Visually Guided Grasping In Unstructured Environments

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

Journal of Robotics and Autonomous Systems (Invited Paper)

Abstract

We present simple and robust algorithms which use uncalibrated stereo vision to enable a robot manipulator to locate, reach and grasp unmodelled objects in unstructured environments. In the first stage, an operator indicates the object to be grasped by simply pointing at it. Next, the vision system segments the indicated object from the background, and plans a suitable grasp strategy. Finally, the robotic arm reaches out towards the object and executes the grasp. Uncalibrated stereo vision allows the system to continue to operate in the presence of errors in the kinematics of the robot manipulator and unknown changes in the position, orientation and intrinsic parameters of the stereo cameras during operation.

1 Introduction

When humans grasp objects, they usually do so with the aid of vision. Visual information is used to locate and identify things, and to decide how they should be grasped. Visual feedback is used to guide the hand to the target. In a similar manner, machine vision can be used to coordinate the manouevres of a robotic arm. Here we describe a system which combines uncalibrated stereo vision with a robotic manipulator to enable it locate, reach and grasp unmodelled objects in an unstructured environment.

Calibrated stereo vision has been used before in robotic applications. A well-calibrated stereo rig can accurately determine the positions of things to be grasped [19]. Calibration, however, is a non-trivial process, and if it is inaccurate or the cameras are disturbed, the system fails gracelessly. An alternative approach in hand-eye applications, where a manipulator moves to a visually-specified target, is to use visual feedback to match the manipulator and the target positions in the image. Exact spatial coordinates are not required, and a well-chosen feedback architecture can correct for quite serious inaccuracies in camera calibration as well as inaccurate kinematic modelling of the robot arm. However, visual feedback alone is inefficient and can lead to unstable behaviour. It is better to exploit the relationship between the robotic kinematics and the vision system.

The system described here is based upon uncalibrated stereo vision. No assumptions are made on accurately knowing the parameters of the cameras. Instead, the system is self-calibrating: an approximate relation between the cameras and the robot's coordinate frame is computed automatically by observing the robot gripper as it follows a preprogrammed sequence of manouevres. This approximate calibration provides the coarse control, and visual feedback provides the fine control. The system is both robust and user-friendly. The system is robust because it continues to operate successfully in the presence of errors in the kinematics of the robot manipulator, uncertainties in the camera parameters, uncertainties in the position of the object and even during unknown changes in the position, orientation and intrinsic parameters of the stereo cameras. The system is easy to set-up and adjust according to the specific application because there are no fixed constraints on the camera geometry.

Operation comprises three distinct stages. In the first stage, an operator indicates the object to be grasped by simply pointing at it. Next, the vision system segments the identified object from the background (by grouping edges into planes) and plans a suitable grasp strategy. Finally, the robotic arm reaches out towards the object and executes the grasp. Here we present the three vision algorithms used for these three separate stages.

Figure 1 shows a typical set-up. The robot arm has 5 degrees of freedom and a parallel-jawed gripper. The robot has its own controller for the low-level control and provides a Cartesian kinematic model. Two cameras are placed 1.5 to 3 metres from the robot's workspace. The angle between the cameras is in the range of 15–30 degrees.

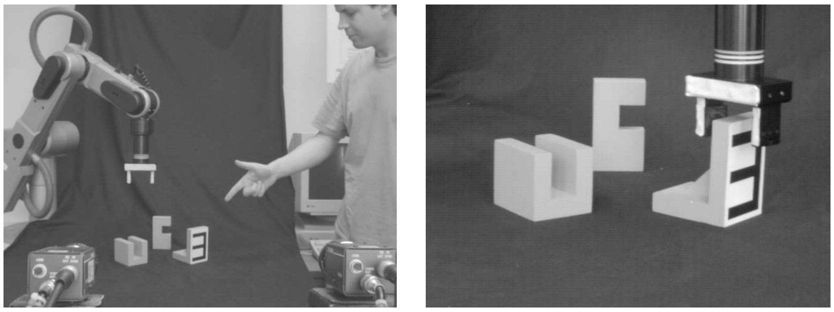


Figure 1: The arrangement of the uncalibrated stereo cameras and the robot. The operator points at an object, and the robot picks it up under visual control.

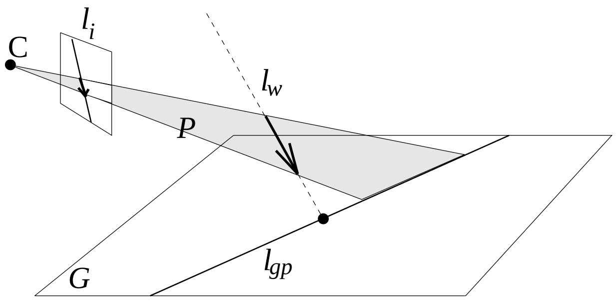
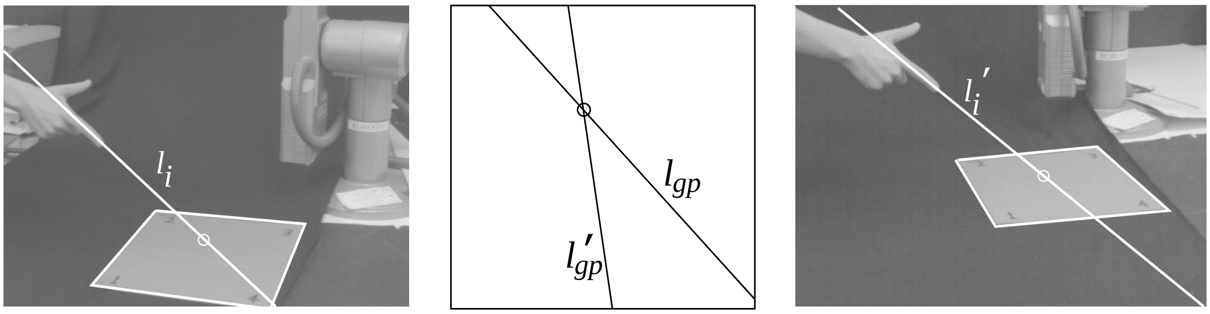


Figure 2: Relation between lines in the world, image and ground planes

2 Indicating the target by pointing

For the first part of the experimental system, we use machine vision in conjunction with a human operator who indicates the object to be grasped by pointing at it. A pair of uncalibrated cameras view the pointing hand in stereo. Unlike most existing gesture-based interfaces (e.g. [9, 22]) our system requires no special gloves or markers. Active contours are used to track the hand in real time. A simple result from projective geometry allows us to estimate where the hand is pointing to on the robot's work table, using just four coplanar reference points. The cartesian co-ordinates of the reference points are also unknown and camera calibration is not required.



(a) Left view (b) Canonical view (c)Right view

Figure 3: Combining constraints from two views of a pointing hand

2.1 Geometrical framework

A single view of a pointing hand is ambiguous: its distance from the camera cannot be determined, and the *slant* component of its orientation cannot be measured with any accuracy. This means that the 'piercing point', where the line defined by the hand intersects the work surface, is constrained to a line, which is the projection of the hand's line in the image (see figure 2). A second view is needed to fix its position in two dimensions [18].

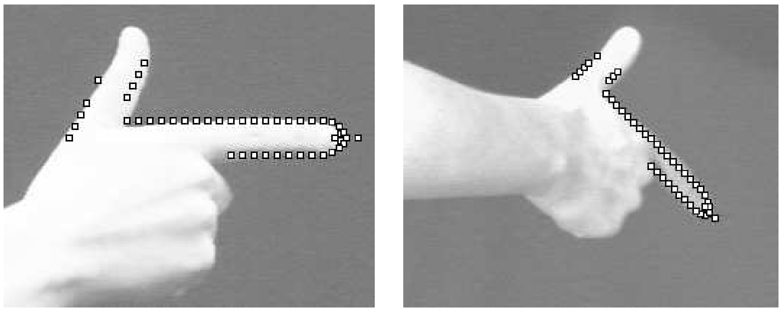
Consider a pair of pinhole cameras viewing a planar surface (such as the robot's work surface). The viewing transformations can be modelled by plane projectivities, and there exists a projective transformation (homography) that maps one image to the other. This transformation can be computed by observing a minimum of four points on the plane. We exploit this to transform the constraint lines into a common 'canonical' view of the plane, and hence find their intersection (figure 3).

The piercing point can then be projected back into the two images; if the four reference points are known its world coordinates can be calculated, otherwise visual feedback can be used to guide a robot manipulator to this point (section 4).

2.2 Tracking mechanism

We use an active contour tracker [12, 3] to track the image of a hand in the familiar 'pointing' gesture. The tracker is based on a template, representing the shape of the occluding contours of an index finger and thumb (figure 4).

The tracker's motion is restricted to 2D affine transformations in the image plane, which ensures that it keeps its shape whilst tracking the hand



(a) (b)

Figure 4: The finger-tracking active contour (a) in its canonical frame (b) after an affine transformation in the image plane (to track a rigid motion of the hand in 3D).

in a variety of poses [2]. This approach is best suited to tracking small planar objects, but also works well with fingers, which are cylindrical.

2.3 Implementation

The four reference points were defined by either observing the robot as it moved to four known points 50mm above the table-top or by tracking a coplanar surface. Hand-trackers are initialised by the operator moving his hand so that it fits into the templates displayed on the left and right monitor screens. The presence of high contrast edges activate the deformable templates which then track the hand continuously in both left and right images (Figure 4).

Figure 5 shows the system in operation. The white quadrilateral represents the images of the reference points in the two views, and the overlaid square shows the position of the indicated point on the plane. Movements of the operator's hand cause corresponding movements of this point in real time, indicating to the system which object is to be grasped.

The accuracy of the indicated point can be estimated by calculating the image errors in localising the hand and reference points. These uncertainties are then propagated through to the computation of the constraints. A stereo vergence angle of about 20° produces an indicated point with accuracy of about 1cm at a distance of 1m [5].

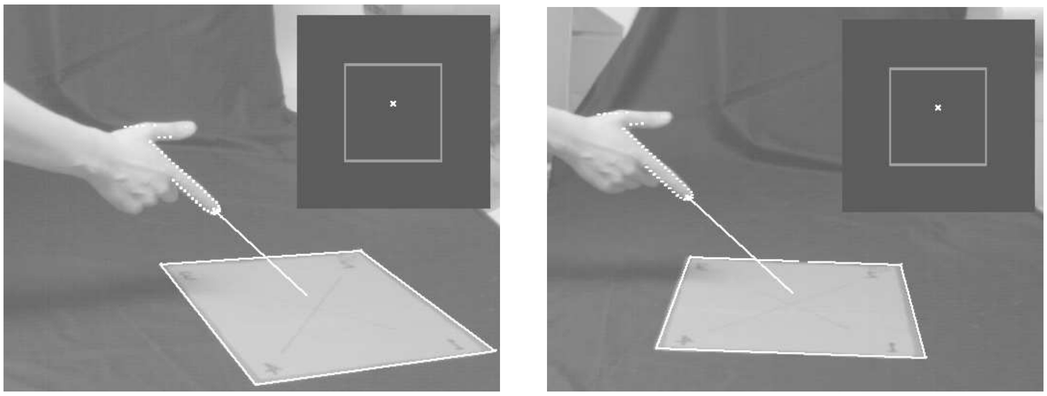


Figure 5: **Stereo views of a pointing hand**. The two views are shown side by side. In each view an active contour is tracking the hand. The inlaid square is the *canonical view* and the cross shows the the indicated point in working plane coordinates.

3 Segmentation and grasp planning

In this section, we describe how the uncalibrated stereo vision system extracts the structure of the object. Many robotic grippers consist of two parallel jaws, such a mechanism is well suited to grasping objects with parallel planar surfaces. Hence, a suitable first step to grasp planning is to search for planar facets. There are well-known algorithms for detecting planar regions in a scene [8, 20]. We describe the adaptation of these algorithms to this application.

Faugeras and Lustman [8] describe an algorithm for identifying planar surfaces in a scene from just two views. They considered the case of a single calibrated camera with unknown motion between two viewpoints. We extend their approach to the case of uncalibrated stereo vision.

3.1 Geometrical Framework

Two views of a planar surface are related by a two-dimensional projective transformation. Under weak perspective they are related by an affine transformation. Features are grouped according to coplanarity by searching for features which follow the same transformation between the two images. In general the search space is prohibitively large because it is also necessary to search for the correspondence between the images. In our system approximate epipolar geometry is used to match line segments.

A hypothesis consists of a basis set of matching line features thought to be coplanar. This defines the affine 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.

In a geometric computational approach [11], the correctness of a prediction is determined by a statistical test, such as the chi-squared test on the Mahalanobis distance between a transferred feature and its predicted match. The uncertainty in the position of 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.

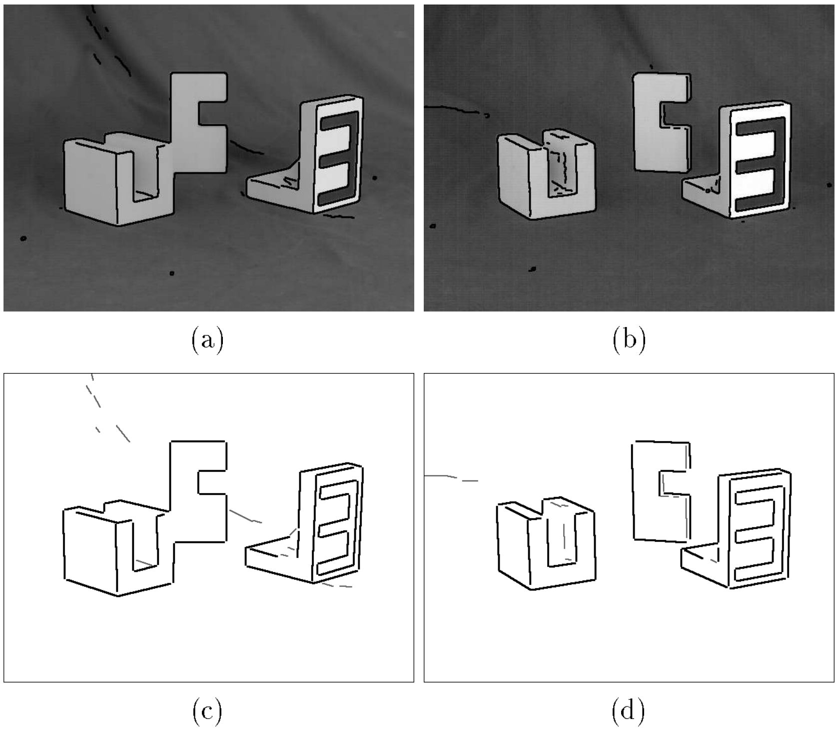


Figure 6: Stereo correspondences: (a,b) stereo images of the workspace with edges superimposed; (c,d) unmatched (light) and matched (dark) line segments.

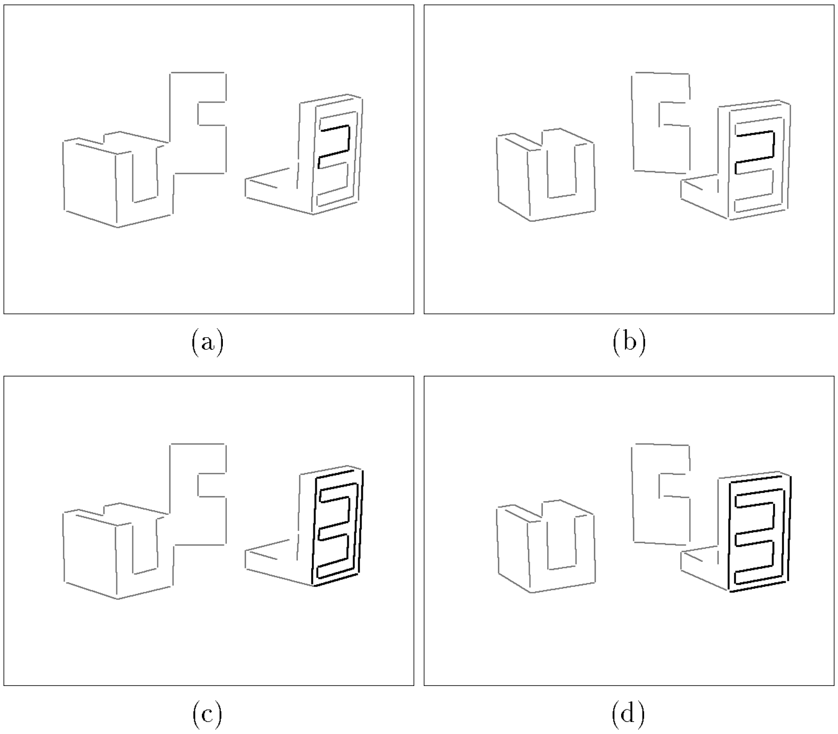


Figure 7: Plane grouping by consensus: (a,b) segments chosen by heuristics to form a plane hypothesis, which is an affine transformation between views; (c,d) segments consistent with the hypothesis.

3.2 Experimental Results

Line segments are detected using Canny's edge detector, followed by chaining, then recursive splitting [1]. Straight lines are fitted to each chain of edgels by an orthonormal regression [7]. Figure 6 stereo pair of images and the line segments detected. 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 [17].

A basis set, consisting of three line segments, is selected automatically, shown in Figure 7. The affine transformation between views, together with the uncertainty of the transformation, is computed by the propagation of errors. The uncertainty of a transferred line is computed from the uncertainty of the original line and the uncertainty in the affine 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 7 c and d shows the line edges which are consistent with the affine transformation defined by the hypothesis and hence are grouped into the same plane.

4 Executing the grasp under visual control

In this section we describe a robust algorithm for executing a grasp. World coordinates are calculated using a linear model of stereo vision which, though of limited open-loop accuracy, is robust to camera disturbances and is easy to calibrate automatically. We assume that the robot's kinematics are known, at least approximately. Closed-loop control is achieved by tracking the gripper's movements across the two images to estimate its position and orientation relative to the target object. The offset is used as a feedback term to guide it into its grasping configuration.

4.1 Affine stereo

Perspective camera model

The conventional perspective camera model is a projective transformation between world coordinates (X, Y, Z) and image coordinates (u, v), and is represented by a 4×3 matrix $\mathbf{M}$ for each camera [21]. This encodes the camera's position and orientation, as well as the intrinsic parameters of the image sensor. Using homogeneous coordinates (with a tilde to symbolize equality up to a scale factor):

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} \mathrm{m}_{11} & \mathrm{m}_{12} & \mathrm{m}_{13} & \mathrm{m}_{14} \\ \mathrm{m}_{21} & \mathrm{m}_{22} & \mathrm{m}_{23} & \mathrm{m}_{24} \\ \mathrm{m}_{31} & \mathrm{m}_{32} & \mathrm{m}_{33} & \mathrm{m}_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{1}$$

Calibration of the camera is necessary to fix the twelve parameters m_{ij}. This can be done by observing at least six points of known position, not all coplanar. The system is homogeneous, so we can constrain $\mathrm{m}_{34} = 1$, and solve for the other coefficients using linear techniques.

In practice, solving for $\mathbf{M}$ is somewhat ill-conditioned, and a large number of reference observations are needed. Non-linear minimization methods can improve the accuracy of calibration. However, if the cameras are disturbed after calibration, stereo reconstruction based on this model is degraded – its nonlinear structure means that errors in some directions are greatly magnified [4].

Affine camera model

Consider a camera viewing a compact scene from a distance several times its maximum diameter. The scaling effect of depth variations (represented by m_{31}, m_{32}, m_{32}) becomes insignificant, and the relation between world and image coordinates can be written very simply as a linear mapping:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \mathrm{p}_{11} & \mathrm{p}_{12} & \mathrm{p}_{13} & \mathrm{p}_{14} \\ \mathrm{p}_{21} & \mathrm{p}_{22} & \mathrm{p}_{23} & \mathrm{p}_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{2}$$

This is the *affine camera* model [14]. It can be shown that this approximate model is more robust to errors in both image and world coordinates [4].

The affine stereo formulation

We assume that the cameras do not move relative to the scene during each period of use. Combining information from a pair images, we have four image coordinates (u, v, u', v') for each point, all functions of the three world

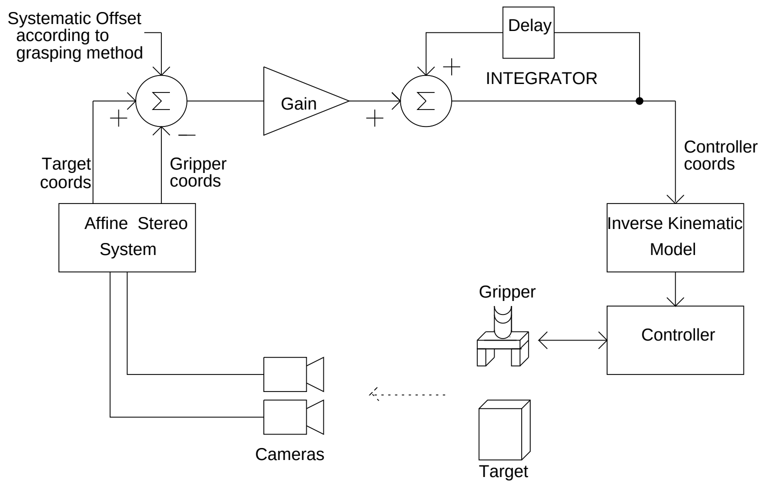


Figure 8: The control structure of the system showing the use of visual feedback.

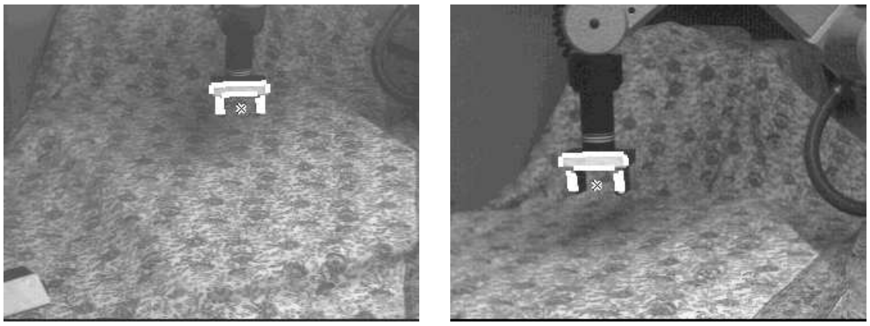


Figure 9: A stereo pair showing the robot gripper at one of the four reference points used for calibration. Active contours are overlaid in white.

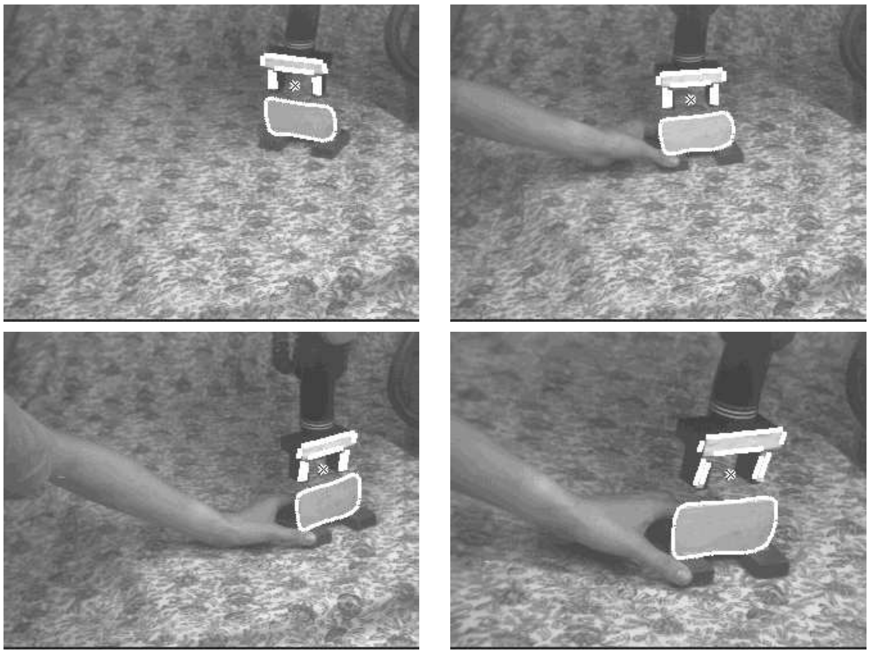


Figure 10: The robot is tracking its quarry, guided by the position and orientation of the target contour (view through left camera). On the target surface is an *affine snake* — an affine tracker obtained by 'exploding' a B-spline snake from the centre of the object. A slight offset has been introduced into the control loop to cause the gripper to hover above it. Last frame: one of the cameras has been rotated and zoomed, but the system continues to operate successfully with visual feedback.

coordinates (x, y, z):

$$\begin{bmatrix} u \\ v \\ u' \\ v' \end{bmatrix} = \mathbf{Q} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{3}$$

$\mathbf{Q}$ is a 4×4 matrix, formed from the p_{ij} coefficients of a pair of cameras. To calibrate the system it is necessary to observe a minimum of four non-coplanar *reference points*, yielding sixteen simultaneous linear equations from which $\mathbf{Q}$ may be found.

It can be shown in practice that *calibration is better conditioned* than with full-perspective stereo, because the system has fewer parameters and is amenable to solution by linear techniques (full projective stereo can be represented by 24 linear coefficients but there are nonlinear constraints on those coefficients. With noisy image data, greater accuracy may be obtained by observing more than four points.

Once the coefficients are known, world coordinates can be obtained by inverting (3), using a least-squares method to resolve the redundant information. Errors in calibration will manifest themselves as a linear distortion of the perceived coordinate frame.

Note:

1. It is *not* essential to calibrate a stereo vision system to obtain useful 3-D information about the world. Instead, four of the points observed may be given arbitrary world coordinates (such as $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$ and $(1, 0, 0)$). The appropriate solution for $\mathbf{Q}$ will define a coordinate frame which is an arbitrary 3-D affine transformation of the 'true' Cartesian frame, preserving affine shape properties such as ratios of lengths and areas, collinearity and coplanarity. This is in accordance with Koenderink and van Doorn's *Affine Structure-from-Motion Theorem* [13].

2. In hand–eye applications, it might instead be convenient to calibrate the vision system in the coordinate space in which the manipulator is controlled (assuming this maps approximately linearly to Cartesian coordinates). This can be done by getting the robot manipulator to move to four points in its workspace.

3. The integration of information from more than two cameras to help avoid problems due to occlusion is easily accommodated in this frame-

work. Each view generates two additional linear equations in 3 which can be optimally combined.

4.2 Visual Feedback for Hand–Eye Coordination

Affine stereo is a simplified stereo vision formulation that is very easily calibrated. Conversely, it is of rather low accuracy. Nevertheless, it gives reliable *qualitative* information about the relative positions of points and can, of course, indicate when they are in precisely the same place and when two surfaces are at the same orientation. (The disparity of the two points will be equal. Similarly the disparity gradient (affine transformation) for two visible surfaces will be equal when they are at the same depth and orienation.) We therefore use a feedback control mechanism to help to guide the gripper to the target, using affine stereo to compute the relative position and orientation of their respective tracked surfaces.

Since the reference points used to self-calibrate are specified in the *controller's* coordinate space, linear errors in the kinematic model are effectively bypassed. The system must still cope with any nonlinearities in control, as well as those caused by strong perspective effects. An integral feedback control architecture is employed. The feedback term is the difference between the vectors that describe the position and orientation of the target and gripper, as seen by the vision system. This term is integrated by summing at each time step, and the resulting vector used to position the robot.

The manipulator moves in discrete steps, through a distance proportional to the difference between the gripper's perceived coordinates and those of the target plane. The gain is below unity to prevent ringing or instability, even when the vision system is miscalibrated. This process is repeated until the perceived positions of the two coincide (or, for grasping, we can introduce a fixed offset).

4.3 Implementation and Experiments

Implementation

When the system is started up, it begins by opening and closing the jaws of the gripper. By observing the image difference, it is able to locate the gripper and set up a pair of affine trackers as instances of a 2-D template. The trackers will then follow the gripper's movements continuously. The robot moves to four preset points to calibrate the system in terms of the controller's coordinate space.

The orientation of the gripper of a 5-DOF manipulator is constrained by its 'missing' axis, and this constraint changes continuously as it moves. To avoid this problem, the present implementation keeps the gripper vertical, reducing the number of degrees of freedom to four. Its orientation is then described by a single *roll angle*. It is assumed that the target surface is also vertical.

By introducing modifications and offsets to the feedback mechanism (which would otherwise try to superimpose the gripper and the target), two 'behaviours' have been implemented. The *grasping behaviour* causes the gripper to approach the object from above (to avoid collisions) with the gripper turned through an angle of 90 degrees, to grasp it normal to its target surface. The *tracking behaviour* causes it to follow the target continuously, hovering a few centimetres above it (figure 10).

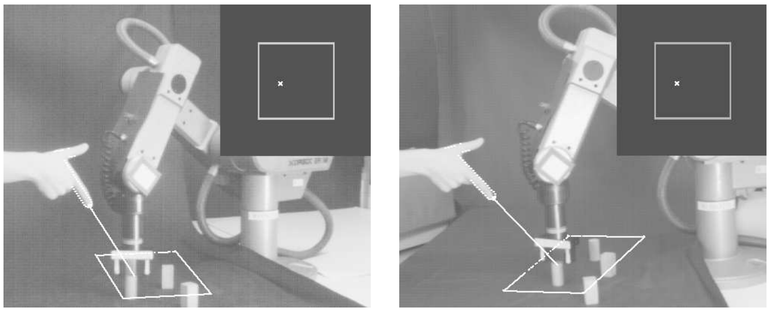
Results

Without feedback control, the robot locates its target only approximately (typically to within 5cm in a 50cm workspace). With a feedback gain of 0.75 the gripper converges on its target in three or four control iterations. If the system is not disturbed it will take a straight-line path. The system has so far demonstrated its robustness by continuing to track and grasp objects despite:

Kinematic errors. Linear offsets or scalings of the controller's coordinate system are absorbed by the self-calibration process with complete transparency. Slight non-linear distortions to the kinematics are corrected for by the visual feedback loop, though large errors introduce a risk of ringing and instability unless the gain is reduced.

Camera disturbances. The system continues to function when its cameras are subjected to small translations, rotations and zooms, even after it has self-calibrated. Large disturbances to camera geometry cause the gripper to take a curved path towards the target, and require more control iterations to get there.

Strong perspective. The condition of weak perspective throughout the workspace does not seem to be essential for image-based control and the system can function when the cameras are as close as 1.5 metres (the robot's reach is a little under 1 metre). However the feedback gain must be reduced to below 0.5, or the system will overshoot on motions towards the cameras.



(a) (b)

Figure 11: Gestural control of robot position for grasping, seen in stereo.

Figure 10 shows four frames from a tracking sequence (all taken through the same camera). The cameras are about two metres from the workspace. Tracking of position and orientation is maintained even when one of the cameras is rotated about its optical axis and zoomed.

Figure 11 shows the robot grasping an object under gestural control.

5 Discussion and Conclusions

In all three stages of the system we have exploited *uncalibrated* stereo. This is attractive because it makes the system easy to set up and robust to disturbances in camera positions. Another advantage of the uncalibrated approach is that it can easily be extended to trinocular and multi-camera systems.

Uncalibrated stereo algorithms have been described for the following tasks:

- To allow the operator to specify what is to be grasped
- Grouping edges into planar surfaces of the object
- Control of the robot as it aligns the gripper with a planar surface and grasps the object.

Pointing can be used successfully to specify positions on a two-dimensional workspace for a robot manipulator. The system is simple and intuitive for an operator to use, and requires practically no training. There is no need for camera calibration because no 3D calculations are involved. By tracking at least four points on the plane it could be made invariant to camera movement.

Uncalibrated stereo vision can also be used to detecting planar regions in a scene and for guiding the robot manipulator to the planar facet to be grasped. An important component of the hand-eye coordination system is affine stereo. Affine stereo provides a simple and robust interpretation of image position and disparity that degrades gracefully when the cameras are disturbed. Also by defining the working coordinate system in terms of the robot's abilities, linear errors in its kinematics are bypassed. The remaining non-linearities can be handled using image-based control and feedback. We have shown that this can be achieved cheaply and effectively using active contours to track planar features on the gripper and target.

For the complete man-machine interface, a wide field of view is necessary for the pointing and the robot manouevre stages, but for the segmentation stage, precision is of more importance. These conflicting requirements of wide field of view and high precision could be achieved by replacing the current static pair of cameras with a *stereo-head* [15]. The performance of the system is considerably improved by ensuring that the hand and gripper contour trackers operating on the left and right image are coupled. Epipolar constraints improve the tracker's performance in the presence of clutter.

Acknowledgements

We gratefully acknowledge discussions with Dr Gabriel Hamid. This work was supported by EPSRC research grant GR/K08635. The robot manipulator was donated by the Olivetti Research Laboratory, Cambridge.

References

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

[2] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *Int. Journal of Computer Vision*, 11(2):127–146, 1993.

[3] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *Int. Journal of Computer Vision*, 9(2):83–112, 1992.

[4] R. Cipolla and N.J. Hollinghurst. Visual robot guidance from uncalibrated stereo. In C.M. Brown and D. Terzopoulos, editors, *Realtime Computer Vision*, pages 169–187, CUP, 1994.

[5] R. Cipolla and N.J. Hollinghurst. Human–robot interface by pointing with uncalibrated stereo vision. *Image and Vision Computing*, 14(2):171–178, 1996.

[6] R. Cipolla, Y. Okamoto, and Y. Kuno. Robust structure from motion using motion parallax. In *Proc. 4th Int. Conf. on Computer Vision*, Berlin, pages 374–382, 1993.

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

[8] O.D. 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.

[9] J. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics: Principles and Practice.* Addison-Wesley, 1990.

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

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

[12] M. Kass, A. Witkin, and D. Terzopoulus. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.

[13] J.J. Koenderink and A.J. van Doorn. Affine structure from motion. *J. Opt. Soc. America*, 8(2):377–385, 1991.

[14] J.L. Mundy and A. Zissermann. editors, *Geometric Invariance in Computer Vision.* MIT Press, 1992.

[15] D. Murray and I. Reid. Active tracking of foveated feature clusters using affine structure. *International Journal of Computer Vision*, 18(1):1–20, 1996.

[16] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(7), pages 629–639, 1990.

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

[18] L. Quan and R. Mohr. Towards structure from motion for linear features through reference points. In *Proc. IEEE workshop on visual motion*, pages 249–254, New Jersey, 1991.

[19] M. Rygol, S. Pollard, and C. Brown. A multiprocessor 3D vision system for pick-and-place. In *Proc. British Machine Vision Conference*, pages 169–174, 1990.

[20] D. Sinclair and A. Blake. Quantitative planar region detection. *International Journal of Computer Vision*, 18(1):77, 1996.

[21] R.Y. Tsai. 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*, RA-3(4):323–344, 1987.

[22] B. Wirtz and C. Maggioni. Imageglove: a novel way to control virtual environments. In *Proc. Virtual Reality Systems*, New York, 1993.