UNCALIBRATED STEREO AND HAND–EYE COORDINATION

Nicholas John Hollinghurst
Trinity Hall
January 1997

dissertation submitted to the University of Cambridge
towards the degree of Doctor of Philosophy

Department of Engineering
University of Cambridge

Uncalibrated Stereo and Hand–Eye Coordination

Nicholas John Hollinghurst

Trinity Hall, Cambridge

January 1997

Summary

This dissertation describes new applications of uncalibrated and weakly calibrated stereo vision to facilitate pick-and-place operations by a robot manipulator.

A ‘weakly calibrated’ stereo rig is one for which only a small number of reference observations have been made (for instance, by observing the robot itself making deliberate motions) and which might be subject to vibrations and small movements during use. Thus the epipolar geometry and camera parameters will be known only approximately. In such an environment, it is shown that an approximate linear model (the affine camera) is well suited to estimating both the epipolar constraint, and the relation between image measurements and the robot's coordinate system (the hand–eye relation).

The stereo system is used to track a pointing hand, implementing a vision-based user interface which allows the operator to specify objects to be grasped and to guide the robot's motion around the workspace. By considering only the plane projectivities between the images and a ground plane, it is shown that points on the plane may be indicated without calibration.

A novel stereo algorithm is developed to match line segments in weakly calibrated views and recover a description of the planar surfaces of objects in the robot's workspace. These can then be reconstructed in an approximate metric frame for grasp planning.

The tracking system employed in this project is a novel type of edge-seeking active contour, based on a template which can deform only affinely in the images. This can be used for tracking the operator's hand, the robot's gripper, and planar facets of objects in the workspace.

By tracking the robot itself, visual feedback can be employed to align the robot's gripper accurately with the surface to be grasped, even in the face of disturbances to the stereo cameras or the robot's control systems. Visually guided grasping is implemented in real time on standard hardware.

Uncalibrated Stereo and Hand–Eye Coordination

Nicholas John Hollinghurst

Trinity Hall, Cambridge

January 1997

Declaration

- This dissertation is submitted to the University of Cambridge towards the requirements for the degree of Doctor of Philosophy.
- This dissertation was composed entirely by myself. Except where otherwise noted, it describes my own original research and contains nothing which is the result of work done in collaboration.
- No part of this dissertation has been or is currently being submitted for any other university degree or diploma.
- The dissertation contains 52 figures and approximately 33000 words.

Acknowledgements

I am most grateful to my supervisor, Roberto Cipolla, for many ideas and suggestions, sound advice, and his constant encouragement throughout the project.

Over the last 4 years I have gained insight from conversations with several other people at CUED, particularly Antranig Basman, Tat-Jen Cham, Andrew Gee, Gabriel Hamid, Jonathan Lawn, Joan Lasenby and Jun Sato. I thank them for many interesting discussions and helpful suggestions.

Thanks to Elizabeth Guild, who helped me plan a path around a number of obstacles.

I would also like to thank my parents, and Mark, for all their love and support whilst I was grappling with this dissertation.

This project was funded by an EPSRC studentship and research grant GR/K08635. The robot manipulator was kindly donated to the Speech, Vision and Robotics group by Olivetti Research Ltd.

Contents

List of Figures

List of Tables

Glossary of symbols

Chapter 2

X, Y, Z	World coordinates of a point in space
X_c, Y_c, Z_c	Camera-centred coordinates of a point in space
x_i, y_i	Coordinates of a point on the image plane
u, v	Image coordinates in pixels
$\mathbf{R}$	Orthogonal 3×3 matrix representing a rotation
$\mathbf{t}$	Vector representing a 3-D translation
$\mathbf{A}$	Coefficients of plane affine transformation (2×2)
k_u, k_v, k_{uv}	Intrinsic camera parameters relating (x_i, y_i) to (u, v)
u_0, v_0	Intrinsic camera parameters denoting pixel coordinates of the origin
$\mathbf{P}$	3×4 matrix of projective camera coefficients
$\mathbf{C}$	World coordinates of the camera's optical centre
$'$	Marks symbols relating to second camera in stereo pair
$\mathbf{F}$	Fundamental matrix (3×3) encoding epipolar constraint
$\mathbf{o}$, $\mathbf{o}'$	Homogeneous coordinates of the epipole in each image
$\mathbf{M}$	2×3 matrix of affine camera coefficients
$\mathbf{Q}$	4×3 matrix of affine stereo coefficients
$\mathbf{c}$, $\mathbf{c}'$	Vectors parallel to camera viewing directions in space
$\mathbf{e}$	4-vector of coefficients for the linear epipolar constraint
x, y	Rectified image coordinates so that $y = y'$ for corresponding points

Chapter 3

$\mathbf{X}_P$	World coordinates of a point on the robot gripper
$\boldsymbol{\Theta}$	Joint angles specifying robot configuration
$\mathcal{K}$	Kinematic relation: $\mathbf{X}_P = \mathcal{K}(\boldsymbol{\Theta})$
$\hat{}$	Denotes a modelled or estimated quantity

$\mathbf{J}_{\mathcal{K}}$	Jacobian of $\mathcal{K}$, or matrix of partial derivatives of $\mathbf{X}_P$ wrt Θ
$\mathbf{u}_P$	Image coordinates of a point on the robot gripper
$\mathbf{u}_S$	Image coordinates of set point
$\mathbf{Q}$	4×3 matrix relating world to image coordinates
$\mathbf{Q}^+$	Left pseudo-inverse of $\mathbf{Q}$, such that $\mathbf{Q}^+\mathbf{Q} = \mathbf{I}$
$\mathbf{X}_P^*$	'Controller space' coordinates sent to kinematic model and robot
$\mathbf{o}_P$	Vector encoding gripper orientation in image-based terms
$\mathbf{o}_S$	Vector encoding target orientation in image-based terms
$\mathcal{F}$	Function relating image-based orientation to controller's parameterisation

Chapter 4

u, v	Image coordinates in pixels
X, Y	Canonical frame coordinates of points on the plane
$\mathbf{T}$, $\mathbf{T}'$	Transformations between plane and image coordinates
l_i	Line in the image defined by one view of pointing hand
l_{gp},l'_{gp}	Constraint lines on the plane, which intersect at the indicated point

Chapter 5

u, v	Image coordinates in pixels
x, y	Approximately rectified image coordinates
$y_{i[0]}$, $y_{i[1]}$	The y coordinates of each endpoint of the ith line segment
L_i	Length of segment in a 'vertically stretched' frame, used for matching
θ_{ij}	Angle between segments in the 'vertically stretched' frame
δ	Disparity along the epipoplar lines, $(x' - x)$
σ_{ij}	Intrinsic support for the match between segments i and j

$\mathcal{O}(n^a)$	Proportional to n^a as n becomes large
n	Number of features (straight line segments) in each image
r	Number of figurally related features per feature $(r < n)$
m	Number of candidate matches per feature $(m \leq n)$

Chapter 1

Introduction

This chapter sets out the motivation for the project, and introduces robot hand–eye coordination with a survey of existing robot vision systems. The contributions of the dissertation are summarised.

1.1 Motivation

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 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 yet match.

Robot manipulators have traditionally been restricted to performing repetitive tasks in highly ordered environments. Reliable and flexible computer vision would enable them to operate in less structured environments containing displaced or unfamiliar objects; to overcome operational errors using visual tracking and feedback; and to be programmed more easily via novel user interfaces such as gestures and pointing.

Because most robots need to move in all three dimensions, we exploit *stereo vision*, the use of two (or more) cameras to obtain 3-D information about the robot and its workspace. The scope of stereo vision applications is generally limited by the need to *calibrate* the vision system — the camera geometry must be measured to a high level of precision [129]. A well-calibrated stereo rig can accurately determine the position and shape of things to be grasped; however, if calibration is erroneous or the cameras are disturbed, the system will often fail gracelessly.

Here we explore the use of robust algorithms for stereo vision and hand–eye coordination which require minimal calibration and can tolerate some uncertainty in camera and robot positions and orientations. We develop a novel **visual grasping system** which uses vision to help plan and execute grasps of unmodelled objects placed at unknown positions in its workspace. The user indicates an object by a pointing gesture, and uncalibrated stereo vision is used to reconstruct its surfaces. Finally, the object is grasped by the robot under visual control.

Two strategies are employed to reduce dependence on calibration: firstly, by the use of *invariant* cues and representations of scene structure which are independent of camera geometry; secondly by the use of *image-based feedback* to correct for errors and align the robot with a visible target. Implementation is based on monochrome CCD cameras and a standard workstation environment.

1.2 Robot vision hardware

1.2.1 Configurations

A number of systems have been proposed using machine vision to help robot manipulators perform pick-and-place operations. The vision hardware may consist of one or more monochrome or colour video cameras, or more sophisticated devices such as structured light or laser rangefinders [16, 116, 75]. Figure 1.1 shows two common configurations:

Eye-in-hand systems have a camera mounted on the last link of the robot manipulator. This gives a detailed view of objects to be grasped, and facilitates visual servoing to align the gripper prior to grasping [21, 31, 27]. It also permits dynamic inspection of the target object from multiple viewpoints, affording a 3-D reconstruction of their surfaces [22, 125].

Eye-in-hand cameras typically suffer from a limited field of view and depth of field, so do not provide an overall picture of the workspace. They require camera calibration and *a priori* knowledge of the camera pose relative to the gripper, since these parameters cannot be recovered by self-calibration [48]. The entire visual field also moves whenever the robot does, which can increase image-processing overheads [134].

External camera or 'independent-eye' systems view both the manipulator and its workspace using one or more distant cameras. By observing the manipulator making known motions, self-calibration is possible [61], and feedback may be used to drive the gripper to a visually-specified target configuration regardless of camera position [134, 48].

With static external cameras, objects in the workspace tend to be viewed at a lower resolution than with a robot-mounted camera. The cameras may instead be mounted on pan/tilt heads, or an integrated stereo head [89]; zoom lenses may be employed for detailed inspection of parts as well as a broader view of the workspace [124]. Additional flexibility may be gained by mounting one or more cameras on an independent robot arm to allow dynamic control of viewpoints [19, 90], although this is clearly more expensive to implement.

1.2.2 Experimental setup

For this project, an external camera system is employed, with a pair of monochrome CCD cameras arranged for stereo vision (figure 1.2). The cameras view a robot manipulator and its workspace from a distance of about 2m; their field of view can also accommodate an operator, who can communicate with the system by means of pointing gestures. The angle between the cameras is in the range of 15–30 degrees. There is some flexibility in the positioning of the cameras, which are mounted on free-standing tripods: since these tend to be disturbed frequently, accurate calibration data are not available.

The experimental system is based around a Sun SPARCstation 20 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 dynamic control loop and provides a Cartesian kinematic model: the computer controls the robot and supplies it with visual feedback by means of a serial interface.

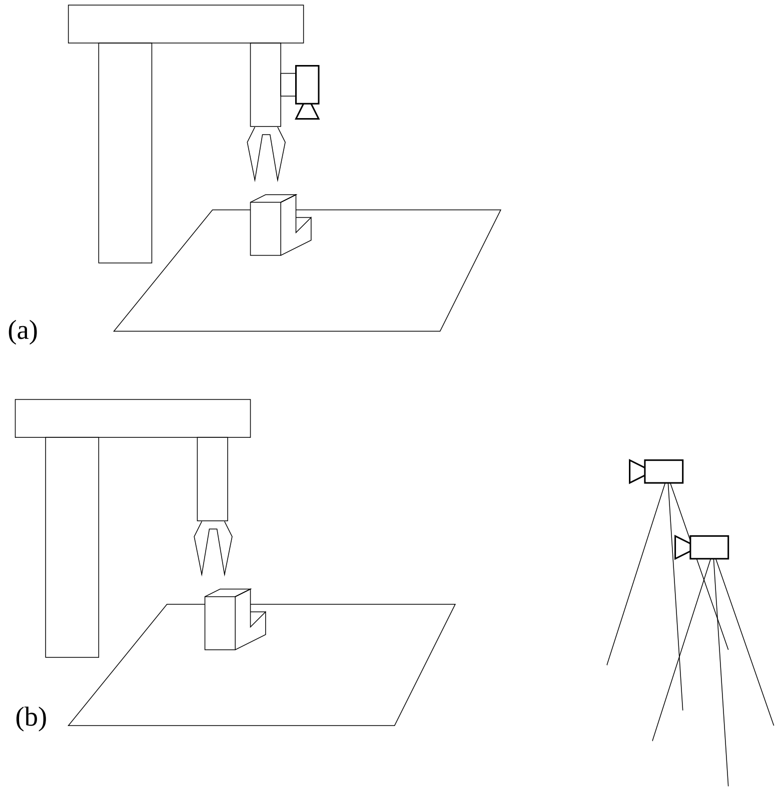


Figure 1.1: Robot/camera configurations: (a) eye-in-hand (b) external cameras

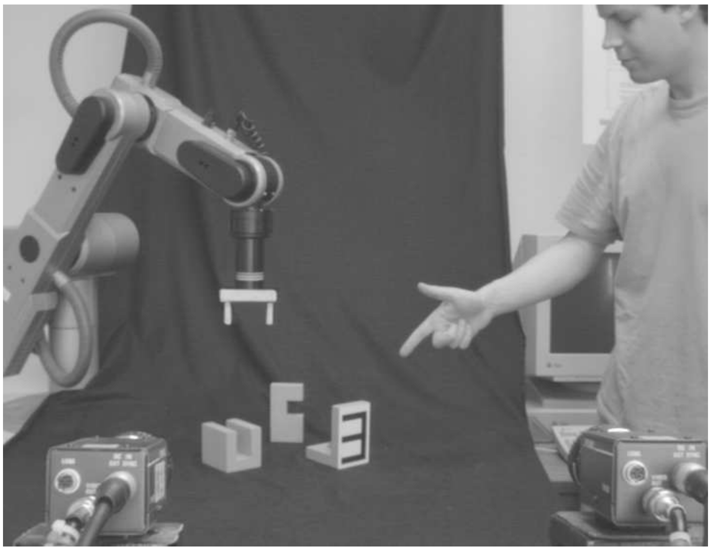


Figure 1.2: The experimental setup showing the robot, its workspace, stereo cameras and operator.

1.3 Existing systems

Here previous work in computer vision for robot manipulator guidance is reviewed (particular techniques and relevant theory will be surveyed in more detail in later chapters).

1.3.1 Look and move

The earliest paradigm for hand–eye coordination has become known as the *'look-and-move'* approach [48]. Vision is used only in the planning of motions, which are executed without visual assistance.

In two dimensions. Early robot vision systems [15, 97, 51], and many still in widespread use, have a single overhead camera to extract two-dimensional information about the positions of features on a part to be grasped, to recover its 2-D pose or to select different robot actions based on object recognition. This approach minimises the computation spent on vision, since the camera is used only once per operation, to analyse a static scene. Frequent recalibration is required to maintain satisfactory operation [15].

Binocular stereo vision. A recent hand–eye system [112] based on the Sheffield TINA stereo vision algorithms of Pollard et al. [103] uses a pair of calibrated cameras which view straight-edged objects taken from a modest repertoire. The system constructs a wire-frame model of the objects' edges [100] which is matched against stored models of the objects [99]. Objects are identified and picked up by an RTX robot using pre-determined grasps.

Other ranging techniques. An experimental system of Ikeuchi et al. [62] reconstructs the contents of a workspace using *photometric stereo*, in which a camera takes multiple images of the scene under different lighting conditions, to recover local surface orientation. This information is supplemented by range data provided by the PRISM stereo system [93], which projects random texture onto the scene and matches the resulting views by a multi-scale algorithm. These techniques allow it to reconstruct, recognise and grasp objects with smooth, featureless surfaces which would otherwise be difficult to see.

Robot planning systems have been proposed using other specialised sensors, such as *laser rangefinders* which reconstruct surface shapes from a single 'view'

[16, 75, 111]. Range imaging can recover the shapes of arbitrary surfaces more accurately than stereo [16]; however the sensors are expensive and require precise calibration.

Tracking and interception of moving targets. Vision can also be used to perform dynamic tasks involving moving objects. Allen et al. [1] describe a stereo vision system which can track a single target at frame rate, using a Kalman filter to estimate and predict its motion. The system is demonstrated using a robot arm to intercept and grasp a model train. Other high-speed stereo tracking systems have been used to perform tasks such as striking a ping-pong ball [2, 107]. As with all of the above systems, there is no visual feedback of the grasping operation, which is not robust to physical disturbances.

Open loop robot vision has become very sophisticated, and has been demonstrated successfully in pick-and-place and other applications. However, it requires accurate calibration so that the imaging and kinematic processes can be inverted without error, and demands high repeatability from the robot manipulator. The need for precision is most acute in the 3-D case, due to the increased number of parameters to be known and the added complexity of both robot and vision systems [119].

1.3.2 Visual feedback in two dimensions

In these systems, a single camera observes a manipulator from above, to guide the gripper's motion in two dimensions. The third dimension of movement is assumed to be constrained or controlled by an independent mechanism, as in many '2-and-a-half dimensional' robots which manipulate objects on a flat table, and whose vertical motion is limited and independent of the main X–Y motion.

Visual feedback for gripper alignment. The seminal work of Shirai and Inoue [117] reported the use of visual feedback to align a square prism over a box into which it is then fitted. The dimensions and heights of the objects were given, but the initial position of the box was unknown, and there was some uncertainty in the alignment of the prism within the gripper. Vision was used to estimate the two-dimensional position and orientation of the box, to place the prism over it. The system then observed the prism, estimated the error in its position and orientation, and made corrective motions.

A similar system was presented in [19] as a 'behavioural module' for an existing model-based manipulation system (the Edinburgh SOMASS system [77]). Here the manipulator moves across a horizontal *approach plane* to align itself vertically with a target, in preparation for grasping. Vision is used to track markers on the two fingers of the gripper, to provide visual feedback. It is noted that there is a 1-to-1 mapping between the approach and image planes, so that feedback can be based directly on the difference between observed and desired image positions.

Dynamic visual control. Most visual feedback systems use a hierarchy of two control loops: an inner one using joint sensors to control the robot's dynamics and an outer, slower loop incorporating vision. However, a few systems attempt to integrate the two using field-rate tracking of simple features on the manipulator [48]. In one experimental setup, a manipulator moves across a flat table and is viewed from above by a single camera [134]. The same camera is used to locate the target object during the planning phase. A point on the end-effector is marked by a beacon which allows it to be tracked at 50Hz to provide position-based feedback during execution. It is shown that the integration of visual feedback into the controller permits efficient operation and fast convergence despite significant errors in camera calibration or kinematic modelling.

Because of the simple 1-to-1 mapping between world and camera coordinates, visual feedback is an effective way to null positioning errors in two dimensions [136]. For fast, efficient operation, visual *tracking* of the end-effector (and/or its target) is required, to continuously update the estimate of the error between the manipulator's actual and desired pose.

1.3.3 Single camera feedback for 3-D tasks

These systems deal with the positioning of a robot in three dimensions under visual control, using either an eye-in-hand or external camera.

Hybrid system with 2-D vision. Harrel et al. [54] describe an eye-in-hand system to guide a fruit-picking robot. This system is notable by its use of *colour* vision to segment citrus fruits from the background and track them. The vision system provides two-dimensional feedback, controlling two degrees of freedom of the arm to keep the camera fixated on the fruit as it approaches; the distance is measured independently by ultrasonic ranging.

Single camera pose estimation. Espiau et al. [31] consider the use of visual feedback to place a calibrated camera in a given pose relative to visible features. They derive analytically the *image Jacobian* (that is, the matrix encoding the differential relation between camera motions and changes in image measurements) as a function of image feature positions. Inverting this relation allows the robot to make appropriate movements to bring the image features into a specified configuration, constraining the camera pose with respect to the target. This is demonstrated for the alignment of an 'eye-in-hand' camera with respect to a known target object.

Affine visual servoing. In the case where the target features are confined to a plane, the interaction between image and world motion is simplified. Colombo and Crowley [27] present a system which tracks features on a target surface and positions a camera at a given pose relative to the surface, deriving the gains for image-based control from a *weak perspective* [108] approximate pose estimation.

Spratling and Cipolla [121] present a similar system which requires no calibration but continuously re-estimates the image Jacobian from recent motions, to bring the camera into the pose corresponding to a goal image. They track the target surface using an active contour, and estimate the affine transformation between observed and goal configurations from area moments, making it correspondence-free [113].

The construction and attainment of an image-based goal requires a model of the camera and of the object to be manipulated [133, 31]; and pose estimation from a single view is ill-conditioned when the camera is distant [53]. Therefore, single-camera servoing is best suited to calibrated eye-in-hand systems.

1.3.4 Stereo visual feedback

Systems have also been proposed using *stereo* visual feedback to improve the accuracy of 3-D manipulation.

Image-specified manipulation. Skaar et al. [119, 18] consider the case in which known points on a manipulator are sporadically observed by two or more cameras, but continuous stereo tracking is not possible. They introduce a simple orthographic camera model and show that the estimated camera parameters also absorb linear errors in the kinematics: this allows the system to predict the configuration which will bring the gripper to a visually-specified target in two or more views. By appropriately weighting a set of observations, they are able to solve for the local hand–eye relation in any region of the workspace, allowing 6-DOF[1] alignment of a gripped object with a visually-specified target.

Stereo image-based feedback. Hollinghurst and Cipolla [61] demonstrated the use of stereo tracking of a robot manipulator whose kinematics are (approximately) known, using visual feedback to align it with a target. A linear camera model is assumed. An extension of this system is described in chapter 3.

Hager et al. [50] present a similar system, using stereo image-based feedback for 6-DOF positioning. Approximate camera calibration is used to estimate the image Jacobian, but the system is shown to be insensitive to calibration errors. Hager then considers the use of visual feedback to enforce one or more constraints (with 6 DOF or less) between the end-effector pose and that of another object, using least-squares solutions in both the underconstrained and overconstrained cases [49]. Visual constraints are used to assist dextrous tasks such as the insertion of a floppy disk into a drive.

Multiple cameras simplify the problems of setting and attaining visually-specified goals for 3-D positioning, and allows the manipulation of *unmodelled* objects (whose pose cannot be determined in a single view). Such systems are robust to small errors in the robot's kinematic model and allow precise manipulation tasks to be performed.

1.3.5 Learning systems for hand–eye coordination

Some systems deal with unknown robot kinematics as well as unknown camera parameters by considering the *visual kinematic relation* between actuator settings and parameters extracted from the image. Since robot kinematics are usually highly nonlinear, the structure of this relation must be *learnt* either before or during operation.

Mel's MURPHY. Mel [85, 86] took inspiration from human learning to devise a vision-guided control and planning system that *learns by doing*. It controls

[1] That is, control of both *position* and *orientation* in three dimensions.

a 3-DOF planar arm and guides it to a visible target whilst avoiding obstacles. MURPHY learns the forward visual kinematic relation, taking an unusual approach by learning to 'envisage' an entire 64×64 image of the arm in any configuration. It is this whole-image-based approach that is the key to its collision-avoiding behaviour. It also learns the inverse differential kinematic relation by observing how the gripper position responds to changes in actuator settings. Learning takes place in an initial 'random flailing' stage in which it views about 17000 of the 3 million legal arm configurations. These models are used by a path planner, which is based upon heuristic depth-first search. A trajectory is planned in joint space, to reach the target avoiding obstacles. Despite promising initial results and a refreshingly simple approach, MURPHY is slow, and scales badly to higher degrees of freedom. Its neural network architecture could not efficiently model the simple geometry underlying the camera and kinematic relation.

3-D visual kinematic learning. Hervé et al. [58, 59] take a qualitative approach to visual kinematic learning by identifying the *singularities* in the joint space / sensor space transformation (points where $|\mathbf{J}| = 0$, i.e. the inverse differential relation is not defined). Away from these singularities, the hand–eye relation is smooth and can be navigated using feedback. The robot makes experimental motions to determine the gradient of its *Perceptual Control Surface*. It builds up a qualitative model of the PCS by noting when it encounters a singularity in the Jacobian, and plans paths which avoid these singularities.

Visual memory-based control can be used to control manipulation by a multi-fingered hand, whose kinematics are difficult to model analytically, by tracking the position and orientation of a grasped object. The system of Jägersrand et al. [63] estimates the Jacobian of the *visual kinematics relation*; that is, the matrix of coefficients relating the movement of each joint to movements in the image of the grasped object [48], using exploratory movements to obtain its components in various directions. As it moves, it builds up a piecewise linear model of this relation, with uncertainty analysis used to ascertain the region of trust for each linear patch.

Learning-based control can be useful when controlling a redundant or multi-fingered manipulator (which would otherwise be difficult to model [63]), but in general this is unnecessary and inefficient. Qualitative modelling of the hand–eye relation can also be used in conjunction with visual feedback [59].

1.4 The approach

Traditional robot vision systems have attempted accurate reconstruction, using metric information to plan and execute motions in an open loop [112, 1, 107]; but these require calibration and are not robust to disturbances. Systems have also been proposed using image-based feedback with varying calibration requirements [117, 134, 31]. Sometimes the characteristics of the robot itself are learnt along with the parameters of the vision system [59, 133, 63].

Here we address the case in which the cameras are uncalibrated but the robot's kinematics are known (perhaps imperfectly), allowing the end-effector to be controlled in terms of Cartesian coordinates with a small, smooth error function. In the absence of accurate calibration, it is reasonable to resort to an approximate linear model of stereo vision. The Cartesian hand-eye relation is monotonic and can be modelled by a linear relation.

The use of a kinematic model simplifies the learning of the hand–eye relation (a linear estimator will suffice), whilst the use of visual feedback retains robustness against small kinematic errors and even non-stationary camera parameters. Such an approach has been used very successfully for visual robot control in two dimensions. Here it is applied to stereo vision for three-dimensional control of position and orientation. We track the robot's gripper in stereo with active contours, and use visual feedback to servo its image position in the two views.

To exploit visual feedback in grasping operations, the manipulator's goal configuration must be specified in terms of *image* measurements. Indication of the target object must therefore be image-based, and this can be achieved using a visual user interface. We use stereo vision to form an affine reconstruction of the facets of the target object in an image-based coordinate frame. This representation is used in conjunction with visual feedback to align the robot gripper with a suitable facet of the target object, so that it may be grasped.

Thus the entire grasping operation is to be facilitated by uncalibrated stereo vision.

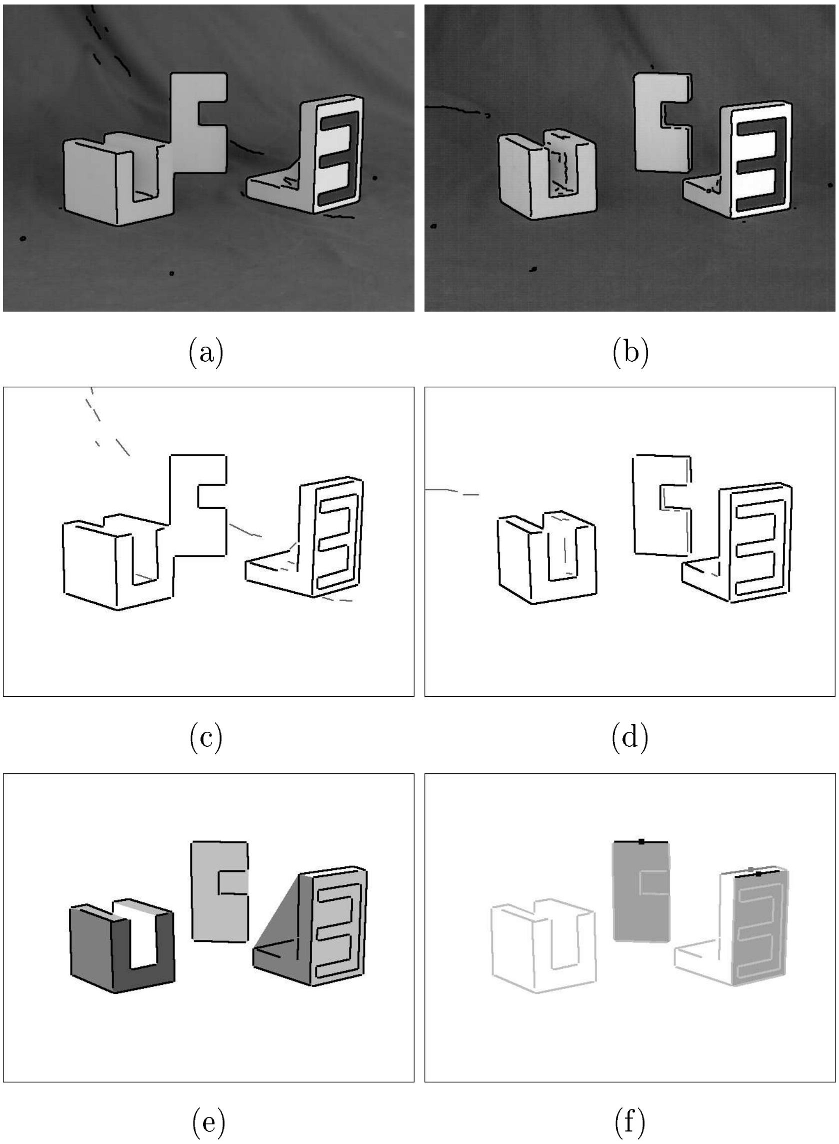


Figure 1.3: Surface reconstruction for grasping: (a,b) stereo images of the workspace with edges superimposed; (c,d) unmatched *(light)* and matched *(dark)* line segments; (e) cyclopean view with planar facets identified; (f) proposed grasping sites.

1.5 Contributions

1.5.1 Affine stereo

In this dissertation, it is argued that a linear approximate camera model is well suited to practical uncalibrated stereo, both for solving the correspondence problem and for modelling the relation between world and image motions, whenever the camera configuration resembles the typical 'parallel' arrangement with equidistant cameras fixating on a compact scene. It is noted that the epipolar geometry of a stereo rig is qualitatively different from that found in many navigation/structure-from-motion applications in which the camera motion is largely along the optical axis. The restricted form of the *affine camera* makes it easier to compute approximate camera parameters from a small number of measurements, than the projective camera model estimated in the traditional manner. Affine stereo is shown to be more robust to image coordinate noise and disturbances to the cameras.

1.5.2 Pointing interface

A novel form of human–robot interface is presented, based on real time stereo vision tracking of the operator's pointing hand. We do not use a full 3-D reconstruction of the hand in space, but consider only plane projectivities between a ground plane and the images. This formulation allows objects on a plane to be indicated by pointing, without the need for camera calibration. Simulations and experiments measure the accuracy of the system, both in open loop and as a means for the operator to servo the position of the robot's gripper.

1.5.3 Weakly calibrated stereo reconstruction

A new stereo matching algorithm is developed for matching line segment images in weakly calibrated stereo pairs (in which the epipolar geometry is only approximately known because only a few reference correspondences have been observed) under weak perspective. Integrated into the system is the grouping of line segments into planar facets. This provides a model of the scene which is suitable for grasp planning with a parallel gripper (figure 1.3). It also allows the visible surfaces to be reconstructed accurately despite uncertainty in the epipolar geometry, which is not generally possible for individual line segments. The reconstruction is used to select a suitable grasp of the target object.

1.5.4 Visual feedback for grasping

The linear approximation to the 'hand–eye relation' between the robot's movements and motion in the images is used as the basis of a visual feedback control loop, allowing the robot to be guided in three dimensions towards a visually specified target. The robot is aligned so that a surface of its gripper is near to and coplanar with a given surface of the object; then rotated into the grasping configuration. It is shown that such an approach is robust to calibration errors of either the robot or vision system, and even to disturbances to the system occurring during operation.

1.6 Overview of the dissertation

Chapter 2 gives a general introduction to the geometry and modelling of stereo vision systems, and derives the camera models which are referred to later in the dissertation.

- Conventional projective and affine models of video camera imaging are introduced. The theory of camera calibration and the epipolar geometry of stereo vision are reviewed for each model.
- The affine and projective camera models are compared in the context of parallel-camera stereo vision, and it is concluded that the affine camera is more robust to errors and more easily calibrated.
- This is supported by experiments and simulations comparing projective and linear models degraded by noisy data. It is shown that the systematic error due to the linear approximation is of comparable magnitude to other errors, e.g. noise in image in feature localisation.

Chapter 3 describes the use of the affine stereo formulation developed above to achieve alignment of the robot with a visually-specified target, using stereo visual feedback.

- A visual feedback scheme is developed for the affine stereo formulation. It is noted that, for point alignment, image-based and position-based servoing are equivalent under this model.
- We extend the feedback scheme to align both the *position* and *orientation* of planar features on the robot and target object.

- The system is implemented using affine active contours to track a surface of the robot's gripper. By tracking the target facet as well as the robot, we close the visual control loop and enable the system to track and grasp objects despite movements and disturbances to the cameras.
- Experiments show that this system is robust to camera motions and small errors in the robot's kinematic model.

Chapter 4 describes a novel human–robot interface based on pointing. This is the proposed means for indicating to the system which object is to be grasped.

- The geometry of pointing at a ground plane is analysed, and it is shown that this does not require camera calibration, apart from 4 matching reference points on the plane.
- It is shown how this method may be used to indicate points on a single plane or in an environment containing multiple planes.
- Methods for tracking a pointing hand are summarised, and a novel implementation is developed using a pair of affine active contours to track the thumb and index finger.
- Experimental results and accuracy evaluation are presented.

Chapter 5 discusses the stereo correspondence and reconstruction of a scene composed mainly of straight edges and planar surfaces.

- Previous approaches to solving the correspondence problem in stereo vision are reviewed, and their shortcomings discussed in the context of uncalibrated or weakly calibrated setups of the kind used in this project.
- A stereo matching algorithm is developed, based on existing work for line segment matching but explicitly allowing for a bounded error in epipolar constraint estimation.
- Uncalibrated plane grouping is incorporated into the system, exploiting the geometry of weak perspective views of coplanar features.
- The groupings are used to extract a description of the planar surfaces of objects for reconstruction, and to improve the accuracy of uncalibrated reconstruction.

Chapter 6 is concerned with the implementation of the complete visual grasping system.

- The theory of grasping is briefly reviewed, with particular emphasis on grasp synthesis for a parallel-jawed gripper.
- A scheme is devised for choosing grasping sites on a stereo reconstruction of the surfaces of the target object, and demonstrated on real images of ‘blocks world’ scenes.
- The algorithms described in the dissertation are integrated to form a complete system.

Chapter 7 reviews the findings and contributions of the dissertation, and concludes with an outline of future work.

Appendix A describes the novel type of active contours used in the project. These are based on a template and are able to deform only affinely. They are suitable for the real time tracking of planar objects or facets under weak perspective, as well as for tracking the index finger and thumb of a pointing hand.

Chapter 2

Perspective and Affine Stereo

In this chapter we review the geometry of monocular and stereo cameras, and show that an approximate linear model of stereo vision is robust and well-suited to uncalibrated and weakly calibrated systems.

2.1 Introduction

In order to make geometrical use of stereo vision we must model the relation between the three-dimensional world and two-dimensional images. Specifically, we will need to use stereo to reconstruct the shapes of objects in the robot's workspace in order to grasp them successfully, and to associate relative image positions with 3-D motions to drive the robot to its target configuration.

This chapter reviews the geometrical modelling of the perspective camera; its generalisation, when full calibration data are not available, to the projective camera; and a useful linear approximation, the affine camera. Essential theory for stereo vision is summarised in each case, describing the relation between a stereo pair of views, and the use of calibrated and uncalibrated stereo systems to reconstruct points and surfaces.

Numerical experiments will be used to demonstrate the superiority of the affine camera for the estimation of the epipolar constraint and the reconstruction of relative positions in three dimensions, when calibration data are noisy and few — it is this *weakly calibrated stereo* model that is used in chapter 3 to control a robot and in chapter 5 to facilitate stereo correspondence and the reconstruction of planar facets.

2.2 The perspective camera

2.2.1 Pinhole camera

Video cameras are conventionally analysed using the *pinhole camera model*, in which an image is projected onto a retinal plane by rays passing through a single point called the *optical centre* [32]. This point forms the origin of a camera-centred coordinate frame, (X_c, Y_c, Z_c) such that the retinal plane has the equation $Z_c = f$, where f is a constant, the *focal length*. Image coordinates (x_i, y_i) on the retina are ratios of world coordinates (X_c, Y_c, Z_c) thus: $x_i = fX_c/Z_c \;\; y_i = fY_c/Z_c$. This simple model is a good approximation to the optics of most types of camera, although it neglects effects such as lens distortion which are significant in some high-accuracy applications such as aerial photogrammetry [128].

The relation between the camera-centred and some other world frame (such as that defined by a robot or another camera) is a rigid motion, encoding the camera's orientation and position. It can be represented by an orthogonal rotation matrix $\mathbf{R}$, and a translation vector $\mathbf{t}$. Using homogeneous coordinates [8] with a tilde to symbolise equivalence up to a scale factor,

$$\begin{bmatrix} x_i \\ y_i \\ f \end{bmatrix} \sim \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{2.1}$$

2.2.2 Projective camera

Measurements on the image plane are not made directly, because the image is sampled into *pixels*. The relation between retinal positions (x_i, y_i) and pixel addresses (u, v) is modelled by an affine transformation (to represent offsets, scaling and shearing) [32]. Aligning the pixel and retinal coordinate systems so that the v and y_i directions coincide,

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} fk_u & fk_{uv} & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{2.2}$$

The 5 coefficients[1] fk_u, fk_v, fk_{uv}, u_0 and v_0 are the camera's *intrinsic parameters*, and the $\mathbf{R}$ and $\mathbf{t}$ components can be expressed in terms of 6 *extrinsic parameters*. Combining these relations, we obtain the *direct linear transformation* (DLT) form of the camera model [128]:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{2.3}$$

This is the usual camera model for many vision systems where the camera intrinsics and pose are not initially known [32]. The transformation matrix is defined up to a scale factor, thus there are 11 degrees of freedom.

2.2.3 Camera calibration

Calibration of the camera is necessary to fix the 11 unknowns in the 12 parameters p_{ij}. This can be done by observing at least 6 points of known position, not all coplanar. Each observation generates two homogeneous equations in terms of p_{ij}. The system is homogeneous, so we can constrain $p_{34} = 1$ and solve using linear least squares estimation. If image positions are noisy, the results can be improved by observing more than 6 points using a recursive linear estimator. Often a special *calibration object* with very accurate grids is used [6].

In practice, the linear method is somewhat ill-conditioned, and a large number of reference points are needed, which must be localised to sub-pixel accuracy [129]. This is because the error measure, when formulated linearly in p_{ij}, is not geometrically meaningful; the last row and column have different numerical dimensions and play different roles in the model. A number of calibration methods have been proposed based on nonlinear (iterative) optimisation and reparameterisations of $\mathbf{P}$, and these give somewhat better results [37, 130].

Having obtained the DLT form, the intrinsic and extrinsic parameters can be extracted if required. For any 3×4 matrix of rank 3, scaled so that $||p_{31}\ p_{32}\ p_{33}|| = 1$, it can be shown [32] that there exist four sets of camera parameters satisfying equation (2.2), the four solutions being trivially related by changes of sign. The optical centre can be recovered directly, by solving $\mathbf{PC} = \mathbf{0}$.

[1] Often, k_{uv} is taken to be zero [32]. This assumes rectangular pixels in the camera and complete decoupling of horizontal and vertical coordinates in the frame capture hardware. Thus there will be only 4 intrinsic parameters, and an additional constraint will be imposed on camera calibration.

Once the intrinsic parameters of the camera are known, pixel coordinates can be converted back to *normalised* image coordinates $(x_i/f, y_i/f, 1)$: these would be the image-plane coordinates for a pinhole camera of unit focal length, and are equivalent up to a scale factor to the camera-centred world coordinates. Hence directions and angles may be measured at the optical centre.

Camera calibration must be repeated whenever the camera lens is replaced, zoomed or refocused (change of f), or the camera position is disturbed (change of $\mathbf{R}$, $\mathbf{t}$). This project was motivated by a desire to avoid full camera calibration, and explores the use of formulations that work satisfactorily with few or no calibration measurements.

2.2.4 Viewing a plane

Consider the case in which several observed points line on a single plane. Thus in some world coordinate system they will all have $Z = 0$, and equation 2.2 loses one column of the camera transformation to become:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} fk_u & fk_{uv} & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} . \tag{2.4}$$

We see that the relation between plane and camera coordinates is a 2-D projectivity, preserving projective invariants of features on the plane [88]. If the intrinsic parameters are known, $\mathbf{R}$ and $\mathbf{t}$ (the camera pose relative to the plane) can be computed up to a two-fold ambiguity from just 3 known points, by exploiting the nonlinear constraints among elements of the rotation matrix [53]. If camera parameters are not known, a minimum of 4 reference points are needed to fix the 2-D projective relation between the world plane and the image coordinate system [109], allowing points and structures on the plane to be reconstructed from a single view.

2.3 Full perspective stereo

In general, a single camera gives only two-dimensional information about scene structure. In the absence of other constraints, two or more views are required for reconstruction.

2.3.1 The epipolar constraint

The image coordinates of a world feature in two images are not independent, but are related by an *epipolar constraint*. This comes about from the fact that 4 image coordinates are derived from only 3 degrees of freedom in world positions. Consider a family of planes passing through the optical centres of both cameras. These project to a family of *epipolar lines* in each image (figure 2.1). 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. Most stereo systems exploit this constraint, which reduces the search for matching features to a single dimension [98].

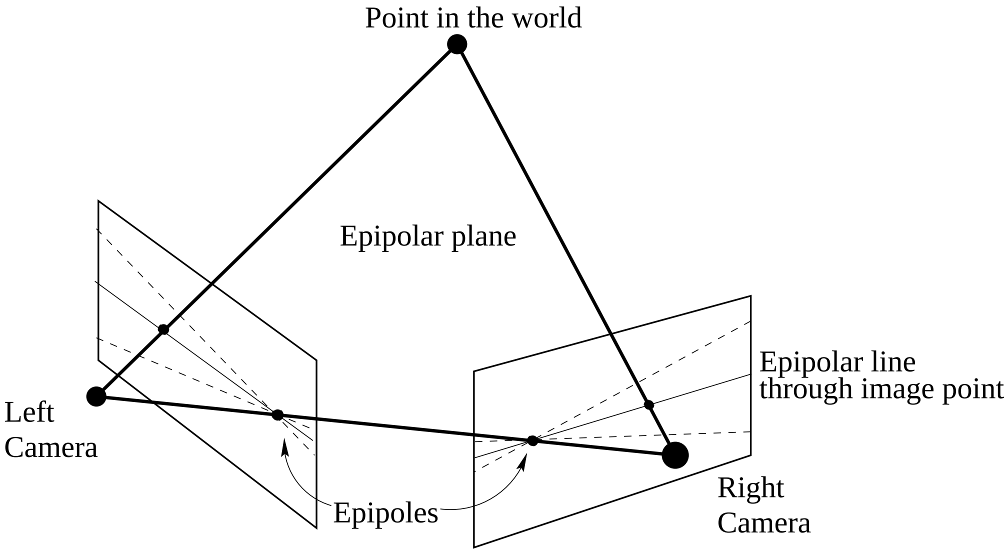


Figure 2.1: The epipolar geometry of stereo vision

Fundamental matrix

The epipolar constraint is represented algebraically by a 3×3 matrix $\mathbf{F}$ called the *fundamental matrix* [32] such that corresponding points (u, v) and (u', v') satisfy:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \mathbf{F} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0. \tag{2.5}$$

This is a generalisation of Longuet-Higgins' *essential matrix* [72], which encoded the relation between camera-centred coordinates in two views, to the case where intrinsic parameters are not known. $\mathbf{F}$ has rank 2 and is defined up to a scale factor, i.e. the constraint has 7 degrees of freedom.[2] The epipole in each view is the image of the other camera's optical centre (i.e. $\mathbf{o} \sim \mathbf{PC}'$, using homogeneous coordinates for $\mathbf{o}$ and $\mathbf{C}'$). The epipoles can be extracted from the fundamental matrix itself: $\mathbf{Fo} = \mathbf{0}$ and $\mathbf{F}^T\mathbf{o}' = \mathbf{0}$; that is, they are in the right and left nullspaces of $\mathbf{F}$.

In calibrated systems, $\mathbf{F}$ can be recovered from the camera matrices [32], otherwise it may be obtained up to a threefold ambiguity by observing 7 corresponding points [127], or estimated by linear least squares given 8 corresponding points [72]. Epipolar geometry can be estimated from image coordinates alone without reference to world coordinates; however, degeneracy occurs when the points all lie on a *critical surface* such as a plane, cone or cylinder [73, 36]. As with camera calibration, the solution is sensitive to errors and may require more than 8 points and/or nonlinear optimisation [76].

Linear form

In the general case, epipolar lines will meet at a single point in each image plane, the *epipole*, which is the image of the other camera's optical centre [32]. However, if the cameras' *focal planes* ($Z_c = 0$, $Z'_c = 0$) coincide, the epipoles will be points at infinity and the epipolar lines parallel. In this case, the first 2×2 elements of $\mathbf{F}$ become zero, and the epipolar constraint is a single linear equation in u, v, u', v' and a constant term [115].

In practical stereo rigs each camera is usually far outside the other camera's field of view, and the linear form of the epipolar constraint is often valid (at least as a first approximation when only a small number of correspondences have been found).

[2] The loss of rank can be explained by considering $\mathbf{F}$ as a projective *correlation* between points in one image and lines in the other. To be an epipolar constraint, all points on an epipolar line must yield the same line when multiplied by $\mathbf{F}$: the matrix is therefore singular.

The linear approximation to the epipolar constraint can be recovered from just 4 corresponding points in uncalibrated stereo. This form of the constraint is incorporated into the *affine stereo* model introduced in section 2.5.

Image rectification

The simplest possible form of the epipolar constraint occurs when the cameras have the same intrinsic parameters and are separated by a pure translation in the X_c direction (*parallel cameras*). The constraint becomes: $v' = v$, i.e. corresponding points must lie on the same horizontal scan line in each image, and object depth is encoded by horizontal *disparities* along the scan lines. This simplifies the problems of stereo correspondence and reconstruction, and many stereo vision algorithms require images in this form [103, 6]. The purpose of *image rectification* is to transform images (or the image coordinates of features) so that the epipolar constraint takes this form, even when they were taken through non-parallel cameras.

If the cameras are calibrated, rectification is achieved by projective transformations of image points into a new coordinate system (x, y) so that $y' = y$ for all matching points. The rectification transformations simulate the rotation of each camera until they are parallel, and the scaling and shifting of one image to bring the scan lines into agreement [32]. If the epipolar geometry is known but not the camera intrinsics, the rectification transformations are defined up to 9 free parameters,[3] usually chosen for numerical convenience [6].

2.3.2 Reconstruction

With calibration

Assume that both cameras have been calibrated for the same world coordinate frame. By rearrangement of (2.3), each measurement of (u, v) yields two simultaneous linear equations in (X, Y, Z), which represent the line of sight from a camera to a point in the world. Two views of the same point give us four linear equations which can be solved, e.g. by a least squares method. Numerical optimization can be used to improve robustness to noise (at the expense of speed), by minimising the offsets in image coordinates between observed and backprojected features [32].

[3] The epipolar constraint has 7 DOF, but a general pair of projective transformations on the two images would have $8 + 8 = 16$ DOF.

Intrinsically calibrated cameras

For cameras with calibrated optics but unknown pose, the rotation and direction of translation between the views may be estimated. The *essential matrix* (which is the fundamental matrix defined in terms of normalized image coordinates [72]) is computed from the intrinsic parameters and the image coordinates of at least 7 correspondences. It can then be decomposed [32] into the product of an antisymmetric matrix **T** and an orthogonal rotation matrix **R**. **T** encodes the translation and is defined up to a scale factor; thus the scene may be reconstructed up to a similarity.

Uncalibrated cameras

The extraction of non-metric and viewpoint-invariant information from completely uncalibrated cameras is a rapidly developing field in machine vision [88, 3, 10].

For instance, given two uncalibrated views of 8 corresponding points (from which the fundamental matrix can be recovered), it is possible to reconstruct the scene up to a 3-D projective transformation[4] [35, 33]. 5 of the points are used as a projective basis in space, i.e. they are assigned the coordinates $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$ and $(1, 1, 1, 1)$. Likewise, 4 of them form a projective basis in each of the images. Using these coordinate systems, each camera transformation matrix takes the form:

$$\begin{bmatrix} \mu a - \nu & 0 & 0 & \nu \\ 0 & \mu b - \nu & 0 & \nu \\ 0 & 0 & \mu c - \nu & \nu \end{bmatrix}$$

where (a, b, c) are the projective image coordinates of the fifth point. The coordinates of the optical centre can also be expressed in terms of a b, c, μ and ν. Thus, each camera model is fixed up to one degree of freedom, the ratio $\mu : \nu$. Faugeras shows how this may be eliminated using the epipolar constraint between views [35], exploiting the relation between the epipoles and the optical centres of the cameras.

This result can be extended by noting that we are within 3 degrees of freedom of an *affine* reconstruction of the scene. The 3 missing parameters encode a representation of the plane at infinity within the above projective basis, and these may be recovered by observing e.g. 3 vanishing points of parallel lines [105, 33].

[4] A projective representation of a 3-D scene is 9 DOF from Euclidean structure, allowing quite serious distortions of the reconstructed scene. It is therefore most useful for applications such as *recognition* of objects based on projective invariants [109].

2.4 Weak perspective and the affine camera

The projective camera model has many parameters and is nonlinear in form, making it difficult to calibrate. We now consider a simpler first-order approximation, the affine camera, as an alternative camera model for stereo vision.

2.4.1 Weak perspective

Let us assume that, within some region of the scene, the relative depth $|\Delta Z_c / Z_c|$ is bounded by a small value (*weak perspective* [108]). Equation (2.1) becomes:

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \frac{f}{h} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{2.6}$$

where $h = \mathbf{r}_3 \cdot \mathbf{p} + t_3$, the normal distance between the focal plane ($Z_c = 0$) and a point $\mathbf{p}$ in the region of interest. We assume that h is constant across this region, i.e. that the relation between world and image coordinates is linear. With a camera whose intrinsic parameters are known, the X_c and Y_c components of feature positions can be recovered up to scale from a single view; and the camera pose can be estimated from ≥ 3 points in known configuration [53]. This approximation to the camera model is useful in *tracking* applications, where a compact object is observed moving around a three-dimensional space [56]. Under weak perspective, any image of a planar facet will be an *affine transformation* of the plane, encoding its depth and orientation relative to the camera, and images of planes will deform affinely under motion [67, 21].

2.4.2 Affine camera

Now if the depth of the entire scene is small compared to the camera distance, h can be assumed constant. Consider the images in terms of pixel coordinates (u, v). Without a knowledge of the intrinsic parameters, camera pose cannot be determined, but the camera model is further simplified:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} + \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{2.7}$$

where (u_0, v_0) is the image of the world origin. This is equivalent to parallel projection followed by an arbitrary affine transformation in the image. It is known as the *affine camera* model [88].

The affine camera can be calibrated by observing just 4 reference points. All 8 coefficients are independent. Its linear form makes it less sensitive to calibration noise, since it can be optimised to minimise errors in the image coordinates themselves. Where the assumption of weak perspective throughout the scene can be made, it allows a more accurate camera model to be constructed from limited calibration data [23].

2.5 Affine stereo

With the affine camera model, image coordinates are linear functions (plus a constant offset) of the 3-D coordinates of points in the world. This simplifies the epipolar constraint, as well as calibrated and uncalibrated stereo reconstruction.

2.5.1 The affine stereo formulation

Combining information from a pair of images, we have four image coordinates (u, v), (u', v') for each point, all linear functions of the three world coordinates (X, Y, Z):

$$\begin{bmatrix} u \\ v \\ u' \\ v' \end{bmatrix} = \begin{bmatrix} u_0 \\ v_0 \\ u'_0 \\ v'_0 \end{bmatrix} + \mathbf{Q} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} . \tag{2.8}$$

$\mathbf{Q}$ is a 4×3 matrix formed from the m_{ij} coefficients of (2.7) for the two cameras. It should be noted that the integration of information from more than two cameras is easily accommodated within this framework: each additional view generates two extra linear equations which can be represented by extra columns to $\mathbf{Q}$.

2.5.2 The epipolar constraint in affine stereo

When a point is viewed in stereo, there are 4 image coordinates, all linear functions of 3 world coordinates. These cannot be independent, but are related by a single linear constraint: the epipolar constraint thus takes the linear form [115], and can be estimated from a minimum of 4 corresponding points.

To analyse the constraint, consider the 4-vector $\mathbf{e}$ satisfying $\mathbf{Q}^T\mathbf{e} = \mathbf{0}$, i.e. the direction orthogonal to the three *rows* of $\mathbf{Q}$. This is the annihilator for vectors of the form $\mathbf{Q}[XYZ]^T$. Thus the epipolar constraint may be written:

$$\mathbf{e} \cdot \begin{bmatrix} u - u_0 \\ v - v_0 \\ u' - u'_0 \\ v' - v'_0 \end{bmatrix} = 0. \tag{2.9}$$

Geometrically, the epipolar planes are the family of planes parallel to both viewing directions $\mathbf{c}$ and $\mathbf{c}'$ (the nullspace vectors of $\mathbf{M}$ and $\mathbf{M}'$), so that the epipolar line direction in the first image is parallel to $\mathbf{Mc}'$, and in the second image to $\mathbf{M}'\mathbf{c}$. Furthermore, it follows that $[e_1\, e_2]{\cdot}\mathbf{Mc}' = 0$ and $[e_3\, e_4]{\cdot}\mathbf{M}'\mathbf{c} = 0$, since motion along the epipolar lines does not violate the constraint.

Image rectification

To rectify a pair of images, each point must be represented in terms of linearly independent coordinates (x, y) such that $y = y'$ for all matching points. This condition is satisfied when:

$$\begin{aligned} y &= -Ae_1(u - u_0) - Ae_2(v - v_0) + B, \\ y' &= Ae_3(u' - u'_0) + Ae_4(v' - v'_0) + B \end{aligned} \tag{2.10}$$

for some scale factor A and offset B. We can then use y and y' values to find or test for matching features in stereo. The rectified x coordinate is most conveniently defined as the component parallel to the epipolar lines in each image, so that basis vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ are orthogonal: thus rectification may be achieved using plane similarity transformations (rotation, translation and scaling) in each image.

2.5.3 Reconstruction

Calibrated cameras

If all the coefficients are known, world coordinates can be obtained by inverting (2.8). Since the model is linear in both the world and image coordinates, least-squares minimisation gives an optimal solution from (uncorrelated) noisy image data. Errors in calibration will manifest themselves as an affine distortion of the perceived coordinate frame [68].

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 tracking the position of a robot gripper as it visits four predefined reference points [61].

Uncalibrated cameras

In the absence of camera calibration, any four (non-coplanar) points may be given arbitrary world coordinates (such as the canonical affine basis $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$ and $(1, 0, 0)$). The appropriate solution for $\mathbf{Q}$ yields an *affine* reconstruction of the scene, which preserves affine shape properties such as collinearity, coplanarity and ratios of parallel lengths. This is in accordance with Koenderink and van Doorn's *Affine Structure-from-Motion Theorem* [68].

2.5.4 Recovery of surface orientation from two views

Any two views of the same planar surface will be affine-equivalent: there will exist an affine transformation that maps one image to the other. This transformation can be used to recover surface orientation [21]. Let the linear mapping between the views be represented by transformation matrix $\mathbf{A}$ and a 2-D translation vector.

It is the $\mathbf{A}$ component which encodes orientation. Consider the standard basis vectors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ in one image and suppose they were the projections of some vectors tangent to the surface. The columns of $\mathbf{A}$ itself will be the corresponding vectors in the second image. By inspection, the epipolar constraint requires that:

$$\begin{aligned} e_1 + e_3 a_{11} + e_4 a_{21} &= 0, \\ e_2 + e_3 a_{12} + e_4 a_{22} &= 0. \end{aligned} \tag{2.11}$$

Two degrees of freedom remain. For purposes of visual servoing on surface orientation, such transformations can simply be parameterised by the pair (a_{11}, a_{12}). For reconstruction, we can form a surface normal vector $\mathbf{n}$ from the cross product of two world-space vectors on the plane:

$$\mathbf{n} = \mathbf{Q}^{+} \begin{bmatrix} 1 \\ 0 \\ a_{11} \\ a_{21} \end{bmatrix} \wedge \mathbf{Q}^{+} \begin{bmatrix} 0 \\ 1 \\ a_{12} \\ a_{22} \end{bmatrix} \tag{2.12}$$

where $\mathbf{Q}^{+}$ is the pseudo-inverse $(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T$ [122].

2.6 Comparison of perspective and affine stereo

A series of experiments and simulations were performed to compare the accuracy of perspective and affine stereo models in cases where only a small number of calibration measurements were available, or the camera positions were perturbed after calibration. Two tasks were considered:

- recovery of the *epipolar constraint*, to facilitate stereo correspondence of two images of an unknown object;
- estimation of the *relative positions* of points, to facilitate reconstruction of an object and visual servoing to align a manipulator with it.

For the numerical simulations, two ideal pinhole cameras were simulated, facing the origin from a distance of 3–24 units, displaced by a rotation of 20° about a vertical axis (figure 2.2). They observed reference and test points within a unit cube centred about the origin. The focal length of the cameras varied with distance, so as to keep a constant image size (for a vertical unit vector at the origin) of 320 pel. Figure 2.3 shows the appearance of the unit cube for camera distances of 3, 8 and 24 units.

2.6.1 Epipolar constraint recovery

These experiments measure the accuracy with which the epipolar constraint may be estimated from a small number of reference points, in both the linear and fundamental matrix forms.

I. Accuracy of the linear model in noiseless simulations

With noiseless images, the fundamental matrix could be calculated with complete accuracy from 8 corresponding points. The linear approximate epipolar constraint was estimated, from 4 correspondences within the cube.[5]

For any point in the left view, an epipolar line may be predicted in the right. The normal distance between this line and the corresponding point gives us a measure of the error in the epipolar geometry model. Figure 2.4 shows the maximum and RMS error for a grid of points filling the unit cube. It can be seen that the errors due to the linear model decrease with increasing camera distance.

[5] Coordinates (-.3,-.3,-.3), (-.3,.3,.3), (.3,-.3,.3), (.3,.3,-.3): a regular tetrahedron.

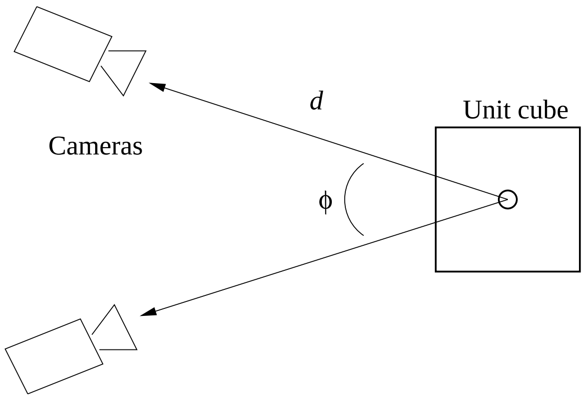


Figure 2.2: The camera geometry used in the numerical simulations. $\phi = 20^{\circ}$ and d varies from 3 to 24 units.

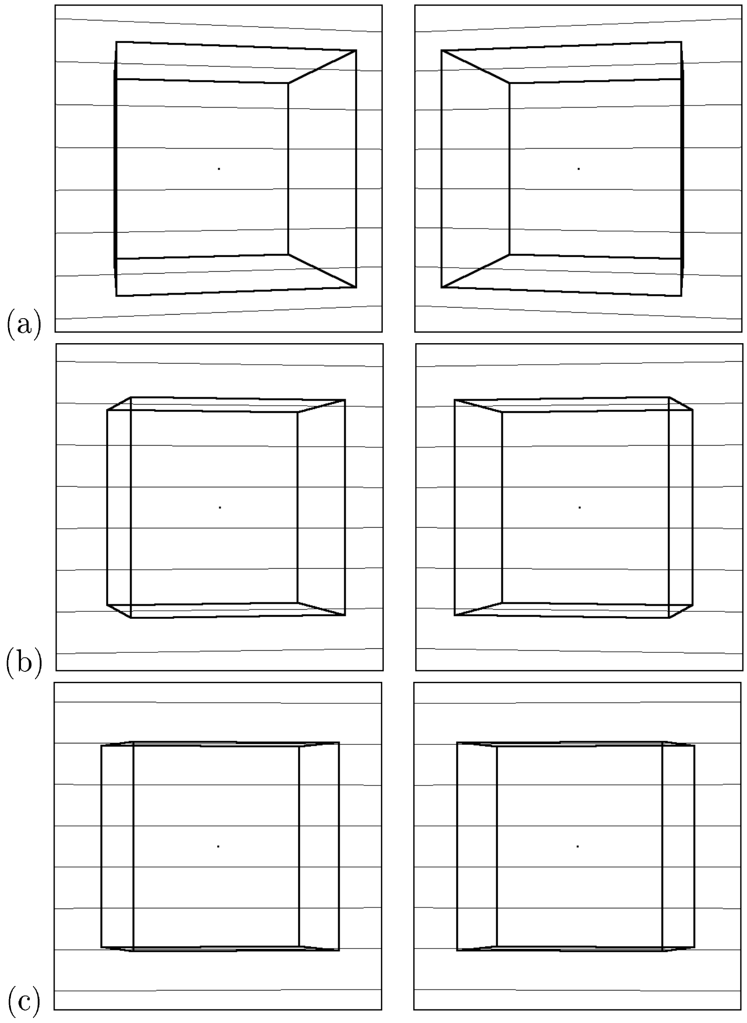


Figure 2.3: The appearance of the unit cube and epipolar lines viewed with the simulated cameras from a distance of (a) 3 (b) 8 (c) 24 units. Image size 512 pel.

II. Linear and fundamental-matrix models from noisy images

We now consider the case in which the epipolar geometry is estimated from noisy correspondences. Image coordinates of the reference points had Gaussian noise ($\sigma = 2.0$ pel) added to each axis. Linear epipolar constraints were estimated from 4 and 8 points, and a fundamental matrix from 8 points,[6] which is the minimum number for an unambiguous solution. Their accuracy was measured as above, using a grid of noiseless correspondences. Figure 2.5 shows the RMS error over 512 trials, for camera distances ranging from 3 to 24. By constraining the epipolar geometry to the linear form, greater robustness to noise is achieved.

III. From noisy image points and known world coordinates

If the world coordinates of the reference points are known, epipolar geometry may be estimated more accurately by first solving for a pair of camera models (calibration).

Reference point image coordinates had 2.0 pel noise as before, but accurate world coordinates were also available. This allowed affine and projective camera models to be estimated from 4 and 6 points respectively. The models were then rearranged to recover epipolar constraints, which take the fundamental-matrix and linear forms respectively. Figure 2.6 shows the RMS error over 512 trials, for camera distances ranging from 3 to 24. The use of world coordinates improves the estimate of the fundamental matrix, but makes no difference to the linear form in the 4-point case.

IV. Real data

For this experiment we used images of a robot to define 8 corresponding points, whose world coordinates were also known. Affine and projective camera models were estimated using linear least squares. A real scene was them observed in stereo, and a number of points of interest selected by hand in the left image. Figure 2.7 compares the epipolar line structure predicted by both affine and full perspective stereo models for matching these points. In this setup, in which the camera distances are about 2 metres, both models gave comparable accuracy — the RMS perpendicular error of the points in the right image from their predicted epipolar lines was 3.6 pel in each case. Furthermore, the affine model can predict epipolar lines using just 4 reference points with sufficient accuracy to allow matching (RMS error 4.4 pel); perspective stereo requires a minimum of 6 points.

[6] (-.3,-.3,-.3), (-.3,.3,.3), (.3,-.3,.3), (.3,.3,-.3), (.3,-.3,-.3), (.3,.3,.3), (-.3,-.3,.5), (.1, -.4, -.2).

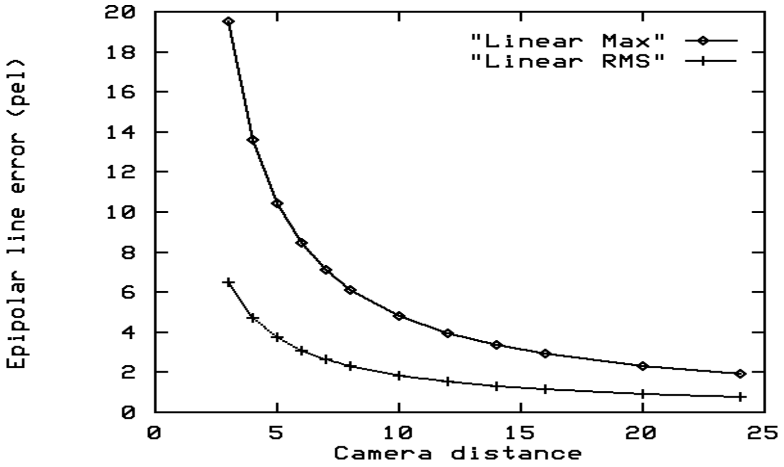


Figure 2.4: Worst-case and RMS error for the linear epipolar constraint.

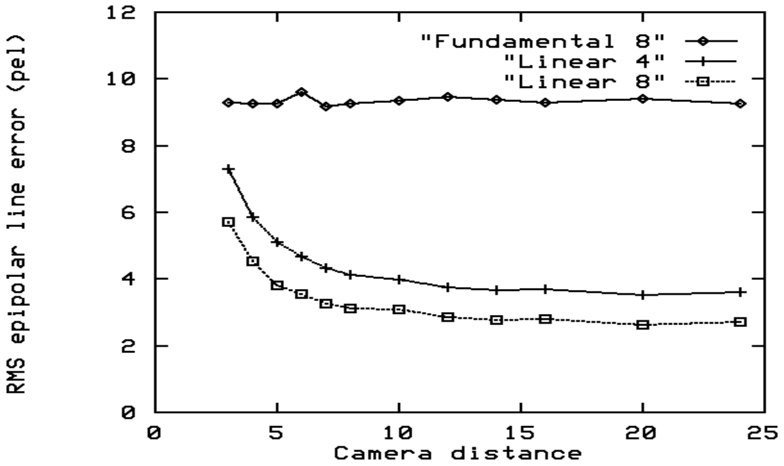


Figure 2.5: RMS error for linear and fundamental-matrix constraints, estimated from noisy correspondences ($\sigma = 2.0$ pel) by least squares.

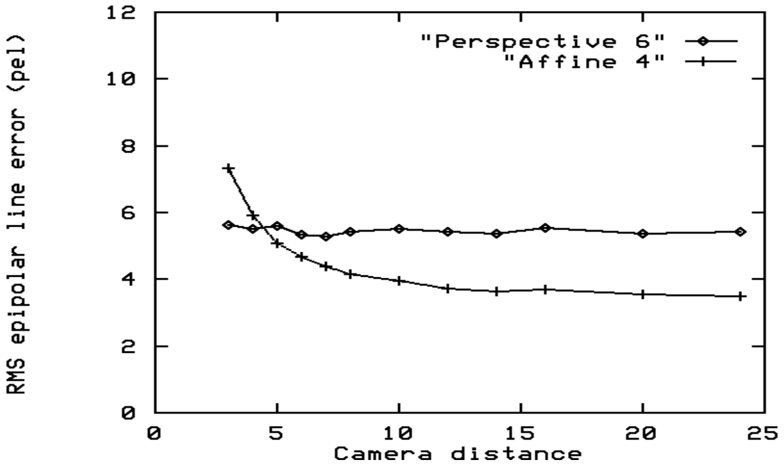
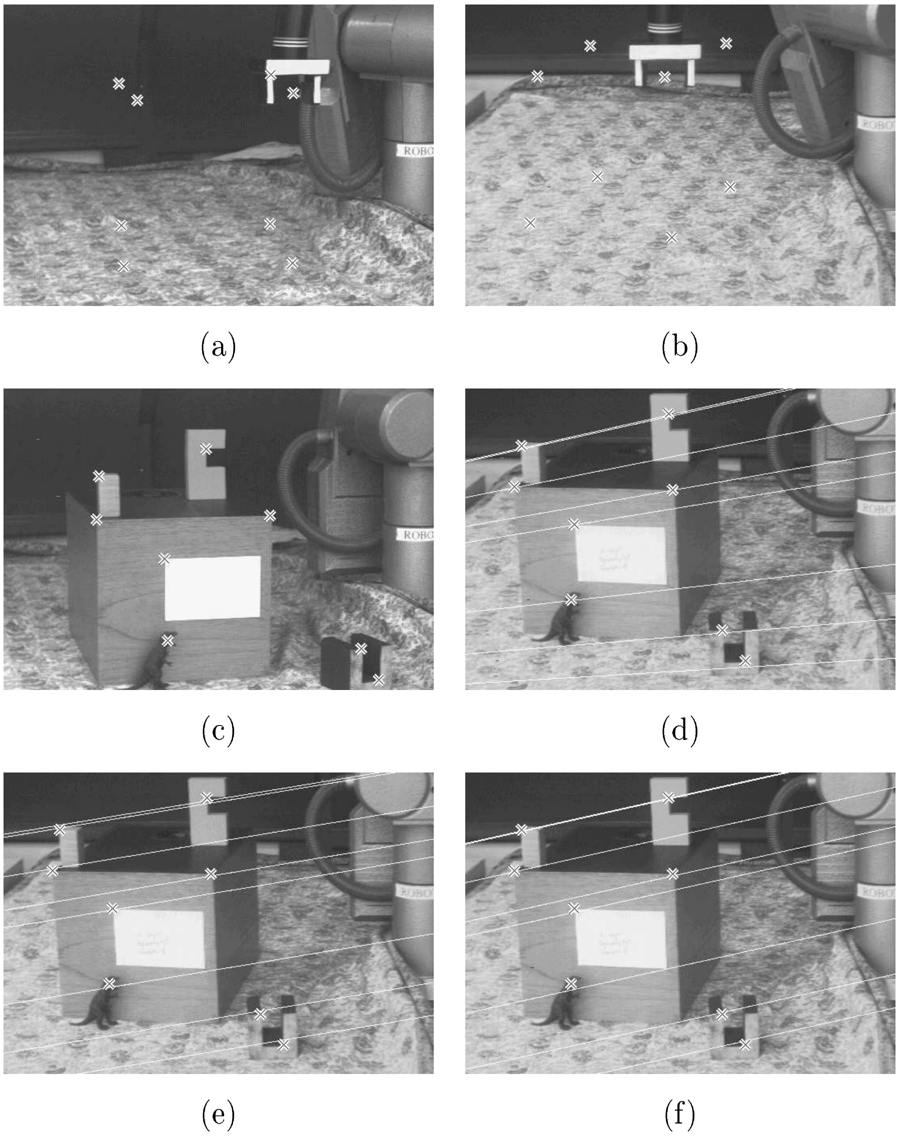


Figure 2.6: RMS error for linear and fundamental-matrix constraints, after noisy calibration of affine and perspective camera models.



(a) (b)

(c) (d)

(e) (f)

Figure 2.7: Estimation of epipolar lines: (a,b) two views of 8 reference points defined by the robot; (c) selected points in the left image; (d) epipolar lines estimated by the projective camera model after calibration with 8 points; (e,f) epipolar lines estimated by affine camera model with 8 points and 4 points respectively.

2.6.2 Accuracy of reconstruction

To compare affine and full perspective stereo reconstruction, simulations were performed measuring their ability to estimate the *relative* positions of points within the unit cube.

I. Under ideal conditions

Without noise or other disturbances, perspective stereo estimates absolute and relative positions with complete accuracy (in our 'pinhole camera' simulations, at least). An affine stereo model was calibrated using 6 reference points. At close range it performs poorly due to strong perspective distortion, but the error decreases in inverse proportion to camera distance. Figure 2.8 shows the RMS error for estimating the vector between a random pair of points within the unit cube (the average length of such a vector is 0.707).

II. With noisy calibration

Adding 2.0 pel noise to the image coordinates of the reference points causes both stereo models to lose accuracy (figure 2.9). Perspective stereo is more sensitive to noise because of its nonlinearity and greater degrees of freedom, and is *less* accurate than the affine stereo approximation at larger camera distances (viewing an increased number of reference points reduces the effects of noise and restores the accuracy of perspective stereo).

III. With noisy image coordinates after calibration

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. The two models converge for camera distances above ≈ 10 units (figure 2.10).

IV. Camera disturbances 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 2.1 shows the average change in perceived relative position when one camera is rotated or translated a small distance around/along each principle axis.

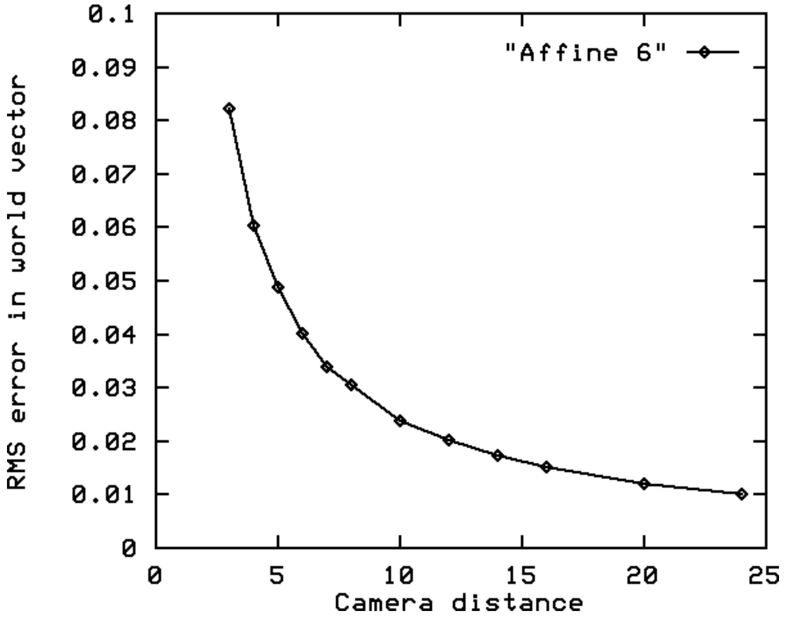


Figure 2.8: RMS relative positioning error (for random point pairs in the unit cube) as a function of camera distance, for the affine stereo model.

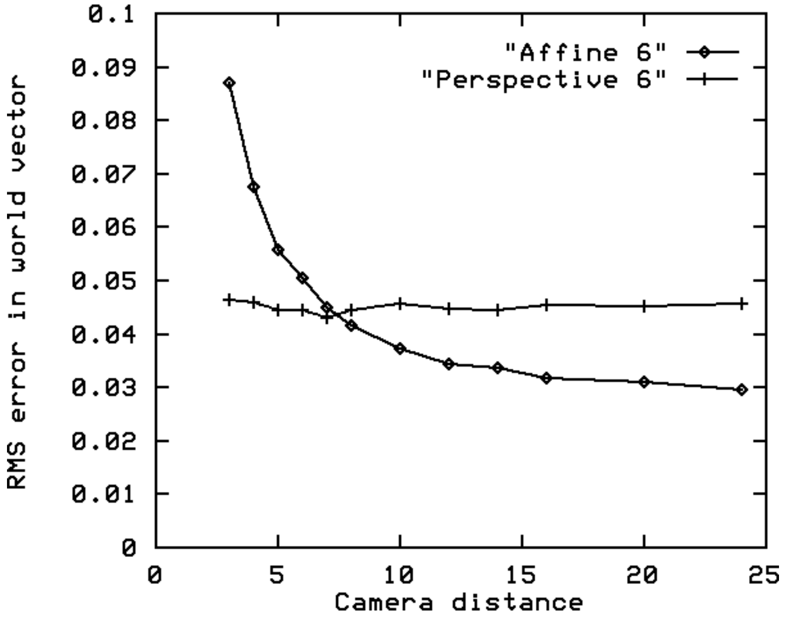


Figure 2.9: RMS relative positioning error as a function of camera distance, after calibration with 6 noisy reference points ($\sigma = 2.0$ pel).

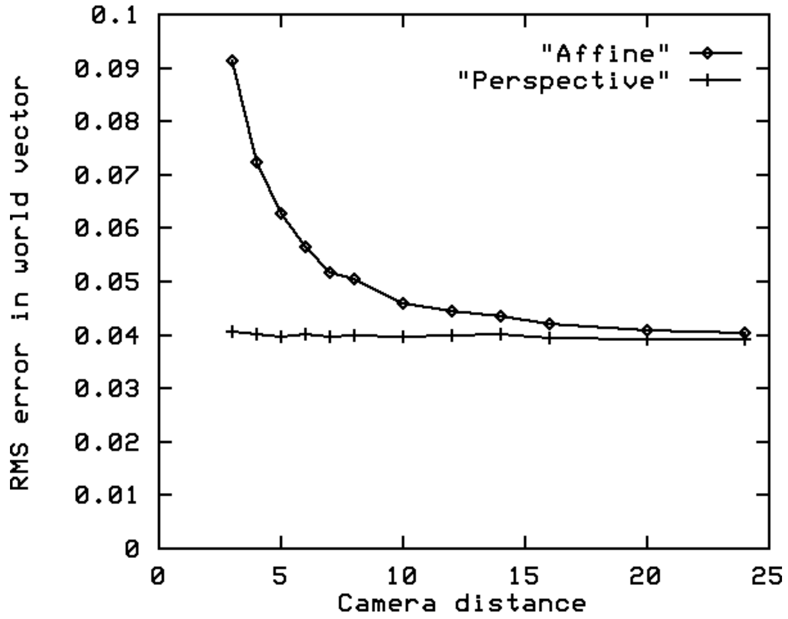


Figure 2.10: RMS relative positioning error from noisy images ($\sigma = 2.0$ pel) of world points after accurate calibration with 8 points.

Disturbance	Change (Affine)	Change (Perspective)
X_c:Y_c (roll) rotation 1°	.0214	.0214
X_c:Z_c (pan) rotation 1°	.0007	.0468
Y_c:Z_c (tilt) rotation 1°	.0006	.0049
X_c:Y_c (roll) rotation 5°	.1069	.1068
X_c:Z_c (pan) rotation 5°	.0095	.1867
Y_c:Z_c (tilt) rotation 5°	.0056	.0769
X_c (epipolar) translation 0.1	.0119	.0207
Y_c (vertical) translation 0.1	.0020	.0007
Z_c (distance) translation 0.1	.0119	.0119
X_c (epipolar) translation 0.5	.0596	.1168
Y_c (vertical) translation 0.5	.0102	.0139
Z_c (distance) translation 0.5	.0574	.0572

Table 2.1: RMS *change* to relative position estimates of world points, caused by disturbing one of the cameras after calibration. Camera distance 10 units.

The two models are affected similarly by small movements, the worst of which is X_c:Y_c rotation about the optical axis (this is the only motion which, to first order, changes the $\mathbf{Q}$ matrix of world–image coefficients).

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.

2.7 Discussion

For a typical stereo setup with two cameras fixating on a compact scene, perspective effects are small, and the epipoles will be far outside the image frames. In this case, a linear model of the epipolar constraint is valid, and the errors due to the linear approximation become comparable to other sources of error such as 'noisy' image measurements from trackers or feature detectors. It should be noted that the conditions required for linear epipolar geometry are *weaker* than those for the affine stereo model itself, which is accurate for camera distances more than $\approx$ 10 times the size of the scene.

Calibration is easier with affine stereo because the system has fewer parameters and is amenable to solution by linear techniques. Even if it could be calibrated accurately, the projective model is still more sensitive to errors and unexpected camera movements after calibration. The linear form of the affine stereo model makes it quite robust to calibration errors and changes. Even without calibration, it affords an approximate affine reconstruction of any scene with more than 4 corresponding points.

We do not attempt to use affine stereo to reconstruct *absolute* positions of points in the scene (as would be used by a look-and-move manipulation system). That would require accurately calibrated perspective camera models. Instead, we propose to use the affine model to match and reconstruct small objects in the scene, and to estimate the relative positions of nearby structures.

In chapter 3 the affine stereo formulation introduced here is used at the heart of a visual feedback controller for executing a grasp operation specified in terms of a pair of images of a graspable surface.

Chapter 3

Uncalibrated Stereo Visual Feedback

The core task in hand–eye coordination is to align a robot with a visually specified target. This chapter describes the use of visual feedback of gripper position and orientation to align it with the target object. The system does not require calibration, but estimates the affine stereo coefficients by making three deliberate motions. It is even robust to small camera motions during operation.

3.1 Introduction

If a stereo vision system were calibrated precisely, then the robot's gripper could be sent directly to the coordinates of a visually-specified target. However, this open-loop approach is sensitive to errors in calibration and kinematic modelling. Instead, we *track* the robot's gripper as it approaches the target, using *visual feedback* to correct the errors in its trajectory.

Affine stereo is a simplified stereo vision formulation that is very easily calibrated, but it is of limited open-loop 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 it as part of a visual feedback loop to align the robot gripper with its target, which is a planar facet of the object to be grasped. Image-based feedback is used to null the error in the images, so as to align their position and orientation despite camera modelling errors.

3.2 Theory

3.2.1 Point to point alignment

First, we consider aligning a point attached to the robot (or defined in terms of an affine coordinate system based on the robot) with a point specified in the images.

Let the point on the robot be P. Its position is determined by a vector of (at least 3) joint settings, $\mathbf{\Theta}$, which are related to Cartesian coordinates by the *kinematic* function $\mathcal{K}$:

$$\mathbf{X}_P = \mathcal{K}(\mathbf{\Theta}). \tag{3.1}$$

$\mathbf{X}_P = [X_P\, Y_P\, Z_P]^T$, its world coordinates in a Euclidean frame. We wish to align the robot with a visually-specified 'set point' S, specified by image coordinates $\mathbf{u}_S = [u_S\, v_S\, u'_S\, v'_S]^T$. Using the affine stereo model, we estimate its position:

$$\hat{\mathbf{X}}_S = \hat{\mathbf{Q}}^+(\mathbf{u}_S - \hat{\mathbf{u}}_0), \tag{3.2}$$

This is the inverse of equation (2.8), where $\hat{\mathbf{Q}}^+$ models the left pseudo-inverse of $\mathbf{Q}$. Suppose that we also have an inverse model $\hat{\mathcal{K}}^{-1}$ of the robot's kinematic function, (if there are more than 3 joints, assume that the redundant degrees of $\mathbf{\Theta}$ are constrained in an appropriate way). We could attempt to send the robot directly to the configuration corresponding to $\mathbf{u}_S$:

$$\mathbf{\Theta}_{OL} = \hat{\mathcal{K}}^{-1}(\hat{\mathbf{X}}_S) \tag{3.3}$$

This is the 'look-and-move' approach. It fails to compensate for inaccuracies in the inverse kinematic model $\hat{\mathcal{K}}^{-1}$ and in the camera model ($\hat{\mathbf{Q}}$, $\hat{\mathbf{u}}_0$) as well as for errors due to strong perspective distortion.

Visual feedback

By tracking the robot's gripper, we can also obtain from its image position $\mathbf{u}_P$ an estimate $\hat{\mathbf{X}}_P$ of its world coordinates,

$$\hat{\mathbf{X}}_P = \hat{\mathbf{Q}}^+(\mathbf{u}_P - \hat{\mathbf{u}}_0). \tag{3.4}$$

Feeding back the relative position term $\hat{\mathbf{X}}_P - \hat{\mathbf{X}}_S$, a simple proportional control law [120] may be devised to null the error:

$$\dot{\mathbf{\Theta}} = -g\hat{\mathbf{J}}_{\mathcal{K}}^{-1}(\hat{\mathbf{X}}_P - \hat{\mathbf{X}}_S). \tag{3.5}$$

$\hat{\mathbf{J}}_{\mathcal{K}}^{-1}$ models the inverse differential kinematic relation [48] at the current robot configuration, and g is an appropriate gain constant. The use of a term such as $\hat{\mathbf{X}}_P - \hat{\mathbf{X}}_S$ is known as *position-based feedback*. We can also express the control law in terms of the image coordinate error term $\mathbf{u}_P - \mathbf{u}_S$ (*image-based feedback*). We use our estimates of the camera and kinematic models to provide a suitable gain:

$$\dot{\boldsymbol{\Theta}} = -g\hat{\mathbf{J}}_{\mathcal{K}}^{-1}\hat{\mathbf{Q}}^{+}(\mathbf{u}_P - \mathbf{u}_S). \tag{3.6}$$

We note that, according to the affine stereo model, position-based and image-based feedback are equivalent. This is because the world–image relation is modelled as linear (cf. [49]). The combined kinematic-and-vision relation $\mathbf{Q}\mathbf{J}_{\mathcal{K}}$ (inverted in (3.6)) is sometimes called the *image Jacobian* [48].

Discrete implementation

In practice, due to the limited bandwidth between the computer vision system and the robot controller, visual feedback is implemented as a *discrete* series of relative motions of the gripper:

$$\mathbf{X}_P^*|_{t+1} = \mathbf{X}_P^*|_t - k\hat{\mathbf{Q}}^{+}(\mathbf{u}_P - \mathbf{u}_S), \tag{3.7}$$

where $\mathbf{X}_P^*$ is the vector of world coordinates passed to the inverse kinematic model; that is, $\boldsymbol{\Theta} = \hat{\mathcal{K}}^{-1}(\mathbf{X}_P^*)$. The gain term, k, governs the rate of convergence.

Convergence criteria

When does the visual feedback loop converge to the set point and when is it unstable? Define $\mathbf{X}_{err}^* = \mathbf{X}_P^* - \mathbf{X}_S^*$ where $\mathbf{X}_S = \mathcal{K}(\hat{\mathcal{K}}^{-1}(\mathbf{X}_S^*))$, and suppose that the error is small, so that a first order model of $\mathcal{K}$ may be used. Equation (3.7) becomes:

$$\mathbf{X}_{err}^*|_{t+1} = (\mathbf{I} - k\hat{\mathbf{Q}}^{+}\mathbf{Q}\mathbf{J}_{\mathcal{K}}\hat{\mathbf{J}}_{\mathcal{K}}^{-1})\, \mathbf{X}_{err}^*|_t\,. \tag{3.8}$$

The error term will vanish and the system converge to the set point[1] only if all the eigenvalues of $(\mathbf{I} - k\hat{\mathbf{Q}}^{+}\mathbf{Q}\mathbf{J}_{\mathcal{K}}\hat{\mathbf{J}}_{\mathcal{K}}^{-1})$ have absolute magnitude below unity [122]. For a perfectly modelled system, $\hat{\mathbf{Q}}^{+}\mathbf{Q}\mathbf{J}_{\mathcal{K}}\hat{\mathbf{J}}_{\mathcal{K}}^{-1} = \mathbf{I}$ and the set point is reached in one step by setting $k = 1$.

Setting $0 < k < 2$ also leads to convergence, but values above 1 will cause overshoots and ringing (which, in a robotic application, could lead to collisions!) To prevent this whilst allowing for some inaccuracy in kinematic and camera modelling, k should be set significantly below unity, e.g. $k = 0.75$.

[1] We assume here that the set point is stationary.

3.2.2 Position and orientation alignment

Suppose now that we wish to align a planar surface on the robot's end effector with one specified in the image. Alignment of position and surface orientation is a 5-DOF constraint; additionally, if a vector on the robot is to be aligned with a distinguished image direction, there are constraints on all 6 components of robot pose.

Recall from section 2.5, that the orientation of a surface is encoded by its *affine transformation between views*, $\mathbf{A}$. This has only two degrees of freedom and may be represented by two components (a_{11}, a_{12}); the other two components can be obtained using the epipolar constraint. An image-based representation of surface orientation[2] is thus the vector $\mathbf{o} = [a_{11}\, a_{12}]^T$. The surface normal direction itself may easily be obtained from $\mathbf{o}$ and an estimate of the $\mathbf{Q}$ matrix.

Image-based feedback of surface orientation

Let the robot now be controlled in terms of a desired position and orientation, where the orientation is expressed in image-based terms:

$$\mathbf{\Theta} = \hat{\mathcal{K}}^{-1}(\mathbf{X}^*_P,\, \hat{\mathcal{F}}(\mathbf{o}^*_P)), \tag{3.9}$$

where $\hat{\mathcal{K}}^{-1}$ is an inverse kinematic model for both position and orientation control, and $\hat{\mathcal{F}}$ is a function to convert image-based orientations into some other parameterisation used by the robot.[3]

A suitable control law to align the gripper with a target is:

$$\dot{\mathbf{\Theta}} = -g(\frac{\partial \mathbf{\Theta}}{\partial \mathbf{X}^*_{\mathbf{P}}} \hat{\mathbf{Q}}^+(\mathbf{u}_P - \mathbf{u}_S) + \frac{\partial \mathbf{\Theta}}{\partial \mathbf{o}^*_{\mathbf{P}}}(\mathbf{o}_P - \mathbf{o}_S)). \tag{3.10}$$

Again, in practice it is convenient to use a discrete implementation, in which a sequence of position and orientation demands are made:

$$\begin{aligned} \mathbf{X}^*_P|_{t+1} &= \mathbf{X}^*_P|_t - k_1 \hat{\mathbf{Q}}^+(\mathbf{u}_P - \mathbf{u}_S), \\ \mathbf{o}^*_P|_{t+1} &= \mathbf{o}^*_P|_t - k_2(\mathbf{o}_P - \mathbf{o}_S). \end{aligned} \tag{3.11}$$

As before, the gain parameters k_1, k_2 should be set to values between 0 and 1.

[2] A third image-based parameter may be added to this vector to specify all 3 DOF of orientation.

[3] $\hat{\mathcal{F}}$ depends on an estimate of $\mathbf{Q}$.

3.3 Simulations

A simple articulated robot was simulated, its origin at coordinates (−2.0, 0, −.5), with two links of length 1.5 units (figure 3.1). The position of the end-effector was governed by three angles: *waist* (θ_1), *shoulder* (θ_2) and *elbow* (θ_3); the kinematic function was:

$$\begin{aligned} X_P &= 1.5\cos\theta_1(\cos\theta_2 + \cos(\theta_2 - \theta_3)) - 2.0, \\ Y_P &= 1.5\sin\theta_1(\cos\theta_2 + \cos(\theta_2 - \theta_3)), \\ Z_P &= 1.5(\sin\theta_2 + \sin(\theta_2 - \theta_3)) - 0.5. \end{aligned} \tag{3.12}$$

The same simulated cameras were used here as in section 2.6, facing the origin from a distance of 4.0 units. The inter-camera angle was 20°. Affine stereo coefficients were estimated by observing 4 known points in a tetrahedron within the unit cube.

Setpoints were enumerated on a dense grid of points within the unit square and the robot aligned with those points:

- **Open loop**, by inverting kinematic and camera models;
- **Closed loop**, using visual feedback with $k = 1$;
- **Closed loop**, using visual feedback with $k = 0.5$.

When using visual feedback, the initial position of the end effector in each trial was the world origin in the centre of the unit cube.

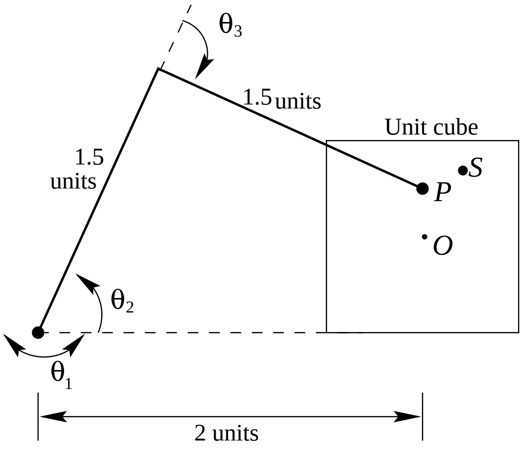


Figure 3.1: Articulated robot model used in the simulations

I. Ideal case

An inverse kinematic model [120] was derived analytically from equation (3.12), and the camera coefficients estimated using noiseless reference points. In open loop, the RMS positioning error for a point within the cube was .068 units, and the maximum error .157 units. These errors are due to perspective distortion. With visual feedback, however, the errors are reduced practically to zero (results are summarised in table 3.1). Figure 3.2(a) shows the trajectory of the robot when the set point is $(.5, .5, .5)$ with $k = 0.5$. It is almost a straight line.

II. With erroneous kinematic model

The simulations were repeated, using a modified inverse kinematic model which moved θ_1 through 1.5 times the desired angle and added a 10° offset to θ_3. This seriously degraded open-loop positioning accuracy; however visual feedback with $k = 0.5$ was able to correct the errors. In this case, better performance was obtained with $k = 0.5$ than with $k = 1$, which lead to 'ringing' and failure to converge in some regions of the robot's configuration space. See figure 3.2(b).

III. After camera disturbances

This time the correct kinematic model was used, but the camera pose was changed between observation of the reference points and alignment with the set points. One of the cameras was translated 0.25 units upwards, the other rotated 10° about its optical axis. Again, visual feedback was able to null the errors. See figure 3.2(c).

	Open loop		$k = 1$		$k = 0.5$	
	RMS	Max	RMS	Max	RMS	Max
No disturbance	.068	.157	.0001	.0007	.013	.034
Kinematic error	.263	.401	.036	.162	.012	.026
Camera disturbance	.197	.361	.003	.023	.025	.071

Table 3.1: Results of simulations in which the end-effector was aligned with points on a dense grid within the unit cube. RMS and maximum positioning errors after 6 iterations are shown.

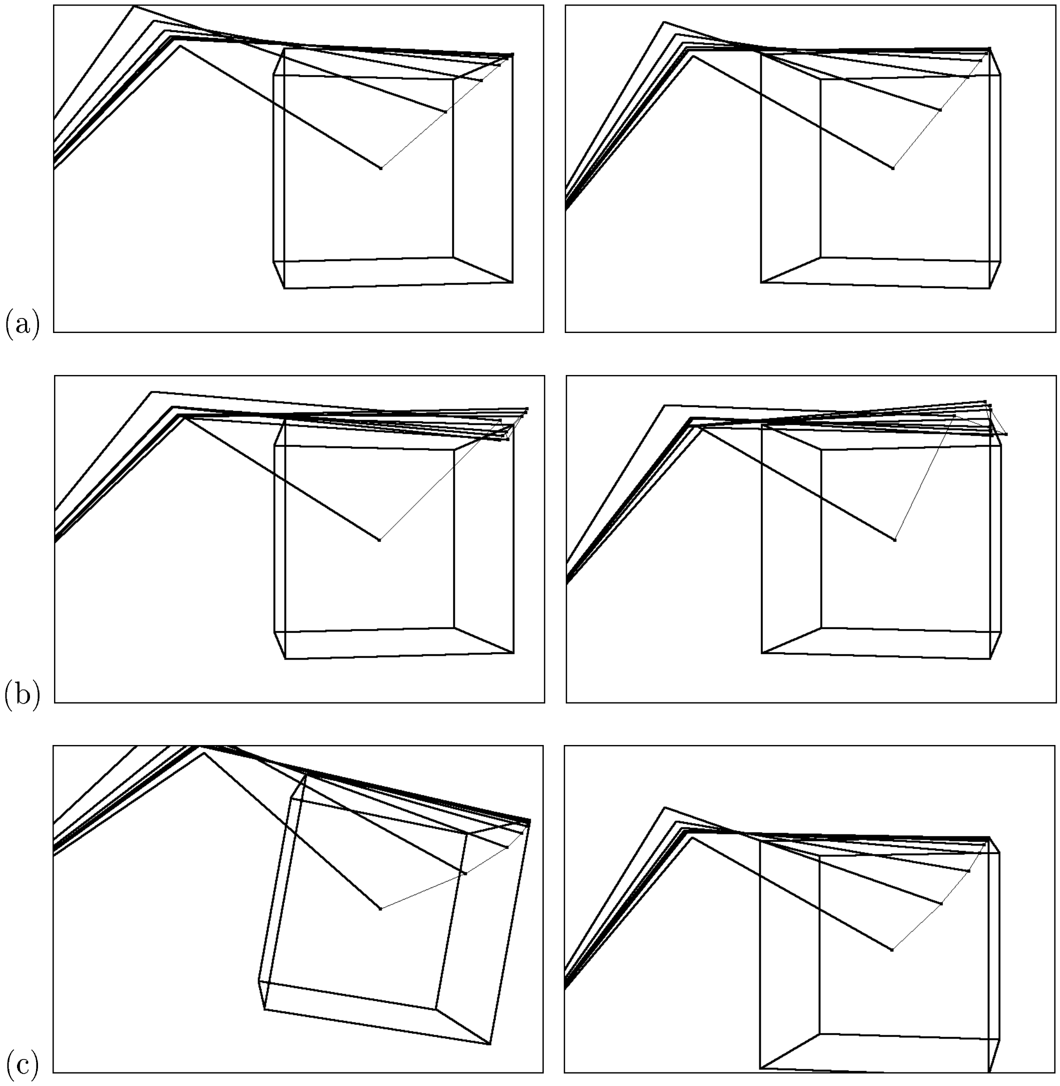


Figure 3.2: Simulated robot trajectories under visual feedback. The end effector is converging on one corner of the unit cube: (a) ideal case ($k = 0.5$); (b) with erroneous kinematic model ($k = 1$) showing 'ringing' behaviour; (c) after camera disturbances ($k = 0.5$).

3.4 Experiment

3.4.1 Setup

When the system was started up, it began by opening and closing the jaws of the robot's gripper. By observing the image difference, it was able to locate the gripper and set up a pair of affine trackers as instances of a hand-made 2-D template. The trackers could then follow the gripper's movements continuously. Stereo tracking was implemented on the Sun at over 10 Hz. The robot then made a series of deliberate motions, moving to four preset points to estimate the coefficients matrix $\mathbf{Q}$.

Since the reference points used to self-calibrate were specified in the *controller's* coordinate space ($\mathbf{X}^*$), linear errors in the kinematic model were effectively bypassed. The system must still cope with any nonlinearities in control, as well as those caused by strong perspective effects.

A target object was located by similar means — by observing the image changes when it was placed in the manipulator's workspace. Alternatively it could be selected from a monitor screen using the mouse. There was no pre-defined model of the target shape, so a pair of 'exploding' B-spline snakes [21] were used to automatically locate the contours delimiting the target surface in each of the images. The snakes were converted into a pair of affine trackers, by re-expressing their sampling points in terms of an affine basis (see appendix).

The target surface was then tracked along with the gripper, to compensate for unexpected motions of either the target or the cameras during operation.

3.4.2 Visual feedback loop

The orientation of the gripper of a 5-DOF manipulator is constrained by its lack of a 'yaw' axis, and the constraint changes continuously as it moves. To avoid this problem, the test implementation kept the gripper vertical, reducing the number of degrees of freedom to four. Its orientation could then be described by a single *roll angle*. It was assumed that the target plane was also vertical. Their image orientations were therefore described by a single quantity, a_{11}.[4]

The gains for position and orientation control are set well below unity at 0.75, to prevent instability, even when the vision system is miscalibrated. The control structure of the system is shown in figure 3.4.

[4] It is assumed that the camera baseline is roughly horizontal, so that a_{11} varies with roll angle.

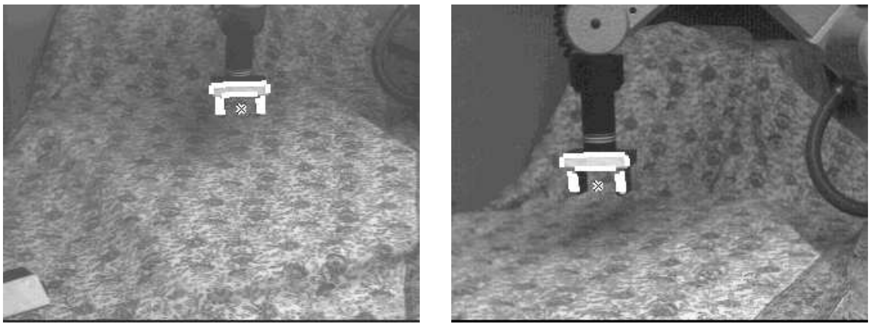


Figure 3.3: A stereo pair showing the robot gripper at one of the four reference points used for calibration. Active contour models are overlaid in white.

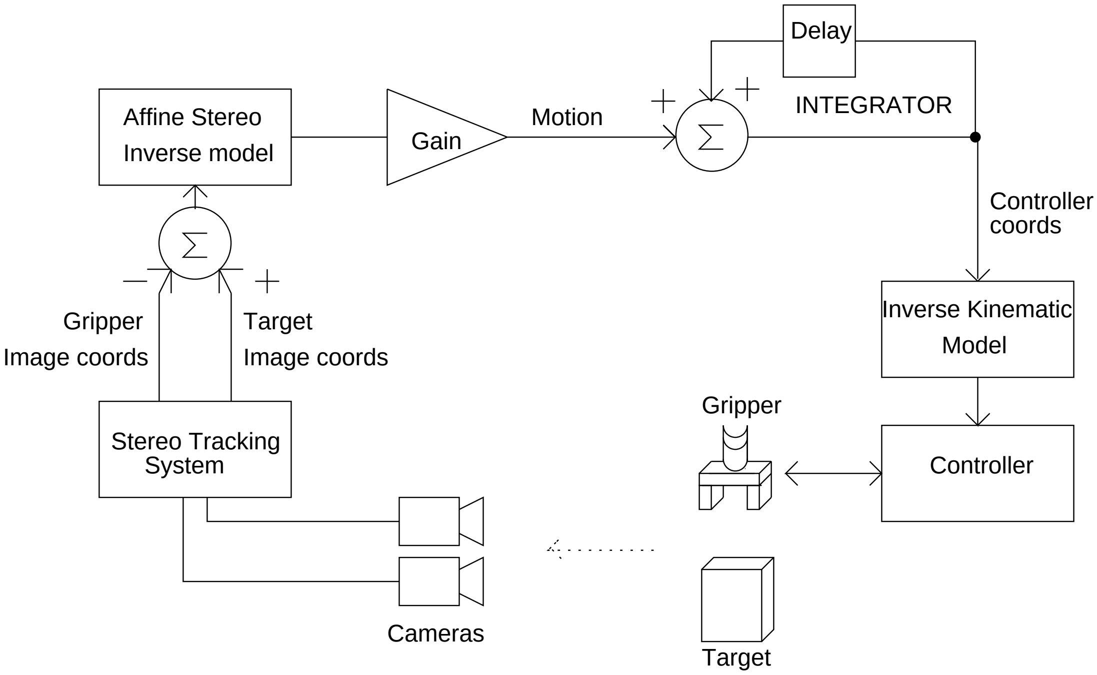


Figure 3.4: The control structure of the system, showing the use of visual feedback.

3.4.3 Tracking and grasping behaviours

Without modification, the visual feedback loop would attempt to superimpose the robot gripper and target object in the images. By offsetting $\mathbf{u}_P$ from the gripper's centre, we introduce a constant offset between gripper and target in space; the offset is defined in terms of a coordinate system attached to the gripper (in fact, the affine basis of the tracking mechanism), so that it will be invariant to motions of the cameras. We set the offset so that the robot tracks the target object continuously, hovering a few centimetres above a point on its top surface (figure 3.5).

Once this *pre-grasp* position has been achieved, the object may be grasped reliably using a pre-programmed motion, which consists of rotating the gripper through 90° and translating downwards (figure 3.6). Depending on the type and shape of object to be grasped, some other grasping motion could be substituted here.

3.4.4 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 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 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 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

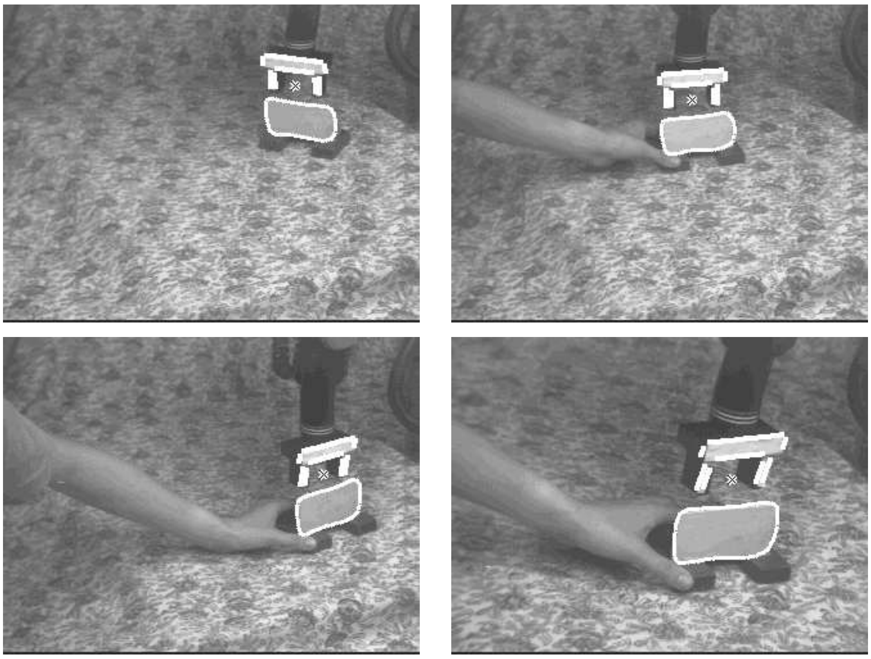


Figure 3.5: The robot is tracking its quarry, guided by the position and orientation of the target contour (view through left camera). On the target surface is an *affine snake* — an affine tracker obtained by ‘exploding’ a B-spline snake from the centre of the object. Last frame: one of the cameras has been rotated and zoomed, but the system continues to operate successfully with visual feedback.

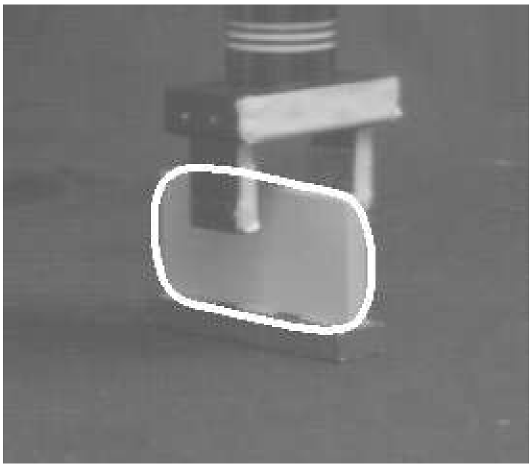


Figure 3.6: Robot grasping a planar target, using an active contour to recover its size and orientation. The gripper is not tracked during the grasping manoeuvre.

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.

Figure 3.5 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.

3.4.5 Why not track Q?

Since the visual feedback system has been designed to be robust to changes in the camera parameters (caused by movement of the cameras) during operation, an obvious question is whether or not efficiency can be improved by tracking these changes. This was attempted in a version of the above experiment, using a Kalman filter [42] whose state vector encodes the camera model $(\hat{\mathbf{Q}}, \hat{\mathbf{u}}_0)$, which is updated from subsequent observations of the robot gripper. However, this conferred little or no detectable benefit to the performance of the system.[5] This is because it is impossible, from a single observation of the gripper, to determine if an error in the image location of the gripper is due to:

1. Strong perspective (temporary change in $\mathbf{Q}$, $\mathbf{u}_0$),

2. Change in $\mathbf{u}_0$ caused by small camera rotations or translations,

3. Change in $\mathbf{Q}$ caused by large camera translations, zooming, or rotation about the optical axis.

Only the last of these warrants tracking, and this was the least frequent change to be observed. The errors due to perspective could to some extent be modelled by ‘observation noise,’ but there were no obvious values for ‘process noise’ to enable the other parameter changes to be distinguished. It was concluded that attempting to track camera motions was not only ill-conditioned but also unnecessary.

[5] Except for a validation gate on the gripper’s image coordinates, which was very useful for detecting failures of the trackers and reinitialising them.

3.5 Discussion

Here the effectiveness of affine stereo has been demonstrated for the task of aligning a robot with a visually specified target, in both position and orientation. In a discrete-time implementation, rapid convergence is achieved with a gain of unity; though if the system is disturbed from its initial configuration, the gain should be reduced to maintain stability and prevent overshoots which could lead to collisions.

The visual servoing system does not require camera calibration, but makes a small number of deliberate motions to actively estimate the relation between hand and eye. Even these are not always necessary, for instance if the cameras have been rotated and then realigned by hand, the previous estimate of $\mathbf{Q}$ will normally still be valid. It is not necessary, or even practical, to track these coefficients over time.

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 visual 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).

Chapter 4

Indicating the Target Object

This chapter describes a human–computer interface which tracks a pointing hand, in order to specify objects and locations for robotic pick and place operations. The system is implemented using uncalibrated stereo vision.

4.1 Introduction

In order to make use of visual feedback in uncalibrated stereo, the target object must be indicated to the system in terms of *image* measurements. If there is more than one object visible in the scene, some means must be chosen to select the desired object for grasping, and to indicate the place to which it is to be moved.

This could be accomplished using a mouse to indicate points in one or both images. This is reliable if somewhat inelegant, and requires a workstation, or similar user-interface hardware, in close proximity to the work area. Alternatively, the operator could interact with the cameras already in place to indicate the target directly. The latter approach is explored here. An interface based on *pointing* is developed, to select objects on a planar table top.

We use a pair of monochrome cameras to observe the robot's work space and pointing hand in stereo. Active contours are employed to track the hand in real time. Using a simple result from projective geometry, the system can calculate where the hand is pointing to on the plane, without camera calibration, to an accuracy of about 10mm.

4.2 Geometrical framework

A single view of a pointing hand is ambiguous: its distance from the camera cannot be determined, and the 'slant' 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. A second view is needed to fix its position in two dimensions [106].

4.2.1 Viewing the plane

Consider a pinhole camera viewing a plane. The viewing transformation is a plane collineation between some world coordinate system (X, Y), and image plane coordinates (u, v), thus:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \mathbf{T} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \tag{4.1}$$

where $\mathbf{T}$ is a 3×3 transformation matrix. The full perspective form of the transformation is used in this case because the workspace will generally be large and possibly foreshortened in one or both images.

The system is homogeneous, so we can fix $t_{33} = 1$ without loss of generality, leaving 8 degrees of freedom. To solve for $\mathbf{T}$ we must observe at least four points. By assigning arbitrary world coordinates to these points (e.g. $(0, 0)$, $(0, 1)$, $(1, 1)$, $(1, 0)$), a new coordinate system on the plane is defined, which we call *working plane coordinates*.

Now, given the image coordinates of a point anywhere on the plane, along with the image coordinates of the four reference points, it is possible to invert the relation and recover the point's working plane coordinates, which are invariant to the choice of camera location [88]. The same set of reference points in the world can be observed in a stereo pair of views, to compute two transformations $\mathbf{T}$ and $\mathbf{T}'$, one for each camera.

4.2.2 Recovering the indicated point in stereo

With natural human pointing behaviour, the hand is used to define a line in space, passing through the base and tip of the index finger. This line will not generally be in the ground plane but intersects the plane at some point. It is this point (the *'piercing point'* or *'indicated point'*) that we aim to recover. Let the pointing finger lie along the line l_w in space (see figure 4.1). Viewed by a camera, it appears on line l_i in the image, which is also the projection of a *plane*, $\mathcal{P}$, passing through the image line and the optical centre of the camera. This plane intersects the ground plane $\mathcal{G}$ along line l_{gp}. It can be seen that l_w lies in $\mathcal{P}$, and the indicated point in l_{gp}, but from one view we cannot see exactly where.

Note that the line l_i is an image of line l_{gp}; that is, $l_i = \mathbf{T}(l_{gp})$, where $\mathbf{T}$ is the projective transformation[1] from equation (4.1). If the four reference points are visible, this transformation can be inverted to find l_{gp} in terms of the working plane coordinates. The indicated point is constrained to lie upon this line on the plane.

Repeating the above procedure with the second camera C$'$ gives us another view l'_i of the finger, and another line of constraint l'_{gp}. The two constraint lines will intersect at a point on the ground plane, which is the indicated point. Its position can now be found in terms of the projective basis formed from the four reference points. This is similar to a construction used by Quan and Mohr [106], who present an analysis based on cross-ratios. Figure 4.2 shows the lines of pointing in a pair of images, and the intersecting constraint lines in a 'canonical' view of the working plane (in which the reference point quadrilateral is transformed to a square).

By transforming this point with matrices $\mathbf{T}$ and $\mathbf{T}'$, the indicated point can be projected back into image coordinates. Although the working plane coordinates of the indicated point depend on the configuration of the reference points, its back-projections into the images do not. Because all calculations are restricted to the image and ground planes, explicit 3-D reconstruction is avoided and no camera calibration is necessary. By tracking at least four points on the ground plane, the system can be made insensitive to camera motions.

4.2.3 Projective versus affine transformations

Assuming a weak perspective view of the plane, we could substitute an affine transformation between views for the projective one: this would require only 3 reference points. However, in this case there is little gain in robustness or simplicity using the affine model that would offset the loss of accuracy caused by perspective distortion.

[1] This is a slight abuse of notation, since for the standard representation of a line the appropriate transformation matrix is $\mathbf{T}^{-1}$. Here $\mathbf{T}()$ refers abstractly to a plane projective transformation which may be applied to points, lines or other image features.

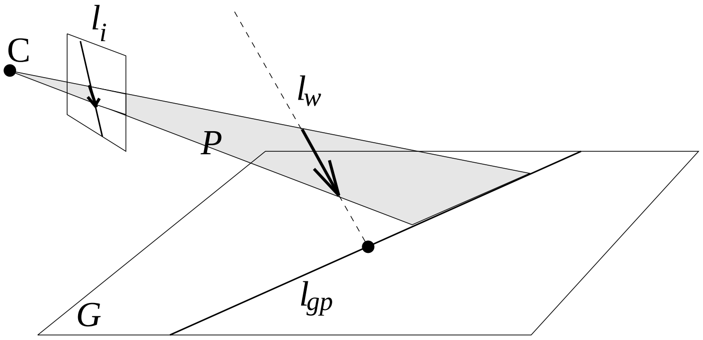


Figure 4.1: Relation between lines in the world, image and ground planes

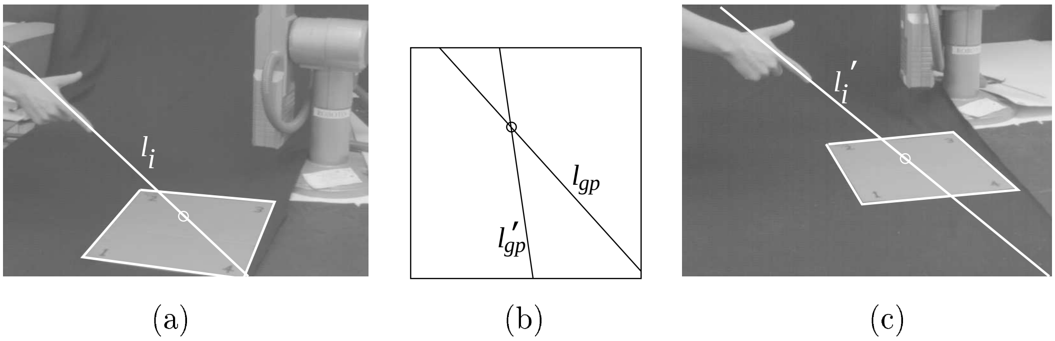


Figure 4.2: Pointing at the plane. By taking the lines of pointing in left and right views (a, c), transforming them into the canonical frame defined by the four corners of the grey rectangle (b), and finding the intersection of the lines, the indicated point can be determined; this is then projected back into the images.

This is because we are considering points on a *plane*, using only 2-D projective transformations: these do not suffer the same sensitivity to noise as would a full 3-D reconstruction.[2] Errors in localising the 4 reference points result in only local inaccuracies in the projective transformation (see table 4.1, page 65).

We are interested, in the first instance, in the *open-loop* accuracy with which the indicated point may be recovered. With the camera setup used in these experiments, the ground plane is large and significantly foreshortened, and this would cause significant errors in a formulation based on affine transformations.

4.2.4 Pointing in a multi-faceted environment

The above geometrical framework relies on the target surface being planar in order to estimate the constraint lines l_{gp}, l'_{gp} and their intersection. This can be extended to environments consisting of more than one plane.

For each planar surface, we need 4 corresponding points, and a description of the surface's boundary, e.g. as a polygon, in either view (recall that the 4 points define a transformation between views, allowing the boundary to be 'transferred' into the other image). Given two views of a pointing hand, we can now ascertain which facet is being pointed to as follows:

- For each facet, test if the pointing line in each view intersects the facet's image boundary in that view.
- If so, solve for the *piercing point* and test that it too lies within the boundary of the facet.
- Where the pointing line intersects more than one facet, choose the one nearest to the fingertip. Distances to the fingertip of points along this line may be compared in either image.

Note that whilst this requires at least 4 correspondences per facet,[3] and *a priori* models of the surfaces and their boundaries in the images, the entire process is image-based and does not rely on a 3-D reconstruction of the hand or the environment.

[2] This is partly due to our choice of working plane coordinates and the use of four reference points in a rectangle, resulting in a well-conditioned **T** which is close to an affine transformation.

[3] For smaller facets which are not strongly foreshortened, 3 correspondences may suffice and an affine stereo model can be used.

4.3 Tracking a pointing hand

4.3.1 Background

There has been a lot of interest lately in the use of hand gestures for human–computer interfacing: they are intuitive for the operator, and provide a rich source of information to the machine. This type of interface is particularly appropriate in applications such as virtual reality, multimedia and teleoperation [123, 40, 9]. Most current commercial implementations rely on sensors that are physically attached to the hand, such as the 'DataGlove' [39]. More recently, systems have been proposed using *vision* to observe the hand. Some require special gloves with attachments or markings to facilitate the localisation and tracking of hand parts [135, 26], but others operate without intrusive hardware. This is attractive because it is convenient for the user and potentially cheaper to implement.

A large number of systems have been proposed for visual tracking and interpretation of hand and finger movements without gloves. These systems can broadly be divided into:

- those concerned with gesture identification (e.g. for sign language), which compare the image sequence with a set of standard gestures using correlation and warping of the templates [29], or classify them with neural networks [13];
- those which try to reconstruct the pose and shape of the hand (e.g. for teleoperation) by fitting a deformable, articulated model of the palm and finger surfaces to the incoming image sequence [69].

Common to many of these systems is the requirement to calibrate the templates or hand model to suit each individual user. They also tend to have high computational requirements, taking several seconds per frame on a conventional workstation, or expensive multiprocessor hardware for real time implementation.

4.3.2 Approach

Our approach differs from these general systems in an important respect: **we wish only to recover the line along which the hand is pointing**, to be able to specify points on a ground plane. This considerably reduces the number of degrees of freedom which we need to track. Furthermore, because the hand must be free to move about as it points to distant objects, it will occupy only a relatively small fraction of the pixel area in each image, reducing the number of features that can be distinguished.

In this case it is not unreasonable to insist that the user adopt a rigid gesture. For simplicity, the familiar 'pistol' pointing gesture was chosen. The pointing direction can now be recovered from the image of the index finger, although the thumb is also prominent and can be usefully tracked. The rest of the hand, which has a complicated and rather variable shape, is ignored. This does away with the need to calibrate the system to each user's hand.

4.3.3 Tracking mechanism

A form of edge-seeking active contour model [64, 22, 56] was used to track the image of a hand in the familiar 'pointing' gesture, in real time. The tracker is an active contour, resembling a B-Spline snake [22], but constrained to deform only affinely in the images. It is based on a template, representing the shape of the occluding contours of an extended finger and thumb (see figure 4.3).

The tracker's motion is restricted to 2-D affine transformations in the image plane, which ensures that it keeps its shape whilst tracking the fingers in a variety of poses. This approach is suitable for tracking planar objects under weak perspective [12]; however it also works well with fingers, which are approximately cylindrical.

A first-order temporal filter is incorporated into the tracker, to predict the future position of the contour, improving its real-time tracking performance. The filter is biased to favour rigid motions in the image, and limits the rate at which the tracker can change scale — these constraints represent prior knowledge of how the hand's image is likely to change, and increase the reliability with which it can be tracked. The dynamics of the tracker are described in more detail in appendix A. It is similar to the trackers we use to track the robot's gripper in stereo images, to provide visual feedback.

To extract the hand's direction of pointing, we estimate the orientation of the index finger by fitting a pair of parallel lines to its image edges. The base of the thumb is also tracked to define the length of the index finger, and to resolve an *aperture problem* [131] induced by the finger's long thin shape.

4.4 Pointing experiment

The above geometrical framework and tracking mechanism were implemented, to indicate points on a planar table top with a pointing hand. The two cameras were about 2m from the scene, angled about 20° apart.

4.4.1 Setup

In this experiment, the corners of a coloured rectangle on the table-top were used to define the working coordinate system. A pair of finger-trackers (one for each camera) were initialised, one after the other, by the operator holding his or her hand up to a template in the image and waiting a few seconds while it 'moulded' itself to the contours of the finger and thumb. Once both trackers were running, the hand could be used as an input device by pointing to places on the table-top. In this implementation, the position and orientation of the finger trackers, and the indicated point on the plane, were updated about 10 times per second.

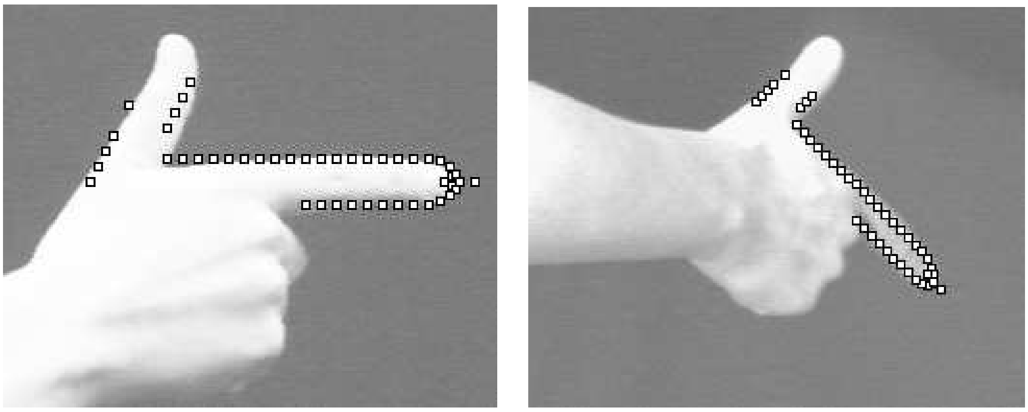
4.4.2 Performance

Figure 4.4 shows the system in operation. The corners of the white rectangle are the four reference points, and the overlaid square shows the position of the indicated point. Movements of the operator's hand caused corresponding movements of this point in real time.

Visual tracking can follow the hand successfully for several minutes at a time; however, abrupt or non-rigid hand movements could cause one or both of the trackers to fail. Because it samples the image only locally, a failed tracker will not correct itself unless the user makes a special effort to recapture it.

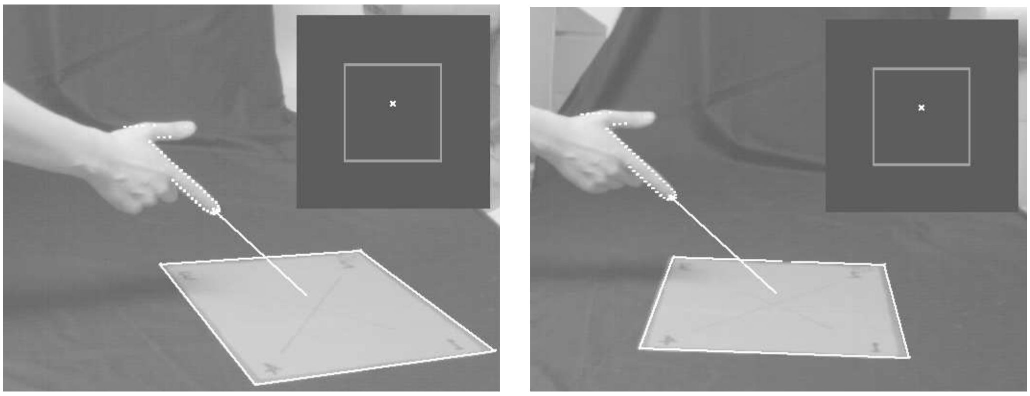
Users reported that the recovered point did not always correspond to their subjective pointing direction, which is related to the line of sight from *eye* to fingertip as well as the orientation of the finger itself. Initial subjective estimates of accuracy were in the order of 20–40mm. If the user received feedback by viewing the system's behaviour on a monitor screen, a resolution within 10mm could be achieved. It is a natural human skill to servo the motion of one's hand to control a cursor or other visual indication.

The system was also tested in a multi-planar environment (figure 4.5). The planes were represented by 9 given correspondences, which also defined bounding quadrilaterals. The user could then indicate points on 3 surfaces: transition between planes occurred automatically as the piercing point crossed their boundaries.



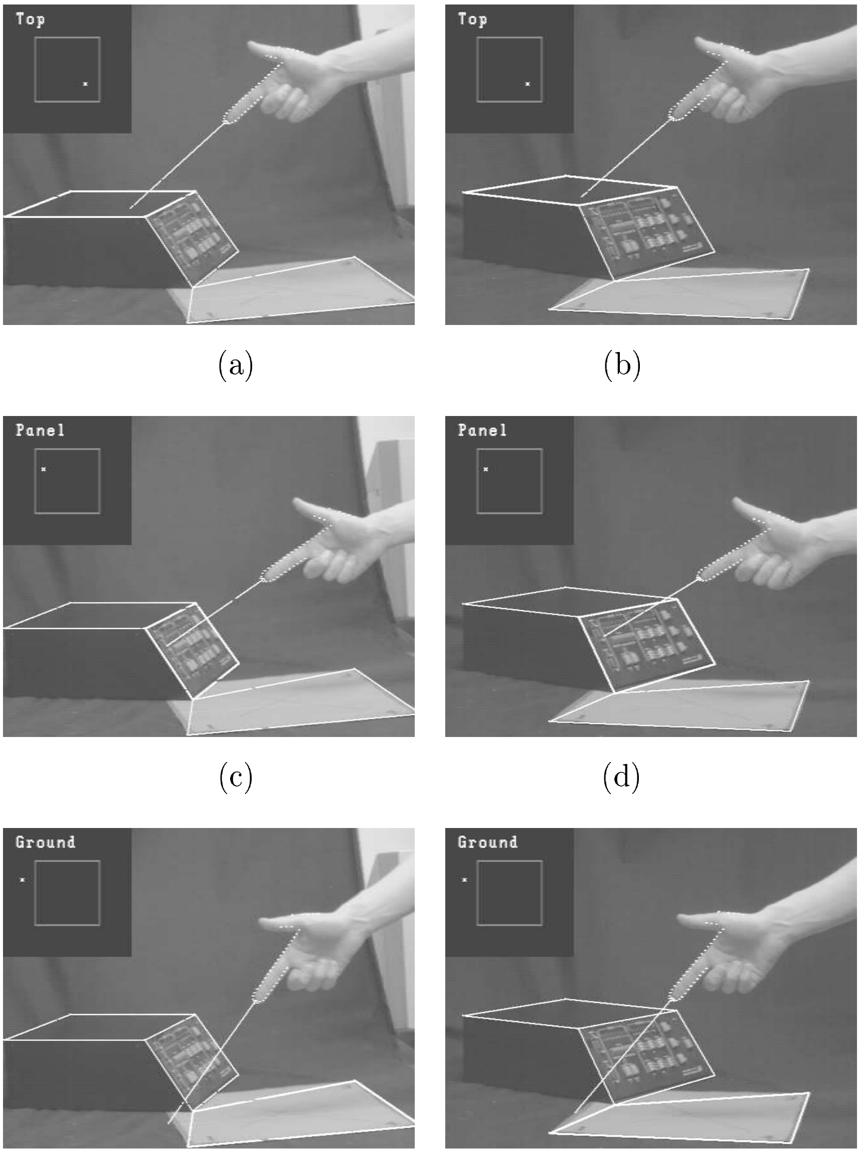
(a) (b)

Figure 4.3: The finger-tracking active contour (a) in its canonical frame (b) after an affine transformation in the image (to track a rigid motion of the hand in 3-D).



(a) (b)

Figure 4.4: 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 a representation of the indicated point in working plane coordinates.



(a) (b)

(c) (d)

(e) (f)

Figure 4.5: Pointing in a multi-planar environment: (a,b) pointing to the top surface of the object; (c,d) pointing to the sloping panel; (e,f) if the pointing line intersects neither of the above surfaces, it defaults to the ground plane.

4.4.3 Accuracy evaluation

To evaluate our system, we calculated the uncertainty of the image coordinates of the hand and reference points in our experimental setup. Using Monte Carlo methods, these were propagated into working plane coordinates, to assess the accuracy of the indicated point.

I. Finger tracker uncertainty

We can obtain a measure of uncertainty for the finger's position and orientation in the image by considering the *residual offsets* between modelled and observed image edges. These are the components of the normal offsets that remain after fitting a pair of parallel lines to model the index finger's occluding edges, with least-squares perpendicular error. They take into account the effects of image noise and occlusion, as well as pixel quantisation effects, and mismatches between the model and the actual shape of the index finger.

These offsets indicated that the image position of the finger's mid-line could be determined to sub-pixel accuracy (standard deviation typically $\sigma = 0.3$ pixels), and the orientation to an accuracy of 0.6°. From this uncertainty measure $\pm 2\sigma$ bounds were calculated for the lines l_i and l'_i; and, by projecting these onto the ground plane, the uncertainty in the indicated point could be estimated.

Figure 4.6 shows the results for three different configurations of the cameras, with a 95% confidence ellipse drawn around the indicated point. The constraint line uncertainties were much the same in each trial, but the uncertainty on the indicated point varied according to the separation between the stereo views: when the cameras were close together, the constraint lines were nearly parallel and tracker uncertainty became very significant (figure 4.6a); as the baseline was increased and the stereo views become more distinct, the constraint lines met at a greater angle and accuracy was improved (figure 4.6c).

II. Reference point uncertainty

In the above experiments, reference points were identified in the images by hand, and we assume an uncertainty of 1 pixel standard deviation (in an application, techniques exist to allow points or lines to be localised to higher accuracy, and errors may be reduced by observing more than 4 corresponding points – this is therefore a rather conservative estimate of accuracy).

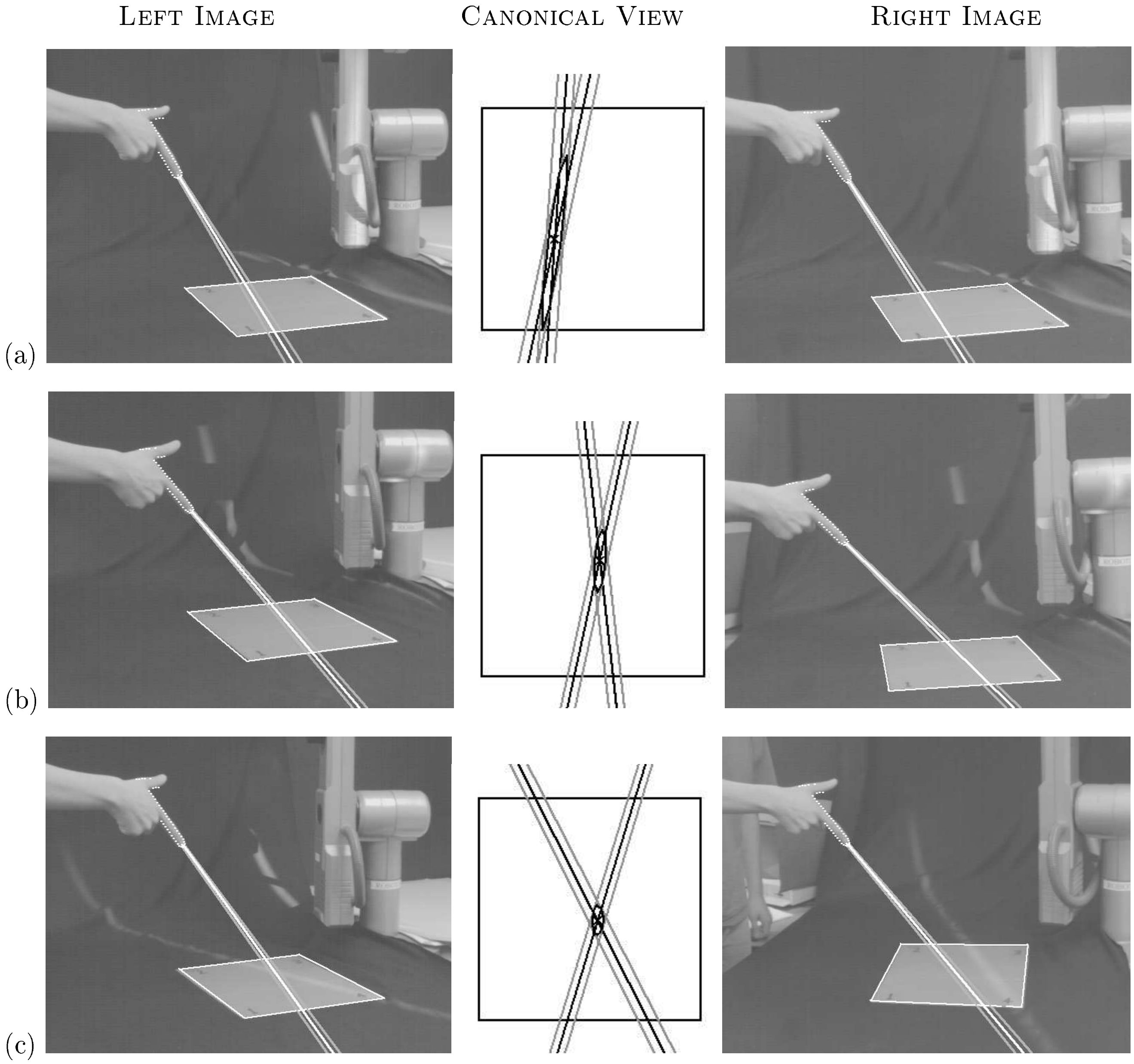


Figure 4.6: Indicated point uncertainty for 3 different camera configurations: 2σ bounds for the pointing lines, their projections into working plane coordinates, and error ellipses for the indicated point, when the angle between stereo views is (a) 7° (b) 16° (c) 34°. The uncertainty is greatest when the camera angle is small and the constraint lines nearly parallel.

We used Monte Carlo simulations (based around real-world configurations of cameras, hand and table) to assess the impact of this uncertainty on the coordinates of the indicated point. The results (table 4.1) show that this source of error is less significant than the tracker uncertainty, and confirm that the system is not especially sensitive to errors in the reference point image coordinates. Again, the errors were most significant when the camera separation angle was small.

Angle between the cameras	(i) Working plane coordinate error (with tracker noise)	(ii) Working plane coordinate error (with ref. point noise)	(iii) Working plane coordinate error (with both)
7°	.119	.040	.124
16°	.044	.019	.047
34°	.020	.008	.022

Table 4.1: Simulated RMS error in working plane coordinates, due to (i) tracker uncertainty derived from 'residual offsets' as detailed above; (ii) reference point image noise, $\sigma = 1$ pixel in each image; (iii) both. A value of 1.0 would correspond to a positioning uncertainty of about 40cm (the width of the reference point rectangle).

III. Experimental accuracy

Ground truth about the position and orientation of a human finger is, of course, very difficult to measure without intrusive equipment that could interfere with the stereo vision system. We therefore tested the accuracy of the pointing system using an artificial pointing device (figure 4.7). The test pointer was a white cylinder, about 15cm long, bounded by black end stops and wrapped around a rod which could be positioned by the robot arm to an accuracy of about 3mm. Whilst not identical to a human hand, it had approximately the same dimensions and was tracked in a similar manner.

A number of trials were carried out with the vision system tracking the rod as it was aligned with points on a grid on the target surface. The RMS error was 2.3% of the working plane coordinates, or 9mm in a 40cm workspace. The maximum reported error was 3.7% (15mm).

4.5 Robot control application

The proposed application for this stereo pointing system is to control a robot manipulator as it grasps and places small objects on a flat table-top. This time the four reference points were defined automatically by the robot itself in a plane a few centimetres above the table.

4.5.1 Setup

The reference points were defined by observing the robot gripper itself as it visited 4 known points in a plane. The robot began by opening and closing its gripper, and using the resulting image motion to initialize a pair of affine active contours (similar to those used to track the pointing hand, described in Appendix A). It was then tracked as it made deliberate motions across the plane. This not only defined the working coordinate system but related it to the robot's own world coordinate system. Finger-trackers were then initialised as before.

4.5.2 Performance

The robot was now instructed to move repeatedly to where the hand was pointing, in the horizontal working plane raised 50mm above the table-top. By watching the robot's motion, the operator was provided with a source of direct feedback of the system's output, allowing him or her to correct for systematic errors between subjective and observed pointing direction, and align the gripper over objects in the robot's workspace.

When the distance between hand and workspace is large, the system is sensitive to small changes in index finger orientation (as one would expect). To reduce this sensitivity, the operator maintains a steep angle to the horizontal, and points from a distance of less than 50cm from the plane, whilst still keeping his or her hand clear of the robot. One can then comfortably position the gripper with sufficient accuracy to pick up small objects (figure 4.8).

4.5.3 Using the interface to grasp objects

In experiments, it was found that two simple classes of object could be grasped reliably without any further planning:

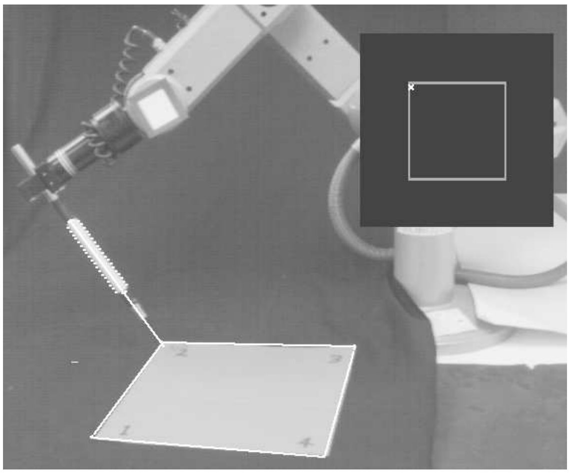
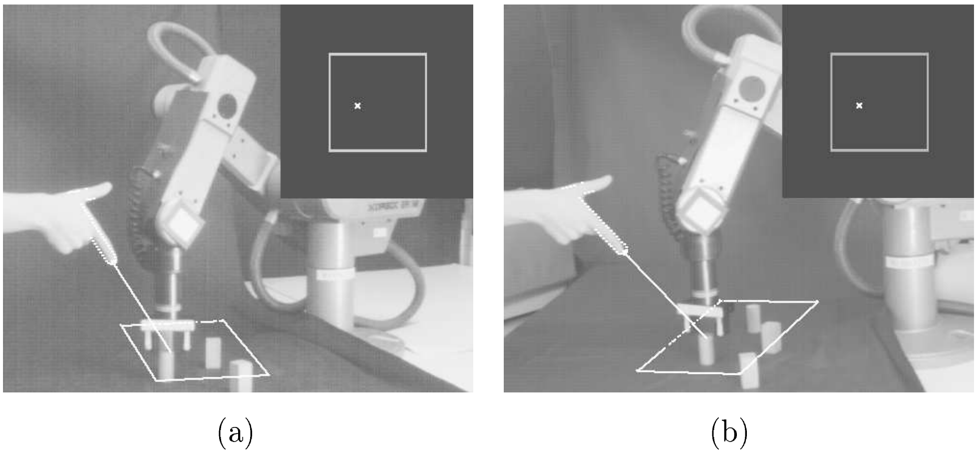


Figure 4.7: Mechanical pointing device used to test the accuracy of the system. We aligned the rod with known points on the workspace, and recorded its coordinates as recovered by the vision system.



(a)

(b)

Figure 4.8: Gestural control of robot position for grasping, seen in stereo.

Small cylinders. For small upright objects on the plane, the grasping operation is trivial and can take place without any further image processing (the grasp configuration being a function only of the target's position in two dimensions). Using visual feedback or under the direct control of the user's gestures (figure 4.8), the robot could be aligned with the target and the grasp executed.

Flat targets. The outer contours of the target's image were localised automatically using a stereo pair of 'expanding' B-spline snakes [21] initialised at the indicated point, enabling both the *position* and *orientation* of the graspable surface to be estimated using affine stereo. They could then be grasped using visual feedback as described in chapter 3.

For successful grasping of more complex objects, it is necessary to incorporate some sort of automatic *grasp planning* based on a stereo *reconstruction* of the target object, to analyse the shapes of its visible surfaces. This is dealt with in chapters 5 and 6.

4.6 Discussion

This algorithm for resolving the direction of pointing proves to be usable and stable in the presence of normal image noise. It does not require camera calibration because all calculations take place in the image and ground planes. By tracking 4 points on the plane it can be made invariant to camera motions.

The system presented here can be extended to situations in which more than one surface can be pointed at; however, this requires an image-based model of those surfaces and is harder to implement with moving cameras (because a large number of world features would have to be tracked to maintain invariance).

The main challenge to this system is the real time tracking of a pointing hand reliably in stereo. At present, this is only possible in an environment where there is a strong contrast between the hand and the background. Tracking is currently implemented on a standard workstation, and could be made more responsive using specialised hardware. Colour vision might also be useful for segmenting the hand in a cluttered scene.

Although subjective pointing direction depends on eye as well as hand position, it is not necessary to model this phenomenon. Instead, by providing the operator with feedback about the *objective* pointing direction (e.g. having a robot follow the pointing hand in real time), objects and locations may be specified for pick-and-place operations. However, in all but the simplest of robotic applications, this will need to be combined with visual reconstruction of objects so that they can be appropriately grasped.