Using a Combined Stereo/Temporal Matcher to Determine Ego-motion

Neil A. Thacker, Ying Zheng and Robert Blackbourn

AI Vision Research Unit, University of Sheffield
Western Bank, Sheffield S10 2TN, UK

A system has been developed for integrating information from sequences of stereo images suitable for use in visual control. The method exploits multiple sources of information to obtain a subset of correctly matched corner features in temporal pairs of stereo images. These have been used to determine the ego-motion of a stereo camera system and to improve position estimates of these features. The algorithm is demonstrated on images of a real scene which would be expected to present stereo or temporal matching algorithms with matching difficulties.

We are constructing a mobile vision platform COMODE with which we intend to address difficulties of visual guidance and path planning. Although information is available from the position of wheels at the front of this vehicle it cannot be relied upon to deliver accurate motion estimates for the moving vehicle. We intend to use stereo vision to compute a more accurate estimate of the motion of the vehicle suitable for closed loop control. For this we have combined several established vision algorithms to construct a system which can provide an estimate of ego-motion while maintaining a 3D point based world model.

A system we have developed uses a corner detector [6,1] to obtain well located features in 3D (but see also [7]). Corners are first matched in stereo to obtain positional information which is used to help temporal matching across pairs of images.

Determination of camera motion has been reported by many authors [1,2,5] and a feature fundamental to the success of these methods is the determination of a correctly matched set of correspondences. There are methods for protecting such methods from outliers caused by a small fraction of incorrect matches, but it is necessary to ensure that this fraction is not exceeded. We describe in this paper a corner matching algorithm which identifies those matches which are likely to be incorrect. A subset of correctly matched points are selected using a reliability heuristic or by exploiting the redundancy in the information available. These data can then be used to determine the ego-motion and for this we combine the transformation estimation suggested by Faugeras [1] with the data weighting method suggested by Kiang [4]. Once the camera motion is determined it is possible to combine positional information of each feature to increase their localisation accuracy.

1. Stereo/Temporal Matching

The corner detector we use is that suggested by Harris and Stephens [3] which calculates an interest operator defined according to an auto-correlation of local patches of the image.

$$M_{uv} = \begin{bmatrix} (\partial I/\partial u)^2 * w & \partial I/\partial u \; \partial I/\partial v * w \\ \partial I/\partial u \; \partial I/\partial v * w & (\partial I/\partial v)^2 * w \end{bmatrix}$$

where u and v are image coordinates and $*w$ implies a convolution with a gaussian image mask. Any function of the eigenvalues α and β of the matrix M will have the property of rotation invariance. What is found is that the trace of the matrix $Tr(M) = \alpha+\beta$ is large where there is an edge in the image and the determinant $Det(M) = \alpha\beta$ is large where there is an edge or a corner. Thus edges are given when either α or β are large and corners can be identified where both are large. Corner strength is defined as

$$C_{uv} = Det(M) - kTr(M)^2$$

Corners are identified as local maxima in corner strength which are fitted to a two dimensional quadratic in order to improve positional accuracy which has been estimated as 0.3 pixels (Thacker & Mayhew [8]).

Image tokens can be matched in some cases using the following heuristics;

(a) restricted search strategies (eg epipolars in the case of stereo).

(b) local image properties (eg image correlation).

(c) uniqueness.

(d) disparity gradient (or smoothness) constraints.

For stereo matching potential matches are sought in a variable epipolar band, with a width determined by the accuracy of stereo calibration. As the corner detector finds local maxima in an auto-

BMVC 1990 doi:10.5244/C.4.23

correlation measure it makes sense to compare possible matches between points on the basis of local image cross correlation. Lists of possible matches are generated, for corners in the left image to the right and right to left, and ordered in terms of the local image correlation measure;

$$M = \int_{-\infty}^{\infty} A^{-2} w_{uv} I_{uv} I'_{uv} \, du \, dv$$

with

$$A = \int_{-\infty}^{\infty} w_{uv} I^2_{uv} \, du \, dv \int_{-\infty}^{\infty} w_{uv} I'^2_{uv} \, du \, dv$$

where w is a gaussian weighting function. This measure varies between 0 and 1 (close to 1 for good agreement), again the assumption has been made that there is little rotation about the viewing axis. This measure is invariant to the scale of the registered image intensity (assuming that no prior knowledge of the lighting conditions and individual camera aperture settings is available). Weak dependence on the absolute image intensity can be reintroduced using an asymmetry cut on the relative corner strength.

$$\frac{|C_1 - C_2|}{C_1 + C_2} > \eta$$

A value of 0.85 is generally chosen for η, this will allow a difference of 12 in relative corner strength or a factor of 1.8 in image intensity.

Only if the absolute value of the correlation measure is high ($M_{\max} > \rho$) is the match accepted and added to the list of possible matches. ρ can be set arbitrarily high to ensure that the underlying images are essentially identical and a value of 0.99 is generally used. We accept that this will inevitably result in some bias in matching ability for front-to-parallel surfaces. Candidate matches are only considered further if they involve the best correlation measure $M_{\max}$ found for that pair of points matched both ways between the left and right images. This algorithm implicitly enforces one to one matching and also eliminates incorrect matches resulting when a feature has only been detected in one image.

Due to the sparseness of corner data in many regions of an image it is difficult to impose a smoothness or disparity gradient constraint. However, it may be possible in future to constrain possible matching using the results from less sparse matching primatives such as edges.

On real images corner detection can be very noisy and setting a generic threshold for corner detection is problematic. Also high frequency textured regions generally give rise to many corners which, on the basis of the above heuristics, are unmatchable, as there are many similar candidate matches for each feature. Thus in real images it is difficult to automate the generation of a reliable set of correspondences, potentially preventing successful ego-motion calculation. What is required is a method of identifying those features which may be unreliably matched.

Unreliable features can be defined as those which have many candidate matches and consequently may be expected to be ambiguously matched. Ambiguous matches can be excluded by selecting matches where neither list of other candidate matches has an entry which is above a value of $M_{\max} - \delta$. The required value of δ is defined by the expected variability of the cross-correlation value for correct matches and can be expected to be relatively constant for all images. δ can be defined so that only very unique matches are accepted as good, a value of 0.005 has been found generally to be sufficient. Such a reliability heuristic reduces the consequence of changing feature detection thresholds on the matching of high frequency features and so allows these thresholds to be lowered. We have successfully used this stereo matching algorithm to provide data for camera calibration (Thacker & Mayhew [8]). If we have temporal match information, a more direct method of selecting reliable matches can be used, as explained below.

Temporal matches are sought using three dimensional positions of corner features combined with odometry information specifying the expected motion of COMODE. Match lists are generated between temporal pairs of images in exactly the same way as for the stereo matcher. The result is a set of possible matching lists for each point in each image to its stereo and temporal counterpart. A subset of correct matches is then selected by checking that the matching between all sets of stereo and temporal images is consistent.

2. Motion Parameter Extraction

After 3D feature point extraction and correspondence processes are completed, motion parameters can be estimated using the corresponding 3D feature points. Consider a set of N feature points extracted from two stereo images taken at time t_1 and t_2, denoted as $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ and $\{\mathbf{p}'_1, \ldots, \mathbf{p}'_N\}$ respectively. The motion between them is defined by the motion equation

$$\mathbf{p}'_i = R\mathbf{p}_i + T + \mathbf{e}_i, \qquad i=1, \ldots, N \qquad (2.1)$$

where R and T are the rotation matrix and translation vector, $\mathbf{e}_i$ is the isotropic error vector associated with the measurements $\mathbf{p}'_i$ and $\mathbf{p}_i$. The obvious way of finding R and T is to use the least squares algorithm to minimise the cost function

$$S = \sum_{i=1}^{N} \mathbf{e}_i^t \mathbf{e}_i \tag{2.2}$$
$$= \sum_{i=1}^{N} (\mathbf{p}'_i - R\mathbf{p}_i - T)^t (\mathbf{p}'_i - R\mathbf{p}_i - T)$$

However, associated with each calculated 3D position, there is always a triangulation uncertainty arising from the limitations imposed by the camera resolution, and this uncertainty grows with distance to the camera. Hence the above cost function is modified by multiplying each term by a weighting factor w_i to yield

$$S = \sum_{i=1}^{N} w_i (\mathbf{p}'_i - R\mathbf{p}_i - T)^t (\mathbf{p}'_i - R\mathbf{p}_i - T) \tag{2.3}$$

There are various ways of calculating the weighting factors. Moravec [6] modelled the position uncertainty as inversely proportional to the distance from the camera. Kiang et al [4] have shown that such an approximation is not adequate as far as the triangulation error is concerned. Instead they proposed modelling the triangulation uncertainty as a 1D model which is determined by the geometric properties of the stereo configuration [Appendix I]. Such error modelling has been adapted here in calculating the weighting factors. Thus the task is to search for R and T which will minimise the cost function (2.3). The solution for this can be obtained [Appendix II] from

$$\min_{R} S = \sum_{i=1}^{N} w_i \| \mathbf{n}'_{i,w} - R\mathbf{n}_{i,w} \|^2 \tag{2.4}$$

and

$$\hat{T} = \mathbf{p}'_w - \hat{R}\mathbf{p}_w \tag{2.5}$$

This optimisation problem, as it is stated, is nonlinear due to the rotation matrix R. Equation (2.4) differs by only the weighting term w_i to the cost function obtained by Faugeras [1] which was solved in a linear manner by reparameterising the rotation as a quaternion. We can still do that here and the optimal quaternion $\mathbf{q}$ is found by solving

$$\min_{\mathbf{q}} \mathbf{q}^t A_w \mathbf{q} \tag{2.6}$$

under the condition $\mathbf{q}^t\mathbf{q} = 1$. A_w is a symmetric positive definite matrix calculated from $\mathbf{n}_{i,w}$, $\mathbf{n}'_{i,w}$ and w_i. The solution $\hat{\mathbf{q}}$ is the eigenvector of unit length of matrix A_w corresponding to the smallest eigenvalue.

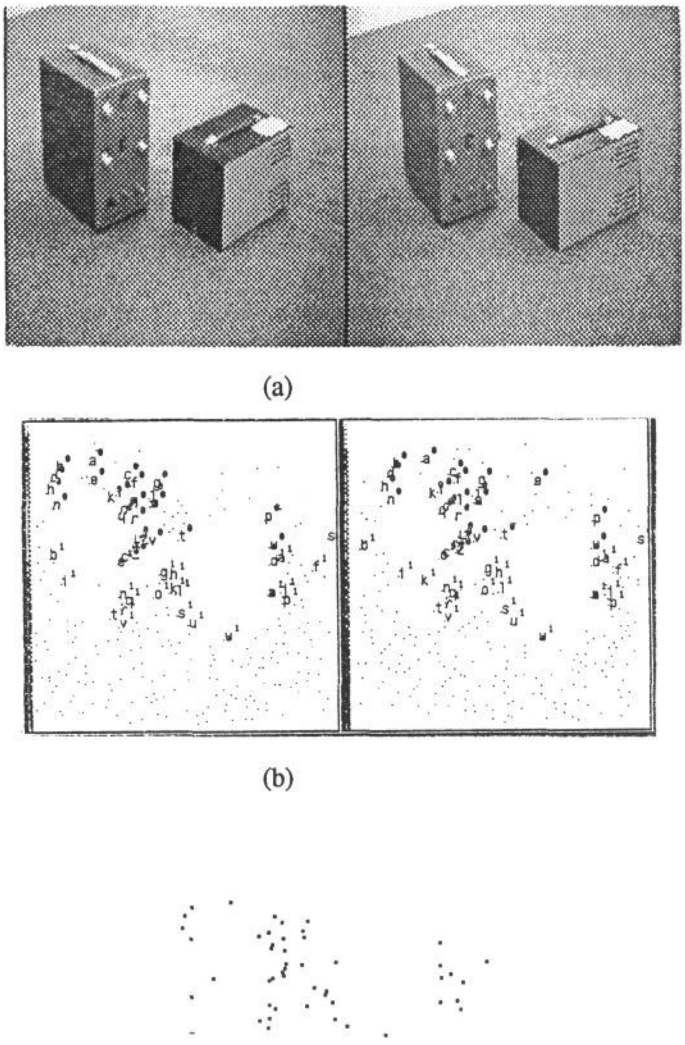
The motion parameters thus obtained are optimal in the sense of least squares provided the measurement errors are independent isotropic Gaussian vectors. Unfortunately in the case of position estimates from stereo this is clearly untrue as depth errors dominate. However, results have shown that by applying weighting in the solution for such data , the estimates of R and T are reliable.

Under the assumption of ego-motion, if the estimated motion parameters are accurate enough, a transformed feature point at time t_1 should lie in the vicinity of its corresponding feature point at t_2. A discrepancy is expected because of measurement errors and parameter estimation errors. However if this discrepancy is unreasonably large for a small subset of the extracted feature points, it is very likely that either the ego-motion assumption is incorrect or the 3D position estimates for those points are subject to anomalous large errors. Such errors may corrupt the estimates of the motion parameters, therefore these points should be excluded from motion estimation process. This can be done by iterative removal of those points which are not consistent with the transformation according to their noise model, which, as mentioned in section 2, is related to the depth of the point.

After eliminating wild points, we have, at time t_2, a set of measured 3D feature points and a corresponding set of the same feature points $\{\mathbf{p}_{1,proj}, \ldots, \mathbf{p}_{N,proj}\}$ projected to t_2 from t_1 using the estimated motion parameters. Better estimates of the true positions of those feature points can now be obtained by combining the above two sets according to their covariances.

3. Results

The results from a stereo match of sets of corners from two stereo images are shown in Figure 1. It should be noticed that many corners have been identified in the image due to the high frequency texture of the carpet. Many of these are not even reliably identified as they approach the threshold for corner detection. Raising this threshold to remove all corner points in the carpet removes nearly all corners in the image. We could not expect to unambiguously match this region of the image using the heuristics identified in part 1. However, the uniqueness heuristic has reduced the set of possible matches to those which would be expected to be reliable. Temporal matches are sought and a set of consistent matches are identified in an eight way matching process between the four images. The set of data is restricted by the efficiency of the corner detection and matching algorithms leaving typically only 50% of the number of matched stereo corners (Figure 2). These few remaining matches can however be used for reliable ego-motion determination using the weighted least squares method.

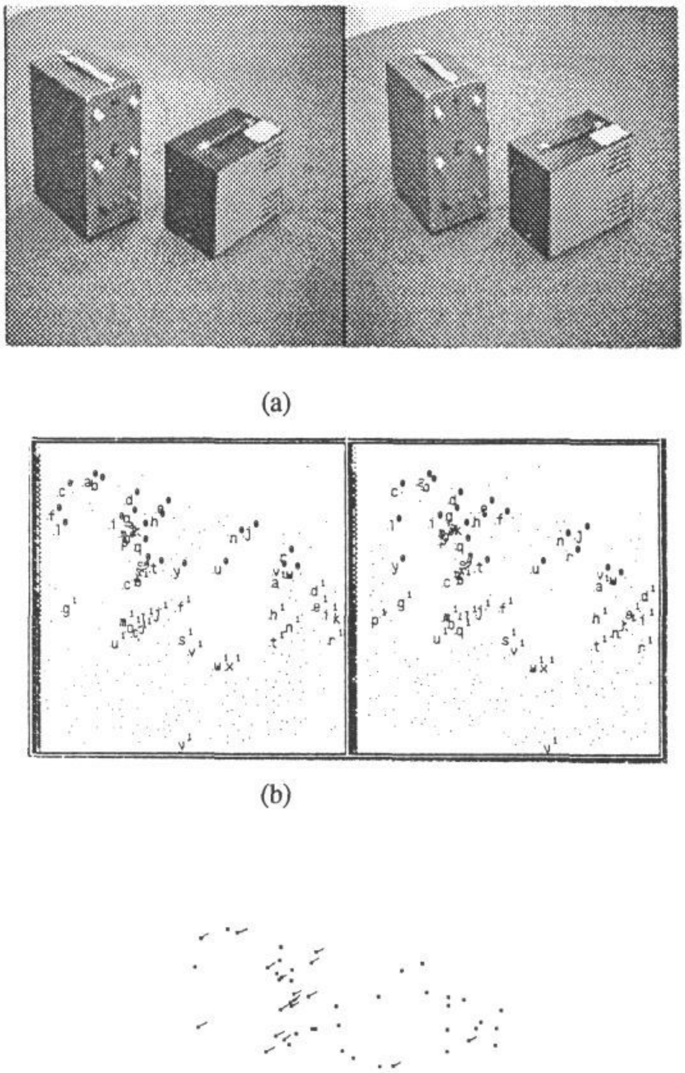


(a)

(b)

(c)

Figure 1. The figure shows the generation of 3D data (c) from raw image (a)through corner detection and matching (b). Notice the large number of noisy corners detected, particularly on the carpet. These have not been matched due to the reliability constraint, the corners that have been matched involve only a few false matches.



(a)

(b)

(c)

Figure 2. This figure shows raw data (a), stereo matched corners (b) and finally 3D data with the reconstructed transformation obtained from temporal matches (line segments)(c). These temporal matches are found to be very reliable for ego-motion calculation.

4. Conclusion

We have descibed a matching algorithm for stereo and temporal corner matching. This algorithm uses reliability constraints to identify those features which may be used for ego-motion determination. The algorithm is robust and can be used to obtain correspondances on arbitrary sets of images without undue parameter tuning. The method is suitable for between-frame estimates of transformation and therefore useful for maintaining temporal coherence across sequences of scenes and for closing the motion control loop with visual feedback. The limit of applicability of this system would be due to the extent to which the rigidity assumption underlying this method was justified in the real world. Reliable motion estimates over long time scales may require knowledge of which objects in the world may be expected to be stationary. In this respect beacon tracking [7] has particular advantages and we intend to use the work described here in a comparative study, of the relative accuracy and merits of these methods, with a view towards system integration.

Acknowledgements

Thanks are due to Dr. Stephen Pollard, Dr John Porrill, Dr. John E.W. Mayhew, Prof. John Frisby and Dr. Stephen Billings for their contributions to this work. We gratefully acknowledge the support of RSRE Malvern and the ESRC/MRC/SERC in the funding of this project.

Appendix I. Weighting Factor Calculation

The weighting factor for each 3D feature point should be related to the 3D error model of the feature point. The uncertainty about a 3D point is mostly contributed by the uncertainty in the direction aligned with the line of sight to the point. Based on this observation, Kiang et al [4] constructed the error model for each 3D point as a 1D line segment. By denoting P as a nonlinear mapping which maps a 2D point with disparity into a 3D point, i.e.,

$$P\colon (i, j, d) \to (x, y, z) \qquad \text{(A1.1)}$$

the two end points of the line segment of the point $P(i, j, d)$ are defined as

$$P_{fr} = \frac{P(i, j, d-1) + P(i-1, j, d-1)}{2} \qquad \text{(A1.2)}$$

$$P_{cl} = \frac{P(i, j, d+1) + P(i+1, j, d+1)}{2} \qquad \text{(A1.3)}$$

Hence given two sets of 3D feature points corresponding to each other, the weighting factor w_i is defined as

$$w_i = \frac{1}{\|P_{fr,i} - P_{cl,i}\|^2 + \|P_{fr,i}' - P_{cl,i}'\|^2} \qquad \text{(A1.4)}$$

A modification on calculating the line segment can be made due to the fact that image feature points and disparities are obtained to sub-pixel acuity, thus eq.s(A1.2) and (A1.3) can be modified yielding

$$P_{fr} = \frac{P(i, j, d-\sigma) + P(i-\sigma, j, d-\sigma)}{2} \qquad \text{(A1.5)}$$

$$P_{cl} = \frac{P(i, j, d+\sigma) + P(i+\sigma, j, d+\sigma)}{2} \qquad \text{(A1.6)}$$

where σ is the standard deviation of the error associated with each 2D feature point and disparity calculations and is usually less than unity.

Appendix II

To summarise the optimisation procedure, define weighted centroids as

$$\mathbf{p}_w = \sum_{i=1}^{N} w_i\mathbf{p}_i / \sum_{i=1}^{N} w_i\ , \qquad \mathbf{p}_w' = \sum_{i=1}^{N} w_i\mathbf{p}_i' / \sum_{i=1}^{N} w_i$$

and let

$$\mathbf{n}_{i,w} = \mathbf{p}_i - \mathbf{p}_w\ , \qquad \mathbf{n}_{i,w}' = \mathbf{p}_i' - \mathbf{p}_w'$$

Then R and T are given by

$$\begin{aligned}\min_{R,T} S &= \min_{R,T} \sum_{i=1}^{N} w_i \|\mathbf{p}_i' - R\mathbf{p}_i - T\|^2 \\ &= \min_{R,T} \Big(\|\mathbf{p}_w' - R\mathbf{p}_w - T\|^2 \sum_{i=1}^{N} w_i \\ &\quad + \sum_{i=1}^{N} w_i \|\mathbf{n}_{i,w}' - R\mathbf{n}_{i,w}\|^2\Big)\end{aligned}$$

because $\sum w_i\mathbf{n}_{i,w} = \sum w_i\mathbf{n}_{i,w}' = 0$. As in the unweighted version of this algorithm [1] the first term can be set to zero for any rotation $\hat{R}$ using

$$\hat{T} = \mathbf{p}_w' - \hat{R}\mathbf{p}_w$$

which is the least squares optimal estimate of T. Thus the total expression can be minimised simply by minimising the second term with respect to R.

References.

[1] **Faugeras, O.D. and M. Hebert** "The representation, recognition, and localisation of 3D shapes from range data" *Int. J. Robotics Res.* Vol.5, Pt.3, pp.27-52. 1986.

[2] **Harris, C.G. and J.M. Pike.** "3D positional integration from image sequences." *Proceedings of the Third Alvey Vision Conference.* pp.223-236 September 1987.

[3] **Harris, C. and M.Stephens** "A Combined Corner and Edge Detector." *Proceedings of the*

Fourth Alvey Vision Conference. pp.147-151 August 1988.

[4] **Kiang, S.M., R.J. Chou, J.K. Aggarwal** "Triangulation errors in stereo algorithms", *Proc. IEEE Comp. Vision Workshop* Miami Beach, Dec. 1-2, 1987, pp.72-78.

[5] **Matthies, L. and T. Kanade.** "The cycle of uncertainty and constraint in Robot Perception" *Robotics Research.* Vol.4, Pt.6, pp 327-335. 1988.

[6] **Moravec, H.P.** "Obstacle avoidance and navigation in the real world by a seeing robot rover" Ph.D Thesis, Stanford Univ., Sept., 1980.

[7] **Pollard, S.B., J. Porrill and J.E.W.Mayhew.** "Experiments in vehicle control using predictive feed-forward stereo." *Image and Vision Computing.* vol 8, no 1, pp.63-70,1990.

[8] **Thacker, N.A. and J.E.W.Mayhew.** "Optimal Combination of Stereo Camera Calibration from Arbitrary Stereo Images." *These Proceeedings.*