, D, and A are then performed in that order between capture of subsequent image pairs. Figure 2 illustrates the processes occurring during this cycle, and is referred to in the text. The four stages are:

A. Stereo matching of unmatched points
B. Matching from the penultimate left image to the latest left image
C. Estimation of motion
D. Verification of matches and matching to, and within, the new stereo pair

As described in the previous section, and more fully in Section 5, a set of stereo matches (SM_{t-1}) between the last stereo image pair should exist for many points before the next pair arrives. This will not be the case for all points however (especially when the scene is difficult to match or when the program is initiated) and attempts are first made to match these unmatched points in stereo (Stage A). Stage B matches between one of the images of this pair and the corresponding image in the latest image pair. Stage C estimates camera motion between the two stereo pairs using the matches made so far and the motion estimation method described in Section 2. These motion calculation stages (A–C) are described fully in Section 4. Stage D, described in Section 5, then verifies and corrects the matches made against the camera motion estimate and provides matches to the remaining image in the latest stereo pair. The method is heavily influenced by the strategy for efficient motion calculation decided upon in the last section.

3.3. Corner Detection and Correlation

The Harris corner detector (Harris and Stephens, 1988) was used to detect image corners. The detector is renowned for being consistent in *detection* (corners do not flicker on and off in subsequent images). Potential matches are evaluated by calculating a correlation measure between the patches surrounding the two corners. This is a normalised measure of intensity difference with a moderate penalty for highly differing intensities, to allow for topological differences in part of the image patches (Molton, 1999).

Because of the slow capture rate available with the hardware used, search windows of up to 200 pixels square were necessary. With such large areas to calculate correlations over, a multi-scale correlation approach (for example, the coarse-to-fine approach of Moravec, 1983) may have been more suitable, as it would reduce the number of comparisons required significantly. This was not done because it was envisaged that with more suitable hardware the search area would be much smaller, which would significantly reduce the suitability of such an approach.

4. Structure from Motion: Initial Matching for Motion

This section describes practicalities of the processes required to calculate motion when a new image pair arrives. The section is divided into three parts, corresponding to the three stages, A to C: matching in stereo for unmatched points; temporal matching to the new image pair when it arrives; and calculation of incremental camera motion to the new images.

Medit Ex-1014, Page 6

4.1. Stereo Matching of Unmatched Points (Stage A)

Towards the end of processing for the latest image pair, a set of stereo matches for the pair are known (calculated by the recently completed stage D). These matches will usually be sufficient to provide the SM_{t-1} matches required by motion calculation when the next image pair has arrived. However, points which were detected, but which have not been stereo-matched in stage D are now processed in the hope of establishing further stereo matches. This is done because:

- Adding to the matches forming SM_{t-1} provides more information for motion determination. This is essential when the system is first started, as these new point matches will be the only source of matches for SM_{t-1}.
- The set of possible stereo matches for each point must be established for the following stage D, and there is no benefit in delaying the calculation.

The potential match from each unmatched left image point, to each of those corners detected in the right image which lie near its epipolar line, is assessed using the correlation method described in the previous section. As well as these viable right image corners, the image positions in a very small area around each are also assessed. This is done in the hope of finding a visually more similar image position, even when this does not correspond to a detected corner. This can arise due to corner location inaccuracy, or the detector failing to identify a corner which is very close to other corners.

The strongest *few* right image matches to each left image point are recorded. With reference to Fig. 2, the last time this was done (at time $t-1$), p_L was matched to p_{R1}, p_{R2}, and p_{R3}, and q_L to q_{R1} and q_{R2}. The most important benefit in allowing the stereo match to have multiplicity is that *all* plausible matches to a point are retained until one of the matches can be subsequently considered reliable. In this way points are not lost in the reconstruction through being written off as mismatched.

When contemplating matches, a minimum match correlation threshold is set, along with a maximum value for the number of multiple matches. The motion calculation stage uses a random sampling and support method (RANSAC), described fully in Section 4.3. The figures written next to the lines of Fig. 3 show (for an indoor scene) the effect of varying this maximum stereo multiplicity between 1 and 8 on the number of consistent points (inliers) found in the motion calculation. To enable a fair comparison, the number of random samples used in the RANSAC calculations was set to 10000 throughout. The graph shows that increasing the maximum multiplicity above 3 gives little increase in the number of inliers supporting the motion calculation, whereas lowering the multiplicity below this value results in a significant loss of inliers. As a result, a multiplicity of 4 was used in all experiments in a range of domains.

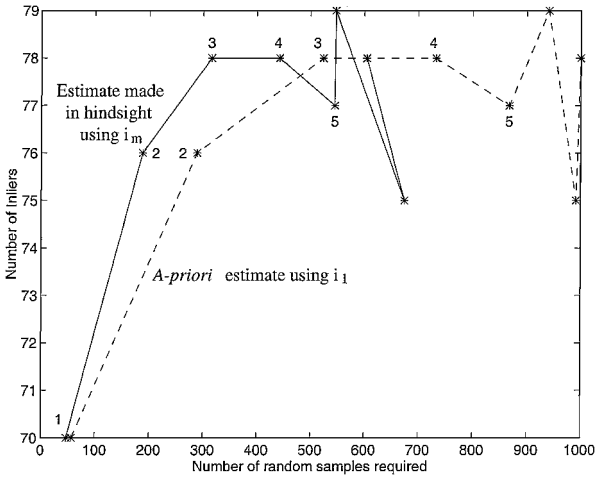


Figure 3. Inliers after motion calculation against an estimate of the number of random samples required, under different stereo match multiplicity. Each '*' represents a different stereo match multiplicity.

The experiments described in Section 7 show that there is little need for stereo match multiplicity for indoor scenes: almost all verified point matches come from the stereo match hypothesis of highest correlation value. For such scenes, differences in the visual appearance of image points tends to be high. When there is more visual ambiguity, for example, from vegetation in outdoor scenes, multiplicity is more necessary, and greater use is made of it.

Section 4.3 demonstrates that the motion calculation performed at stage C takes a time proportional to the cube of the average multiplicity. Fortunately, as the process gets under way the number of points which are not verified by stage D, and so must be multiply matched, reduces significantly (in practice to about half the total, depending on the nature of the scene).

4.2. Matching Between the Most Recent Two Left Images (Stage B)

When a new pair of images is captured, the first thing to be attempted is to match points to them. Since it was

determined in Section 2.1 that, without loss of generality, points should be matched between two consecutive left images, this was done, based again on image patch correlation. A search is made for the match in the newer image, within a neighbourhood of the position of the point in the previous image, or, if a movement model (for example of human walker motion, such as that described in Molton and Brady (accepted for publication)) is available, it is used to guide a more efficient search. Compactness of the search area, hence efficiency, is defined by the motion model accuracy. Since no movement model was used for the experiments described in Section 7 feature motions are large.

Figure 2 illustrates matching between left images: p_L is matched to p'_L, and q_L to q'_{L1}. The matches will be imperfect but, provided camera motion between the two images is found sufficiently confidently at the next stage, they are later ignored. The possibility of multiple matches across time is not entertained because, after motion is known, temporal matches are redefined by the stereo matches found in stage A.

4.3. Estimation of Motion (Stage C)

Matches established in the two previous stages, though crude in stage B, should contain sufficient information for the camera motion between the latest two stereo pairs to be calculated using the three point method described in Section 2.2. Robust statistics are used to deduce the motion, since a significant proportion of matches tend to be incorrect and because multiple matches can be accommodated. RANSAC (Fischler and Bolles, 1981) was chosen for this, resulting in a method which randomly samples sets of three points from those available, and chooses the result consistent with the movement of the largest number of the remaining points. An example of a random three point set from Fig. 2 might be the point p_L (with matches $p_L - p_{R1}$ and $p_L - p'_L$), the point q_L (with matches $q_L - q_{R2}$ and $q_L - q'_{L1}$), and a third point, with one of its stereo matches randomly selected if necessary. Another stage might be applied after the robust RANSAC stage, to maximise accuracy, as suggested by Torr and Murray (1997); but to date this has not been done.

For motion estimation, the use of *multiple* hypothesised stereo matches is better than the use of a single best match because more correct matches will be available to provide support. The additional false matches will correspond to random movements which will not generally form a consistent motion set. The number of random samples required by RANSAC is (Fischler and Bolles, 1981):

$$\Upsilon = 1 - \left(1 - \sigma_m^3\right)^k, \tag{4}$$

where Υ is the probability that a sample containing only correctly matched points (inliers) is chosen, σ_m is an estimate of the proportion of correctly matched points in the dataset when multiple stereo matches are allowed, and k is the number of sample sets taken. The number of elements required in a sample set, in the current case 3, has been substituted into the equation. Equation (4) can be rearranged to give k, the number of samples required:

$$k = \frac{\log(1 - \Upsilon)}{\log\left(1 - \sigma_m^3\right)}$$

Since σ_m is a function of average match multiplicity, M, which will be different for each scene, σ_m should be approximated in practice by σ_1/M, where σ_1 is an estimate of the proportion of correctly matched points in the dataset if each point were allowed only one stereo match. Once the process is underway, σ_1 can be estimated from the fraction of inliers found at the end of stage C during the previous cycle. This gives the modified form:

$$k = \frac{\log(1 - \Upsilon)}{\log(1 - (\sigma_1/M)^3)} \tag{5}$$

In practice, the number of inliers under multiple stereo matching should be greater than the number under single matching, and the number of samples required will be *overestimated*, which is satisfactory. Figure 3 illustrates this for a real example; for maximum stereo multiplicity from 1 to 8, an estimate of the number of samples required from Eq. (5) is compared to the actual number required given the true number of inliers (i_m). Equation (5) can be closely approximated by:

$$\begin{aligned} k &\simeq \vartheta M^3 \\ \vartheta &= \frac{\log(1 - \Upsilon)}{m_0^3 \log\left(1 - \left(\frac{\sigma_1}{m_0}\right)^3\right)} \end{aligned} \tag{6}$$

Where m_0 is an over-large stereo multiplicity value. If m_0 is set to 50, Eq. (6) overestimates the number of random samples required by between 3 and 5, for multiplicity values from 1 to 10 and proportion of outliers from 2 to 90%. Over this range the actual number of samples required varies from 3 to 20000, so the approximation is valid. Equation (6) demonstrates that the number of samples required is proportional to the

cube of the multiplicity, so doubling the multiplicity implies an order of magnitude increase in processing. For this reason it is unwise to allow excessive multiplicity when it is not beneficial.

If, at the end of the RANSAC process, a significant proportion of inliers exist (a similar proportion to that found in the last successful cycle, or a proportion based on σ_m, in the case of the first few cycles), and a minimum number of inliers are present, motion estimation is deemed a success, otherwise stages B and C are repeated with a wider search area in the hope of finding a sufficiently supported motion.

Ouliers (points with no correct stereo match, or an incorrect temporal match) will always arise in practice, even when the correct motion is found. A third set of points will also exist: points which leave the scene under the motion, and therefore cannot provide support for it. Intuitively these are more supportive than outliers but less so than inliers. It was found however that if they are given value in evaluating the support for a motion under RANSAC, and the number of inliers is very small, motions resulting in *all* points leaving the scene can become more favourable than the true motion.

5. Structure from Motion: Structure Calculation and Verification

This section describes stage D of the structure from motion process, which determines a consistent set of reconstructed feature points given the motion calculated by stages A to C. The stage attempts to match points over all four images shown in Fig. 2: the previous stereo pair (numbered 1 and 2), and the latest pair (numbered 3 and 4). At stage A an attempt has been made to define stereo matches between points in images 1 and 2, in many cases resulting in multiple matches for each point. In this stage, the reliable matches among these are identified, using the additional information content of the new image pair and the estimate of camera motion between the pairs. Stereo matches SM_t are also found as a result, for use in the next cycle.

5.1. Match Verification

Each of the multiple hypothesised stereo matches made at stage A has associated with it: a point position in image 1; a match disparity between image 1 and 2; and a match correlation, α_{12}. From the position and disparity, the point's location in space at the earlier time step can be determined, the motion estimate used to transform this location to the latest time step, and the new location projected to give predicted position in images 3 and 4. The point should appear close to these positions if the stereo match hypothesis is correct. A search for the point is conducted very near to the predictions (how close is defined by a *maximum prediction deviation* parameter), not necessarily at the location of detected corners, and the best correlation found to image 3 (α_{13}) and image 4 (α_{14}) is recorded.

A correlation threshold, defining the amount of image deformation allowed between views of a point, is applied to α_{13} and α_{14}, in order to classify the attempted matches as a success or failure. The same threshold as was applied to the image 1-image 2 match correlation (α_{12}) is used, despite the fact that, through greater camera displacement, α_{14} will generally be lower than α_{12} or α_{13}. If the matches to both new images are sufficiently strong, a match strength, Ψ, is calculated to rate the stereo hypothesis, based on the similarity of the matched image points and their accordance with the rigid motion:

$$\Psi = \frac{\sqrt[3]{\alpha_{12}\ \alpha_{13}\ \alpha_{14}}}{d_L + d_R + \varepsilon_d} \tag{7}$$

The values d_L and d_R are the distances between the predicted and matched point position in image 3 and 4 respectively. The *prediction error influence* parameter, ε_d, balances the importance placed on visual similarity (match correlation) over rigidity (deviation of matched points from their predicted positions) in the reconstruction. The value chosen is a function of camera calibration accuracy and the effectiveness of the correlation measure used. The value of Ψ given by Eq. (7) is inversely proportional to deviation from motion-predicted image position and inversely proportional to pixel intensity difference. The numerator is a product to limit the effect of excessively high correlation values. The denominator is a sum because one of the values may be zero. If the value of Ψ is above a *match strength threshold*, the point becomes *verified*. The threshold chosen should be a function of the error generated in the processes used, i.e. of error in the α and d values of correct hypotheses.

Figure 4 (described fully in Section 5.5) shows that with Ψ defined by Eq. (7), when all processed corners are above a certain corner strength (100), no corners are falsely verified. The same result can be achieved with a lower corner strength threshold if the threshold on Ψ is increased. As a result, a more appropriate

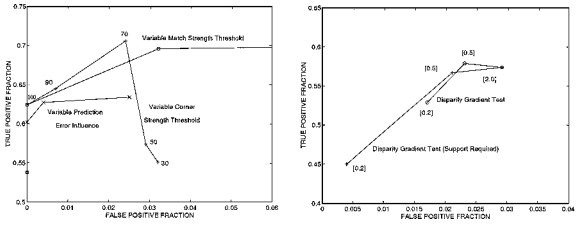


Figure 4. ROC curves showing the effect of various program parameters.

match strength threshold is one which places a higher requirement on weaker corners. The form of Ψ given by Eq. (8) is appropriate, and was used successfully for some experiments. The modifier is limited to values between 0 and 0.16 (a typical value for the match strength threshold is 0.5).

$$\Psi = \frac{\sqrt[3]{\alpha_{12}\,\alpha_{13}\,\alpha_{14}}}{d_L + d_R + \varepsilon_d} + \left(0.2 - \frac{20}{\text{Corner Strength}}\right) \quad (8)$$

5.2. *Further Elimination of Multiple Hypotheses*

Many of the multiple hypotheses for a point will have been eliminated by the match strength threshold during the first part of the verification stage, but for some points, more than one verified match may still exist. Two actions, described below, are performed in an attempt to isolate a single match for these. If, after these tests, a point remains multiply verified it is not used in the reconstruction.

Supremacy of a Particular Match. Where multiple verified matches exist for a point, the *supremacy* of the strongest hypothesis is calculated by:

$$\text{supremacy} = \Psi_1 / \Psi_2 \quad (9)$$

where Ψ_1 is the match strength of the strongest hypothesis for a point and Ψ_2 is that of the second strongest hypothesis for the same point. If the value of supremacy is greater than a *supremacy threshold*, the leading hypothesis is chosen and all others disregarded. If the threshold is set too low, hypotheses may be favoured at random because they carry favourable noise; if it is set too high it will not be recognised that, despite the noise in all Ψ values, a hypothesis is superior. The threshold was set to 5 for all the experiments reported here.

Proximity Test. The proximity test is related to resolvability of sampled and distorted scene texture. Sometimes when trying to match two features, because of pixelation and viewpoint induced deformation, there may be more than one place close to the true feature match position for which correlations are very similar and are close to the maximum. This is particularly likely in natural scenes containing complicated repeated patterns, such as the leaves on a hedge.

After calculating match strength values, Ψ, if the strongest two stereo match hypotheses for a point correspond to right image (image 2) positions which are separated by only a few pixels (by less than a *proximity threshold*) the hypothesis with the highest value of Ψ is (arbitrarily) taken since the reconstruction error in choosing this over the other is likely to be small. This test allows extra points to be reconstructed, but with a slight loss in the accuracy of their reconstructed position. Since the addition of reconstructed points is a principal aim of reconstruction (Section 3), and the accuracy loss is not critical, the proximity test is included.

The maximum acceptable separation is that resulting in the maximum permissible error in reconstructed position. For the experiments described in Section 7 the threshold is set to three pixels, which corresponds to a reconstruction error of up to 24 mm in the x-direction for an obstacle at 5 m.

5.3. *Disparity Gradient Check*

A global continuity problem occasionally arises; sometimes the order of nearby verified points is different in the left and right images (images 3 and 4). This is particularly problematic if it occurs in the image x-direction. Therefore, the *disparity gradient* test is used after verification in order to enforce some local consistency. The definition of disparity gradient, and its use is described fully in Pollard et al. (1991b).

The disparity gradient check is applied to the image 3 to image 4 stereo-matches for each singly verified point. The disparity gradient support and disparity gradient failure for the point are calculated. If more local points reject the match than support it, the point's reconstruction is rejected. For the work described here, the disparity gradient threshold (Pollard et al., 1991a) is set to 0.5 and applied in both image directions, which results in a general rejector of surfaces whose tangent is nearly parallel to the camera optical axes. A small amount of correct structure may be rejected under these

conditions but the density of features should ensure that some correct structure remains, enabling obstacle avoidance.

In densely featured image areas, the disparity gradient check identifies and rejects incorrectly matched points. Single point objects which occur against the background of a deeper surface will also be rejected by the disparity gradient check. The check will have no effect on points which are isolated in the images, since rejection and support for these will be zero. If, in addition, it is insisted that a hypothesis has *some* support, this type of isolated point will also be rejected. The practical effect of the disparity gradient test is shown in Fig. 4(b) (described more fully in Section 5.5), for the case where *some* support is, and is not insisted upon (disparity gradient threshold values are shown on the two curves). When the threshold is set to 0.5 the test is seen to be successful in removing falsely reconstructed points without significantly affecting the number of correctly reconstructed points.

5.4. *Transfer of the Verification Stage Results to the Next Cycle*

If, at the end of the verification stage tests, a single verified hypothesis remains for a point, the stereo match in images 3 and 4 is recorded for use in the next cycle (becoming a SM_{t-1} match). In Fig. 2, the match between p_L, p_{R2}, p'_L, and p'_R has been confirmed at the expense of the other hypothesised matches, $p_L - p_{R1}$ and $p_L - p_{R3}$. The match $q_L - q'_{L1}$ is found to be inconsistent, and q_L matched to q'_{L2} instead, as well as to q'_R.

Once a set of high confidence matches have been produced by the verification stage it is possible (although this was not done) to recalculate motion from the verified points only, using either least squares (which will now not be affected by gross outliers), or RANSAC with a tighter consensus threshold.

5.5. *The Effect of Program Parameters*

The program uses various parameters in stage D, the effects of which are discused briefly here. Parameter effect is evaluated by looking at the corner matches between images 3 and 4 at the end of stage D. This allows points in image 1 to be classified into one of five groups, defined by Table 1. Point are considered reconstructible if they appear in all four images 1–4.

Table 1. Classification of reconstruction results by hand.

	Reconstructible	Not reconstructible
Reconstructed correctly	*True positive,* (TP)	—
Reconstructed incorrectly	*False positive(a),* (FP_a)	*False positive(b),* (FP_b)
Not reconstructed	*False negative,* (FN)	*True negative,* (TN)

The success of the verification stage, and of the reconstruction as a whole, can be measured in relation to the objectives defined in Section 3, namely that no points should be reconstructed incorrectly (no *false positives*) and that as many points as possible are reconstructed correctly (many *true positives*). False negative and true negative numbers are also significant as they allow presentation of the results obtained for a particular scene against the result of a perfect structure recovery method, or a perfectly bad method.

Figure 4 compares the effect of altering various program parameters (while the others parameters are kept constant) on the primary objectives via Receiver Operating Characteristic (ROC) curves (Highnam and Brady, 1999), which give a standardised indication of relative performance. The ROC plot is of *True Positive Fraction* (*TPF*) against *False Positive Fraction* (*FPF*), where these are defined for point reconstruction as:

$$TPF = \frac{TP}{TP + FP_a + FN}$$
$$FPF = \frac{FP}{TP + FP + TN + FN}, \tag{10}$$

with $FP = FP_a + FP_b$. These equations correspond to the ratio of the figure obtained to its theoretical maximum. The curves are incomplete because data containing lots of false positives was not taken.

The curves all intersect the zero false positive fraction axis at a critical point, and continue towards zero true positives. The parameters should ideally be set at these critical points. Since the parameters are interrelated however it is difficult to do this in practice for each new scene. All parameters offer more true positives at the expense of additional false positives although some changes come at the expense of speed (a secondary objective). All parameters are critical, but those which can be altered to most benefit over the particular range plotted are the corner strength threshold and disparity gradient limit.

6. Structure Tracking

The process described in the previous section provides point positions relative to the cameras from one instant to the next. This section describes the nature of uncertainty in this measurement process, and the use of this uncertainty to track the point positions in a fixed coordinate frame.

Uncertainty in image feature location and in the camera parameters results in error in the estimates of motion and of structure relative to the camera, which in turn produce error in the estimate of structure in a static coordinate frame. The effect and correction of camera calibration errors are not discussed further here. Discussion is limited to the effect of uncertain image position.

6.1. Image Feature Localisation Errors

The true positions of image corner features cannot be determined perfectly because of limitations in the imaging and feature location processes. The resulting errors are mostly non-systematic and include the following:

- Sensing in a CCD camera is performed by an array of sensor elements mounted on a planar surface. The elements do not fully cover the imaging plane and the resulting gaps in the sensor limit the accuracy at which feature position on it can be determined.
- Noise will be present in the intensity values recorded by the cameras. This may produce error in the apparent locations of image features.
- Because a digital image is a matrix of samples of, rather than a continuous representation of, a view, it is difficult to precisely locate positions on it. Common approaches are either to take the position of the nearest pixel as the feature position or to interpolate position from the pixel 'cornerness' values. Interpolated measurements tend however to be heavily skewed towards the step measurement which would be seen without the interpolation (Molton, 1999), and, as a result, position uncertainty will be similar to its non-interpolated value.
- Many corner and line detectors use Gaussian smoothing as a preprocessing step to reduce noise on the image intensity values. It was shown by Wang and Brady (1995) however, that this moves the position of a corner as a linear function of the value of σ, creating significant uncertainty.

The overall effect of these uncertainties depends on the camera and corner detector used. In some cases, for example Gaussian smoothing, the same error will occur repeatedly, so stereo disparity, for example, will be unaffected. It is difficult to isolate and quantify the different components of image position uncertainty. For this reason, the combination of these uncertainties cannot be accurately described numerically. For the experiments described in Section 7, image position is approximated to a Gaussian distribution with a standard deviation of one pixel.

6.2. The Effect of Image Position Error on Reconstruction Error

The reconstruction process uses stereo vision to produce an estimate of the positions in space of points. This position is represented by (X, Y, Z), where X is horizontal displacement, Y is vertical displacement, and Z is depth. However, the image positions upon which this structure estimate is based contain an error (e_{xL}, e_{yL}) in the left image, and an error (e_{xR}, e_{yR}) in the right.

If the errors, e_{xL}, e_{yL}, e_{xR}, and e_{yR} arise from Gaussian distributions of zero mean, and variance σ_i^2, the probability distributions of δX, δY, and δZ (the error in the reconstructed positions) can be shown to be given by Eqs. (11–13) respectively (Molton, 1999). The first and second are valid if point depth is not too large—in the case of the set-up used here, if Z is less than 27 m. Since structure further away than 10 m is unlikely to be hazardous to the partially sighted wearer of the camera rig, the equations are useful in practice.

$$f_{\delta X} = \frac{Bf}{Z\sigma\sqrt{2}\tilde{X}\sqrt{2\pi}} \frac{2\tilde{X}^2(X-B) - B\delta X(X-\delta X)}{2\tilde{X}^2(X-B+\delta X)} \times e^{-\frac{1}{2}\left(\frac{Bf\delta X}{\sqrt{2}Z\sigma_i\tilde{X}}\right)^2} \quad ; \; X < B/2 \tag{11}$$

$$f_{\delta Y} = \frac{Bf}{\sqrt{2}\tilde{Y}Z\sigma_i\sqrt{2\pi}} \frac{\tilde{Y}^2 - \Delta Y\bar{Y}}{\tilde{Y}^2} e^{-\frac{1}{2}\left(\frac{Bf\delta Y}{\sqrt{2}\tilde{Y}Z\sigma_i}\right)^2} \tag{12}$$

$$f_{\delta Z} = \frac{Bf}{\sqrt{2}(Z-\delta Z)^2\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Bf\delta Z}{\sqrt{2}(Z-\delta Z)Z\sigma_i}\right)^2}, \tag{13}$$

where $\tilde{X} = \sqrt{2}^{-1}\sqrt{(X-\delta X)^2 + ((X-B)+\delta X)^2}$, $\bar{Y} = Y + \delta Y$, $\tilde{Y} = \sqrt{\bar{Y}^2 + \frac{1}{4}B^2}$, B is camera separation, and f is camera focal length. A very similar expression for $f_{\delta X}$ is obtained when $X > B/2$. The factor $\sqrt{2}$ in the definition of $\tilde{X}$ is chosen so that if the point is midway between the two cameras, or is far out to the

side, the value of $\tilde{X}$ becomes close to X. In practice it is likely that δX, δY, and δZ will be small in comparison to $\tilde{X}$, $\tilde{Y}$, and Z respectively. In this case the measurement error distributions approximate to:

$$f_{\delta X} = \frac{Bf}{\sqrt{2}\tilde{X}Z\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Bf\delta X}{\sqrt{2}\tilde{X}Z\sigma_i}\right)^2} \quad (14)$$

$$f_{\delta Y} = \frac{Bf}{\sqrt{2}\tilde{Y}Z\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Bf\delta Y}{\sqrt{2}\tilde{Y}Z\sigma_i}\right)^2} \quad (15)$$

$$f_{\delta Z} = \frac{Bf}{\sqrt{2}Z^2\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Bf\delta Z}{\sqrt{2}Z^2\sigma_i}\right)^2} \quad (16)$$

The approximation for $f_{\delta X}$ is now valid for all values of X. These equations have become simpler because the skew present in the more accurate uncertainty measure of Eqs. (11) to (13) is removed. The simplified equations suggest that the measurement error distributions can be approximated to Gaussian distributions with zero mean and standard deviations given by:

$$\sigma_{\delta X} = \frac{\sqrt{2}\tilde{X}Z\sigma_i}{Bf} \quad \sigma_{\delta Y} = \frac{\sqrt{2}\tilde{Y}Z\sigma_i}{Bf} \quad \sigma_{\delta Z} = \frac{\sqrt{2}ZZ\sigma_i}{Bf}, \quad (17)$$

where $\tilde{X} \simeq \sqrt{2}^{-1}\sqrt{X^2 + (X-B)^2}$, and $\tilde{Y} \simeq \sqrt{Y^2 + \frac{1}{4}B^2}$. The three error functions are very similar, the only difference being in the representative measure used ($\tilde{X}$, $\tilde{Y}$, or Z). This measure in the X direction is the root mean square X-direction distance to the two cameras, divided by the square root of two—which becomes close to X if the point is midway between the two cameras, or if X is large in comparison to camera separation. The representative distance for Y is Y when comparatively large, but tends to half of the camera separation as Y goes to zero. The representative distance for Z-direction error is simply Z. Matthies uses a linear analysis to get to a similar approximation (Matthies and Shafer, 1987).

In practice, it is likely that image errors will be correlated to some degree for image points produced by the same position in the world. Each image point error might be the sum of an image-wide error distribution and a non-zero-mean error distribution specific to image points derived from the same world entity.

6.3. *A Stationary Reference Coordinate Frame*

The output of stereo triangulation is position relative to the left camera. Points tracked in this frame will follow the motion of the cameras and as a result tracking would be extremely difficult. For purposes of experimental validation, tracking of feature position for a static scene is far simpler when it is done in a coordinate frame which is fixed in space. The conversion of point position from a camera to a fixed world coordinate frame is determined by the position and orientation of the camera in the world frame. This is defined at iteration k by a rotation, $\mathbf{R}_k$, of the camera from the world coordinate system, followed by a translation, $\mathbf{t}_k$, from the world origin relative to the new orientation. The camera position is updated from the relative movement of the camera between state $k-1$ and state k, defined by a rotation $\mathbf{R}$ and translation $\mathbf{t}$. The update is then:

$$\mathbf{R}_k = \mathbf{R}\mathbf{R}_{k-1} \qquad \mathbf{t}_k = \mathbf{R}^{-1}\mathbf{t}_{k-1} + \mathbf{t} \quad (18)$$

In implementation, camera position is defined by a translation vector and rotation quaternion. Rotation quaternions are used because a combination of rotation matrices can easily become degenerate through numerical error. Point positions, and their uncertainty ellipsoids, can then be transferred from the camera frame ($\mathbf{p}_c$, $\mathbf{u}_c$) to the world frame ($\mathbf{p}_w$, $\mathbf{u}_w$) by:

$$\mathbf{p}_w = \mathbf{R}_k(\mathbf{p}_c + \mathbf{t}_k) \qquad \mathbf{u}_w = \mathbf{R}_k\mathbf{u}_c \quad (19)$$

Uncertainty in the motion estimate should also be added to the point position uncertainties, and this is done by various authors. It was found that error in the motion estimate is proportional to image feature location error (Molton, 1999); but the motion uncertainty was not quantified fully and so is not included in the process.

6.4. *Structure Filtering*

Given observations of point position in the static world frame, a Kalman Filter is used to update a position estimate for each point. The use of a Kalman Filter is justified by the statement in Section 6.2 that if feature location error is Gaussian, position measurement uncertainty is close to Gaussian. Since each component of position can be measured independently from one of the observations, a different Kalman Filter is used to track each component. A validation gate is also used to remove points for which the reconstruction is unstable. This results in a simple and fast method, although it is suboptimal, as uncertainty in the three directions components will be correlated in practice.

7. Structure Recovery Experiments

This section describes some results from using the structure and motion process described in this paper, and discusses some practical issues arising from them. Three scenes, with different characteristics, were used in the experiments: an indoor scene, a scene outside a house, and a street scene. In each case, a stereo image sequence, captured using the ASMONC backpack, was used to make a 3D point map reconstruction. Because image capture speed was limited by the portable hardware used (to about 1 Hz), there are considerable jumps (of up to 200 pixels) in the positions of features between successive image pairs. As the C40 processor proved to be too slow to process in real time, the images were processed off-line. However, it is already clear that, performance approaching real time could be achieved with optimised code, smaller images, and a faster processor (Molton, 1999).

7.1. An Indoor Scene

The reconstruction of a laboratory scene was made from a sequence of 49 image pairs. Some consecutive images from the left-half of the sequence are shown in Fig. 5, and two of the images are shown in more detail in Fig. 6.

The world map reconstructed by the program is shown in Fig. 7. The image is a top-down view of the reconstructed scene, where each square represents 1 m. Each grey blob is an output point from the program. The shade of grey indicates height above the ground, lighter being higher. The observer's start position, on which the world coordinate system is defined, is at the bottom of the map image, facing up. The actual layout of the room, and approximate viewing direction for the images of Fig. 6, are superimposed onto the computer generated map. Points reconstructed to a position which is at least 10 cm from any other point are omitted by the program.

The reconstruction is quite reliable, no obstacles are missed, and no false structure blocks movement into unobstructed parts of the room. False structure is generated behind the mirror, but this is an acceptable consequence of observation by computer vision. The laboratory scene provides a baseline for evaluating the algorithm. It is full of planar surfaces and has relatively little texture. The reconstructed scene that results from

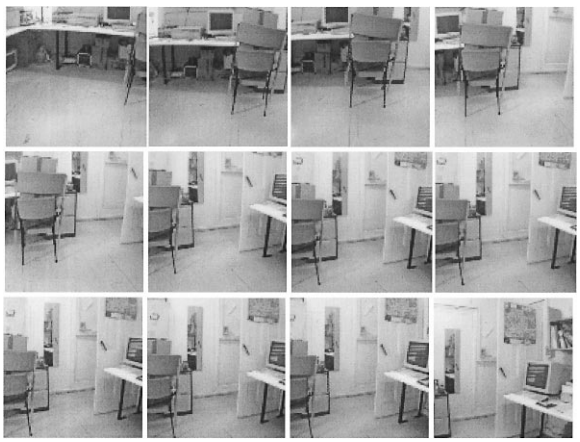


Figure 5. Consecutive left images from the indoor image sequence.

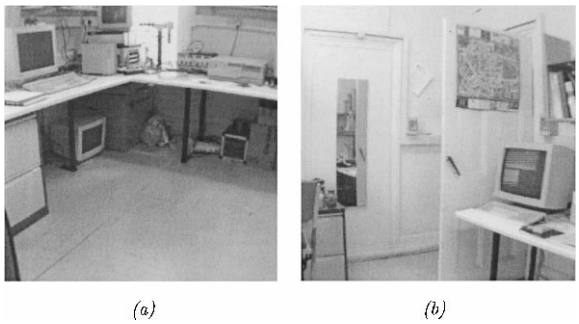


Figure 6. Images from the indoor scene.

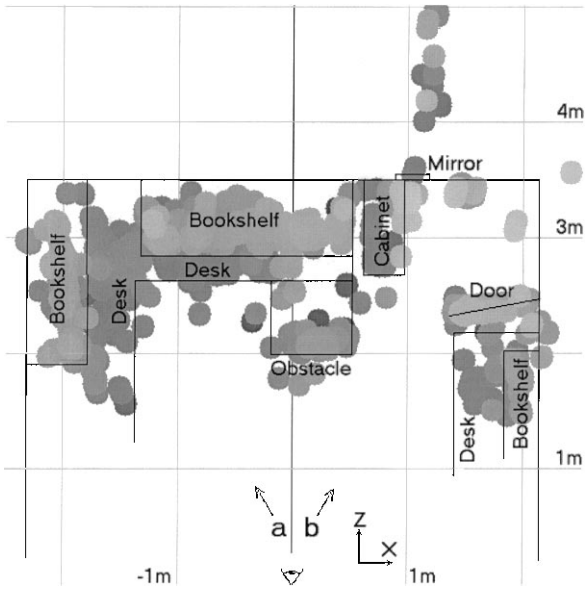


Figure 7. Point reconstruction of the indoor scene.

the algorithm is sufficiently detailed and accurate to allow collision-free walking by a partially-sighted wearer of the ASMONC backpack.

For this scene, the matches which are verified at stage D almost all come from the strongest of the multiple stereo matches made at stage A (about 98%). Not much use is made of stereo match multiplicity because most of the features in the scene are well defined corners, for which it is relatively easy to identify the true feature match by looking at visual similarity alone.

7.1.1. Position Prediction Error. Section 6 of this paper described how error in camera calibration generates error in estimates of motion and structure position. However, the combined effect of the errors is small; predictions made in the verification stage are accurate. Typical average errors (d_L and d_R values) for verified hypotheses are about 1.3 pixels for the left image and about 2.4 for the right. The left image side is favoured in the process because the motion calculation uses matches between consecutive images on this side, and between the left and right images.

During the motion estimation stage, any consistent error in triangulation (from camera calibration) will result in an opposing error in the motion estimate, such that the motion estimate gains maximum support. When position predictions are made in the verification stage, the errors again oppose, resulting in good position prediction. Those components of the error in calibration which cannot be accommodated in the motion contribute to the error in the predictions.

7.2. A Scene Outside a House

Some of the left images from a 36-pair sequence of the outside of a house are shown in Fig. 8. An earlier image from the sequence is shown at a large scale in Fig. 10. This time the result of interlacing, which was not corrected for, can be made out in some images, for example, on the drain pipe of Fig. 10. This is a practical effect which could fairly easily be removed.

The world map produced by the program is shown in Fig. 9. A few points which lie on the ground are seen, shown as the darkest dots. It proved quite difficult to observe a large rosebush at the start of the sequence; it is represented by a handful of points in the reconstruction despite the large number of point features it produces in the images.

For this scene a stereo match multiplicity up to 4 was allowed. Verified points came from the first to fourth

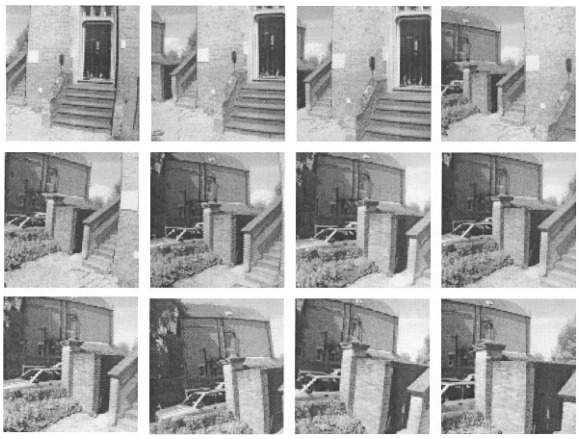


Figure 8. Consecutive left images from the house image sequence.

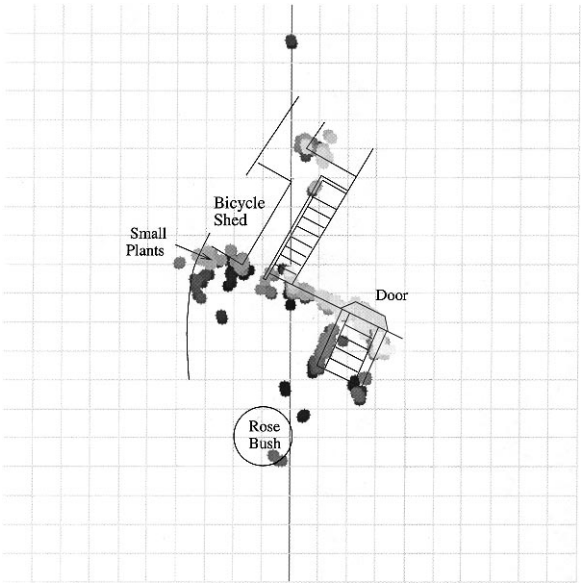


Figure 9. Point reconstruction of the scene outside a house.

best match as follows: 93.7139%, 4.1794%, 1.4951%, 0.6116%. The values demonstrate that the multiplicity is made more use of for this outdoor scene, which contains some features which are difficult to match, i.e. regions of plant material. Again, sufficient information is available to navigate a partially sighted person through the scene.

7.2.1. Distant Features. For this scene, objects which are distant at the start of the sequence later become close. This required additional steps to be taken. Firstly, points in view, but further than a distance threshold

away (8 m) were used to calculate motion but were not made part of the world map. Secondly, if a point was reconstructed earlier and should now be in view, but is not, the point is removed from the map. This prevents points being reconstructed twice: once while far away, and again when near, if the two observations cannot be associated with each other. A similar problem was encountered by Moravec (1983).

7.2.2. Corner Number. Outdoor scenes such as this show substantial variation in the number of points detected, depending on the nature of what is seen. Areas of vegetation produce many more corner features than do artificial objects such as the door or brickwork. When a constant corner detection threshold is used the number of corners can vary by up to an order of magnitude (for this scene the number varied between 300 and 2000). Control of corner number is critical, as a lack of corners makes motion calculation unreliable, whereas too many makes processing time excessive, in an exponential way, and results in the presence of more weak, unreliable corners. An approximately constant number of corners is desired, so an adaptive threshold must be adopted.

It is not sufficient to apply a threshold to corner number over the whole image when different types of object are present in the scene. For example, Fig. 10(a) shows corner detection with a single threshold in an image from the outdoor scene. Valid features on the right of the image are cut out by the artificially high threshold imposed by the abundance of features on the rose bush to the left of the image. If, following capture of image 10(a) the camera panned to the right by a large amount, there may not be enough corners matched through time to calculate motion.

Ideally, features which are locally exceptional should be extracted, regardless of their variation in



Figure 10. Corner detection: (a) Single threshold; (b) Threshold for each image quadrant.

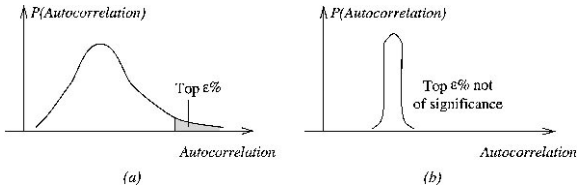


Figure 11. Probability density functions of local image autocorrelation.

comparison to the scene as a whole. The most varied values of image correlation, which varies locally, are desired; i.e. the highest ε% from the probability density function of the autocorrelation values in a local area is sought. Ordinarily the probability density function will be bell shaped, as in Fig. 11(a); if it is tower-shaped, as in Fig. 11(b), then no features are locally exceptional, so no features should be extracted. This will occur in untextured regions.

In practice, more even corner distribution was achieved by dividing the image into areas, and taking the strongest n corners from each area. The approach adopted for this scene, was to split the image into four quadrants. This was found to be sufficient to overcome the corner density problem present with a single threshold. The result of this is shown on Fig. 10(b), where the overall number of corners is similar to that in Fig. 10(a) but they are more evenly spread.

7.2.3. Limitations of the Camera System's Field of View. Another potential problem is that the system can only see things in the direction in which the cameras face. This may be inadequate if the camera platform turns a sharp corner into an unseen area or if something moves into view from the side. These are intrinsic limitations of the field of view provided by the camera setup.

The problem is exacerbated by the use of parallel camera stereo, because the parts of the scene which are present in both cameras can be limited. An example is seen in the house scene when the user walks down the side of the house; each camera sees a side of the alley; but neither side is seen by both, so the stereo system is essentially blind to the alley sides. Verging the cameras should reduce the possibility of this. Another problem can occur if a person walks next to a wall, then stops and faces the wall; the cameras may be too close for stereo matching, and the presence of a wall or passage could not then be said with certainty. To remedy this, the person carrying the camera platform



Figure 12. Some images from a street scene.

should be encouraged to move in such a way as to avoid the problem, if the possibility of this can be anticipated by the system.

7.3. A Street Scene

A third sequence was used, shot from a pavement next to a road. Some sample (non-consecutive) images from the sequence are shown in Fig. 12. Reconstruction was difficult for this scene because a large amount of vegetation was present. This type of scenery is difficult to process in the presence of large feature position jumps, because it is very difficult to match. Distortion of the foliage texture, relative to its characteristic size is so marked that it is often impossible to find the correct match even by hand. Support for the camera motion, coming predominantly from man-made objects in the remainder of the scene, is as low as 20%. This is just sufficient to avoid a break down in the structure and motion recovery process.

Since reliable match information is so scarce, it is again important not to let the plant material dominate the detected corners. For this scene the image was broken down into 16 areas, to further increase the evenness of corner detection. Ideally those regions which are plant material (random, heavily featured texture) should be recognised, and points arising from them not used in the point matching system. This would dramatically reduce the number of outliers in motion estimation from points for outdoor scenes of this type.

8. Conclusions and Further Work

This paper illustrates a stereo vision implementation designed to calculate motion from a limited set of matches, for maximal efficiency, and thereafter to fully exploit the redundancy of information in stereo vision to match points reliably. The design is achieved through multiple stereo matches, a motion calculation using three points in three images, and a consistency check which uses image similarity and conformity to the motion estimate.

The implementation detailed demonstrates the feasibility of a stereo vision system providing motion and sparse structure from corner features, for the purpose of obstacle detection. The system has the following characteristics:

- The system works under *any* movement of the cameras. No model of the motion of the cameras was used for the work described here. A human walker gait model was implemented (Molton and Brady, accepted for publication) and its use would greatly improve efficiency.
- The system works for scenes dominated by *structural* corners, or *textural* corners which are not part of texture which is unstable (changes topologically) under moderate changes in viewpoint. In practice this means that the system is suitable for indoor scenes and urban outdoor scenes, but may break down when the scene comprises mostly plant material, or when features do not exist (for example, on a smooth tarmacked path).
- Currently, the system is not real-time, due to hardware limitations. The system is designed in such a way such that it could be used in real time given better hardware.

8.1. System Improvements

The method of structure recovery described suffers from the effects of unknown camera calibration error. The accuracy and robustness of the process might be significantly improved by using the redundancy present in stereo vision to provide partial dynamic calibration.

The reconstruction produced by the method is reliable but incomplete for two reasons: some parts of the images used cannot be processed reliably when described as a set of point features (for example vegetation); and the feature based method is not continuous, that is, does not attempt to reconstruct every image pixel. The first of these shortcomings could be lessened through the use of less common feature types such as lines or deformable areas. The latter could be characterised by their perimeter curve, and the use of colour or statistical pattern descriptors to describe their interior. Colour might also be considered as an aid to point matching. Both shortcomings might be remedied under an infinitesimal motion method, for example, incorporating analysis of optic flow. This could not be attempted because of hardware performance limitations.

A nonlinear uncertainty analysis for stereo vision is described but not made full use of here. The accuracy of

the method could be improved significantly by making more use of uncertainty.

The structure from motion method described can be extended to cope with non-static, rigid objects within a scene. This basically requires the motion estimation to be carried out on the outliers remaining after the first RANSAC motion estimating stage. This extension, and the tracking and avoidance of moving scene objects, is discussed theoretically in Molton (1999).

8.2. Potential of the System

The system described might be shrunk for the ASMONC application. The cameras could in principle be connected rigidly via something small, such as a pair of glasses, and electronics carried in a pocket. The structure recovery methods described are also applicable to obstacle avoidance for autonomous machines moving in an unconstrained or complex way, for example, a flying, or walking robot. Some of the practical methods could be applied to other feature-based or area-correlation stereo vision sequence analysis systems.

Acknowledgment

Financial Support for the ASMONC project came from European Union's TIDE initiative. Nicholas Molton's research is funded by the EPSRC.

Notes

1. If M translates both cameras, fundamental matrices for $\underline{LIM}$ and $\underline{RIM}$ are determinable. If M also contains a rotation, M can be determined from the fundamental matrices.
2. In practice isolated points in the reconstruction are removed, so a very small number of erroneous points could be tolerated.

References

Bouguet, J. and Perona, P. 1995. Visual navigation using a single camera. In *Proc. 5th International Conference on Computer Vision, Boston MA*, Los Alamitos, CA, USA. IEEE Computer Society Press, pp. 645–652.

Ferrari, F., Grosso, E., Sandini, G., and Magrassi, M. 1990. A stereo vision system for real time obstacle avoidance in unknown environment. In *Proceedings of IEEE International Workshop on Intelligent Robots and Systems IROS '90*, pp. 703–708.

Fischler, M.A. and Bolles, R.C. 1981. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, 24:381–395.

Harris, C. 1987. Determination of ego-motion from matched points. In *Proc. 3rd Alvey Vision Conference*, Cambridge, UK, pp. 189–192.

Harris, C. and Pike, J.M. 1987. 3d positional integration from image sequences. In *Proc. 3rd Alvey Vision Conference*, Cambridge, UK, pp. 233–236.

Harris, C. and Stephens, M. 1988. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, Manchester, UK, pp. 147–151.

Highnam, R. and Brady, M. 1999. *Mammographic Image Processing*. Kluwer Academic Publishers.

Konolige, K. 1998. Small vision systems: Hardware and implementation. In *Robotics Research: Proceedings of the 8th International Symposium*, Springer-Verlag, pp. 203–212.

Li, F., Brady, J.M., and Hu, H. 1995. Visual guidance of an AGV. In *Proceedings of ISRR '95*, Germany.

Matthies, L. and Grandjean, P. 1994. Stochastic performance modeling and evaluation of obstacle detectability with imaging range sensors. *IEEE Trans. on Robotics and Automation*, 10(6):783–792.

Matthies, L., Kelly, A., Litwin, T., and Tharp, G. 1996. Obstacle detection for unmanned ground vehicles: A progress report. In *Robotics Research: Proceedings of the 7th International Symposium*, G. Giralt and G. Hirzinger (Eds.). Springer-Verlag, pp. 475–486.

Matthies, S.A. and Shafer, L. 1987. Error modeling in stereo navigation. *IEEE Journal on Robotics and Automation*, 3(3):239–248.

Molton, N.D. 1999. *Computer vision as an aid for the visually impaired*. DPhil Thesis, Dept of Engineering Science, University of Oxford.

Molton, N.D. and Brady, J.M. Accepted for publication. Modelling the motion of a sensor attached to a walking person. *Robotics and Autonomous Systems*.

Molton, N.D., Se, S., Brady, J.M., Lee, D., and Probert, P. 1999. Robotic sensing for the partially sighted. *Robotics and Autonomous Systems*, 26(3):185–201.

Moravec, H.P. 1983. The stanford cart and the cmu rover. *Proceedings of IEEE*, 71(7):872–884.

Murray, D. and Little, J. 1998. Using real-time stereo vision for mobile robot navigation. In *Proceedings IEEE Workshop on Perception for Mobile Agents*, pp. 19–27.

Pollard, S.B., Mayhew, J.E.W., and Frisby, J.P. 1991. Implementation details of the pmf stereo algorithm. In *3D Model Recognition from Stereoscopic Cues*, J.E.W. Mayhew and J.P. Frisby (Eds.). MIT Press, pp. 33–39.

Pollard, S.B., Porrill, J., Mayhew, J.E.W., and Frisby, J.P. 1991. Disparity gradient, Lipschitz continuity, and computing binocular correspondences. In *3D Model Recognition from Stereoscopic Cues*, J.E.W. Mayhew and J.P. Frisby (Eds.). MIT Press, pp. 25–32.

Se, S. 1998. *Computer vision aids for the partially sighted*. DPhil Thesis, Dept of Engineering Science, University of Oxford.

Se, S. and Brady, M. 1998. Stereo vision-based obstacle detection for partially sighted people. In *Proc. Third Asian Conference on Computer Vision ACCV '98*, pp. 152–159.

Stein, G. and Shashua, A. 1998. Direct estimation of motion and extended scene structure from a moving stereo rig. In *Proc. Computer Vision and Pattern Recognition CVPR'98*, Santa Barbara, California, USA, June 1998, IEEE Computer Society Press, pp. 211–218.

Torr, P.H.S. and Murray, D.W. 1997. The development and comparison of robust methods for estimating the fundamental matrix. *Int Journal of Computer Vision*, 24(3):271–300.

Tsai, R.Y. 1987. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4): 323–344.

Wang, H. and Brady, M. 1995. Real-time corner detection algorithm for motion estimation. *Image and Vision Computing*, 13(9):695–703.

Wolfe, W.J., Mathis, D., Sklair, C.W., and Magee, M. 1991. The perspective view of three points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(1):66–73.

Young, G.J. and Chellappa, R. 1990. 3-d motyion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(8):735–759.

Zhang, Z. and Faugeras, O.D. 1992. Estimation of displacements from two 3-d frames obtained from stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(12):1141–1155.