

ADVANCES IN ROBOT VISION: MECHANISMS AND ALGORITHMS

Alexander Zelinsky and John B. Moore

Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, Australia
Email: Alex.Zelinsky@anu.edu.au

Abstract: This paper presents recent results from the Robotics Systems Laboratory at the Australian National University in the area of Robot Vision. Progress has been made in the design and construction of novel mechanisms for both active and panoramic vision, together with computer systems that facilitate real-time vision processing. The robot vision systems are being used to develop human-friendly robots, for guidance in mobile robot navigation and human-machine interaction.

1. ACTIVE VISION

Robot vision can be pursued from two contexts; active and passive systems. The biological world has numerous examples of both classes of visual sensors. Active systems have shown that by centering the fovea on the area of interest vision processing tasks such as visual attention, range estimation and tracking can be significantly simplified (Blake & Yuille 1992). A popular approach to building mechanisms for active vision systems has been to provide motor actuation to every degree of freedom. Direct drive systems are fast and have the advantage of zero backlash due to elimination of the gear-train. We are exploring a design philosophy based on two principles: (1) **parallel actuation** is preferable to serial actuation because it has the potential to reduce the driven inertia thus reducing motor sizes. Figure 1 shows a typical serial mechanism, where some motors must carry other motors. (2) **cable-drive transmission** technology is attractive because it provides a stiff, backlash-free reduction that is relatively low-tech and therefore economical. In this paper we describe a prototype pan-tilt mechanism that utilises both principles.

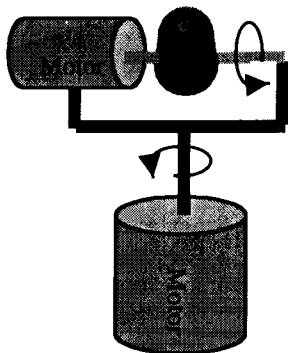


Fig. 1. Dependent Actuation Paths

The schematic of our prototype active vision system is shown in Figure 2. The independent actuation paths are realised using cable drive technology. The usefulness of cable drive mechanisms for robotics has been proven by (Townsend and Salisbury 1993) in the development of the MIT-WAM robot manipulator. We have used the same cable drive principles to develop our active vision system. A full description of the mechanism is given in (Truong et.al. 1999).

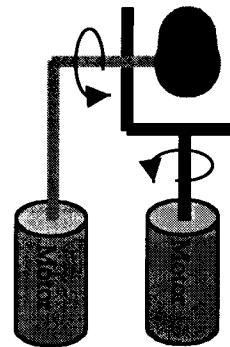


Fig. 2. Independent Actuation Paths

Figure 3 shows the fully assembled mechanism with mounted camera. The two horizontal spools are driven by separate motors through a cable-wrapped pinion. The ends of each cable are fixed to the driven spool. The Capstan effect ensures that the cable does not slip on the pinion.

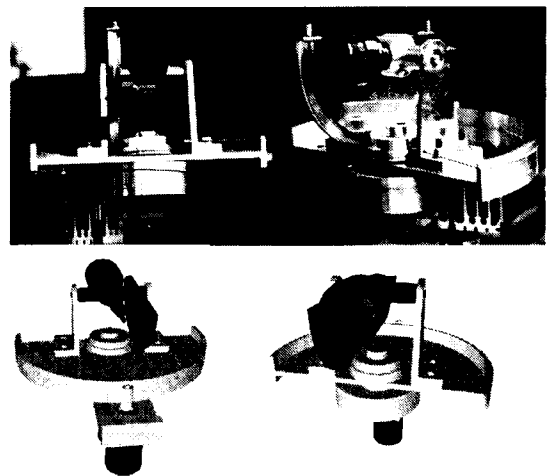


Fig. 3. The Mechanism

The performance of our mechanism summarised in Table 1 is competitive with the best of current high-speed designs.

Specification	Pan	Tilt
Max Velocity (deg/s)	530	520
Max Acceleration (deg/s ²)	10400	10000
Time for 60 saccade (s)		0.166
Time for 90 saccade (s)	0.233	
Resolution (deg)	<0.01	<0.01

Table 1: Summary of Performance

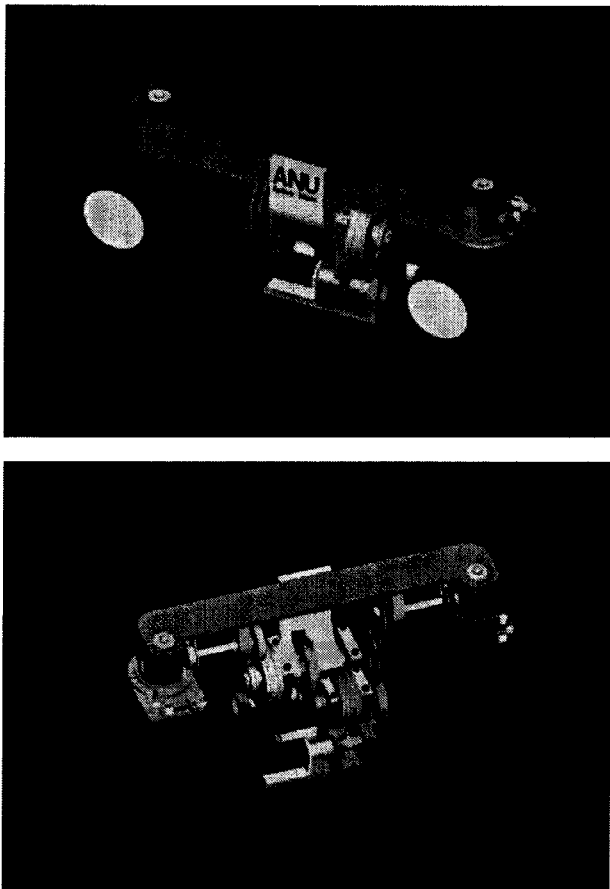


Fig 4. Active Stereo Vision Mechanism

We are presently developing a stereo active mechanism based on similar concepts. Figure 4 illustrates our new design.

2. PANORAMIC VISION

While active vision mechanisms seek to address fundamental problems, their biggest drawback is that they are usually mechanically complicated and require sophisticated software for the vision processing to overcome the problems of image blurring and optical flow

that are introduced from moving the sensor. An alternate approach is to use panoramic vision (Chahl and Srinivasan 1997). Panoramic systems are built using reflecting surfaces that capture wide aspects of the environment. Figure 5 shows the principle of operation of panoramic vision sensors together with a sensor built in our laboratory. If the correct profile is chosen for the shape of the reflecting surface very wide angle imaging is possible. Cones and hemispherical mirrors are popular shapes.

In our research we are investigating optimising the mirror shapes to address the problem of lack of pixel resolution at larger angles from the central axis of the reflector. We are use hyperbolic and parabolic shaped mirror surfaces.

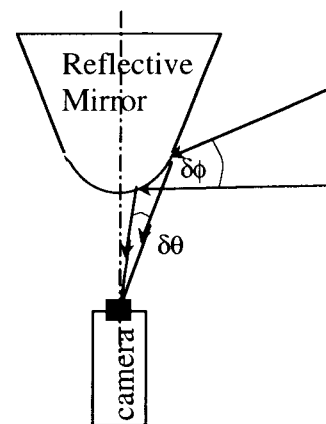
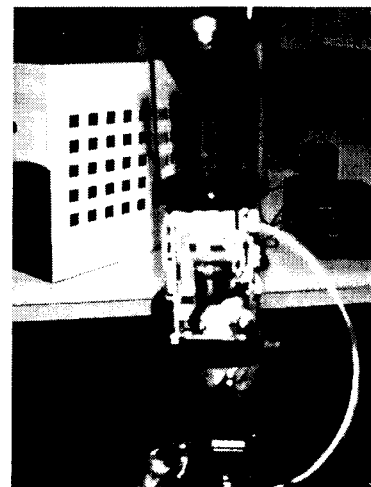


Fig 5. Panoramic Vision Sensor and Principle of Operation

Figure 6 shows a panoramic image and the associated unwarped image. A drawback of panoramic vision sensing is that the polar image is captured on a cartesian array of pixels in the CCD camera. Therefore the image quality varies over the unwarped image as there are less pixels at

the centred of the polar image compared with the regions on the outer edge of the image.

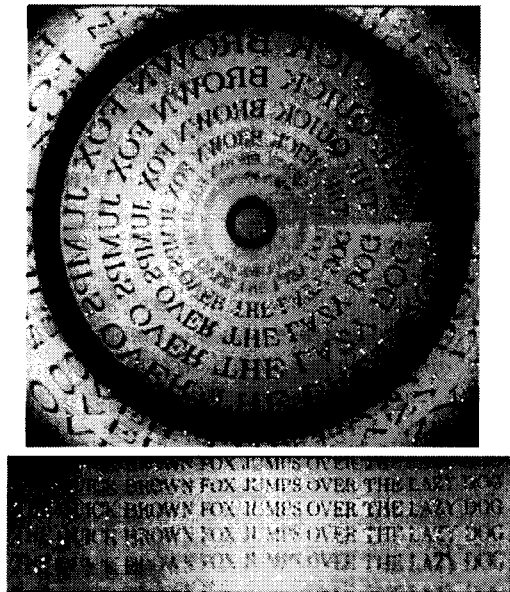


Figure 6. Panoramic Image and Unwarped Image

One of the goals of our research has been to improve the pixel resolution of panoramic vision sensors. Image resolution invariance is achieved by adjusting the mirror profile to image relatively less of the scene in the centre of the image and relatively more at the perimeter, that is, select a mirror profile to maintain a constant relationship between pixel density and the angle of elevation in the scene. Our research goal has been to develop a variable gain mirror. The mirror gain α , is the relationship between the change in elevation of rays incident on the mirror and the change in the angle of rays reflected into the camera is as follows:

$$\alpha = \frac{\delta\phi}{\delta\theta}$$

Where $\delta\phi$ is the change in vertical elevation in the scene and $\delta\theta$ is the change in angle of incidence. See Figure 5 for an illustration. In the work of (Chahl and Srinivasan 1997) α is of constant value. This has the effect of maximising the field of view. Figure 7 shows an example of the images produced with Chahl and Srinivasan mirror. These images suffer from not having the property of image resolution invariance.

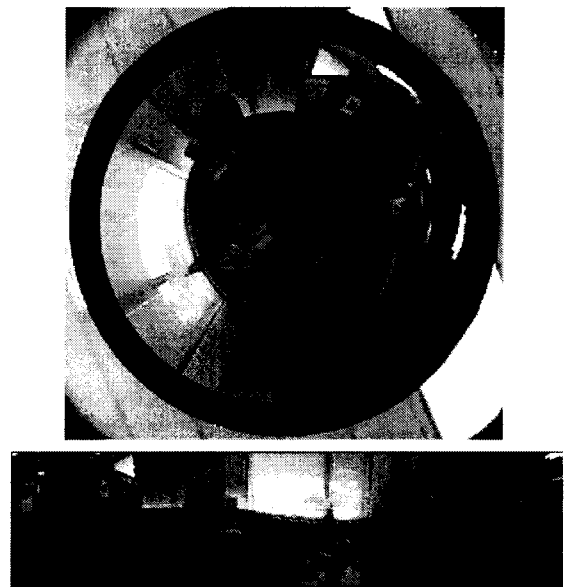


Figure 7. Panoramic Image with constant α

Image resolution invariance α becomes a function of image angle θ that is related to the radial coordinate in the image. Full details of the calculation of image resolution invariance α are given in (Conroy and Moore 1999). Using a variable α we are able to generate improved panoramic images. Figure 8 shows a panoramic image generated using a image resolution invariance mirror. Visually comparing Figures 7 and 8, we can see that Figure 8 has a greater resolution in the regions that are closest to the origin of the polar image.

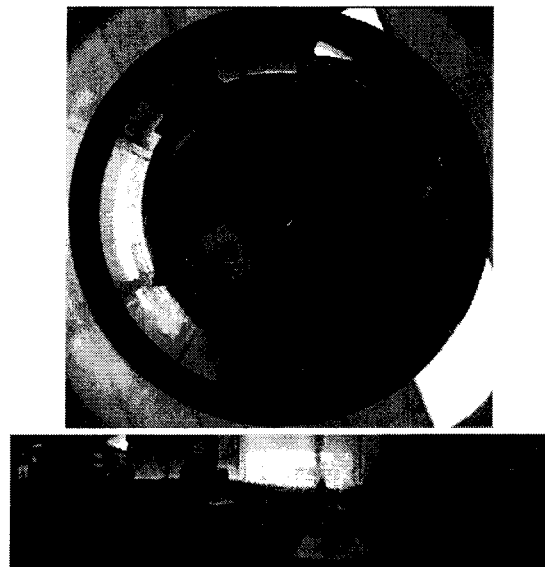


Figure 8. Panoramic Image with variable α

The image resolution invariant property of our mirror shapes has allowed us to design a new family of stereo panoramic vision sensors based on the principle shown in Figure 9. Image resolution invariant stereo images have the property of making stereo correspondence matching a relatively straightforward task. Figure 10 shows images from our image resolution invariant stereo panoramic sensor.

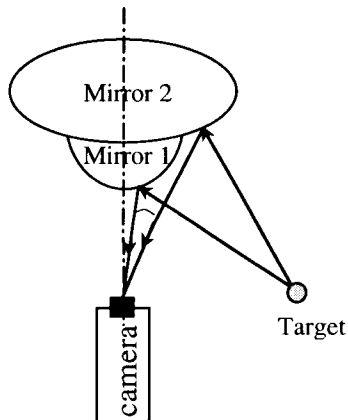


Figure 9. Stereo Panoramic Image Sensor

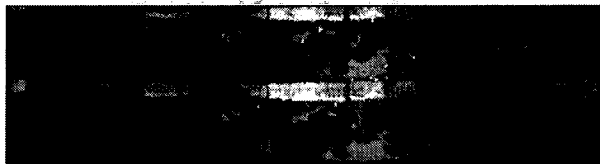


Figure 10. Stereo Panoramic Images

Stereo panoramic imaging using our image resolution invariant approach shows considerable promise. We are presently investigating building new sensors with an increased vertical offset to improve the range of the sensor's distance calculation. In future we plan to apply

the sensor to mobile robot navigation.

3. VISION ALGORITHMS IN HUMAN-MACHINE INTERACTION

We have implemented a visual interface that tracks a person's facial features in real time (30Hz) and estimates users gaze point. Our system does not require special illumination nor facial makeup. The system is user independent and insensitive to fluctuations in lighting conditions. We use the Hitachi IP5000 vision processing system to implement our visual interface. This vision processing system is designed to implement up to 10 image convolution operations, such as smoothing, erosion, template correlation in each frame of a colour NTSC video stream. The hardware consists of a single PCI-bus card, running on an Pentium processor running LINUX. We have created a stereo vision system by using the video interlace on a single channel to transmit alternating images from the left and right cameras. The processing steps we use in our visual interface are as follows:

1. *Skin Detection.* To find a person's face skin colour in the YUV colour space is used. Human skin maps to a small region in the hue(Y) and saturation(U) space. The skin detector binarises the image into skin and non-skin pixels.
2. *Face Detection.* The face detector assumes the largest area skin blob that is closest to the camera platform is a human face. The detector zeros all other pixels outside the face blob, and all the pixels inside the face blob are set to one. This creates a mask that is applied to the original image to segment the face.
3. *Feature Detection.* The feature detector uses the intensity information in the YUV colour space, to identify the eyes and mouth inside the segmented face. This results in a binarised image with holes in the image for the eyes and mouth.
4. *Selecting Tracking Features.* The feature tracking selector uses the binary image to find the corners of the eyes and mouth. The selected features are then masked against the segmented face image to generate image correlation templates.
5. *3D Facial Model.* The tracking features that were selected in one camera view are then matched against the corresponding features in the other camera. The stereo data is used to build a 3D model of the face, as well as a 3D feature tracking network, shown in Figure 11.

6. *Stereo Feature Tracking.* In each video frame the selected tracking features left and right camera images are matched by the vision processor to determine their 3D positions.
7. *3D Pose Estimation.* After feature tracking has completed an iterative fitting approach is used to estimate the motion of the 3D model from the previous frame to its new state in the current video frame. Spring-like connections between features in the 3D feature network are used to help estimate the 3D pose and incrementally estimate the gross motion. Features that are tracking well help estimate the positions of features that are tracking poorly. The best fitting 3D model is then back projected into 2D and used to update the search areas for the tracked features to be used in step 6.



Figure 11. Identifying Features for Face Tracking

Using this approach we are able to accurately predict and robustly track facial features. Our recent work has concentrated on estimating the gaze point direction of the user. Knowing where a person is looking is an important cue for human-robot interaction, and for building machines that can evaluate the way people perform tasks. The processing for determining the gaze point is as follows:

1. *Determine 2D Eye positions.* From the facial feature tracking data by determining the 2D position of the pupils of each eye, using the distances from the corners of the eye and the eyebrow.
2. *3D Eye Pose Estimation.* Use a rotating spherical model with the 2D eye position data to calculate the two angles of rotation α and β .
3. *Smooth the 3D Eye Pose Estimate.* Using the constraints the eyes are fixated on a single point in a single plane, the estimates of α and β can be smoothed and weighted by the eye which is tracking best.

4. *Estimate 3D Gaze point.* The user's gaze point is estimated by fusing the local 3D pose of the eyes with global 3D pose of the facial model.

Figure 12 shows results from our face tracking and pupil extraction scheme for gaze estimation system. Presently our system tracks with less than 3 degrees rotation error and less than 1mm in translation error. The gaze point can be estimated with an average error of less than 5 degrees. We are using our system to control the actions of a human friendly robot. A full description of this visual interface can be found in (Zelinsky et.al. 1999).



Figure 12. Pupil detection and Gaze Point Estimation

5. VISION ALGORITHMS IN MOBILE ROBOT NAVIGATION

The aim of our research is to construct an autonomous mobile robot that can navigate in dynamic and unknown indoor environments, without using explicit geometric models of the environment. For the robot to survive it must be capable of reacting in real-time to dynamic situations. This real-time constraint led to the development of the Behaviour-based approach pioneered by (Brooks 1986). This approach was inspired by studies of biological systems which showed that intelligent behaviours are achieved by only using simple mechanisms. Our laboratory has also adopted the behaviour-based approach to build vision-guided mobile robots {Cheng-and Zelinsky 1998} and multiple cooperating robots for cleaning (Jung-and Zelinsky:1999).

To operate in dynamic indoor environments a robot must be capable of performing a number of basic behaviours. These include:

- moving in free space,
- contour following,
- collision avoidance,
- goal seeking and
- visual servoing.

The robot must have the ability to move freely through the environment. This can be done by keeping the robot in areas that are predominately free space. This works reasonably well in environments that are sparsely populated with small-sized obstacles. Contour following

behaviour that causes the robot to follow the contour outline of an obstacle. Collision avoidance is a behaviour that halts the robot before a collision occurs and turns it way until the robot can move forward again. The robot must have the ability to seek and detect visual targets or goals such as landmarks. Goal seeking can act as cue to trigger other behaviours such as visual servoing. Visual servoing is a behaviour that moves the robot to a location in the environment identified by the vision system. The basic behaviours need not specifically consider the dynamic nature of the environment. If the robot can process vision data at camera frame rates then the dynamic aspects of the environment can be treated in the same manner as the static cases.

One of the great challenges in robotics is to provide robots with appropriate sensing capabilities that can supply sufficient information to allow a task to be achieved within a reasonable response time. By using efficient techniques, real-time performance can be achieved. Our vision processing methods are based on cross-correlation template matching. Template matching is a simple and effective method for image processing that allows us to achieve real-time performance. The free space detection process can be summarised as follows. Each image is segmented into a grid of 8 columns and 7 rows, producing 56 individual cells of 64x64 pixels in size. Each of these cells in the live video is correlated against a stored image of the floor taken at the beginning of each experiment. Using the matrix of correlation values the visual behaviours described earlier are implemented.

The Move in Free-Space behaviour causes the robot to keep to the areas of the environment that are free of obstacles. This is done by searching for finding the direction of the greatest free space region within the robot's field of view. The vision processor marks the free space cells of the robot's 8x7 retina. A search from the bottom of the grid looking for a region of free cells which the robot can pass through is performed. Once the region has been found the robot steers toward the highest free space cell. The highest cell represents free space which is furthest way. This behaviour guides the robot into open areas and favours moving in straight lines. Figure 13 shows a sequence of images taken from the vision processor during a navigation experiment.

6. ACKNOWLEDGEMENTS

The authors extend their thanks to the students and research colleagues at the Department of Systems Engineering who contributed to the systems described in this paper. Particularly Tanya Conroy, Gordon Cheng, David Jung, Jochen Heinzmann, Son Truong, Jon Kieffer, Yoshio Matsumoto and Rhys Newman.

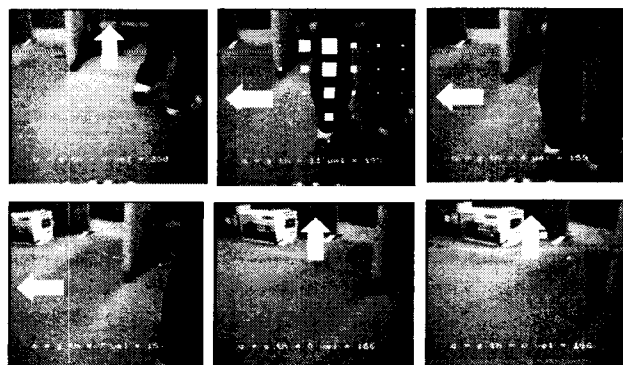


Figure 13 Obstacle Avoidance with Move Free-Space Behaviour

7. REFERENCES

- Blake, A. and Yuille A. (1992) *Active Vision*, MIT Press, ISBN 0-262-02351-2, Cambridge, Massachusetts.
- Brooks. R.A. (1986) A Robust Layered Control System for a Mobile Robot, *IEEE Journal of Robotics and Automation*, RA: Vol 2 No 1, pp14-23, April 1986.
- Conroy, T. and Moore, J.B. (1999) Resolution Invariant Surfaces for Panoramic Vision Systems, submitted to the 1999 International Conference on Computer Vision.
- Chahl, J.S. and Srinivasan, M.V., (1997), Reflective Surfaces for Panoramic Imaging, *Applied Optics*, Vol. 36, No. 31, pp 8275-8285, November 1997.
- Cheng, G. and Zelinsky, A. (1998) Goal-oriented Behaviour-based Visual Navigation, *Proceedings of IEEE International Conference on Robotics and Automation*, Leuven, Belgium, pp3431-3436, May 1998.
- Jung, D., and Zelinsky, A. (1999) An Architecture for Distributed Cooperative-Planning in a Behaviour-based Multi-robot System", *Robotics and Autonomous Systems*, Vol 26, No. 2-3, pp149-174, February 1999.
- Townsend, W.T. and Salisbury, J.K. (1993), Mechanical design for whole-arm manipulation, *Robots and Biological Systems: Toward a New Bionics*, pp 153-164, 1993.
- Truong, S.N., Kieffer, J. and Zelinsky A. (1999), A Cable-driven Pan-tilt Mechanism for Active Vision, *Proceedings of the Australian Conference on Robotics and Automation*, pp 172-177, Brisbane, Australia, March 1999.
- Zelinsky, A. Matsumoto, Y. Heinzmann, J. and Newman, R., (1999) Towards Human Friendly Robots: Vision-based Interfaces and Safe Mechanisms, *Proceedings of the International Symposium on Experimental Robotics*, pp 431-442, Sydney, Australia, March 1999.