# A Multimodal Approach to Real-Time Active Vision

Andrew Brooks†, Samer Abdallah‡, Alexander Zelinsky† and Jon Kieffer‡

† Robotic Systems Laboratory
Research School of Information Sciences and Engineering
The Australian National University
Canberra ACT 0200, Australia
email: {zoz,alex}@syseng.anu.edu.au

‡ Department of Engineering
Faculty of Engineering and Information Technology
The Australian National University
Canberra ACT 0200, Australia
email: {samer,jon}@faceng.anu.edu.au

**Abstract**

*The increase in hardware performance versus component cost has brought vision firmly into the realm of practical robotic sensors. We present an overview of an integrated research program to create a general-purpose robotic vision system from an expansive rather than linear perspective, incorporating a multimodal approach to real-time visual interaction with the environment. Design and construction of a high-performance active camera platform using primarily off-the-shelf components has proceeded in parallel with algorithm design and implementation towards a suite of low-level behaviours suitable for feeding quantised data to a multitude of visually guided tasks, from mobile robot navigation and visual servoing to human-robot interaction and instruction. We describe these initiatives, including performance specifications and experimental results obtained under real-world conditions.*

## 1 Introduction

This research is concerned with working towards the development of a general-purpose active visual sensor system for real world, real time robotic applications. It is intended that the system be fit for use with minimal modification on a diverse variety of platforms, and performing a range of tasks, so it must have multiple capabilities of basic functions. In the context of these tasks, the real world ultimately will include both indoor and outdoor environments, and real time involves a rate of change of the environment equal to that generated by the movement of humans and other robotic agents.

The development of this system has proceeded in the form of two parallel initiatives. A hardware project has concentrated on the design and construction of a new active camera platform, and concurrent software development has been concerned with algorithms and processing techniques for application on the physical apparatus. The hardware platform has been conceived with the purpose of having a high level of mechanical performance to allow rapid motion and short reaction times, as well as being highly modular and reconfigurable so that it can be readily applied to the various areas of operation envisaged for it [2]. The software strategy has focused on the development of a collection of behaviour primitives using pipelined real time image processing hardware.

## 2 Active Camera Platform

The fundamentals of our design have been heavily influenced by the mechanics of the human visual system. The human eye achieves extraordinary performance through its low weight and inertia and its use of muscles for actuation. Muscles are themselves lightweight, have high acceleration and do not suffer from the common problem encountered in active head design whereby each degree of freedom requires sufficiently powerful actuators to move all previous degrees of freedom including their actuators. This has previously led to large, heavy, over-engineered camera platforms [12] or systems with reduced degrees of freedom such as coupled vergence [6]. It has been shown that it is possible to build high performance heads using this traditional *serial* mechanism, but by incurring corresponding disadvantages in cost, complexity or weight [14, 7, 11]. We have found that by adopting a *parallel* mechanism, relocating the motors to a fixed base and thereby minimising the active component to a small, low-inertia 'eye', we can achieve acceptably high performance without such significant penalties. This is therefore the major guiding principle for our design, among several other important issues which have influenced our project, the list of which is as follows:

- Actuation via high performance cable drive.
- Light weight of moving and fixed components.
- Modularity of the visual apparatus.
- Colour image acquisition and processing.
- Easy reconfigurability of the platform.

### 2.1 Performance Specifications

The abilities of our platform expressed in the traditional performance metrics are given in Table 1. The angular resolution has been selected with the aim of allowing the platform to perform meaningfully small camera movements with a wide variety of lenses, including those having foveal fields of view. For example, a 6° fovea occupying a full RS-170 video frame has a resolution of 0.0117°/pixel in the horizontal axis and 0.0124°/pixel in the vertical axis. Thus the 0.01° angular resolution of our apparatus allows single pixel movements in both pan and tilt for even such narrow angle or zoom lenses.

| Specification | Pan | Tilt |
|---|---|---|
| Maximum velocity | $600^{\circ s^{-1}}$ | $600^{\circ s^{-1}}$ |
| Maximum acceleration | $72000^{\circ s^{-2}}$ | $72000^{\circ s^{-2}}$ |
| Angular resolution | 0.01° | 0.01° |
| Axis range | 120° | 180° |

Table 1: Standard performance specifications of the camera platform.

We propose four additional specifications for an active vision system which not only involve the speed and acceleration of the axes but also express the usage intention of the system in the form of a functional requirement (Table 2).

| | |
|---|---|
| Minimum time for 90° saccade in pan | 0.15s |
| Minimum time for 90° saccade in tilt | 0.15s |
| Maximum pan angle change achievable from stationary start to stationary stop within the time of one video frame | 15° |
| Maximum tilt angle change achievable from stationary start to stationary stop within the time of one video frame | 15° |

Table 2: Task-oriented performance specifications of the camera platform.

Our platform achieves these values in order to satisfy constraints related to its desired abilities with respect to its intended visual input, 30Hz RS-170 interlaced video. The minimum saccade time allows the camera to perform up to three 90° saccades, each preceded by four target location video frames and succeeded by one stabilization video frame, per second, while the maximum angular change within one video frame equates to the ability to predictively track an object moving past the cameras at up to four metres per second at a distance of one metre.

### 2.2 Drive System

For the actuation mechanism of our platform we have selected a cable drive scheme of a type that has shown much promise in robotic applications such as manipulator arms [16] and other active vision systems [3]. The cable drive principle involves the transfer of power from the motor to the final moving component via a series of cable-wrapped pinions and pulleys. The cables are fixed at both ends and rely on the wrapping friction to turn the intermediate drive train components.

The cable drive transmission offers many advantages over gear trains, resulting in several significant benefits for our design:

- No backlash is contributed to the mechanism.
- Cable drives are subject only to rolling friction.
- Freedom from problems with compliance.

The drive mechanism for our system consists of two separate cable drive trains, one for each of pan and tilt, coupled at one point near the camera (Figure 1). The coupling is in the form of a cable bevel gear inspired by the one presented in [15], and this makes it possible for both motors to remain stationary during movement about both axes. The pan motor is connected to a pinion which winds a cable onto a larger drum to provide speed reduction. This drum is connected to the vertical wheel of the bevel cable gear, which is a complete circle. The tilt motor also drives a pinion which winds a larger drum portion, onto which are attached the camera mount holders, which remain immobile relative to the tilt drum. On the camera mount itself, which is free moving, is the horizontal wheel of the bevel cable gear. This wheel is semicircular in order to provide a larger radius without adding excessive weight. As the system is coupled at the bevel system, pan motion is performed by driving the pan motor alone, and both motors are driven together to achieve an offset in tilt.

## 3 Processing Hardware

The image processing is performed by a MaxTD system from DataCube, consisting of one each of their MaxVideo 200 and DigiColor boards in a VME rack controlled by a MC68040 running the LynxOS realtime kernel (Figure 2). Motor control is provided by a single Motion Engineering PCX/DSP eight-axis programmable motion controller board. Although a single installation of our vision platform requires only two axes of motor control, this setup allows the control of up to four active eyes, or a binocular stereo head and a three degree-of-freedom neck, with the single card.

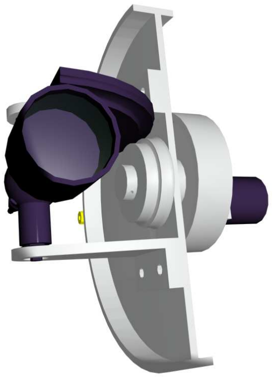


Figure 1: Front view of the camera platform showing the cable bevel gear.

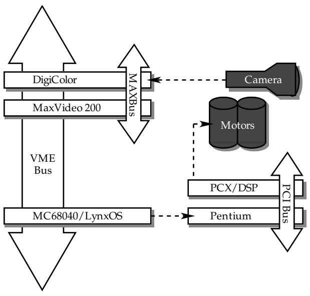



Figure 2: Processing architecture of the active vision system.

# 4 Visual Algorithms

We are of the opinion that a comprehensive collection of basic visual behaviours may be more suitable for investigation of an unknown, complex, unstructured environment such as the real world than a high-level processing technique specifically constructed for a certain situation. One aim of this research was the implementation of such a suite of elementary competencies. As low-level mechanisms, the algorithms should attempt to minimise assumed knowledge in the form of explicit world or object models and high-level decision making. However, they should be designed in such a way as to be bases, suitable for having these abstractions applied on top of them. They should be general-purpose in order to maximise their applicability to real world layouts and situations, and should be suitable for use either in series or in parallel. Finally, the algorithms should be designed with real time performance in mind, meaning a target which is close as possible to 30Hz video frame rate.

Three major categories of behaviour emerged as directions to be pursued. These are learning the correspondence between field of view (FOV) and motive ability, direction of attention, and maintenance of gaze fixation. The first two of these are dealt with in individual sections below. The latter, although a fundamental part of our project, has to date only been implemented as the zero-disparity filter (ZDF, see for example [4]) -based virtual horopter technique described in [13]. A fourth type, exploration or searching behaviour, was considered to be simple enough not to warrant isolated experimentation. As a low-level visual behaviour, exploration is merely the process of periodically changing the viewpoint and applying the other algorithms.

## 4.1 Attention Fixation

In order to gain its processing advantage in the unstructured, often cluttered, and dynamic real world, the active vision system must direct its attention to fixate on that which is important, simultaneously disregarding background irrelevancies, within the time frame of relevance of the meaningful subject. Attentional mechanisms are therefore necessary to identify from the input, as fast as possible, objects, agents and locations that are useful for the completion of goal-directed behaviour and that cause the environment to be dynamic by interacting with it. However, when viewed with foveated or wide angle lenses, objects appear vastly different depending upon their location within the FOV and object recognition becomes nebulous at best. Rather than object comparison at the initial attention level, we believe the active vision processor should focus on the detection of basic visual cues and direct its attention to likely areas by generating an appropriate saccade. At this point the target will be fixated in the optimum position for higher level processing as its location in the visual field is not only known but also the least distorted.

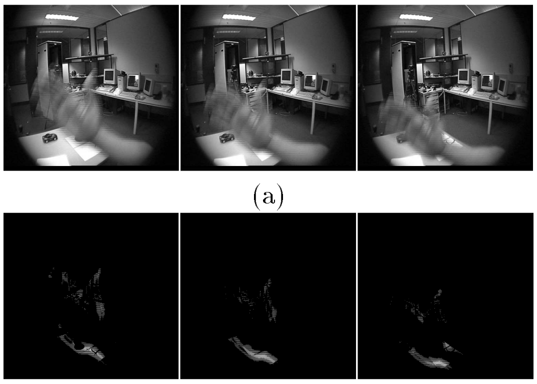
The visual cues driving attention concentrated on by the work presented here were signs that aided in the detection of other agents, such as humans or other robots, for monitoring and interaction purposes. These essentially consisted of *motion*, for separating agents from the background by their activity, and *colour*, for identifying humans based on the characteristic chromaticity of their skin.

### 4.1.1 Motion Segmentation

With a desire to exploit the strengths of the available processing hardware, a system based on frame differencing has been implemented in this research as the basic perceptual unit with respect to motion. As the mechanism is designed to be comfortable with fast-moving objects, it was necessary to be able to ignore object extension and duplication caused by both object translation and the resulting disocclusion appearing in the difference result, so an additional strategy for disregarding apparent motion caused by disocclusion was included based on frame history more distant than just the previous image. This algorithm is given in the equation

$$D_n = \left[T(F_n - F_{n-1}) \ \wedge \ \overline{T(D_{n-1})}\right] \times (F_n - F_{n-1}) \quad (1)$$

for frames $F$ and differences $D$ subject to some binary threshold $T$. The degree to which a moving object is detected by this method varies according to the object's speed – the faster the motion, the more of the object appears, beginning with its leading edge.



(a)

(b)

Figure 3: (a) A sequence of consecutive frames showing a moving object. (b) Gross motion images generated from the data in (a). Distinct areas of motion are identified and localised to be passed to higher behaviours.

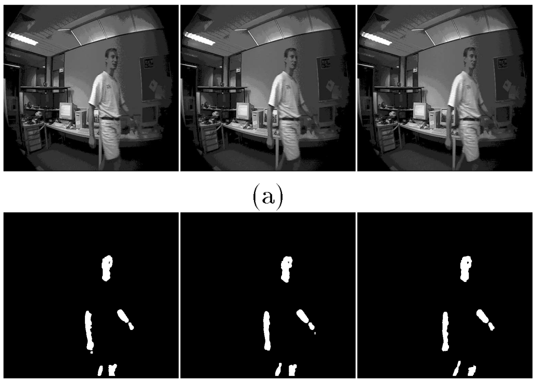
The motion segmentation system proved satisfactory for its intended task. An example motion sequence, and the corresponding raw output from the implementation, is shown in Figure 3. A parallel output accumulation technique consisting of grouping nearby positive regions effectively recognises the output as a single object and passes this quantised representation to other processes. In Figure 3(b) the disocclusion removal strategy can clearly be seen favouring the leading edges of the object, although this aspect could be even more effective if non-interlaced video were to be used, as portions of leading and trailing edges often occupy alternate interlaces. Overall, the motion segmentation system is able to reliably detect movement on a cluttered background. As this is a coarse, lowest level algorithm, future efforts, once appropriate hardware is procured, will concentrate on robust computation of optical flow to allow motion detection to continue when the platform itself is undergoing independent movement.

#### 4.1.2 Skin Detection

Human skin colour is made up of sufficiently distinctive chrominances that we believe it is an effective means of directing a robot's attention towards humans without requiring knowledge of human shape or other high-level model. Even variations in racial characteristics and illumination tend only to modify luminance, so if the colour space is normalised as in the equations

$$\begin{aligned} R_N(r,g,b) &= r/(r+g+b) \\ G_N(r,g,b) &= g/(r+g+b), \end{aligned} \quad (2)$$

then human skin colour in general becomes an identifiable constant. As the intensity normalised colour space sums to 1, only two channels of data need to be processed. Chrominances of source pixels are matched against known skin colour data modified by an appropriate error threshold to histogram image data into a binary stream indicating the presence or absence of skin at each pixel. This technique has previously been used as an underlying basis for high-level functions such as room monitoring and face tracking (see for example [1] and [17]).



(a)

(b)

Figure 4: (a) A sequence of frames of a moving human in a typical laboratory environment. (b) Classification of the images from (a) showing regions of human skin.

This processing technique (Figure 4) was used as the basis for a demonstration of the application of a higher-level, limited model-based abstraction over a low-level behaviour. An example system for possible human-robot interaction based on skin colour has been implemented, with the function of identifying and labeling separate human body parts based on quantised color region output from the low-level behaviour. The object list is limited to a maximum of five visible body parts, comprising the head, two arms and two legs. The head and arms are then identified solely by position heuristics, including discrimination between the left and right arm, bounded and labelled. The model is simple, but it has proved capable of robustly identifying and tracking the appendages of a single human, assumed to be facing the camera, at any location in the visual field and at 15 Hz. It is therefore considered a successful demonstration of the concept of low-level behaviours with an absence of abstract model information feeding parallel higher-level computational processes with the capability to perform useful and specific tasks.

### 4.2 Self-Calibration

A generalised modifier was desired for adjusting between image and motor co-ordinates, in order to cope with the pronounced deviation in scene-to-pixel representation observed in wide-angle lenses and the kinematic dependency exhibited by most pan-tilt devices. We prefer a generated mapping to a manually calculated exact model of the lenses and kinematics, as it is intended to use a variety of lenses on possibly varying mechanical platforms. We have therefore developed a method of reducing the error associated with movement from an initial to a target direction to an appropriately small tolerance on line, from visual feedback, in a self-calibration behaviour.

Our algorithm involves observation of the changing representation of the viewed scene under eye motion, in much the same conceptual way as the continuous adjustment process of biological systems. However, the time frame required for biological perceptual adaptation to converge is impractically long when considered in the context of a computer vision system, so we are content to assume that the optics and kinematics will not change during the operation of the active vision system, and to allow the self-calibration behaviour to execute independently during the initial phase instead of continuously in parallel with other behaviours. The behaviour has therefore been designed to complete rapidly, and to use real-world scene detail rather than artificial calibration tools such as measured grids. Our method operates by fixating on a target and then systematically moving the camera around while monitoring the corresponding movement of the target across the image. At each target image location the angular distances moved from the zero or "home" position are known and can be inserted directly into the coordination function without the need for any weighted averaging or repeat visits. A similar, but opposite in cause-effect inference and somewhat less efficient, method is presented in [9], a system which exhibits parallels in operation but prefers to extract formal camera parameters such as the focal length is given in [10], and a formal geometric treatment of active head kinematics can be seen in [8].

Target recognition has been implemented using the parameterised sum of absolute/squared differences (SAD/SSD) greyscale template matching algorithm introduced in [5]. The self-calibration operation proceeds in two stages, computing mappings for the lens distortion and platform kinematics separately, as the motion of the target within the image at the axis extremes is dependent on the lens distortion as well as the kinematics. Our system therefore only examines the lens vertically and horizontally from the home position, to discern the $x$ and $y$ lens distortion components in terms of pan and tilt movements, and generalises the observed results to the whole lens according to some regularity assumptions. This is due to the fact that certain problems, such as increased execution time, the compounding of difficulties related to matching in the darkened periphery, and the introduction of kinematic effects into the observed movement, make visiting every spot in the visual field undesirable.

All match locations are treated as polar co-ordinates in the construction of the distortion "map", and initially the axis skew is determined by averaging the deviation along each match track from the correct multiple of $\pi/2$ radians. The distances between consecutive matches are then analysed to plot the angle-to-pixel transfer function from the centre to the periphery of the lens. The match distances are first normalised to accommodate an unassumed rectangular pixel aspect ratio, and locations between successful matches are calculated by interpolation. Once this process has completed, the results are made available to all vision processes, allowing any movement-triggering behaviour to map a target's location anywhere in the FOV to the appropriate pan and tilt values of the camera platform in order to adjust it to directly fixate upon that point from the home position.

The second stage of self-calibration is to create a mapping for the kinematic effects on visual field movement, namely the proportions of translation and rotation that will be encountered under motion from any given position in the independent axis. A similar algorithm to that used in the lens distortion map creation procedure is used, but from each axis extreme rather than the home position, and correcting dynamically for the lens distortion effects on each match. The results from each extreme are extrapolated to determine the amount of dependency that can be expected at all axis positions. This method is applicable to the entire range of axis dependencies from completely independent to fully co-dependent.

The performance of the calibration routine naturally depended almost entirely on the speed and accuracy of the image correlation algorithm. The entire self-calibration procedure was able to complete within 7

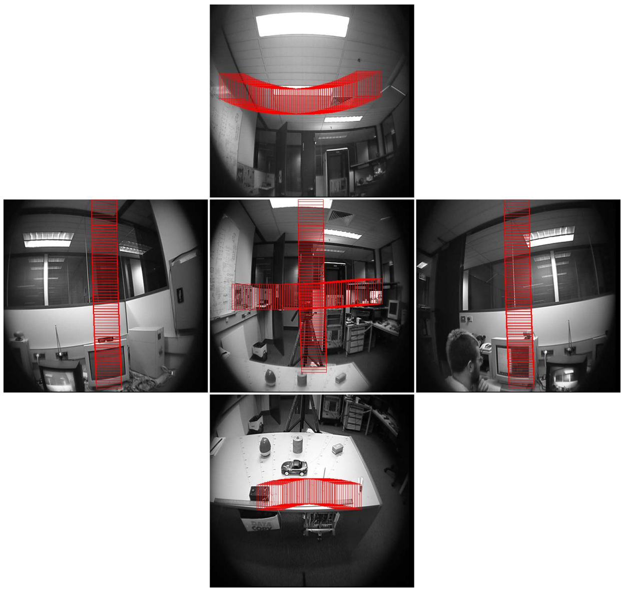


Figure 5: The results of self-calibration, showing match tracks for the lens distortion in the centre surrounded by those for the axis extremes.

minutes despite requiring over 1 second for each correlation. Figure 5 shows images of all five positioning stages of the calibration process for an example lens, along with the overlay tracks marking the location of each successful target match. Correct performance of the overall calibration scheme was verified by observing the system's ability after calibration to execute error-free saccades to motion using various combinations of lenses and mounting arrangements. The behaviour was able to cope with the significantly different properties of three test lenses (with FOVs of $33°$, $63°$ and $108°$), and with rotational mounting errors of up to $90°$ in either direction.

## 5 Conclusion

We have presented an integrated program of active vision research in which complementary hardware and software mechanisms have been developed with a wide range of areas and methods of operation in mind. A modular, flexible, high-performance hardware platform has been designed with major reliance upon standard, low-cost components. We have described our philosophy for vision computation based on low-level image processing behaviours, and have demonstrated implementations of these algorithms for application to principal areas of visual function such as direction of attention, gaze holding and visual adaptation. In addition, a higher level head and limb tracking process has been successfully implemented as an exemplar of the interacting multilevel behaviour scheme. Further research will concentrate on the expansion and refinement of the behavioural repertoire and its use in complex real-world tasks involving terrestrial and submersible mobile robots and force-controlled manipulators.

## References


[1] Appenzeller G., Lee J.-H., and Hashimoto H. Building topological maps by looking at people: An example of cooperation between intelligent spaces and robots. In *Proc. International Conference on Intelligent Robots and Systems*, Grenoble, France, November 1997.

[2] Brooks A., Dickins G., Zelinsky A., Kieffer J., and Abdallah S. A high-performance camera platform for real-time active vision. In *Proc. International Conference on Field and Service Robotics*, pages 559–564, Canberra, Australia, 8–10 December 1997.

[3] Brooks R. and Stein L. Building brains for bodies. *MIT AI Laboratory AI Memo No. 1439*, 1993.

[4] Brown C., Coombs D., and Soong J. Real-time smooth pursuit tracking. In Andrew Blake and Alan L. Yuille, editors, *Active Vision*, pages 123–136. MIT Press, 1993.

[5] Cheng G. *Vision-based Qualitative Mobile Robot Navigation.* PhD thesis, Research School of Information Sciences and Engineering, Australia, 1998.

[6] Clark J. and Ferrier N. Attentive visual servoing. In Andrew Blake and Alan L. Yuille, editors, *Active Vision*, pages 137–154. MIT Press, 1993.

[7] Kuniyoshi Y., Kita N., Rougeaux S., and Suehiro T. Active stereo vision system with foveated wide angle lenses. In *Proc. Asian Conference on Computer Vision*, Singapore, 1995.

[8] Li M., Betsis D., and Lavest J.-M. Kinematic calibration of the KTH head-eye system. Technical Report KTH/NA/P–94/07–SE, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Sweden, February 1994.

[9] Marjanovic M., Scassellati B., and Williamson M. Self-taught visually-guided pointing for a humanoid robot. In *Proc. 4th International Conference on Simulation of Adaptive Behaviour*, 1996.

[10] McLauchlan P. and Murray D. Active camera calibration for a head-eye platform using the variable state-dimension filter. *IEEE Transactions on PAMI*, 18(1):15–22, January 1996.

[11] Murray D., Du F., McLauchlan P., Reid I., Sharkey P., and Brady M. Design of stereo heads. In Andrew Blake and Alan L. Yuille, editors, *Active Vision*, pages 155–172. MIT Press, 1993.

[12] Pahlavan K. and Eklundh J.-O. A head-eye system – analysis and design. *Computer Vision, Graphics and Image Processing: Image Understanding*, 56(1):41–56, 1992.

[13] Rougeaux S., Kita N., Kuniyoshi Y., and Sakane S. Tracking a moving object with a stereo camera head. In *Proc. 11th Annual Conference of the Robotics Society of Japan*, pages 373–376, 1993.

[14] Sharkey P., Murray D., Vandevelde S., Reid I., and McLauchlan P. A modular head/eye platform for real-time reactive vision. *Mechatronics*, 3(4):517–535, 1993.

[15] Townsend W. The effect of transmission design on force-controlled manipulator performance. PhD thesis published as AI-TR-1054, MIT Artificial Intelligence Laboratory, April 1988.

[16] Townsend W. and Salisbury J.K. Mechanical design for whole-arm manipulation. *Robots and Biological Systems: Toward a New Bionics?*, pages 153–164, 1993.

[17] Yang J. and Waibel A. A real-time face tracker. In *Proc. 3rd Workshop on Applications of Computer Vision*, pages 142–147, Sarasota, Florida, USA, 1996. IEEE.