Visual robot guidance from uncalibrated stereo

Roberto Cipolla and Nicholas J. Hollinghurst

Department of Engineering,
University of Cambridge,
Cambridge CB2 1PZ, England.

1 Introduction

When humans grasp and manipulate objects, they almost invariably do so with the aid of vision. Visual information is used to locate and identify things, and to decide how (and if) they should be grasped. Visual feedback helps us guide our hands around obstacles and align them accurately with their goal. *Hand–Eye Coordination* gives us a flexibility and dexterity of movement that no machine can match.

Most vision systems for robotics usually need to be *calibrated*. Camera geometry — the focal length, principal point and aspect ratio of each camera [17], the relative position and orientation of the cameras (epipolar geometry) [15] and their relation to the robot coordinate system [16] — must be measured to a high degree of accuracy. A well-calibrated stereo rig can accurately determine the position and shape of things to be grasped in all three dimensions [14]. However, if calibration is erroneous or the cameras are disturbed, the system will usually fail gracelessly.

An alternative approach in hand–eye applications where a manipulator moves to a visually-specified target, is to use visual feedback to match manipulator and 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) [19]. Visual feedback alone without exploiting or learning the relationship between the robot kinematics and the stereo cameras, however, can lead to inefficient motions [10].

In this paper we describe a system that combines stereo vision with a robotic manipulator to enable it to efficiently locate and reach simple unmodelled objects in an unstructured environment. The system is initially uncalibrated; it "calibrates" itself automatically by tracking the gripper during four deliberate exploratory movements in its workspace and is able to operate successfully 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.

The system exploits an *affine stereo* algorithm – a simple but robust approximation to the geometry of stereo vision – (described in section 2) which, though of modest accuracy, requires minimal calibration and can tolerate small camera movements. We show that, in some circumstances, this simplified camera model is less sensitive to image measurement error since it avoids computing parameters required in the full perspective stereo which are inherently ill-conditioned [4]. 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. This is done with a form of *active contour model* resembling a B-spline snake [3] but constrained to deform only affinely (described in section 3) to produce a more reliable tracker which is less easily confused by background contours or partial occlusion. Inevitable errors in aligning the gripper and target object position and orientation are corrected by an image-based feedback mechanism (section 4). Preliminary results of a realtime implementation are presented (section 5) and show the system to be remarkably immune to unexpected movements of the cameras and focal lengths even after the initial self-calibration.

2 Affine Stereo

2.1 Perspective and projective camera models

The full perspective transformation between world and image coordinates is conventionally analysed using the *pinhole camera* model, in which image-plane coordinates (u, v) are ratios of world coordinates (x_c, y_c, z_c) in a camera-centred frame, thus: $u = fx_c/z_c$, $v = fy_c/z_c$. The relation between the camera-centred and some other world coordinate frame consists of *rotation* ($\mathbf{R}$) and *translation* ($\mathbf{t}$) components representing the camera's orientation and position. Using homogeneous coordinates (with scale factor s for convenience),

$$\begin{bmatrix} su \\ sv \\ sf \end{bmatrix} = \begin{bmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \mathrm{r}_{13} & \mathrm{t}_1 \\ \mathrm{r}_{21} & \mathrm{r}_{22} & \mathrm{r}_{23} & \mathrm{t}_2 \\ \mathrm{r}_{31} & \mathrm{r}_{32} & \mathrm{r}_{33} & \mathrm{t}_3 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}. \tag{1}$$

The relation between image plane coordinates (u, v) and pixel addresses (X, Y) can be modelled by an affine transformation (to represent offsets, scaling and shearing). Combining this with (1) yields a general 3D to 2D

projection, with 11 degrees of freedom:

$$\begin{bmatrix} \sigma X \\ \sigma Y \\ \sigma \end{bmatrix} = \begin{bmatrix} \mathrm{m}_{11} & \mathrm{m}_{12} & \mathrm{m}_{13} & \mathrm{m}_{14} \\ \mathrm{m}_{21} & \mathrm{m}_{22} & \mathrm{m}_{23} & \mathrm{m}_{24} \\ \mathrm{m}_{31} & \mathrm{m}_{32} & \mathrm{m}_{33} & \mathrm{m}_{34} \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}. \tag{2}$$

This is the usual camera model for many stereo vision systems. Although it neglects effects such as lens distortion which are significant in some high-accuracy applications [15], it correctly predicts image distortion due to perspective effects e.g. parallel 3D lines projecting to intersect a vanishing point and the cross ratio (not ratio) of lengths is invariant to this transformation.

2.2 Weak perspective and affine camera models

Consider a camera viewing a compact scene of interest from distance h. For convenience, we can translate the world coordinate system so that the scene lies close to the world origin. The component of $\mathbf{t}$ along the optical axis, t_3, will then equal h. As distance increases relative to the extent of the scene, sf/h will tend to unity for all points, and equation (1) becomes:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{f}{h} \begin{bmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \mathrm{r}_{13} & \mathrm{t}_1 \\ \mathrm{r}_{21} & \mathrm{r}_{22} & \mathrm{r}_{23} & \mathrm{t}_2 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}. \tag{3}$$

This formulation assumes that images are not distorted by variations in depth, and is known as *weak perspective* [13]. It is equivalent to orthographic projection scaled by a factor inversely proportional to the average depth, h. It can be shown that this assumption results in an error which is, at worst, $\Delta h/h$ times the scene's image size.

The entire projection, again incorporating scaling and shearing of pixel coordinates, may now be written very simply as a linear mapping:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mathrm{p}_{11} & \mathrm{p}_{12} & \mathrm{p}_{13} & \mathrm{p}_{14} \\ \mathrm{p}_{21} & \mathrm{p}_{22} & \mathrm{p}_{23} & \mathrm{p}_{24} \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}. \tag{4}$$

The eight coefficients p_{ij} efficiently represent all intrinsic and extrinsic camera parameters [15]. This simple approximation to the projection transformation — the affine camera [11] — will be used as the camera model throughout the paper. Its advantages will become clear later when it leads to efficient calibration and reduced sensitivity to image measurement error. Note that parallel lines project to parallel lines in the image and ratios of lengths and areas are invariant to the transformation.

2.3 Motion of planar objects under weak perspective

There are many situations in computer vision where an object must be *tracked* as it moves across a view. Here we consider the simple, but not uncommon, case where the object is small and has planar faces.

We can define a coordinate system centred about the object face itself so that it lies within the xy plane. If the object is small compared to the camera distance, we again have weak perspective, and a special case of (4):

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{5}$$

We see that the transformation from a plane in the world to the image plane is a 2D *affine* translation. As the camera moves relative to the object, parameters a_{ij} will change and the image will undergo translation, rotation, change in scale (divergence) and deformation, but remain *affine-invariant* [8, 2] (figure 1).

This is a powerful constraint that can be exploited when tracking a planar object. It tells us that the shape of the image will deform only affinely as the object moves, and that there will exist an affine transformation between any two views of the same plane.

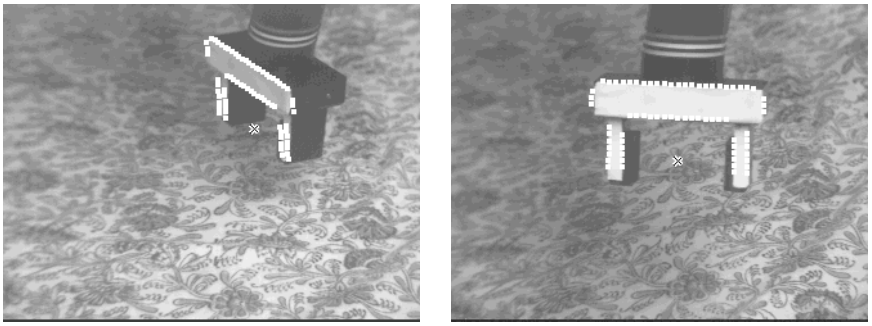


Figure 1: The gripper being tracked as it translates and rotates under *weak perspective*. The origin and sampling points of the tracker are shown in white. The front of the gripper is approximately planar, and its image shape distorts affinely as it moves under *weak perspective*.

2.4 The affine stereo formulation

In stereo vision two calibrated views of a scene from known viewpoints allow the Euclidean reconstruction of the scene. In the following two uncalibrated views under weak perspective projection are used to recover relative 3D positions and surface orientations.

Recovery of relative position from image disparity

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 (X, Y, X', Y') for each point, all linear functions of the three world coordinates (x_w, y_w, z_w):

$$\begin{bmatrix} X \\ Y \\ X' \\ Y' \end{bmatrix} = \mathbf{Q} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}, \tag{6}$$

where $\mathbf{Q}$ is a 4×4 matrix formed from the p_{ij} coefficients of (4) for the two 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. 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 (6), 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* [9].

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 framework. Each view generates two additional linear equations in 6 which can be optimally combined.

Recovery of surface orientation from disparity gradients

Under weak perspective any two views of the same planar surface will be related by an affine transformation that maps one image to the other. This consists of a pure 2D translation encoding the displacement of the centroid and a 2D tensor – the disparity gradient tensor – which represents the distortion in image shape. This transformation can be used to recover surface orientation [2]. Surface orientation in space is most conveniently represented by a surface normal vector $\mathbf{n}$. We can obtain it by the vector product of two non-collinear vectors in the plane which can of course be obtained from three pairs of image points. There is, however, no redundancy in the data and this method would be sensitive to image measurement error. A better approach is to exploit all the information in available in the affine transform (disparity field).

Consider the standard unit vectors $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ in one image and suppose they were the projections of some vectors on the object surface. If the linear mapping between images is represented by a 2×3 matrix $\mathbf{A}$, then the first two columns of $\mathbf{A}$ itself will be the corresponding vectors in the other image. As the centroid of the plane will map to both image centroids, we can easily use it and the above pairs of vectors to find three points in space on the plane (by inverting (6)) and hence the surface orientation.

3 Tracking using affine active contours

An *active contour* (or *'snake'*) [7] is a curve defined in the image plane that moves and deforms according to various 'forces'. These include *external forces*, which depend on local image properties and are used to guide the active contour towards the image features, and *internal forces* which depend on the contour shape and are used to enforce smoothness. Typically, a snake will be attracted to maxima of image intensity gradient, and used to track the edges of a moving object.

3.1 Anatomy

Our model-based trackers are a novel form of active contour. They resemble B-spline snakes [3] but consist of (in the order of 100) discrete sampling points, rather than a smooth curve [6]. We use them to track planar surfaces bounded by contours, on the robot gripper and the object to be grasped. Pairs of trackers operate independently in the two stereo views. The trackers can deform only affinely, to track planes viewed under weak perspective [1]. This constraint leads to a more efficient and reliable tracker than a B-spline snake, that is less easily confused by background contours or partial occlusion.

Each tracker is a 2D model of the image shape it is tracking, with sampling points at regular intervals around the edge. At each sampling point there is a *local edge-finder* which measures the offset between modelled and actual edge positions in the image, by searching for the maximum of gradient along a short line segment [5]. Due to the so-called *aperture problem* [18], only the normal component of this offset can be recovered at any point (figure 2).

The positions of the sampling points are expressed in affine coordinates, and their image positions depend upon the tracker's *local origin* and two *basis vectors*. These are described by six parameters, which change over time as the object is tracked. The contour tangent directions at each point are also described in terms of the basis vectors.

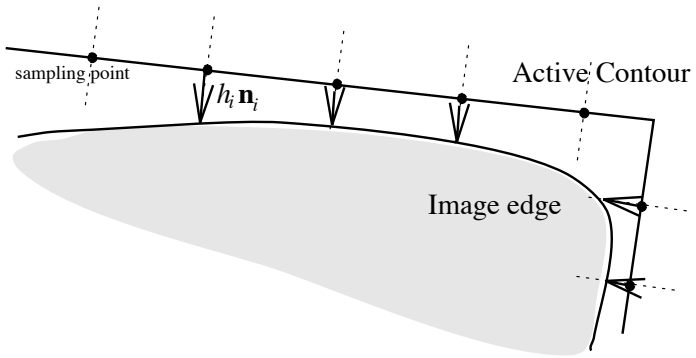


Figure 2: An active contour: The image is sampled in segments normal to the predicted contour (dotted lines) to search for the maximal gradient. The offsets between predicted and actual edges (arrows) are combined globally to guide the active contour towards the image edge.

3.2 Algorithm

At each time-step the tracker moves and deforms to minimise the sum of squares of offsets between model and image edges (h_i). In our implementation this is done in two stages. First the optimal translation is found, then the deformation, rotation, scale (divergence) components are calculated. Splitting the task into these two stages was found to increase stability, as fewer parameters were being estimated at once. To find the optimal translation $\mathbf{u}$ to account for normal offset h_i at each sampling point whose image normal direction is $\mathbf{n}_i$, we solve the following equation:

$$h_i = \mathbf{n}_i \cdot \mathbf{u} + \epsilon_i. \tag{7}$$

ϵ_i is the error term, and we solve the whole system of equations using a least-squares method to minimise $\sum \epsilon_i^2$.

Once the translation has been calculated, the other components are estimated. It is assumed that the distortion is centred about the tracker's local origin (normally its centroid, to optimally decouple it from translation). The effects of translation ($\mathbf{n}_i \cdot \mathbf{u}$) are subtracted from each normal offset, leaving a residual offset. We can then find the matrix $\mathbf{A}$ that maps image coordinates to displacement:

$$(h_i - \mathbf{n}_i \cdot \mathbf{u}) = \mathbf{n}_i \cdot (\mathbf{A}\mathbf{p}_i) + \epsilon'_i, \tag{8}$$

where $\mathbf{p}_i$ is the sampling point's position relative to the local origin and ϵ'_i is again the error term to be minimised.

In practice this formulation can lead to problems when the tracked surface moves whilst partially obscured (often, a tracker will catch on an occluding edge and become 'squashed' as it passes in front of the surface). It can also be unstable and sensitive to noise when the tracker is long and thin. We therefore use a simplified approximation to this equation that ignores the aperture problem (equating the normal component with the whole displacement):

$$(h_i - \mathbf{n}_i \cdot \mathbf{u})\mathbf{n}_i = \mathbf{A}\mathbf{p}_i + \mathbf{e}_i. \tag{9}$$

$\mathbf{e}_i$ is an error vector, and our implementation solves the equations to minimise $\sum |\mathbf{e}_i|^2$. This produces a more stable tracker that, although sluggish to deform, is well suited to those practical tracking tasks where motion is dominated by the translation component. The tracker positions are updated from $\mathbf{u}$ and $\mathbf{A}$ using a real time first-order predictive filter. This enhances performance when tracking fast-moving objects.

4 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. 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.

We take an iterative approach, based upon *relative* positions. The manipulator moves in discrete steps; each motion is proportional to the difference

between the gripper's perceived position and orientation, and those of the target plane. This is equivalent to an *integral* control architecture, in which the error term is summed at each time step (see figure 3).

The gain is set below unity to prevent instability, even when the vision system is miscalibrated. The process repeats until the perceived coordinates of the gripper coincide with those of the target. Alternatively, to implement a particular grasping strategy, an offset can be introduced into the control loop to specify the final pose of the gripper relative to the target plane. Irrespective of the accuracy of stereo and hand-eye calibration visual feedback will ensure that target and gripper are aligned.

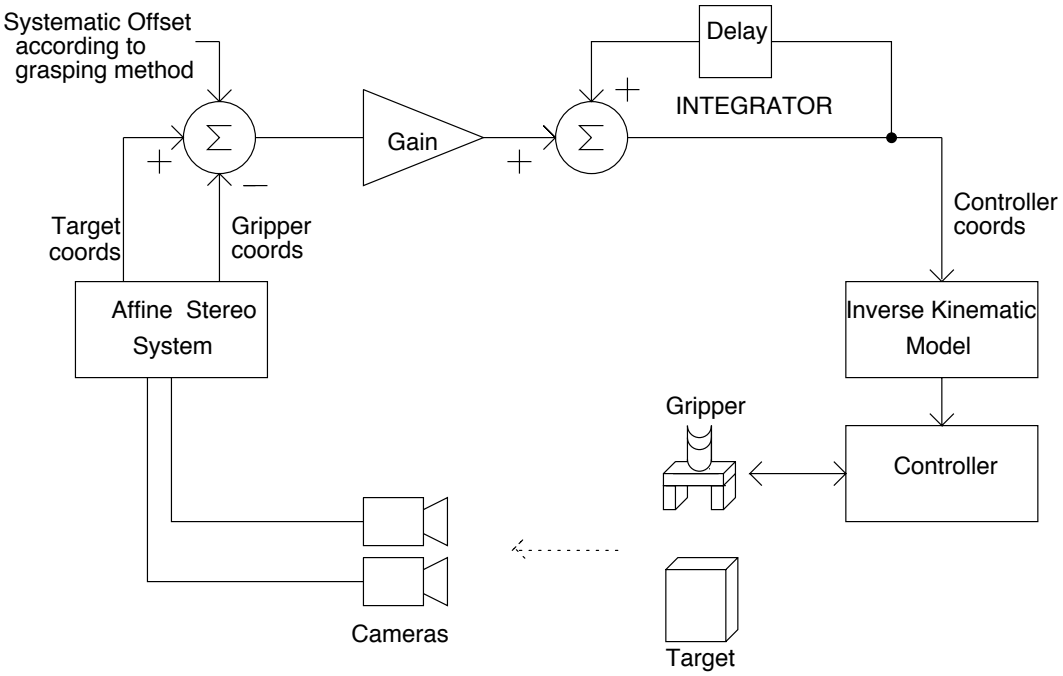


Figure 3: The control structure of the system, showing the use of visual feedback.

5 Implementation and Experiments

5.1 Equipment

The system was implemented on a Sun SPARCstation 10 with a Data Cell S2200 frame grabber. The manipulator is a Scorbot ER-7 robot arm, which has 5 degrees of freedom and a parallel-jawed gripper. The robot has its own 68000-based controller which implements the low-level control loop and provides a Cartesian kinematic model. Images are obtained from two inex-

pensive CCD cameras placed 1m–3m from the robot's workspace. The angle between the cameras is in the range of 15–30 degrees (figure 4).

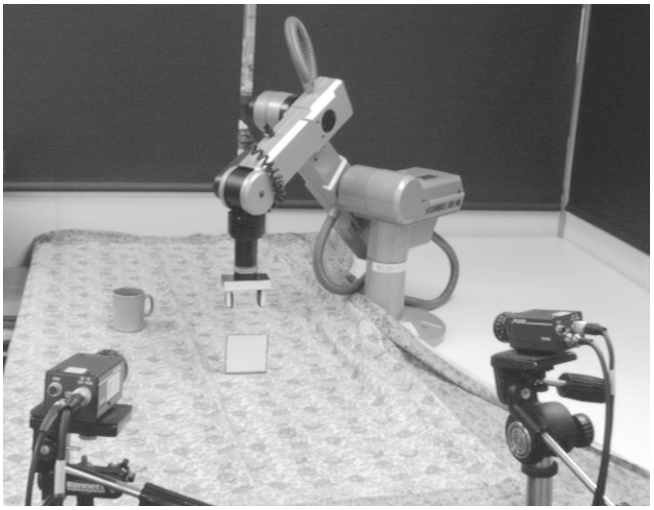


Figure 4: The experimental setup. Uncalibrated stereo cameras viewing a robot gripper and target object.

5.2 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 hand-made 2D template. The trackers will then follow the gripper's movements continuously. Stereo tracking can be implemented on the Sun at over 10 Hz. The robot moves to four preset points to calibrate the system in terms of the controller's coordinate space.

A target object is found by similar means — observing the image changes when it is placed in the manipulator's workspace. Alternatively it may be selected from a monitor screen using the mouse. There is no pre-defined model of the target shape, so a pair of 'expanding' B-spline snakes [2] are used to locate the contours delimiting the target surface in each of the images. The snakes are then converted to a pair of affine trackers. The target surface is then tracked, to compensate for unexpected motions of either the target or the two cameras.

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 *tracking behaviour* causes it to follow the target continuously, hovering a few centimetres above it (figure 5). The *grasping behaviour* causes the gripper to approach the target from above (to

avoid collisions) with the gripper turned through an angle of 90 degrees, to grasp it normal to its visible surface (figure 6).

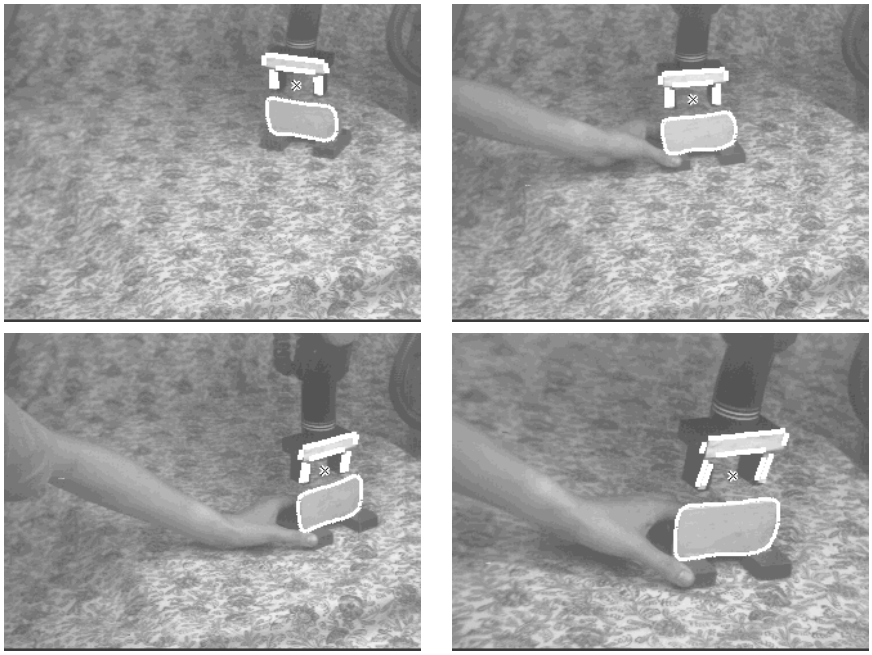


Figure 5: The robot is tracking its quarry, guided by the position and orientation of the target contour (view through left camera). On the target surface is an *affine snake* — an affine tracker obtained by 'expanding' a B-spline snake from the centre of the object. A slight offset has been introduced 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.

5.3 Results

Without feedback control, the robot locates its target only approximately (typically to within 5cm in a 50cm workspace) reflecting the approximate nature of affine stereo and calibration from only four points. 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 nonlinear distortions to the kinematics are corrected for

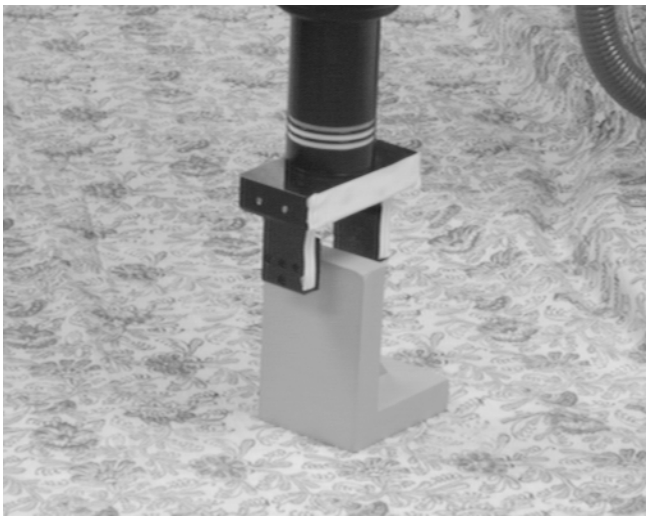


Figure 6: Affine stereo and visual feedback used to grasp a planar face.

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 (e.g. 20cm), rotations (e.g. 30 degrees) and zooms (e.g. 200% change in focal length), 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 robot's 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 or the system will overshoot on motions towards the cameras.

Figure 5 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 5, bottom right).

6 Future Developments

The current system is based upon matching the positions and orientations of two *planes* – the gripper and target. Since the gripper and target planes must

be visible in both images at all times, neither can rotate through more than about 120 degrees. We intend to develop the system to track the gripper using a three-dimensional rigid model [6], drawing information from both images simultaneously. We also aim to identify and track more than one of the surfaces of the object to be grasped, both for 3-D tracking and also for analysis of its 3D shape.

We plan to equip such a system with a grasp planner that uses the relative positions, sizes and orientations of the visible surfaces on the object, to direct the robot to grasp unmodelled objects in a suitable way.

7 Conclusion

An important component of the system presented in this paper has been affine stereo. The appendix contains a quantitative comparison with perspective stereo. Affine stereo provides a simple and robust interpretation of image position and disparity that degrades gracefully when cameras are disturbed. Calibration is not only easier (fewer parameters and amenable to linear techniques) but also less sensitive to small perspective effects. It is suitable for uncalibrated and self-calibrating systems and therefore the preferred stereo method for our visual servoing application.

By defining the working coordinate system in terms of the robot's abilities, linear errors in its kinematics are bypassed. The remaining nonlinearities can be handled using image-based control and feedback. We have shown that this can be achieved cheaply and effectively using a novel form of active contour to track planar features on the gripper and target.

Such a system has been implemented and found to be highly robust, without unduly sacrificing performance (in terms of speed to converge on the target).

Appendix: Comparison of full-perspective and affine stereo

Correspondence and the epipolar constraint

In the affine stereo formulation it was assumed that two sets of image coordinates were available for each world point. The task of identifying pairs of image features which correspond to the same point in space is known as the *correspondence problem*.

The image coordinates of a world feature in two images are not independent, but related by an *epipolar constraint*. Consider the family of planes passing through the optical centre of each camera. These project to a family

of *epipolar lines* in each image. If a feature lies upon a particular line in the left image, the corresponding feature must lie upon the line in the right image, which is the projection of the same plane. The constraint reflects the redundancy inherent in deriving four image coordinates from points in a three-dimensional world. Most correspondence algorithms exploit this constraint, which reduces the search for matching features to a single dimension, and identifying it is an important aspect of any calibration scheme.

In affine stereo, the epipolar planes are considered to be parallel, and the constraint takes the form of a single linear relation among the four image coordinates. With the full perspective model, the lines need not be parallel, and converge to a point called the *epipole* (the projection of one camera centre on the other camera's image plane). The constraint may be obtained from calibration data, for instance by rearranging the model to predict one image coordinate as a function of the other three.

Figure 7 compares the epipolar line structure predicted by both affine and full perspective stereo models (after calibration using linear least squares). In this setup, in which the camera distance is about 2 metres, both models give similar epipolar accuracy. Furthermore, the affine model can predict epipolar lines using just 4 reference points; perspective stereo requires a minimum of 6.

Accuracy of reconstruction

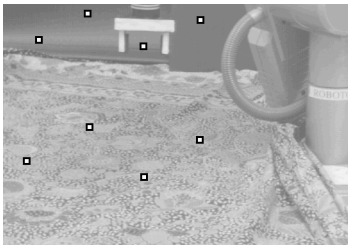
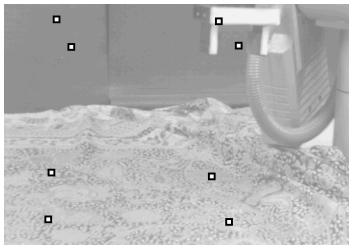
To compare affine and full perspective stereo, we performed a series of numerical simulations, measuring their ability to estimate the relative positions of points within a workspace, viewed by a pair of pinhole cameras.[1]

Under ideal conditions: Without noise or other disturbances, perspective stereo estimates absolute and relative positions with complete accuracy. At close range affine stereo performs poorly, but the error decreases in inverse proportion to camera distance (figure 8).

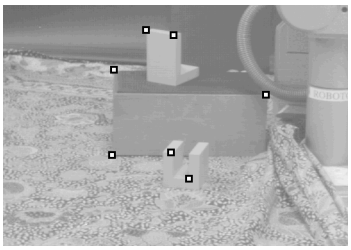
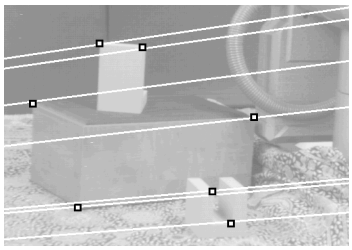
Accuracy is also somewhat dependent on the number and configuration of the reference points used in calibration, and there is a limited improvement as the unit cube is sampled more regularly.

With noisy calibration data: Adding 1% Gaussian noise to the image coordinates of the reference points causes both systems to lose accuracy. Perspective stereo is more sensitive to noise because of its nonlinearity and greater degrees of freedom, and is *less* accurate than the affine stereo

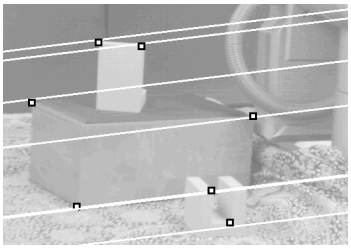
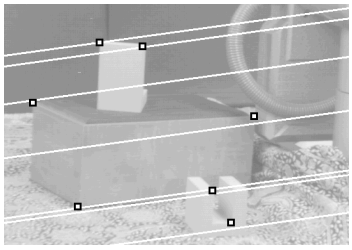
[1]Reference and test points are confined to a unit cube centred about the origin. There are 6 reference points within the unit cube. Test points are distributed uniformly within the cube. The cameras face the origin from a distance of 3–24 units, angled 20° apart (their focal length is proportional to distance, to normalize image size).

(a,b) Two views of 8 reference points defined by the robot

(c,d) Selected points in the left image, and epipolar lines estimated by the perspective camera model after calibration with 8 points

(e,f) Epipolar lines estimated by affine camera model after calibration with 8 points and with 4 points

Figure 7: Estimation of epipolar lines. Although it considers the epipolar lines to be parallel, the affine camera model (e) is almost as accurate as perspective in this experiment (RMS perpendicular error 4.1 pixels). Even with only 4 reference points, it produces a reasonable solution (f) from which stereo correspondence could be performed (RMS error 6.2 pixels).

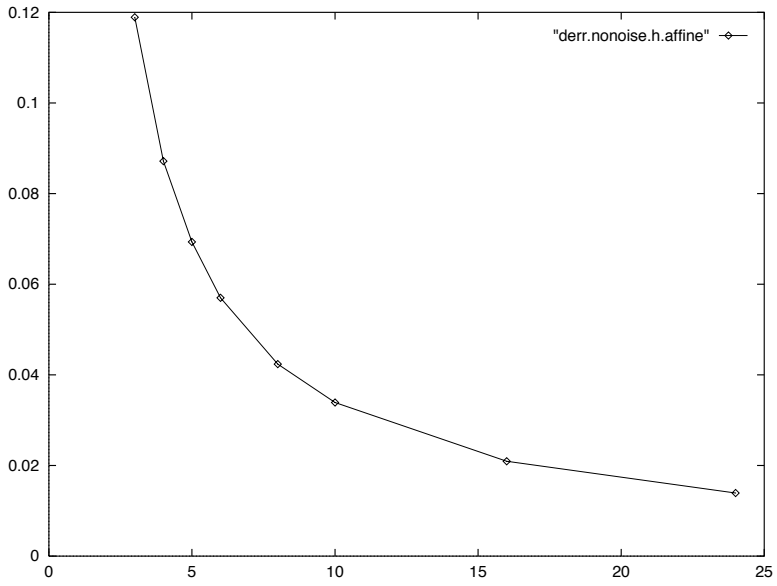


Figure 8: RMS relative positioning error (for random point pairs in the unit cube) as a function of camera distance. The error is due to the approximate nature of the affine stereo model and drops as camera distance increases.

approximation at large viewing distances (figure 9). (viewing a larger number of reference points reduces the effects of noise and restores the accuracy of perspective stereo).

Camera movements after calibration: In a laboratory or industrial environment it is possible for cameras to be disturbed from time to time and subject to small rotations and translations. If this happens after calibration, it will give rise to a corresponding error in stereo reconstruction.

Table 1 shows the average change in perceived relative position when one camera is rotated or translated a small distance around/along each principle axis. The two systems degrade comparably with small movements, the worst of which is rotation about the optical axis. Perspective stereo is more sensitive to larger movements, and to rotations and translations in the epipolar plane (in which a small error can induce large changes of perceived depth), because it distorts nonlinearly.

With noisy image coordinates: When gaussian noise is added to the image coordinates of the points whose relative position is to be estimated (after accurate calibration), the effect is comparable on both systems, and their performance converges at large camera distance (figure 10).

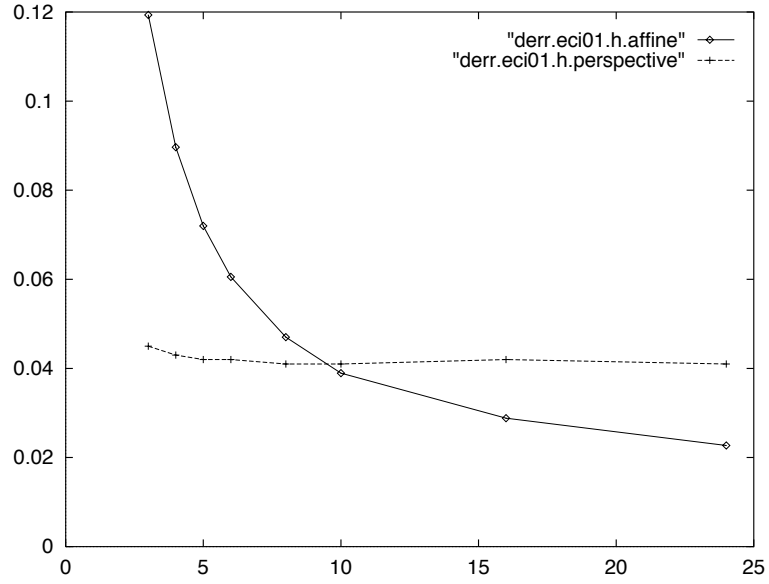


Figure 9: RMS positioning error as a function of camera distance, after calibration with noisy reference point images (standard deviation 1% of image size). The error suffered by the perspective model (dotted) is comparable in magnitude to the affine stereo systematic error.

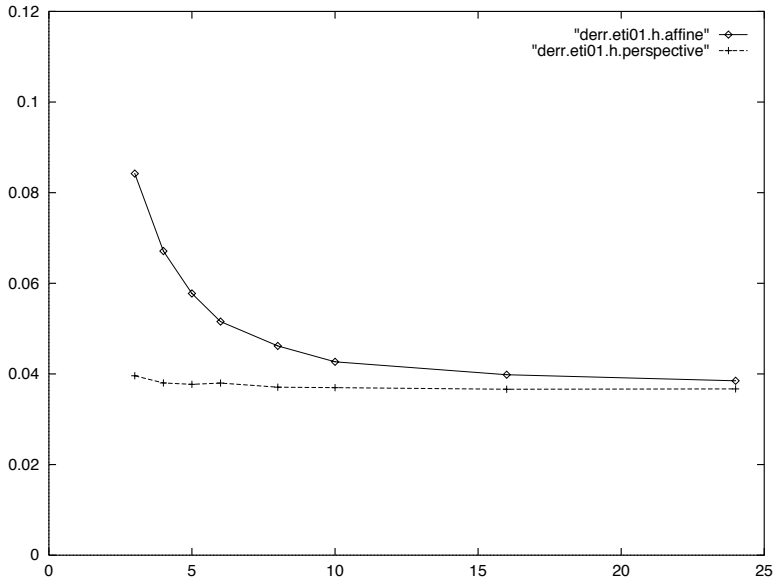


Figure 10: RMS relative positioning error from noisy images (standard deviation 1% image size) of world points after *accurate* calibration with 27 reference points. The two models converge for camera distances above ~ 10 units.

Disturbance	Change (Affine)	Change(Perspective)
xy (cyclic) rotation 1^o	.0214	.0214
xz (epipolar) rotation 1^o	.0007	.0468
yz (vertical) rotation 1^o	.0006	.0049
xy (cyclic) rotation 5^o	.1069	.1068
xz (epipolar) rotation 5^o	.0095	.1867
yz (vertical) rotation 5^o	.0056	.0769
x (epipolar) translation 0.1	.0119	.0207
y (vertical) translation 0.1	.0020	.0007
z (distance) translation 0.1	.0119	.0119
x (epipolar) translation 0.5	.0596	.1168
y (vertical) translation 0.5	.0102	.0139
z (distance) translation 0.5	.0574	.0572

Table 1: RMS *change* to relative position estimates of world points, caused by disturbing one of the cameras after calibration

References

[1] A. Blake, R. Curwen, and A. Zisserman. Affine-invariant contour tracking with automatic control of spatiotemporal scale. In *Proc. 4th Int. Conf. on Computer Vision*, pages 66–75, 1993.

[2] R. Cipolla and A. Blake. Surface orientation and time to contact from image divergence and deformation. In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision*, pages 187–202. Springer–Verlag, 1992.

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

[5] R. Curwen and A. Blake. Dynamic contours: real-time active splines. In A. Blake and A. Yuille, editors, *Active Vision*, pages 39–58. MIT Press, 1992.

[6] C. Harris. Tracking with rigid models. In A. Blake and A. Yuille, editors, *Active Vision*, pages 59–74. MIT Press, 1992.

[7] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. In *Proc. 1st Int. Conf. on Computer Vision*, pages 259–268, 1987.

[8] J.J. Koenderink. Optic flow. *Vision Research*, 26(1):161–179, 1986.

[9] J.J. Koenderink and A.J. van Doorn. Affine structure from motion. *J. Opt. Soc. America*, pages 377–385, 1991.

[10] B.W. Mel. *Connectionist Robot Motion Planning.* Academic Press, San Diego, 1990.

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

[12] G.F. Poggio and T. Poggio. The analysis of stereopsis. *Annual review of neuroscience*, vol. 7, pages 379–412, 1984.

[13] L.G. Roberts. Machine perception of three-dimensional solids. In J.T. Tippet, editor, *Optical and Electro-optical Information Processing.* MIT Press, 1965.

[14] M. Rygol, S. Pollard, and C. Brown. A multiprocessor 3D vision system for pick-and-place. In *Proc. British Machine Vision Conf.*, BMVC90 pages 169–174, 1990.

[15] R.Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses In *IEEE Journal of Robotics and Automation*, RA-3(4) pages 323–344, 1987.

[16] R.Y. Tsai and R.K. Lenz. A new technique for fully autonomous and efficient 3D robotics hand-eye calibration. In *4th International Symposium on Robotics Research*, volume 4, pages 287–297, 1987.

[17] R.Y. Tsai and R.K. Lenz. Techniques for calibration of the scale factor and image center for high accuracy 3D machine vision metrology. *IEEE Trans. Pattern Analysis and Machine Intell.*, 10(5):713–720, 1988.

[18] S. Ullman. *The interpretation of visual motion.* MIT Press, Cambridge,USA, 1979.

[19] S.W. Wijesoma, D.F.H. Wolfe and R.J. Richards. Eye-to-hand coordination for vision-guided robot control applications. In *Int. J. Robotics Research*, 12(1) pages 65–78, 1993.