Computer Vision in Interactive Robotics

Roberto Cipolla, Nicholas Hollinghurst, Andrew Gee and Robert Dowland

University of Cambridge
Department of Engineering
Trumpington Street
Cambridge CB2 1PZ
England

Abstract

Computer vision provides many opportunities for novel man–machine interfaces. Pointing and face gestures can be used as a simple, passive means of interfacing with computers and robots. In this paper we describe two novel algorithms to track the position and orientation of the user's hand or face in video images. This information is used to determine where the hand or face is pointing. This can be used in interactive robotics to allow a user with manipulation disabilities or working in hazardous environments to guide a robot manipulator to pick up simple object of interest.

Keywords: stereo vision, interactive robotics, human-computer interaction, visual tracking.

1 Introduction

There has been a lot of interest lately in the use of hand and face gestures as a means for interfacing with computers and robots: 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 [2, 12, 23]. It is also important as a means of helping individuals with manipulation disabilities.

Most current commercial implementations rely on sensors that are physically attached to the hand, such as the 'DataGlove' [11]. More recently, systems have been proposed using *computer vision* to observe the hand. Some require special gloves with attachments or markings to facilitate the localization and tracking of hand parts [5, 24], but others operate without intrusive hardware. This is attractive because it is convenient for the user and potentially cheaper to implement.

In this paper we describe two novel algorithms to track the position and orientation of the user's hand or face in video images. This information is used to determine where the hand or face is pointing. This can be used in *interactive robotics* to allow a user with manipulation disabilities or working in hazardous environments to guide a robot manipulator to pick up simple object of interest.

We describe preliminary implementations of these algorithms. We also describe an interactive robotic inspection system being designed at Cambridge to provide a means for a physically disabled person to perform the job of visual inspection of hybrid micro-circuits.

2 Pointing interfaces

Here we present an experimental system that uses *pointing* gestures to guide a robot to pick and place objects on a table top [6, 7, 8]. A pair of cameras view the table and pointing hand in stereo. Active contours are used to track the hand in real time. By simple geometry we can calculate where the hand is pointing in the robot's workspace, without camera calibration, to an accuracy of about 10mm.

2.1 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 (see figure 1). A second view is needed to fix its position in two dimensions [22].

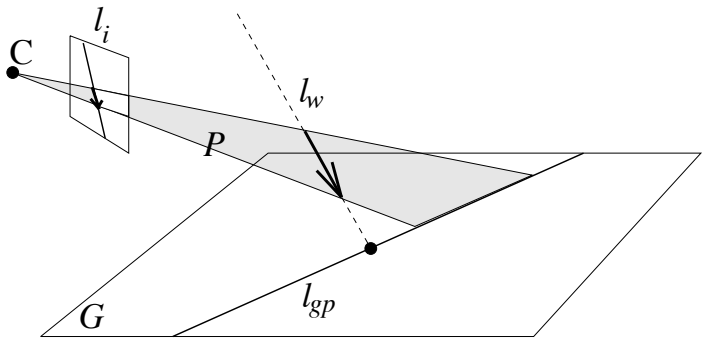


Figure 1: Relation between lines in the world, image and ground planes

Consider a pair of pinhole cameras viewing a planar surface (such as the robot's work surface). The viewing transformations can be modelled by plane projectivities, and there exists a projective transformation that maps one image to the other. This transformation can be computed by observing a minimum of four points on the plane. We exploit this to transform the constraint lines into a common 'canonical' view of the plane [21], and hence find their intersection (figure 2).

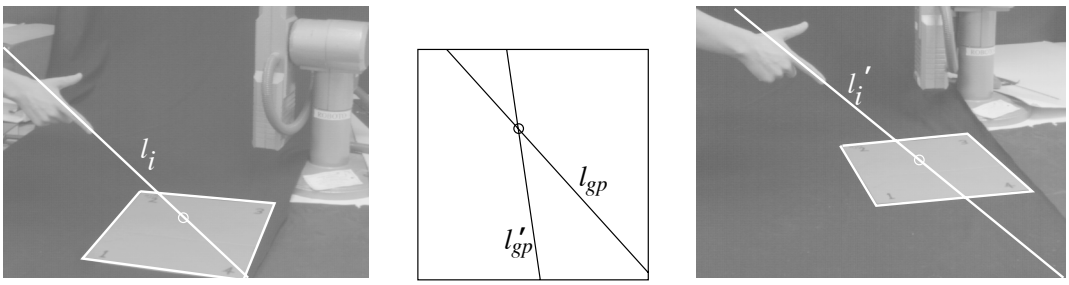


Figure 2: Combining constraints from two views of a pointing hand

The piercing point can then be projected back into the two images; if the four reference points are known its world coordinates can be calculated, otherwise visual feedback can be used to guide a robot manipulator to this point [16].

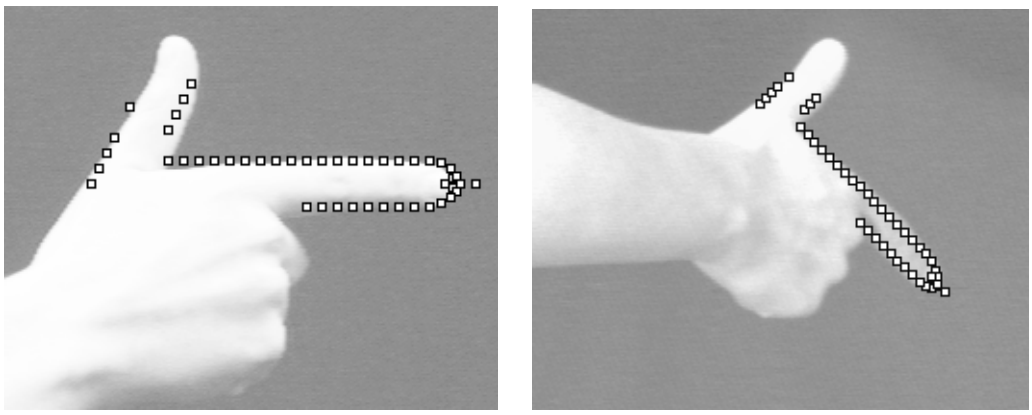
2.2 Tracking mechanism

We use a type of active contour tracker [18, 4] to track the image of a hand in the familiar 'pointing' gesture. The tracker is based on a template, representing the shape of the occluding contours of an index finger and thumb (figure 3).

The tracker's motion is restricted to 2D affine transformations in the image plane, which ensures that it keeps its shape whilst tracking the hand in a variety of poses [3] (this approach is best suited to tracking small planar objects, but also works well with fingers, which are cylindrical).

2.3 Implementation

The system is implemented on a Sun SPARCstation 10 with a Data Cell S2200 frame grabber. Images are provided by two monochrome CCD cameras, which view the operator's hand and the working area from a distance of about 1.6 metres. The angle between the cameras is about 30°. A Scorbot ER-7 robot arm is also controlled by the Sun (figure 4).



(a) (b)

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

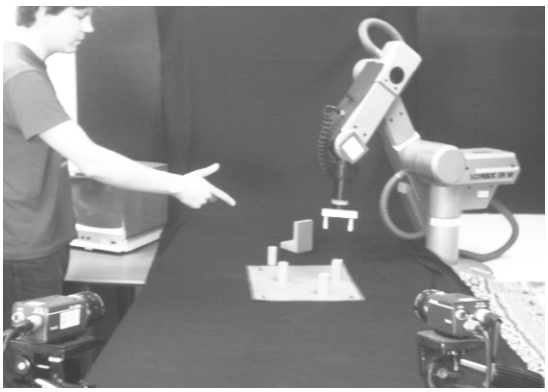
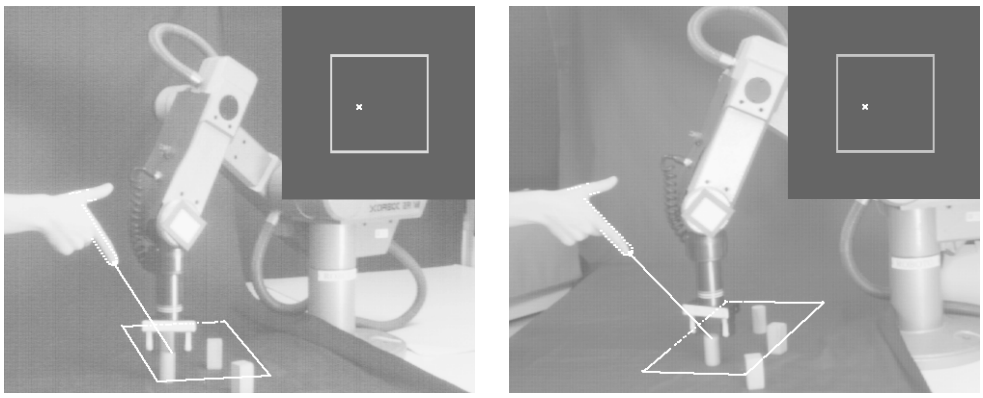


Figure 4: Arrangement of the stereo cameras, table and robot

The four reference points were defined by observing the robot as it moved to four known points 50mm above the table-top. Hand-trackers were initialised by the operator.

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



(a) (b)

Figure 5: Gestural control of robot position for grasping, seen in stereo.

The robot is instructed to move repeatedly to where the hand is pointing. One can thus position the gripper with sufficient accuracy (about 10mm) to pick up small objects in the workspace.

3 Faces and HCI

The ability to detect and track a person's face is potentially very powerful for human-computer interaction. For example, a person's gaze can be used to indicate something, in much the same manner as pointing (section 2). Gaze aside, the head orientation can be used in virtual holography applications [1], or to guide a remote, synthesised "clone" face for low bandwidth video conferencing [19, 20]. In addition, a head tracker could provide a very useful computer interface for physically handicapped people, some of whom can only communicate using head gestures. We have developed techniques for inferring the gaze direction of arbitrary faces in static images [13], and also tracking familiar faces as they move at high speed in image sequences [14, 15].

3.1 Face tracking

The face tracking algorithm has been specially developed to cope with the poorly defined features and unpredictable motion associated with faces. In each frame, the tracker tries to detect six features on the subject's face — see figure 6. Dark-pixel finders (which simply return the location of the darkest pixel within their local search windows) are used to locate the eyes, nostrils and the shadow below the lower lip. A coarse correlation-based detector is used to detect the centre of the eyebrows using a dark-light-dark correlation template.

In theory, the pose of the face can be calculated up to a two-fold ambiguity using just three of these features, a geometric model of the face and the "alignment" algorithm [17]. In practice, the feature detectors are unreliable, and we have to draw on more than three measurements. Typical non-stationary measurement noise is illustrated in figure 7, which shows the tracking error of one of the nostril detectors. The detector performs well when the nostril is visible, finding the nostril position to within, typically, three pixels. When the head is turned so that the nostril is less prominent, between cycles 500 and 1100, the feature detector starts to show a bias as it finds instead the shadow around the outline of the nose. There are also many isolated outliers (gross errors) in the x and y traces. This sort of behaviour is typical of simple visual feature detectors,

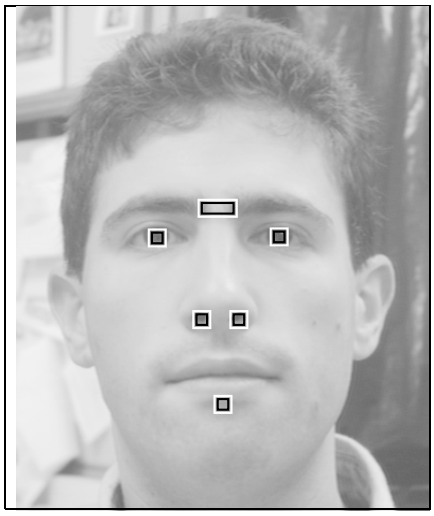


Figure 6: Features and search windows for face tracking.

and demonstrates that it is unwise to assume zero-mean Gaussian noise on the detector outputs. A standard least-squares fit to the six measurements is therefore inappropriate.

Instead we use RANSAC [10], a robust regression scheme, to accurately estimate the face's pose. All combinations of three measurements out of the six available are considered in each frame. For each set of three measurements, the two solutions proposed by the alignment algorithm are assessed on the basis of consensus: if the pose is also consistent with some of the other measurements, this constitutes an extended consensus which lends support to that pose hypothesis.

The notion of consensus is illustrated in figure 8, which shows a frame from a hypothetical image sequence. A set of facial features is detected in the image using imperfect feature detectors whose outputs are displayed as crosses. All measurements are good except at the left corner of the mouth (measurement 4). If measurements 1, 2 and 3 are used to estimate the pose, the model back-projects into the image as shown on the left (back-projected model points are displayed as circles). Four of the measurements are consistent with this pose. If measurements 1, 2 and 4 are used to estimate the pose, then the consensus set is only three, as shown on the right. Thus consensus can be used to select the best pose from the candidate solutions, reliably rejecting outlying measurements.

The candidate pose solutions are also assessed on the basis of motion smoothness: a pose which implies smooth motion (based on previous frames) is more likely to be correct than one which implies a large jump. The consensus and motion scores are combined using Bayesian inference, and the pose solution with the highest combined score is selected. The entire process is repeated for each frame in the image sequence. Since the algorithm is tolerant of noisy, computationally cheap feature detectors, real-time operation is comfortably achieved on standard hardware.

Figure 9 shows some sample frames from a test sequence. A range of facial expressions is represented, and the face is partially occluded about 20 seconds into the sequence. The motion is fairly unconstrained, with angular velocities up to 1.4 rad/s, and angular accelerations up to 14 rad/s^2. By properly combining consensus scores and inter-frame motion constraints, tracking is maintained through high velocity manoeuvres and partial occlusion of the target. The tracker's orientation estimate is shown as a drawing pin in the top left hand corner of each frame.

4 Interactive robotics in workshop activities

Employment opportunities for individuals with manipulation disabilities can be severely restricted. Interactive robotics has been investigated by a number of research groups as a means of expanding

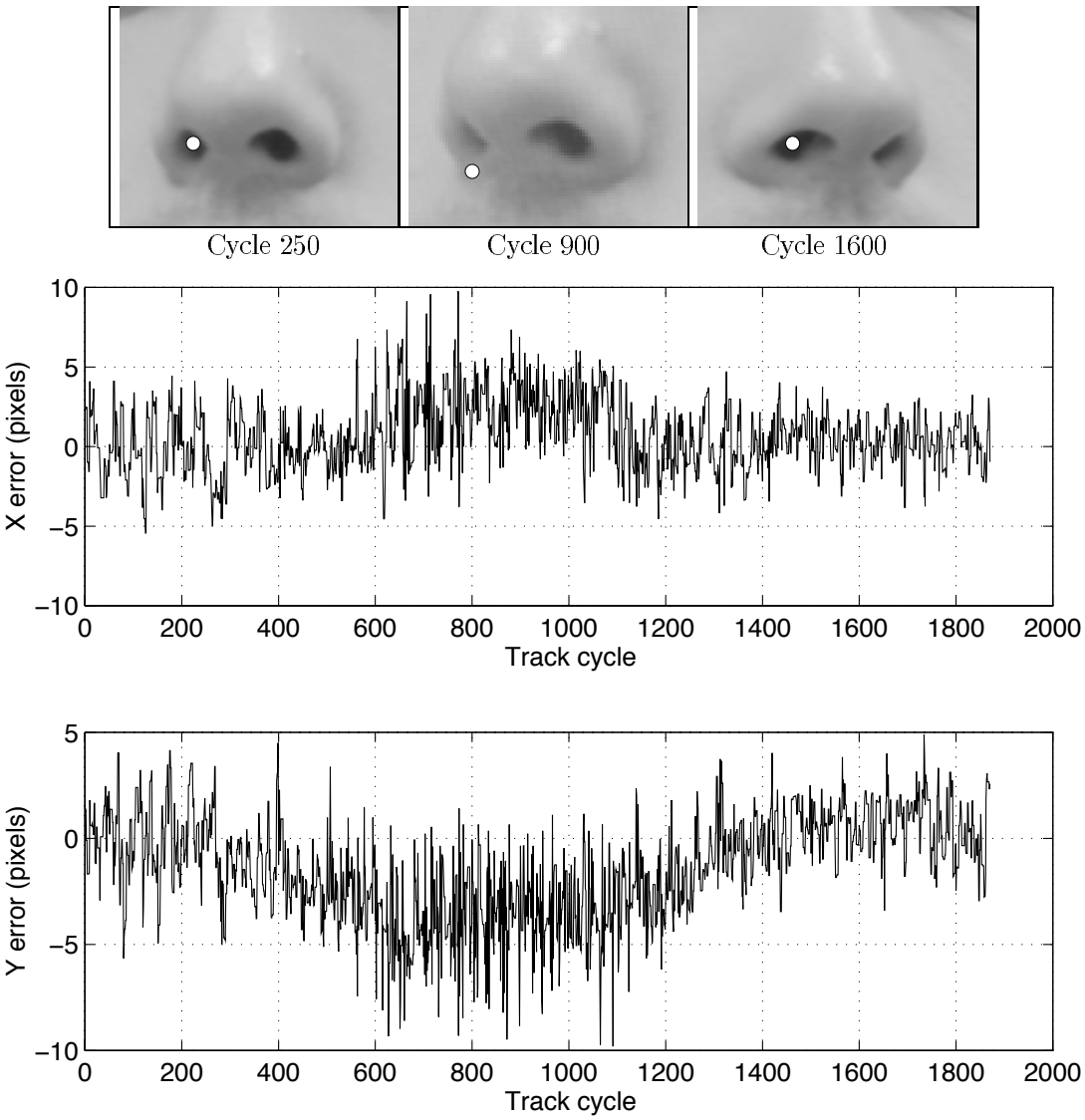


Figure 7: Tracking error of a typical feature detector.

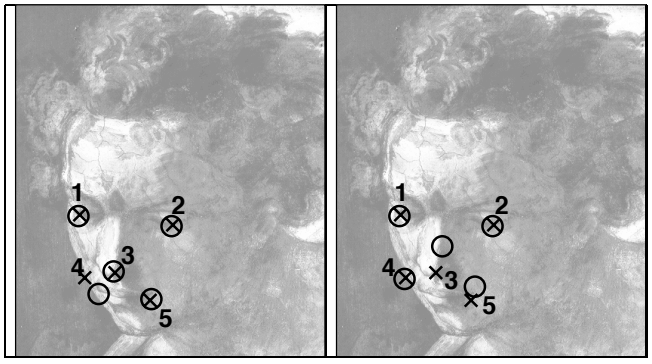


Figure 8: Using consensus to select the best pose.

Figure 9: Performance of the face tracker.

these opportunities. While the majority of work to date has focused on facilitating access for persons working in office environments, recent research within the Department of Engineering, University of Cambridge led to the design and implementation of an Interactive Robotic Inspection System (IRVIS) [9]. This was to provide the means for a physically disabled person to perform the job of visual inspection of hybrid microcircuits.

The mechanical system, illustrated in Figures 10 and 11, offers 5 degrees of freedom facilitating a variety of viewpoints. These include, rotation and translation in the horizontal plane, orientation of a gantry mounted camera for variation of pitch and a fifth degree of freedom used for focusing. A CCT camera with fixed focus zoom lens provides on screen magnifications in the range 30-200X.

Following extensive trials within a manufacturing environment the system is currently undergoing a second stage of development in which research is more closely focused on issues surrounding human computer interaction and the implementation of assistive technology from the area of machine vision. This new project entitled Interactive Robotics in Workshop Activities (IRWACS) builds on the experience of IRVIS and introduces some key modifications afforded by recent developments in multimedia technology and the enhanced DSP capabilities of the current range of high-end PC processors.

The Windows based computer interface, currently being developed, includes a live video display (Figure 12) which has been integrated with the system to allow for interaction with the subject of inspection in a more direct and intuitive manner. In the new system it is intended that the video display should become the focal point for interaction with mechanical hardware. This approach is intended to provide a form of interaction with an inspection rig which will be both easy to use and hardware independent.

Features currently under investigation include the implementation of software controlled autofocus to allow for focusing on individually selected areas of the image, realignment of circuits under view through image-centred mouse input and registration of the current viewpoint with respect to a circuit-specific frame of reference. The last of these is eventually intended to provide a basis for alignment with an associated circuit diagram for quick cross reference between circuit and component level documentation as well as comparison with an image database. Pointing and face gestures may also be used as part of the interface.

5 Conclusions

Pointing can be used successfully to specify positions on a two-dimensional workspace for a robot manipulator. The system described in section 2 is simple and intuitive for an operator to use, and requires practically no training. There is no need for camera calibration because no 3D calculations are involved. By tracking at least 4 points on the plane it could be made invariant to camera movement.

Although subjective pointing direction depends upon 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), the hand can be aligned with any desired object un the workspace. Points can then be indicated with sufficient accuracy to guide simple pick-and-place operations that might otherwise have been specified using

Figure 10: The IRVIS system for circuit board inspection in use

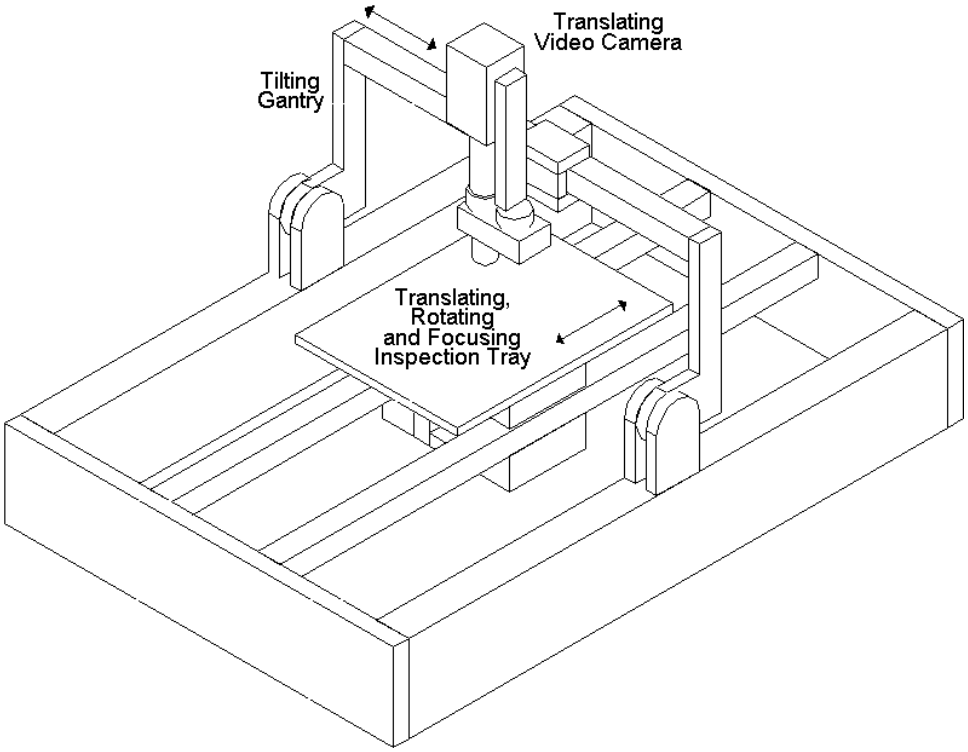


Figure 11: Schematic of the IRVIS system for circuit board inspection

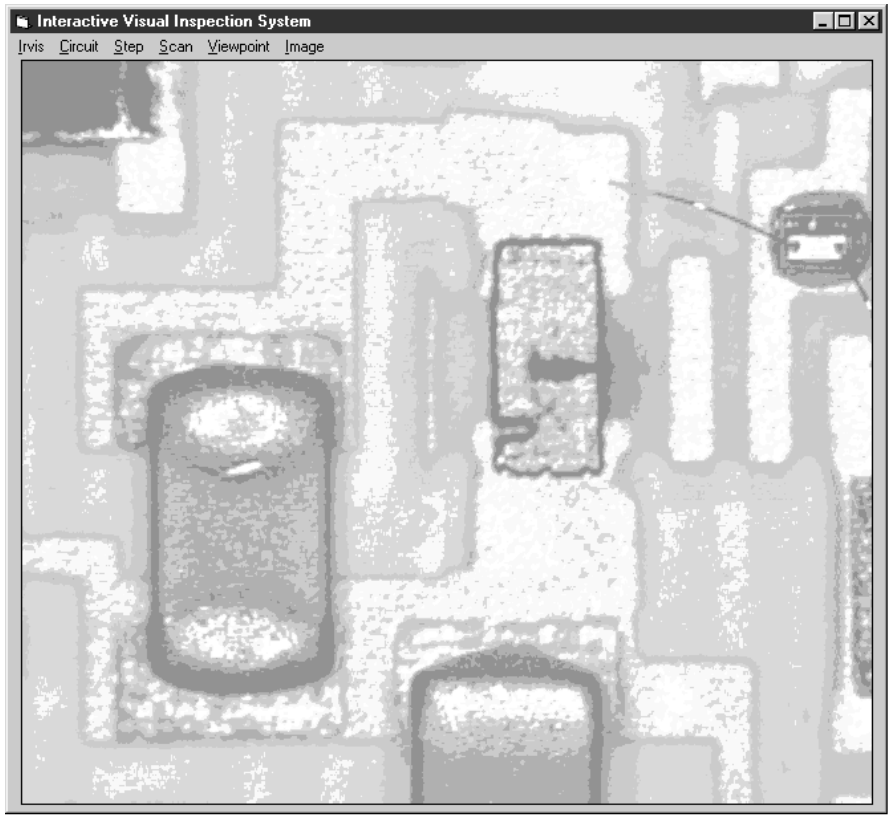


Figure 12: Interactive Visual Inspection System main application window showing integration of a live video field.

a teach pendant.

A person's gaze can be used to indicate something, in much the same manner as pointing. The algorithm presented in section 3 is capable of estimating the pose of the user's face in realtime. We hope to integrate these into the Interactive Robotics Inspection System to expand the opportunities of individuals with manipulation disabilities.

Acknowledgements

We gratefully acknowledge the financial support of EPSRC, and the donation of a robot manipulator by the Olivetti Research Laboratory, Cambridge. This research has benefited from discussions with Mr. Masaaki Fukumoto and Dr. Yasuhito Suenaga of the NTT Human Interface Laboratories and Dr John Dallaway.

References

[1] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland. Visually controlled graphics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):602–605, 1993.

[2] T. Baudel and M. Beaudouin-Lafan. Charade: remote control of objects with freehand gestures. *Communications of the ACM*, vol. 36 no. 7 pp 28–35, 1993.

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

[4] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *Int. J. Computer Vision*, vol. 9 no. 2 pp 83–112, 1992.

[5] R. Cipolla, Y. Okamoto and Y. Kuno. Robust Structure from Motion using Motion Parallax. *Proc. 4th Int. Conf. on Computer Vision*, pp 374–382. Berlin, 1993.

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

[7] R. Cipolla and N.J. Hollinghurst. Man-machine interface by pointing with uncalibrated stereo. Technical Report CUED/F-INFENG/TR199, University of Cambridge, November 1994.

[8] R. Cipolla and N.J. Hollinghurst. Human–Robot interface by pointing with uncalibrated stereo vision. *Image and Vision Computing* 1996.

[9] J. L. Dallaway, R. M. Mahoney, and R. D. Jackson. The application of rehabilitation robotics within manufacturing industry. In *Proceedings of the Fourth International Conference on Rehabilitation Robotics (ICORR '94)*, pp 145–149.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

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

[12] M. Fukumoto, K. Mase, and Y. Suenaga. Realtime detection of pointing actions for a glove-free interface. In *Proceedings of the IAPR Workshop on Machine Vision Applications*, pages 473–476, 1992.

[13] A. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647, December 1994.

[14] A. Gee and R. Cipolla. Fast visual tracking by temporal consensus. Technical Report CUED/F-INFENG/TR 207, Cambridge University Department of Engineering, February 1995. To appear in *Image and Vision Computing*.

[15] A. Gee and R. Cipolla. Adaptive visual tracking by temporal consensus. In *Proceedings of the 2nd Asian Conference on Computer Vision*, volume 1, pages 44–48, Singapore, December 1995.

[16] N. J. Hollinghurst and R. Cipolla. Uncalibrated stereo hand–eye coordination. *Image and Vision Computing*, vol. 12 no. 3 pp 187–192, 1994.

[17] D. P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

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

[19] R. Koch. Dynamic 3-D scene analysis through synthesis feedback control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):556–568, 1993.

[20] H. Li, P. Roivainen, and R. Forchheimer. 3-D motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, 1993.

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

[22] L. Quan and R. Mohr. Towards structure from motion for linear features through reference points. *Proc. IEEE Workshop on Visual Motion*, pp 249–254. Princeton, New Jersey, 1991.

[23] D. Sturman, D. Zelter and S. Pieper. Hands-on interaction with virtual environments. *UIST: Proc. ACM SIGGRAPH Symposium on User Interfaces*, pp 19–24. Williamsburg, Virginia, November 1989.

[24] B. Wirtz and C. Maggioni. ImageGlove: a novel way to control virtual environments. *Proc. Virtual Reality Systems*. New York City, April 1993.