# Identifying planar regions in a scene using uncalibrated stereo vision.

Gabriel Hamid, Nick Hollinghurst and Roberto Cipolla
Department of Engineering, University of Cambridge,
Cambridge, CB2 1PZ.
{gh,njh,cipolla}@eng.cam.ac.uk

**Abstract**

We describe the use of well-known uncalibrated stereo algorithms for detecting planar regions in a scene from the transformation of feature locations between views. Simulations indicate that a typical set-up would have a resolution of the order of one centimetre. A fully operational system is not yet complete, but here we present a number of steps towards achieving this goal.

**Keywords:** uncalibrated stereo vision, segmentation, planar.

## 1 Introduction

We are developing a system which combines stereoscopic vision with a robotic manipulator to enable it to locate, reach and grasp unmodelled objects in an unstructured environment. Part of the system has been built. The algorithm for indicating the object of interest is described in [2] and the algorithm for visually guiding the robot arm to the object is described in [6]. Both these algorithms use uncalibrated stereo vision. The advantages of using uncalibrated stereo are that it is easier to set up, more robust to disturbances of the cameras and insensitive to uncertainties in the camera parameters.

In order to complete the system an uncalibrated stereo algorithm is required for grasp planning. Many robotic grippers consist of two parallel jaws, such a mechanism is well suited to grasping objects with parallel planar facets. Hence, a simple paradigm for grasp planning is to search for planar facets. There are well-known uncalibrated stereo algorithms for detecting planar regions in a scene [4, 11]. Here we describe for the first time the implementation of these algorithms to this application. A fully operational system is not yet complete, but here we present a number of steps taken towards achieving this goal.

## 2 Theoretical framework

The principle underlying all uncalibrated stereo algorithms for segmenting a scene into planar regions is the same and is summarised here.

Two views of a planar surface are related by a two-dimensional projective transformation. Features are grouped according to coplanarity by searching for

BMVC 1996 doi:10.5244/C.10.20

features which follow the same transformation between the two images. The search space is large because it is also necessary to search for the correspondence between the images. The search is performed by a strategy of hypothesis, prediction and testing, see figure 1. A hypothesis consists of a basis set of matching features thought to be coplanar. This defines a projective transformation between the two stereo views. A prediction consists of the mapping of a feature from one image to the other according to this transformation. If the transformation correctly predicts how other features transfer between the images then the hypothesis is accepted and the features are grouped as a plane. Whereas if no consensus can be found with any other features then the hypothesis is discarded, and another one must be tried. The correctness of a prediction is determined by a statistical test, such as the chi-squared test on the Mahalanobis distance between the features. The uncertainty in the positions of both the transferred feature and its predicted match are computed by the *propagation of errors* through all the computations starting from the initial image measurements. If the Mahalanobis distance between a transferred feature and its predicted match is below a specified confidence level then the match is deemed correct, otherwise not.

## Computation of uncertainty by the propagation of errors

**Method** The approach is adapted from [3]. Let $\mathbf{x}$ be a vector in $R^n$ of some measurement data with associated covariance matrix $\Lambda_x$, and let $\mathbf{y}$ be a vector in $R^m$ computed from those measurements. If $\mathbf{y} = \mathbf{f}(\mathbf{x})$ then to a first order approximation the uncertainty on $\mathbf{y}$ is:

$$\Lambda_y = \mathbf{Df}(\mathbf{x})\Lambda_x\mathbf{Df}(\mathbf{x})^T \tag{1}$$

where $\mathbf{Df}(\mathbf{x})$ is the derivative of $\mathbf{f}$ evaluated at $\mathbf{x}$, an $m \times n$ Jacobian matrix.

**An example: Computation of the uncertainty on the projective transformation between two views of a planar surface.**

Let $(A_i, B_i, C_i)$ and $(a_i, b_i, c_i)$ be the homogeneous coordinates of a basis set of four corresponding pairs of lines, $i = 1$ to 4. The projective transformation between the two views can be written as:

$$k \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} = \begin{pmatrix} t_0 & t_1 & t_2 \\ t_3 & t_4 & t_5 \\ t_6 & t_7 & 1 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} \quad \text{where } k \neq 0$$

The eight parameters of the projective transformation are computed directly from the four pairs of corresponding lines [8]. The Jacobian of the function mapping the line data to the transformation parameters is an $8\times24$ matrix of the form:

$$\mathbf{Df} = \begin{pmatrix} \frac{\partial t_0}{\partial A_1} & \cdots & \frac{\partial t_0}{\partial c_4} \\ \vdots & \ddots & \\ \frac{\partial t_7}{\partial A_1} & & \end{pmatrix} \tag{2}$$

The individual elements of the Jacobian matrix are estimated numerically by finite differences. The measurement data consists of the four pairs of corresponding

lines, because the lines are all independent, the 24×24 covariance matrix of the measurement data is of the form:

$$\mathbf{\Lambda_X} = \begin{pmatrix} \Lambda_0 & \mathbf{0} & \cdots & \\ \mathbf{0} & \Lambda_1 & & \\ \vdots & & \ddots & \\ & & & \Lambda_7 \end{pmatrix} \tag{3}$$

where $\Lambda_0, \ldots, \Lambda_7$ are the 3×3 covariance matrices of the eight lines, and $\mathbf{0}$ is the 3×3 null matrix. The uncertainty in the positions of each line is computed from the residuals of the line fit to the edge data. Finally, the uncertainty of the projective transformation is computed by inserting equations (2) and (3) into (1).

# 3 Experimental results

Typical values for the parameters of the stereo rig are: 0.6 metres baseline between the optical centres of the cameras, 30 degree angle between the principal axes of the cameras, and a depth of 1.5 metres from the cameras to the object. We define the resolution of the system as the minimum distance between a point and a plane, such that the point can be distinguished from the plane. At this stage we do not know what value of resolution is required for the task of grasp planning – but about a centimetre seems a reasonable guess.

## Simulations

Each camera is represented by a perspective projection matrix [3]. The intrinsic parameters of the two cameras are set as follows: focal length 8.5 mm, pixel coordinates of principal point (256, 256), pixel width 12.7 microns, pixel height 8.3 microns. The scene consists of two parallel planes separated by a perpendicular distance of 1cm. Each plane is marked with a uniform grid of 50 points, see figure 2. The scene is projected onto the two cameras. Gaussian noise is added to the image points. A basis set of four corresponding points is chosen and all the points are tested for coplanarity with this basis. A Mahalanobis distance between a transferred point and its corresponding match is defined as $\mathbf{x}^T C^{-1} \mathbf{x}$ where, $\mathbf{x}$ is the difference between a transferred point and its match, and $C$ is the sum of the covariance matrices for the transferred point and its match. An image point has two degrees of freedom so the Mahalanobis distance follows a chi-squared probability distribution with two degrees of freedom. A threshold of 9.21 on the Mahalanobis distance sets the confidence level to 95% [10]. If the Mahalanobis distance between a transferred point and its predicted match is less than 9.21 then the point is deemed as belonging to the plane. Results are shown in figure 3. Similar results were obtained when the experiment was repeated for different sets of bases points.

The two planes are only correctly discriminated when the level of noise on the image points was down to 0.1 pixels. From these results we conclude that it is necessary to localise image features to a precision of 0.1 pixels in order to segment the scene into planar regions to a resolution of one centimetre. Corner detectors

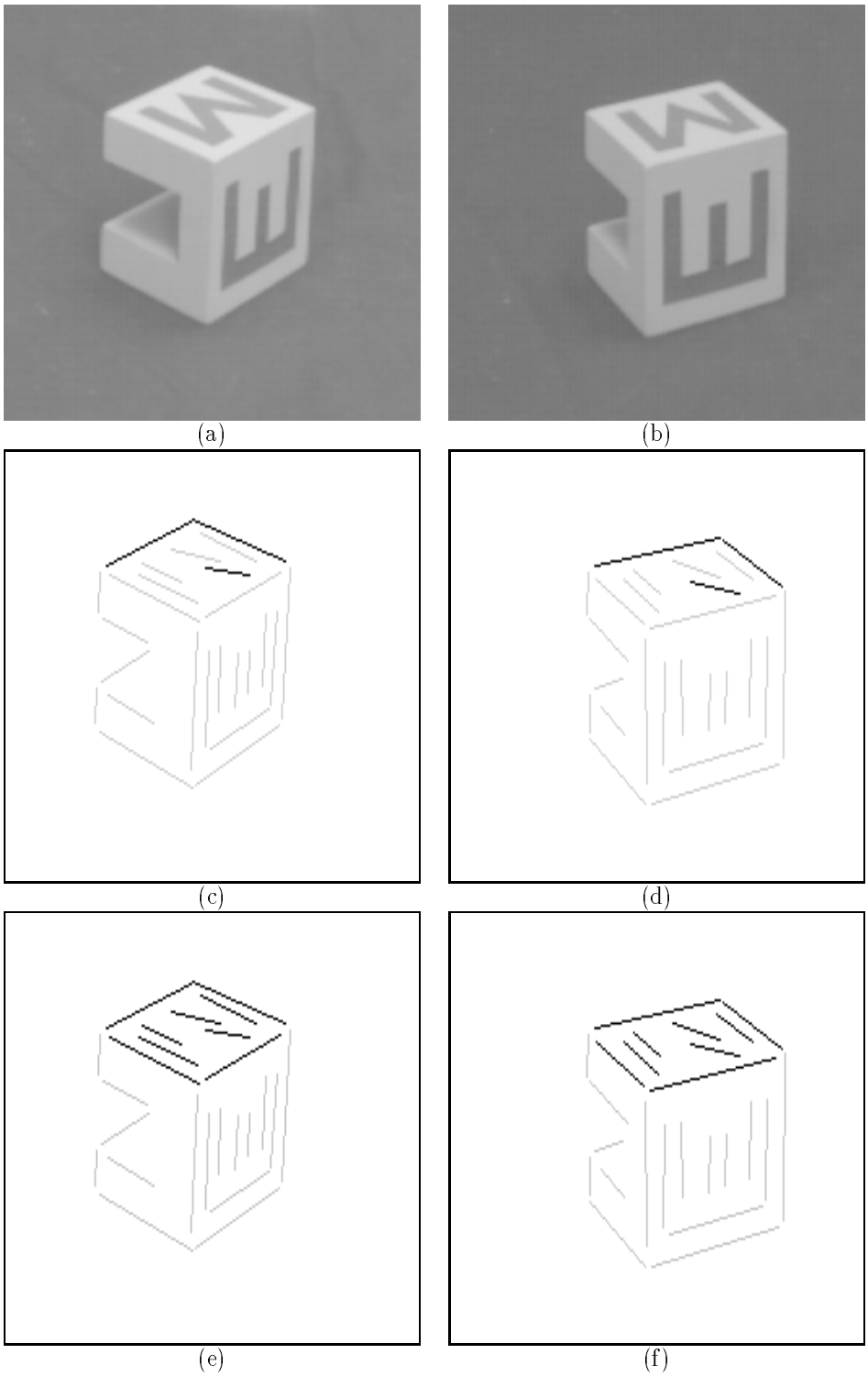



Figure 1: The greyscale images of a stereo pair are shown in figures (a) & (b). A hypothesis is made of a basis set of coplanar line segments, as highlighted in figures (c) & (d). For the affine case, three corresponding pairs of line segments define the transformation between the two views. Line segments that follow the same transform are grouped as being coplanar with the basis set, as shown in figures (e) & (f).

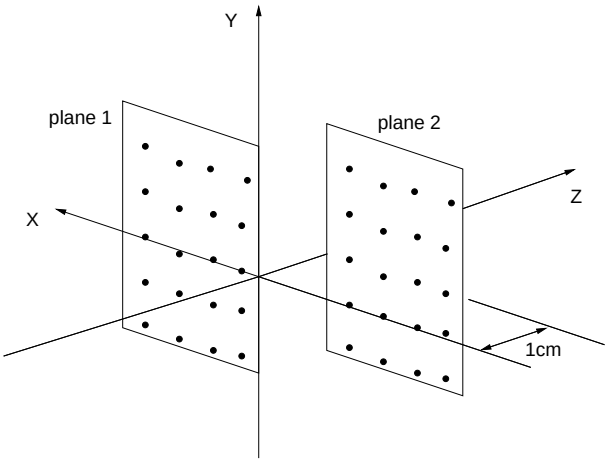



Figure 2: The simulated scene consists of two parallel planes separated by a 1cm gap

have been quoted to sub-pixel accuracy [13]. The position and orientation of a line segment can generally be measured with more precision than the position of an isolated point [1]. So by using line features the system should be capable of segmenting a scene into planar facets to a resolution of about a centimetre.

## Real data

Line segments are detected using Canny's edge detector, followed by chaining, then recursive splitting [1]. The Canny edge detector performs an isotropc smoothing of the image, which can introduce a bias error when two edges are nearby to one another. In order to avoid this bias only uncluttered scenes are used. Image blur is another cause of isotropic smoothing, hence it is necessary to ensure the cameras are well-focussed. A straight line is fitted to each chain of edgels by an orthonormal regression [3]. Lines are represented by an equation of the form $Ax + By + C = 0$. The uncertainty in the position of each line is computed from the residuals of the best-fit line to the edge data [10].

A basis set, consisting of four line segments, is manually selected. The projective transformation between views, together with the uncertainty of the transformation, is computed by the method outlined in section 2. The uncertainty of a transferred line is computed from the uncertainty of the original line and the uncertainty in the projective transformation. The line representation is converted to the form (m,c), where y=mx+c or x=my+c depending on the line orientation, hence the same Mahalanobis distance criteria described above can be used to test predicted matches.

Figure 4 comprises a stereo pair of images and the line segments detected. In figure 4(d) two transferred lines are shown. The upper of the two transferred lines is correctly matched by the algorithm, the Mahalanobis distance is 1.6. Despite appearances (remember the computations are to sub-pixel accuracy) the lower of

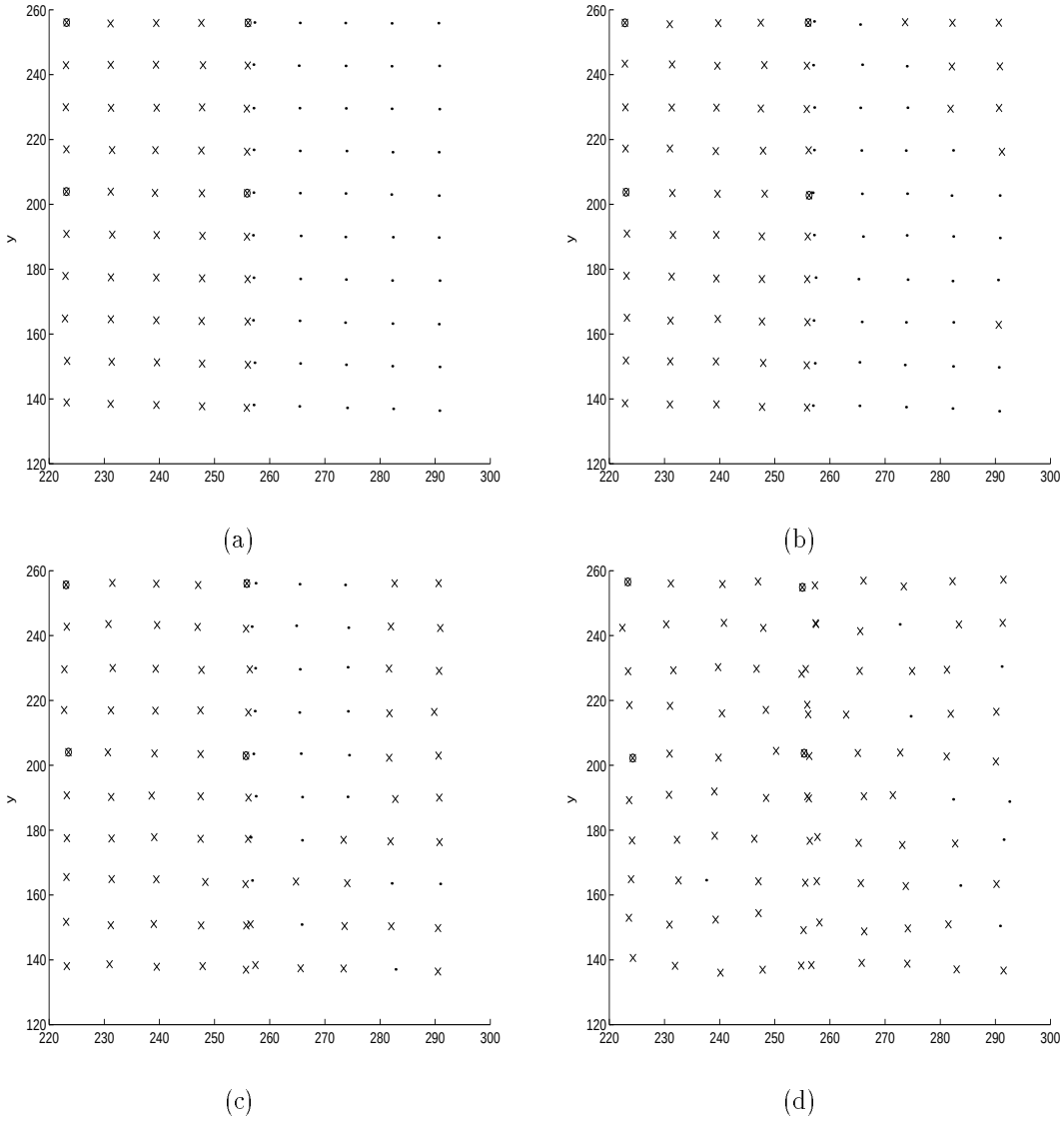



Figure 3: Each figure represents the image in the left camera. Figures a,b,c and d are for four different levels of Gaussian noise, 0.1, 0.2, 0.3, 1.0 pixels standard deviation, respectively. An 'x' marks image points that are computed as being coplanar with the basis set, and a '.' marks image points that are not. The basis set of points is highlighted with a small circle. Correct discrimination of the two planes only occurs in (a).

the two transferred lines is not sufficiently close to its corresponding match in order to be deemed correct, the Mahalanobis distance is 15. In order to understand why one line segment is accepted and the other is not, the experiment was repeated for one hundred trials. For each trial a new stereo pair of images is taken of the same static scene. The graphs in figure 5 plot the parameters of the same transferred line and its corresponding match for the hundred trials. The ellipses shown in the graphs are the computed 95% confidence regions from the first trial alone. The size of the computed confidence regions are of the same size as the scatter of the measurement data, this shows that the uncertainty computed by the propagation of errors from one trial is to the right order of magnitude as the true uncertainty measured over a hundred trials. There is greater uncertainty in the position of a transferred line than its corresponding line because it is dependent on more image measurements. The shift between the centre of each ellipse and the the scatter of data is because each ellipse is centred on the line parameter computed in that one particular trial and not on the mean value averaged over the hundred trials.

The lower of the two transferred lines in figure 4 was rejected during the first trial because it happened to be on the outer limits of the distribution, as shown in the graph in figure 5(b). If, however, the uncertainty values are measured over the hundred trials instead of being computed from one trial then the transferred line would be correctly matched. The Mahalanobis threshold was set to yield a 95% confidence level, so on average one would expect to incorrectly reject 5% of the matches, but in practice we find that we lose a much higher proportion. Here we suggest that it is more reliable to measure the uncertainty over a hundred trials than to compute the uncertainty by the propagation of errors from one trial.

# 4 Conclusion

In this paper we have investigated the feasability of using uncalibrated stereo vision for detecting planar regions in a scene. The results from simulations indicate that the resolution of a typical system is of the order of one centimetre. In practice such resolution has not yet been achieved, but initial results are encouraging.

It is important to minimise any systematic error when implementing an algorithm based on a geometric computational framework [7]. For example, it was necessary to restrict experiments to uncluttered scenes because the Canny edge detector introduces a bias error when two edges are nearby to one another. We are currently replacing the Canny edge detector with one based on anisotropic diffusion [9]. Other potential sources of systematic error include lens distortion and non-Lambertian reflection causing variations in visual features between viewpoints. We are also currently assessing the merits of measuring the uncertainty values over successive frames rather than computing the uncertainty values by the propagation of errors. For a hardware implementation, measuring the uncertainty is clearly preferable because no extra circuitry is required for the propagation of errors. For future work we plan to incorporate the estimation of the Fundamental matrix [12] in order to provide a further constraint on the allowable transformations between views, and to extend the approach to a trinocular system.

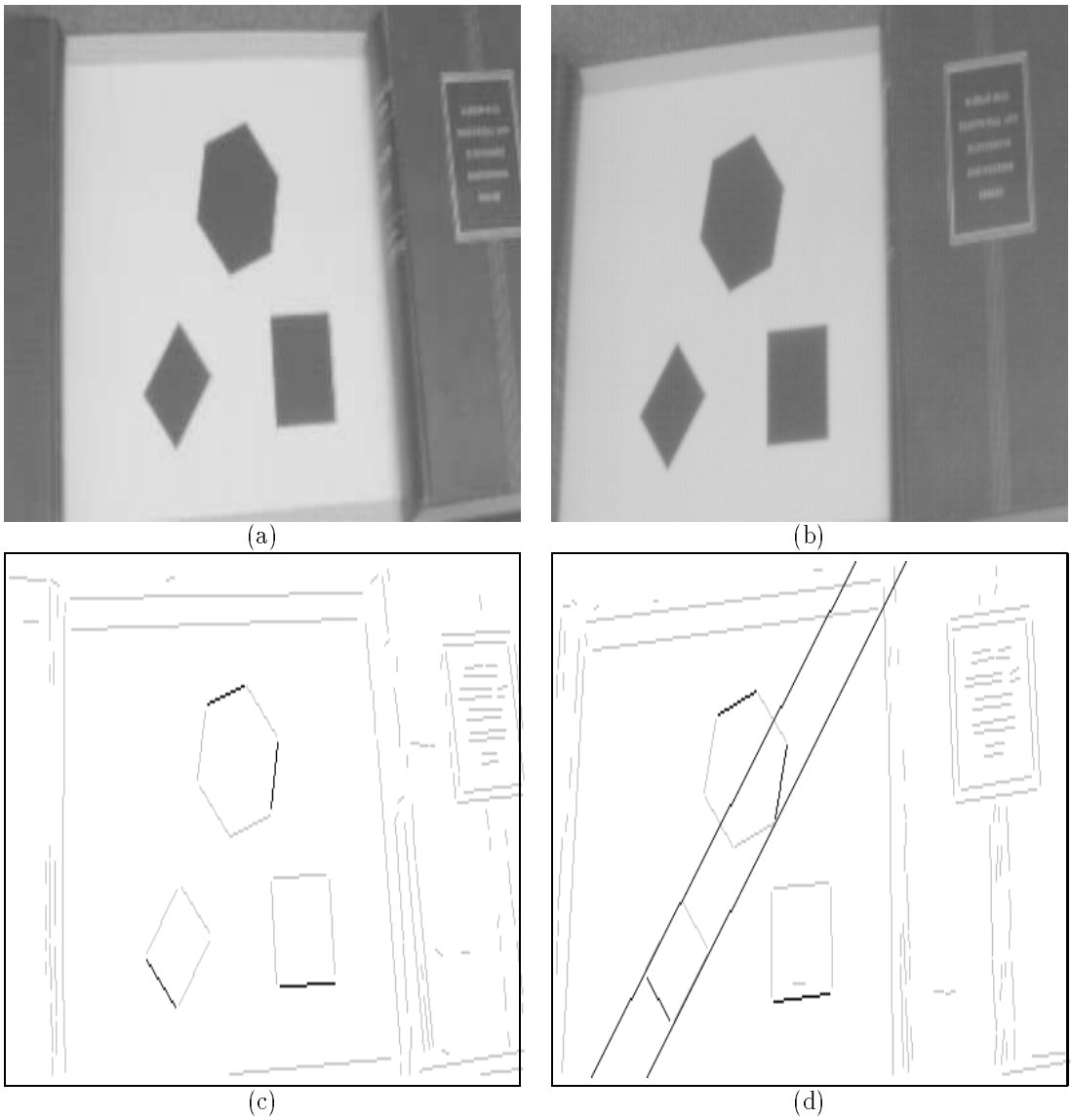
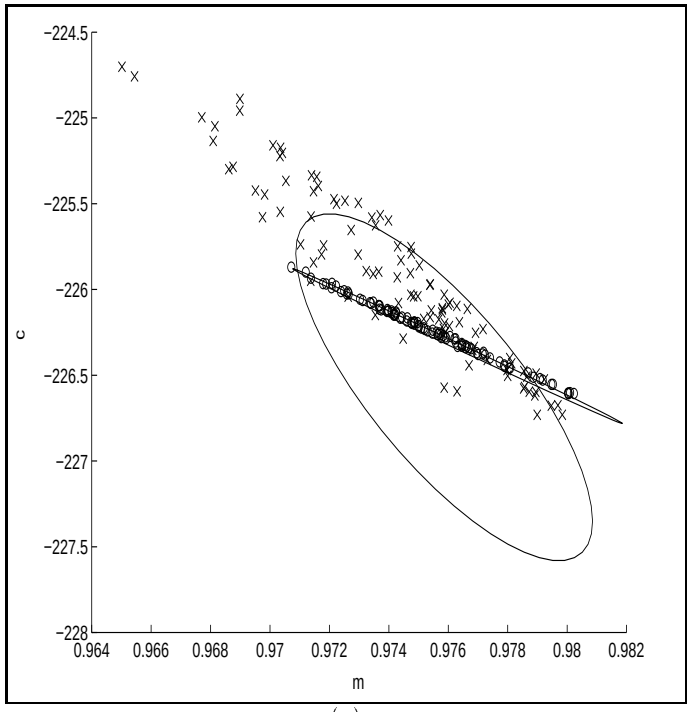
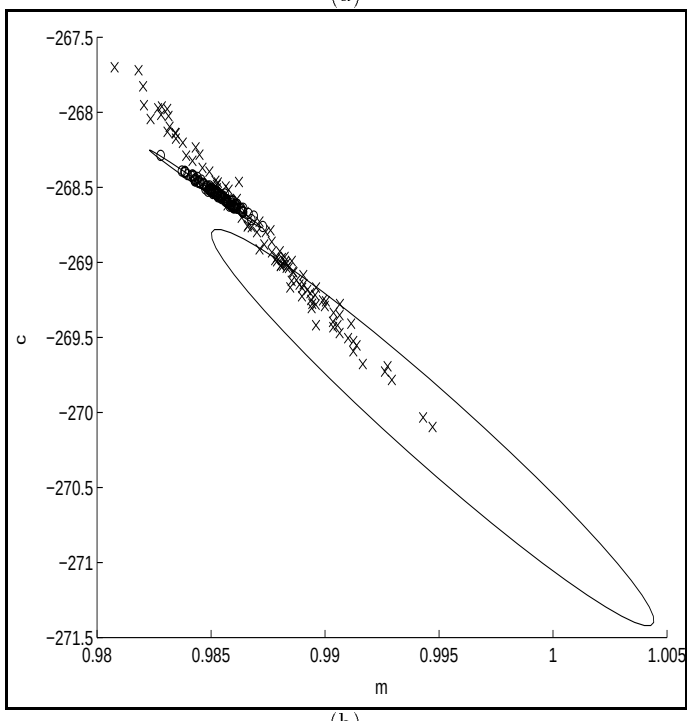



Figure 4: A stereo pair of images are shown in (a) & (b). The line segments detected are shown in (c) & (d). The four highlighted line segments define a projective transformation between the two views. Two lines are transferred across from the left image to the right image, as shown by the continuous lines in (d).




(a)




(b)

Figure 5: Each graph plots the parameters of a transferred line and its corresponding match for 100 trials. In both graphs, the intercept 'c' of the line is plotted against the gradient 'm', where the line is of the form y=mx+c. A transferred line is marked with an 'x' and its corresponding line segment is marked with an 'o'. The ellipses show the confidence regions computed by the propagation of errors from the first trial alone, hence the ellipses are centered on the value of the parameters computed during that first trial. The scatter of the data from the hundred trials reveals the true uncertainty.

# References


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

[2] R. Cipolla, P. Hadfield, and N. Hollinghurst. Uncalibrated stereo vision with pointing for a man-machine interface. In *Proc.IAPR Workshop on Machine Vision Applications*, pages 163–166, Kawasaki, Japan, 1994.

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

[4] O. Faugeras and F. Lustman. Motion and structure from motion in a piecewise planar environment. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 2(3):458–508, 1988.

[5] M. Fischler and R. Bolles. Random sample consensus. *Graphics and Image Processing*, 24(6):381–395, 1981.

[6] N. Hollinghurst and R. Cipolla. Uncalibrated stereo hand-eye coordination. *Image and Vision Computing*, 12(3):187–192, 1994.

[7] K. Kanatani. *Geometric computation for machine vision.* Oxford University Press, 1993.

[8] J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision.* MIT Press, 1992.

[9] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Analysis and Machine Intell.*, 12(7):629–639, 1990.

[10] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C.* Cambridge University Press, 1992.

[11] D. Sinclair, A. Blake, S. Smith, and C. Rothwell. Planar region detection and motion recovery. In *Proc. British Machine Vision Conference*, pages 59–68, 1992.

[12] P. Torr. *Motion segmentation and outlier detection.* PhD thesis, University of Oxford, 1995.

[13] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Technical Report 2273, INRIA, France, 1994.