An Uncalibrated Stereo Visual Servo System

Alistair Conkie and Prabhas Chongstitvatana

Dept. of Artificial Intelligence, University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL, Scotland

We describe a robotic system composed of a manipulator and two cameras. We use the vision system to guide the robot hand to a visible target. The camera positions are known only approximately. Our system does not use the details of the kinematics of the manipulator. There is no common frame of reference linking vision system, workspace and robot hand. The stereo vision system gives information in terms of picture coordinates. This information is used to control a three degree of freedom robot manipulator in a straightforward and robust fashion, in terms of lines and points visible in the images. We describe the implementation with which we tested this idea.

In a traditional robot vision system, to work out how to move a robot to manipulate an object, the location of that object must be found in terms of a common reference frame between the vision system, the workspace and the robot hand, (the so called "world coordinate frame"). Stereo vision can be used to derive depth information. In doing this, the location of the two cameras must be known to do triangulation calculations. For example, edge-based binocular stereo can provide accurate depth data, provided that the camera geometry is known [1]. The method of calibration to get an accurate camera geometry is well established [2,3].

Also, a traditional robot system works in its reference frame (the so called "base coordinate frame"). A model of the kinematics of the manipulator is used to translate between the Cartesian frame and the measured joint space [8,7]. The measurements come from the information from the encoders in the joints of the manipulator. While the ultimate accuracy is limited to the resolution of the joint encoders in a particular robot configuration, in practice the accuracy of going to a location depends mainly on the accuracy of this model. It is true that many experimental systems do use external measurement to determine the position of the robot hand, for example the system by Inoue [4], but rarely is this measurement used to close the control loop.

Our system is very different. We use external measurement to close a control loop around the end-effector position relative to the target position. We do not calculate the absolute position or orientation of the target object so the geometry of the camera setup need not be known. This has the great advantage of not relying on an accurate geometric model of the world, including the cameras and robot, and allows the robot to be controlled in a straightforward and robust manner, in terms of lines and points visible in the image planes of the cameras.

STEREO VISUAL SERVOING

Sanderson & Weiss [5] proposed the visual servoing of a manipulator system based on position or image data, and later an adaptive control scheme based on image feature reference was developed, where the measurement of image features was used in the control loop [6]. Shepherd [13] has shown a simple tactic for making use of active cooperation of camera and hand for an insertion task, with one camera at the same height as the assembly, and an indeterminate distance from the object. The camera is moved through 90 degrees to provide two perpendicular views. The robot is first moved at right angles to the camera in each case to calibrate the setup.

An uncalibrated stereo vision system has been used to navigate a mobile robot [10] and more recently to determine the size and shape of a room [9]. The advantage of an uncalibrated vision system is that it does not rely on precise modelling or precisely calibrated equipment.

Our system measures positional displacements in the picture coordinates (2D pixel coordinates) of object relative to target. We use two cameras to derive the information to control 3 degrees of freedom of a robot. Because we measure the relative displacements and map them directly into the joint motion commands, we do not need to know the camera positions or optical parameters. The measurement is done in terms of the difference between the hand and the target in the images. Very little prior knowledge is required. The accuracy of the system does not depend on knowing accurately the kinematics of the robot.

STEREO GEOMETRY

The cameras are set up so that the field of view of both cameras covers the area of interest. A point in the field of

BMVC 1990 doi:10.5244/C.4.49

view of both cameras, $\mathbf{P}$, can be represented as (x_L, y_L) and (x_R, y_R) in the left and right images respectively. Both y axes are assumed to point upwards, and the two optical axes are assumed not to be collinear. The elevation angles of the two cameras are arbitrary, but approximately equal. In an arbitrary but fixed global cartesian space, $\mathbf{P}$ can be represented uniquely as $\mathbf{x} = (x, y, z)$. Each point $\mathbf{P}$ can also be represented uniquely by a vector in the space of points $\mathbf{p} = (x_L, x_R, y_L)$. The transformation between cartesian coordinates and image coordinates, $\mathbf{J}_v$ is a function of the geometry of the cameras and lenses. Approximate values can be calculated by considering pinhole cameras in a specific configuration [1]. Consequently,

$$\mathbf{p} = \mathbf{J}_v \mathbf{x} \tag{1}$$

and

$$\mathbf{x} = \mathbf{J}_v^{-1} \mathbf{p} \tag{2}$$

We consider only 3 translation degrees of freedom for the robot. There is a transformation from the robot joint space represented by the set of points $\mathbf{m} = (m_1, m_2, m_3)$ and Cartesian space, as follows:

$$\mathbf{x} = \mathbf{J}_r \mathbf{m} \tag{3}$$

This matrix can be shown to be invertible for a 3 degree of freedom robot when it is restricted to cover a connected region with no singularities, and a function of $\mathbf{x}$ or $\mathbf{m}$

So

$$\mathbf{p} = \mathbf{J}_v \mathbf{J}_r \mathbf{m} = \mathbf{J}\mathbf{m} \tag{4}$$

and since $\mathbf{J}$ is invertible if we restrict the domain of interest to a connected region of space where both $\mathbf{J}_v$ and $\mathbf{J}_r$ are invertible (distant from any singularities of the robot).

$$\mathbf{m} = \mathbf{J}^{-1} \mathbf{p} \tag{5}$$

That is, given the uniqueness of the mapping we can attempt to control the robot using image data directly for the feedback and reference signals by observing the vector from the robot hand to the target in the images.

For control purposes, we found that $\mathbf{J}$ could be approximated sufficiently by measuring the result of a sequence of independent small motions of each of the motors m_1, m_2, m_3 and noting the resultant changes in $\mathbf{p}$.

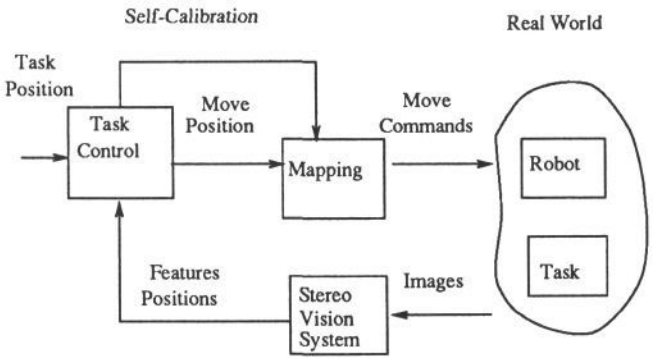

Figure 1: *Control Scheme*

$$\mathbf{J} \doteq \begin{bmatrix} \frac{\Delta x_L}{\Delta m_1} & \frac{\Delta x_L}{\Delta m_2} & \frac{\Delta x_L}{\Delta m_3} \\ \frac{\Delta x_R}{\Delta m_1} & \frac{\Delta x_R}{\Delta m_2} & \frac{\Delta x_R}{\Delta m_3} \\ \frac{\Delta y_L}{\Delta m_1} & \frac{\Delta y_L}{\Delta m_2} & \frac{\Delta y_L}{\Delta m_3} \end{bmatrix} \tag{6}$$

This worked sufficiently well to allow us to use only an initial sequence of three movements to serve in calculating a value of $\mathbf{J}$ used in subsequent control. This method was used in our first experiment. Clearly on a global scale this approximation is unjustified, but it worked over the control area of interest. With the above equation (6) and with the target point and the robot end effector in the field of view, the measured parameters are the vectors from the robot hand to the target.

It is possible to update the estimate of $\mathbf{J}$ as a form of adaptive control [6]. Our second experiments used repeated measurement of the observed motion to update $\mathbf{J}$ (Fig 1). The "reference" is described in terms of the features in the images. The "task control" handles the identification of features and the motion strategies. The "self-calibration" updates the mapping $\mathbf{J}$.

THE EXPERIMENTS

We developed two experiments, the first to control the robot hand so as to align its fingertips with a visible target, the second to pick up and stack three cubes.

We used an RTX robot arm from Universal Machine Intelligence Ltd. The first camera was a Sony DXC-101P with a 16 mm lens; the second was a Panasonic F10 with a 10.5mm-84mm zoom lens (both have CCD sensors). We adjusted the focal length so that the scale of the images from the two cameras did not differ by more than about a third, which could be done easily by hand.

The First Experiment

The targets for the stereo vision system to track were simple, two white triangles. One was fixed to the robot

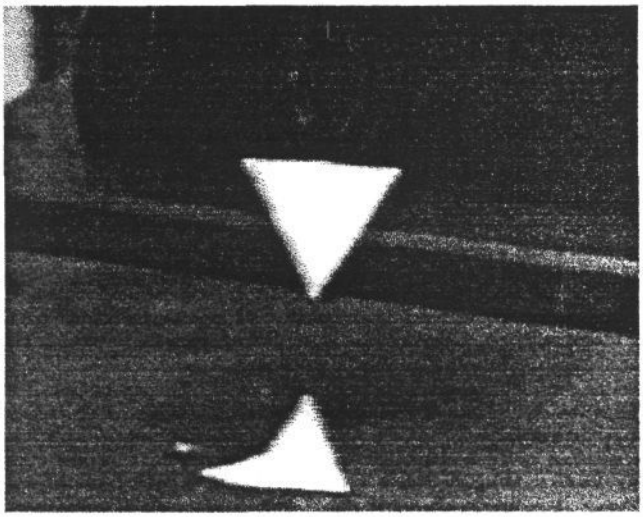

Figure 2: *Aligning two triangles*

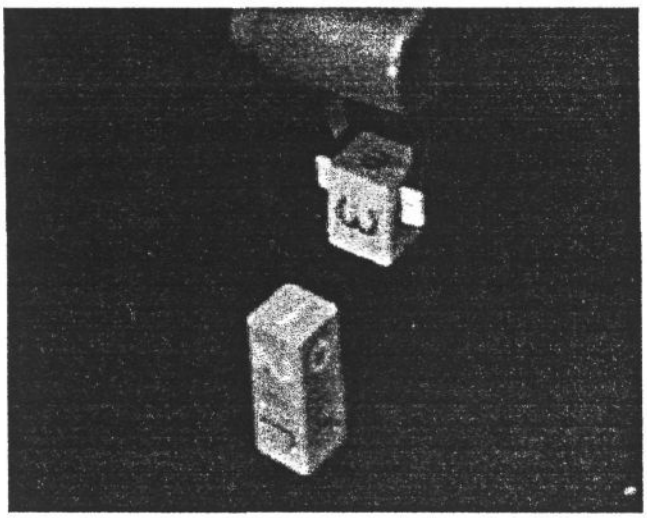

Figure 3: *Stacking three blocks*

hand the tip pointing downward, the other lay on the surface table in the working region with the tip pointing upward (Fig 2).

The tips of two triangles were detected in both cameras. The goal was to align them in 3D space. The matrix of equation (6) was constructed as described above and then the motion strategy adopted in this case was to move the hand so as to align the two triangles vertically and then move downward half the calculated goal distance. The motion was decoupled into two separate error reducing processes. The goal position was achieved by repeating these processes several times.

The variation of the target position is in the area 10 x 10 cm. and the height of the hand is 20 cm. We found that the target was reached within 1-2 mm in every run, with approximately eight iterations.

The Second Experiment

The aim was to pick up and stack three blocks (Fig 3). This experiment involved an additional degree of freedom of motion of the robot end effector, the hand being able to rotate about the vertical axis in order to pick up a cube. The visual features that we used for control were the fingertips of the manipulator, which are painted white and the centroid of the blocks. The hand position was taken as the point midway between the two fingertips with the hand empty (not holding the block) and the centroid of the block when there was a block in the hand.

Three blocks were laid in a line without touching each other. Their individual orientations were arbitrary and they were allowed to be anywhere inside the region visible to both cameras.

The strategy for picking up a block is similar to that in the first experiment except that in addition the rotation of the hand about the vertical axis is adjusted to enable grasping. This is done using the edge of the block and at present uses the view from only one camera. The last move is downward to pick up the block without using visual feedback. We do not use visual feedback here because the present vision processing routines cannot separate the fingertips and the block when occlusion occurs.

The vertical axis component of the mapping was continuously updated because accuracy is needed for the estimation of the final approach motion to the grasping position.

For stacking the blocks, the bottom edge of the block in the hand and the top edge of the target block are used to servo the vertical motion. The visible vertical edges of both blocks are used to guide alignment.

VISION PROCESSING

We need only 2D information from two cameras. To measure the error vector, both the robot hand and the target must be clearly seen. The error vector is the vector from a point on the robot hand to a point on the target. We get this information by tracking the selected features of the robot hand and the object. Presently, we use the boundaries of object silhouettes against the background, and get the corners by the technique of polygon approximation of the boundary [11].

CONCLUSIONS

In machine vision, the process usually starts with early vision processing that produces a 2 1/2 D sketch. Then follows higher level vision processing to do image segmentation and so on until a scene is described. Most current machine vision systems are model-based: they use a geometric model of the objects in understanding

the image in a 3D reference frame.

Our system, on the other hand, does early vision processing only to get to the level of point and line descriptions in 2D, after which we depart from the main stream. Our system does not build a model-based description of the scene nor recover the positions and orientations of the objects in 3D according to a fixed reference frame. Using only measurements in the 2D image planes of the cameras we derive local and approximate mappings between robot joint motion and changes in the image and use this in a control loop. Using this method we can control the robot in a straightforward and robust fashion using lines and points visible in the image planes.

Our experiments show the applicability of this scheme, but we have not yet completed a full mathematical analysis. This method is being incorporated into the SOMASS robotic assembly system [12] for part acquisition.

We would like to thank Chris Malcolm and Bob Fisher for encouraging our work,and Akis Petropoulakis who helped develop our ideas. A. Conkie works under SERC/ACME grant GR/E 68075.

References

[1] **Pollard, S.B., Pridmore, T.P., Mayhew, J.E.W., Frisby, J.P.,** "Geometrical Modelling from Multiple Stereo Views", *Int. J. of Robotics Research,* Vol. 8, No.4, Aug. 1989, pp 3-32.

[2] **Chang, Y.L., Liang, P.,** "On Recursive Calibration of Cameras for Robot Hand-eye Systems", *Proc. of 1989 IEEE Int. Conf. on Robotics and Automation,* Vol. 2, pp 838-843.

[3] **Tsai, R.Y.,** "A Versatile Camera Calibration Technique for High Accuracy 3D Machine Vision Metrology Using Off-the-shelf TV Cameras and Lenses", *IEEE Trans. on Robotics and Automation,* Vol. RA-3, Aug. 1987, pp 323-344.

[4] **Inoue, H., Inaba, M.,** "Monitoring 3D Pose of Robot Hand by Real Time Vision", *2nd Int. Conf. of Advanced Robotics,* Tokyo, 1985.

[5] **Sanderson, A.C., Weiss, L.E.,** "Adaptive Visual Servo Control of Robots", *Robot Vision,* Pugh, A., (ed.), Springer-Verlag 1983, pp 107-116.

[6] **Weiss, L.E., Sanderson, A.C., Neuman, C.P.,** "Dynamic Sensor-based Control of Robots with Visual Feedback", *IEEE J. of Robotics and Automation,* Vol. RA-3, No.5, Oct. 1987, pp 404-417.

[7] **Hollerbach, J.M.,** "A Recursive Lagrangian Formulation of Manipulator Dynamics and a Comparative Study of Dynamics Formulation Complexity", *IEEE Trans. Syst., Man, Cybern.,* Vol. SMC-10, Nov. 1980, pp 730-736.

[8] **Paul, R.P.,** *Robot Manipulators: Mathematics, Programming, and Control,* MIT Press, Cambridge, MA., 1981.

[9] **Sarachik, K.B.,** "Characterising an Indoor Environment with a Mobile Robot and Uncalibrated Stereo", *Proc. of 1989 IEEE Int. Conf. on Robotics and Automation,* pp 984-989.

[10] **Brooks, R.A., Flynn, A.M., Marill, T.,** "Self Calibration of Motion and Stereo Vision for Mobile Robot Navigation", Artificial Intelligence Lab., MIT, Cambridge, MA,*AI-Memo 984,* Aug. 1987.

[11] **Malcolm, C.A.,** "The Outline Corner Filter", *Proc. of the 3rd Int. Conf. on Robot Vision and Sensory Controls,* Nov. 1983, pp 61-68.

[12] **Malcolm, C.A., Smithers, T.,** "Symbol Grounding via a Hybrid Architecture in an Autonomous Assembly System", *Robotics and Autonomous Systems,* Vol. 6, 123-144, 1990.

[13] **Shepherd, B.,** "Performing hand-eye operations in an unstructured environment using active cooperation between a robot and a mobile camera", private communication.